A Model for Addressing Quality Issues in Big Data

Grace Amina Onyeabor[1,2](✉) and Azman Ta'a[1]

[1] School of Computing, College of Arts and Sciences, Universiti Utara
Malaysia, 06010 Sintok, Kedah, Malaysia
grace_amina@ahsgs.uum.edu.my, azman@uum.edu.my
[2] Federal University of Technology, Minna, Nigeria
grace.onyeabor@futminna.edu.ng

Abstract. Big Data (BD) is everywhere and quite a lot of benefits have been derived from its usage by different organizations. Notwithstanding, there are still numerous technical and research challenges that must be tackled to comprehend and gain its full potential. The major challenges of BD are not just its processing, storage and analytics, there are also challenges associated with it that run across the BD value chain such as the data collection phase, integration and the enforcement of quality. This paper propose a DQ transformation model to evaluate BD quality from the data collection phase through to the visualization phase involving both data-driven and process-driven quality evaluation by assessing the quality of data itself first then assessing the process quality. This is still an ongoing research and hopefully will be experimented using specific Data Quality Dimensions (DQDs) like completeness, consistency, accuracy and timeliness with process quality dimensions such as Throughput, response time, latency with their corresponding metrics.

Keywords: Big Data · Data Quality · Data Quality Dimensions

1 Introduction

Going by the wide acceptance and usefulness of BD by several organizations, it comes with some challenges. These challenges are recognized in [1] namely: DQ, adequate characterization of data, interpreting correctly the results, visualization of data, real time data view versus retrospective view and determination of the importance of projects and results. The challenge that is fiddlier is DQ, which is, ensuring quality in the BD. Realizing high quality of data has always been a significant element of data management in business organizations due to the fact that it could help the organizations in articulating or enhancing their business strategies for better decision making. Organizations have suffered a lot from the effects of poor DQ such as wrong decision making, increase in production cost and the ability to satisfy their various customers [2]. Besides, with the vast amount of BD with unknown quality [3], the speed at which they are being exchange and in the various formats and structure they are created, DQ is far from being fetch. Business organizations such as online retailers are gradually taking advantage of data to increase discernibility into expenditures. The recent

F. Saeed et al. (Eds.): IRICT 2018, AISC 843, pp. 65–73, 2019.
https://doi.org/10.1007/978-3-319-99007-1_7

research works such as [4–8] proposed some kinds of good initiatives for the integration of DQ in BD, but, most of the initiatives did not evaluate DQ across the BD value chain (from data collection to the visualization phase). Only a few initiatives proposed all-inclusive solutions to guarantee BD quality and there's still requirements for that in organizations such as the online retail. It is vital to ensure the construction and implementation of DQ across the BD value chain. Therefore, this paper propose DQ model to implement quality of data across the BD value chain.

2 Big Data and Data Quality

According to [9], BD is a used to describe massive data sets that are of diverse format created at a very high speed, the management of which is near impossible by using traditional database management systems. According to [10], BD does not only concern the large volume of data but it also includes the ability to search, process, analyze and present meaningful information obtain from huge, varied and rapidly moving datasets. These three attributes lead to the foundational definition of BD regarding volume, variety, and velocity. Data are created from an extensive range of sources such as social media, the internet, databases, websites, sensors, and so on. But before these data are stored, processing and cleansing with the help of numerous analytical algorithms are performed on them [11]. However, because of the nature of BD, oftentimes, organizations encounter issues and challenges. These concerns and challenges need to be looked into to make proper business decisions prospectively. The researchers in [12] identified the challenges and the most risky of them all is DQ.

It has not been easy to define the concept of DQ [13]. Notwithstanding, the definition has been given by various researchers since the days of the traditional data. DQ is defined as fitness for use and being able to meet the set purpose by the user [14, 15]. The Authors on recent researches in DQ such as [4, 6, 16–18] have proposed some ideas relating to DQ. Some of them even delivered all-inclusive solutions guaranteeing quality across the BD value chain which is a good significant progress in this research area. Nonetheless, there's still room for improvement of DQ. According to [19], it was pointed out that improvement of DQ involves approaches of data-driven and process-driven. Data-driven handles the data the way it is, making use of methods and actions like cleansing to enhance the quality of the data while Process-driven strategy tries to detect originating poor DQ sources then redesigns the way the data is produced. DQ problems exist, right before the introduction of BD in the field. Because of these joint issues, the processes of BD cleaning and sifting are phases to be implemented before the analyses of data with quality that is unknown. [20] pointed out that DQ problems are more pronounced when dealing with data from multiple data sources. This problem obviously multiplies the data cleansing needs. Also, the huge amount of data sets that comes in at an unprecedented speed creates an overhead on the cleansing processes [17]. With the magnitude of data generated, the velocity at which the data arrives, and the huge variety of data, the quality of these data has left so much to be desired which cost organizations billion dollars yearly [21, 22].

3 Data Quality Requirements for Big Data

This section provides the description of DQ requirements for carrying out effective and efficient evaluation and assessment of DQ across BD value chain. These requirements include Data Quality Dimensions (DQDs) and their associated metrics. Data process quality and their metrics as introduced by [6].

3.1 Data Quality Dimensions

DQ is measured using multiple dimensions. A DQD is a characteristics or information part use for data requirements. DQD affords us the way to measure and manage DQ [6, 16, 23–25]. DQDs sometimes in the literature are referred to as characteristics, or attributes [26]. There are many DQDs in the literature but they are usually categorized into intrinsic and contextual. Intrinsic DQDs deals with the schema of the data and contextual deals with the values of the data [27, 28]. The commonly used DQDs for BD are the intrinsic DQDs which includes: Completeness DQD, Consistency DQD, Accuracy DQD and Timeliness DQD [29, 30].

In addition, for the evaluation of data process-driven quality, [7] introduced some DQDs which are associated with BD quality across the chain of BD. Some of these DQDs are namely: 1. Latency: which measures the delay in receiving the results of the processed data. 2. Response time: refereeing to the maximum time it takes to process the data records. 3. Throughput: refers to the number of processed records over a period of time. 4. Capacity: means the maximum number the processes run concurrently and to add the fifth one – Scalability: This has to do with the hardware- showing how much the hardware can scale the processing of the data.

3.2 Data Quality Metrics

DQ metrics measures specific properties of a DQD. This evaluates the extent of the presence of the DQD within the dataset. All DQDs are associated with one or more than one metric [7]. Table 1 shows different DQDs with their associated quality metrics.

Table 1. Data quality dimensions and metrics

Data quality dimension	Description	Metrics functions
Completeness	Null or missing values in the data set	$Comp = \left(\frac{Nnv}{N}\right)$
Consistency	Consistency in the format and structure of data values	$Cons = \left(\frac{Nvc}{N}\right)$
Accuracy	Error free data	$Acc = \left(\frac{Ncv}{N}\right)$
Timeliness	Currency and volatility of the data	TMa = (1 − CMa / VMb)

3.3 Metadata

Metadata is referred to as data about data [31]. It provides information that are pertinent to data like the quality or provenance of the data. Using metadata is a stress-free and faster way of data features extraction and process. The description of metadata is represented by making use of vocabularies which are of definite models and standard [32]. According to [7], Java Script Object Notation (JSON) [31], is a standard of metadata that is used to signify large dataset into a structure built property graph models.

4 Related Research Works

Studies have been carried out to address DQ assessment and evaluation on other organizational data. The researchers in [20] pointed out there's increment in DQ issues when dealing with data from numerous sources. Furthermore, the authors in [33] proposed a classification model and categorized the DQ issues during data preprocessing according to (i) errors correction, (ii) unstructured data to structured conversion and (iii) integrating data from various sources of data. With the inception of BD, [34] looked at data provenance as a pertinent data source to enable the evaluation of its quality. Data provenance refers to a concept being used for databases that are distributed with business and scientific data for the evaluation of the DQ. Tracing of data from its gathering through any transformation until the data visualization can be done by provenance data. The authors in [35] did something like that by proposing a multiple layer framework for data provenance collection. Another proposal is by [36] providing data cleaning tools for BD. The researchers called it NADEEF and this idea was extended in [30] handling the quality issues of data streaming. Furthermore, [37] proposed data semantics to guarantee the consistency of DQ dimension for BD. The replication of data with consistency below an efficient network bandwidth optimization was achieved. Literature also include extensive explanations on the identification and discussions on DQ issues and challenges [38, 39]. In fact, [38] proposed a more inclusive process for DQ assessment for the evaluation of BD quality after identifying the main challenges of DQ in BD. Nevertheless, [40] distinguished objective DQ assessment from subjective DQ assessment for identifying quality inconsistencies and suggest actions for enhancement.

None of the authors from the above research work conducted addressed BD quality across the BD value chain. Only a few studies in the literature attempted to do that. For example, [41] proposed a framework for the integration of different DQ areas and find out the assessment of DQ dimensions. Also, [42] proposed a cross layer approach the assessment of DQ. The approach was applied to trustworthiness evaluation of sensory data. Furthermore, in the efforts to provide an all-inclusive DQ assessment and evaluation, a recommendation system for repairing data was proposed by [43]. This scheme is to handle degradation in DQ. And the authors in [44] proposed a framework for the evaluation and management of social media DQ all through the BD value chain. The framework was extended and evaluated in [44] by implementing a reference architecture for DQ management in social media data. So far so good, there's still room for more research work in the literature to address product DQ issue of DQ in BD.

To address the above challenges relating with BD quality across the BD value chain. This paper is proposing a model for quality evaluation and assessment of BD from its collection until visualization enforcing DQ at all phases.

5 Proposed Research Model

Figure 1 below gives the description of the key processes constituting the BD transformation model for assessing DQ. The processes consists of about six phases namely: BD collection, the pre and post BD quality evaluation, BD processing and analytics evaluation and the finally the quality evaluation of the BD visualization phase. The phases will work together for the achievement of a comprehensive assessment of quality across the chain. The dataset that will be used for this study will be online retail product data from different heterogeneous sources online.

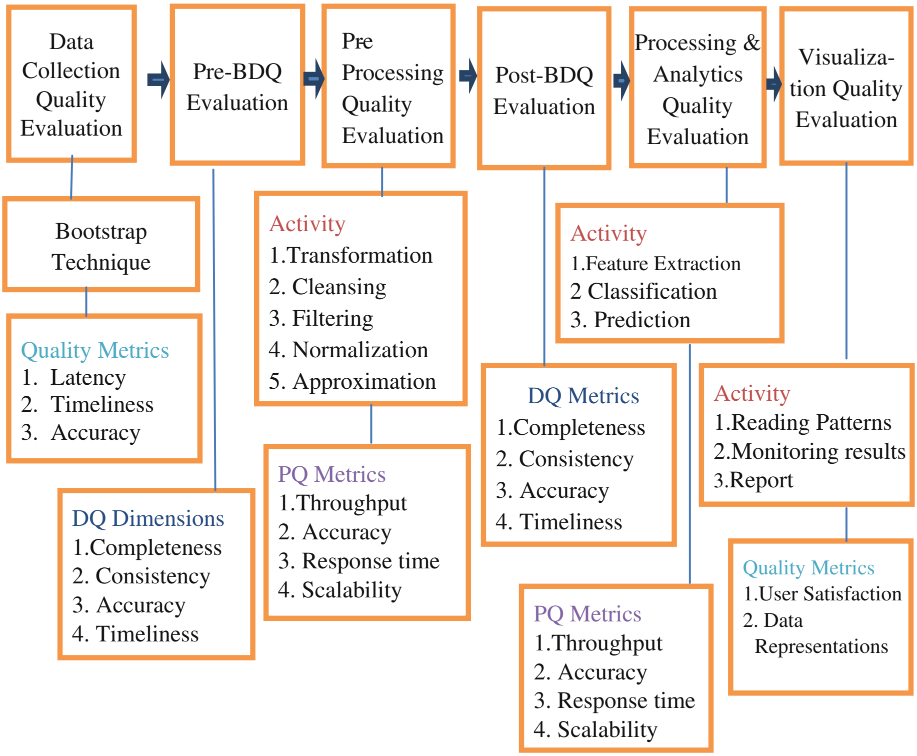

Fig. 1. A model for assessing quality issues in Big Data value chain

Phase 1: Data Collection

This is the phase where the data will be collected from the sources and stored on a temporary storage location. Due to the volume of the BD, Bootstrap sampling technique will be used to sample the data. The DQ metrics expected to be measured here are timeliness, latency and accuracy of the process of the data extraction and/or collection.

Phase 2: Pre-Big Data Quality Evaluation

This evaluation is going to be performed on the collected data itself (data-driven quality evaluation) before data preprocessing. The essence of this is to have a knowledge of the percentage of DQ to be able to choose the appropriate preprocessing activity that would be applicable. The DQDs that will be measured here will be completeness, consistency, accuracy using their metrics.

Phase 3: Data Pre-Processing Quality Evaluation

In this phase, the result of the previous phase will determine which of the preprocessing activity will be selected and performed on the dataset. Owing to the diversity of the data sources, the data might have been collected with diverse quality levels containing for example, redundancies or noise. Also, some data analysis applications and methods could involve requirements that are stringent or specific on DQ. Hopefully, in this phase, we could improve on throughput, accuracy and response time of the processes and the reliability and scalability of the systems.

Phase 4: Post-Big Data Quality Evaluation

The post DQ evaluation will be carried out after the preprocessing activities to prove that the percentages of the DQ has improved compared to the results obtained in the phase before the preprocessing. It is also going to be a data-driven quality evaluation measuring the same DQDs, that is, completeness, consistency, accuracy with their associated metrics.

Phase 5: Big Data Processing and Analytics Quality Evaluation

This phase involves applying machine learning procedures and methods of data mining leading to a required results of DQ. Here, the same preprocessing metrics (throughput, accuracy and response time) will be measured for the evaluation processing and analytics quality.

Phase 6: Big Data Visualization Quality Evaluation

To view and validate data result of data from the BD value chain is the process of visualization. The presentation of data will be done here using diverse sorts of views such as graphs, monitoring results summary, reading patterns. DQDs metrics such as quality of data representation, user satisfaction will be used for evaluating this phase.

6 Conclusion and Future Work

There is an urgent need for researcher in both industry and academia to give more attention to evaluation and assessment of BD quality because there are only few initiatives so far that addressed this vital aspect of BD. Thus, this paper proposed a model to tackle this problem. The model will handle the evaluation and assessment of

BD from the data collection phase all through to visualization phase using both data-driven and process driven quality. This is research in progress and the future work will be the implementation of the proposed model.

References

1. Tee, J.: The Server Side (2013). http://www.theserverside.com/feature/Handling-the-four-Vof-big-data-volume-velocity-varietyand-Veracity
2. Levitin, V., Redman, T.C.: Data as a resource: properties, implications, and prescriptions. Sloan Manag. Rev. **40**, 89–101 (1998)
3. Izham Jaya, M., Sidi, F., Ishak, I., Suriani Affendey, L.I.L.L.Y., Jabar, M.A.: A review of data quality research in achieving high data quality within organization. J. Theor. Appl. Inform. Technol. **95**(12), 2647–2657 (2017)
4. Hu, H., Wen, Y., Chua, T.-S., Li, X.: Toward scalable systems for Big Data analytics: a technology tutorial. IEEE Access **2**, 652–687 (2014)
5. Idi, F., Shariat Panahy, P. H., Affendey, L.S., Jabar, M.A.H., Ibrahim, H., Mustapha, A.: Data quality: a survey of data quality dimensions. In: 2012 International Conference on Information Retrieval Knowledge Management (CAMP), pp. 300–304 2(012)
6. Glowalla, P., Balazy, P., Basten, D., Sunyaev, A.: Process-driven data quality management-an application of the combined conceptual life cycle model. In: 47th Hawaii International Conference on System Sciences (HICSS), pp. 4700–4709 (2014)
7. Serhani, M.A., El Kassabi, H.T., Taleb, I., Nujum, A.: An hybrid approach to quality, evaluation across Big Data value chain. In: IEEE International Congress on Big Data (BigData Congress), pp. 418–425. IEEE. (2016)
8. Pääkkönen, P., Pakkala, D.: Reference architecture and classification of technologies, products and services for Big Data systems. Big Data Res. (2015). https://doi.org/10.1016/j.bdr.2015.01.001
9. Maier, M., Serebrenik, A., Vanderfeesten, I.T.P.: Towards a Big Data Reference Architecture. University of Eindhoven, Eindhoven (2013)
10. Malik, P.: Governing Big Data: Principles and Practices. IBM J. Res. Dev. **57**, 1–13 (2013)
11. Soares, S.: Big Data Governance: An Emerging Imperative. MC Press, Boise (2012)
12. Feldman M.: The Big Data challenge: intelligent tiered storage at scale. White Paper (2013)
13. Strong, Y.W., Lee, Y.E., Wang, R.Y.: Data quality in context. Commun. ACM **40**(5), 103–110 (1997)
14. Wang, R.Y.: A product perspective on total data quality management. Commun. ACM **41** (2), 58–65 (1998)
15. Fürber, C., Hepp, M.: Using SPARQL and SPIN for data quality management on the semantic web. In: International Conference on Business Information Systems pp. 35–46 Springer, Berlin, Heidelberg (2010)
16. Sidi, F., Shariat Panahy, P.H., Affendey, L.S., Jabar, M.A., Ibrahim, H., Mustapha, A.: Data quality: a survey of data quality dimensions. In: International Conference on Information Retrieval Knowledge Management (CAMP) (2012)
17. Taleb, I., Dssouli, R., Serhani, M.A.: Big Data pre-processing: a quality framework. In: 2015 IEEE International Congress on Big Data (BigData Congress), pp. 191–198. IEEE (2015)
18. Hazen, B.T., Boone, C.A., Ezell, J.D., Jones-Farmer, L.A.: Data quality for data science, predictive analytics, and Big Data in supply chain management: an introduction to the problem and suggestions for research and applications. Int. J. Prod. Econ. **154**, 72–80 (2014)

19. Loshin, D.: Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSql, and Graph. Elsevier, Amsterdam (2013)
20. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. IEEE Data Eng. Bull. **23**(4), 3–13 (2000)
21. Eckerson, W.W.: Data Quality and the Bottom Line: Achieving Business Success Through a Commitment to High-Quality Data. Data Warehousing Institute, Chatsworth (2002)
22. Fan, W., Geerts, F.: Foundations of Data Quality Management. Morgan & Claypool, San Rafael (2012)
23. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. **41**(3), 1–52 (2009)
24. McGilvray, D.: Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information. Morgan Kaufmann, Burlington (2008)
25. Jayawardene, V., Sadiq, S., Indulska, M.: An analysis of data quality dimensions, pp. 1–32 (2015)
26. Loshin, D.: The Practitioner's Guide to Data Quality Improvement. Elsevier, Amsterdam Morgan Kaufmann OMG Press (2011)
27. Batini, C., Cappiello, C., Francalanc, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. (CSUR) **41**(3), 16 (2009)
28. Taleb, I., Dssouli, R., Serhani, M.A.: Big Data pre-processing: a quality framework. In: IEE International Congress on Big Data (2015)
29. Saha, B., Srivastava, D.: Data quality: the other face of Big Data. In: IEEE 30th International Conference on Data Engineering (ICDE), pp. 1294–1297 (2014)
30. Tang, N.: Big Data cleaning. In: Chen, L., Jia, Y., Sellis, T., Liu, G. (eds.) Web Technologies and Applications, pp. 13–24. Springer, Berlin (2014)
31. Introducing JSON. http://www.json.org/
32. Understanding Metadata. NISO Press, Bethesda, MD, USA, (2004)
33. Oliveira, P., Rodrigues F., Henriques, P.R.: A formal definition of data quality problems. In: IQ (2005)
34. Glavic, B.: Big Data Provenance: Challenges and Implications for Benchmarking. In: Specifying Big Data Benchmarks, pp. 72–80 Springer, Berlin Heidelberg (2014)
35. Cheah, Y-W., Canon, R., Plale, B., Ramakrishnan, L.: Milieu: lightweight and configurable Big Data provenance for science. In: 2013 IEEE International Congress on Big Data (BigData Congress) pp. 46–53 (2013)
36. Ebaid, A., Elmagarmid, A., Ilyas, I.F., Ouzzani, M., Quiane-Ruiz, J.-A., Tang, N., Yin, S.: NADEEF: a generalized data cleaning system. Proc. VLDB Endow. **6**(12), 1218–1221 (2013)
37. Recuero, A.G., Esteves, S., Veiga, L.: Towards quality-of-service driven consistency for Big Data management. Int. J. Big Data Intell. **1**(1/2), 74 (2014)
38. Juddoo, S.: Overview of data quality challenges in the context of Big Data. In: International Conference on Computing, Communication and Security (ICCCS), pp. 1–9 (2015)
39. Rao, D., Gudivada, V.N., Raghavan, V.V.: Data quality issues in Big Data. In: IEEE International Conference on Big Data (Big Data) (2015)
40. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Commun. ACM **45**(4), 211–218 (2002)

41. Cheah, Y.-W., Canon, R., Plale, B., Ramakrishnan, L.: Milieu: lightweight and configurable Big Data
42. Monga, M., Sicari, S.: Assessing data quality by a cross-layer approach. In: IEEE International Conference on Ultra Modern Telecommunications & Workshops (ICUMT 2009) (2009)
43. Ding, X., Wang, H., Zhang, D., Li, J., Gao, H.: A fair data market system with data quality evaluation and repairing recommendation. In: Web Technologies and Applications, pp. 855–858 (2015)
44. Immonen, A., Pääkkönen, P., Ovaska, E.: Evaluating the quality of social media data in Big Data architecture. In: IEEE Access, vol. 3, pp. 2028–2043 (2015)