DATA QUALITY EVALUATION FRAMEWORK FOR BIG DATA

Article *in* i-manager's Journal on Cloud Computing · January 2018

DOI: 10.26634/jcc.5.2.15692

2 authors:

Grace Onyeabor
Federal University of Technology Minna
2 PUBLICATIONS **1** CITATION
SEE PROFILE

Azman Taa
Universiti Utara Malaysia
44 PUBLICATIONS **188** CITATIONS
SEE PROFILE

REVIEW PAPER

DATA QUALITY EVALUATION FRAMEWORK FOR BIG DATA

By

GRACE AMINA ONYEABOR * AZMAN TA'A **

* Lecturer, Department of Information and Media Technology, School of Information and Communication Technology (SICT), Federal University of Technology, Minna Nigeria.
** Senior Lecturer, School of Computing, Universiti Utara Malaysia (UUM), Sintok, Malaysia.

Date Received: 11/01/2019 Date Revised: 06/02/2019 Date Accepted: 18/03/2019

ABSTRACT

Data is an important asset in all business organizations of today. Thus the results of its poor quality can be very grievous leading to erroneous insights. Therefore, Data Quality (DQ) needs to be evaluated before the analysis of any Big Data (BD). The evaluation of DQ in BD is challenging. Given the enormous datasets that are of varied format fashioned at a rapid speed, it is impossible to use the traditional methods of evaluating DQ in BD. Rather, there is a requirement of strategies and devices for the assessment and evaluation of DQ in BD in a rapid and more efficient manner. However, assessing the quality of data on the whole BD can be very expensive. In addition, there is also a need for improvement in data transformation activities of BD. This paper proposes a framework for DQ evaluation with the application of data sampling technique on BD sets from different data sources reducing the size of the data to samples representing the population of the BD sets. The Bag of Little Bootstrap (BLB) sampling technique will be used. The target Data Quality Dimensions (DQDs) to be used in this paper are completeness, consistency, and accuracy. In addition, the DQDs will be measured using different metric functions relevant to the DQDs. This will be done before and after an improved data transformation techniques to check the improvement of DQ in BD.

Keywords: Big Data, Data Sampling, Data Transformation, Data Quality Evaluation.

INTRODUCTION

Data is considered today as an asset to various organizations. This is due to the fact that business strategic decisions are founded on the insights from the generated data. Data from wherever originally holds irregularities and discrepancies, such as dirty, incomplete, or missing data triggered by several elements like readings from sensory devices, data entry by humans just to mention a few. The term BD can be described as vast datasets with variety of formats generated at a fast speed, which is almost impossible to be managed using traditional or the classical database management systems. Nowadays, there are a lot of commercial businesses and organizations producing huge datasets in same way, where vast number of data is being acquired, shared, and stored in various sources (Maier, Serebrenik, & Vanderfeesten, 2013). Quality in itself is complex, multidimensional and with continuous process, referring to diverse facets ranging from service quality to software quality to DQ (Taleb, El Kassabi, Serhani, Dssouli, & Bouhaddioui, 2016). The authors went on to say that quality is related to domain, it is well-defined via a set of attributes and depends on dimension and methods of assessment. This shows that a good knowledge of the domain of the data and DQDs are some of the main requirements for a DQ evaluation. Consequently, DQ can be captured by making use of multiple measures and assessment tools for numerous domain activities. In BD context, a vital problem exist within the data itself and thus in its quality.

Going by the wide acceptance and usefulness of BD by several organizations, it comes with some challenges. These challenges are recognized in (Merino, Caballero, Rivas, Serrano, & Piattini, 2016) 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 fiddler 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 (Jaya, Sidi, Ishak, Affendey, & Jabar, 2017). Besides, with the vast amount of BD with unknown quality, the speed at which they are being exchange and in the various formats and structure they are created, DQ is far from being fetched. The recent research works, such as (Hu, Wen, Chua, & Li, 2014; Sidi et al., 2012; Glowalla, Balazy, Basten, & Sunyaev, 2014; Serhani, El Kassabi, Taleb, & Nujum, 2016; Pääkkönen & Jokitulppo, 2017) proposed some initiatives for the integration of DQ in BD, but the point of view of their initiatives are from different perspectives in the sense DQ is only improved by just data cleaning or evaluating DQ only at the point of data analysis instead of evaluating DQ right from the point of data collection. Furthermore, some of these initiatives are all inclusive solutions. In addition, it is vital to ensure the construction and implementation of DQ across the BD value chain. Different from the prior studies, this paper proposes DQ framework for a rapid DQ evaluation using a few DQDs important to online retail business data by the application of sample strategies on the BD set. Reducing the size of the data to a representative samples of the data size enabling a rapid DQ evaluation. Furthermore, the framework will entail some data transformational activities essential for BD. This will provide a well-constructed information of the DQ concerning the data attributes and the statistics in selected DQDs. The information offers a start in preparation for a BD analytics task targeting the dataset and paramount attributes and dimensions so that the evaluation of quality attains an adequate level of confidence.

1. Data Quality (DQ)

It has not been easy to define the concept of DQ (Fürber & Hepp, 2011). 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 (Strong, Lee, & Wang, 1997; Lee & Strong, 2003; Levitin & Redman, 1998; Wang, 1998; Sebastian-Coleman, 2012). This definition shows that DQ is extremely reliant on the context of the usage of the data and the interactions with the customer requirements, capability to use, and access the data (Jaya et al., 2017). Authors on recent researches in DQ, such as (Hu et al., 2014; Sidi et al., 2012; Glowalla et al., 2014; Serhani et al., 2016; Pääkkönen, & Jokitulppo, 2017; Taleb, Dssouli, & Serhani, 2015; Hazen, Boone, Ezell, & Jones-Farmer, 2014; Immonen, Pääkkönen & Ovaska, 2015) 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 significant progress in this research area. Nonetheless, there is still room for improvement of DQ especially in the area of online retail data where much has not yet been done. According to Loshin (2013), it was pointed out that two approaches are significant for the improvement of DQ. They are data-driven approach and process-driven approach. The first approach deals with the data the way the data is, using methods and transformation tasks such as cleansing for the DQ enhancement. While the Process-driven strategy makes effort to identify poor DQ sources and then redesigns the method of the data production. DQ problems have been in existence even before the introduction of BD era. The researchers in (Serhani et al., 2016) categorized the issues and challenges with DQ as errors correction, conversion of unstructured data to structured, and data integration from various data sources. In addition, there is existence of a number of specific BD issues, including enormous data volume generated by Web2.0 with unprecedented velocity, and with structures that are schema-less. Other BD quality issues are also identified related with BD features in (Wang & Strong, 1996; Juddoo, 2015; Cai & Zhu, 2015; Krogstie & Gao, 2015). BD cleansing and sifting are important phases to be applied on BD before the analyses of the data, especially when the quality is unknown due to the above mentioned issues. Also, (Rahm & Do, 2000) revealed that DQ hitches are often more noticeable

when handling data from multiple sources of data. This obviously multiplies the need of data cleansing. Needless to say, the vast amount of datasets pouring in at an unparalleled speed causes overhead during the cleansing process (Serhani et al., 2016). With the greatness of the generated data, the unprecedented speed with which the data moves, and the different data structure, the quality of BD has left so much questions unanswered. There has been an estimation of erroneous data costing the US business organizations the sum of $600 billion yearly (Eckerson, 2002) . Recorded error rate in business data is probably between 1% and 5%. A few business organizations' data error rate rose above 30% (Fan & Geerts, 2012; Fan, Geerts, Jia, & Kementsietsidis, 2008). Data cleaning amounts usually takes 30% to 80% of the developmental time including the budget for improving the DQ, which is in contradiction to building the system in most of the data warehouse developments. And concerning web data, about 58% of the files that are available are XML, out of which, only one-third of the XML documents with associated XSD/DTD are of good quality (Grijzenhout & Marx, 2013; Addo-Tenkorang & Helo, 2011).

2. Big Data (BD)

According to (Maier et al., 2013; Immonen et al., 2015, the term BD can be described as vast datasets with variety of formats generated at a fast speed, which is almost impossible to be managed using traditional or classical database management systems. There are nowadays, lots of commercial businesses and organizations producing huge datasets. In the same way, vast number of data is being acquired, shared, and stored in various sources. This is the era of BD which started to be recognized a few years back. The initial definition only gave the term a scanty meaning of what it actually represent, conveying an idea of a massive volume of data too huge for the management of the existing computer processor (Loshin, 2013; Chang, 2015).

However, according to Malik (2013) and Mahanti (2014), the idea of BD is beyond just the massive volume of data. It also comprises of the ability of BD analytics in searching, processing, analyzing, and presenting valuable information obtained from enormous, wide-ranging, and fast moving datasets. These three characteristics are the foundational definition of BD concerning volume, variety, and velocity popularly referred to as 3Vs. Furthermore, BD with huge volume, unprecedented velocity and different variety assets of information demands cost-effective ground-breaking methods of data processing for improved insights to aid decision making. Data are generated from a wide range of sources like internet websites, social media, databases, sensors just to mention a few. Nevertheless, prior to the storage the data, they go through some processing stages using several analytical algorithms (Soares, 2012). Still, due to the nature of BD, most times, business organizations end up with some concerns and challenges resulting in poor decision making. As mentioned earlier, the large volume, varieties of data formats, and with the rapid speed of the incoming data makes it puzzling to manage the data. The study in (Feldman, 2016; Chen, Mao, Zhang, & Leung, 2014) identified some of these challenges and concerns and the most risky of them is DQ.

3. Review of Related Literature

The evaluation of BD is described by several challenges that require to be addressed from different ways. The challenges are in the size of data, data generation speed, data attributes, and DQDs together with the metrics. There are only few research works that have been done in the area of BD quality (Serhani et al., 2016). More so, few research works have been on different point of view addressing quality from diverse perspectives. There are those that provided overall definitions for DQ (Juddoo, 2015), while some others gave the definition based on data domain and on different points of view (Sneed & Erdoes, 2015). Glowalla et al. (2014) mentioned that there is an agreement in the literature that DQ is related to data life cycle processes. For example, DQ is closely related to the phases of data generation and its origin. Furthermore, there are adapted and adopted traditional methods or strategies of approaches for the assessment of BD quality. Also the evaluation metrics are affected by the type of data that is going to be evaluated. It is usually based on the content, context or rating-based (Taleb et al., 2016). It is content based when the DQ indicators are the data

itself, context-based metrics comes from the metadata as quality indicators, while the metrics based on rating make use of clear ratings of the data and the data sources (Immonen et al., 2015).

DQ issues are grouped into error correction, conversion of unstructured, data and integration of data from heterogeneous sources. Other issues that were also deliberated upon in the literature concerning BD quality are the enormous volume of data, the unprecedented speed, and the schema-less structures of BD. In other words, the initial 3Vs foundational characteristics of BD are Volume, Velocity, and Variety. These research works also identified some problems of BD quality that are correlated to some features of BD (Immonen et al., 2015; Juddoo, 2015; Cai & Zhu, 2015; Krogstie & Gao, 2015).

Pipino, Lee, and Wang (2002) discussed DQ assessment earlier in the literature by categorizing DQ assessment into two major groups: the first is called subjective and the second objective. The authors further provided the approach of combining these two groups offering organizations with DQ metrics that are functional in evaluating their data. Unfortunately, this particular approach is not applicable to BD. Furthermore, a framework was proposed by (Immonen et al., 2015) for the evaluation and management of BD quality in social media domain across BD value chain. The limitation of this framework is that it was BD domain specific with limited attributes of quality and the researchers did not put data sources such as product data and market analysis and feedback into consideration. The quality metrics in the approach recommended by (Floridi, 2014) were based on the categorization of the domain for which the data is to be processed.

A comprehensive study was presented by Zhou et al. (2015) relating BD quality issues with computing infrastructure, such as code defects, defaults of hardware, human errors, configurations, and the solutions. But the authors (Han, Nie, Ghanem, & Guo, 2013) targeted the BD computations under limited resources. Their work also involved designing an algorithm for elastic mining for the approximation of quality results when varying time, cost, and the distribution of resources.

So far so good, the major issue of the BD quality has not been fully addressed. Especially, the literature have not been able to fully answer the questions of what to evaluate, how to evaluate, and the drive of the evaluation. Believing that the evaluation of BD quality should be tackled as early as possible, there should be mechanisms in place for good results which leads to particular tasks in improving DQ.

This paper is an ongoing research that proposes a framework for BD evaluation. It is an effort to collect vital insights concerning DQ attributes and profile. The resulting information is used to recommend some DQ rules, which should be taken into consideration for the BD evaluation during the preparation of the data for analysis. The DQ rules are mined from the results of evaluating the DQDs, which will also help in enhancing the BD sets by amending and removing attributes or the data that have the possibility of hurting the data analytics.

4. Framework for Big Data Quality Evaluation

Addressing the DQ before going into data analytics is the main aim of this framework. Achieving this is by the estimation of data attributes quality by the application of the metric of DQD to measure the quality. The DQ dimensions to be measured in this study are completeness, consistency, and accuracy. The result of these measurements is the DQ assessment suggesting the DQ constraints that will either worsen or improve the DQ. This evaluation is very important for the assurance of specific levels of DQ for the associated processes with an ideal cost. BD quality is vital for the fact that it is impossible to generate robust estimations of the cost of data analytics. The dataset to be used for this study is Amazon products details dataset.

Figure 1 illustrates the events in the BD quality evaluation framework, where the data will go through some sections to determine its quality. The main sections of the framework includes: Data extraction, sampling and profiling, selection of DQDs and metric, transformation and Quality evaluation of sampled data.

4.1 Data Extraction, Sampling, and Profiling

Online store product data of Amazon will be extracted

from multiple data sources after which data sampling and profiling follows. The authors (Gadepally et al.2015; Cormode & Duffield, 2014) discussed some strategies for sampling that are applicable to BD. They went on to evaluate the effects of methods of sampling on BD. They are of the view that sampling enormous data is effective

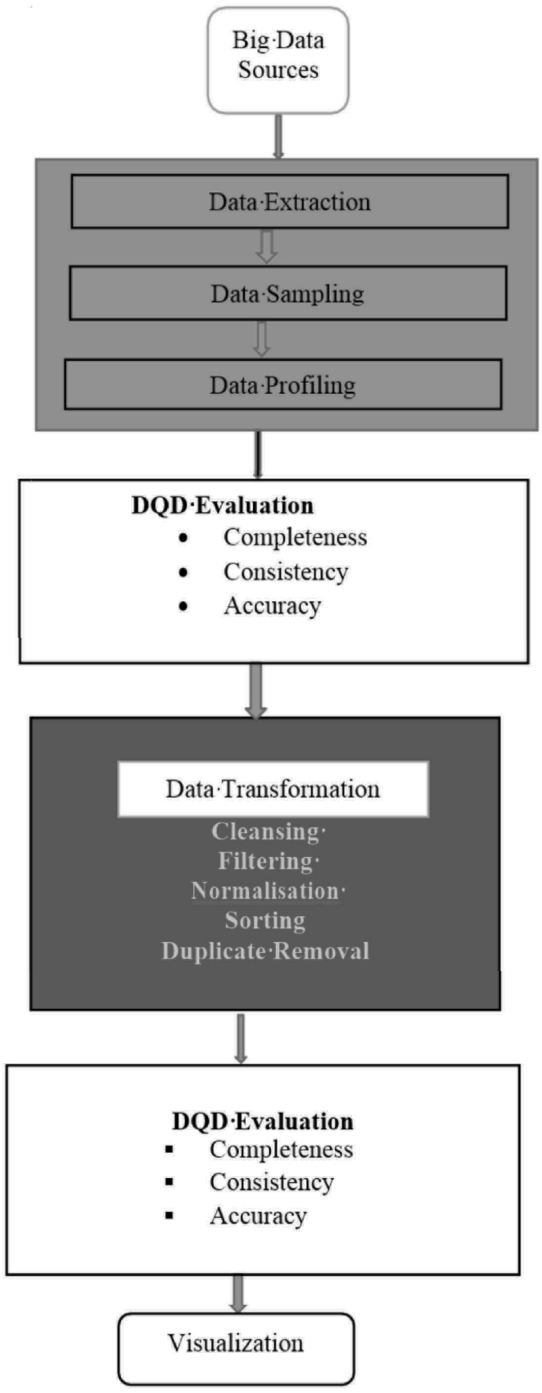

Figure 1. Framework for BD Quality Evaluation

in reducing run time and computational footmark of link prediction algorithms by the maintenance of adequate prediction performance. Bootstrap Sampling Technique does the evaluation of the sample distribution of the estimator. It carries out the sampling with replacement from the original sample. Liang, Kim, and Song (2016); Satyanarayana (2014); Kleiner, Talwalkar, Sarkar, and Jordan (2012) tackled the Bootstrap sampling technique in the perspective of BD. Bag of Little Bootstrap (BLB) (Kleiner et al., 2012) algorithm will be used in the framework of the DQ evaluation. It will work by combining results of bootstrapping several bits of subsets of the BD. The algorithm will use original BD set used for generating bits of samples without replacement. Then, additional set of samples will be produced by resampling with replacements for each sample that is generated.

The section of data profiling does DQ selection based on the summary of information and statistics to ascertain data features from the sources of data. This process is visualized as the assessment of data that provides DQ first summary information, such as various attributes with their types and values, the description of data format, range of data (maximum or minimum), and data constraints. The summary information will be extracted from the data itself devoid of any extra information such as making use of its metadata, which might be too costly for a BD project. Therefore, to cut cost accrue to the large volume of the data, BLB will also be used for the reduction of the data size to sample representing the population together with the addition of results of the profiling.

4.2 Data Transformation

Data transformation is a significant step in the BD quality framework. BD comes from numerous data sources that are heterogeneous in nature. Therefore, the dataset are inconsistent, inaccurate, and contains records that are incomplete records with most of them in the unstructured format. The data needs to be cleansed, filtered, normalized, aggregated, sorted, and the outliers need to be identified and general restructuring must take place for improvement in the quality.

An example of a usual problem with BD inconsistency is

date format problem. The dataset from different sources usually contain various types of date formats such as diverse data separators (slashes, dots, and hyphens) and diverse order of writing date such as: (day.month.year, month.day.year). Example of incomplete date format includes (day and month only) and so on and so forth. Therefore, data transformation activities such as cleansing, normalization, and other activities as earlier stated will be performed on the dataset.

4.3 Data Quality Evaluation

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. DQDs sometimes in the literature are referred to as characteristics, or attributes. There are many DQDs in the literature, but they are usually categorized into intrinsic and contextual. Intrinsic DQDs deal with the schema of the data and contextual deal with the values of the data. The commonly used DQDs for BD are the intrinsic DQDs which includes: 1. Completeness DQD (Null or missing values in the dataset), 2. Consistency DQD (Consistency in the format and structure of data values), and 3. Accuracy DQD (Error free dataset). DQ metrics measures specific properties of a DQD. This evaluates the extent of presence of the DQD within the dataset. All DQDs are associated with one or more than one metric. The equations in section 4.4 below shows the metric functions associated with each DQD.

Based on the previous information provided by profiling of the data, metrics functions will be used to evaluate each DQD. It should be noted that the metric functions would have been applied on the dataset before the transformation activities and the results of the DQ evaluation before and after the data transformation will be compared at the end to see the improvements in the DQ of the dataset. The researchers will come up with an algorithm to that effect. Furthermore, the result of the DQ evaluation will be visualized in the form of graphs in future.

4.4 Metric Functions

$$Acc_i = \sum_{i=1}^{n} \frac{Nav_i}{N_i} \quad (1)$$

$$Comp_i = \sum_{i=1}^{n} \frac{Nnv_i}{N_i} \quad (2)$$

$$Cons_i = \sum_{i=1}^{n} \frac{Ncv_i}{N_i} \quad (3)$$

where i = Number of Data Sources

Nav = Number of Accurate values

Nnv = Number of Null values

Ncv = Number of Consistent values

N = Total Number of the Sampled dataset

Conclusion and Future Work

This paper proposes Data Quality Evaluation Framework for Big Data towards the generation of actions for the improvement of DQ in BD. The framework is showcased in the paper and discussion on the stages of the framework from the data extraction, sampling and profiling through to the transformation stage, and the DQ evaluation to be carried out were given. This is an ongoing research. The future work entails the implementation of the framework.

References

[1]. Addo-Tenkorang, R., & Helo, P. (2011). Enterprise resource planning (ERP): A review literature report. In *Proceedings of the World Congress on Engineering and Computer Science* (Vol. 2, pp. 19-21).

[2]. Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal,* 14, 2. DOI: http://doi.org/10.5334/dsj-2015-002.

[3]. Chang, W. L. (2015). *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy* (No. Special Publication (NIST SP)-1500-4).

[4]. Chen, M., Mao, S., Zhang, Y., & Leung, V. C. (2014). *Big Data: Related Technologies, Challenges and Future Prospects.* Springer.

[5]. Cormode, G., & Duffield, N. (2014). Sampling for big data: A tutorial. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 1975). ACM.

[6]. Eckerson, W. W. (2002). Data quality and the bottom

line: Achieving business success through a commitment to high quality data. *The Data Warehousing Institute*. Retrieved from http://download.101com.com/pub/tdwi/Files/DQReport.pdf

[7]. Fan, W., & Geerts, F. (2012). *Foundations of Data Quality Management - Synthesis Lectures on Data Management*. Morgan & Claypool.

[8]. Fan, W., Geerts, F., Jia, X., & Kementsietsidis, A. (2008). Conditional functional dependencies for capturing data inconsistencies. *ACM Transactions on Database Systems (TODS)*, 33(2), 6:1-6:48.

[9]. Feldman, M. (2016). *The Big Data Challenge: Intelligent Tiered Storage at Scale - Actionable Market Intelligence for High Performance Computing* [White Paper]. Retrieved from https://www.cray.com/sites/default/files/resources/Integrated_Tiered_Storage_Whitepaper.pdf

[10]. Floridi, L. (2014). Big Data and information quality. In *The Philosophy of Information Quality* (pp. 303-315). Springer, Cham.

[11]. Fürber, C., & Hepp, M. (2011). Towards a vocabulary for data quality management in semantic web architectures. In *Proceedings of the 1st International Workshop on Linked Web Data Management* (pp. 1-8). ACM.

[12]. Gadepally, V., Herr, T., Johnson, L., Milechin, L., Milosavljevic, M., & Miller, B. A. (2015). Sampling operations on big data. In *Signals, Systems and Computers, 2015 49th Asilomar Conference on* (pp. 1515-1519). IEEE.

[13]. Glowalla, P., Balazy, P., Basten, D., & Sunyaev, A. (2014). Process-driven data quality management--An application of the combined conceptual life cycle model. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on* (pp. 4700-4709). IEEE.

[14]. Grijzenhout, S., & Marx, M. (2013). The quality of the XML web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 19, 59-68.

[15]. Han, R., Nie, L., Ghanem, M. M., & Guo, Y. (2013). Elastic algorithms for guaranteeing quality monotonicity in big data mining. In *Big Data, 2013 IEEE International Conference on* (pp. 45-50). IEEE.

[16]. Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). 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. *International Journal of Production Economics*, 154, 72-80.

[17]. Hu, H., Wen, Y., Chua, T. S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access*, 2, 652-687.

[18]. Immonen, A., Pääkkönen, P., & Ovaska, E. (2015). Evaluating the quality of social media data in big data architecture. *IEEE Access*, 3, 2028-2043.

[19]. Jaya, M. I., Sidi, F., Ishak, I., Affendey, L. S., & Jabar, M. A. (2017). A review of data quality research in achieving high data quality within organization. *Journal of Theoretical & Applied Information Technology*, 95(12), 2647-2657.

[20]. Juddoo, S. (2015). Overview of data quality challenges in the context of Big Data. In *Computing, Communication and Security (ICCCS), 2015 International Conference on* (pp. 1-9). IEEE.

[21]. Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. (2012). The big data bootstrap. *arXiv preprint arXiv:1206.6415*.

[22]. Krogstie, J., & Gao, S. (2015). A semiotic approach to investigate quality issues of open big data ecosystems. In *International Conference on Informatics and Semiotics in Organisations* (pp. 41-50). Springer, Cham.

[23]. Lee, Y. W., & Strong, D. M. (2003). Knowing-why about data processes and data quality. *Journal of Management Information Systems*, 20(3), 13-39.

[24]. Levitin, A. V., & Redman, T. C. (1998). Data as a resource: properties, implications, and prescriptions. *Sloan Management Review*, 40(1), 89-102.

[25]. Liang, F., Kim, J., & Song, Q. (2016). A bootstrap Metropolis–Hastings algorithm for Bayesian analysis of big data. *Technometrics*, 58(3), 304-318.

[26]. Loshin, D. (2013). *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques,*

NoSQL, and Graph. USA: Elsevier.

[27]. Mahanti, R. (2014). Critical success factors for implementing data profiling: The first step toward data quality. *Software Quality Professional,* 16(2), 13-26.

[28]. Maier, M., Serebrenik, A., & Vanderfeesten, I. T. P. (2013). *Towards a big data reference architecture* (Master's Thesis, University of Eindhoven).

[29]. Malik, P. (2013). Governing big data: Principles and practices. *IBM Journal of Research and Development,* 57(3/4), 1-13.

[30]. Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A data quality in use model for big data. *Future Generation Computer Systems,* 63, 123-130.

[31]. Pääkkönen, P., & Jokitulppo, J. (2017). Quality management architecture for social media data. *Journal of Big Data,* 4(1), 6. DOI: https://doi.org/10.1186/s40537-017-0066-7

[32]. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM,* 45(4), 211-218.

[33]. Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.,* 23(4), 3-13.

[34]. Satyanarayana, A. (2014). Intelligent sampling for big data using bootstrap sampling and Chebyshev inequality. In *Electrical and Computer Engineering (CCECE), 2014 IEEE 27th Canadian Conference on* (pp. 1-6). IEEE.

[35]. Sebastian-Coleman, L. (2012). *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*. USA: Newnes.

[36]. Serhani, M. A., El Kassabi, H. T., Taleb, I., & Nujum, A. (2016). An hybrid approach to quality evaluation across big data value chain. In *Big Data (BigData Congress), 2016 IEEE International Congress on* (pp. 418-425). IEEE.

[37]. Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data quality: A survey of data quality dimensions. In *Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on* (pp. 300-304). IEEE.

[38]. Sneed, H. M., & Erdoes, K. (2015). Testing big data (Assuring the quality of large databases). In *Software Testing, Verification and Validation Workshops (ICSTW), 2015 IEEE Eighth International Conference on* (pp. 1-6). IEEE.

[39]. Soares, S. (2012). *Big Data Governance: An Emerging Imperative*. MC Press.

[40]. Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM,* 40(5), 103-110.

[41]. Taleb, I., Dssouli, R., & Serhani, M. A. (2015). Big data pre-processing: A quality framework. In *Big Data (BigData Congress), 2015 IEEE International Congress on* (pp. 191-198). IEEE.

[42]. Taleb, I., El Kassabi, H. T., Serhani, M. A., Dssouli, R., & Bouhaddioui, C. (2016). Big data quality: A quality dimensions evaluation. In *Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ ATC/ ScalCom/ CBDCom/ IoP/ SmartWorld), 2016 Intl. IEEE Conferences* (pp. 759-765). IEEE.

[43]. Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM,* 41(2), 58-65.

[44]. Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems,* 12(4), 5-33.

[45]. Zhou, H., Lou, J. G., Zhang, H., Lin, H., Lin, H., & Qin, T. (2015). An empirical study on quality issues of production big data platform. In *Proceedings of the 37th International Conference on Software Engineering* (Vol. 2, 17-26). IEEE Press.

ABOUT THE AUTHORS

Grace Amina Onyeabor is working as a Lecturer. She is currently a PhD Research Student with Univeristi Utara Malaysia (UUM) in Malaysia. A member of Computer Professionals in Nigeria (CPN). She received her Bachelor of Technology degree in Physics/Computer Science from Federal University of Technology, Minna, Nigeria in 2002 and Master degree in Information Science from the University of Ibadan, Nigeria in 2006. Her main research interests include data quality, semantic web technologies, data engineering, knowledge management and big data.

Azman Ta'a is currently a senior lecturer supervising undergraduate and postgraduate (Masters and PhD) students in areas such as software engineering, data warehouse, business intelligence, knowledge management, ontology, and data analytics., and PhD in Software Engineering (Data Warehouse) in 2012 from UUM. He received his Bachelor of Computer Science in 1988 from UniversitiTeknologi Malaysia (UTM), Master of Information Technology in 1999 from Universiti Utara Malaysia (UUM). He is a member of the IEEE, and International Association Software Architecture (IASA). He is also a Certified Professional Trainer from Malaysian Institute of Management (MIM), Certified Data Science Specialization from Johns Hopkins University, and Certified Professional Requirement Engineering Foundation Level from Malaysian Software Testing Board (MSTB). His research interests focus on business intelligence, data warehouse, knowledge management, ontology, and data analytics.