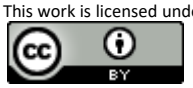
Available online at: https://ijact.in

Date of Submission	19/05/2020
Date of Acceptance	28/07/2020
Date of Publication	31/08/2020
Page numbers	3785-3790 (6 Pages)

This work is licensed under Creative Commons Attribution 4.0 International License.



ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

BIG DATA QUALITY FACTORS, FRAMEWORKS AND CHALLENGES

[1]Mohammad Abdallah, [2]Mohammad Muhairat, [3]Ahmad Althunibat, [4]Ayman Abdalla
[1,2]Assistant Professor, [3,4]Associate Professor
[1,2,3]Software Engineering Department, [4]Computer Science Department
Al-Zaytoonah University of Jordan

Abstract: Big Data applications are widely used in many fields such as artificial intelligence, marketing, commercial applications and health care, as demonstrated by the role of Big Data in coping with the COVID-19 pandemic. Therefore, it is essential to ensure the quality of the generation and use of Big Data applications. Consequently, Big Data applications must satisfy quality factors suited for these applications. Furthermore, quality frameworks need to be applied and tested for the quality factors of Big Data applications. Nevertheless, the quality measurement process needs to overcome some challenges for it to become applicable and trustworthy. This research lists different quality factors and dimensions and describes quality frameworks that are commonly used to measure the quality of Big Data. Furthermore, it lists the frequent challenges that researchers and data scientists face throughout the Big Data quality measurement process. Finally, it outlines the solutions that need to be developed for confronting the challenges of Big Data quality.

Keywords: Big Data, Quality Dimension, Quality Factors, Quality Frameworks, Quality Challenges

I. INTRODUCTION

The ideas of Big Data started to become commonly known several decades ago. In 1944, scientists began to realize the explosion of information [1, 2], but the term "Big Data" in information technology was first used in 1980 by Charles Telly. Then, in 2005, the term Big Data was described by Tim O'Reilly as the "huge amount of data," and this term entered the Oxford Dictionary in 2013. Nowadays, this term is used in defining modern concepts, applications, technologies, instruments, and performance measurements [3, 4].

One study estimated the size of Big Data to reach 40 zettabytes (40 trillion gigabytes) by the end of 2020 [5]. However, the actual number will probably exceed this prediction due to the heavy data exchange over the internet during the COVID-19 pandemic.

Big Data is a term often used to describe the large information packages accessed or processed frequently by the users [5]. Dumbill [6] described Big Data in terms of database capacity; i.e., Big Data is identified as it requires storage size, processing speed, or architecture that exceeds the limits of a conventional database system. Consequently, it requires unconventional solutions for its processing.

The above descriptions of Big Data have lead to the classifications of Big Data characteristics, also known as the V's of Big Data. The first three V's of Big Data were introduced by Laney [7] in 2001 as Volume, Velocity, and Variety. Since then, the V's of Big Data have been increasing and they reached 51 V's in 2019 [8].

When a business deals with Big Data, it needs to determine how the data have been collected, understood, processed, cleaned, analyzed, visualized, and utilized. This emphasizes the importance of how the quality of Big Data is going to be measured and on which factors this measurement will be based. These quality factors are referred to as the Big Data characteristics or the V's of Big Data [9].

The utilization of big data has increased significantly. For example, in Jordan [10], the government has developed and published more than 20 mobile and website applications to facilitate people's urgent needs while coping with the COVID-19 pandemic. This website and its applications cover necessities and lively sectors such as food supply,

health, education, constructions, and labor. These applications can provide a valuable source of data for these important sectors.

Big Data research and applications take numerous directions. They do not only consider computerized applications, but they also have an impact on society and people's daily activities [11]. Therefore, it has become essential to ensure that the data have been collected, processed, and utilized correctly with precision.

Overall, Big Data users, producers, visualizers, and analysts have to measure the quality of the data to guarantee that the data serve their intended purposes [12]. Therefore, there is a demand for quality factors to be used in the measurement of Big Data and a need for quality frameworks for testing and measuring the quality of Big Data. However, Big Data testing and quality measurements face many challenges.

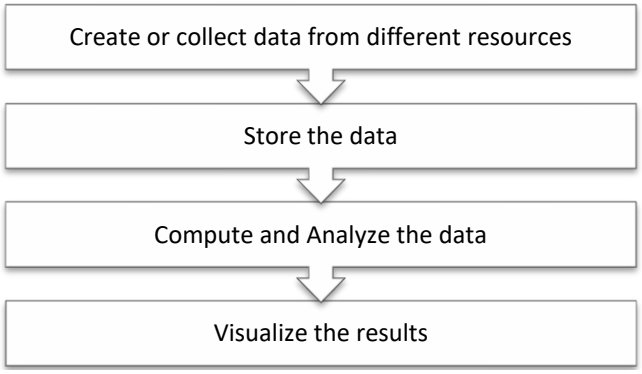


FIGURE 1. Big Data Life Cycle [13]

As seen in Figure 1 [13], the Big Data life cycle consists of four main steps; collect or create, store, analyze, and then visualize the data. The measures of these life-cycle steps of Big Data are the main categories discussed in this paper. They will be used to identify the challenges of Big Data quality, quality factors, and quality frameworks while discussing the consistency metrics that can be used to assess the Big Data systems across the entire process. Furthermore, this paper will highlight the methods used to assess the consistency of Big Data systems as it tackles the key challenges at each stage.

This paper is organized as follows. Section 2 describes the quality factors used in evaluating the quality of Big Data applications. Section 3 describes the commonly used frameworks used in Big Data quality measurements. Section 4 describes the challenges that face building, measuring, and applying the Big Data quality frameworks. Finally, Section 5 presents conclusions and future directions.

II. BIG DATA QUALITY FACTORS

Data quality defines the standards of data that are fit for use by data users [14]. To decide how the data quality factors can be extended to Big Data, it is necessary to review the common measurements of data quality [15]. Data quality is considered as a measurable concept describing the level of a set of qualitative and quantitative dimensions, factors, and metrics describing this quality [16]. Initially, the data must come from a source, and this source is the best place to start the data quality investigation. To ensure the authenticity of the data resource, some questions need to be asked such as:

- Who creates the data?
- How were the data created?
- How can the data be useful for the applications?

These three questions provide a start for the data quality measurements of any kind of data including Big Data. To answer these questions and further questions related to the data store, process, analysis, and visualization, data quality dimensions must be considered.

The data quality dimensions and characteristics are the concepts used by data processing experts to characterize a data attribute that can be calculated or analyzed against established evaluation criteria, and that determines the reliability of data for making decisions. These dimensions are as follows [17-21]:

- Completeness: The values of all single data components can be completed accurately. It determines that the appropriate details are rendered accessible in the details platform to achieve current and potential market objectives.
- Uniqueness: The data are recorded only once.
- Timeliness: Determine the time-period required for data acquisition, analysis, and usage to accomplish a defined business purpose. Failure to obtain the information in an acceptable timeframe can impair its usefulness for making decisions.
- Validity: The data validity is determined based on rules for the format, type, and range.
- Accuracy: Determine how the data fits the dataset. This metric is a test of how well the data represent the subject of the "real world," entity, or case.
- Consistency: There should be no contrast between the descriptions of two or more depictions of any subject.

In addition to the previous data quality dimensions, which are the most common, other dimensions include Volume, Believability, Ease of manipulation, Free-of-error, Interpretability, Objectivity, Relevance, Reputation, Security, Understandability, and Value-added [16].

In the context of data analytics, most of the quality dimensions are identified to coincide with the traditional dimensions of data quality in the field of database management. In the context of data quality, on the other hand, metrics are traditionally required for each dimension.

The above dimensions are the general quality characteristics of the data. However, the quality of Big Data requires additional characteristics to be measured. Therefore, the V's of Big Data have been derived mainly from data quality dimensions and some other V's were added as well.

The first three V's of Big Data were Volume, Velocity, and Variety. Volume is related to the size of Big Data and should have enormous size. Velocity gives an indication of the data generation speed, which must be relatively fast. Variety demands the data to be of different types that can be structured, semi-structured, and unstructured.

More recently, researchers have added more V's to Big Data such as Veracity, which considers the data bias or noise, and Value, which indicates the data usefulness [22, 23]. Research by [21] provided an alternative definition of

Validity as the correctness and accuracy of data concerning the intended usage, and they defined Volatility as the ease of recalling the retention policy of structured data implemented in the businesses on daily basis [24]. In another research[25], three more V's have been added; Viscosity, Variability and Viability. Viscosity discusses data complexity. Variability measures the change and inconsistency of data flow over time. Viability indicates the ability of Big Data to remain live and active perpetually.

The V's kept on increasing and reached 51 V's in 2019 [8] as researchers tried to do comprehensive overview studies of the Big Data domain characteristics. However, these characteristics and quality factors remain unusable unless they are employed in a quality model, i.e., a framework to measure the Big Data and its applications. Some quality frameworks will be listed and discussed in the next section.

III. BIG DATA QUALITY FRAMEWORKS

The success of a Big Data project depends on its effective utilization of the organizational, technological, and analytical aspects [26]. When using Big Data and functions on Big Data, the accuracy of the data needs to satisfy the required criteria and to confirm the application uses. Therefore, it is essential to introduce quality frameworks to measure the quality of Big Data.

In [27, 28], the researchers defined the dimensions of data quality accepted as Big Data quality standards and they redefined their concepts to be convenient for the business needs. Each standard they presented includes multiple elements linked to it where each element has its indicators of quality. Moreover, they introduced a framework consisting of Big Data quality dimensions, quality characteristics, and quality indexes, which were mentioned earlier in this paper, and then they produced a dynamic assessment using this framework.

Catarci et al. [29] discussed the problem of data consistency in policies by demonstrating how Big Data consistency is managed in the multiple phases of transmission systems and how Big Data sources are combined with conventional sources. They introduced the Big Data processing pipeline, which shows the process that Big Data goes through, starting from recording until the end, while interpreting the data. The processing pipeline goes along with the Big Data quality pipeline where each step in processing goes through a step in the quality pipeline.

The researchers in[21] focused on the impact of Big Data quality on the business process. They introduced an eight-step methodology that maintains Big Data quality while concentrating on the team and the business process.

Taleb et al. [28] suggested a comprehensive management quality paradigm that identifies essential facets of quality and examines how Big Data quality dealt across the whole development cycle. They also identified processes for handling and addressing data quality issues and provided a quality evaluation scheme to ensure its efficient management. They concluded that the most important steps, in order, are: (1) data creation, (2) data source (3) data collection, (4) data transport, (5) data storage, (6) data preprocessing, (7) processing and analytics, and (8) visualization.

In [30], a warning system was proposed. They suggested a consistency model that would enable the user to become aware of data quality problems during the review process.

Ridzuan and Zainon[31] tried to use a data cleansing method to improve the quality of Big Data. The method of data cleansing is primarily focused on finding and removing errors. Even though the data could be quickly analyzed, the data cleaning process remains complicated and time-consuming as the method attempts to ensure better data quality. Since clarification and evaluation are the key features of the cleaned results, it is essential to employ the domain expert on the results of the cleaning method [32].

Other frameworks tried to assess, manage, and maintain data quality. The twelve general-purpose data quality frameworks were discussed and compared in detail in [33] where the comparison included data quality definitions, assessment, and improvement processes. These additional data quality frameworks are used for data assessment and management in general, but they can be used for Big Data application as well.Figure 2 lists these twelve frameworks and illustrates their timeline.

As seen in this section, the data quality frameworks are strongly related to the Big Data application types and purposes. Therefore, the quality factors and their weights in the quality measurement may vary between applications.

FIGURE 2. Additional Data Quality Frameworks and Their Timeline

IV. BIG DATA QUALITY CHALLENGES

Many different factorsmay causethe generation of low-quality data in general such as human errors, invalid information, machine/deviceerroneous processing, unstructured results, and missing values[18].

Big Data applications face some testing challengesthat do not arise in other applications. Numerous functional and non-functional quality attributes need to be considered, such as performance while dealing with huge data, scalability

that fits with the volume of Big Data, and the accuracy of data analysis obtained by applying the proper algorithms. Moreover, data currency, backup, and recovery capabilities have to be checked to ensure good data quality[34].

In addition, Big Data application systems consist of complex software paired with advanced hardware preparation. Consequently, they require a combined quality model that measures both software and hardware [35].

One of the main characteristics of Big Data is Volume where the volume of collected data exceeds the possibilities of the system's vertical growth. Alternatively, the system should expand horizontally with more servers for dealing with data collection tasks [36]. Time is another challenge, not only in collecting data but also in processing and utilizing it. Time is directly related to Velocity or the speed of data generation. Data may be imported in two ways: batch data, in which the dataset loads all data at the same time, and stream data, which imports and processes data flows continuously as they are generated. Stream processing is the basis for choosing the Big Data analytics tool since the most common time-sensitive real-time method demands a faster and more accurate analytic performance[25, 37].

Defective data are inaccurate, poor, or infinite information that may be unreliable, incomplete, uncertain, latent, false, or approximated data, where the overall data value may be affected negatively by this defective data [38, 39]. Moreover, Data loads become challenging to manage, particularly when the usage of social media increases, which typically makes the data charge rise for certain events[40, 41]. For example, the data may not be accurate in the COVID-19 crisis because the data extracted from social media will be loaded with many COVID-19 related posts or fake news about it. Other information, such as the personal posts of social media users, will be immersed in the pandemic-related materials.

According to Abdullah et al. [21], three steps need to be applied to create valuable Big Data. First, start with the right Big Datastore and the technology that can fit the business problem or opportunity. Then, add deep knowledge, which is the human intelligence that accumulates within a certain practice or process. Finally, apply the right analysis and reporting tool that yields the maximum benefits from the data.

The previously addressed challenges have several impacts on organizations and businesses[42, 43]. The main impacts from the organizational point-of-view are as follows[42, 44]:

- Operational view, which discusses the employee and customer dissatisfaction about the data or its use. It also considers the increased cost of operation such as resources, time, and tools used to rectify data errors.
- Tactical view, which is concerned with the negative effects that defective data have such as poor decision-making, increased difficulty of reengineering, and lack of trust among members of the organization.
- Strategic view, which is concerned with how the defective data creates more difficulties in defining and executing organizational strategies and data proprietorship. The effects may also divert management attention away from key factors such as customers and competition.

The effects of data quality from a business perspective are mainly as follows[45, 46]:

- Impact of confidence and satisfaction, which is to the degree of satisfaction of stakeholders, such as clients, employees, and suppliers. Lower confidence or satisfaction may result in lowering organizational trust, cause incoherent management and operational reporting, and yield inappropriate decisions.
- Productivity, which is concerned with increased processing time and decreased production. This may cause higher workloads and ultimately hinder product delivery.
- Risk and compliance, which indicate the potential hazards that Big Data may cause, and the violation of policies and regulations of the government, the industry, or the business establishment.
- Finance, which detects lower profit, reduced cash flow, and increased costs of operation. This may increase penalties and waste opportunities.

In addition to the previous challenges, Kolajo et al. [47] discussed the evaluation of data scalability, integration, fault-tolerance, heterogeneity, completeness, load balancing, high throughput, privacy, and ethical aspects.

Overall, these challenges are not the only challenges that face Big Data applications; however, they are the most common. Therefore, Big Data application developers and users need to have quality factors and frameworks to measure their applications. As a result, quality assurance companies and organizations have customized, modified, and adapted the existing quality factors of data to suit Big Data.

V. CONCLUSIONS AND FUTURE DIRECTIONS

This paper has focused on the quality of Big Data applications while describing the similarities and contrasts with common data quality measurements. With Big Data, data collection, cleansing, management, and visualization are quite different from their counterparts in typical databases. Due to its nature, Big Data requires more work and imposes more conditions to become useful for its users.

Consequently, it is recommended for future research on Big Data quality to be aimed at developing the following solutions:

- The measurement of Big Data quality must start as early as possible in the Big Data life cycle; i.e., in the data collection stage, so that some extra conditions can be introduced to the data being gathered.
- Big Data storage should follow some specialized protocols that ensure easier and faster storage, restoration, and retrieval.
- Specialized Big Data quality metrics need to be defined to meet the complex nature of Big Data and its unconventional nature.
- Big Data qualitycontrol, continuous quality enhancement, and compliance frameworks need to be developed.

- There is a need for developing new measurements for Big Data consistency with unique unstructured data measuring characteristics and with fewer data schemes.
- Effective compliance, with effective report generation and reviews are needed to promote evaluation activities.
- More Big Data quality reporting and monitoring should be achieved by building automated real-time dashboards.
- The quality evaluation of a representative data set must be performed to produce a quality model that can be applied to the entire data. This will offer an insight into the data quality and provide findings that can be implemented fairly on various types of data.
- The privacy of data must be considered. One of the biggest issues that face Big Data is how people may handle sharing their data with governmental, health, and commercial institutions and companies. Consequently, one of the challenges that Big Data application developers must keep in mind is persuading people to share their data.

VI. REFERENCES

[1] M. van Rijmenam,"A Short History Of Big Data," 2013. Available: https://datafloq.com/read/big-data-history/239. Last Accessed: April 25, 2020.

[2] F.Rider, *The Scholar and theFuture of the Research Library: A Problem and Its Solution*: Hadham Press, NewYork, 1944.

[3] S. Sagirogluand D. Sinanc,"Big data: A review," in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, IEEE, 2013. doi: 10.1109/CTS.2013.6567202.

[4] M.Kataria and M.P. Mittal, "Big data: a review,"*International Journal of Computer Science and Mobile Computing*, vol. 3, no. 7,pp. 106-110, July 2014. https://ijcsmc.com/docs/papers/July2014/V3I7KJ06.pdf.

[5] D. Reinsel and J. Gantz,"The Digital Universe in 2020,"Dec. 2012. Available: https://www.emc.com/leadership/digital-universe/2012iview/index.htm. Last Accessed: April 28, 2020.

[6] E.Dumbill, "Making Sense of Big Data,"*Big Data*, vol. 1, pp. 1-2,2013, doi: 10.1089/big.2012.1503.

[7] D.Laney, "3D Management: Controlling Data Volume, Velocity, and Variety," in *Application Delivery Strategies, META Group*, 2001. Available:https://blogs.gartner.com.

[8] N. Khan, A. Naim, M. Rashid Hussain, Q.N. Naveed, N. Ahmad, and S.Qamar,"The 51 V's Of Big Data: Survey, Technologies, Characteristics, Opportunities, Issues and Challenges," in *Proceedings of the International Conference on Omni-Layer Intelligent Systems (COINS '19)*, Association for Computing Machinery, New York, NY, USA, 2019, pp. 19–24, doi: 10.1145/3312614.3312623.

[9] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics,"*International Journal of Information Management*, vol. 35, no. 2,pp. 137-144,2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[10] https://one.gov.jo [Last Accessed July 13, 2020].

[11] A.Coelho VazHenriques, F. Meirelles, and M.A. Cunha, "Big data analytics: achievements, challenges, and research trends,"*Independent Journal of Management & Production (IJM&P)*,vol. 11, no. 4,pp. 1201-1222,2020, doi: 10.14807/ijmp.v11i4.1085.

[12] M. Abdallah, "Big Data Quality Challenges," in *2019 International Conference on Big Data and Computational Intelligence (ICBDCI)*, Mauritius,2019, pp. 1-3,doi: 10.1109/ICBDCI.2019.8686099.

[13] C.Batini, A. Rula, M. Scannapieco, and G. Viscusi, "From Data Quality to Big Data Quality,"*Journal of Database Management*, vol. 26,pp. 60-82, 2015, doi: 10.4018/JDM.2015010103.

[14] D.M.Strong, Y.W. Lee, and R.Y. Wang, "Data quality in context,"*Commun. ACM*, vol. 40, no. 5,pp. 103-110,1997.

[15] A. Ramasamy and S. Chowdhury, "Big Data Quality Dimensions: A Systematic Literature Review,"*Journal of Information Systems and Technology Management – Jistem USP*,vol. 17,pp. 1-13,2020, doi: 10.4301/S1807-1775202017003.

[16] L.L.Pipino, Y.W. Lee, and R.Y. Wang, "Data quality assessment,"*Commun. ACM*, vol. 45, no. 4,pp. 211-218,2002, doi: 10.1145/505248.506010.

[17] F. Sidi, P. H. ShariatPanahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, "Data quality: A survey of data quality dimensions," in*2012 International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, 2012, pp. 300-304,doi: 10.1109/InfRKM.2012.6204995..

[18] F.I.Salih, S.A. Ismail, M.M. Hamed, O.M.Yusop, A.Azm, and N.F.M.Azmi,"Data Quality Issues in Big Data: A Review," in *3rd International Conference of Reliable Information and Communication Technology (IRICT 2018)*,in F. Saeed, N.Gazem, F. Mohammed, A.Busalim, Eds., *Recent Trends in Data Science and Soft Computing*, Cham: Springer International Publishing, 2019.

[19] M.Mirzaie, B. Behkamal, and S. Paydar, "State of the Art on the Quality of Big Data: A Systematic Literature Review and Classification Framework," 2019. arXiv preprint arXiv:1904.05353.

[20] M.Mirzaie, B. Behkamal, and S. Paydar, "Big Data Quality: A systematic literature review and future research directions,"2019, arXiv preprint arXiv:1904.05353.

[21] N.Abdullah, S.A. Ismail, S.Sophiayati, and S.M. Sam,"Data quality in big data: A review,"*Int. J. Advance Soft Compu. Appl.*,vol. 7, no. 3,pp. 16-27,2015. Available: http://home.ijasca.com/data/documents/IJASCA-SI-070302_Pg16-27_Data-Quality-in-Big-Data-A-Review.pdf

[22] H.J.Hadi, A.H. Shnain, S. Hadishaheed, and A.H. Ahmad, "BigDataand Five V'sCharacteristics,"*International Journal of Advances in Electronics and Computer Science*, vol. 2, no. 1,pp.16-23,2015.Available: http://www.iraj.in/journal/journal_file/journal_pdf/12-105-142063747116-23.pdf.

[23] Ishwarappa and J. Anuradha, "A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology," *Procedia Computer Science*, vol. 48,pp. 319-324,2015, doi: 10.1016/j.procs.2015.04.188.

[24] M.A. Khan, M.F. Uddin, and N. Gupta, "Seven V's of Big Data: Understanding Big Data to extract value,"in *Proc. of the 2014 Zone 1 Conf. of the American Society for Engineering Education*, Bridgeport, CT, 2014, pp. 1-5, doi: 10.1109/ASEEZone1.2014.6820689.

[25] N.Khan, M.Alsaqer, H. Shah, G.Badshah, A.A.Abbasi, S.Salehian, "The 10 Vs, Issues and Challenges of Big Data,"in *2018 International Conference on Big Data and Education (ICBDE '18)*, 2018, pp.52-56,doi:10.1145/3206157.3206166.

[26] Z.Al-Sai, R. Abdullah, and H. Husin, "Critical Success Factors for Big Data: A Systematic Literature Review,"in *IEEE Access*, vol. 8,pp.118940-118956,2020, doi: 10.1109/ACCESS.2020.3005461.

[27] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era,"*Data Science Journal*, vol. 14,2015, doi:10.5334/DSJ-2015-002.

[28] I. Taleb, M.A. Serhani, and R. Dssouli,"Big Data Quality: A Survey," in *2018 IEEE International Congress on Big Data (BigData Congress)*,2018, pp. 166-173, doi: 10.1109/BigDataCongress.2018.00029.

[29] T.Catarci, M. Scannapieco, M. Console, and C. Demetrescu, "My (fair) big data," in*2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, 2017, pp. 2974-2979, doi: 10.1109/BigData.2017.8258267.

[30] E.Gyulgyulyan, J. Aligon, F. Ravat, and H. Astsatryan, "Data Quality Alerting Model for Big Data Analytics,"inWelzer T. et al., Eds,*New Trends in Databases and Information Systems, ADBIS 2019, Communications in Computer and Information Science*, vol. 1064, Springer, Cham, pp. 489-500,2019, doi: 10.1007/978-3-030-30278-8_47.

[31] F. Ridzuan and W.M.N.W. Zainon, "A Review on Data Cleansing Methods for Big Data,"*Procedia Computer Science*, vol. 16,pp. 731-738,2019, doi: 10.1016/j.procs.2019.11.177.

[32] C.S. Rao, J. Rajanikanth, V. Chandra Sekhar, and B. Raju, "Data Cleaning: A Framework for Robust Data Quality In Enterprise Data Warehouse,"*International Journal of Computer Science and Technology*, vol. 3, no. 3,pp. 36-41,2012.

[33] C. Cichy and S. Rass, "An Overview of Data Quality Frameworks,*"IEEE Access*, vol. 7,pp. 24634-24648,2019, doi: 10.1109/ACCESS.2019.2899751.

[34] G.Bath, "The Next Generation Tester: Meeting the Challenges of a Changing ITWorld," in S. Goericke,, Ed.,*The Future of Software Quality Assurance*, pp. 15-26, 2020, doi: 10.1007/978-3-030-29509-7_2.

[35] D.Staegemann, M. Volk, A. Nahhas, M. Abdallah and K. Turowski, "Exploring the Specificities and Challenges of Testing Big Data Systems," in*2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Sorrento, Italy, 2019, pp. 289-295, doi: 10.1109/SITIS.2019.00055.

[36] NIST, *NIST Big Data Interoperability Framework: Volume 1, Definitions*, 2018, [Online]. doi: 10.6028/NIST.SP.1500-1r1.

[37] A.Gani, A.Siddiqa, S. Shamshirband, and F. Hanum, "A survey on indexing techniques for big data: taxonomy and performance evaluation,"*Knowl. Inf. Syst.*,vol. 46,pp. 241–284, 2016, doi: 10.1007/s10115-015-0830-y.

[38] TechAmerica Foundation's Federal Big Data Commission, "Demystifying big data: A practical guide to transforming the business of government," 2012.Available: https://bigdatawg.nist.gov/_uploadfiles/M0068_v1_3903747095.pdf.

[39] A.Katal, M. Wazid, and R.H. Goudar, "Big data: Issues, challenges, tools and Good practices." in *2013 6th International Conference on Contemporary Computing (IC3)*,IEEE, Nodia,2013, 404-409, doi: 10.1109/IC3.2013.6612229.

[40] N. Elgendy and A. Elragal, "Big Data Analytics: A Literature Review Paper,"in Perner P., Ed,*Advances in Data Mining. Applications and Theoretical Aspects, ICDM 2014, Lecture Notes in Computer Science*, Springer, Cham,vol. 8557,pp. 214-227,2014, doi:10.1007/978-3-319-08976-8_16.

[41] D.Loshin, "Evaluating the business impacts of poor data quality,"*Information Quality Journal*, 2011.Available: http://dataqualitybook.com/kii-content/BusinessImpactsPoorDataQuality.pdf.

[42] T.C. Redman, "The Impact of Poor Data Quality on the Typical Enterprise,"*Commun. ACM*, vol. 41, no. 2,pp. 79-82,1998, doi: 10.1145/269012.269025.

[43] S.N. Samsudeen and A. Haleem, "Impacts and Challenges of Big Data: A Review,"*International Journal of Psychosocial Rehabilitation*,vol. 24, no. 7,2020. Available: http://ir.lib.seu.ac.lk/handle/123456789/4348.

[44] A.Haug, F. Zachariassen, and D. Van Liempd, "The costs of poor data quality,"*Journal of Industrial Engineering and Management (JIEM)*, vol. 4, no. 2,pp. 168-193,2011. Available: https://www.jiem.org/index.php/jiem/article/view/232/130.

[45] G. Press, "12 Big Data Definitions: What's Yours?"2014. [Online]. Available: https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#497bbf8713ae. Last Accessed: May 02, 2020.

[46] P.Geczy, "Big data characteristics," *The Macrotheme Review*, vol. 3, no. 6,pp. 94-104,2014. Available: http://macrotheme.com/yahoo_site_admin/assets/docs/8MR36Pe.97110828.pdf.

[47] T.Kolajo, O. Daramola, and A. Adebiyi, "Big data stream analysis: a systematic literature review,"*Journal of Big Data*, vol. 6,pp. 47,2019, doi: 10.1186/s40537-019-0210-7.