Volume XI Issue 3 Year 2026 | Page 1123-1135 ISSN: 2527-9866
Received: 26-06-2026 | Revised: 17-07-2026 | Accepted: 25-08-2026

DRASTIC: Big Data Quality in Big Data Integration

Muhammad Noor[1], Fauziah Baharom[2], Haslina Mohd[3]
[1,2,3] University Utara Malaysa, Sintok, Kedah, Malaysia, 06010
e-mail: m_noor@ahsgs.uum.edu.my[1], fauziah@uum.edu.my[2], haslina@uum.edu.my[3]
*Correspondence: m_noor@ahsgs.uum.edu.my

Abstract: *In today's digital era, organizations are increasingly relying on data-driven decision-making to enhance operational efficiency and strategic planning. Data from multiple sources should be integrated to support this movement. However, this process is complex due to the emergence of big data. Consequently, it significantly increases the challenges of managing and integrating data, which can degrade data quality and lead to poor decision outcomes. In fact, existing data quality characteristics are no longer adequate in big data era. Therefore, this paper conducted a comprehensive literature review of peer-reviewed articles retrieved from electronic databases published between 2010 and 2025 to examine existing data quality characteristics and identify gaps related to the 5V's big data characteristics. Moreover, this paper compares and evaluates existing data quality characteristics and their sufficiency for assessing the quality of data in big data integration. Based on these evaluations, this paper proposes DRASTIC, a set of 14 data quality characteristics, with dependency and scalability introduced as new characteristics in the context of big data integration because these characteristics are underexplored in existing literature. The findings contribute to the literature by extending current data quality characteristics and addressing the challenges posed by big data's unique characteristics in data integration.*
Keywords: *DRASTIC, Big Data Quality, Big Data Quality Evaluation, Big Data Quality Characteristics, Big Data Integration.*

1. Introduction

Numerous technological advancements have occurred in the information technology sector since the start of the 21st century [1], [2]. This trend became particularly prominent in the mid-2000s, as many companies restructured their organizations to respond to the demands of an evolving digital landscape and the Industrial Revolution 4.0 [3]. Trierveiler et al. [4] characterized this movement as digital transformation, through which organizations have developed information systems to improve their operations and increase efficiency. In addition, the fundamental transformation of this movement lies in data-driven business models, in which data is considered one of the most valuable assets, and nearly all strategic business decisions are made based on insights derived from data [5]. It enables the extraction of actionable insights and facilitates accurate decision-making grounded in empirical data. Therefore, data has become the backbone of modern business strategies, innovation and performance.

The study by Zaki [6] highlighted that as more organizations concentrate on digital transformation initiatives, data integration becomes increasingly crucial, depending on their ability to integrate disparate datasets and extract value. Golshan et al. [7], Ma et al. [8], and Schildhauer [9] describe data integration as the process of merging two or more data sets from multiple data locations and sources, often represented in different formats, then compiling them into a unified and accessible storage. It provides users with a unified and comprehensive view of the data. The ability to integrate disparate datasets is essential for organizations, as it impacts operational efficiency, customer loyalty, and overall market competitiveness.

Despite its importance, data integration remains challenging because organizational data is complicated, as it is generated in different structures, formats, and types, including structured, semi-structured, and unstructured data [10], [11]. These various data forms originate from diverse and often disparate sources, known as heterogeneous data, creating a substantial challenge for integration. Sadeghi et al. [12] further emphasize that integrating diverse data sources into a single system can be challenging because the system must be capable of understanding and accommodating these different formats. It requires complex transformations that can introduce inconsistencies, errors, and data quality issues if not carefully controlled.

The main issue addressed in this paper is the insufficiency of current data quality characteristics for evaluating data quality in the big data integration environment. Although big data creates substantial opportunities for organizational value creation, however, these benefits depend on the ability to maintain data quality throughout the integration process. Cai and Zhu [13] and Firmani et al. [14] noted that there remains a lack of consensus on specific criteria for measuring big data quality due to limited data quality standards for big data and the complexity and diversity of big data. Existing data quality characteristics do not encompass all necessary big data characteristics, as the concept of big data quality is still new, making it difficult to establish a uniform or standard definition of big data quality and its criteria.

Accordingly, this paper aims to address the research gap arising from the absence of comprehensive big data quality characteristics that explicitly reflect the complexities of data integration in big data environments. This paper has three objectives: (1) to examine existing data quality characteristics and identify gaps related to 5V's big data characteristics; (2) to evaluate the sufficiency of existing characteristics for assessing data quality in big data integration; and (3) to propose DRASTIC, a new comprehensive set of data quality characteristics for big data integration. The subsequent sections of this paper are organized as follows: Section 2 provides a brief explanation of data integration, big data and data quality. Followed by Section 3, which explains the literature review methodology employed. Section 4 presents and discusses the findings and the proposed DRASTIC data quality characteristics. Finally, Section 5 concludes the paper, provides insights for researchers, clarifies the limitations, and identifies directions for future research.

2. Literature Review

This literature review section provides a brief explanation of previous studies related to the main topic and objective of this paper, with particular emphasis on data integration, big data, and data quality. It establishes the conceptual foundation for examining why conventional data quality characteristics may be insufficient in big data integration environments.

A. Data Integration

Data are typically created, stored, and managed independently across departments, teams, or systems within an organization, resulting in what is known as data silos [15]. These silos emerge when data are not adequately shared across organizational boundaries, producing fragmented information and limiting the ability to obtain an integrated view [16]. Consequently, data integration is necessary to connect these isolated sources and unlock their potential value [17]. Hence, effective integration of various data sources enables organizations to overcome data silos and provides users with more accessible and analyzable information.Moreover, data integration is a fundamental capability in the current data-driven landscape because it plays a vital role in combining information from diverse sources to construct a coherent and unified dataset [18]. More specifically, data integration refers to the process of combining, merging, or joining data from distinct or multiple data objects to create a single, unified data object [7], [8], [9]. Sreemathy et al. [19] emphasize that the data integration process enables organizations to derive valuable and meaningful information from diverse sources, thereby supporting decision-making and operational efficiency. Therefore, the purpose of data integration is to provide seamless access to multiple sources of data and present them as a single resource.

However, data integration is a complex process because it must deal with heterogeneous data in terms of managing, processing, transforming, and analyzing data [20]. Data originating from different sources may vary in type, structure, and format. These differences can introduce data quality problems during integration. Consequently, the effectiveness of data integration depends not only on successful data consolidation but also on maintaining data quality throughout the integration process.

B. Big Data

Big data refers to large volumes of data that are generated at unprecedented speed and consists of diverse types, formats, and sources [21], [22]. Moreover, Katal et al. [23] argued that the significance of big data is not only in its size but also in its complexity. Typically, the characteristics of big data are summarized by the 5Vs: volume, velocity, variety, value, and veracity, which are commonly used to describe the features of big data [10], [24] as illustrated in Figure 1 below.

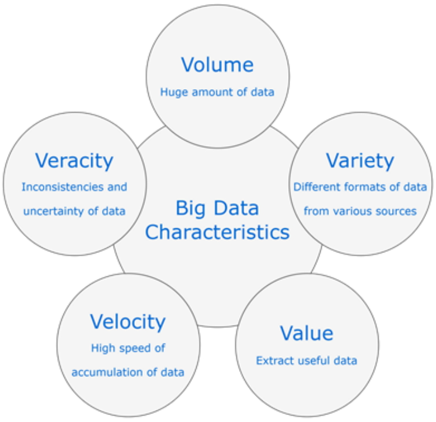

Figure 1. Big Data Characteristics

Figure 1 illustrates the “5Vs” of big data characteristics, which are explained below.

1. Volume refers to the massive scale of data generated, collected, and stored. In the big data environment, the amount of data produced is unprecedented, reaching terabytes, petabytes, and even exabytes. It creates substantial requirements for storage, processing, and management.
2. Velocity relates to the speed of data created, transmitted, transformed, and processed. Recently, data are produced continuously and may require near-real-time or real-time processing to remain useful.
3. Variety refers to the diversity of data types, structures, formats, and sources. Big data may include structured, semi-structured, and unstructured data forms such as text, video, images, social media content, and sensor data.
4. Veracity relates to the accuracy, quality, or trustworthiness of data. As data are generated from numerous and heterogeneous sources, big data environments are more exposed to inaccurate, incomplete, or misleading information. It makes veracity essential to trustworthy analysis and decision-making because the reliability of data-driven insights depends on the quality of the underlying data.
5. Value represents the potential benefits that organizations can derive from data. The value of big data lies in its ability to generate meaningful insights from large and complex datasets that support analysis and decision-making.

Therefore, Ghasemaghaei [25] argued that the characteristics of big data provide a comprehensive understanding of its complexities and potential of big data. These characteristics are essential for managing and utilizing big data effectively, enabling accurate analysis and strategic decision-making.

C. Data Quality

Nowadays, the importance of data quality and trustworthiness has garnered considerable attention, particularly in the context of the exponential growth of big data. Papastergios and Gounaris [26] defined data quality as the degree to which data characteristics meet requirements and are fit for use by humans and/or systems. In the data integration process, maintaining data quality is challenging due to the diverse nature of data in terms of structure, format, semantics, and processing mechanisms [27], [28]. These differences can create conflicts during the data integration process and can degrade the overall quality of the integrated data. Consequently, poor data quality can lead to inaccurate analyses, inappropriate decisions, and operational inefficiencies [29]. Furthermore, Rødseth et al. [30] highlighted that existing data quality characteristics become insufficient when applied to the complex and diverse nature of big data. These obstacles are closely associated with the distinctive characteristics of big data, including volume (i.e., the scale of data), velocity (i.e., the speed of data processing), variety (i.e., the diverse forms of data), and veracity (i.e., the uncertainty of data).

Therefore, the literature indicates a lack of consensus on the criteria for measuring big data quality. This gap stems from the complexity and diversity of big data, the limited availability of quality characteristics for big data, and the fact that the inability of existing data quality characteristics to address the unique characteristics of big data. This issue arises because the concept of big data quality is still new, making it difficult to establish a uniform or standard definition of big data quality and its quality criteria [13], [14]. Moreover, the absence of unified and widely accepted data quality characteristics, along with increased requirements for big data processing technologies, further complicates the situation [13], [31]. Consequently, as data proliferates in the big data era, assessing and standardizing data quality characteristics have become increasingly necessary and increasingly important.

3. Methods

This paper uses a literature review approach to identify, evaluate, and synthesize the existing big data quality characteristics, with particular attention to characteristics that remain underexplored in big data integration. The methodology is designed to provide a transparent basis for identifying gaps in the literature and deriving the proposed DRASTIC characteristics. The research process is adapted from Kitchenham [32]. It has three stages, starts with planning, conducting, and ends with reporting, as illustrated in Figure 2.

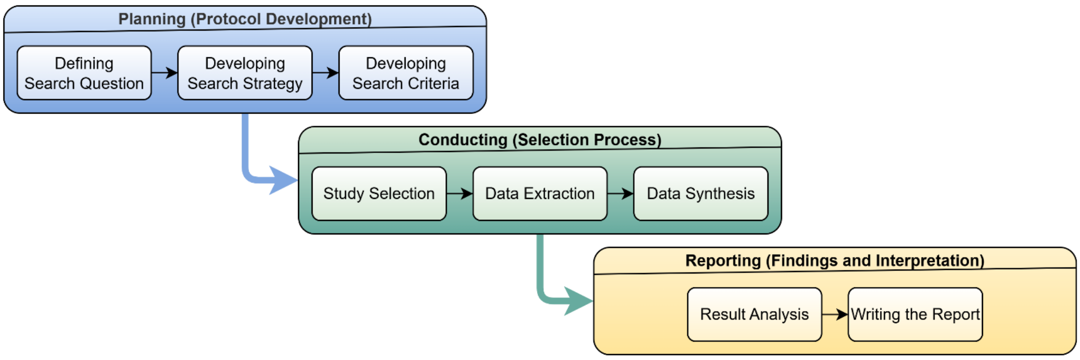

Figure 2. Research Process

A. Planning

The planning stage consists of three activities. It starts with defining the search question that guides the review, as follows:

1. How does the complexity of big data affect data quality?
2. How have existing studies related to big data quality characteristics?
3. What are the gaps remaining unaddressed in the literature?

A formal search strategy was developed after defining the research questions to identify empirical resources that are relevant to the objective of this paper. Searches were conducted across prominent digital libraries, including Web of Science, IEEE Xplore, ACM Digital Library, ScienceDirect, and Google Scholar. It covers peer-reviewed journal articles and high-impact conference proceedings published between 2010 and 2025. Furthermore, search criteria employed combinations of the following search string:

("data" OR "big data") AND ("data quality" OR "quality characteristic").

Retrieved studies were screened using predefined inclusion and exclusion criteria. The inclusion and exclusion criteria were applied to determine study eligibility:

Inclusion criteria:
- The study is published in a peer-reviewed journal;
- The study is written in English;
- The study should be relevant to the search criteria;
- The study should be an empirical research paper, an experience report, or a workshop paper;
- The study was published between 2010 and 2025.

Exclusion criteria eliminated studies that did not discuss data quality in big data integration or were inconsistent with the inclusion criteria.

B. Conducting
The conducting stage consists of three activities, including study selection, data extraction, and data synthesis. Study selection was conducted to ensure the credibility and relevance of the included studies. A formal quality assessment was conducted for all of them. Each study was evaluated against predefined criteria covering methodological accuracy, clarity and transparency of reporting, relevance to the review objectives, and contribution to the data quality literature. Studies that failed to meet the minimum quality criteria were excluded to improve the reliability and validity of the results. Following study selection, data were extracted using a comprehensive literature review approach adapted from Daraojimba [33].

C. Reporting
The reporting stage focuses on presenting the results of the review process. It has two activities, including result analysis and writing the report, as presented in the following results and discussion section.

4. Results and Discussions
Recently, digital transformation has reshaped organizational processes and business models [34], while big data has transformed how organizations collect, process, and analyze information [35]. Companies and organizations are increasingly developing information systems that leverage advances in information technology to support operational activities and improve performance [36]. In this context, digital transformation has elevated data to one of the most valuable assets, with nearly all strategic organizational business decisions made based on insights derived from data [37]. Although it is recognized that the value of such data can be influenced by the processes of data collection, analysis, and interpretation, Gandomi and Haider [38] consider data quality as the most significant factor in determining the value of data. Similarly, Ardagna et al. [39] argued that the true value of such a large amount of data can be realized only when the data are reliable, accurate, and complete.

Nevertheless, Cai and Zhu [13] and Firmani et al. [14] argued that the concept of big data quality remains an emerging concept, and there is currently no universally accepted definition or standardized data quality or quality criteria specifically tailored to the big data context.

Furthermore, the lack of standardized data quality and quality criteria in big data is further intensified by the 5V's unique characteristics of big data, including its volume, variety, veracity, velocity, and value, which fundamentally complicate conventional data quality assessment methods [40]. Accordingly, the increasing complexity of big data environments has posed significant implications for existing data quality characteristics, necessitating the re-examination and extension of conventional data quality characteristics to adequately address the scale, heterogeneity, and dynamic nature of big data, as summarized in Table 1.

Table 1. The Complexity of Big Data and Its Consequences to Data Quality

Data Quality Characteristics	Big Data Characteristics	Justification
Dependency	Variety	The variety of data (types, formats, structures, and sources) increases the challenges of data dependency across datasets [41].
Relevance	Value	Diverse data types and multi-source data can make it more difficult to identify relevant information, reducing the usefulness and relevance of data for decision-making, analytics, and business intelligence [41].
Accuracy	Veracity	The nature of big data, as data is generated from uncertain or noisy sources, can reduce accuracy and produce inconsistent analytical results [42].
Accessibility	Value	Accessible data can be used more efficiently for organizational analytics and decision-making, thereby increasing the value realized from big data [41].
Availability	Variety	Data are often distributed across multiple sources and systems in big data environments, making it challenging to ensure data availability [38].
Scalability	Volume	Scalability is essential for handling the rapid increase in data volume and ensuring system performance remains stable as datasets grow [43].
Timeliness	Velocity	Timeliness is directly associated with the rate at which data is generated, processed, and delivered, thereby reflecting the velocity characteristics inherent to big data [41].
Traceability	Veracity	Traceability supports veracity by enabling stakeholders to verify data origins, transformations, and provenance [42].
Integrity	Veracity	Integrity helps ensure that data remain accurate and unaltered throughout their lifecycle, thereby supporting the veracity or trustworthiness of data [40].
Completeness	Veracity	Data are often collected from unreliable sources, resulting in missing or incomplete records. This undermines the trustworthiness of analytics outcomes, a core aspect of data veracity [41].
Consistency	Veracity	Consistent data across systems and datasets reduces ambiguity and errors, strengthening the reliability and veracity of big data [44]
Compliance	Veracity	Compliance with legal, ethical, and regulatory requirements supports trustworthy data handling and, in turn, contributes to data veracity [23].
Confidentiality	Veracity	Protecting sensitive information from unauthorized access supports trustworthy and verifiable data management, thereby strengthening overall data veracity [42].
Credibility	Veracity	Credibility reflects users' confidence in data sources and analytical outputs, making it a fundamental element of data veracity [45].

Table 1 demonstrates the complex interrelationships between big data characteristics and their consequences for data quality. It reveals that each big data characteristic introduces distinct and interconnected challenges to various data quality characteristics. These interrelationships underscore the need to extend existing data quality characteristics and develop more comprehensive criteria for assessing data quality in big data integration environments.

A. The DRASTIC, Proposed Big Data Characteristics in Big Data Integration

This study proposes DRASTIC, an acronym for 14 big data quality characteristics, including dependency, relevance, accuracy, accessibility, availability, scalability, timeliness, traceability, integrity, completeness, consistency, compliance, confidentiality, and credibility. DRASTIC is intended to address the research gap identified in Table 1, as there remains an unaddressed gap in the literature, providing a comprehensive set of big data quality characteristics tailored to the requirements of big data integration. The 14 characteristics were selected using three complementary criteria. First, each characteristic that appears in at least one existing data quality characteristic established by researchers. It aims to ensure relevance to the data quality context, as shown in Table 2. Second, each characteristic is supported by the literature linking it to challenges arising from big data characteristics to ensure theoretical grounding, as presented in Table 1. Third, the proposed dependency and scalability characteristics were included because they remain underexplored within existing data quality characteristics but are particularly important for data integration in heterogeneous and big data environments. The rationale for each DRASTIC characteristic is presented below.

Dependency

Dependency refers to the degree to which data elements are related or connected to maintain their meaning and integrity [46]. In data integration, where data is collected from multiple sources and systems, maintaining aligned relationships among interdependent elements is crucial to prevent inconsistencies. This characteristic is relatively underexplored in prior data quality research. Recently, dependency has gained attention with the rise of big data. Reflecting its growing importance within complex and interconnected data systems. As organizations move toward integrated data ecosystems, dependence becomes a prominent feature, underscoring the complex relationships among datasets, systems, and business processes [46]. The recognition of dependency emphasized a shift in the discussion of data quality from viewing data elements as isolated units to understanding them as parts of interconnected systems. This means that high data quality is not only about the accuracy of individual values but also about how those values depend on one another within the overall contextual meaning [47]. Hence, dependency highlights the importance of properly aligning interconnected elements and improves interoperability in multi-domain integration, preventing inconsistencies and ensuring that data from various systems aligns logically [48], [49].

Relevance

Relevance refers to the extent to which data is appropriate and valuable for its intended use [13]. In data integration, relevance means the resulting data aligns with specific organizational objectives, operational processes, and user requirements. Data may be technically accurate, however, they may still offer limited value if they are not contextually appropriate for the intended analysis or decision. Therefore, relevance captures the growing awareness of the contextual dimension of data quality by emphasizing whether integrated data are meaningful and applicable to their intended use.

Accuracy

Accuracy refers to the degree to which data correctly reflects the real-world objects or events it is intended to represent [50]. It assesses the degree to which observed data correspond to valid or accepted values. Accuracy is a fundamental characteristic of data quality and is closely associated with the veracity dimension of big data. Maintaining accurate data is essential for producing trustworthy analyses and supporting data-driven decisions.

Accessibility

Accessibility refers to the degree to which data can be accessed easily and efficiently when the user needs it [13]. It emphasizes the practical usability of integrated data and the ease with which authorized users can obtain and use relevant information. Accessibility is also linked to the value dimension of big data because timely and efficient access enables organizations to derive greater analytical and operational value from integrated datasets.

Availability

Availability refers to the extent to which data can be retrieved by authorized users and/or applications when required [51]. It ensures continuity and reliability of data access, which is especially important for time-sensitive operations. In big data integration environments, maintaining availability across heterogeneous and distributed sources can be difficult because data may be dispersed across systems with varying operational and technical characteristics.

Scalability

Scalability refers to the degree of the system's ability to handle increasing data volumes, diverse data sources, and growing user demands without significant performance degradation [52]. It is closely associated with the volume characteristic of big data and recognizing its growing importance as datasets increase exponentially. Scalability emphasizes the system's ability to handle increasing data volumes, diverse data sources, and growing user demands without significant performance degradation [52]. In distributed storage and processing systems, inadequate scalability can lead to performance deterioration, processing bottlenecks, and, ultimately, data quality degradation.

Timeliness

Timeliness refers to the degree to which data is available within the expected timeframes for decision-making or operational needs [50]. It reflects how current or up-to-date the data is for its intended use and ensures that decisions are based on the most recent available information. It highlights the relevance of real-time analytics and dynamic decision-making. Timeliness is closely associated with the velocity dimension of big data because rapid data generation increases the need for timely or real-time processing. In contrast, outdated data may reduce the relevance and effectiveness of the decisions.

Traceability

Traceability refers to the degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data [51]. It enables users to determine where data originated, how they were transformed, and how it moved across integrated systems. Therefore, traceability is important for auditing, compliance, and accountability in analytical processes. Without adequate traceability, it becomes more difficult to verify the provenance and reliability of integrated data and analytical results.

Integrity

Integrity refers to the degree to which data are free from corruption during storage or transfer [13]. It supports the preservation of data correctness throughout the data lifecycle and helps prevent unauthorized or accidental changes. In big data environments, integrity is closely related to veracity because data transformations across multiple systems must preserve data integrity and the reliability of the resulting information.

Completeness

Completeness refers to the degree to which all required data is present [50]. This characteristic is closely related to the veracity aspect of big data because missing or incomplete records can reduce the reliability of analytical and decision-making. In big data integration, data are often collected from unreliable or heterogeneous sources, leading to missing or incomplete records. Consequently, maintaining completeness is essential for producing integrated datasets that are sufficiently comprehensive for analysis and decision-making.

Consistency

Consistency refers to the uniformity of data across different datasets and systems [50]. Consistent data across systems and datasets reduces ambiguity and prevents conflicting representations, thereby strengthening the reliability and veracity of big data. In big data environments, heterogeneous sources may consist of different definitions, standards, and formats. Therefore, consistent representations and shared rules are essential for interoperability and reliable integration.

Compliance

Compliance refers to the degree to which data has attributes that adhere to standards, conventions, or regulations in force, as well as similar rules relating to data quality [51]. This characteristic contributes to the trustworthiness of data by ensuring that data are handled in accordance with relevant legal, ethical, and organizational requirements. In the context of big data, compliance also strengthens accountability and supports confidence in how integrated data are collected, processed, stored, and utilized.

Confidentiality

Although confidentiality is often considered a security issue rather than a quality concern, it is acknowledged as essential to the trustworthiness of data handling, especially in privacy-sensitive environments. Confidentiality refers to the degree to which data attributes ensure that it is accessible and interpretable only by authorized users [51]. Protecting integrated data from unauthorized access and disclosure helps preserve trust and supports the responsible use of data. This relationship illustrates the close connection between data quality, security, and trustworthy information management.

Credibility

Credibility refers to the degree to which data has attributes that are regarded as true and believable by users [51]. It concerns how believable and trustworthy data appear to users, particularly when addressing data bias or misinformation. Credibility is closely associated with veracity because stakeholders must be able to trust both the underlying data and the analytical outputs. Collectively, the 14 DRASTC data quality characteristics contribute a unique and complementary role in determining the usability, validity, and overall value of big data. The diversity of these characteristics underscores the evolving and increasingly complex landscape of the big data environment.

B. Comparison of Existing Data Quality Characteristics

Numerous methodologies exist that have been proposed to identify and manage data quality, each addressing different aspects of quality assessment by proposing appropriate dimensions, measures, and tools [53]. Several works have proposed initiatives to incorporate data quality characteristics, such as Cai and Zhu [13], Taleb et al. [50], Elouataoui et al. [54], Jansen et al. [55], Papastergios and Gounaris [26], and ISO 25012 [51]. These sources provide the basis for the comparative analysis presented in Table 2.

Table 2. The Comparison of Existing Data Quality Characteristics

Data Quality Characteristics	Big Data Characteristics	DRASTIC	ISO/IEC 25012	Cai & Zhu (2015)	Taleb et. al. (2016)	Elouataoui et al. (2022)	Jansen et al. (2024)	Papastergios and Gounaris (2024)
Dependency	Variety	√						
Relevance	Value	√		√		√		
Accuracy	Veracity	√	√	√	√		√	√
Accessibility	Value	√	√	√		√		√
Availability	Variety	√	√			√		
Scalability	Volume	√						
Timeliness	Velocity	√	√	√	√	√	√	√
Traceability	Veracity	√	√					
Integrity	Veracity	√		√		√		
Completeness	Veracity	√	√	√	√	√	√	√
Consistency	Veracity	√	√	√	√	√	√	√
Compliance	Veracity	√	√					√
Confidentialiy	Veracity	√	√					
Credibility	Veracity	√	√					

Table 2 presents a comparison of existing data quality characteristics established by researchers. This comparison highlights researchers' agreement and disagreement regarding data quality characteristics. These characteristics serve as standards for evaluating whether data fulfill the criteria of their stated purpose. It provides an essential foundation for understanding how researchers have approached the concept of data quality. Most notably, Table 2 reveals that dependency is entirely absent from all existing data quality characteristics introduced by researchers despite its importance to maintaining relationships among heterogeneous datasets during integration. Similarly, scalability is not represented in any existing data quality characteristics despite its relevance to the increasing volume and processing demands of big data environments. The comparison indicates that DRASTIC is the only data quality characteristic in Table 2 that includes both dependency and scalability. These characteristics address two important challenges in big data integration, including preserving relationships from heterogeneous and interconnected data and the ability to handle exponentially increasing data volumes.

5. Conclusions

Digital transformation has elevated data to one of the most valuable organizational assets, with all strategic business decisions made based on insights derived from data. However, one notable consequence of this transformation is the massive volume and continuous generation of data at an unprecedented rate, commonly referred to as big data. As more organizations focus on digital transformation initiatives, data integration becomes increasingly crucial, as it depends on their capability to integrate diverse data and derive value. This paper examined the challenges associated with the data integration process in the complex big data environment. Ensuring high-quality data in the big data era has not only become more important but also more challenging. A comprehensive literature review of peer-reviewed articles retrieved from electronic databases published between 2010 and 2025 revealed that the existing data quality characteristics become insufficient when applied to the complexity of a big data environment, particularly in the data integration process and left the underexplored dimension. Based on the synthesis, this paper proposes DRASTIC, a set of 14 data quality characteristics for assessing integrated data in big data environments which are acronym of dependency, relevance, accuracy, accessibility, availability, scalability, timeliness, traceability, integrity, completeness, consistency, compliance, confidentiality, and credibility. In addition, this paper also introduces dependency and scalability, which are particularly important contributions because they address the relational complexity and increasing scale in big data integration. Furthermore, this paper is expected to provide guidance to stakeholders on assessing data quality in big data, particularly during the data integration process. However, the limitations of this paper were derived from the concept identified through a comprehensive review of existing literature. Although this paper review provides theoretical grounding, empirical validation through case studies and expert review is needed to confirm DRASTIC's sufficiency in real-world settings. Hence, future research is expected to conduct empirical studies to validate DRASTIC in various data integration scenarios, operationalize the characteristics through measurable metrics, and develop practical guidelines and implementation strategies.

References

[1] S. A. Al Wahid, N. Mohammad, R. Islam, M. H. Faisal, and M. S. Rana, "Evaluation of Information Technology Implementation for Business Goal Improvement under Process Functionality in Economic Development," *Journal of Data Analysis and Information Processing*, vol. 12, no. 2, pp. 304–317, 2024, doi: 10.4236/jdaip.2024.122017.

[2] R. L. V. Nyoto, N. Nyoto, and N. Renaldo, "Information Technology Advancements for A Digital Economy," *Informatica and Digital Insight Journal*, vol. 1, no. 1, pp. 1–7, 2024.

[3] I Made Sukarsa *et al.*, "Evaluation of E-Government Maturity Models in Sub-District Public Services in Indonesia Using the SPBE Framework," *RESTI*, vol. 4, no. 2, pp. 243–253, Apr. 2020, doi: 10.29207/resti.v4i2.1825.

[4] H. J. Trierveiler, D. Sell, and N. dos Santos, “The benefits and challenges of digital transformation in Industry 4.0,” *Global Journal of Management and Business Research*, vol. 19, no. A12, pp. 27–39, 2019.
[5] M. K. Saggi and S. Jain, “A survey towards an integration of big data analytics to big insights for value-creation,” *Information Processing & Management*, vol. 54, no. 5, Art. no. 5, Sep. 2018, doi: 10.1016/j.ipm.2018.01.010.
[6] M. Zaki, “Digital transformation: harnessing digital technologies for the next generation of services,” *JSM*, vol. 33, no. 4, pp. 429–435, Sep. 2019, doi: 10.1108/JSM-01-2019-0034.
[7] B. Golshan, A. Halevy, G. Mihaila, and W.-C. Tan, “Data Integration: After the Teenage Years,” in *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, Chicago Illinois USA: ACM, May 2017, pp. 101–106. doi: 10.1145/3034786.3056124.
[8] B. Ma, T. Jiang, X. Zhou, F. Zhao, and Y. Yang, “A Novel Data Integration Framework Based on Unified Concept Model,” *IEEE Access*, vol. 5, pp. 5713–5722, 2017, doi: 10.1109/ACCESS.2017.2672822.
[9] M. Schildhauer, “Data Integration: Principles and Practice,” in *Ecological Informatics*, F. Recknagel and W. K. Michener, Eds., Cham: Springer International Publishing, 2018, pp. 129–157. doi: 10.1007/978-3-319-59928-1_8.
[10] H. B. Alwan and K. R. Ku-Mahamud, “Big data: definition, characteristics, life cycle, applications, and challenges,” *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 769, no. 1, p. 012007, Feb. 2020, doi: 10.1088/1757-899X/769/1/012007.
[11] M. S. Kettouch, C. Luca, M. Hobbs, and A. Fatima, “Data integration approach for semi-structured and structured data (Linked Data),” in *2015 IEEE 13th International Conference on Industrial Informatics (INDIN)*, Cambridge, United Kingdom: IEEE, Jul. 2015, pp. 820–825. doi: 10.1109/INDIN.2015.7281842.
[12] M. Sadeghi, A. Carenini, O. Corcho, M. Rossi, R. Santoro, and A. Vogelsang, “Interoperability of heterogeneous Systems of Systems: from requirements to a reference architecture,” *J Supercomput*, vol. 80, no. 7, pp. 8954–8987, May 2024, doi: 10.1007/s11227-023-05774-3.
[13] L. Cai and Y. Zhu, “The Challenges of Data Quality and Data Quality Assessment in the Big Data Era,” *CODATA*, vol. 14, no. 0, Art. no. 0, May 2015, doi: 10.5334/dsj-2015-002.
[14] D. Firmani, M. Mecella, M. Scannapieco, and C. Batini, “On the Meaningfulness of ‘Big Data Quality’ (Invited Paper),” *Data Sci. Eng.*, vol. 1, no. 1, Art. no. 1, Mar. 2016, doi: 10.1007/s41019-015-0004-7.
[15] M. Wibowo, S. Sulaiman, and S. M. Shamsuddin, “Machine Learning in Data Lake for Combining Data Silos,” in *Data Mining and Big Data*, vol. 10387, Y. Tan, H. Takagi, and Y. Shi, Eds., in Lecture Notes in Computer Science, vol. 10387. , Cham: Springer International Publishing, 2017, pp. 294–306. doi: 10.1007/978-3-319-61845-6_30.
[16] J. Patel, “Bridging data silos using big data integration,” *International Journal of Database Management Systems*, vol. 11, no. 3, pp. 01–06, 2019, doi: 10.5121/ijdms.2019.11301.
[17] Z. Adams and A. Nelson, “Breaking Down Silos: Integrating Big Data Across Organizational Functions,” *International Journal of Network and Communication Research*, vol. 7, no. 1, pp. 81–89, 2023.
[18] K. Zhang, Z. Chu, J. Xing, H. Zhang, and Q. Cheng, “Urban Traffic Flow Congestion Prediction Based on a Data-Driven Model,” *Mathematics*, vol. 11, no. 19, p. 4075, Sep. 2023, doi: 10.3390/math11194075.
[19] J. Sreemathy, I. Joseph V., S. Nisha, C. Prabha I., and G. Priya R.M., “Data Integration in ETL Using TALEND,” in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Coimbatore, India: IEEE, Mar. 2020, pp. 1444–1448. doi: 10.1109/ICACCS48705.2020.9074186.
[20] M. Asch *et al.*, “Big data and extreme-scale computing: Pathways to Convergence-Toward a shaping strategy for a future software and data ecosystem for scientific inquiry,” *The International Journal of High Performance Computing Applications*, vol. 32, no. 4, pp. 435–479, Jul. 2018, doi: 10.1177/1094342018778123.
[21] W. Jiang, X. Ning, and Y. Xu, “Review on Big Data Fusion Methods of Quality Inspection for Consumer Goods,” in *2018 5th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2018 4th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, Shanghai: IEEE, Jun. 2018, pp. 95–102. doi: 10.1109/CSCloud/EdgeCom.2018.00025.

[22] M. Oliveira and D. Afonso, "Industry Focused in Data Collection: How Industry 4.0 is Handled by Big Data," in *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, Seoul Republic of Korea: ACM, Jul. 2019, pp. 12–18. doi: 10.1145/3352411.3352414.
[23] A. Katal, M. Wazid, and R. H. Goudar, "Big data: Issues, challenges, tools and Good practices," in *2013 Sixth International Conference on Contemporary Computing (IC3)*, Noida, India: IEEE, Aug. 2013, pp. 404–409. doi: 10.1109/IC3.2013.6612229.
[24] I. Taleb, M. A. Serhani, and R. Dssouli, "Big Data Quality Assessment Model for Unstructured Data," in *2018 International Conference on Innovations in Information Technology (IIT)*, Al Ain: IEEE, Nov. 2018, pp. 69–74. doi: 10.1109/INNOVATIONS.2018.8605945.
[25] M. Ghasemaghaei, "Understanding the impact of big data on firm performance: The necessity of conceptually differentiating among big data characteristics," *International Journal of Information Management*, vol. 57, p. 102055, Apr. 2021, doi: 10.1016/j.ijinfomgt.2019.102055.
[26] V. Papastergios and A. Gounaris, "A survey of open-source data quality tools: shedding light on the materialization of data quality dimensions in practice," 2024, *arXiv*. doi: 10.48550/ARXIV.2407.18649.
[27] F. Nargesian, A. Asudeh, and H. V. Jagadish, "Responsible Data Integration: Next-generation Challenges," in *Proceedings of the 2022 International Conference on Management of Data*, Philadelphia PA USA: ACM, Jun. 2022, pp. 2458–2464. doi: 10.1145/3514221.3522567.
[28] A. Bin Mahfoodh, M. Ibrahim, M. Hawi, and K. Hakami, "Introducing a Big Data System for Maintaining Well Data Quality and Integrity in a World of Heterogeneous Environment," in *Day 2 Tue, April 25, 2017*, Dammam, Saudi Arabia: SPE, Apr. 2017, p. D023S012R002. doi: 10.2118/188082-MS.
[29] V. Gudivada, A. Apon, and J. Ding, "Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations," *International Journal on Advances in Software*, vol. 10, no. 1, pp. 1–20, 2017.
[30] Ø. J. Rødseth, L. P. Perera, and B. Mo, "Big data in shipping-Challenges and opportunities," 2016, Accessed: Aug. 01, 2024. [Online]. Available: https://sintef.brage.unit.no/sintef-xmlui/handle/11250/2390646
[31] E. T. Bekar, P. Nyqvist, and A. Skoogh, "An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study," *Advances in Mechanical Engineering*, vol. 12, no. 5, p. 168781402091920, May 2020, doi: 10.1177/1687814020919207.
[32] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," 2007.
[33] E. C. Daraojimba, C. N. Nwasike, A. O. Adegbite, C. A. Ezeigweneme, and J. O. Gidiagba, "COMPREHENSIVE REVIEW OF AGILE METHODOLOGIES IN PROJECT MANAGEMENT," *Comput. sci. IT res. j.*, vol. 5, no. 1, pp. 190–218, Jan. 2024, doi: 10.51594/csitrj.v5i1.717.
[34] K. Schwertner, "Digital transformation of business," *TJS*, vol. 15, no. Suppl.1, pp. 388–393, 2017, doi: 10.15547/tjs.2017.s.01.065.
[35] I. Yaqoob *et al.*, "Big data: From beginning to future," *International Journal of Information Management*, vol. 36, no. 6, pp. 1231–1247, Dec. 2016, doi: 10.1016/j.ijinfomgt.2016.07.009.
[36] S. Tams, V. Grover, and J. Thatcher, "Modern information technology in an old workforce: Toward a strategic research agenda," *The Journal of Strategic Information Systems*, vol. 23, no. 4, pp. 284–304, Dec. 2014, doi: 10.1016/j.jsis.2014.10.001.
[37] A. Sestino, A. Kahlawi, and A. De Mauro, "Decoding the data economy: a literature review of its impact on business, society and digital transformation," *European Journal of Innovation Management*, 2023, doi: 10.1108/EJIM-01-2023-0078.
[38] 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, Apr. 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.
[39] D. Ardagna, C. Cappiello, W. Samá, and M. Vitali, "Context-aware data quality assessment for big data," *Future Generation Computer Systems*, vol. 89, pp. 548–562, Dec. 2018, doi: 10.1016/j.future.2018.07.014.
[40] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi, "From Data Quality to Big Data Quality:," *Journal of Database Management*, vol. 26, no. 1, pp. 60–82, Jan. 2015, doi: 10.4018/JDM.2015010103.

[41] A. Wahyudi, A. Farhani, and M. Janssen, “Relating Big Data and Data Quality in Financial Service Organizations,” in *Challenges and Opportunities in the Digital Era*, vol. 11195, S. A. Al-Sharhan, A. C. Simintiras, Y. K. Dwivedi, M. Janssen, M. Mäntymäki, L. Tahat, I. Moughrabi, T. M. Ali, and N. P. Rana, Eds., in Lecture Notes in Computer Science, vol. 11195. , Cham: Springer International Publishing, 2018, pp. 504–519. doi: 10.1007/978-3-030-02131-3_45.

[42] J. Stefanowski, K. Krawiec, and R. Wrembel, “Exploring complex and big data,” *International Journal of Applied Mathematics and Computer Science*, vol. 27, no. 4, pp. 669–679, Dec. 2017, doi: 10.1515/amcs-2017-0046.

[43] C.-W. Tsai, C.-F. Lai, H.-C. Chao, and A. V. Vasilakos, “Big data analytics: a survey,” *Journal of Big Data*, vol. 2, no. 1, p. 21, Dec. 2015, doi: 10.1186/s40537-015-0030-3.

[44] I. Taleb, M. A. Serhani, and R. Dssouli, “Big Data Quality: A Survey,” in *2018 IEEE International Congress on Big Data (BigData Congress)*, San Francisco, CA, USA: IEEE, Jul. 2018, pp. 166–173. doi: 10.1109/BigDataCongress.2018.00029.

[45] R. Y. Wang and D. M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, Mar. 1996, doi: 10.1080/07421222.1996.11518099.

[46] B. Thalheim, *Dependencies in relational databases*. Springer-Verlag, 2013. Accessed: May 13, 2025. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=oTS2BgAAQBAJ&oi=fnd&pg=PA17&dq=Dependency+refers+to+the+degree+of+relationships+or+connections+between+different+datasets+or+data+elements&ots=Sg3cBSI5Cc&sig=cOsR3ncUF84dfRutkkESrIHVuOo

[47] S. Canali, “Towards a Contextual Approach to Data Quality,” *Data*, vol. 5, no. 4, p. 90, Sep. 2020, doi: 10.3390/data5040090.

[48] P. Pernot and F. Cailliez, “A critical review of statistical calibration/prediction models handling data inconsistency and model inadequacy,” *AIChE Journal*, vol. 63, no. 10, pp. 4642–4665, Oct. 2017, doi: 10.1002/aic.15781.

[49] S. Sandhya Kona, “Bridging Data Silos: Enhancing Business Operations through Advanced Data Integration and System Interoperability,” *IJSR*, vol. 12, no. 12, pp. 2100–2104, Dec. 2023, doi: 10.21275/sr24529180313.

[50] I. Taleb, H. T. E. Kassabi, M. A. Serhani, R. Dssouli, and C. Bouhaddioui, “Big Data Quality: A Quality Dimensions Evaluation,” in *2016 Intl IEEE Conferences on 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)*, Toulouse: IEEE, Jul. 2016, pp. 759–765. doi: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122.

[51] “ISO/IEC 25012:2008,” ISO. Accessed: Nov. 12, 2024. [Online]. Available: https://www.iso.org/standard/35736.html

[52] H. Hu, Y. Wen, T.-S. Chua, and X. Li, “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial,” *IEEE Access*, vol. 2, pp. 652–687, 2014, doi: 10.1109/ACCESS.2014.2332453.

[53] C. Batini and A. Rula, “From Data Quality to Big Data Quality: A Data Integration Scenario.,” in *SEBD*, 2021, pp. 36–47. Accessed: Jan. 03, 2025. [Online]. Available: https://ceur-ws.org/Vol-2994/paper4.pdf

[54] W. Elouataoui, I. El Alaoui, S. El Mendili, and Y. Gahi, “An Advanced Big Data Quality Framework Based on Weighted Metrics,” *BDCC*, vol. 6, no. 4, p. 153, Dec. 2022, doi: 10.3390/bdcc6040153.

[55] B. J. Jansen, K. K. Aldous, J. Salminen, H. Almerekhi, and S. Jung, “Data Quality Assessment,” in *Understanding Audiences, Customers, and Users via Analytics*, in Synthesis Lectures on Information Concepts, Retrieval, and Services. , Cham: Springer Nature Switzerland, 2024, pp. 55–64. doi: 10.1007/978-3-031-41933-1_5.

Acknowledgements

This work was supported by the Ministry of Higher Education (MoHE) of Malaysia under the Fundamental Research Grant Scheme (FRGS) number FRGS/1/2022/ICT01/UUM/02/2.