Scientific Research Publishing

Journal of Computer and Communications, 2022, 10, 55-73
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Short Review of the Literature on Automatic Data Quality

Deepak R. Chandran[1], Vikram Gupta[2]

[1]President & CTO, Iris Energy LLC, Edison, NJ, USA
[2]Sr. Director, Head of AWS Cloud Practice-Emerging Technology, CGI, New York, NY, USA
Email: cdr22@me.com

How to cite this paper: Chandran, D.R. and Gupta, V. (2022) A Short Review of the Literature on Automatic Data Quality. *Journal of Computer and Communications*, **10**, 55-73.
https://doi.org/10.4236/jcc.2022.105004

Received: April 16, 2022
Accepted: May 28, 2022
Published: May 31, 2022

Abstract

Several organizations have migrated to the cloud for better quality in business engagements and security. Data quality is crucial in present-day activities. Information is generated and collected from data representing real-time facts and activities. Poor data quality affects the organizational decision-making policy and customer satisfaction, and influences the organization's scheme of execution negatively. Data quality also has a massive influence on the accuracy, complexity and efficiency of the machine and deep learning tasks. There are several methods and tools to evaluate data quality to ensure smooth incorporation in model development. The bulk of data quality tools permit the assessment of sources of data only at a certain point in time, and the arrangement and automation are consequently an obligation of the user. In ensuring automatic data quality, several steps are involved in gathering data from different sources and monitoring data quality, and any problems with the data quality must be adequately addressed. There was a gap in the literature as no attempts have been made previously to collate all the advances in different dimensions of automatic data quality. This limited narrative review of existing literature sought to address this gap by correlating different steps and advancements related to the automatic data quality systems. The six crucial data quality dimensions in organizations were discussed, and big data were compared and classified. This review highlights existing data quality models and strategies that can contribute to the development of automatic data quality systems.

Keywords

Data Quality, Monitoring, Toolkit, Dimension, Organization

1. Introduction

In handling data within an organization, realizing high data quality has become

a crucial element. Organizations having high data quality would have better decision-making processes, improved business strategy plans, and unveil excellent patterns for better problem-solving. Inability to provide high data quality has led to organizations having issues such as customer dissatisfaction, high operational cost, and wrong decisions due to incorrect data [1]. Several studies in data quality have unraveled numerous advancements in defining data measurements, dimensions, techniques, quality, evaluation, and improvement models. With the background of data quality being wide, scientists and researchers have decided that high quality of data is characterized by data that is fit for use and satisfies all the conditions set by the users, based on its field of application [2].

Data quality is a significant concern in numerous application fields. The efficiency of the model forecast is hugely dependent on the data quality employed in model implementation [3]. The valuation of data quality is essential and plays a crucial role in assessing the practicality of data gathered from the Software Process [4] and empirical software [5]. Some fields that have studied data quality include multimedia data, the internet of things [6], smart cities, deep learning, and machine learning [7], big data analytics and management [8], drug dataset, and living systems [9].

While several studies are published in relation to different dimensions of data quality, there are no papers that capture comprehensively all the dimensions of an automated data quality systems. This gap in the literature is because the field of automated data quality systems is still under development. This paper seeks to address the identified gap by collating all the important dimensions of data quality management, which has a bearing on the development of automated data quality management systems. The methodology used in this qualitative study is the narrative review of the existing literature, primarily from the perspective of practitioners who are currently engaged in developing the automated data quality systems.

This introductory section is followed by a review of the selected literature having a bearing on the development of automated data quality systems. The experience of the author in developing the system is drawn for a subjective selection of the existing literature. After the literature review, the methodology used for this study is briefly explained. The methodology section is followed by a discussion of the findings from the study and drawing conclusions, including for the future research.

2. Literature on Data Quality

Numerous advancements have been recorded in data quality research. Data quality has tremendous importance; accurate data provide excellent results and ensure a deep understanding of the research data of the research establishment, which are updated. Data quality dimensions such as timeliness, completeness, and correctness are vital for successful, effective processes. Errors recorded in data extend across numerous areas and weaken the entire research tasks of an

organization. This study captures and review different aspects of data quality present in organizations and various other platforms, as discussed in the existing literature.

2.1. Data Quality Procedures

Different strategies, process-driven and data-driven, can improve data quality. Batini *et al.* [10] reported that individual strategies used different techniques. Nevertheless, each strategy is aimed at enhancing the quality of data.

Process-driven strategy

Process-driven strategy is aimed at redesigning all the processes, or processes and alters the data to enhance quality. This type of strategy has two additional techniques, namely, process redesign and process control. Process redesigns eliminate the causes of low data quality and thereby add a new process to generate high quality, while Process controls check and manage the manufacturing process [10]. Process-driven strategy per-forms better in the long period; it obliterates the causes of quality problems.

Data-driven strategy

A data-driven strategy is intended at improving the data quality through adjustment of the data value, nonstop. In this strategy, there are some related improvement methods of data-driven strategy, and they are data error correction and localization, normalization, standardization, source trustworthiness, cost optimization, and schematic integration [10]. Data-driven strategy is more expensive than process-driven, and it is effective in the short term.

2.2. Types of Data

Different strategies, process-driven and data-driven, can improve data quality. Batini *et al.* [10] reported that individual strategies used different techniques. Nevertheless, each strategy is aimed at enhancing the quality of data.

Data are the new oil; it can be stored, elaborated, and retrieved through a soft-ware-based process and can communicate through a network [11]. Researchers have classified data into different types in different are-as. **Table 1** presents three different data types according to their classifications, implicitly or explicitly. Another classification of data is considering data as a product and is classified into three types as presented in **Table 2**.

Table 1. Data as an implicit or explicit.

Data Types	Descriptions
Structured Data	Generalization of objects defined by basic features defined within a field
Unstructured Data	A generic sequence of symbols typically coded in natural language
Semi-Structured Data	Data that possess a structure with some level of flexibility

Adapted from Sidi *et al.* [12].

Table 2. Data as a product.

Data Types	Descriptions
Raw Data Objects	Smaller data ties are used to generate information and mechanisms data objects.
Component Data Objects	Data are generated from raw data objects and stored momentarily until the final product is manufactured
Information Products	Data that is the concern of performing manufacturing action on data

Adapted from Sidi *et al.* [12].

Data can also be classified based on austerity to measure and achieve quality data, and this can be classified into two classes, namely, aggregated and elementary data. Data managed using active processes and that represent atomic occurrences of the world are referred to as elementary data, and they are age, sex, and others. Data gathered from elementary data and applied with an aggregate function is called aggregate data, tax, income, and others.

2.3. Data Quality Methods

Data quality is a multidimensional idea as data quality is assessable by considering a range of scopes applicable to a specific field. There are three different approaches to assess the data quality dimensions: empirical, theoretical, and intuitive.

Empirical method

The empirical approach is employed to analyze data gathered from consumers; it is used to determine the dimensions they use to evaluate if the data are fit for use in their frameworks. The empirical method is a general term for research techniques that conclude from observable evidence [13].

Theoretical method

This approach focuses on data becoming incomplete during the built-up process. The theoretical approach can provide a complete set of data quality dimensions. This data quality approach assumes that an information system represents a real-time world system as the users see. The quality of data dimension is derived based on possible inconsistencies between how the user sees the practical system as inferred from the information system and the view collected by openly observing the practical system.

Intuitive method

The data quality dimension in an intuitive approach is chosen based on the researcher's experience and knowledge about what features are vital in a specific framework. The merit of using this method is that it allows the individual study to choose the dimensions important to the specific goals of the framework.

2.4. Data Quality Problem

Different processes obstruct data; most of these processes affect data quality to a

certain extent. The quality of data deteriorates differently based on the situation they are found. There are thirteen kinds of processes that lead to data problems; Figure 1 presents some of the causes of data problems which are gathered into three advanced categories.

The left side of Figure 1 presents the processes in bringing data from the outside through manual means or using data integration methods and some interfaces. The introduction of incoming data will bring forth errors during the data loading, transformation, and extraction processes; these problems are further magnified through the traffic of high data volumes. The right side contains the processes that handle the data manipulation. Some procedures in data manipulation are routine, while some are due to database restructure, periodic, mass data updates, and some other ad hoc actions. However, in the implementation stage, some of these procedures lack the resources, time, and reliable metadata necessary to realize data quality effects.

The bottom part highlights the chain of processes that cause correct data to be incorrect over time, without any physical alteration applied to it. It simply means the data values are intact, but their accuracy drops. Such a situation arises when a real-time object described by the data changes, but the data gathering processes do not capture the change, turning the old data inaccurate and obsolete.

Sidi *et al.* [12] further classified data quality problems into two different groups: single-source and multi-source problems. Previous research also categorizes data quality into four different forms, as illustrated in Table 3.

2.5. Roles in Data Quality

There are three vital roles in the framework of data quality; they are data custodian, data producers, and data consumers [15] [16]. Data custodians are the individuals managing the resources for impending use. Data producers are individuals, sources, or groups that produce the data. Data consumers are the data users who have access to the collected data. These roles in the context of data quality are related to a task. Producers of data are associated with the production of the data process. Data custodians are saddled with the securing, maintenance, and storage of data and give access to consumers that want to employ data for a specific process. Data quality works gauge its quality from the consumers' perception of the data. In some dispensation data, custodians and producers are often categorized as the same unit.

2.6. Data Quality Toolkit

It is a known fact from previous literature that the performance of machine learning and deep learning model is bounded by the quality of data used for model training and validation [17] [18]. There have been continuous efforts to improve the quality of models, but little effort has been recorded toward refining the quality of data. One of the essential requirements before employing a specific

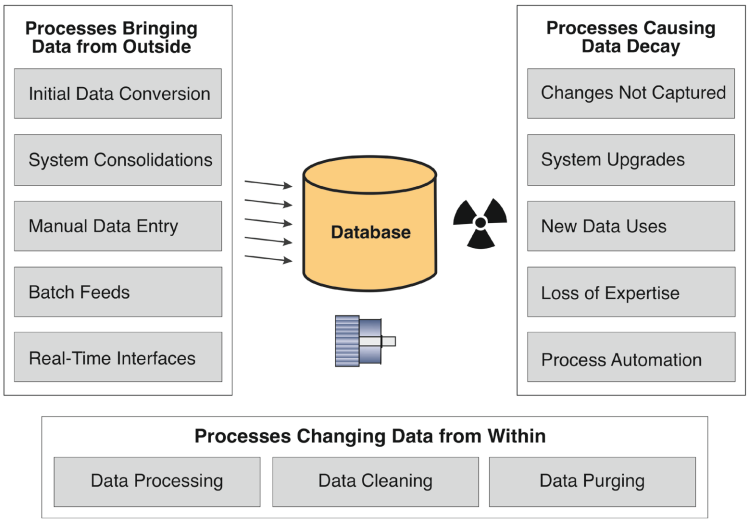

Figure 1. Processes affecting data quality [14].

Table 3. Data quality problems and classification.

Data Quality Problem	Type	Description
Single-source Problem	Instance Level	Misspelling
		Inconsistent values
		Data entry errors
		Redundancy Replicas
		Referential Reliability
	Schema Level	Poor schema designer
		Lack of integrity checks
		Uniqueness checks
Multi-source Problem	Instance Level	Overlapping, challenging and erratic timing, data, and aggregating
	Schema Level	Naming Clashes
		Mixed data models and schema design

Adapted from Sidi *et al.* [12].

dataset is to understand the dataset and its compatibility with the intended research work. Failure to do so will result in erratic inferences and incorrect analytics. Evaluating the data quality over intelligently developed metrics and developing matching transformation operations to solve the quality gaps helps ease the effort of data scientists for iterative mending of the machine learning algorithm to enhance the performance of the model. Due to these facts, it is imperative to design an algorithm or tool to reduce the execution time for data preparations. The data quality toolkit is the systematic measure of data quality to develop machine learning models. Presented in **Figure 2** is the general structure of the positional stand of the data quality toolkit as a part of the typical data science pipeline. Gupta *et al.* [19] reported that a data quality toolkit was developed to address four primary objectives:

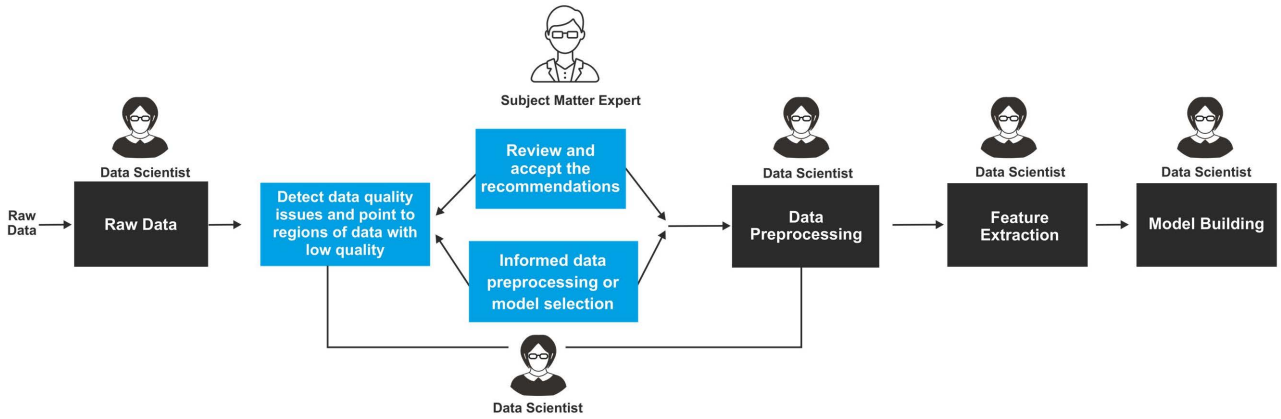

Figure 2. Positioning of data quality toolkit (Gupta *et al.* [19]).

1) Analyze data quality at step 0 of a data science development, such that the data scientist can make informed inferences at the later stages of model development.

2) Offer explanations for cases of low data quality by highlighting the region responsible for the low quality.

3) Suggest actions to enhance the data quality and provide a smooth way to implement the suggestions.

4) In the presence of an auto-generated report highlighting the operations history on the dataset for smooth reference to all the changes made to the original data.

The data toolkit can serve as a decision support system (DSS) to ensure robust decision-making processes. These decisions can involve selecting a model that identifies issues such as features dropping and noise irrelevant labels, determining the quality of data, and providing a response to the data acquisition procedure. The data quality toolkit contains a set of metrics that measure data problems serving as a pointer of the consumed data readiness for downstream machine learning tasks. The data quality dimensions are measured on a scale of 0 - 1, whereby a score specifies the absence of practical issues for the individual quality feature. The data quality toolkit comprises three key components of data remediation, data quality measurement, and data readiness report.

2.7. Data Remediation

The data remediation mechanisms can be invoked to enhance and adjust the data quality issues acknowledged by the individual quality metrics. Based on the type of quality measurement, this step may include a human in the loop process to integrate human review and response before data transformation. After using all the data remediation functions on tasks, this step writes the updated data to the system file. Data remediation also helps to confirm whether the user wants to include a new class tag for data in updated data or not [19].

2.8. Data Quality Measurement

Data quality measurement is critical and remains the most significant compo-

nent in the toolkit. A data quality measurement takes in the data and assesses the data quality based on different quality features. The quality scores measure quality measurement's interest as an actual number between 1 and 0. There is no issue when a score of 1 is recorded in the ingested data for that metric. A textual explanation is provided to help users interpret the score meaning. Also, recommendations are offered to protect the user's technique that can be applied to fix the issues acknowledged.

2.9. Data Readiness Report

The data readiness report is a shareable asset that is the one-stop-shop of the standard quality of the input data and a record of operations and remediation carried out [20]. It is fundamentally a single source for understanding the data quality and the series of transformations that has been undertaken. Over time data scientists spend much time tackling and exploring different data quality issues. The data readiness report is a complete report of all data properties and quality issues, such as the lineage of the data that illustrates how the data evolve. Furthermore, a data readiness report is more of a dynamic and evolving document that can be generated at any time during the interaction with the toolkit.

2.10. Data Quality Monitoring

Data quality monitoring is a framework that incessantly controls the data quality present in an information system through metric reports or periodically conducting data profiling. Data quality measurement is different from data quality assessment; the base is to evaluate a single data quality score from a range of observations. Data quality can be monitored through an application that monitors various information systems over some time. Presented in **Figure 3** is the structure of the data quality monitoring application; it consists of four main components, namely, 1) Metadata gathering and computation of data quality assessment results; 2) data quality depository; 3) Time-series data analysis; and 4) a user-friendly interface to monitor and track the data quality development of the observed system.

2.11. Data Quality Dimensions

Data quality dimension is a characteristic of information used for organizing information and data requirements. Over time, the most frequently mentioned data quality dimensions are completeness, accuracy, consistency, and timeliness, with different definitions from researchers. Wang and Strong [22] proposed different categories of data quality dimensions. **Figure 4** presents six categories of data quality dimensions, namely, accuracy, completeness, timeliness, constituency, uniqueness, and validity.

Completeness

Completeness can be defined as the degree to which data are of sufficient depth, breadth, and scope for the operation at hand [23].

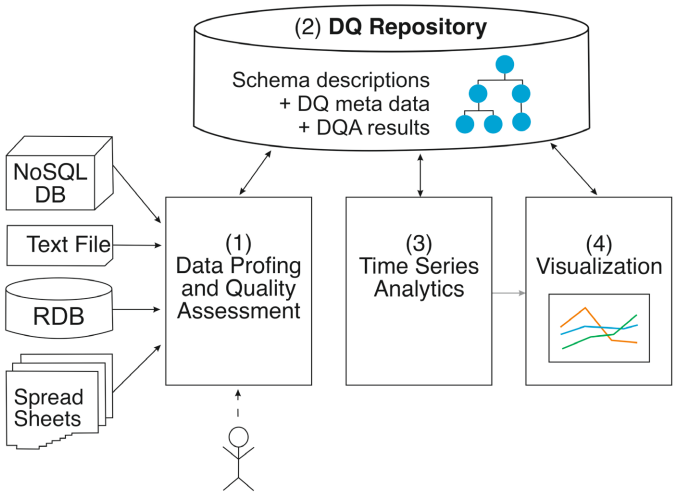

Figure 3. Data quality monitoring application [21].

Figure 4. Six data quality dimensions (Askham *et al.* [25]).

Accuracy

Accuracy is the degree to which data are error-free, reliable, and accurate [23]. It can be further defined as the closeness between a value v and v', deliberated as the correct representation of the real-time occurrence that v represents [11].

Timeliness

Timeliness shows how to present data for the job at hand. It is inspired by the fact that it is possible to have current data that are ineffective because they are late for a specific usage [11]. It can be further defined as the degree to which data

are satisfactorily updated for the operation at hand [22]. Timeliness is defined as the datedness associated with volatility and age; the age of information or an object is the measure of how old that piece of information or object is [24]. Volatility is a measure of information unpredictability with the incidence of change of the value for an entity characteristic.

Consistency

Consistency is about how data is achieved in a system with specified restrictions. Consistency, as addressed in a database system, is seen as an integrity restriction, and it is highly regarded as crucial data quality problem.

2.12. Proposals for Data Quality Dimensions

Different proposals report the classification of data quality dimensions instead of conservative data quality aspects [26]. The most widely considered proposals for data quality dimensions are highlighted below:

Proposal by Nauman

In the proposal of Nauman [27], data quality dimensions are defined as specific to incorporated web information systems where quality is considered an accumulated value of numerous data quality standards. Under this proposal, four different categories of data quality dimensions play a vital role in web information systems. These categories are intellectual, content-related, instantiation-related criteria, and technical. From the mentioned criteria, some criteria will be vital for a particular application domain. The two domains are stock information systems and search engines.

Proposal by Wang and Strong

In this proposal, data quality is determined by cross-examining data consumers. The data quality category for Wang and Strong proposal was intrinsic, accessibility, contextual, and representational data quality [22]. This proposal was used for many data quality studies to develop diverse data quality models and structures in different contexts.

Proposal by Bovee et al.

Proposal by Bovee *et al.* [24] considers data quality dimensions by examining the view of data consumers and developed a theoretical model consisting of four different features. This model was adapted from Wang and Strong [22], and the features have been compared with this proposal to assess data quality with all its essential dimensions or attributes that determine the quality in any field; this proposal has four different attributes, namely, relevance, integrity, interpretability, and accessibility [24]. The integrity attribute is classified into four sub-attributes completeness, existence, consistency, and accuracy.

2.13. Related Works

Decision-making in an institution or organization is centered on information and data retrieved through data analysis that offers different insights to generate an accurate and dependable process. The way data is essential in some institu-

tions is the same way its quality must be checked. Sidi *et al.* [12] focus on a review of data quality dimensions to be employed for a proposed framework that combines statistical methods and data mining to evaluate dependencies present between dimensions; this study further illustrates how retrieving knowledge can increase process quality.

Previous and existing studies agree that data quality is a cyclic process that must be continuously carried out. However, most data quality tools permit the assessment of data sources only at a certain point in time. Therefore, it is imperative to schedule and automate data by the user to support an excellent data evaluation. The quality of data monitoring ensures the assessment of applied data quality developments and compares various system states. Ehrlinger and Wöß [21] developed a data quality monitoring tool and investigated correct data quality metrics for continuous monitoring and to enable the development of the consistent technique of storing data quality assessment results.

Nowadays, companies and institutions have invested a lot in data, and most of their decisions are guided by data collected from the field or sales. Therefore, inaccurate data information will ultimately compromise the decision process. It is essential to verify the source, integrity, and quality of their data to ensure smooth decision making. Schelter *et al.* [28] proposed a system for automating the processes in verifying data quality. The verification system provides a declarative application program interface that merges quality limitations with user-defined authentication code and permits unit tests for data. The system further leverages the machine learning algorithms and incremental data quality validation on increasing datasets.

The validation of the quality of data is essential in present-day data-enabled applications. Unexpected behaviors in downstream services like machine learning models and production pipelines are recorded due to errors in the founding data. Over the years, unforeseen data quality problems have been resolved through manual and monotonous restoring processes volatilely. Growing usage of a large dataset must be ingested into non-relational stores such as data lakes. It is even worse when the features of the to-be ingested dataset change periodically, and field expertise is not available to define the data quality limitations. Redyuk *et al.* [29] proposed a data-centric technique that automates data quality assessment in scenarios painted earlier on. Here, the proposed technique does not require field expert to define limitations and rules; the technique is built such that the data self-adapt to periodic and temporal changes in the data features. The technique was improved and evaluated on five real-time datasets having both artificial and real generated errors.

Günther *et al.* [30] reported that not all organizations check data quality before any decision-making processes. Therefore, a simplified methodology was proposed to execute data quality evaluations and enhance the understandability of its results. This work was aimed at making data quality usable and accessible to small and medium-sized enterprises. The technique uses context-related

semi-structured and structured data as the input and employs a range of generic test criteria applicable to various fields and tasks. The methodology employed is assessed using data gathered from the manufacturing execution system and enterprise resource planning.

2.14. Big Data Quality

The growing importance of big data has introduced new challenges to data quality. Accordingly, various research studies have sought to identity the quality dimensions of big data. **Table 4** presents an overview of the big data quality classification.

The existing literature on various dimensions of the data quality management has been discussed in this section. The following section will briefly explain the methodology used in identifying the literature and criteria used for selecting specific papers for inclusion in this review.

3. Methodology

This paper is based on a limited and targeted narrative review of the existing literature on data quality. The purpose of the review is not an extensive survey of the existing literature, but a limited exploration of the current knowledge to understand various dimensions and practices of data quality that have a bearing on the development of automated data quality systems. Hence, all the steps of a systematic review of the literature [38] have not been adhered to in this study. Instead, the study is conducted using the qualitative paradigm and a narrative approach to the literature review [39].

A search on the Google scholar using the string "automated data quality" returns nearly 1250 results. However, all these results are discussions on proposals for developing or implementing automated data quality systems for specific industries or sectors. The string "automated data quality systems" returns only two results [40] [41]. Both these papers were also dealing with the use of automated data quality systems for specific purposes, rather than dealing with the system itself. Therefore, the search was extended to all the dimensions of data quality, its monitoring, and assurance. The results were analyzed from the perspective of their usefulness in developing automated data quality systems and the selection and inclusion in this study of specific papers was based on a subjective and qualitative assessment based on the experiences of the author [39].

4. Results

The literature review section captured the existing knowledge on various aspects of data quality. The review results also showed that there is no existing literature that addresses all the comprehensive knowledge necessary for developing the automated data quality systems. Different models are developed for data quality assessment and management, but the designing an appropriate model for automated data quality management is yet to take place.

Table 4. Big data quality classification.

Study	Architecture	Terminology
Big Data Pre-Processing: Closing the Data Quality Enforcement Loop [31]	Three Dimensions	Accuracy, Completeness, Consistency
A Hybrid Approach to Quality Evaluation Across Big Data Value Chain [32]	2 Types, 4 Dimension	Accuracy, Timeliness, Completeness, Consistency
Big Data Quality: A Survey [33]	4 Types 18 Dimensions	Accuracy, Timeliness, Consistency, Completeness Reputation, Relevancy, Accessibility, Quantity, Value-added, Believability Interpretability, Representational, Conciseness of representation, Consistency, Manipulability, Ease of understanding Access, Security
Context-Aware Data Quality Assessment for Big Data [34]	7 Dimensions	Accuracy, Completeness, Consistency, Distinctness, Precision, Timeliness, Volume
Data Quality in Big Data Processing: Issues, Solutions, and Open Problems [35]	4 Dimensions	Availability, Usability, Reliability, Relevance
Big Data Quality: A Quality Dimensions Evaluation [36]	3 Dimensions	Accuracy, Completeness, Consistency
Big Data, Big Data Quality Problem [37]	7 Dimensions	Accuracy, Precision, Completeness, Consistency, Timeliness, Lineage/Pedigree and Relevance

4.1. Data Quality Assessment

Assessment of data quality is a process of evaluating collected data to determine whether the program's objectives are met, and the data is the right type that has the quality to support the intended objectives. The results obtained from the data quality assessment will determine the relevancy, accessibility, completeness, accuracy, currency, and consistency of the data. Data quality is highly reliant on the underlying data reporting and management systems. To generate quality data, functional components must be available in all levels, ranging from data sources, gathering, encoding, transfer, data checking, reviewing, and processing. The data quality assessment tools are intended to 1) evaluate the sources of the data; 2) validate the collected data quality; 3) design strategic actions to enhance the system and data. The data quality assessment intends to evaluate the quality of gathered data, thereby aiming to a) ensure that the information retrieved reflects the authenticity on the field and has all the six properties of data quality dimension; b) assess the capacity of the dataset system by managing, capturing, reporting, and processing quality data; and lastly c) to design and implement measures to strengthen data gathering and management at all levels.

Components of data quality assessment

The components of data quality assessment are divided into three main groups, namely, a) agreement with data conversion and submission criteria; b) data authentication; and c) system evaluations.

Component 1: *Agreement with data conversion and submission criteria*: Providing an excellent database that offers complete and appropriate data to part-

ners and management is an important objective that enhances an informed and better decision-making process. This first component deals with the timelessness and completeness of the submitted data. Data reporting standards and requirements are pre-set and provided at all levels.

Component 2: *Data authentication*: In ensuring data authentication, templates, forms, and documents are employed to capture activities available in the database. Data verification helps confirm the consistency and accuracy of the data from different sources it was gathered from through cross-checking the reported data and information. This component will detect, track, and resolve discrepancies and errors in the database.

Component 3: *System evaluations*: Data quality is highly reliant on the systems put in place. Systems with solid backgrounds produce improved data quality. The system evaluation will address the following facts:

1) Abilities of the staff involved in gathering and management of data.

2) Technical aid and capacity building

3) Data gathering, management, and processing involve using patterns and paper-based forms, conversion and submission, data quality mechanism, reporting and application, recovery, and storage.

The data quality assessment is intended to help aid the management or researchers to understand the fundamental limitations and problems faced during data gathering, management, and processing, also to determine possible sources of error in data gathering, detect measures to enhance the capabilities of users involved in the process and reinforce data management at all levels.

4.2. Data Quality Management

To ensure the continued quality of data over time, improving data quality dimensions through an efficient process is vital. Owing to this reason, numerous researchers have worked to propose a model and methodologies for organized data quality management. Wang [42] proposed a model named Total Data Quality Management (TDQM). TDQM was intended to support the idea of data being a product. This idea will ensure the production of high data quality through duplication of the physical production of a high-quality product. The TDQM extends the total quality management framework that employs physical production. The approaches of TDQM start with a description of the production information. The information production (IP) has its features and requirements to achieve a high-quality data state. The second approach is the quality of information metrics developed and employed to measure the information product. The result obtained from the measurement is analyzed and used the pattern recognition, Pareto chart, and statistical process control. Time data quality management was developed to manage the quality of data present in the database, and present technologies involving big data analytics may limit the practice.

Another data quality management model is the Aim Quality model by Lee *et*

al. [43], which comprises of service and product performance model for information quality. This tool measures information quality and analyses the gap within the information quality to enhance the quality of information retrieved. The medium of information retrieval and evaluation in this method is a questionnaire. Further inferences are unraveled through statistical analysis, which involves identifying information quality problems.

5. Discussion

The critical factor that differentiates an automatic data quality system from the existing data quality models is that the former should obviate the need for human intervention and associated subjectivity, in ensuring data quality. All the identified dimensions of data quality should be attained on a real-time basis, through well-defined protocols of the automated system. This review shows the critical need for developing a comprehensive model for automated data quality systems, taking into account the already identified dimensions and objectives of data quality management.

Future Research

Technological advancement in data quality has enhanced various problem-solving platforms addressing real-life problems. Most previous works were centered on using structured data in applying data quality by researchers and organizations. There should be more research works employing unstructured data, its types, and big data technology in addressing some of the existing problems. Much discussion of data quality dimension is based on structured data in achieving high data quality, especially in data timeliness, completeness, accuracy, and consistency. The evaluation methods of these crucial dimensions would differ in the context of big data; big data produces huge data volume with high velocity and range of features. Data quality assessments, management, and methodologies are essential deliverables in data quality research work. Efficient data quality can be achieved by adopting appropriate data quality assessment and management models, as discussed previously.

6. Conclusions

This review answered questions related to data quality types, dimensions, assessments, and data quality management. Advancement in technology is making data in organizations cut across other domains, unlike being limited to the database. Numerous sources of data such as social media and websites are becoming crucial to organizations, researchers, and building relationships with their customers. This review discussed the expansion of data quality to support different areas in new technology. Six critical data quality dimensions were also highlighted in this paper; these dimensions are vital in data quality and must be accorded with the highest priority in evaluating and managing data quality. In handling data quality within an organization or by researchers, data quality

management must be adopted. Incorporating this management model and methodologies may require enhancement to support different data sources in unlabelled data.

Furthermore, data quality assessment is a vital tool for efficient data quality management. In this review, different components of data quality assessment to ensure data quality were highlighted. It was noted that the data quality of unlabelled data could be enhanced by engaging in more studies directed toward unstructured data.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Strong, D.M., Lee, Y.W. and Wang, R.Y. (1997) Data Quality in Context. *Communications of the ACM*, **40**, 103-110. https://doi.org/10.1145/253769.253804

[2] Lee, Y.W. and Strong, D.M. (2003) Knowing-Why about Data Processes and Data Quality. *Journal of Management Information Systems*, **20**, 13-39. https://doi.org/10.1080/07421222.2003.11045775

[3] Bosu, M.F. and MacDonell, S.G. (2013) Data Quality in Empirical Software Engineering: A Targeted Review. *Proceedings of the* 17*th International Conference on Evaluation and Assessment in Software Engineering*, Porto de Galinhas, 14-16 April 2013, 171-176. https://doi.org/10.1145/2460999.2461024

[4] Shirai, Y., Nichols, W. and Kasunic, M. (2014) Initial Evaluation of Data Quality in a TSP Software Engineering Project Data Repository. *Proceedings of the* 2014 *International Conference on Software and System Process*, Nanjing, 26-28 May 2014, 25-29. https://doi.org/10.1145/2600821.2600841

[5] Shepperd, M. (2011) Data Quality: Cinderella at the Software Metrics Ball? *Proceedings of the* 2*nd International Workshop on Emerging Trends in Software Metrics*, Waikiki, 24 May 2011, 1-4. https://doi.org/10.1145/1985374.1985376

[6] Qin, Y., Sheng, Q.Z., Falkner, N.J., Dustdar, S., Wang, H. and Vasilakos, A.V. (2016) When Things Matter: A Survey on Data-Centric Internet of Things. *Journal of Network and Computer Applications*, **64**, 137-153. https://doi.org/10.1016/j.jnca.2015.12.016

[7] Abu-Mostafa, Y.S., Magdon-Ismail, M. and Lin, H.T. (2012) Learning from Data. AML Book, Albany.

[8] Gudivada, V.N., Jothilakshmi, S. and Rao, D. (2015) Data Management Issues in Big Data Applications. *Proceedings of the* 1*st International Conference on Big Data*, *Small Data*, *Linked Data and Open Data*, Barcelona, 19-24 April 2015, 16-21.

[9] McNaull, J., Augusto, J.C., Mulvenna, M. and McCullagh, P. (2012) Data and Information Quality Issues in Ambient Assisted Living Systems. *Journal of Data and Information Quality (JDIQ)*, **4**, Article No. 4. https://doi.org/10.1145/2378016.2378020

[10] Batini, C., Cappiello, C., Francalanci, C. and Maurino, A. (2009) Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys* (*CSUR*), **41**, Article No. 16. https://doi.org/10.1145/1541880.1541883

[11] Batini, C. and Scannapieca, M. (2006) Data Quality: Concepts, Methodologies and

Techniques. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-33173-5

[12] Sidi, F., Panahy, P.H.S., Affendey, L.S., Jabar, M.A., Ibrahim, H. and Mustapha, A. (2012) Data Quality: A Survey of Data Quality Dimensions. *Proceedings of the* 2012 *International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, 13-15 March 2012, 300-304. https://doi.org/10.1109/InfRKM.2012.6204995

[13] Madnick, S.E., Wang, R.Y., Lee, Y.W. and Zhu, H. (2009) Overview and Framework for Data and Information Quality Research. *Journal of Data and Information Quality (JDIQ)*, **1**, Article No. 2. https://doi.org/10.1145/1515693.1516680

[14] Steve H. (2007) Data Quality Assessment. GPL Ghostscript.

[15] Wang, R.Y., Ziad, M., Lee, Y.W. and Wang, Y.R. (2001) Data Quality of the Kluwer International Series on Advances in Database Systems. Kluwer Academic Publishers, Dordrecht.

[16] Gertz, M., Özsu, M.T., Saake, G. and Sattler, K.U. (2004) Report on the Dagstuhl Seminar. *ACM SIGMOD Record*, **33**, 127-132. https://doi.org/10.1145/974121.974144

[17] Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S. and Munigala, V. (2020) Overview and Importance of Data Quality for Machine Learning Tasks. *Proceedings of the* 26*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event, 6-10 July 2020, 3561-3562. https://doi.org/10.1145/3394486.3406477

[18] Bandyopadhyay, B., Bandyopadhyay, S., Bedathur, S., Gupta, N., Mehta, S., Mujumdar, S. and Patel, H. (2021) 1st International Workshop on Data Assessment and Readiness for AI. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Delhi, 11 May 2021, 117-120. https://doi.org/10.1007/978-3-030-75015-2_12

[19] Gupta, N., Patel, H., Afzal, S., Panwar, N., Mittal, R.S., Guttula, S. and Saha, D. (2021) Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. arXiv preprint arXiv:2108.05935.

[20] Afzal, S., Rajmohan, C., Kesarwani, M., Mehta, S., and Patel, H. (2021) Data Readiness Report. *Proceedings of* 2021 *IEEE International Conference on Smart Data Services* (*SMDS*), Chicago, 5-10 September 2021, 42-51. https://doi.org/10.1109/SMDS53860.2021.00016

[21] Ehrlinger, L. and Wöß, W. (2017) Automated Data Quality Monitoring. *Proceedings of the* 22*nd MIT International Conference on Information Quality* (*ICIQ* 2017), Little Rock, 6-7 October 2017, Article No. 19.

[22] Wang, R.Y. and Strong, D.M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, **12**, 5-33. https://doi.org/10.1080/07421222.1996.11518099

[23] Fugini, M.G., Mecella, M., Plebani, P., Pernici, B. and Scannapieco, M. (2002) *Data Quality in Cooperative Web Information systems*. Kluwer Academic Publishers, Dordrecht.

[24] Bovee, M., Srivastava, R.P. and Mak, B. (2003) A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality. *International Journal of Intelligent Systems*, **18**, 51-74. https://doi.org/10.1002/int.10074

[25] Askham, N., Cook, D., Doyle, M., Fereday, H., Gibson, M., Landbeck, U. and Schwarzenbach, J. (2013) The Six Primary Dimensions for Data Quality Assessment. DAMA UK, Bristol, 432-435.

[26] Knight, S.A. and Burn, J. (2005) Developing a Framework for Assessing Informa-

tion Quality on the World Wide Web. *Informing Science*, **8**, 159-172. https://doi.org/10.28945/493

[27] Naumann, F. (2003) Quality-Driven Query Answering for Integrated Information Systems. Vol. 2261, Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45921-9

[28] Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F. and Grafberger, A. (2018) Automating Large-Scale Data Quality Verification. *Proceedings of the VLDB Endowment*, **11**, 1781-1794. https://doi.org/10.14778/3229863.3229867

[29] Redyuk, S., Kaoudi, Z., Markl, V. and Schelter, S. (2021) Automating Data Quality Validation for Dynamic Data Ingestion. *Proceedings of the* 24*th International Conference on Extending Database Technology* (*EDBT*), Nicosia, 23-26 March 2021, 61-72.

[30] Günther, L.C., Colangelo, E., Wiendahl, H.H. and Bauer, C. (2019) Data Quality Assessment for Improved Decision-Making: A Methodology for Small and Medium-Sized Enterprises. *Procedia Manufacturing*, **29**, 583-591. https://doi.org/10.1016/j.promfg.2019.02.114

[31] Taleb, I. and Serhani, M.A. (2017) Big Data Pre-Processing: Closing the Data Quality Enforcement Loop. *Proceedings of* 2017 *IEEE International Congress on Big Data* (*BigData Congress*), Honolulu, 25-30 June 2017, 498-501. https://doi.org/10.1109/BigDataCongress.2017.73

[32] Serhani, M.A., El Kassabi, H.T., Taleb, I. and Nujum, A. (2016) A Hybrid Approach to Quality Evaluation Across Big Data Value Chain. *Proceedings of* 2016 *IEEE International Congress on Big Data* (*BigData Congress*), Francisco, 27 June-2 July 2016, 418-425. https://doi.org/10.1109/BigDataCongress.2016.65

[33] Taleb, I., Serhani, M.A. and Dssouli, R. (2018) Big Data Quality: A Survey. *Proceedings of* 2018 *IEEE International Congress on Big Data* (*BigData Congress*), Francisco, 2-7 July 2018, 166-173. https://doi.org/10.1109/BigDataCongress.2018.00029

[34] Ardagna, D., Cappiello, C., Samá, W. and Vitali, M. (2018) Context-Aware Data Quality Assessment for Big Data. *Future Generation Computer Systems*, **89**, 548-562. https://doi.org/10.1016/j.future.2018.07.014

[35] Zhang, P., Xiong, F., Gao, J. and Wang, J. (2017) Data Quality in Big Data Processing: Issues, Solutions, and Open Problems. *Proceedings of* 2017 *IEEE SmartWorld*, *Ubiquitous Intelligence* & *Computing*, *Advanced* & *Trusted Computed*, *Scalable Computing & Communications*, *Cloud* & *Big Data Computing*, *Internet of People and Smart City Innovation* (SmartWorld/SCALCOM/UIC/ATC/CBDCom /IOP/SCI), Francisco, 4-8 August 2017, 1-7. https://doi.org/10.1109/UIC-ATC.2017.8397554

[36] Taleb, I., El Kassabi, H.T., Serhani, M.A., Dssouli, R. and Bouhaddioui, C. (2016) Big Data Quality: A Quality Dimensions Evaluation. 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, 18-21 July 2016, 759-765. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0122

[37] Becker, D., King, T.D. and McMullen, B. (2015) Big Data, Big Data Quality Problem. *Proceedings of* 2015 *IEEE International Conference on Big Data* (*Big Data*), Santa Clara, 29 October-1 November 2015, 2644-2653. https://doi.org/10.1109/BigData.2015.7364064

[38] Massaro, M., Dumay, J. and Guthrie, J. (2016) On the Shoulders of Giants: Under-

taking a Structured Literature Review in Accounting. *Accounting, Auditing & Accountability Journal*, **29**, 767-801. https://doi.org/10.1108/AAAJ-01-2015-1939

[39] Rother, E.T. (2007) Systematic Literature Review X Narrative Review. *Acta Paulista de Enfermagem*, **20**, v-vi. https://doi.org/10.1590/S0103-21002007000200001

[40] Rukat, T., Dustin, L., Sebastian, S. and Felix, B. (2020) Towards Automated Data Quality Management for Machine Learning.

[41] Fernandez, M.C. (2017) Exploring Deep Computing in CMS for Automated Data Validation in DQM. No. CERN-STUDENTS-Note-2017-185.

[42] Wang, R.Y. (1998) A Product Perspective on Total Data Quality Management. *Communications of the ACM*, **41**, 58-65. https://doi.org/10.1145/269012.269022

[43] Lee, Y.W., Strong, D.M., Kahn, B.K. and Wang, R.Y. (2002) AIMQ: A Methodology for Information Quality Assessment. *Information & Management*, **40**, 133-146. https://doi.org/10.1016/S0378-7206(02)00043-5