big data and cognitive computing

MDPI

Article

An Advanced Big Data Quality Framework Based on Weighted Metrics

Widad Elouataoui [1,*], Imane El Alaoui [2], Saida El Mendili [1] and Youssef Gahi [1]

[1] Laboratory of Engineering Sciences, National School of Applied Sciences, Ibn Tofail University, Kenitra 14000, Morocco

[2] Telecommunications Systems and Decision Engineering Laboratory, Ibn Tofail University, Kenitra 14000, Morocco

* Correspondence: widad.elouataoui@uit.ac.ma

Abstract: While big data benefits are numerous, the use of big data requires, however, addressing new challenges related to data processing, data security, and especially degradation of data quality. Despite the increased importance of data quality for big data, data quality measurement is actually limited to few metrics. Indeed, while more than 50 data quality dimensions have been defined in the literature, the number of measured dimensions is limited to 11 dimensions. Therefore, this paper aims to extend the measured dimensions by defining four new data quality metrics: Integrity, Accessibility, Ease of manipulation, and Security. Thus, we propose a comprehensive Big Data Quality Assessment Framework based on 12 metrics: Completeness, Timeliness, Volatility, Uniqueness, Conformity, Consistency, Ease of manipulation, Relevancy, Readability, Security, Accessibility, and Integrity. In addition, to ensure accurate data quality assessment, we apply data weights at three data unit levels: data fields, quality metrics, and quality aspects. Furthermore, we define and measure five quality aspects to provide a macro-view of data quality. Finally, an experiment is performed to implement the defined measures. The results show that the suggested methodology allows a more exhaustive and accurate big data quality assessment, with a more extensive methodology defining a weighted quality score based on 12 metrics and achieving a best quality model score of 9/10.

Keywords: big data quality metrics; weighted big data quality; weighted metrics; big data quality aspects

check for updates

Citation: Elouataoui, W.; El Alaoui, I.; El Mendili, S.; Gahi, Y. An Advanced Big Data Quality Framework Based on Weighted Metrics. *Big Data Cogn. Comput.* **2022**, *6*, 153. https://doi.org/10.3390/bdcc6040153

Academic Editor: Salvador García López

Received: 30 October 2022
Accepted: 7 December 2022
Published: 9 December 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

In the last decade, data analytics has shown great potential for supporting organizations to improve their business and to get closer to their customers. Indeed, data are considered as one of the most valuable resources that industries rely on for decision-making. However, the benefits of data could not be reached if the data are of low quality. Using unstructured and inaccurate data may bias data analytics and lead managers to make wrong decisions. Therefore, data quality has gained wide attention from both academics and organizations, and several approaches have been suggested and adopted in this regard [1–3]. With the rise of Big Data, ensuring data quality has become more challenging. Indeed, managing big data involves handling heterogeneous and messy data that traditional data quality tools could not manage. It is worth mentioning that the big data issues are not only related to data volume but also to other big data properties, known as the Big Data V's. Over the last few years, big data characteristics have risen to more than 50 V's [4]. Figure 1 presents the most common ones, the 7V's of Big Data.

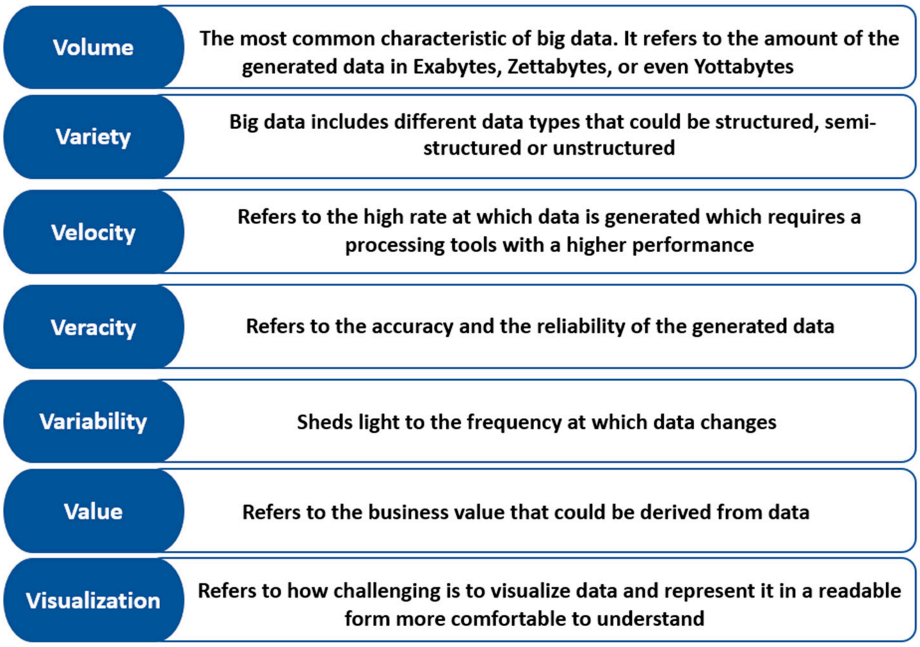
Volume
The most common characteristic of big data. It refers to the amount of the generated data in Exabytes, Zettabytes, or even Yottabytes
Variety
Big data includes different data types that could be structured, semi-structured or unstructured
Velocity
Refers to the high rate at which data is generated which requires a processing tools with a higher performance
Veracity
Refers to the accuracy and the reliability of the generated data
Variability
Sheds light to the frequency at which data changes
Value
Refers to the business value that could be derived from data
Visualization
Refers to how challenging is to visualize data and represent it in a readable form more comfortable to understand

Figure 1. The 7V's of Big Data.

Thus, because of big data characteristics, ensuring a high big data quality is considered as one of the most challenging and critical phases of the Big Data Value Chain (BDVC) [5,6]. In a recent review of big data quality challenges [7], the authors have shown that the leading big data quality challenges relate to data quality measurement, data cleaning, and data quality rules. Therefore, big data quality issues have been addressed by both academics and professionals in recent studies, suggesting new methodologies to enhance the quality of big data [8–10]. However, despite its high importance, there has not been much work regarding big data quality assessment. Most of the existing approaches apply the traditional data quality assessment measures to big data and do not consider the particular issues related to big data. Indeed, the number of defined dimensions is estimated to have reached more than 50 dimensions [11]. However, the measured dimensions are limited to 11 metrics. This large gap between the defined and measured dimensions points out a great need to implement the quality dimensions and express them in a more tangible form to allow them to be effectively used. In addition, despite their high impact on the accuracy of measurements, data weights are less considered by the existing quality assessment frameworks, which challenges the precision and the correctness of the performed measures. To address the raised issues, we aim to enhance big data quality assessment with four main contributions:

- Extending the number of the measured dimensions by defining four new data quality metrics: Integrity, Accessibility, Ease of manipulation, and Security.
- Defining a comprehensive Big Data Quality Assessment Framework based on 12 metrics: Completeness, Timeliness, Volatility, Uniqueness, Conformity, Consistency, Ease of manipulation, Relevancy, Readability, Security, Accessibility, and Integrity.
- Improving the precision of the measures by considering the weights of data fields, data quality metrics, and aspects.
- Providing a macro-view of Big Data Quality by defining and measuring five quality aspects: Reliability, Availability, Usability, Pertinence, and Validity.

The remainder of this paper is organized as follows. Section 2 highlights the importance of considering the big data quality metrics (BDQM) in big data context. Section 3 describes the research approach performed for the literature review. In the Section 4, we review the most relevant studies addressing data quality measurement. Then, we introduce in the Section 5 big data quality aspects and define the selected big data quality metrics. We also highlight how data weights could be applied to data quality assessment in this section. Section 6 presents an implementation of the suggested assessment framework and discusses the observed results by comparing existing frameworks. Finally, we conclude by highlighting the main findings of this paper and discuss some future research outlooks.

2. The Big Importance of BDQM for Big Data Use Cases

Nowadays, considering data quality metrics is a standard of data processing in the different contexts and domain applications of big data. With the emergence of Big Data, data quality metrics are gaining more attention than ever. Indeed, enhancing data quality has always been a priority in big data environments due to the poor quality of big data. Some of the benefits gained from considering the BDQM are listed below:

- Provide measurable insights into data quality: Assessing data quality metrics provides excellent insights into the health of data and, thus, allows data managers to better anticipate and address quality issues. In a big data environment where data is exposed to continuous changes, such as preprocessing and analysis, it is essential to diagnose the quality state of data using BDQM to assess the impact of the performed changes on data quality [12].
- Lead to further data quality enhancement: Assessing BDQM allows locating data quality weaknesses and addressing them effectively. A recent study [13] investigating the impact of data quality on business has shown that less than 50% of companies are confident in their internal data quality, and only 15% are satisfied with the data provided by third parties. These numbers reveal the criticality of the issue of data quality, especially when companies rely mainly on data to make their business decisions.
- Identify data quality defects: Sometimes, the interest in measuring metrics is not necessarily to address them but to be aware of them. Indeed, data may contain hidden weaknesses that data managers do not even realize. This may severely impact the rest of the data value chain, as fuzzy data lead to fuzzy and inaccurate results. A study [14] performed to assess the impact of BDQM using sentiment analytical approaches for big data has shown that ignoring big data quality metrics entirely biases the prediction results with a sentiment analytical accuracy of 32.40%; thus, this study concludes that not considering the quality metrics for big data has a powerful influence on the predicted results.
- Separate high-quality data from low-quality data: This will allow the processing of data differently depending on the quality level. This approach is time- and cost-effective, especially in big data environments where data preparation is the most complex and critical part of data processing.

Research Questions:

Given the importance of considering data quality metrics in big data context, this study aims to contribute to the ongoing discussion regarding big data quality assessment by addressing the following research questions:

- What are the data quality metrics defined in the literature?
- Given the large gap between the defined dimensions and the measured dimensions, do existing metrics really cover all of the data quality aspects?
- Are there any big data quality requirements that need to be captured in terms of quality metrics?
- Are there any new metrics introduced in the literature that fit big data requirements?
- How could big data quality assessment be enhanced in terms of accuracy and precision?
- How could big data quality assessment be enhanced in terms of exhaustivity?

In the next section, we describe the research methodology performed for the literature review to address these research questions.

3. Research Methodology

A systematic literature review was conducted to capture and synthesize the relevant and available studies addressing data quality measurement. This literature review was performed following the guidelines stated in [15], where the authors proposed a three-step literature review methodology as shown in Figure 2.

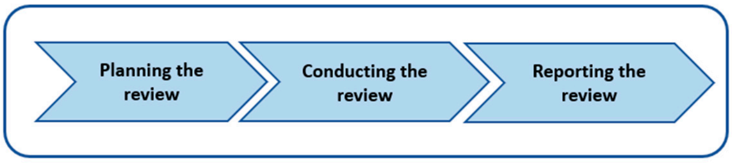

Figure 2. Literature Review Research Methodology.

3.1. Planning the Review

This first step consists of preparing a review methodology and identifying the aim and the scope of the literature research.

Aim and Scope

To address the research questions raised in the previous section, we selected in this review two main kinds of contributions:

- Studies that have defined new data quality metrics or used existing metrics in a big data context.
- Studies that have defined new quality metrics in non-big data context (small datasets).

3.2. Conducting the Review

In this step, we performed the review by searching and selecting studies according to the defined scope.

3.2.1. Research Keywords

A primary search was conducted by first using the following generic expressions: «Weighted Data Quality», «Data Quality Metrics», «Data Quality Measurement», and «Data Quality Aspects». Then, to capture studies about big data, specific expressions, such as «Big Data Quality Metrics» and «Big Data Quality Measures», were used. First, only papers with titles corresponding to the keywords mentioned above were selected. Then, the abstracts were reviewed, and irrelevant papers were excluded. This primary search yielded 47 articles. Table 1 shows the number of obtained articles by digital library and keyword. Most of articles are duplicated (found twice) in Research gate and Scopus Libraries.

Table 1. Number of obtained articles by digital library and keyword.

Keywords/Digital Libraries	Springer	Scopus	IEEE Xplore	Research Gate	Science Direct
"Data Quality Metrics"	8	9	7	21	2
"Data Quality Measurement"	2	3	2	7	1
"Data Quality Aspects"	2	5	0	5	1
"Big Data Quality Metrics"	5	4	5	10	2
"Big Data Quality Measures"	1	1	0	3	0
"Weighted Data Quality"	1	0	0	1	0
Total	19	22	14	47	6

3.2.2. Digital Libraries

The search was limited to articles published in journals and conference proceedings and was conducted using the following digital libraries:

- Springer (http://www.springer.com/gp/ (accessed on 23 September 2021))
- IEEE Xplore (http://ieeexplore.org/ (accessed on 23 September 2021))

- Scopus (https://www.scopus.com/ (accessed on 23 September 2021))
- Science Direct (http://www.sciencedirect.com/ (accessed on 23 September 2021))
- Research Gate (https://www.researchgate.net/ (accessed on 23 September 2021))

3.2.3. Inclusion and Exclusion Criteria

After a literature search, the next step is narrowing down the papers regarding pertinence, availability, and contents. For this, a diagonal reading was performed on the captured studies by filtering out irrelevant studies based on the following inclusion criteria:

- Defining new data quality metrics or using the existing quality metrics in a big data context.
- Available in the digital libraries.
- Recent (from 2015 to 2022).
- Written in English.

A total of 32 articles were selected, following a deep and detailed analysis. Further, a deeper search was performed on the references of the selected studies. This led to the selection of three other papers considered as pertinent to the scope of our research. Then, the selected articles were thoroughly read and carefully examined, and 17 studies were deemed relevant to the scope of our research. Finally, the articles' descriptive details were checked and filed in a Zotero database. Figure 3 represents the methodology followed for the literature search.

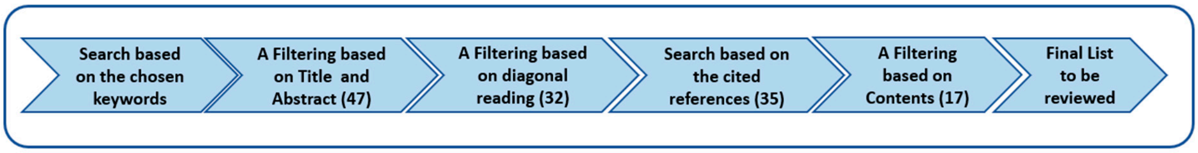
Search based on the chosen keywords
A Filtering based on Title and Abstract (47)
A Filtering based on diagonal reading (32)
Search based on the cited references (35)
A Filtering based on Contents (17)
Final List to be reviewed

Figure 3. Research Methodology.

3.3. Reporting the Review

This last step consists of reporting the main findings of the literature review and making conclusions. Figure 4 shows the number of selected papers corresponding to each type of content. This literature review shows a significant lack of works defining new metrics that fit big data requirements, which motivates us to conduct a deep analysis of the current state of the art to frame the need and make a significant contribution in this regard. The following section reviews the 17 papers selected for our study and highlights the main findings of this literature review.

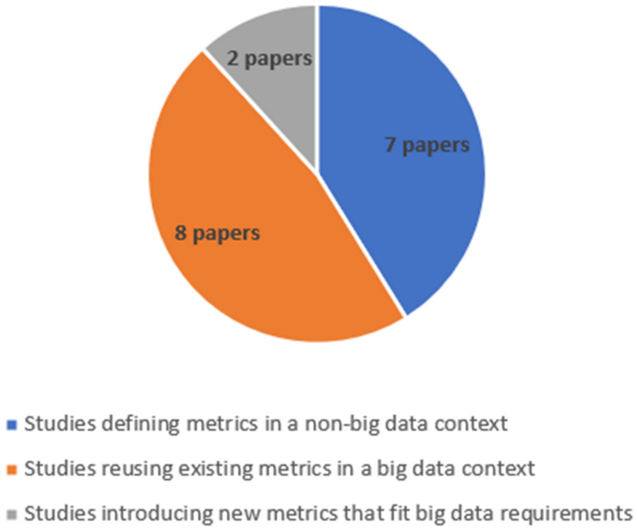

Figure 4. Number of papers corresponding to each type of content.

4. Related Work

Data quality assessment has long been addressed in the literature. Indeed, the first paper introducing data quality assessment dates back to 1998 [16], when Wang showed an analogy between data quality and product quality and defined a cyclical methodology called TDQM that allows a continuous improvement of data quality. Likewise, Lee et al. introduced in [17] a new method to assess information quality called AIMQ, which uses a questionnaire to measure information quality dimensions. Later, several data quality assessment methodologies were introduced in the literature, such as DQA, DWQ, TIQM, and ISTAT. All of these methodologies have added a significant value to data quality assessment. However, these works have focused more on the methodology used for data quality assessment and have not suggested detailed data quality measurements. Later, specific research studies were conducted to develop a more explicit definition for data quality metrics. Thus, dimensions, such as completeness, uniqueness, and consistency, were implemented and measured in several studies [18–20]. With the emergence of big data, novel data quality assessment approaches were introduced, and new data quality dimensions were defined, such as ease of manipulation, storage penalty, and normalization [14]. Thus, to review all the available research that has contributed to the ongoing discussion on data quality measurement, we conducted a literature review following the research methodology presented in the previous section. A total of 17 studies were selected for our review. These studies are shown in Table 2 with the following information: year of publication, the main idea of the study, whether the study is about big data, the defined metrics, whether new metrics are defined in the study, and the techniques used for the measurement.

Table 2. Approaches addressing data quality metrics.

Ref	Year	Main Idea	For Big Data	The Defined Dimensions	New Metrics	Techniques Used
[2]	2022	The authors defined a normalized double entropy (NDE) method to assess image data quality using probability entropy and distance entropy. This work focuses on healthcare data quality and data redundancy.	No	Uniqueness	No	Normalizing probability entropy and distance entropy
[21]	2022	The authors suggested a novel data quality assessment framework based on image classification. The data are represented as pixel prototypes, and disturbed entropy is then measured to assess data redundancy.	No	Uniqueness	No	Disturbed entropy
[22]	2021	Taleb et al. defined a framework to manage the quality of big data that consists of storing valuable information, including project parameters and requirements, data quality rules proposals, and data quality profiling. The framework is applied to the whole big data value chain. It was implemented and evaluated.	Yes	Generic	No	Used the definitions of Completeness, Consistency, Accuracy, and Uniqueness. Used a sampling and profiling algorithm.
[23]	2021	The authors implemented a machine learning model for data quality prediction. The suggested model consists of 3 steps: data noise detection, data noise impact assessment based on Generative Mixture Methods, and, finally, data quality prediction using the sequential learning of a deep-learning network.	Yes	Data Quality Rating functions (DQRi)	No	Used Recurrent Neural Network and Long Short-Term Memory techniques.

Table 2. *Cont.*

Ref	Year	Main Idea	For Big Data	The Defined Dimensions	New Metrics	Techniques Used
[24]	2019	The authors presented some data quality measuring approaches, such as laissez-faire, reactive, and proactive approaches. Also, they have described the different steps of each approach allowing to improve the quality of data.	No	Generic	No	Presented some data quality measuring approaches, such as laissez-faire, reactive, and proactive approaches
[25]	2019	The authors presented a framework for data quality measurement based on rule-based measurement. The suggested framework allows the handling of uncertainty. It was implemented using survey data.	No	Generic	Yes	Provided formulas for rule-based data quality measurement
[26]	2019	Goutam et al. suggested a big data accuracy assessment tool. The suggested model consists of comparing datasets to choose the optimal one using record linkage and word embeddings.	Yes	Accuracy	No	Word embeddings and record linkage, K-NN, Logistic Regression, and Decision Trees
[27]	2019	The authors suggested a data quality profiling model for big data based on the following modules: sampling, exploratory quality profiling, profiling, quality profile repository, and data quality profile.	Yes	Generic	No	Sampling and profiling algorithms
[28]	2018	The authors presented the requirements that should fulfill the metrics: Interval-Scaled Values, Existence of Minimum and Maximum, Quality of the Configuration Parameters, Efficiency of the Metric, and Sound Aggregation of the Metric Values. In addition, the authors showed the applicability of these requirements by evaluating five data quality metrics.	Yes	Timeliness, Completeness, Reliability, Correctness, and Consistency	Yes	Defined the requirements without formulas. Used the definitions of Timeliness, Completeness, Reliability, Correctness, and Consistency
[18]	2018	The authors developed an environment in which users can interactively customize data quality metrics to meet the requirements.	No	Completeness, Validity, Plausibility, Time Interval Metrics, and Uniqueness	No	Provided a generic definition of quality metrics. Defined Completeness, Validity, Plausibility, Time Interval Metrics, and Uniqueness.
[29]	2018	The authors highlighted the meanings of domain model, metrics, and weights of metrics. In this study, the authors built a system that allows users to seek tweets using a keyword search.	Yes	Readability, Completeness, and Usefulness	No	Used Readability, Completeness, and Usefulness for big data.
[19]	2017	The authors defined new data quality metrics that consider data weights (the importance of the information contained within the data). The suggested metrics were also implemented.	No	Completeness, Relevancy, Accuracy, Timeliness, and Consistency	Yes	Provided formulas to measure weighted Completeness, Relevancy, Accuracy, Timeliness, and Consistency.

Table 2. *Cont.*

Ref	Year	Main Idea	For Big Data	The Defined Dimensions	New Metrics	Techniques Used
[30]	2017	The authors reviewed some of the existing data quality approaches. They classified them based on three criteria: the degree of heterogeneity, the scope of the assessment approach, and the techniques used.	No	Completeness, Semantic Consistency, Structural Consistency, and Uniqueness	No	Used the definitions of Completeness, Semantic Consistency, Structural Consistency, and Uniqueness.
[31]	2016	The authors suggested a data quality assessment approach based on data proofing and sampling, allowing the optimization of processing time. The suggested approach comprises data quality evaluation using data profiling, data sampling, and data quality analysis.	Yes	Accuracy, Completeness, and Consistency	No	Used the definitions of Accuracy, Completeness, and Consistency. Used a sampling and profiling algorithm.
[32]	2016	The authors proposed a novel approach that combines data-driven and process-driven quality assessments throughout the whole data chain value. For each phase of the suggested approach, the authors defined the quality metrics that should be considered.	Yes	Timeliness, Currency, Volatility, Accuracy, Completeness, and Consistency	No	Used the definitions of Timeliness, Currency, Volatility, Accuracy, Completeness, and Consistency
[33]	2015	The authors highlighted how quality dimensions can be defined in a big data context. In addition, they showed that there are many quality notions that can be applied depending on the data types, which should be considered.	Yes	Redundancy, Consistency, Freshness, Accuracy, Copying, Spread, Completeness, and Trustworthiness	Yes	Defined Redundancy, Consistency, Freshness, Accuracy, Copying, Spread, Completeness, and Trustworthiness
[7]	2015	This paper aims to present the data quality challenges raised by the emergence of big data. The authors also highlighted various activities and components of data quality management, such as measuring metrics, data profiling, sampling, and data quality rules.	Yes	Generic	No	Provided a generic definition of quality metrics.

Based on the above research, a review of the existing data quality assessment approaches leads us to make the following points:

- In a big data context, ensuring data quality requires a specific and in-depth study that could not be limited to a few metrics. Indeed, while the number of metrics considered by the existing approaches does not exceed 11 metrics, more than 50 dimensions have been defined in the literature.
- Despite the high impact of data weights on the accuracy of measurements, most existing studies do not consider the weight of data elements when measuring data quality.
- Even if some studies have proposed classifications of the data quality dimensions, no studies have measured the quality aspects.

To overcome the raised issues, this paper aims to extend the measured dimensions by defining new four data quality metrics: Integrity, Accessibility, Ease of manipulation, and Security. Thus, we propose a comprehensive Big Data Quality Assessment Framework based on 12 metrics: Completeness, Timeliness, Volatility, Uniqueness, Conformity, Consistency, Ease of manipulation, Relevancy, Readability, Security, Accessibility, and Integrity.

Moreover, to provide a macro-view of data quality, we group the measured metrics into five quality aspects, namely Reliability, Availability, Usability, Validity, and Pertinence, which allows an easy understanding of data quality. Furthermore, to enhance the accuracy and precision of the quality measurements, data quality aspects and metrics are assessed based on the data weights applied to the data fields, the quality metrics, and the quality aspects. In the next section, we define the data quality aspects considered in this study. In addition, we describe and provide the measures for 12 quality metrics. Then, we explain the concept of weighted data quality and highlight how weighted quality could be applied at several levels.

5. Big Data Quality Aspects and Metrics

5.1. Big Data Quality Aspects

To easily understand and interpret data quality, the quality metrics could be grouped into specific categories called data quality aspects, which encapsulate metrics with shared traits. The data quality aspects add an abstraction layer to the quality metrics and, thus, provide more relevant and understandable insights about data quality which could support data managers during data analysis. Many classifications of the quality metrics have been suggested in the literature [34,35]. One of the most common classifications was introduced in [36], where the authors grouped the quality metrics into four categories:

- Contextual: Refers to the context of use of data and points to the extent to which data are pertinent to the intended use.
- Representational: Data should be represented in a readable and understandable format.
- Intrinsic: Implies that data values are accurate and in conformance with real world values.
- Accessibility: Data should be available and reachable in an appropriate way by data consumers.

Different criteria could be used for classifying and grouping the quality metrics, such as the metrics' nature, the metrics' meaning, and even the context of the study. As shown in Figure 5, we define in this paper 12 quality metrics that we gather into 5 quality aspects, namely Pertinence, Reliability, Validity, Availability, and Usability. We define these quality aspects as follows:

- Reliability: Refers to the trustworthiness and credibility aspect of data
- Availability: Refers to the accessibility and shareability of data while maintaining the appropriate level of data security.
- Usability: Refers to the relevancy and the ease of use of data.
- Validity: Assures that data conform to a specific format and comply with the defined business rules.
- Pertinence: Refers to what make data appropriate and suitable for the context of use.

It is worth mentioning that data quality metrics are related to and dependent on each other. Indeed, data quality dimensions could not be enhanced separately as improving a specific quality metric may negatively impact other metrics. For example, completing all data field values to enhance data completeness may lower data relevancy, as not all the stored data would be pertinent and relevant for the intended use.

Even if many classifications have been suggested in the literature, no study has measured the quality aspects. Indeed, the classification of metrics is generally used to highlight the particular properties of the metrics or to define the general aspects of data quality, and not for measurement or assessment purposes. Thus, we assess the quality aspects presented in Figure 5 using a weighted average of the metrics related to each aspect. In this paper, average factors vary depending on the relevancy of each metric. In the next part, we define the quality metrics associated with the abovementioned aspects. In addition, quality measures are suggested for the metrics.

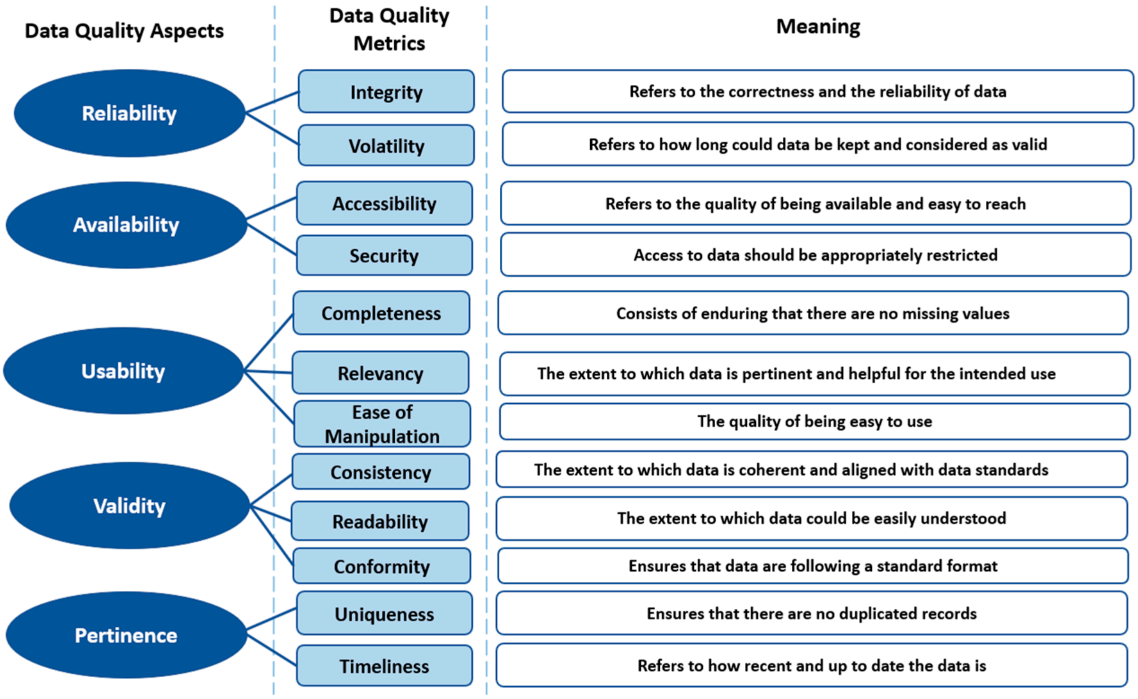
Data Quality Aspects
Data Quality Metrics
Meaning
Reliability
Integrity
Refers to the correctness and the reliability of data
Volatility
Refers to how long could data be kept and considered as valid
Availability
Accessibility
Refers to the quality of being available and easy to reach
Security
Access to data should be appropriately restricted
Usability
Completeness
Consists of enduring that there are no missing values
Relevancy
The extent to which data is pertinent and helpful for the intended use
Ease of Manipulation
The quality of being easy to use
Validity
Consistency
The extent to which data is coherent and aligned with data standards
Readability
The extent to which data could be easily understood
Conformity
Ensures that data are following a standard format
Pertinence
Uniqueness
Ensures that there are no duplicated records
Timeliness
Refers to how recent and up to date the data is

Figure 5. Data Quality Aspects and Metrics.

5.2. Big Data Quality Metrics

Data quality could be defined in terms of its properties called data quality metrics. Arolfo et al. in [29] described a data quality metric as a function that maps a dimension to a value between 0 and 1 and measures the data quality aspect and piece associated with this quality dimension. While it is true that most of the existing quality metrics remain valid for big data, new metrics specific to big data need to be defined. Indeed, with the emergence of big data, several notions of data quality have been challenged, such as the applicability of existing metrics, the performance of assessment tools, and the accuracy of measurements. Thus, we define in this section 12 quality metrics applicable in a big data context and show how big data characteristics impact these metrics. In addition, we suggest the measures for the 12 metrics: Completeness, Timeliness, Volatility, Uniqueness, Conformity, Consistency, Ease of manipulation, Relevancy, Readability, Security, Accessibility, and Integrity.

5.2.1. Completeness

In big data environments, the collected raw data are usually incomplete and lack contextual information. Thus, data completeness is one of the most crucial criteria when assessing data quality. Data completeness could be defined as the extent to which the data are sufficiently complete and fulfill the needed information [37]. Data completeness measurement is usually related to the number of missing values. However, data completeness could also be measured at the attribute level, mainly when some mandatory fields exist. Moreover, completeness may be measured horizontally at the row level. This metric is more relevant for reference data, such as countries and currencies. In this paper, we define completeness as the ratio of non-missing values.

$$\text{Completeness}(\%) = \frac{\text{Number of non empty values}}{\text{Total values}} \times 100$$

5.2.2. Uniqueness

Large-scale datasets are usually redundant since the data are gathered from multiple sources; therefore, the same information can be recorded more than once in a different format. Data uniqueness is required before the analytical phase because duplicated records corrupt the analytical results. Data uniqueness could be defined as the ratio of non-duplicated values [38].

$$\text{Uniqueness}(\%) = \frac{\text{Number of unique rows}}{\text{Total rows}} \times 100$$

5.2.3. Consistency

Because data are collected from different data sources, the captured data in big data environments are usually inconsistent and may contain some anomalies. Thus, converting the raw data into a structured and consistent format is a challenging task when handling big data. Consistent data could be defined as data presented in the same structure and coherent with data schemas and standards. In this paper, the measure of consistency is based on the data types previously defined. Hence, we define consistency as the ratio of values that comply with the expected data type to total values.

$$\text{Consistency}(\%) = \frac{\text{Number of values with consistent types}}{\text{Total values}} \times 100$$

5.2.4. Conformity

Data invalidity issues in big data systems are not just about data types. Indeed, data formats are also heterogeneous as each source has different standards and rules. A typical data conformity problem that could be faced is the Date and Time fields that could have different formats depending on the data source. In addition, as data are unstructured, fields having a specific format, such as email, phone number, and postal code, may not necessarily be correctly filled. Thus, we define conformity as the extent to which data respect the rules and the constraints of their environment. Each field's regular expression should be specified as it represents the field's pattern with a specific syntax. Hence, we define conformity as the ratio of values that comply with the prescribed rules to total values.

$$\text{Conformity}(\%) = \frac{\text{Number of values with consistent format}}{\text{Total values}} \times 100$$

5.2.5. Timeliness

One of the most common characteristics of big data is variability, which refers to the high frequency at which data are updated as new information becomes available. Ensuring that data are up to date in such an environment is very important since outdated data may bias data analysis. The Timeliness metric (currency or freshness) measures recent data and describes real-world values. Timeliness measurement is based on how long the data have been recorded and could be defined as the delay between the current date and the last modification date.

$$\text{Timeliness}(\%) = \frac{\text{Current Date} - \text{Last Modification Date}}{\text{Current Date} - \text{Creation Date}} \times 100$$

5.2.6. Volatility

Before being defined as a data quality metric, volatility was first introduced as a big data characteristic referring to the duration of data usefulness and relevancy. Unlike timeliness, volatility has nothing to do with preprocessing tasks and is somewhat related to data nature. Indeed, processing big data becomes more challenging when data are unstable and continuously updated. Thus, volatility as a quality metric refers to how long data could be kept and considered valid. Volatility could be defined as the delay between the storage and modification dates.

$$\text{Volatility}(\%) = \frac{\text{CreationDate} - \text{Modification Date}}{\text{Current Date} - \text{Creation Date}} \times 100$$

5.2.7. Readability

Data validity is not limited to data format but refers to data semantics as well. Indeed, raw data may contain misspelled words or even nonsense words, especially when the database is overwhelmed by human data entries, as in the case of social media data. Semantic issues could also be related to non-digital data, such as the information contained within pictures and audio. Data readability could be defined as the ability to process and extract information contained within the data. Hence, we define readability as the ratio of processed and non-misspelled values.

$$\text{Readability}(\%) = \frac{\text{Number of processed and non misspelled values}}{\text{Total values}} \times 100$$

5.2.8. Ease of Manipulation

As we have mentioned, the preprocessing phase occurs before big data are used. In this phase, several processes are applied to the data, such as cleaning, data integration, and reduction. Applying all of these transformations to a large amount of data may be costly in terms of money, effort, and time. Thus, we define the ease of manipulation as the extent to which data could be easily used with minimal effort. The measurement of this metric is related to the invested effort to prepare data for manipulation. To quantify this effort, we compare the data in their original schemas and after preprocessing. Thus, we define ease of manipulation as the ratio of differences between the raw data and the preprocessed data to the total data.

$$\text{Ease}(\%) = \frac{\text{Number of differences between original and cleaned table}}{\text{Total data}} \times 100$$

5.2.9. Relevancy

Data relevancy and usability comprise yet another essential quality dimension to consider in big data environments, as not all captured data are relevant to the intended use. Relevancy refers to the level of consistency between the information contained within data and the needed information. Measuring relevancy is very contextual and depends on the intended use of data. Thus, different measures have been defined in the literature [19–29]. However, this paper considers the definition suggested in [19]. In that study, the authors linked data relevancy to the number of accesses to data and considered the most frequently accessed data as the most relevant ones.

$$\text{Relevancy}(\text{Field F}) = \frac{\text{Number of access to F}}{\text{Total access to the table that includes F}} \times 100$$

5.2.10. Security

Data Security refers to the extent to which access to data is appropriately restricted. With increasing large-scale privacy breaches and security attacks, ensuring data confidentiality and security has become a priority. Measuring data security requires a more specific and in-depth study. However, we can highlight some guiding questions allowing us to assign a score to the data security level:

- Is there a security policy restricting data use? (20%)
- Are security protocols used for data transfer? (20%)
- Are there measures for threat detection? (20%)
- Are data appropriately encrypted? (20%)
- Is there a security documentation that accompanies the data? (20%)

5.2.11. Accessibility

This metric ensures that data are available and easy to retrieve. Ensuring data accessibility is a high priority since unreachable data are useless. Data should be accessible and effectively used out of their local repository if they are distributed for external use. We define accessibility as the ratio of accessible values.

$$\text{Accessibility}(\%) = \frac{\text{Number of Accessible values}}{\text{Total values}} \times 100$$

5.2.12. Integrity

Data integrity refers to the accuracy and trustworthiness of data over their lifecycle. In a big data environment, data go through multiple processes before being used. Therefore, it is essential to ensure that data values have not been altered and that data validity is maintained during data processing. Measuring integrity consists of comparing data values before and after data processing. Thus, we define data integrity as the ratio of differences between the original and processed data values to the total values.

$$\text{Integrity}(\%) = \frac{\text{Number of differences between original and processed values}}{\text{Total values}} \times 100$$

5.3. Weighted Quality Metrics

Big data quality measurement could not be significant without considering data weights. Indeed, the information contained within data is not equally important from a business point of view. Actually, in most organizations, some data are more significant than others. Hence, relevant data must have a higher impact on data quality measurements. The example in Table 3 shows a customers' dataset schema with the following fields and their respective completeness scores: First Name, Last Name, Age, Address, Email, Phone Number, Country, and City.

Table 3. Customers' dataset fields and their completeness scores.

First Name	Last Name	Age	Address	Email	Phone Number	City	Country
90%	90%	80%	40%	30%	20%	65%	70%

Data completeness is considered to be the percentage of non-lacking values. Thus, the completeness score is 60.62% (1).

$$\text{Completeness} = \frac{90 + 90 + 80 + 40 + 30 + 20 + 65 + 70}{8} = 60.62\% \tag{1}$$

From a business point of view, if we consider the above dataset is intended to be used for a marketing campaign of a company, which is gathering data about their potential customers. The company will be more interested in contact data, such as email and phone numbers, that can be used to contact customers and promote their products or services. For more accurate data quality measurement, these fields should have a higher impact on the completeness metric. Thus, we suggest using weighted metrics by assigning weights to data fields according to the following methodology:

Step 1: Prioritize the fields according to their relevancy and the intended impact on the quality score. In the previous example, considering that contact information is the most important, the fields are ordered as follows:

1- Email.
2- Phone Number.
3- Address.
4- City and Country.
5- First name, Last Name, and Age.

Step 2: Assign 1 as factor f to the less important fields. In the above example, we assign 1 to the following fields: First name, Last Name, and Age. Then, assign a factor between 2 and 10 to each field according to the intended impact of the field compared to the less important field(s). We suggest using the guiding Table 4 to assign factors. Table 5 shows the assigned factors to the fields of the example above.

Table 4. Factor's impact degree.

Factor Range	1–2	3–4	5	6–7	8–10
Impact Degree	Very Low	Low	Moderate	Significant	Very High

Table 5. Field's factors.

First Name	**Last Name**	**Age**	**Address**	**Email**	**Phone Number**	**City**	**Country**
1	1	1	3	6	4	2	2

Step 3: Calculate the weight of each field which is the ratio of the assigned factor to the total factors. In this example, the sum of factors is 20.

Thus, if we apply weights to the fields as shown in Table 6, we obtain a completeness score of 45.5% (2).

$$\text{Completness} = \frac{90 \times 0.05 + 90 \times 0.05 + 80 \times 0.05 + 40 \times 0.15 + 30 \times 0.3 + 20 \times 0.2 + 65 \times 0.1 + 70 \times 0.1}{8} = 45.5\% \quad (2)$$

Table 6. Field's weights.

First Name	**Last Name**	**Age**	**Address**	**Email**	**Phone Number**	**City**	**Country**
1/20 = 0.05	1/20 = 0.05	1/20 = 0.05	3/20 = 0.15	6/20 = 0.3	4/20 = 0.2	2/20 = 0.1	2/20 = 0.1

Therefore, by ignoring data weights, a dataset containing the required information will be considered incomplete if the other details are missing despite their irrelevancy. On the other hand, by considering data weights, a dataset can have several empty values and be deemed comprehensive if it contains information that meets the business needs. Thus, data weights should be considered more precisely when measuring data quality, especially in big data systems where not all information are relevant. In [19], Vaziri et al. highlighted the value of considering data weights for data quality assessment. They also defined measures incorporating data weights for the following metrics: Completeness, Relevancy, Accuracy, and Timeliness. The suggested measures were then implemented in a case study. Despite the high impact of this approach on measurement accuracy, most of the studies have not considered data weights in their assessment approaches, which challenges the correctness and the precision of the performed measures.

It is worth noting that this approach is not limited to data fields and could also be applied to several levels of data units. Indeed, data rows could also have varying relevancy. From a business point of view, some customers could be more valuable than others in terms of sales and profitability. In this case, organizations would be more interested in having complete information on their best customers than others. Considering the example above, the company may be more interested in customers living in specific regions. In this case, if the required information is missing for the targeted customers, the database will be considered incomplete from a business point of view. Thus, the completeness score should reflect all of these insights by applying data weights to obtain an accurate quality assessment. Likewise, this approach could be applied to other data units, such as tables and data quality dimensions. Thus, for more accurate measurements, we suggest a weighted big data quality score where this approach will be applied at three levels:

- Data Fields: The data attributes are of varying relevancy and, thus, should have different weights when measuring a quality metric.

- Quality Dimensions: The quality metrics are of varying relevancy and, thus, should have different weights when measuring a quality aspect.
- Quality Aspects: The quality aspects are of varying relevancy and, thus, should have different weights when measuring the global quality score.

In the next section, we implement the quality metrics and aspects defined previously in a case study related to Twitter Steam while considering data weights at different levels. Also, a comparison study of the existing assessment approaches is performed, and conclusions are made.

6. Implementation

6.1. Dataset Description

In this section, we describe the implementation of the quality measures defined in previous sections. The data quality measures were applied to a large-scale dataset of tweets from the Twitter Stream related to COVID-19 chatter [39]. This case study was chosen for technical reasons, such as the data size corresponding to big data scale, source type (CSV file), attribute type, and simplicity. Besides the technical reasons, this case study was also chosen for its relevance. It consists of social media data that deal with a topical issue and its flexibility allows us to use it for further work. The dataset contains over 283 million tweets and complementary information, such as language, country of origin, and creation date and time. The gathered tweets were collected from 27 January to 27 March with over four million daily tweets. In our case study, the gathered data were structured and were of reasonably good quality. Thus, to stress our quality assessment approach, the dataset was intentionally scrambled in a way that impacted all the assessed metrics.

6.2. Tools

Our big data quality assessment framework was implemented using Apache Spark, which handles large datasets in big data environments. The quality measures were implemented in Python using Pyspark libraries with the following software and tools:

- Apache SPARK 3.1.2, a big data processing system used to handle large datasets.
- Python 3.8.8
- Jupiter Notebook 6.3.0, a web-based computing platform that serves as a development environment.
- Great Expectation Package [40], a python library that offers multiple functions for validating and profiling data. The library is open-source and appropriate for large datasets (scalable).
- Scientific Python libraries, such as Numpy, Spell Checker, Matplotlib, Pandas, Scipy, and Datetime.

The above tools were chosen based on three criteria: suitability for big data (Scalability), being open-source, and documentation availability.

6.3. Results

This section aims to show how big data quality can be assessed using the metrics and quality aspects presented in the previous sections. Before any assessment, the dataset was loaded and preprocessed. Indeed, the data schema was adjusted to fulfill the measurement requirements. Thus, some fields, such as creation date and time, were formatted. Moreover, additional fields were included in the dataset, such as the last modification DateTime field that allowed the measurement of the Timeliness and Volatility metrics. Additionally, the number of accesses to the data were added to the dataset to calculate the Relevancy metric. As the goal of the implementation was to test the suggested measurements, arbitrary values were assigned to the additional fields while maintaining the case study's consistency and logic. Then, weights were assigned to the data fields (Table 7), the metrics (Tables 8 and 9), and the quality aspects (Table 10). The performed experiments aimed to accomplish the following goals:

- Measuring the following quality metrics: Completeness, Timeliness, Volatility, Uniqueness, Conformity, Consistency, Ease of manipulation, Relevancy, Security, Readability, Accessibility, and Integrity, while considering the attribute weights in Table 7.

Table 7. Fields' Weights.

Tweet	Language	Country	Creation Date Time	Last Modification Date Time
0.4	0.15	0.15	0.15	0.15

- Measuring the following quality aspects: Reliability, Availability, Usability, Pertinence, and Validity, while considering the metric weights in Tables 8 and 9.

Table 8. Metric Weights 1.

Reliability		Availability		Pertinence	
Integrity	**Volatility**	**Security**	**Accessibility**	**Timeliness**	**Uniqueness**
0.7	0.3	0.8	0.2	0.7	0.3

Table 9. Metrics Weights 2.

Validity			Usability		
Consistency	**Conformity**	**Readability**	**Completeness**	**Relevancy**	**Ease of Manipulation**
0.4	0.4	0.2	0.5	0.3	0.2

- Measuring a weighted data quality score while considering the aspect weights in Table 10

Table 10. Aspect Weights.

Reliability	Availability	Pertinence	Validity	Usability
0.3	0.1	0.1	0.3	0.2

We assessed the quality metrics for each field according to the above weights. Then, the global quality metric was measured for the whole data. Based on the obtained results, the data quality aspects were then evaluated. Finally, an overall big data quality score was deducted. The obtained results are presented in the following charts: Figure 6 shows the score in percentage for each quality metric and Figure 7 shows the score of the quality aspects.

It is worth mentioning that the fields' weights were not considered for all the metrics. Indeed, some metrics could not be assessed for each field and, therefore, were not measured for all data attributes, such as Uniqueness, Relevancy, Security, and Accessibility.

To measure the Ease of manipulation metric, data preprocessing should be performed. Thus, in addition to the above adjustments, consistency and conformity were improved by changing data types and formats to meet the expected data model. Then, the original and preprocessed tables were compared, and the similarity rate was assessed.

Once the data quality metrics were measured, the data quality aspects were evaluated. The overall big data quality score was then deducted to get this final score:

$$\text{Quality Score}(\%) = 58.35\%$$

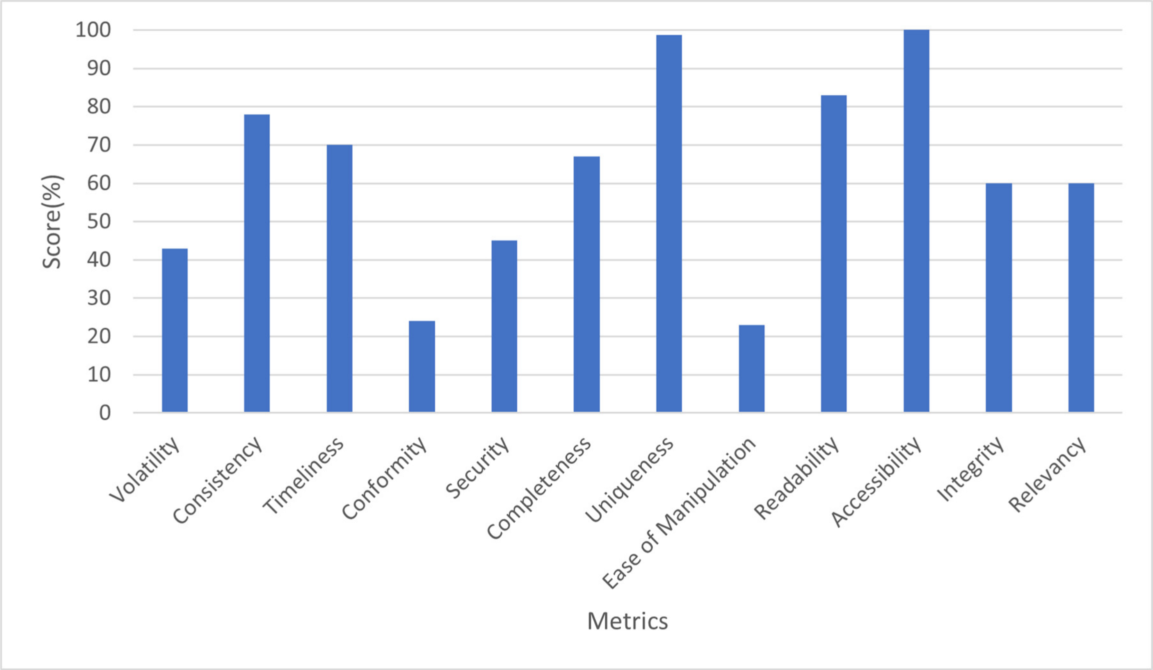

Figure 6. Data Quality Metric Scores.

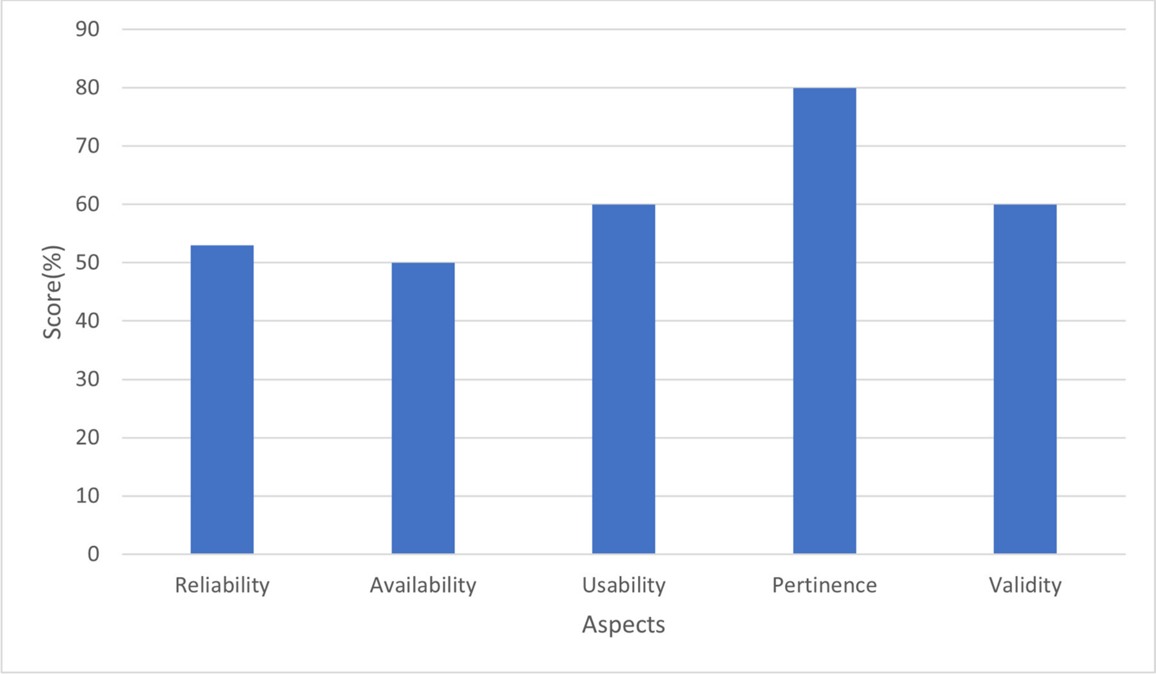

Figure 7. Data Quality Aspect Scores.

6.4. Comparative Study

In this section, we conduct a comparative study to assess the suggested framework against the existing data quality assessment approaches in terms of the framework model, the used existing metrics, the new introduced metrics, and the big data scope and precision.

6.4.1. Framework Model

A recent study [41] about data quality assessment for big data has defined a quality evaluation model that allows attributing a score to the existing quality assessment frameworks for big data. The suggested score is defined as a formula (S) using the seven evaluation criteria specified in Table 11, whose value is either 0 or 1. In the survey, the best-achieved score was 7.5 in the paper [31]. Thus, we evaluated our framework and the existing frameworks using this model.

$$S = D + 2 \times M + 2 \times Mt + 1.5 \times F + 1.5 \times S + A + P \quad (S)$$

Table 11. Definitions of the evaluation criteria.

Evaluation Criteria	Description
D	Providing a data quality definition
M	Proposing a quality model
Mt	Offering an assessment metrics
F	Proposing a methodology or framework
S	Making a simulation or prototype
A	Presenting a state of the art
P	Conducting a poll

We assessed our framework according to the following statements:

- We provided a data quality definition, so D = 1.
- We proposed a new quality model based on 12 metrics and 5 quality aspects, so M = 1.
- We defined 12 evaluation metrics, so Mt = 1.
- We did not conduct polling, so P = 0.
- We outlined the state of the art, so A = 1.
- We conducted a simulation, so S = 1.
- We proposed an assessment framework providing a weighted data quality score, so F = 1.

Table 12 shows the classification and the comparison of the existing data quality frameworks that have defined new or existing quality metrics and the score for each.

Table 12. Comparative table of the frameworks and their scores.

	Ref.	[28]	[18]	[2]	[33]	[19]	[21]	[29]	[32]	[31]	[30]	[26]	**Our Framework**
Score	D	1	1	1	1	1	1	1	1	1	1	1	1
	M	1	1	1	1	1	1	1	1	1	0	0	1
	Mt	1	1	0	1	1	0	1	1	1	1	1	1
	F	1	0	0	0	0	0	0	0	1	1	1	1
	S	0	1	1	0	0	1	1	1	0	0	1	1
	A	1	1	1	1	1	1	1	1	1	1	1	1
	P	0	0	0	0	0	0	0	0	0	0	0	0
	Score	7.5	7.5	5.5	6	6	5.5	7.5	7.5	7.5	5.5	7	9

6.4.2. Used Existing Metrics

Table 13 shows the comparison of the existing data quality frameworks in terms of the used existing metrics.

6.4.3. New Metrics

Table 14 shows the comparison of the data quality assessment frameworks in terms of the new quality metrics introduced by each framework.

Table 13. Comparative table of used existing metrics.

	Ref.	[28]	[18]	[2]	[33]	[19]	[21]	[29]	[32]	[31]	[30]	[26]	**Our Framework**
Metrics	Completeness	*	*		*	*		*	*	*	*		*
	Timeliness	*	*		*	*			*				*
	Volatility								*				*
	Uniqueness		*	*	*		*				*		*
	Accuracy	*			*	*			*	*		*	
	Conformity		*		*								*
	Consistency	*	*		*	*			*	*	*		*
	Correctness	*			*								
	Relevancy					*		*					*
	Readability							*					*
	Spread				*								

*: The mentioned metric is addressed by the referenced study.

Table 14. Comparative table of new introduced metrics.

Ref.	[28]	[18]	[2]	[33]	[19]	[21]	[29]	[32]	[31]	[30]	[26]	**Our Framework**
Introducing New Metrics?	yes	no	no	yes	yes	no	no	no	no	no	no	yes
Number of New Metrics	1			2	1							4

6.4.4. Scope and Precision

Table 15 shows the comparison of the data quality assessment frameworks regarding their precision in terms of weighted metrics and number of considered metrics, as well as their big data scope.

Table 15. Comparative table of precision and big data scope.

Ref.	[28]	[18]	[2]	[33]	[19]	[21]	[29]	[32]	[31]	[30]	[26]	**Our Framework**
Big Data	*			*			*	*			*	*
Weighted Quality					*							*
Number of considered Metrics	5	5	1	8	5	1	3	5	3	3	1	12

*: The mentioned scope and precision criteria are addressed by the referenced study.

6.4.5. Discussion

The obtained overall big data quality score shows the low quality of big data and sheds light on the importance of cleaning and preprocessing big data before using them. Considering big data characteristics, getting very low scores for the Volatility, Conformity, and Ease of manipulation metrics is quite normal. As discussed previously, the data quality scores, as presented in Figures 6 and 7, show the high dependency between the quality metrics. Hence, we obtained a low security score for a high accessibility, and a balanced score between completeness and relevancy. Thus, the suggested framework allows not only the assessment of big data quality in an accurate way, but also the assessment of how the quality metrics impact on each other. Based on the results presented in Tables 12–15, we notice that, even if many efforts have been conducted to assess data quality, there is still a significant lack of works defining new metrics. Indeed, common data quality metrics, such as Completeness, Consistency, and Accuracy, are the most addressed metrics in the literature. On the other hand, new metrics, such as Volatility, Spread, and Readability, are less considered in the literature. Moreover, we notice that there is a confusion in the naming of metrics. In fact, there are some metrics referring to the same meaning but are named differently in each study, such as Freshness that refers to Timeliness and Copying that refers to Conformity. Therefore, while setting up the comparative table, we considered the meanings of the defined metrics instead on their names. The tables also show a lack of consideration of data weights by the existing data quality assessment approaches despite their high importance and impact on the accuracy of the measures. Thus, the obtained results show that the suggested methodology outperforms the current data quality assessment frameworks with a score of 9/10, while including 12 defined metrics, considering data weights, and addressing big data characteristics. On the other hand, with the 12 defined metrics, there is still a large gap between the defined dimensions and the measured metrics, which highlights the need to measure and implement new quality metrics. In addition, some metrics are so contextual, and, therefore, can be measured differently depending on the context of the data, such as Relevancy and Accessibility. For these metrics, generic and non-context-aware approaches need to be implemented. To summarize, based on the results obtained from this implementation, we conclude the following:

- Big data must be preprocessed before any use, as the gathered data are usually not consistent and of low quality.
- Considering data weights is mandatory for an accurate and significant assessment.
- Data managers should be aware of the dependencies between data quality metrics as improving a quality dimension may degrade the other ones.
- The quality assessment should be performed in each stage of the BDVC as every change in the data may degrade data quality.
- There is still a great need to implement new metrics especially for big data.
- Some metrics, such as Relevancy and Accessibility, need to be defined using generic and non-context-aware approaches [42].

7. Conclusions and Future Work

In recent years, big data have shown high capabilities to support companies in improving their business and making informed decisions. However, big data benefits could be exploited only if data quality is improved. Indeed, the collected big data are usually unstructured and contain anomalies that may bias data analysis. Thus, this work suggests a data quality assessment that allows big data quality monitoring. In this paper, we first reviewed all the available studies that have addressed data quality metrics. Then, we suggested measures for 12 quality metrics: Completeness, Timeliness, Volatility, Uniqueness, Conformity, Consistency, Ease of manipulation, Relevancy, Readability, Security, Accessibility, and Integrity. In addition, we defined and measured five quality aspects: Reliability, Availability, Usability, Pertinence, and Validity. These measures were performed while considering data attribute weights, quality metric weights, and quality aspect weights. Finally, the suggested measures were implemented, and an evaluation score was estimated to evaluate the presented framework against the existing data quality assessment frameworks. The obtained results show that the suggested methodology outperforms the current data quality assessment frameworks with a score of 9/10, while including12 defined metrics, considering data weights, and addressing big data characteristics. In future work, we aim to extend our framework by considering more metrics for a more accurate assessment. We also aim to improve the defined metrics to enhance the quality of big data.

Author Contributions: W.E. performed the primary literature review and proposed the framework. W.E. conducted the experiments and wrote the initial draft manuscript. I.E.A. and Y.G. provided guidelines to perform the study and supervised the programming. Y.G. and S.E.M. reviewed and validated the proposed framework and the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset supporting the conclusions of this article is available in the Kaggle repository, https://kaggle.com/adarshsng/covid19-twitter-dataset-of-100-million-tweets (accessed on 7 October 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

BDVC: Big Data Value Chain; BDQM: Big Data Quality Metrics.

References

1. Elouataoui, W.; Alaoui, I.E.; Gahi, Y. Data Quality in the Era of Big Data: A Global Review. In *Big Data Intelligence for Smart Applications*; Baddi, Y., Gahi, Y., Maleh, Y., Alazab, M., Tawalbeh, L., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 1–25. [CrossRef]
2. Li, Y.; Yang, J.; Zhang, Z.; Wen, J.; Kumar, P. Healthcare Data Quality Assessment for Cybersecurity Intelligence. *IEEE Trans. Ind. Inform.* **2022**, *19*, 841–848. [CrossRef]
3. Elouataoui, W.; El Alaoui, I.; Gahi, Y. Metadata Quality Dimensions for Big Data Use Cases. In Proceedings of the International Conference on Big Data, Modelling and Machine Learning (BML), Kenitra, Morocco, 6 August 2022; pp. 488–495. [CrossRef]
4. Kapil, G.; Agrawal, A.; Khan, R.A. A study of big data characteristics. In Proceedings of the 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 21–22 October 2016; pp. 1–4. [CrossRef]
5. Faroukhi, A.Z.; El Alaoui, I.; Gahi, Y.; Amine, A. An Adaptable Big Data Value Chain Framework for End-to-End Big Data Monetization. *Big Data Cogn. Comput.* **2020**, *4*, 34. [CrossRef]
6. Faroukhi, A.Z.; El Alaoui, I.; Gahi, Y.; Amine, A. Big data monetization throughout Big Data Value Chain: A comprehensive review. *J. Big Data* **2020**, *7*, 3. [CrossRef]
7. Juddoo, S. Overview of data quality challenges in the context of Big Data. In Proceedings of the 2015 International Conference on Computing, Communication and Security (ICCCS), Pointe aux Piments, Mauritius, 4–5 December 2015; pp. 1–9. [CrossRef]
8. Eouataoui, W.; El Alaoui, I.; Gahi, Y. Metadata Quality in the Era of Big Data and Unstructured Content. In *Advances in Information, Communication and Cybersecurity*; Maleh, Y., Alazab, M., Gherabi, N., Tawalbeh, L., Abd El-Latif, A.A., Eds.; Advances in Information, Communication and Cybersecurity. Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2021; Volume 357. [CrossRef]
9. Alaoui, I.E.; Gahi, Y.; Messoussi, R.; Todoskoff, A.; Kobi, A. Big Data Analytics: A Comparison of Tools and Applications. In *Innovations in Smart Cities and Applications*; Ben Ahmed, M., Boudhir, A., Eds.; Lecture Notes in Networks and Systems; Springer: Cham, Switzerland, 2018; Volume 37. [CrossRef]
10. Alaoui, I.E.; Gahi, Y.; Messoussi, R. Full Consideration of Big Data Characteristics in Sentiment Analysis Context. In Proceedings of the 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 12–15 April 2019; pp. 126–130. [CrossRef]
11. Sidi, F.; Shariat Panahy, P.H.; Affendey, L.S.; Jabar, M.A.; Ibrahim, H.; Mustapha, A. Data quality: A survey of data quality dimensions. In Proceedings of the 2012 International Conference on Information Retrieval Knowledge Management, Kuala Lumpur, Malaysia, 13–15 March 2012; pp. 300–304. [CrossRef]
12. El Alaoui, I.; Gahi, Y.; Messoussi, R. Big Data Quality Metrics for Sentiment Analysis Approaches. In Proceedings of the 2019 International Conference on Big Data Engineering, New York, NY, USA, 11 June 2019; pp. 36–43. [CrossRef]
13. Wang, R.Y.; Strong, D.M. Beyond Accuracy: What Data Quality Means to Data Consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [CrossRef]
14. Alaoui, I.E.; Gahi, Y. The Impact of Big Data Quality on Sentiment Analysis Approaches. *Procedia Comput. Sci.* **2019**, *160*, 803–810. [CrossRef]
15. Tranfield, D.; Denyer, D.; Smart, P. Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *Br. J. Manag.* **2003**, *14*, 207–222. [CrossRef]
16. Wang, R.Y. A product perspective on total data quality management. *Commun. ACM* **1998**, *41*, 58–65. [CrossRef]
17. Lee, Y.W.; Strong, D.M.; Kahn, B.K.; Wang, R.Y. AIMQ: A methodology for information quality assessment. *Inf. Manag.* **2002**, *40*, 133–146. [CrossRef]
18. Bors, C.; Gschwandtner, T.; Kriglstein, S.; Miksch, S.; Pohl, M. Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics. *J. Data Inf. Qual.* **2018**, *10*, 1–26. [CrossRef]
19. Vaziri, R.; Mohsenzadeh, M.; Habibi, J. Measuring data quality with weighted metrics. *Total Qual. Manag. Bus. Excell.* **2019**, *30*, 708–720. [CrossRef]
20. Batini, C.; Barone, D.; Cabitza, F.; Grega, S. A Data Quality Methodology for Heterogeneous Data. *Int. J. Database Manag. Syst.* **2011**, *3*, 60–79.
21. Li, Y.; Chao, X.; Ercisli, S. Disturbed-entropy: A simple data quality assessment approach. *ICT Express* **2022**, *8*, 3. [CrossRef]
22. Taleb, I.; Serhani, M.A.; Bouhaddioui, C.; Dssouli, R. Big data quality framework: A holistic approach to continuous quality management. *J. Big Data* **2021**, *8*, 76. [CrossRef]

23. Wong, K.Y.; Wong, R.K. Big data quality prediction informed by banking regulation. *Int. J. Data Sci. Anal.* **2021**, *12*, 147–164. [CrossRef]
24. Azeroual, O.; Saake, G.; Abuosba, M. Data Quality Measures and Data Cleansing for Research Information Systems. *arXiv* **2019**, arXiv:1901.06208. Available online: http://arxiv.org/abs/1901.06208 (accessed on 12 November 2021).
25. Timmerman, Y.; Bronselaer, A. Measuring data quality in information systems research. *Decis. Support Syst.* **2019**, *126*, 113138. [CrossRef]
26. Mylavarapu, G.; Thomas, J.P.; Viswanathan, K.A. An Automated Big Data Accuracy Assessment Tool. In Proceedings of the 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), Suzhou, China, 15–18 March 2019; pp. 193–197. [CrossRef]
27. Taleb, I.; Serhani, M.A.; Dssouli, R. Big Data Quality: A Data Quality Profiling Model. In *Services—SERVICES 2019*; Springer: Cham, Switzerland, 2019; pp. 61–77. [CrossRef]
28. Heinrich, B.; Hristova, D.; Klier, M.; Schiller, A.; Szubartowicz, M. Requirements for Data Quality Metrics. *J. Data Inf. Qual.* **2018**, *9*, 1–32. [CrossRef]
29. Arolfo, F.A.; Vaisman, A.A. Data Quality in a Big Data Context. In *Advances in Databases and Information Systems*; Benczúr, A., Thalheim, B., Horváth, T., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11019.
30. Micic, N.; Neagu, D.; Campean, F.; Zadeh, E.H. Towards a Data Quality Framework for Heterogeneous Data. In Proceedings of the 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Exeter, UK, 21–23 June 2017; pp. 155–162. [CrossRef]
31. Taleb, I.; Kassabi, H.T.E.; Serhani, M.A.; Dssouli, R.; Bouhaddioui, C. Big Data Quality: A Quality Dimensions Evaluation. In Proceedings of the 2016 Intelligence 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, France, 18–21 July 2016; pp. 759–765. [CrossRef]
32. Serhani, M.A.; El Kassabi, H.T.; Taleb, I.; Nujum, A. An Hybrid Approach to Quality Evaluation across Big Data Value Chain. IEEE. In Proceedings of the 2016 IEEE International Congress on Big Data (BigData Congress), Washington, DC, USA, 5–8 December 2016; pp. 418–425. [CrossRef]
33. Firmani, D.; Mecella, M.; Scannapieco, M.; Batini, C. On the Meaningfulness of "Big Data Quality" (Invited Paper). *Data Sci. Eng.* **2016**, *1*, 6–20. [CrossRef]
34. Cai, L.; Zhu, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci. J.* **2015**, *14*, 2. [CrossRef]
35. Zhang, P.; Xiong, F.; Gao, J.; Wang, J. Data quality in big data processing: Issues, solutions and open problems. In Proceedings of the 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), San Francisco, CA, USA, 4–8 August 2017; pp. 1–7. [CrossRef]
36. Wand, Y.; Wang, R.Y. Anchoring data quality dimensions in ontological foundations. *Commun. ACM* **1996**, *39*, 86–95. [CrossRef]
37. Gahi, Y.; El Alaoui, I. Machine Learning and Deep Learning Models for Big Data Issues. In *Machine Intelligence and Big Data Analytics for Cybersecurity Applications*; Maleh, Y., Shojafar, M., Alazab, M., Baddi, Y., Eds.; Studies in Computational Intelligence; Springer: Cham, Switzerland, 2021; Volume 919. [CrossRef]
38. Elouataoui, W.; Alaoui, I.E.; Mendili, S.E.; Gahi, Y. An End-to-End Big Data Deduplication Framework based on Online Continuous Learning. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 33. [CrossRef]
39. COVID-19: Twitter Dataset Of 100+ Million Tweets. Available online: https://kaggle.com/adarshsng/covid19-twitter-dataset-of-100-million-tweets (accessed on 7 October 2021).
40. Great Expectations Home Page. Available online: https://www.greatexpectations.io/ (accessed on 24 August 2022).
41. Reda, O.; Sassi, I.; Zellou, A.; Anter, S. Towards a Data Quality Assessment in Big Data. In Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications, New York, NY, USA, 23–24 September 2020; pp. 1–6. [CrossRef]
42. Alaoui, I.E.; Gahi, Y. Network Security Strategies in Big Data Context. *Procedia Comput. Sci.* **2020**, *175*, 730–736. [CrossRef]