A Novel Rigorous Measurement Model for Big Data Quality Characteristics

1st Haochen Zou
Department of Computer Science and Software Engineering
Concordia University
Montreal, Quebec, Canada
haochen.zou@mail.concordia.ca

2nd Kun Xiang
Department of Science and Engineering
Hosei University
Koganei, Tokyo, Japan
kun.xiang.2u@stu.hosei.ac.jp

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020564

***Abstract*—Satisfiable data quality is the basic guarantee for data-based research, decision-making, and service. Today, new trends in the creation, collection, and utilization of data are constantly emerging. With the usage of massive data, the problem of data quality is highlighted. Several studies on the measurement, evaluation, and management of big data quality have been proposed, and the data quality problem in the big data environment has received attention. The big data characteristics Vs model describes the dimensions and attributes information of data sources in detail, which can be implemented in big data quality measurement. In this paper, a novel rigorous big data quality measurement architecture is proposed for automatically and parallelly quantifying the value of six big data Vs, which are Volume, Variety, Velocity, Veracity, Validity, and Vincularity according to the developed algorithms in every big data process step and time phase of the big data pipeline. Thresholds for the six big data Vs are provided correspondingly for analyzing the result values. The hierarchical measurement model is constructed with multiple-based measures, derived measures, and indicators. The model is verified by comparative experiments and experiments results indicate that the designed architecture can improve the outcomes of data source implementation.**

***Index Terms*—big data quality characteristics, measurement hierarchical model, representational theory of measurement, big data measurement framework**

I. Introduction

With the rapid development of information technology such as the Internet, Internet of things (IoT), and cloud computing, a surging amount of massive data are generated anytime and anywhere in various fields such as e-commerce, biomedicine, and financial investment [1]. The scale of these data is growing daily, promoting the advent of the era of big data. As a hot research field, big data brings new challenges and opportunities to the development of information technology on the one hand. Analyzing and processing big data has become a popular technical issue in the industry [2]. Meanwhile, big data itself is still becoming a strategic resource, with which one can become a dominant player in the big data market and even in the field of information technology [3].

In the era of big data, numerous researchers have discussed the problems of big data science and the opportunities as well as challenges in the research. Big data has the characteristics of immense data volume, diverse data content, and complex data structure [4]. As a result, there will inevitably be a series of quality issues such as errors and missing inside the big data. The usage of poor-quality data is bound to reduce the accuracy of research results, increase project operation and maintenance costs, and bring huge risks and economic losses to decision-makers [5]. High-quality data is an important factor to ensure that the value of data can play a normal role, so the research on the value evaluation of data quality is particularly important.

Various researchers have diverse perceptions of the value of data quality, with the in-depth study of the big data quality, numerous scholars believe data quality is a multidimensional concept [6]. Data quality is considered to be the suitability of data consumers for use, and data quality degrees are regarded as a set of quality attributes that represent some aspect or structure of data. The big data quality value dimension provides criteria for assessing data quality. Becker et al. analyzed data quality issues from the quality dimensions of relevance, accuracy, consistency, integrity, security, and so on. It was concluded that decisions made based on the most relevant, complete, accurate, and timely data were obviously more conducive to research development [7]. Cai et al. proposed hierarchical data quality standards from the perspective of users, which are composed of quality dimensions such as usability, reliability, relevance, and representability. Multiple factors are designed in the paper for each dimension for evaluation practice [8]. Abdallah et al. listed different quality factors and dimensions, and described quality frameworks that can be utilized to measure the quality of big data [9].

Big data does not refer to the progressively large multiple datasets, but also to the fundamental improvement of the big data systems architecture need to manage the quality of the data [10]. Publications have proposed frameworks and architectures to address the diverse needs for data quality analysis. The quality assessments of big data have achieved various research results and cover multiple fields. However, due to the characteristics of large data in the current environment, such as enormous numbers, multiple sources and diverse formats, existing research works contain certain limitations in the evaluation of the data [11]. In addition, the main challenge that research and industry face contemporarily is the lack of visibility and transparency of big data's quality. Therefore, in this paper, we address the need by developing a novel big data quality measurement framework. The data issues and characteristics can be identified, analyzed, evaluated, documented, and reported continuously by integrating the big data quality measurement procedures within the steps of the big data pipelines. The research aims to bridge the gap between the industrial usage of big data and the underlying big data quality to verify the quality for the intended purpose. Big data quality information is analyzed and generated before the data propagate into the further decision-making process.

In the research, the intrinsic big data quality characteristics referred to as Vs are being concerned and focused, which are Volume, Velocity, Variety, Veracity, Validity, and Vincularity [12]. Measurement models and algorithms are proposed. A coherent, easy-to-use, and flexible framework in which the measurements can be applied to practical big data systems is designed. The contribution of this paper is the proposed architectural pipeline solution for providing continuous measurement feedback on the six Vs data quality characteristics with visualization, quantification, and interpretation of the results to assess the quality and determine the suitability according to the proposed quality thresholds of the quality characteristics for the further implementation of big data.

The paper is organized as follows: Section II explores existing works. Section III introduces and analyzes the proposed method. Section IV describes the comprehensive experiments. Section V concludes this paper and outlines future research.

II. Literature Review

Big data is generally utilized to describe a considerable amount of structured and unstructured data, which is difficult to be processed with traditional databases and software. Xu et al. consider big data as an information asset with a massive, high growth rate, and diversified attributes, which needs a novel processing mode to adapt [13]. Mikalef et al. believe that big data is a kind of information that can be understood and read by human beings, and its scale is too large to be processed by on-job methods in an acceptable time [14]. The big data V-model characteristics proposed by academic and industry researchers are largely utilized by big data managers for modelling big data sufficiently in application sectors. The big data V-model was first decorated with three Vs which are Volume, Velocity, and Variety by Laney et al. [15], and widely expanded by researchers with the development of the big data study in recent years. Alsaig el al. collected 19 big data Vs from an exhaustive study of previously published works on the Vs to characterize big data [16]. They selected 10 Vs from the overall set which are Volume, Velocity, Variety, Value, Validity, Veracity, Volatility, Vitality, Vincularity and Visualization. The chosen Vs can characterize big data in diverse sectors from the justification they constructed. The above-discussed studies focused on identifying and building the big data's inherent quality Vs characteristics, however, did not publish specific algorithms and concrete models to measure and quantify the value of proposed data quality characters.

Data quality is the basis of big data research. If the quality of big data is fundamentally poor, the results of data mining and analysis will be polluted. 15 big data qualities are established by ISO/IEC/IEEE Standard 15939 software measurement methods and guidelines, which are accuracy, completeness, consistency, credibility, currentness, accessibility, compliance, efficiency, confidentiality, precision, traceability, portability, availability, understandability, and recoverability [17]. Several studies have been conducted to measure the data quality Vs characteristics. They construct frameworks and architectures for evaluating and managing big data quality, and provide decision-making policies. Taleb et al. proposed a big data management framework for enhancing the preprocessing activities while strengthening data [18]. Their work measured four data quality dimensions Vs, which are Volume, Velocity, Variety, and Veracity. Ormandjieva et al. derived a measurement information model for quantifying the most widely utilized three indicators of data quality Vs: Volume, Velocity, and Variety. The framework developed by them complies with the guideline established by ISO/IEC/IEEE 15939 standard [19]. Bhardwaj et al. developed a novel conceptual quality measurement framework for big data with the purpose of assessing the underlying quality characteristics of big data at each step of the big data pipelines [20]. In their research, the theoretical quality measurement models for four of the big data Vs: Volume, Variety, Velocity, and Veracity are automated.

Although the above-discussed research addressed big data quality measurement modelling according to big data V-model dimensions and characteristics, there are limitations to existing studies evaluating big data. First of all, with the explosive growth of the number and format of research data, the three Vs and four Vs models proposed by the studies may not be able enough to fully cover the data attributes in the evaluation. Secondly, the value of the threshold of each data V characteristic, as the boundary limit to justify whether the data quality is qualified, is not clearly provided in the above research. Therefore, in this paper, we designed a big data quality measurement model which is built upon the standard measurement principles specified in ISO/IEC/IEEE 15939 software measurement methods and guidelines. It automates six Vs big data measurement characteristics hierarchical models aimed at evaluating and quantifying six aspects of big data attributes, including Volume, Velocity, Variety, Veracity, Validity, and Vincularity. Meanwhile, the threshold for each aspect of data characteristic is defined clearly for the evaluation and justification of the big data quality. In what follows, more details will be introduced.

III. Methodology

The goal of this section is to construct theoretically valid data measures for six Vs characteristics in the model, namely Volume, Velocity, Variety, Veracity, Validity, and Vincularity. The big data dimensions are the indicators in the measurement, which consist of one to multiple based and derived measures.

A. Big Data Volume

Big data volume measures the number of information bits across all records required to specify the information content of multiple datasets. It refers to the vast amount of data that are generated and consumed at different timed. The based measure for analyzing the indicator Volume Vol is $Ndde$, which is the number of distinct data elements across multiple datasets MDS. The equation for measuring Volume Vol is defined as:

$$Vol(MDS) = Ndde(MDS) \times \log_2(Ndde(MDS)) \quad (1)$$

The measurement goal of measuring big data volume is to count the number of the big data, ensure that the quantity of big data elements inside big datasets is feasible for further industry or academic implementation and not suffer a drastic change, with the threshold of maximum 3 times greater or lower than previous dataset in the quantity among diverse time phases of the pipeline during big data modifications [21].

B. Big Data Velocity

The notion of big data velocity is defined informally in terms of the relative growth of big data over a period of time T. It is considered as the speed of processing of data in any form of handling, recording, and publishing of data. Velocity indicator refers to changes in the big data volume over time of the pipeline and the speed at which data are generated and consumed. The equation for measuring indicator Velocity Vel of big data is formally defined as follows:

$$Vel(MDS) = \frac{Vol(MDS_{T2}) - Vol(MDS_{T1})}{Vol(MDS_{T1})} \times 100 \quad (2)$$

As displayed in the equation above, the based measure of indicator Velocity Vel is time T, and the derived measure as well as indicator of Velocity is Volume Vol. The threshold of the big data velocity is at the limitation of 1:3 of the market regulation [21]. Increasing the size of the velocity of big data deployment in different time phases is a worthwhile challenge.

C. Big Data Variety

Big data variety is defined as the weighted sum of the number of distinct data elements $Ndde$ across multiple datasets MDS, the length of the big dataset Lbd, which is the total number of data records in multiple datasets MDS, and the number of big datasets Nds across multiple datasets MDS. These values are aggregated into a single value for given multiple datasets. Variety refers to the ever-increasing different forms that data elements can come in among various time phases. It reflects correspondingly the diversity of unique data elements, the multiplicity of records, and the difference of datasets in the pipeline. Data variety is a measure of the richness and fruitiness of the data representation. From an analytic perspective, it is presumably the biggest obstacle to effectively utilizing large volume of data. The equation for measuring indicator Variety Var is formally defined as:

$$\begin{gathered} Var(MDS) = Ndde(MDS) \times W_{Ndde} + \\ Lbd(MDS) \times W_{Lbd} + Nds(MDS) \times W_{Nds} \end{gathered} \quad (3)$$

In the equation, W_{Ndde}, W_{Lbd}, and W_{Nds} are the weight of the number of distinct data elements $Ndde$, the weight of the length of the big dataset Lbd, and the weight of the number of datasets Nds respectively. The values of all weights are set to $\frac{1}{3}$ by default and will be adjusted according to the actual situation. The sum of all the weight values is equal to 1. The based measures of the indicator Variaty Var are the number of distinct data elements $Ndde$, the length of the big dataset Lbd, and the number of datasets Nds.

The measurement goal of big data variety is to calculate different formats of the data. The standard variance threshold value is calculated by two Python modules named $sklearn$ and $feature\_selection$, which have the $VarianceThreshold$ function for the purpose [22]. The actual variety value should not be lower than the threshold limitation, which indicates low variance and tiny features attributes of the big data contents.

D. Big Data Veracity

Big data veracity is considered to the degree that data is accurate, trusted, and precise. It is not only the accuracy of the data itself, but the trustworthiness of the data source, type, and processing of it. The derived measures associated with the Veracity indicator are defined as Accuracy, Completeness, Currentness, and Availability which are introduced from the ISO/IEC 25012 standard for quality of big data products [23].

The accuracy of the data reflects whether the data truly describes the application scenario. Let the jth data element of the ith data source be D_{ij}. There are N_i data records in the ith data source. The reference value standard of important attribute set $attr(D_i) = \{A_1, A_2, \cdots, A_l\}$ in this scenario is $M = M_1, M_2, \cdots, M_l$, then there is a mapping function:

$$\mu(D_{ij}) = \prod_{k=1}^{l} \varphi(D_{ij} \cdot A_k) \quad (4)$$

In the equation, $\varphi(\cdot)$ is the accuracy judgment function. If the value of D_{ij} on the attribute A_k meets the reference value standard M_k, then the value of $\varphi(\cdot)$ is 1. Otherwise, its value is 0. Therefore, when the values of data D_{ij} on the attribute set $attr(D_i)$ are all correct, it indicates that the data are true and accurate, and the value of $\mu(\cdot)$ is 1.

The data accuracy measurement model is shown as follows:

$$Accuracy = DAccuracy(D_i) = \frac{\sum_{j=1}^{N_i} \mu(D_{ij})}{N_i} \quad (5)$$

In the equation, the value of $DAccuracy$ ranges from [0, 1]. When the value of $DAccuracy$ is 0, it indicates that all data in the ith data source are not accurate. When the value of $DAccuracy$ is 1, it indicates that all data in the ith data source are extremely accurate.

In this study, Accuracy is defined as the degree to which data have attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use. The equations for measuring the data accuracy $Accu$ are formally developed and shown as follows:

$$H_{Accu}(MDS) = \log_2(Lbd) - \frac{1}{Lbd} \times \sum_{j=\{1,\cdots,k\}} P_j \log_2(P_j) \tag{6}$$

$$H_{Max}(MDS) = \log_2(Lbd) \tag{7}$$

$$Accu(MDS) = \frac{H_{Accu}(MDS)}{H_{Max}(MDS)} \tag{8}$$

In the above equations, MDS refers to the multiple datasets of the big data. Lbd is the length of the big dataset, which is the total number of data records in multiple datasets MDS. j is the serial number of the measured dataset and k is the overall number of the multiple datasets. P_j provides the total number of duplicate items and their specific count in datasets.

Completeness is considered as the degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use. The completeness assessment measures the completeness score value of data based on two evaluation indexes: data non-emptiness and data normalization.

In the process of data analysis and management, if the attribute values required by some analysis models are missing, the data will not have the function of analysis and management, and the usability of data will be directly reduced. Let the jth data element in the ith data source is D_{ij}, its value set on l important attributes is $attrVal_{ij}$, then there is a mapping function defined as follows:

$$Y(D_{ij}) = \begin{cases} 1, len(attrVal_{ij}) = l \\ 0, otherwise \end{cases} \tag{9}$$

Therefore the data non-emptiness measurement model is as:

$$Completeness_1 = DNonEmptiness(D_i) = \frac{\sum_{j=1}^{N_i} Y(D_{ij})}{N_i} \tag{10}$$

In the equation, $Y(D_{ij})$ is the mapping function to determine whether the jth data of the ith data source is non-empty. The value of $DNonEmptiness$ ranges from [0, 1]. If the value of $DNonEmptiness$ is 0, it indicates that all of the important attribute values of the data records from the ith data source are empty. If the value of $DNonEmptiness$ is 1, it indicates that all of the important attribute values of the data records from the ith data source are non-empty.

Data normalization is utilized to evaluate whether the description of each attribute value in the data is standardized, and the degree of normalization has a great influence on the identifiability between datasets. Let the jth data element in the ith data source is D_{ij}, its value set on l important attributes is $attrVal_{ij}$, then there is a mapping function:

$$Z(D_{ij}) = \begin{cases} 1, attrVal_{ij}\, all\, specifications \\ 0, otherwise \end{cases} \tag{11}$$

The data normalization measurement model is as follows:

$$Completeness_2 = DNormalization(D_i) = \frac{\sum_{j}^{N_i} Z(D_{ij})}{N_i} \tag{12}$$

In the equation, $Z(D_{ij})$ is a mapping function for determining whether the jth data element of the ith data source is normalized. The value of $DNormalization$ ranges from [0, 1]. If the value of $DNormalization$ is 0, it means that all the important attribute values of the ith data source are not standardized. If the value of $DNormalization$ is 1, it indicates that all the important attribute values of the data elements in the data source are standardized.

In summary, the completeness assessment model is as:

$$Completeness = \frac{\sum_{i \in Evaluation_{Completeness}} Completeness_i}{k} \tag{13}$$

In the equation, $Evaluation_{Completeness}$ is the subscript set of Completeness evaluation indicators determined for study enrollment, k is the number of elements in the $Evaluation_{Completeness}$ subscript set.

In this study, the equation for measuring the derived measure big data completeness $Comp$ is formally defined as:

$$Comp(MDS) = \frac{RecNoNull(MDS)}{Lbd(MDS)} \tag{14}$$

In the equation, the based measure $RecNoNull(MDS)$ is the frequency of records in multiple datasets MDS without null values. The based measure Lbd is considered as the length of the big dataset, which is the total number of data records across multiple datasets MDS.

Currentness is defined as the degree to which data has attributes that are of the right age in a specific context of use. Currentness evaluation is the timeliness of data update, which calculates the time difference between data generation time and current time.Take the current time as the baseline time and set it as t, then there is a data currentness model:

$$DCurrentness(D_i) = 1 - \frac{\sum_{i=1}^{n}(\frac{t-t_i}{t})}{n} \tag{15}$$

Where, t_i represents the recording time of the ith data. The value range of $DCurrentness$ is (0, 1), and the closer it is to 1, the better the timeliness of data is. In summary, since the currentness dimension only includes the data timeliness evaluation index, the currentness evaluation model is as follows:

$$Currentness = DCurrentness(\cdot) \tag{16}$$

In this paper, the equation for measuring the derived measure big data currentness $Curr$ is formally defined as follows:

$$Curr(MDS) = \frac{RecAccuAge}{Lbd(MDS)} \tag{17}$$

In the equation, the based measure $RecAccuAge(MDS)$ provides the total number of records with ages that fall within the acceptable range based on the upper and lower quantiles of the Box and Whisker. The based measure Lbd is the total number of data records in multiple datasets MDS.

Availability is the degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use. The equation for measuring the big data availability $Avai$ is formally defined as follows:

$$Avai(MDS) = \frac{NumSuccReq(MDS)}{NumReq(MDS)} \tag{18}$$

In the equation, the based measure $NumSuccReq(MDS)$ is the number of successful requests from an API, server, datastore, origins of data. The based measure $NumReq(MDS)$ is the number of requests in multiple datasets MDS.

The equation for measuring the value of indicator Veracity Ver is formally defined as follows:

$$\begin{gathered} Ver(MDS) = Accu(MDS) \times W_{Accu} + Comp(MDS) \times \\ W_{Comp} + Curr(MDS) \times W_{Curr} + Avai \times W_{Avai} \end{gathered} \tag{19}$$

In the equation, W_{Accu}, W_{Comp}, W_{Curr}, and W_{Avai} are the weight of the big data accuracy, the weight of the big data completeness, the weight of the big data currentness, and the weight of the big data availability respectively. The values of all the four weights are set to $\frac{1}{4}$ by default and will be further adjusted slightly according to the actual situation. The sum of total the four weight values is equal to 1.

The measurement goal of big data veracity is to count the ambiguity of the data [22]. The threshold minimum limitation is the $p - value$, which is 0.95 of the overall data [24].

E. Big Data Validity

Validity of big data is defined in terms of its accuracy and correctness for the purpose of usage. Validity of data might be having same ideas with veracity of data, but they do not have same concepts and theories. The derived measures of Validity indicator are Credibility and Compliance which are introduced from the ISO/IEC 25012 standard for big data quality [23].

Compliance is defined as the degree to which data has attributes that adhere to standards, conventions, or regulations in force and similar rules relating to data quality in a specific context of use. The equation for measuring the data compliance $Compli$ is formally defined as follows:

$$Compli(MDS) = \frac{\sum_{\forall DS \in MDS} NumRecCompli(DS)}{Nds(MDS)} \tag{20}$$

In the equation, the based measure $NumRecCompli(DS)$ refers to the number of compliant records in a dataset. The based measure $Nds(MDS)$ is the number of big datasets across overall multiple datasets MDS.

Credibility refers to the degree to which data has attributes that are regarded as true and believable by users in usage. The equation for measuring the data credibility $Cred$ is as follows:

$$Cred(MDS) = \frac{NdsCred(MDS)}{Nds(MDS)} \tag{21}$$

In the equation, the based measure $NdsCred(MDS)$ is the number of credible datasets, and the based measure $Nds(MDS)$ is the number of big datasets of multiple datasets.

The equation for measuring the value of indicator Validity Val is formally defined as follows:

$$\begin{gathered} Val(MDS) = Compli(MDS) \times W_{Compli} + \\ Cred(MDS) \times W_{Cred} \end{gathered} \tag{22}$$

In the equation, W_{Compli} and W_{Cred} are the weight of the big data compliance and the weight of the big data credibility. The values of all the two weights are set to $\frac{1}{2}$ by default and will be further adjusted according to the actual situation. The sum of total two weight values is equal to 1.

The measurement goal of the big data validity is to calculate and improve the value above the minimum threshold of 0.7 of the overall data elements in multiple datasets [25].

F. Big Data Vincularity

Vincularity of big data refers to the connectivity and linkage of data. The derived measure of Vincularity indicator is Traceability, which is introduced from the ISO/IEC 25012 standard for big data quality [23].

Traceability measure provides the degree to which data has attributes that provide an audit trail of access to the data and any changes made to the data in a specific context of use. Traceability is measured both from inherent and system-dependent points of view. The equation for measuring the data traceability $Trac$ is formally defined as follows:

$$Trac(DS) = \frac{RecTrac(DS)}{Ldst(DS)} \tag{23}$$

In the equation, the based measure $RecTrac(DS)$ provides the total number of records that are traceable in datasets. The based measure length of the record $Ldst(DS)$ is the total number of occurences of data elements in datasets.

The equation for measuring the value of indicator Vincularity Vin is formally defined as follows:

$$Vin(MDS) = \frac{\sum_{\forall DS \in MDS} Trac(DS)}{Nds(MDS)} \tag{24}$$

As displayed in the equation above, the based measure for measuring indicator Vincularity Vin is $Nds(MDS)$, which refers to the number of big datasets of multiple datasets. The derived measure for measuring the big data vincularity is the Traceability of each dataset in multiple datasets.

The measurement goal of the big data vincularity is to improve the value of the traceable data proportion into the minimum limitation of the threshold, which is the $p - value$ 0.95 of the overall data for further implementations.

G. Hierarchical Model

The hierarchical model proposed in this paper is shown in “Fig. 1”. The model is a hierarchical structure linking information needs to the relevant entities and attributes of concern, including the based measures, the derived measures, and the indicators. It defines how the relevant attributes are quantified and converted to indicators that provide a bias for decision-making [19]. In the model developed in this paper, the six Vs characteristics are decomposed through three layers, including the external attribute layer, the internal indicators layer, the derived, and based measures layer.

H. Big Data Quality Architecture

In this paper, a big data quality measurement architecture has been developed and implemented to allow for the usage of measuring data quality characteristics, including the derived measures and the based measures associated with the six big data Vs indicators. The proposed model provides feedback at each stage of the big data operation modification process and helps stakeholders related to big data such as big data managers, big data developers, data analysts, and data scientists to incorporate measurements for the Vs quality characteristics of big data into their industry and academic implementation.

As shown in “Fig. 2”, it depicts the overview of the designed big data quality measurement model combined with a big data process pipeline in the architecture. The five data process stages in the pipeline are illustrated, which are the data extraction phase, the data loading and preprocessing phase, the data processing phase, the data analysis phase, and the data loading and transformation phase. The original data sources for implementation come in from the leftmost end of the process and move from left to right, as displayed with arrows.

For each stage in the data process pipeline, the data quality model will analyze the 6 Vs data quality dimensions in parallel according to the algorithms and equations before and after the phase. Analysis results will be provided based on thresholds to help stakeholders obtain timely data quality information at each operational stage. In addition, data elements that affect the overall quality of the data source will be listed continually. Big data stakeholders are able to make decisions about these data elements before and after each data process phase. While helping to improve data quality, data contamination is avoided from affecting further series of processing and analysis steps.

IV. Experiments

A case study based on a public big data source from the Kaggle platform with a four-time frame is designed to verify the feasibility of the proposed model. A comparative experiment based on machine learning and deep learning models in the natural language process is constructed to simulate the practical application of the proposed model in data processing and analysis operations in the academic field. The purpose of the comparative experiment is to verify the effectiveness of the big data quality measurement model proposed in this paper.

A. Big Data Case Study

The feasibility and flexibility of the designed big data architecture are illustrated based on data which are real-time, timestamped, freely available, and traceable to a source for the purpose of reproducibility. Therefore, we opted to utilize the real estate for sale data in a city with four quarters time phases of a year as the data source to construct the big data process pipeline. The data source is from the Kaggle platform and is publicly visible to all users and researchers. Since the experiment data source is supplied on the platform by researchers as a hosting database for subsequent academic and industrial research, it is not raw and has already been through a bunch of complete data processing processes by the uploader such as data extraction, data preprocessing, data processing, data analysis, and data transformation. Therefore, the proposed data quality measurement model is primarily employed to supervise the last and most significant process step, which is whether the experiment data source can meet the quality requirements and be utilized in the following applications.

During the four-quarters time phases, the big data measurement model automatically calculates the big data 6 Vs quality characteristics parallelly according to the algorithms. The model provides the calculation results with a comparison of the thresholds after each time phase as a data quality guide to the stakeholders' decision-making in the approach.

As displayed in tables Table I and Table II below, the 6 Vs data quality characteristics values quantified by the designed big data quality measurement model and the comparison between generated results and standard thresholds about the datasets from the experiment data source are illustrated. All the data quality measurement values locate within the threshold requirements, which indicates that the datasets from the experiment data source have qualified data quality in every time phase and are suitable for further implementation.

TABLE I
6 Vs Data Quality Characteristics Results

Data Characters	**Time Phases**			
	T1	T2	T3	T4
Volume	4057135.68	4059903.76	4063722.51	4058985.33
Velocity	N/A	6.82%	9.41%	-11.66%
Variety	147908.33	155359.11	176430.93	150894.72
Veracity	96.87%	96.80%	96.77%	96.83%
Validity	88.38%	88.25%	88.51%	88.47%
Vincularity	98.85%	99.12%	98.20%	98.96%

The derived measures values of the big data veracity indicator are illustrated in “Fig. 3” as follows. The values of data accuracy, data availability, and data veracity decrease first and then increase slightly during the four-time phases. The values

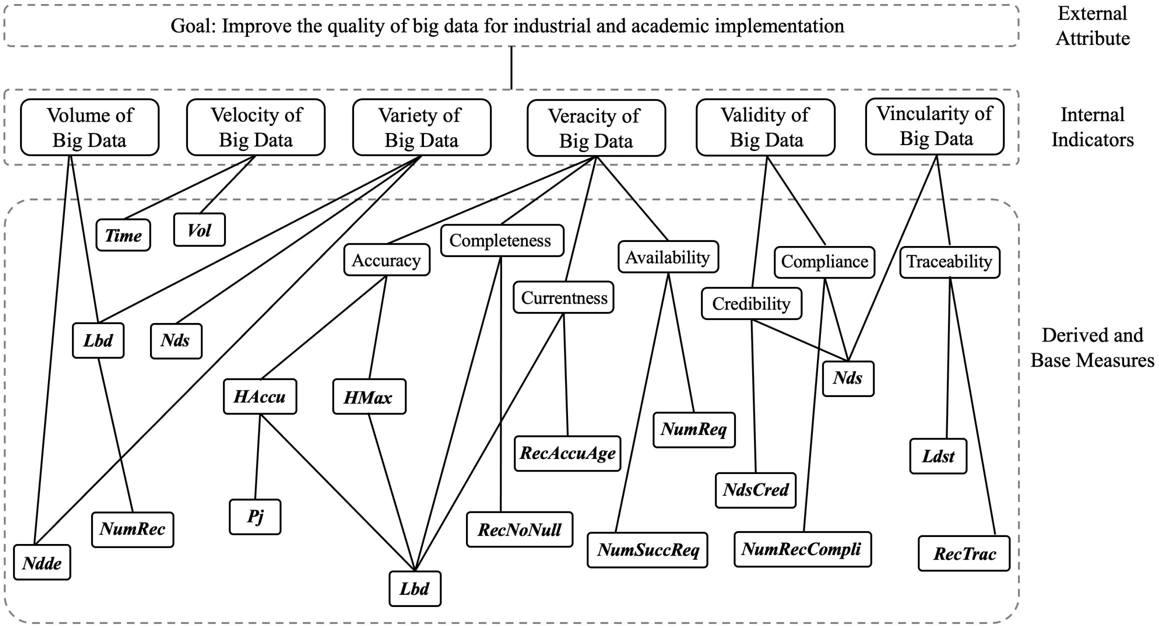

Fig. 1. The big data 6Vs measurement hierarchical model.

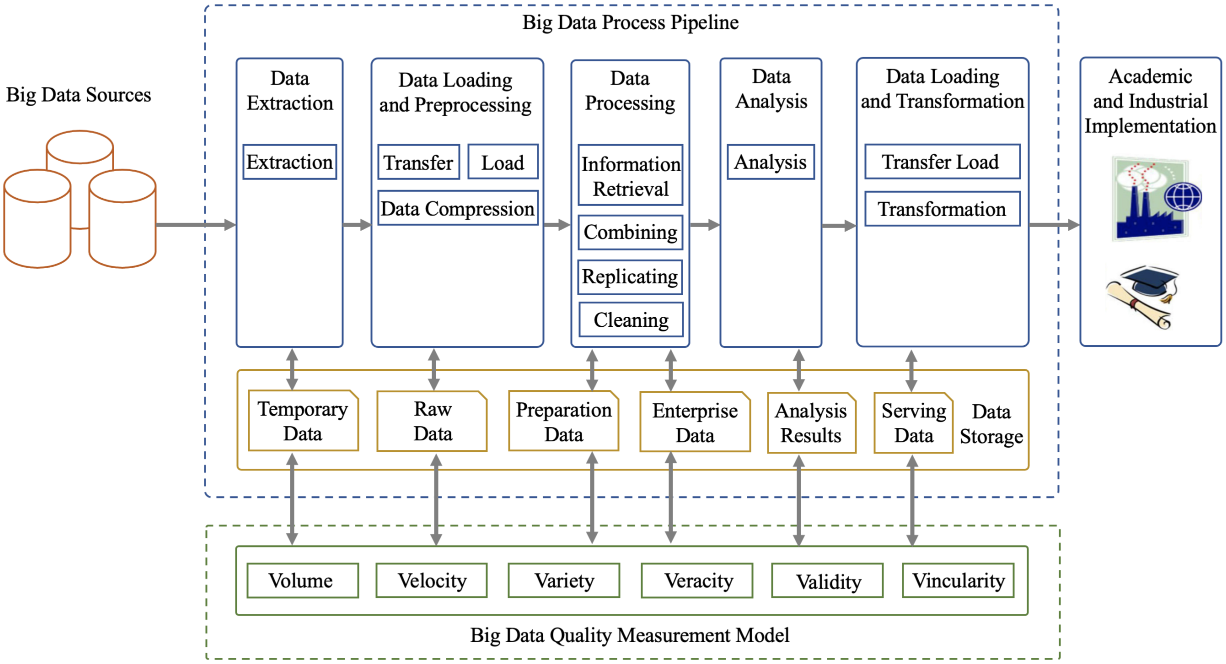

Fig. 2. The designed big data quality measurement architecture.

TABLE II
COMPARISON BETWEEN RESULTS AND THRESHOLDS

Data Characters	**Time Phases**				**Threshold Value**
	T1	T2	T3	T4	
Volume	100%	100.68%	100.94%	98.83%	**33.33%** to **300%**
Velocity	N/A	1:1	1:1.38	1:2.24	**1:3**
Variety	+34505.57	+40960.32	+61575.14	+37983.80	Dynamic
Veracity	+1.87%	+1.80%	+1.77%	+1.83%	Minimum of **95%**
Validity	+18.38%	+18.25%	+18.51%	+18.47%	Minimum of **70%**
Vincularity	+3.85%	+4.12%	+3.20%	+3.96%	Minimum of **95%**

of data completeness and data currentness first experienced moderate growth and then sink in the four phases of time.

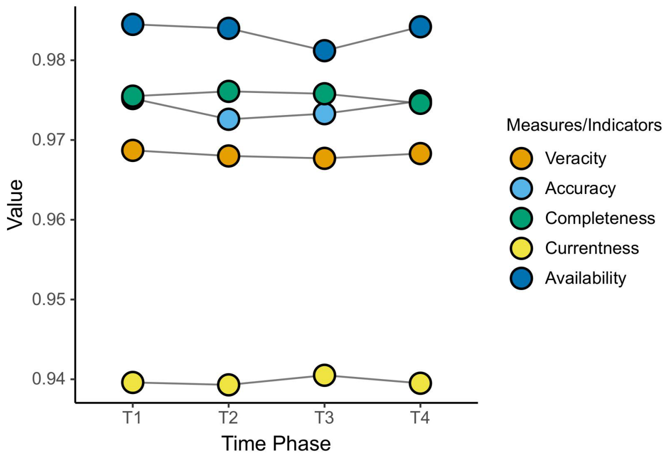

Fig. 3. The derived measures values of the big data veracity indicator.

The derived measures values of the big data validity and vincularity are displayed in “Fig. 4” below. It can be seen that the variation trend of derived measures is basically the same as that of indicators during the four time phases.

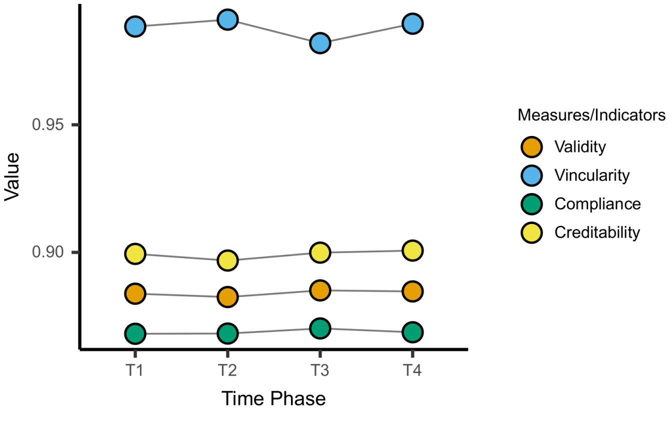

Fig. 4. The derived measures values of validity and vincularity indicator.

B. Comparative Experiment of Model Implementation

We evaluate the practicability and effectiveness of the developed big data quality measurement model on various machine learning and deep learning models for analyzing and classifying short text sentiment values. By comparing the results with the data source before and after data quality check and modification, the impact of the proposed architecture on the results of research in the academic field is dissected.

The experimental data source of the experiment utilize the Twitter Tweets sentiment dataset, a public dataset on the Kaggle platform. There are 24,591 samples in the dataset, of which texts with neutral sentiment value accounted for 40%, texts with positive sentiment value accounted for 31%, and texts with negative sentiment value accounted for 29%.

Five commonly used evaluation indexes are adopted in the experiment, including precision rate, recall rate, accuracy rate, F1-score, and AUC (Area Under Curve). The calculation methods of the first four evaluation indexes are displayed in (25), (26), (27), and (28) as follows.

$$P_{precision} = \frac{TP}{TP + FP} \tag{25}$$

$$P_{recall} = \frac{TP}{TP + FN} \tag{26}$$

$$P_{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{27}$$

$$F1 - score = \frac{2 \cdot P_{precision} \cdot P_{recall}}{P_{precision} + P_{recall}} \tag{28}$$

In the euqations, TP, TN, FP, and FN indicated correct prediction of positive sample number, correct prediction of negative sample number, wrong prediction of positive sample number, and wrong prediction of negative sample number respectively. AUC is the area enclosed under the Receiver Operating Characteristic Curve (ROC). The abscissa of ROC is the false positive class rate, and the ordinate is the true class rate. Their formulas are shown in (29) and (30) as follows. The value of AUC ranges from 0 to 1, and the higher the value, the better the performance of the prediction model.

$$P_{FPR} = \frac{FP}{FP + TN} \tag{29}$$

$$P_{TPR} = \frac{TP}{TP + FN} \tag{30}$$

The comparative experiment results of the model implementation are illustrated in Table III as follows.

TABLE III
COMPARATIVE EXPERIMENT RESULTS OF MODEL IMPLEMENTATION

Model Name	Before Quality Measurement and Improvement				After Quality Measurement and Improvement			
	Precision Rate	Recall Rate	Accuracy Rate	F1-Score	Precision Rate	Recall Rate	Accuracy Rate	F1-Score
Naive Bayes	0.6417	0.6255	0.5232	0.6335	0.7041	0.6693	0.5598	0.6863
Random Forest	0.6283	0.6095	0.5137	0.6187	0.6911	0.6586	0.5548	0.6745
SVM	0.7549	0.7253	0.7795	0.7398	0.8077	0.7839	0.8457	0.7956
CNN	0.8255	0.7981	0.8409	0.8116	0.8985	0.8579	0.9072	0.8777
RCNN	0.8374	0.8032	0.8435	0.8199	0.9076	0.8657	0.9095	0.8862
RNN	0.8501	0.8349	0.8452	0.8424	0.9185	0.9017	0.9124	0.9102
LSTM	0.8595	0.8470	0.8573	0.8532	0.9247	0.9176	0.9275	0.9211
BiLSTM	0.8583	0.8436	0.8583	0.8501	0.9235	0.9106	0.9269	0.9170

Short text sentiment classification experiments based on deep learning models are all designed on the Keras platform, an open-source deep learning library. Keras is a highly modular neural network library written in Python programming language and based on TensorFlow, Theano, and CNTK backends [26]. This deep learning library has the advantage of easy expansibility of modules, through which model prototype design can be carried out simply and quickly. In addition, this paper employs NVIDIA GeForce RTX 3080 graphics cards to train the model in order to speed up the data training process.

In the implementation details, the dropout technique is utilized to prevent the deep learning model neurons from self-adapting and to reduce overfitting. The proportion of dropouts starts from 0.5 and gradually decreases until models perform best, which is 0.2. The number of training epochs is 8. The optimizer used for training is AdamOptimizer [27].

Each model is analyzed twice and the results are recorded, as illustrated in Table III. The first group is the results before data source quality measurement and improvement, which comes from the original multiple datasets in the data source from the platform. The second group is the results after data source quality measurement and improvement, including manual removal and modification of the detected null, incorrect, unacceptable, incredible, and untraceable data.

It can be discovered from the table that after implementing the designed data quality measurement model and modifying the defective data elements, the values of precision rate, recall rate, accuracy rate, and F1-Score will increase in the range of 6% to 10% according to the number of records in datasets. The experimental results indicate that the quality of the big data source has been improved after being altered and updated by the proposed big data measurement framework. A more satisfying outcome can be generated after integrating the constructed big data quality measurement pipeline with academic and industrial implementations. The enhanced big data source with quality improvement can result in better performance in big data academic implementations such as machine learning models for natural language processing.

V. CONCLUSION

This paper proposes a novel rigorous measurement architecture for big data 6 Vs quality characteristics. Algorithms and equations are developed for quantifying based measures, derived measures, and indicators of big data quality. The threshold for each big data quality characteristic is discussed as the guide for big data stakeholders to analyze the results generated from the model. We conduct extensive experiments and precise analyses for evaluating the feasibility, flexibility, practicability, and effectiveness of the architecture. The experiment results demonstrate that the designed model can improve the outcomes of further data source implementation.

REFERENCES

[1] Misra, N.N., Dixit, Y., Al-Mallahi, A., Bhullar, M.S., Upadhyay, R. and Martynenko, A. IoT, big data and artificial intelligence in agriculture and food industry. IEEE Internet of Things Journal, 2020.

[2] Jin, X., Wah, B.W., Cheng, X. and Wang, Y. Significance and challenges of big data research. Big data research, 2(2), pp.59–64, 2015.

[3] Oztemel, E. and Gursev, S. Literature review of Industry 4.0 and related technologies. Journal of Intelligent Manufacturing, 31(1), pp.127–182, 2020.

[4] Gandomi, A. and Haider, M. Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), pp.137–144, 2015.

[5] Ma, G. and Wu, M. A Big Data and FMEA-based construction quality risk evaluation model considering project schedule for Shanghai apartment projects. International Journal of Quality & Reliability Management, 2019.

[6] Taleb, I., Serhani, M.A. and Dssouli, R. Big data quality: A survey. In 2018 IEEE International Congress on Big Data (BigData Congress), pp. 166–173, 2018, July.

[7] Becker, D., King, T.D. and McMullen, B. Big data, big data quality problem. In 2015 IEEE International Conference on Big Data (Big Data), pp. 2644–2653, 2015, October.

[8] Cai, L. and Zhu, Y. The challenges of data quality and data quality assessment in the big data era. Data science journal, 14, 2015.

[9] Abdallah, M., Muhairat, M., Althunibat, A. and Abdalla, A. Big data quality factors, frameworks and challenges. Compusoft, 9(8), pp.3785–3790, 2020.

[10] Ghorbanian, M., Dolatabadi, S.H. and Siano, P. Big data issues in smart grids: A survey. IEEE Systems Journal, 13(4), pp.4158–4168, 2019.

[11] Yu, W., Dillon, T., Mostafa, F., Rahayu, W. and Liu, Y. A global manufacturing big data ecosystem for fault detection in predictive maintenance. IEEE Transactions on Industrial Informatics, 16(1), pp.183–192, 2019.

[12] Alsaig, A., Alagar, V., Chammaa, Z. and Shiri, N. Characterization and efficient management of big data in iot-driven smart city development. Sensors, 19(11), p.2430, 2019.

[13] Xu, X.Application research of accounting archives informatization based on big data. In Data Processing Techniques and Applications for Cyber-Physical Systems (DPTA 2019), pp. 55–62. Springer, Singapore, 2020.

[14] Mikalef, P., Giannakos, M.N., Pappas, I.O. and Krogstie, J. The human side of big data: Understanding the skills of the data scientist in education and industry. In 2018 IEEE global engineering education conference (EDUCON), pp. 503–512, 2018, April..

[15] Laney, D. 3D data management: Controlling data volume, velocity and variety. META group research note, 6(70), p.1, 2001.

[16] Alsaig, A., Alagar, V. and Ormandjieva, O. A critical analysis of the V-model of big data. In 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), pp. 1809–1813, 2018, August.

[17] ISO/IEC/IEEE. ISO/IEC/IEEE 15939:2017 Systems and Software Engineering - Measurement Process. Technical Report. ISO/IEC, 2017.

[18] Taleb, I., Serhani, M.A., Bouhaddioui, C. and Dssouli, R. Big data quality framework: a holistic approach to continuous quality management. Journal of Big Data, 8(1), pp.1–41, 2021.

[19] Ormandjieva, O., Omidbakhsh, M., Trudel, S., Abran, A. and Özcan-Top, O. Measuring the 3V's of Big Data: A Rigorous Approach. In IWSM-Mensura, 2020, October.

[20] Bhardwaj, D. and Ormandjieva, O. Toward a Novel Measurement Framework for Big Data (MEGA). In 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 1579–1586, 2021, July.

[21] Jagadish, H.V. Big data and science: Myths and reality. Big Data Research, 2(2), pp.49–52, 2015.

[22] Shukla, A.K., Yadav, M., Kumar, S. and Muhuri, P.K. Veracity handling and instance reduction in big data using interval type-2 fuzzy sets. Engineering Applications of Artificial Intelligence, 88, p.103315, 2020.

[23] International Organization for Standardization/International Electrotechnical Commission. ISO/IEC 25012: Software engineering-software product quality requirements and evaluation (square)-data quality model. ISO/IEC, 2009.

[24] Majumdar, J., Naraseeyappa, S. and Ankalaki, S. Analysis of agriculture data using data mining techniques: application of big data. Journal of Big data, 4(1), pp.1–15, 2017.

[25] Mikalef, P., Boura, M., Lekakos, G. and Krogstie, J. Big data analytics capabilities and innovation: the mediating role of dynamic capabilities and moderating effect of the environment. British Journal of Management, 30(2), pp.272–298, 2019.

[26] Erickson, B.J., Korfiatis, P., Akkus, Z., Kline, T. and Philbrick, K. Toolkits and libraries for deep learning. Journal of digital imaging, 30(4), pp.400–405, 2017.

[27] Zou, H. and Xiang, K. Sentiment classification method based on blending of emoticons and short texts. Entropy, 24(3), p.398 , 2022.