Measuring the 3V's of Big Data: A Rigorous Approach[1]

Olga Ormandjieva[1] [0000-0001-5641-0976], Mandana Omidbakhsh[2] [0000-0003-0845-6339] and
Sylvie Trudel[2] [0000-0002-4983-1679]

[1] Concordia University, Montreal, Canada
[2] Université du Québec à Montréal, Montreal, Canada

Abstract. Although the success of big data technologies depends highly on the quality of underlying data, no standard measurement model has been yet established for assessing quantitatively the quality of big data. This research aims at investigating thoroughly the quality of big data and laying rigorous foundations for its theoretically valid measurement. We recently proposed a quality measurement hierarchy for methodically selected 10V's of big data, based on the existing ISO/IEC standards, and NIST (National Institute of Standards and Technology) definitions and taxonomies. In this paper, pursuant to our latest research, we derive measurement information model for the most widely used 3V's of big data: Volume, Velocity and Variety. The proposed 3V's measures, declined into a hierarchy of 3 indicators, 2 derived measures and 4 base measures, are validated theoretically based on the representational theory of measurement. Our future research will enhance the theoretical findings presented in this paper with empirical evidences through evaluation of these measures with open-access data.

Keywords: Big Data Quality, Measurement Information Model, ISO/IEC Standards, Volume, Velocity, Variety, Representational Theory of Measurement.

Introduction

Big data refers to the vast amount of digital data stored and originated from different sources of digital and physical environments. Businesses, organizations and more recently governments rely remarkably on the interpretation and analysis of big data to enhance their domain knowledge, make efficient decisions and in consequence, improve their profitability, productivity and performance robustly [1]. Although big data analysis and interpretation depend highly on the quality of the underlying data, there has been no standard measurement model to evaluate the quality of big data. In this paper, we derive a new measurement information model that aims to lessen this gap by proposing valid measures specifically for these 3V's of big data characteristics: (i) Volume, (ii) Variety and (iii) Velocity.

Motivation of this research. Big data has been evolving through three phases along with the generation of the core components ranging from Database Management System (DBMS)-based structured content to web-based, unstructured content and then, in the last decade, to mobile and sensor-based content. Firstly, Relational DBMS (RDBMS) and data warehousing, Extract-Transfer-Loading (ETL), online analytical processing, data mining and statistical analysis have been established. Secondly, information retrieval and extraction, opinion mining, question answering, web/intelligence analytics, social media/network analytics, social network analysis and spatial temporal analysis have been originated. And thirdly, location aware analytics, person centered analysis; content relevant analysis, mobile utilization and human computer interaction have been developed [2].

Why 3V's? The increase of Volume, Velocity and Variety of data contributed majorly to big data evolution; they formed primarily the 3V's of big data characteristics. With the addition of veracity, valence, value, volatility, vitality, validity and vincularity characteristics, the 10v's of big data were shaped.

Challenges. One of the main challenges that research and industry face nowadays is the lack of visibility and transparency of big data's quality: even though researchers and practitioners invest largely in the most sophisticated technologies such as deep learning, big data analytics, training of personal, etc., the statistical and other research findings obtained from big data content are still as good as the underlying data.

Approach. The aim of this research is to bridge the gap between the industrial usage of big data and the underlying big data quality in order to verify its quality for the intended purposes, by bringing practical goal-driven solutions and assessing quantitatively and thus objectively the big data 3V's.

The rest of the paper is organized as follows: in section 2, the background work, including our published hierarchical model for quality measurement of the big data, is briefly described. We define the quality measurements for the big data characteristics of Volume, Velocity and Variety on the basis of NIST (National Institute of Standards and Technology) definitions and taxonomies, and in accordance with ISO/IEC/IEEE Std. 15939 guidelines. In section 4, our quality assessment of the above-mentioned 3V's measures is illustrated on real data. The proposed measures are validated theoretically based on the representational theory of measurement in section 5. The conclusion and future work directions are outlined in section 6.

2 Background and Related Work

2.1 The 3V's of Big Data

Although big data has been commonly characterized by many different criteria known as V's of big data, there are three main characteristics that are generally agreed upon, known as 3V's of big data: Volume, Velocity and Variety, as illustrated in Fig. 1.
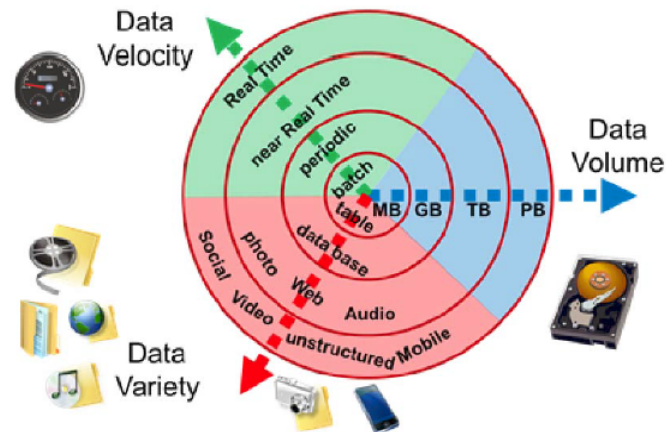
Fig. 1. The 3V's of Big Data, extracted from [7].

Volume. It refers to the magnitude of data [8].

Velocity. It refers to the speed at which data is being generated, which can be done in the following ways:

- Real-time is when the data is generated immediately such as streaming, radar systems, customer service systems, and bank ATMs. There is the continual input, constant processing and steady output.
- Near real-time is when the speed is important but the data is not generated immediately, mostly for the production of operational intelligence, which is a combination of data processing and CEP (Complete Event Processing) combining data from multiple sources in order to detect patterns such as: sensor data processing, IT systems monitoring, financial transaction processing.
- Batch is when the data is generated in delays such as payroll, billing, data analysis from operational data, historical and archived data, data from social media, service data, etc.

Variety. It refers to the ever-increasing different forms of data as in text, images, voice, and geospatial. The different types of data are categorized as structured, unstructured and semi-structured data.

- Structured data refers to the data stored in databases in an ordered manner. The data in library catalogues (such as: date, author, place, subject, etc.), and economic data (such as: GDP, PPI, ASX) are considered structured.

- <u>Unstructured</u> data refers to any data with unknown form/structure. The data in the form of media (mp3, digital photos, audio, video), text (word processing, spreadsheets, presentations) and social media (data from Facebook, twitter, LinkedIn) are considered unstructured.
- <u>Semi-structured</u> data refers to a form of structured data that does not conform with the formal structure of data models as in RDBMS or other data tables, but contains tags or other markers which separate semantic elements and enforce hierarchies of records and fields within the data. It is also known as self-describing structure. An example of this form of data can be the personal data stored in an xml file as:
  ```
  <rec>
  <name>Harry</name>
  <sex>Female</sex>
  <age>23</age>
  </rec>
  ```

2.2 Our Proposed Quality Measurement Model

Big data analysis and interpretation depend highly on the quality of data as an eminent factor for its maturity [3]. Although NIST has developed taxonomy towards to the standardization of big data technology [4], in which the characteristics of data are pivoted at different levels of granularity; there has been no standard measurement model in order to assess quantitatively the quality of big data, its analysis and interpretation techniques.

Therefore, we proposed recently our new hierarchical goal-driven measurement model for the 10V's of big data characteristics [4]. We adopted the NIST taxonomy [4], in which a hierarchy of roles/actors and activities including data elements as the smallest level, records as groups of elements, dataset as groups of records and finally multiple datasets, are provided at different levels and defined as below:

"Data elements are individual data elements with the same definition in the big data paradigm and are populated by their actual value, constrained by its data type definition (e.g.: numeric, string, date) and chosen data format. The actual value can be constrained by a set of allowed values or a specific standard vocabulary for interoperability with others in the field. For example, in the context of unstructured text, a data element would refer to a single token such as a word. Records are groups of data elements that describe a specific entity or event or transaction. Records have structure and they are grouped as structured, semi-structured and unstructured, as with the increase of mobile and web data (e.g.: online texts, images and videos) more emphasis is on unstructured data, as so a data record can refer to phrase or sentence or entire document data in context of unstructured data. Records can be grouped to form datasets. In the context of unstructured text, a record could be a sentence, paragraph, or section and the dataset could refer to the complete data. Multiple datasets are group of datasets with the emphasis on the integration and fuse of data. The variety characteristic of big data is concerned at this level." [11]

In our quality measurement model, we related each of the 10V's of big data characteristics to their corresponding levels in the NIST taxonomy, namely: data element, record, dataset and multiple dataset. We adapted the ISO/IEC DIC 25024 international standard's data quality measures in order to define the quality model on the basis of the 10V's of big data. Fig. 2 shows an overview of our approach to eliciting measurements for assessing the big data characteristics (the 10V's).
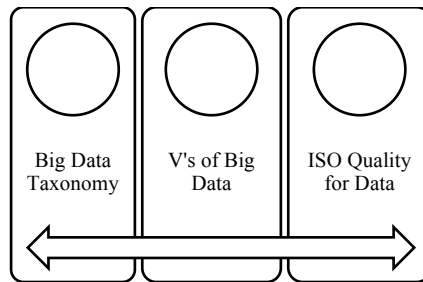

Fig. 2. Overview of our Approach to Big Data Quality Measurement [4]

The quality model is tailored in a way that facilitates the evaluation of such systems in terms of ISO/IEC standard 25024's measures of data's characteristics Availability, Accuracy, Accessibility Credibility, Completeness, Compliance, Currentness, Efficiency, Portability, Traceability and Understandability. The validity of the proposed big data quality measurement model is rooted in the standardization of: i) the NIST big data taxonomy, and ii) data measurements defined in ISO/IEC. For a more detailed explanation of our new hierarchical quality model designed specifically for the purpose of measurement of quality for the selected 10v's of big data, please refer to [4].

According to our research findings in [4], for the big data characteristics of Volume, Velocity and Variety (the 3V's), there is no related measurements found in the existing ISO/IEC standards. In this paper, we propose a new measurement information model for quality assessment of the above-mentioned 3V's.

3 Quality Measurements for 3V's of Big Data

The goal of this section is to derive theoretically valid measures for 3V's characteristics of big data, namely, Volume, Velocity and Variety.

3.1 Mathematical Modeling of the Measurement Entities

Theoretically valid measure is founded on mathematical modeling of the entities of interest. According to the ISO/IEC/IEEE Std. 15939, an object that is to be characterized by measuring its attributes is named "entity" [10]. In this work, the entities of

interest correspond to the hierarchical levels of the NIST hierarchy (data element, record, dataset, and multiple datasets). We undertake set-theoretical approach to modeling these NIST hierarchy elements, as described next:

Data Elements. As explained in section 2, the data elements in big data originate in heterogeneous nature, including attributes from traditional databases, and newer, for instance, text from social media and sensors data. To be able to model a collection of heterogeneous data elements as a set, we first label each data element with a unique identifier (UID_E). This also can be traced to our view of Variety discussed later on in section 3.2. We state as universe a fixed set of all distinct data elements in the multiple datasets and form a set we refer as *DE* of the UID_Es of all distinct data elements. Every reference to a data element below is to be interpreted as an indication of its UID_E.

Record. Data elements are stored in *records*. Informally, a record can be seen as a collection of data elements. Every record is referred to by a unique record ID (UID_R). As explained in section 2, records in big data originate in heterogeneous sources, including traditional databases, and newer, less structured sources like social media and etc. therefore record can refer to a phrase or entire document data in context of unstructured data. We model mathematically a record ***r*** as a *multiset,* which may be formally defined as a 2-tuple (*DEr, m*) where *DEr* is the *underlying set* of the multiset formed from its distinct elements (*DEr* $\subseteq$ *DE*). The *multiplicity m: DEr* $\rightarrow N^+$ is a function from *DEr* to the set of the positive integers, giving the number of occurrences of each element $el \in DEr$ as the number *m*(*el*).

Dataset. The term *dataset* refers to a collection of one or more records. We model a dataset ***DS*** as a set of records' unique identifiers (UID_R). Every unique dataset is referred to by a unique dataset ID (UID_{DST}).

Multiple datasets. A Big data is viewed as multiple datasets and thus can be formally modeled as a set of datasets ***MDS*** (in mathematical terms, as a set of multisets).

The aim of this section 3.1 was to propose a mathematical model based on which to define the 3V's measurement hierarchy. Such approach is justified by the fact that mathematical models would greatly simplify the automation of the measurement procedures. It is to be noted that the automation of the 3V's measurements is out of the scope of this paper and will be tackled in our future work.

The measurements built hierarchically upon the mathematical model are described in section 3.2.

3.2 Proposed 3V's Measurement Information Model

In section 3.2 we closely follow the terminology and guidelines outlined in the International ISO/IEC/IEEE Standard 15939 on measurement processes in software engineering [10]. Fig. 3 illustrates the hierarchical relationships among the key components of

the proposed measurement information model. The model defines three types of measures: base measures, derived measures, and indicators as detailed below.

Base Measures. A base measure is defined in ISO/IEC/IEEE Std. 15939 as functionally independent of other measures [10]. The base measures in the 3V's measurement model are depicted at the lowest level of the hierarchy (see Fig. 3). Their definitions are provided next:

Number of distinct data elements (Ndde). Ndde reflects the Variety of the data elements, as stated in section 3.1 (see the data element mathematical modeling). The measurement method for the base measure Ndde is counting of the number of distinct labels (UID_E) of data elements in the set DE, formally defined as Ndde (DE) = |DE| (cardinality of set DE, which is the total number of UID_E in multiple datasets). The scale type of the measurement is absolute ([9]) and the measurement unit here is a UID_E.

Number of records in a dataset (Nrec). Nrec assesses the Variety of a dataset in terms of diversity of records and their sources. This base measure is defined formally as Nrec (DS) = |DS| (cardinality of set DS formed by unique identifiers of records (UID_R), that is, the number of records in the dataset DS). The scale type of the measurement is absolute ([9]) and the measurement unit of Nrec is a UID_R; the corresponding measurement method is counting.

Number of datasets in big data (Nds). Nds reflects an aspect of Variety in terms of diversity of datasets in multiple datasets (MDS). The measurement method for the base measure Nds is counting the total number of unique identifiers UID_{DST} of datasets in the multiple datasets. Formally, Nds (MDS) = |MDS|. The scale type of the measurement is absolute ([9]) and the measurement unit is a UID_{DST}.

Time (T). T models the absolute time. It is required to accompany the measurement of the Volume at a specific time in order to calculate Velocity and understand big data growth over time.

Derived Measures and Indicators. Derived measure is defined as a measurement function of two or more values of base and derived measures [10]. According to ISO/IEC/IEEE Std. 15939, the indicators of the 3V's are defined here as measures that provide an evaluation of the big data characteristics of Volume, Velocity and Variety derived from the big data measurement needs stipulated earlier (see the introduction section). These indicators serve as a basis for the analysis of big data quality and decision-making based on an interpretation of the results of measurement.

The derived measures and indicators defined in this research are specified next:

Length of big data (Lbd). Lbd is defined informally in this work as the total number of records in MDS. The measurement formula is as follows:

$$Lbd\ (MDS) = \sum \forall DS \in MDS\ \underline{Nrec}\ (DS)$$

The measurement unit of Lbd is a UID_R.

Big Data Volume (Mvol). In this research we define informally Volume in terms of the information content of multiple datasets, that allows us to apply information-theoretic measuring of Volume in mathematical bits, and thus make it possible to define Velocity measure of big data. In information theory, the definition of the amount of information is often expressed with the binary logarithm, corresponding to making the bit the fundamental unit of information.

Assume that there are n elements in the set DE (that is, Ndde (DE)=n). Each record may, or may not, include each individual element from the set DE, thus there are n binary decisions of Yes/No to be made for each record in terms of selecting data elements. This is equivalent to 2^n choices per record. Each binary decision can be represented by an information bit. Thus, binary logarithms can be used to calculate the length of the representation of the number of information bits needed to encode a record in information theory.

In general, in engineering applications we do not take the logarithm of a dimensioned number, only of dimensionless quantities [13]. To avoid this problem, we define a function *bin_dec: Ndde (DE)* $\rightarrow \mathbb{N}$ to convert Ndde (DE) measured in UID_E, into a dimensionless quantity represented by a Natural number.

This approach allows us to apply information-theoretic measuring of Volume in information bits, which is independent from the specific technologies used to store the masses of raw data.

Mvol of multiple datasets (MDS). It is defined formally as:

$$Mvol\ (MDS) = Lbd\ (MDS)\ log_2\ (bin\_dec\ (Ndde\ (DE)))$$

Mvol measures the number of information bits across all records required to specify the information content of multiple datasets. The measurement unit is an information bit; thus, the measurement results of Volume allow to compare objectively multiple datasets in terms of their information content. Hence, the Volume measure proposed in this work gives organizations the ability to objectively assess and compare multiple datasets in terms of their information content, expressed in information bits.

Trend of Mvol depicts graphically Volume of big data over time.

Big Data Velocity (Mvel). We define informally the notion of Velocity of big data (set of multiple datasets) in terms of a relative growth of big data over a period of Time (T), that is, the speed of increase in big data Volume.

The Velocity measure function is defined as:

$$Mvel\ (MDS) = ((Mvol\ (MDST1) - Mvol\ (MDST2))\ /\ Mvol\ (MDST1)\ *100$$

where MDS_{T1} is the multiple datasets at time T1 and MDS_{T2} represents the multiple datasets at time T2 in which T2>T1. Therefore, Mvol (MDS) is expressed as percentage of volume growth over interval of time T2-T1 along with its adequate unit of measure (seconds, minutes, hours, weeks, etc.).

Velocity can be used to assess organizations in understanding the relative growth of their big data's information content over a period of time. Negative values indicate that the data has been archived or removed from the source, or is no longer relevant. In other words, negative values of Velocity show that data's life span is over.

Big Data Variety (Mvar). In our work we consider a 3-fold root cause of Variety: i) the diversity of data elements in multiple datasets (MDS), ii) the diversity of data records in MDS, and iii) the diversity of datasets in MDS.

Mvar (MDS) is defined as 3-tuple (Ndde (DE), Lbd (MDS), Nds (MDS)) that reflects correspondingly the diversity of unique data elements, diversity of records, and diversity of datasets in MDS. As mentioned earlier, the diversity of data elements is inherently captured in the DE set where the data elements are mapped to UID_E, thus abstracted from their diverse sources. Mvar allows to objectively compare different multiple datasets in terms of the above three objective measures.

The measurement information model we proposed in section 3.2 is based on the ISO/IEC/IEEE Std. 15939 terminology and guidelines. The objective of the next section is to present graphically this new hierarchical measurement model tailored specifically to the 3V's of big data Volume, Velocity and Variety.

3.3 Hierarchy of the 3V's Measures

The measurement information model proposed in section 3.2 is a hierarchical structure linking information needs to the relevant entities and attributes of concern, such as number of distinct data elements, number of distinct records, number of distinct datasets, or length. It defines how the relevant attributes are quantified and converted to indicators that provide a basis for decision making. In our approach, the 3V's characteristics were decomposed through three layers as depicted in Fig. 3.
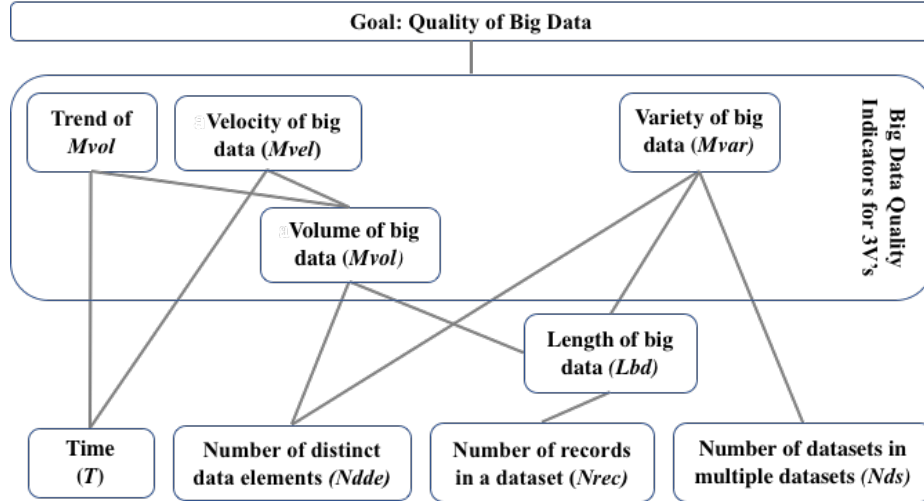
Fig. 3. Hierarchy of Big Data 3V's Measures.

In section 4 we demonstrate the measurement approach with a short example used for illustrative purposes only.

4 Measurement Illustration

A sample dataset of Facebook users is provided in Fig. 4 to illustrate the measurement data collection and analysis procedures.

We first assign unique identifiers to the distinct data elements that needs to be distinguished. For instance, Age21, Age19, Age9, Age33, Age22, Age25, Age5, Age26, Age27, Age33, Age35 will represent the different data elements in the Age category, total of 11. Nationalities can be represented by N_S (for Saudi), N_P (for Pakistan), and N_Y (for Yemen) (3 in total). The 4 different values in the column Own Pic/week can be represented by the labels O7, O2, O3 and O8 correspondingly. The 3 different types of exposure are mapped to E_H, E_M and E_L. The collection St_3, St_12, St_7, St_2, St_8, St_22, St_33 labels the different Status/week data elements, 7 in total. Ratio distinct values can be mapped to R0.1, R5, R4, R3, R1, R7, R2 correspondingly (7 in total). The union of all unique identifiers forms the set DE={Age21, Age19, Age9, Age33, Age22, Age25, Age5, Age26, Age27, Age33, Age35, N_S, N_P, N_Y, O7, O2, O3, O8, E_H, E_M, E_L, St_3, St_12, St_7, St_2, St_8, St_22, St_33, R0.1, R5, R4, R3, R1, R7, R2}. Thus, the value of the base measure Ndde (DE) is 35 UIDs.

Similarly, the 20 records listed in the sample dataset 1 can be labeled as REC1 ... REC20. Thus, the number of the distinct records in this dataset (Nrec) is 20 UID_R. The number of the datasets in this example is 1 (Nds = 1 UID_{DST}). The corresponding value of length of big data (Lbd) is 20 UID_R (the total number of records in MDS).

Age	Nationality	Own Pic/week	Exposure	Status/week	Ratio
21	Saudi	7	High	3	0.1
19	Pakistan	2	High	12	5
33	Yemen	7	High	2	4
5	Saudi	2	High	7	0.1
19	Pakistan	2	Medium	12	3
25	Yemen	3	Medium	3	4
22	Saudi	7	Medium	7	0.1
33	Pakistan	8	Medium	8	5
21	Saudi	7	Medium	22	5
27	Saudi	8	Medium	33	1
22	Saudi	2	Low	22	7
26	Saudi	8	Low	2	2
21	Saudi	7	High	3	0.1
19	Pakistan	2	High	12	5
33	Yemen	7	High	2	4
35	Saudi	2	High	7	0.1
19	Pakistan	2	Medium	12	3
25	Yemen	3	Medium	3	4
22	Saudi	7	Medium	7	0.1
33	Pakistan	8	Medium	8	5

Fig. 4. Sample Data (dataset 1)

Big Data Volume Indicator (Mvol) is calculated as follows:

Mvol (MDS$_{TI}$) = Lbd (MDS$_{TI}$) log$_2$ (bin_dec (Ndde (DE$_{TI}$))) = 20 log $_2$ (35)

which is 103 information bits.

In order to illustrate the quantification of Velocity of big data (Mvel) indicator we need to compare data in two different instances of time, T1 and T2. Assume that the following records were added up to dataset 1 (see Figure 6) by time T2, in addition to the dataset 1 records (see Fig. 5) that were measured in time T1 (T2>T1):

21	Saudi	7	Medium	22	5
27	Saudi	8	Medium	33	1
22	Saudi	2	Low	22	7
26	Saudi	8	Low	2	2

Fig. 5. Additional Entries to the Sample Data

Lbd (MDS $_{T2}$) has increased to 24 records, while the set DE remains the same: (DE$_{T1}$ = DE$_{T2}$). The Volume of big data after the change is:

$$Mvol\ (MDST2) = Lbd\ (MDST2)\ log2\ (bin\_dec\ (Ndde\ (DE\ T2))) = 24\ log\ 2\ (35)$$

which is 124 information bits. Therefore, Velocity of big data (Mvel) shows an increase of:

$$Mvel\ (MDS) = (124 - 103)/\ 103\ *100 = 20.4\%.$$

In this example, the Variety Indicator of big data Mvar (MDS) is (35 data elements, 20 records, 1 dataset).

The major requirement in measurement is that the measurements are defined consistently with the entity's real-world behavior to ensure they accurately measure those attributes they purpose to quantify. Our view of measurement validation is explained in section 5.

5   Validation of the 3V's Measures

Validation is critical to the success of big data measurement. Measurement validation is "the act or process of ensuring that (a measure) reliably predicts or assesses a quality factor" [12]. In other words, a given measure is valid if it reflects the real meaning of the concept under consideration and is based on the representational theory of measurement.

Two approaches to validation have been prescribed and practiced in software engineering: (a) theoretical validation, and (b) empirical validation [9]. These two types of validation are respectively used to demonstrate that a measure is really measuring the attribute it is purporting to measure.

5.1   Theoretical Validation vs. Empirical Validation

The 3V's measures are theoretically validated with respect to *Tracking* and *Consistency* criteria introduced in [12], as described below:

The Tracking Criterion. This criterion assesses whether or not a measurement is capable of tracking changes in product or process quality over the life cycle of that product or process [12]. A change in the attributes at different times should be accompanied by a corresponding change in the measurement data. It can be expressed formally as follows:

If a measure M is directly related to a quality characteristic F, for a given product or process, then a change in a quality characteristic value from F_{T1} to F_{T2}, at times $T1$ and

T2, shall be accompanied by a change in the measurement value from M_{T1} to M_{T2}. This change shall be in the same direction (e.g. if *F* increases, *M* increases). If *M* is inversely related to *F*, then a change in *F* shall be accompanied by a change in *M* in the opposite direction (e.g. if *F* increases, *M* decreases).

The Consistency Criterion. This criterion assesses whether or not there is a consistency between the ranks of the characteristics of big data quality (3V's) and the ranks of the measurement values of the corresponding indicator for the same set. It is used to determine whether or not a measurement can accurately rank, by quality, a set of products or processes [12]. The change of ranks should be in the same direction in both quality characteristics and measurement values, that is, the order of preference of the 3V's will be preserved in the measurement data and can be expressed as follows:

If quality characteristic values F_1, F_2, …, F_n, corresponding to multiple datasets 1 … n, have the relationship $F_1 \succ F_2 \succ … \succ F_n$, then the corresponding indicator values shall have the relationship $M_1 > M_2 > … > M_n$.

This preservation of the relationship means that the measure must be objective and subjective at the same time: objective in that it does not vary with the measurer, but subjective in that it reflects the intuition of the measurer [9].

Tracking and consistency are a way to validate the representational condition without collecting and analyzing large amounts of measurement data, thus can be done manually.

Empirical validation is a process for establishing software measurement accuracy by empirical means.

Ultimately, it is clear that both theoretical and empirical validation, are necessary and complementary.

In this paper, we target manual theoretical validation of the 3V's measurements. The reason is that the empirical validation of big data's measurements would require large amount of measurement data that cannot be collected manually. As stated in the introduction section of this paper, the implementation of the proposed measurement data collection and the empirical validation of the proposed measurement procedures will be tackled in our future work.

5.2 Theoretical validation of the 3V's Indicators

Mvol. Based on the meaning of Volume, the more information big data contains (that is, the higher the information content of the big data), the larger the Mvol indicator value. The perception of 'more' should be preserved in the mathematics of the measure: if we increase the information content of big data, its Volume will increase, as expected.

For instance, Mvol (MDS$_{T1}$) =103 information bits. In the case of MDS$_{T2}$, the dataset is larger than that of MDS$_{T1}$ and consequently Mvol (MDS$_{T2}$) is higher (124 > 103).

Mvel. The above expectation is also confirmed by the indicator Mvel, which shows the anticipated increase (20.4%).

Mvar. The Variety of MDS$_{T2}$ is not expected to differ much: there was no change in the set DE in time T2, also, there is still one dataset only. The is reflected by the indicator values: Mvar (MDS$_{T2}$) is (35 UID$_E$, 24 UID$_R$, 1 UID$_{DST}$), which shows an increase in the MDS length parameter only, as compared to Mvar (MDS$_{T1}$).

These calculations establish the theoretical validity of the big data 3V's measures, as required by the representational theory of measurement. In our future work, the 3V's measures will be validated empirically through controlled experiments.

6   Conclusion and Future Work

In this paper, we proposed a new measurement information model to quantify three aspects of Big Data – Volume, Variety, and Velocity, known as the 3V's indicators. Four levels of entities have been considered, derived from the NIST hierarchy: data element, record, dataset, and multiple datasets. The model elements are compliant with ISO/IEC/IEEE Std. 15939 guidelines for their definitions, where four base measures are first defined, assembled into two derived measures, evolving into three indicators, thus, the 3V's. Theoretical validation of these 3V's have been demonstrated. The model is suitable for big data in any forms of structured, unstructured, and semi-structured.

As anyone may inquire about the relevance of such model for the industry, we can respond with simple examples of usage of these measures and indicators:

- Velocity (Mvel): The data owner can oversight the growth of volume over time. A slower growth than expected might trigger investigation as a source of data could be damaged or unavailable. A faster growth than expected could also be an indication of a corrupted source resending data multiple times. Both cases would affect data quality.
- Volume (Mvol) and its trend: Volume is useful to objectively compare multiple datasets in terms of their information content, as well as to oversight their growth over time.
- Variety (Mvar): Variety reflects the structural heterogeneity at different levels (data elements, records and datasets) thus allows to easily and objectively compare multiple datasets in terms of their Ndde (number of distinct data elements), Nrec (number of records) and Nds (number of datasets) values.

Our future research will enhance the theoretical findings presented in this paper with empirical evidences through evaluation of these measures with open-access data and

industry data. The automation of the 3V's measurements is out of the scope of this paper and will be tackled in our future work.

References

1. Agbo, B. et al.: Big data: the Management Revolution), , https://hbr.org/2012/10/big-data-the-management-revolution (2018)
2. Mcfee, A., Brynjolfsson, E.: of The Second Machine Age, https://www.bigdataframe-work.org/short-history-of-big-data/), October 2012 issue of Harvard Business Review.https://www.amazon.com/Second-Machine-Age-Prosperity-Technolo-gies/dp/0393350649 (2012).
3. Taleb, I., Serhani, M., Dssouli, R.: Big Data Quality: A Survey. 10.1109/BigDataCon-gress.2018.00029 (2018).
4. Omidbakhsh, M., Ormandjieva, O.: Toward A New Quality Measurement Model for Big Data, DATA 2020, The 9th International Conference on Data Science, Technology and Applications (2020).
5. Rouse, M., Agenda, I.: The 4th Industrial Revolution: Hidden Threats to Human and Cyber-security, Servamus Community-based Safety and Security Magazine (2019).
6. Thompson, K., Mattalo, B., The Internet on Things: Guidance, Regulation and the Canadian Approach, CyberLex (2016).
7. Hilbert, M., Lopez, P.: The World's Technological Capacity to Store, Communicate and Compute Information, Science, 332(6025), 60-65 (2011).
8. Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. International journal of information management. 35(2):137-44 (2015).
9. Fenton, N., Bieman, J.: Software Metrics: A Rigorous and Practical Approach, 3rd edn. CRC Press. https://doi.org/10.1201/b1746 (2014).
10. ISO/IEC/IEEE 15939, 2017-04, Systems and Software Engineering-Measurement Process (2017).
11. NIST U.S. Department of Commerce, Special Publication 1500-1, NIST Big Data Interop-erability Framework: Volume1, Definitions. Volume2, Big Data Taxonomies (2018).
12. IEEE Std 1061. IEEE Standard for a Software Quality Metrics Methodology (1998).
13. Al Qutaish, R.E. and Abran, A.: An analysis of the design and definitions of Halstead's metrics, In 15th Int. Workshop on Software Measurement (IWSM'2005). Shaker-Verlag, 337-352 (2005).