Measurement Framework for Assessing Quality of Big Data (MEGA) in Big Data Pipelines

Dave Bhardwaj

A Thesis
In the Department of
Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Computer Science and Software Engineering at
Concordia University
Montreal, Quebec, Canada

August 2021

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Dave Bhardwaj

Entitled: Measurement Framework for Assessing Quality of Big Data (MEGA) in Big Data Pipelines

And submitted in fulfillment of the requirements of the degree

Master of Computer Science

Complies with the regulations of the university and meets the accepted standards with respect to originality and quality

Signed by the final examining committee:

_____ Chair
Dr. Tristan Glatard

_____ Examiner
Dr. Tse-Hsun (Peter) Chen

_____ Examiner
Dr.Tristan Glatard

_____ Supervisor
Dr. O. Ormandjieva

Approved by: _____
Chair of Department or Graduate Program Director

_____ 2021 _____
Dean of Faculty

Abstract For Masters

Measurement Framework for Assessing Quality of Big Data (MEGA) in Big Data Pipelines

Dave Bhardwaj

Big Data is used widely in the decision-making process and businesses have seen just how powerful data can be, especially for areas such as advertising and marketing. As institutions begin relying on their Big Data systems to make more informed and strategic business decisions, the importance of the underlying data quality becomes extremely significant. In our research this is accomplished through studying and automating the quality characteristics of Big Data, more specifically, through the V's of Big Data.

In this thesis, our aim is to not only present researchers with useful Big Data quality measurements, but to bridge the gap between theoretical measurement models of Big Data quality characteristics and the application of these metrics to real world Big Data Systems. Therefore, our thesis proposes a framework (The MEGA Framework) that can be applied to Big Data Pipelines in order to facilitate the extraction and interpretation of Big Data V's measurement indicators. The proposed framework allows the application of Big Data V's measurements at any phase of the architecture process in order to flag quality anomalies of the underlying data, before they can negatively impact the decision-making process. The theoretical quality measurement models for six of the Big Data V's, namely Volume, Variety, Velocity, Veracity, Validity, and Vincularity, are currently automated.

The novelty of the MEGA approach includes the ability to: i) process both structured and unstructured data, ii) track a variety of quality indicators defined for the V's, iii) flag datasets that pass a certain quality threshold, and iv) define a general infrastructure for collecting, analyzing, and reporting the V's measurement indicators for trustworthy and meaningful decision-making.

ACKNOWLEDGMENTS

I would like to express my thanks and gratitude to my thesis advisor Dr. Olga Ormandjieva for her support, guidance, and encouragement throughout my degree. I have learned a lot from her and wish her nothing but the best.

I would also like to acknowledge and express gratitude for all the frontline workers and health care providers of the COVID-19 pandemic. They've risked their lives to make it possible for so many of us to continue to work and succeed.

Lastly, I would like to thank my family and friends for their constant support and motivation throughout my entire life, I'm very lucky to be blessed with so many wonderful people.

Table of Contents

List of Figures

Chapter 1 Introduction

Big data has fundamentally changed how businesses leverage both their own data and that of others. Consequently, this has given businesses the ability to make better and more data-driven decisions, gain a deeper understanding of their customers and focus their resources to improve both productivity and profits (Walker, 2015) (Lee, 2013) Big data isn't just for businesses though. A large variety of sectors and industries benefit greatly from the use of big data, from the healthcare industry to even agriculture and farming. Big data can be used as one of many tools to solve some of the most important problems in industries like healthcare and agriculture (Andreu-Perez, Poon, Merrifield, Wong, & Yang, 2015) (Yu & Song, 2016) (Kelly & Knezevic, 2016). Being able to model and predict health assessment from electronic health records is one the many (Andreu-Perez, Poon, Merrifield, Wong, & Yang, 2015) (Bates, Saria, Ohno-Machado, Shah, & Escobar, 2014) kinds of advancements that can be made, by leveraging data.

While Big Data can allow industries that have access to a vast amount of data the ability to make critical decisions, data isn't always necessarily perfect itself. Data scientists and software engineers spend large amounts of time and effort developing complex architectures to capture, process and store data so that it can be ready to be used for analysis and decision-making. However, models built on even processed data may still be imperfect and even less than ideal if developers can't be sure of the quality of their data (Gudivada, Apon, & Ding, 2017).

The objective of this thesis is to bridge the gap between Big Data Quality Measurements and Big Data Systems. More specifically, there exists known characteristics that help us understand the underlying data quality. We wish to build measures for these existing characteristics and have a framework that can be used to apply them on systems. We present a reference framework that can be applied to a variety of Big Data Systems for on-going Big Data quality measurements. In this framework, developers and users can collect metrics on the underlying data to be evaluated by measures chosen by them. They can then use this information to determine the meaningfulness of the underlying data.

1.2 Motivation of This Research

Big Data scientists and engineers invest large amounts of resources developing complex architectures to capture, process and store Big Data so that it can be used for strategic decision-making. Yet, while Big Data allows critical decisions to be made based on the analysis of vast amounts of data, the underlying data isn't always perfect and thus can lead to costly mistakes (Gudivada, Apon, Ding 2017). Hence, models built on processed Big Data are as good as the quality of the underlying data. Consequently, the visibility of the underlying data quality is becoming increasingly important.

Although there have been advancements in measuring the Big Data V's ((Ormandjieva, Omidbakhsh, & Trudel, 2020), (Ormandjieva, Omidbakhsh, & Trudel, 2021)) we still lack a coherent framework in which these measurements can be applied to real Big Data systems. The main contribution of this thesis is the proposed architectural solution for providing continuous measurement feedback on quality characteristics of Big Data (the V's), including visualization and interpretation of the measurement results, in a manner that is flexible and easy to use. Big Data users can use the V's measurement results in order to assess the quality of their underlying data and determine its suitability for their purposes.

The problems we are attempting to solve here include the ability to: i) process both structured and unstructured data, ii) track a variety of base measures depending on the quality indicators defined for the V's, iii) flag datasets that pass a certain quality threshold, and iv) define a general infrastructure for collecting, analyzing, and reporting the V's measurement results for trustworthy and meaningful decision-making.

1.3 Challenges

According to the U.S. National Institute of Science and Technology (NIST) Big Data Public Working Group (NIST Big Data Interoperability Framework: Volume 1, Definitions. Volume2, Big Data Taxonomies, 2018) Big Data does not only refer to the increasingly large multiple datasets, but also to the fundamental improvements in the architecture needed to manage the

quality of this data. Recent publications have proposed frameworks and architectures to address the different needs for data quality analysis ((Pääkkönen & Ovaska 2015), (Taleb, Dssouli & Serhani 2015), (Ramaswamy, Lawson & Gogineni, 2013), (Veiga, Saraiva, Chapman, et al. 2017), (Merino, Caballero, Rivas, Serrano & Piattini 2017)). However, as Big Data Systems handle larger and larger amounts of data, with more and more automation, the need for monitoring quality becomes indispensable for the users of Big Data.

1.4 Approach and Novelty

Approach. In this thesis we address this need by proposing a novel quality measurement framework for Big Data (MEGA) where data issues can be identified and analyzed continuously by integrating data quality measurement procedures within Big Data Pipelines phases. The goal is to flag data quality issues before they propagate into the decision-making process.
In this research we focus on ten of the intrinsic Big Data quality characteristics referred to as V's, namely: Volume, Variety, Velocity, Veracity, Vincularity, Validity, Value, Volatility, Valence and Vitality (Omidbakhsh & Ormandjieva 2020). Measurement information models were proposed previously for Volume, Velocity, Variety and Veracity ((Ormandjieva, Omidbakhsh, & Trudel, 2020), (Ormandjieva, Omidbakhsh, & Trudel, 2021)). The remaining 6 V's (Vincularity, Validity, Value, Volatility, Valence and Vitality) that lack properly defined valid measurements and will be tackled in our future work.

Novelty. Although there have been advancements in measuring the Big Data V's ((Ormandjieva, Omidbakhsh, & Trudel, 2020), (Ormandjieva, Omidbakhsh, & Trudel, 2021)), we still lack a coherent framework in which these measurements can be applied to real Big Data Systems. The main contribution of this thesis is the proposed architectural solution for providing continuous measurement feedback on quality characteristics of Big Data (the 10V's), including visualization and interpretation of the measurement results, in a manner that is flexible and easy to use. Big Data users can use the V's measurement results to assess the quality of their underlying data and determine its suitability for their purposes. The problems we are attempting to solve here include the ability to: i) process both structured and unstructured data, ii) track a variety of base measures depending on the quality indicators defined for the V's, iii) flag datasets that pass a certain quality

threshold, and iv) define a general infrastructure for collecting, analyzing and reporting the V's measurement results for trustworthy and meaningful decision-making.

1.5 Overview of the Contributions

The specific contributions of this research are listed as follows:

Measurement of Validity. The derived measures include more specifically Accuracy, Credibility and Compliance and used to build the vector for Validity. It’s important to note that the measure of Accuracy was defined in (Ormandjieva, Omidbakhsh & Trudel 2021) however it has been redefined in for Validity and this new definition will be used for Veracity as well, as the old one would yield negative values after testing. The measure for Credibility remains the same as the one introduced in (Ormandjieva, Omidbakhsh & Trudel 2021). Compliance is a new derived measure, first introduced in Validity and for this thesis.

Measurement of Vincularity. The derived measure, Traceability is first introduced here and is originally developed for the use of Vincularity.
The MEGA Framework. Originally, built upon the ideas from (Social Media Data) the MEGA framework is its own framework developed to be compatible with the measurement V’s of Big Data Quality characteristics.

1.6 Outline of the Thesis

This thesis is broken into seven chapters. Chapter 2 focuses on some of the required knowledge needed to fully understand the methods and processes used to develop the measurements found in this paper. In chapter 3, we discuss the approach to the MEGA framework and some background knowledge for the framework. Chapter 4 discusses the development of the Validity indicator. Chapter 5 discusses the development of the Vincularity indicator. Chapter 6 discusses the MEGA framework and provides a Case Study to show it works on real data. Finally, Chapter 7 provides a conclusion and discusses the future work of what we're studying.

Chapter 2 Background

Here we discuss the knowledge and information needed to fully grasp later concepts when we discuss the development of Big Data Indicators. Here we discuss the role measurements have in Software Engineering and how development needs are related to our goals.

2.1 Basics of Software Measurement

2.1.1 5W's of Measurement

The five W's of software engineering are questions that we ask to help us gain a better understanding of why we take measurements and to determine what the goals are for those measurements.

There are many reasons to take measurements, but these measurements may differ depending on each stakeholder. From a Software Engineering perspective, managers may need to understand their business performance, while an engineer may want to understand their productivity and the quality of their product.

The five W's include:

- Why should we measure?
- Who are we measuring for?
- What are we measuring?
- When should we measure?
- What measures should we use?

These questions can help us more concretely understand the reasons for why we take a measurement and it's the approach we use to build measurements for Big Data Quality.

2.1.2 Goal-Driven Approach to Measurement

To build a foundation for the measurement of quality of big data at its different levels of granularity (elements, records, datasets, and multiple datasets), we adopted the Goal Question (Indicator) Model (GQ(I)M) top-down approach to align the measurement process with the business goals of big data. In the GQ(I)M approach, we define the measurement goal, and we generate questions to identify quality characteristics on that basis, then we define the indicators and measurement procedures required for answering those questions. Indicators can be derived from multiple base or derived measures.

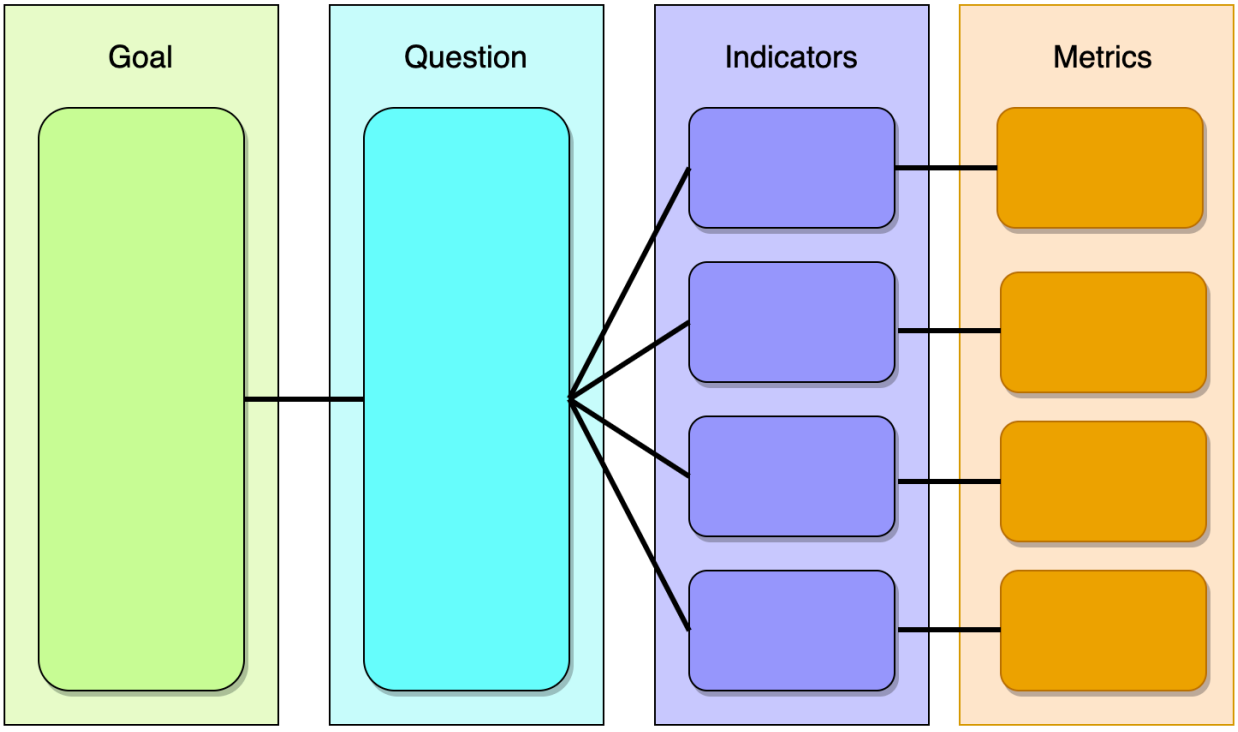

Figure 2.1 GQ(I)M

2.1.3 Representational Approach to Theoretical Validation of Measurement

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". 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 practised in software engineering: (a) theoretical validation, and (b) empirical validation. These two types of validation are respectively used to demonstrate that a measure is really measuring the attribute it is purporting to measure.

The Validity measures are theoretically validated using the Representational Theory of measurement (Fenton, Bieman 2014) with respect to Tracking and Consistency criteria introduced in (IEEE Std 1061, 1998), as described below:

The Tracking Criterion. This criterion assesses whether a measurement is capable of tracking changes in product or process quality over the life cycle of that product or process. 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 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. 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 MDS 1 … n, have the relationship $F_1 \succ F_2 \succ \ldots \succ F_n$, then the corresponding indicator values shall have the relationship $M_1 > M_2 > \ldots > M_n$.

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

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

Ultimately, both theoretical and empirical validation are necessary and complementary.

2.2 Related ISO Standards

2.2.1 ISO/IEC/IEEE 15939

The ISO/IEC/IEEE Std. 15939 establishes many of the common processes and frameworks for the measurement of systems and software. We use this as a basis of building our mathematical models. The document defines many important engineering terminologies that will be used throughout this thesis. (ISO/IEC/IEEE 15939 2017.) These include the following:

Base Measures. A base measure is defined in ISO/IEC/IEEE Std. 15939 as functionally independent of other measures (ISO/IEC/IEEE 15939 2017).

Derived Measures and Indicators. Derived measure is defined as a measurement function of two or more values of base and derived measures (ISO/IEC/IEEE 15939 2017).

Entity. Object that is to be characterized by measuring its attributes. (ISO/IEC/IEEE 15939 2017.)
Attribute: Property or characteristic of an entity that can be distinguished quantitatively or qualitatively by humans or automated means. (ISO/IEC/IEEE 15939 2017.)

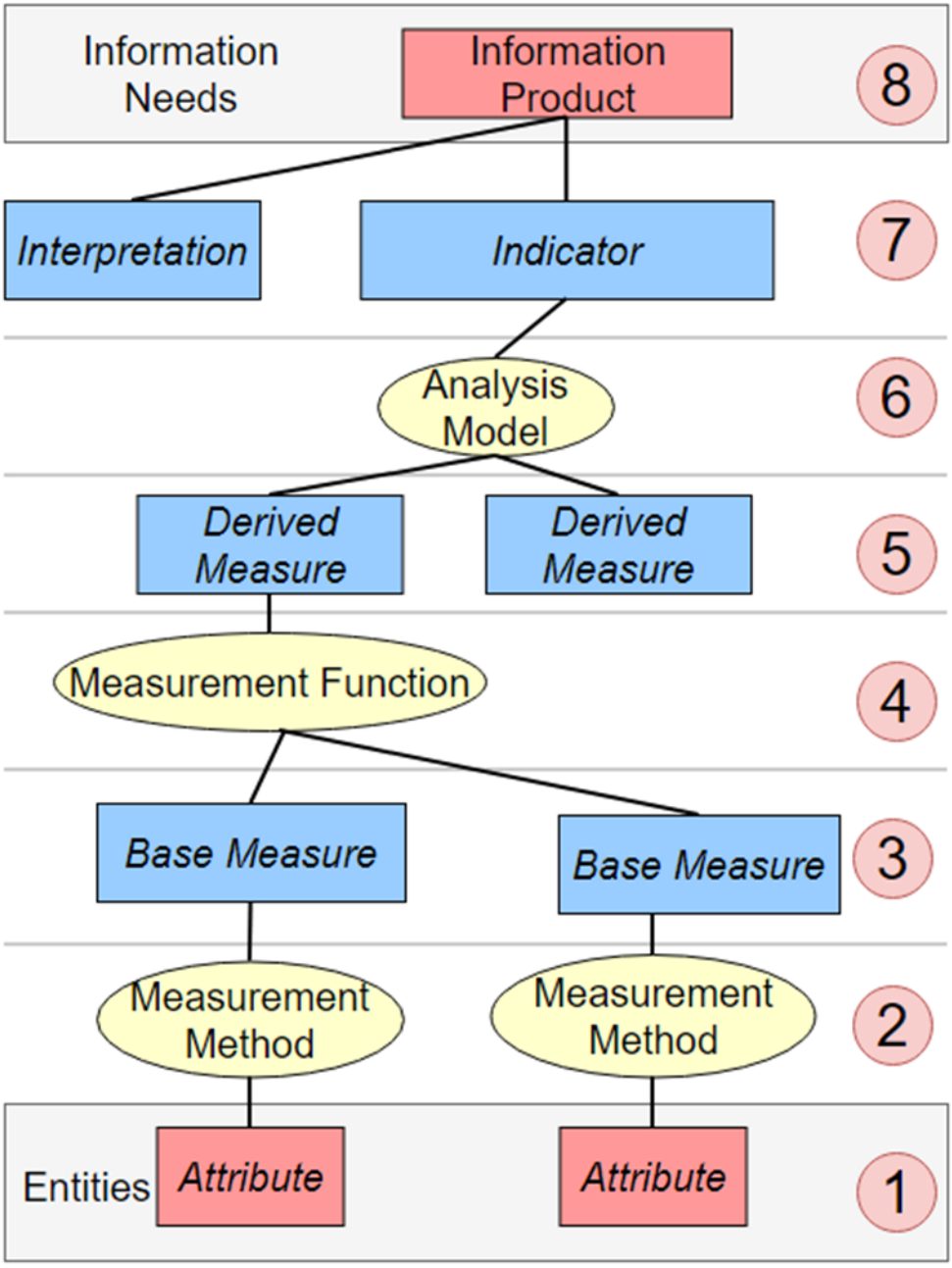

Figure 2.2 Relations in the Measurement Information Model (ISO 15939)

As seen in Figure 2.2. Attributes are properties that are relevant to the information needs. These attributes are qualified against a scale (Measurement Method) and measured (Base Measure). The Measurement Function is an algorithm that combines two or more base measures to produce what is called a derived measure. The Analysis Model is another algorithm that combines measures and decision criteria to produce an Indicator, which itself is an estimate or evaluation that provides the basis for decision-making. (ISO/IEC/IEEE 15939 2019.)

2.2.2 ISO/IEC/IEEE 25012

Quantitative assessment of Big Data V's requires an establishment of data quality characteristics that must be considered when specifying Big Data quality requirements and evaluating data quality. Comprehensive data quality characteristics are proposed in the ISO/IEC international standard ISO/IEC 25012 (ISO/IEC 25012:2008). The data quality model defined in the standard ISO/IEC 25012 is composed of 15 characteristics that reflect two points of view: i) inherent data quality (refers to the degree to which data quality characteristics satisfy data requirements), and ii) system dependent data quality (degree to which data quality is reached and preserved when data is used under specified conditions) (ISO/IEC 25012:2008).

The characteristics proposed in ISO 25012 include the following:

- **Inherent Data Quality**:
    - **Accuracy**: Degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use. (ISO/IEC 25012, 2008.)
    - **Completeness**: 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. (ISO/IEC 25012, 2008.)
    - **Consistency**: Degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities. (ISO/IEC 25012, 2008.)

    - **Credibility**: Degree to which data has attributes that are regarded as true and believable by users in a specific context of use. Credibility includes the concept of authenticity (the truthfulness of origins, attributions, commitments). (ISO/IEC 25012, 2008.)
    - **Currentness**: Degree to which data has attributes that are of the right age in a specific context of use. (ISO/IEC 25012, 2008.)
- **Inherent and System-Dependent Data Quality:**
    - **Accessibility**: degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability. (ISO/IEC 25012, 2008.)
    - **Compliance**: 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. (ISO/IEC 25012, 2008.)
    - **Confidentiality**: Degree to which data has attributes that ensure that it is only accessible and interpretable by authorized users in a specific context of use. Confidentiality is an aspect of information security (together with availability, integrity) as defined in ISO/IEC 13335-1:2004. (ISO/IEC 25012, 2008.)
    - **Efficiency**: Degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use. (ISO/IEC 25012, 2008.)
    - **Precision**: Degree to which data has attributes that are exact or that provide discrimination in a specific context of use. (ISO/IEC 25012, 2008.)
    - **Traceability**: degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use. (ISO/IEC 25012, 2008.)
    - **Understandability**: degree to which data has attributes that enable it to be read and interpreted by users, and are expressed in appropriate languages, symbols and units in a specific context of use. (ISO/IEC 25012, 2008.)
- **System-Dependent Data Quality**

- **Availability**: Degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use. (ISO/IEC 25012, 2008.)
- **Portability**: Degree to which data has attributes that enable it to be installed, replaced or moved from one system to another preserving the existing quality in a specific context of use. (ISO/IEC 25012, 2008.)
- **Recoverability**: Degree to which data has attributes that enable it to maintain and preserve a specified level of operations and quality, even in the event of failure, in a specific context of use. (ISO/IEC 25012, 2008.)

2.2.3 ISO/IEC/IEEE 25024

ISO/IEC 25024 provides measures, including associated measurement methods and quality measure elements for the quality characteristics in the data quality model. The definition of some of the measures from ISO/IEC DIC 25024 is as follows:

Accuracy measures provide the degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use. (ISO/IEC/IEEE 25024, 2015.)

Credibility measures provide the degree to which data has attributes that are regarded as true and believable by users in a specific context of use. Credibility can be measured from the "Inherent" point of view only. (ISO/IEC/IEEE 25024, 2015.)

Compliance measures provide 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. Compliance is measured both from "Inherent" and "System dependent" point of view. (ISO/IEC/IEEE 25024, 2015.)

Traceability measures provide the degree to which data has attributes that provide an audit trail of access to the data and of any changes made to the data in a specific context of use. Traceability is measured both from "Inherent" and "System dependent" point of view. (ISO/IEC/IEEE 25024, 2015.)

However, no specific guidelines or models exist for characterizing the quality of Big Data.
In this research we propose a new hierarchical goal-driven quality model for ten Big Data characteristics (V's) at its different levels of granularity built on the basis of: i) the ISO/IEC standard data terminology and measurements, and ii) NIST (National Institute of Standards and Technology) definitions and taxonomies for Big Data, which is introduced next.

2.3 National Institute of Standards and Technology (NIST)

NIST (National Institute of Standards and Technology) has stimulated collaboration among professionals to secure the effective adoption of Big Data techniques and technology and developed Big Data standards roadmap to this aim. NIST clarified the definitions and taxonomies for Big Data interoperability framework that we will adopt in our study. The taxonomy consists of a hierarchy of roles/actors and activities that visit the characteristics of data at different levels of granularity, namely, element, record which is a group of related elements, datasets which is a group of records and subsequently multiple datasets, as depicted in Figure 2.3.

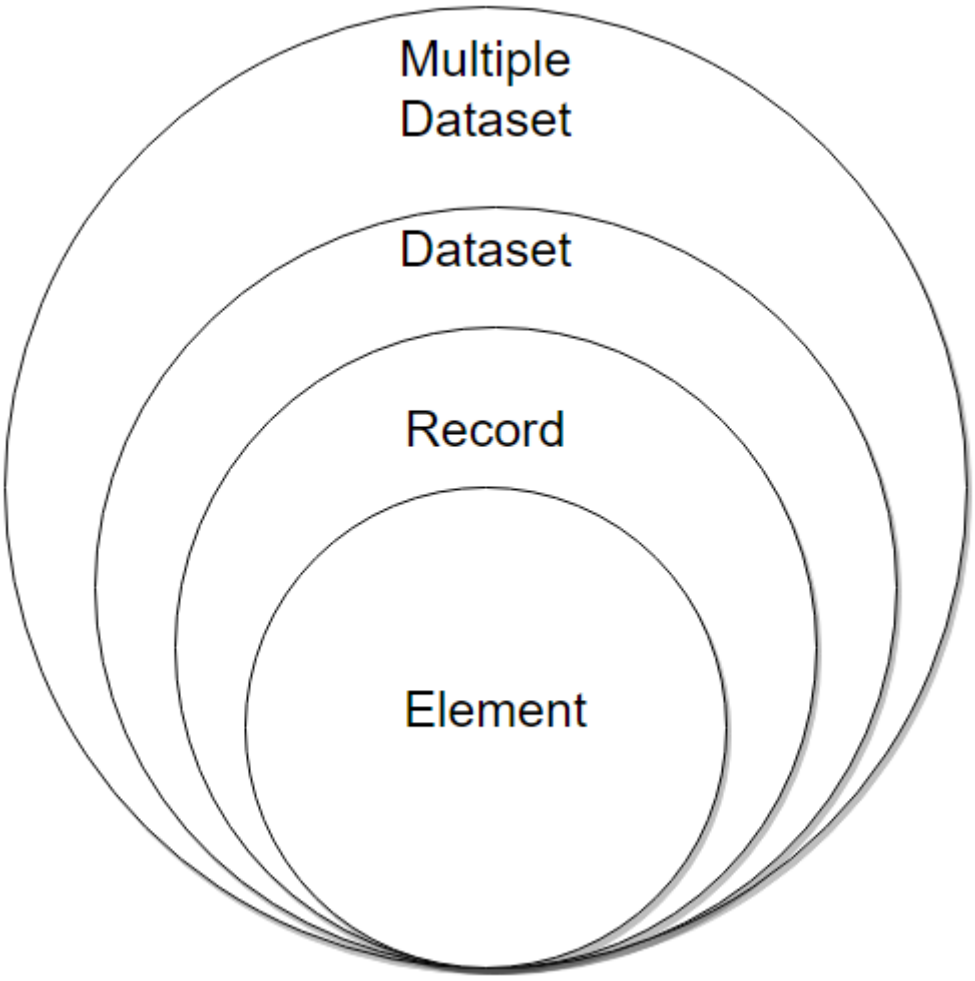

Figure 3.1 Overview of Approach to Big Data Quality Modeling

NIST Taxonomy is used as a foundation for building the proposed new hierarchical measurement model of Big Data's quality at its different levels of granularity (that is, elements, records, datasets and multiple datasets).

Chapter 3 MEGA Approach to Big Data Quality Modeling and Measurement

3.1 Big Data V's

There's a need for developing a standardized quality measurement model in order to achieve measurements that can accurately and objectively model, analyze and interpret the underlying data behind Big Data. (Omidbakhsh & Ormandjieva 2020) outlines such a model by proposing to use a new hierarchical goal-driven quality model for the 10 V's of Big Data, built on the NIST definitions and taxonomies for Big Data as well as the ISO/IEC standard data terminology and measurements.

This approach or one similar is necessary in the development of Big Data because it allows us to more accurately predict the usefulness of the data that's being interpreted. By having a set of indicators, in this case the V's of Big Data, we can, in the future, build models with tools like Deep Learning or Machine Learning in order to effectively understand data.

While it would have been optimal for this thesis to have had mathematical models developed for all 10 V's, there's currently very little research done on progressing the quality characteristics of Big Data. The understanding of the V's started in 2001 with the introduction of the 3 V's defined by Doug Laney in his paper (Doug, D., 2001) which include:

- **Volume**: The vast amount of data generated by the world
- **Velocity**: The speed at which data is being generated and can even include the speed at which data is being processed or handled.
- **Variety**: Refers to the ever-increasing different forms that data can come in. This includes formats such as text, images, sound, videos, 3d models and much more.

Even today, the 3 V's of Big Data are seen as the foundation of Big Data quality and these characteristics are used widely.

But over the years researchers have come to understand more Big Data Quality Characteristics (Gupta, U., Gupta, A., 2016) (Demchenko, Y., et al., 2013) (Soupal, V., 2015) (Staff, B., 2013) (Normandeau, K., 2013) (Mahshewari, R., 2015) and we now have over 10 Big Data Quality Characteristics. These include the 3 V's as well as the following:

- **Veracity**: This quality characteristic refers to the quality of the data, which can vary greatly. This definition, however, isn't that overly general and since the culmination of the characteristics of Big Data refer to the overall quality of Big Data this can make Veracity confusing. Which is why it's expanded in the following paper (Ormandjieva, Omidbakhsh, & Trudel, 2021) where it is more precisely defined.
- **Valence**: Refers to how Big Data can connect with one another.
- **Value**: The insight gained from the processing of Big Data.
- **Volatility**: How long the data is valid for and how much time it should be stored for (Normandeau, K., 2013)
- **Vitality**: Refers to the criticality of the data (Mahshewari, R., 2015)
- **Validity**: The accuracy and correctness of the data for the purpose of usage
- **Vinculairty**: Refers to the connectivity or linkage of data

The aim is to eventually assess (quantitatively) the 10 V's of Big Data and integrate them into the MEGA approach of assessing the Big Data quality. At this moment, however, we've chosen to focus on four quality characteristics (Volume, Velocity, Variety, Veracity) that have been developed in past works (Ormandjieva, Omidbakhsh, & Trudel, 2020) (Ormandjieva, Omidbakhsh & Trudel 2021) and the ones developed in this thesis (Validity and Vincularity). These V's have been fleshed out to have objective measures based on the approach outlined in the paper (Omidbakhsh & Ormandjieva 2020), making them the ideal candidates for a framework that can be applied to an existing Big Data Pipeline and used to continuously take measurements on the underlying data.

3.2 Approach to Modeling the V's.

We can map the notion of a Big Data Quality Characteristic (any of the 10 V's) to the ISO/IEC DIC 25024 data quality characteristics by following what was proposed in the paper (Omidbakhsh & Ormandjieva 2020):

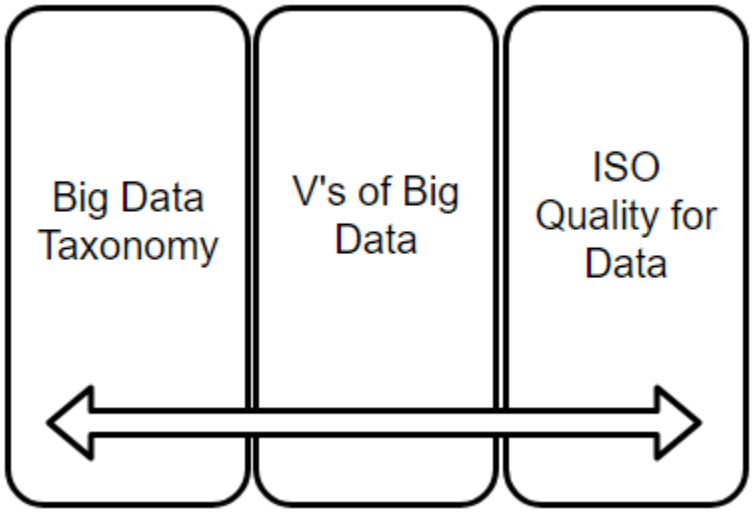

Figure 3.2 Overview of the MEGA Approach

We first need to understand the comprehensive data quality characteristics proposed in ISO/IE international standard ISO/IEC 25012 (ISO/IEC 25012:2008) (see Sections 2.2.2 and 2.2.3).

These characteristics can be mapped to specific Big Data Quality Characteristics and be used to derive them. To do this we use the Goal Question (indicator) Metrics (GQ(I)M, see section 2.1.2) top-down approach discussed in earlier sections of this thesis. In essence, questions are generated and then analyzed to identify the quality characteristics, indicators and measurement procedures needed to answer the questions.

For example, for the Big Data Quality Characteristic, Vincularity, we can generate a question such as: "*What is the Vincularity of Big Data?*" We can then define the Indicator as Vincularity (Mvinc).

The characteristics from ISO/IEC 25012 can then be mapped onto Vincularity based on the GQ(I)M. Vincularity, as previously defined, is the connectivity or linkage of data. Studying the 15 characteristics defined by ISO/IEC 25012, we can map the characteristic of Traceability to Vincularity. In this case, only a single characteristic was mapped to a Vincularity, but as we'll see in future sections, several characteristics can be mapped to the V's. These measures (the characteristics of ISO/IEC 25012) can correspond to any level of the NIST hierarchy of big data, which includes element, records, datasets, and multiple datasets (see section 2.3).

3.3 MEGA Foundation of Big Data V's Measurement: Formal Model of NIST Entities

Theoretically valid measure is founded on the 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" (ISO/IEC/IEEE 15939 2017). 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 a set-theoretical approach to modeling these NIST hierarchy elements, as described next:

Data Elements. 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). 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). Records in big data can originate in heterogeneous sources, including traditional databases, and newer, less structured sources like social media, etc. therefore records 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 two-tuple (DE_r, m) where DE_r is the underlying set of the multiset formed from its distinct elements (DEr Í DE). The multiplicity m: DEr → N + is a function from DEr to

the set of the positive integers, giving the number of occurrences of each element el ∈ 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. 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). An approach like this is justified by the fact that mathematical models would greatly simplify the automation of the measurement procedures.

3.4 MEGA Framework

The aim of the framework is to process and analyze data that follows through each step of the Big Data Pipeline in an attempt to collect important quality metrics and provide insight on the quality status of each step in the pipeline.

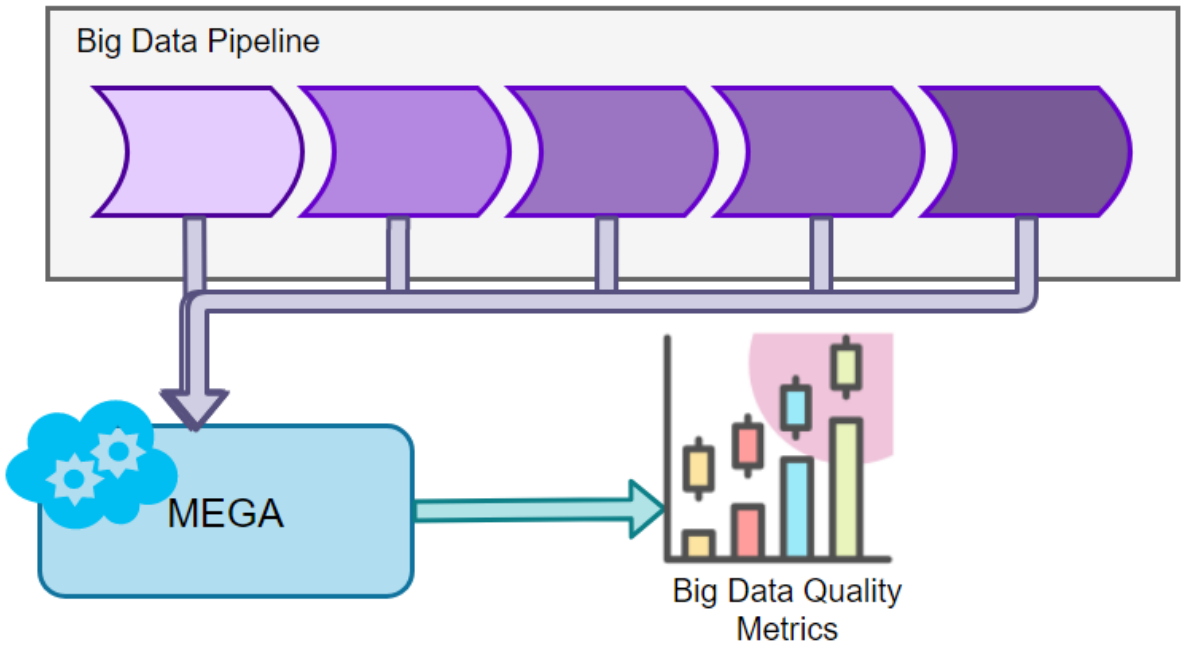

Figure 1 Overview of the MEGA Approach

The processing is done by retrieving data after each step of the pipeline and following a set of rules described by the data practitioners. These include which measurements to take and what thresholds to look for. The measurements themselves, however, are the one's presented in this thesis and is purpose built to objective for the use of data quality collection. Graphs and Illustrations are generated by the MEGA framework as a report to allow stakeholders the ability to gain quick insight into their product and to gain trust on the fact that if their system fails, they can catch it when it does and look back at the reports generated to understand what happened.

3.5 Conclusion

In the following chapters we will be discussing in further details the indicators used and how they have been implemented. This includes the ones that I have developed in collaboration with Dr. Olga Ormandjieva, which are Vincularity and Validity as well the ones developed in past works that include Volume, Velocity, Variety and Veracity. And then a case study will be presented showing how these indicators are used in a simulated user case based on real-world stock data from Yahoo's API.

Chapter 4 Measurement Information Model for Validity of Big Data

4.1 Introduction

In this chapter we propose a new hierarchical measurement of the Big Data quality characteristic referred to as Validity, which was published in (Bhardwaj, Ormandjieva, IDEAS'21, 2021). The proposed measurement model is built upon; i) the NIST (National Institute of Standards and Technology) taxonomy towards to the standardization of big data technology (NIST 2018), ii) measurement principles described in ISO/IEC/IEEE Std. 15939 (ISO/IEC/IEEE 15939, 2017), and iii) the hierarchical measurement models discussed in (Ormandjieva, Omidbakhsh & Trudel 2020) and (Ormandjieva, Omidbakhsh & Trudel 2021). The newly proposed Validity measurements are validated theoretically using the representational theory of measurement (Fenton, Bieman 2014).

4.2 Background and Related Work

4.2.1 Overview of Volume, Velocity, Variety and Veracity Measurements

The MEGA framework automates the 3V's measurement information model proposed to quantify three aspects of Big Data – Volume, Velocity and Variety. Four levels of entities have been considered, derived from the underlying Big Data interoperability framework NIST (National Institute of Standards and Technology) standard hierarchy of roles/actors and activities (NIST 2018). This hierarchy includes data elements, records, datasets and multiple datasets at different levels. The model elements are compliant with ISO/IEC/IEEE Std. 15939 guidelines (ISO/IEC/IEEE 15939, 2017) for their definitions, where four base measures are first defined, assembled into two derived measures, evolving into three indicators - the 3V's. The 3V's measures were validated theoretically based on the representational theory of measurement. For more details, please refer to (Ormandjieva, Omidbakhsh & Trudel 2020). Veracity is one of the characteristics of big data that complements the 3V's of Big Data and refers to availability,

accuracy, credibility, correctness and currentness quality characteristics of data defined in ISO/IEC DIS 25024 (ISO/IEC DIS 25024, 2015). A new measurement information model for Veracity of big data was built upon (Ormandjieva, Omidbakhsh & Trudel 2020) and published in (Ormandjieva, Omidbakhsh & Trudel 2021). The proposed Veracity measurement model is defined as a hierarchy of 6 indicators, 3 derived measures and 13 base measures, as described in (Ormandjieva, Omidbakhsh & Trudel 2021).

4.3 Measurement Information Model for the Validity of Big Data

Validity of Big Data is defined in terms of its accuracy and correctness for the purpose of usage (ISO/IEC DIS 25024, 2015). However, few studies have been done on the evaluation of data validity.

4.3.1 Comparison with Related Work

Big Data validity is measured in (Zhou, Huang & Zhong, 2018) from the perspective of completeness, correctness, and compatibility. It is used to indicate whether data meets the user-defined condition or falls within a user-defined range. The model proposed in (Zhou, Huang & Zhong, 2018) for measuring Validity is based on medium logic. In contrast, in our work we consider a 3-fold root cause of Validity inspired by the notions of ISO/ 25024 data quality characteristic accuracy, credibility and compliance: i) the accuracy of data in MDS, ii) the credibility of DS in MDS, and iii) the compliance of data elements in records, compliance of records in DS, and compliance of DS in MDS.

4.3.2 Mapping of Validity to the ISO/IEC DIC 25024 Data Quality Characteristics

Big Data validity is measured in this thesis from the perspective of accuracy, credibility, and compliance, which are adapted and refined in order to provide an evaluation of the Big Data Validity with respect to defined information needs of its measurement model (see Figure 4.0).

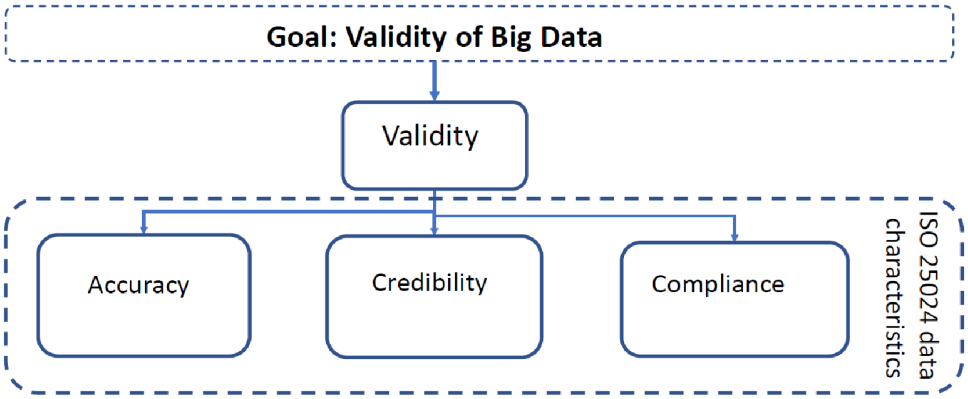

Figure 4.0 Big Data Validity Mapping to ISO25024 Data Quality Characteristics

The measurement information model for the Validity of Big Data defines 3 indicators for measuring accuracy, credibility, and compliance.

4.4 Accuracy Indicator (Acc)

Big data accuracy is essential for the Validity of Big Data. The users of big data sets require the highest validity of their data, but it's a well-known fact that big data is never 100% accurate.

4.4.1 Notion of Accuracy

According to the dictionary definition, accuracy means "the quality or state of being correct or precise". ISO/IEC DIC 25024 defines data accuracy as a degree to which "data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use". It also states that accuracy can be measured from the "inherent" point of view only.
One of the ways to increase the accuracy is to match records and merge them if they relate to the common values of the data attributes. Consequently, we define accuracy as a measure of the common information in DS relationships within MDS.

To measure Accuracy, we propose using the idea presented by Van Emden (Emden, 1971), known as entropy, to determine and calculate the diversity of data that exists in a dataset. This is because, in the case of Big Data, the diversity of data plays a large role in being able to properly analyse data and is one of the avenues in which we can purse accuracy.

4.4.2 Base Measures and Derived Measures

To quantify objectively the common information within the multiple datasets (MDS), we use Emden's information theory model (Emden, 1971): we first abstract the MDS as an Attribute-Record table, where the rows represent all data elements in MDS, and the columns represent the records in MDS. The value of a $cell_{ij}$ of the resulting Attribute-Record table is set to '1', if the data element is included in the record; otherwise, the value of the cell is '0'.

Base Measures *H_acc* and *Hmax*. We use the notion of entropy *H* to objectively quantify the common information (Emden, 1971). The measurement formula for calculating the entropy *H_acc* in the Accuracy measurement model is as follows:

$$H_{acc} = \frac{\log_2(Lbd) - 1}{Lbd * \sum_{j=[1\ldots k]} p_j \log_2(p_j)}$$

where *Lbd* is the count of the total number of records in the MDS, *k* is the number of different columns configurations in the *Attribute-Record* table corresponding to MDS, and p_j is the number of columns with the same configuration so that,

$$Lbd = \sum_{j=[1\ldots k]} p_j$$

The value of *H_acc* varies according to the diversity of the column configurations: common (repeated) configurations in the *Attribute-Record* table (representing duplicated records with the same values of the data attributes) will lower the entropy *H_acc*, while diversity of the records will increase *H_acc* .

H_acc = 0 when all records in MDS contain all values of all data attributes. That is, *k = 1* and $p_1 = Lbd$. The values of *H_acc* for a given DS vary between *0* and *Hmax* calculated for a specific MDS, where *Hmax* represents the maximum entropy for that MDS when all records are different and thus there is no common information within MDS. $Hmax = \log_2(Lbd)$ when all records in MDS are distinct, corresponding to the best-case scenario where there is no need to merge records:

$$Lbd = k\ ,\ p_j = 1,\ and\ \forall j = [1\ldots k]$$

The unit of measurement is the information bit.

Derived Measure *Acc*. In order to measure accuracy independently of the volume of the MDS, we propose to normalize the entropy *H_acc* measure with *Hmax*, Hence, the measurement function for the Accuracy indicator is:

$$Acc\ (MDS) = \frac{H_{acc}}{H\_max}$$

Acc (MDS) normalizes entropy by the best-case scenario *Hmax*, thus normalization will allow data users to objectively compare different DS within the MDS in terms of their common information. *Acc* value is a number between 0 and 1, 0 meaning the worst case (all data elements are common for all records), and 1 corresponding to the best-case scenario when all records in MDS are distinct. When calculating accuracy its important to note that because we meaure Pj and Lbd as our base measures the time complexity of this measure remains linear.

4.4.3 Theoretical Validation of Accuracy

The Accuracy measures are assessed in this section with respect to the Tracking and Consistency criteria introduced in section 2.3. To illustrate the Acc indicator, we use an example of MDS at different time frames T1 and T2. MDST1 shown in Figure 4.1, is a representation of data at T1 that will be used as an example of real-life Big Data. This will also be used to show how we can measure data elements in Big Data.

For the purposes of theoretical validation, we present a modified case of MDS_{T1} where new records were added at time T2 (MDS_{T2}, see Figure 4), T2>T1. In both cases, there are no duplicate records in the multiple datasets MDS_{T1} or $MDS_{T2.}$ All records were mapped to *Attribute-Record* tables similar to the method described in (Emden, 1971) and the entropy was calculated based on the method described in section 4.2.

Dataset 1 (T1)			Dataset 2 (T1)			Dataset 3 (T1)		
Name	**Salary**	**Debt**	**Name**	**Salary**	**Debt**	**Name**	**Salary**	**Debt**
Jill	50,000	10,000	Jessica	55,000	10,000	Chris	50,000	90,000
Eve	40,000	0	Jenella	80,000	400	Beth	40,000	100
Adam	75,000	5,000	Melvin	15,000	3,000	Ela	7000	0

Figure 4.1: **:** Big Dataset Illustration of Accuracy at T1

Intuitively, we expect the value of Accuracy for both multiple datasets MDS_{T1} and MDS_{T2} due to the fact that both correspond to the best-case scenario of maximum accuracy, where there is no need to merge records. From our intuitive understanding, we expect the entropy H_acc_{T2} of the DS depicted in Figure 4.4 to be higher than H_acc_{T1}. We also expect the values of the based measure *Hmax* for MDS_{T2} to be higher than the corresponding values in MDS_{T1} due to the increased size of the DS at time T2.

The values of the variables *Lbd, k* and *p* at time T1 are:

Lbd = 9, k = 9, and $p_i = 1 \ \forall j=[1..k]$. *The entropy value at time T1 for the DS depicted in Figure 4.1 is:* $H\_acc_{T1}=3.1699$

At time T2 the values of the variables are: Lbd = 18, k = 14, $p_i = 2$ *for* $i \in \{1, 6, 12, 13\}$ *and* $p_i = 1$ *for the remaining column configurations. The entropy value at time T2 for the DS depicted in Figure 4.2 is:* $H\_acc_{T2} = 4.1699$. *As expected,* $H\_acc_{T2} > H\_acc_{T1}$. *Similarly* $Hmax_{T2} = log_2(18) > Hmax_{T1} = log_2(9)$

As expected, the value of the Acc measure indicates maximum Accuracy result (ACC(MDS) = 100%) for both MDS_{T1} *and* MDS_{T2}.

Dataset 1 (T2)			Dataset 2 (T2)			Dataset 3 (T2)		
Name	**Salary**	**Debt**	**Name**	**Salary**	**Debt**	**Name**	**Salary**	**Debt**
Jill	50,000	10,000	Jessica	55,000	10,000	Chris	50,000	90,000
Eve	40,000	0	Jenella	80,000	400	Beth	40,000	100
Adam	75,000	5,000	Melvin	15,000	3,000	Ela	7,000	1,000
Jacky	55,000	95,000	Robin	50,000	90,000	Ace	70,000	30,000
Brook	45,000	1500	Luffy	400	100	Rojer	50,000	5100
Cathy	4000	90	Zoro	0	10,000	Odin	70000	0

Figure 4.2 Big Dataset Illustration of Accuracy at T2

To validate Tracking and Consistency criteria of the Accuracy measures on DS with duplicated records, we modify the MDS shown in Figure 4.1 (MDS_{T1}) and Figure 4.2(MDS_{T2}) as depicted in Figure 4.3 below:

	Dataset 1			Dataset 2			Dataset 3		
	Name	Salary	Debt	Name	Salary	Debt	Name	Salary	Debt
T1	Jill	50,000	10,000	Jessica	55,000	10,000	Chris	50,000	90,000
	Jill	50,000	10,000	Jessica	55,000	10,000	Beth	40,000	100
	Adam	75,000	5,000	Melvin	80,000	400	Beth	40,000	100
T2	Jacky	55,000	10,000	Robin	15,000	3,000	Ace	70,000	30,000
	Brook	45,000	1,500	Luffy	50,000	90,000	Ace	70,000	30,000
	Cathy	4,000	90	Zoro	400	10,000	Jenella	6,000	0

Figure 4.3: Illustration of Accuracy Measurement with duplicated records (T1 &T2)

Given that there are duplicated records in Figure 4.3 at both time T1 and time T2 MDS, intuitively we would expect the Accuracy of MDS'_{T2} to be higher than the Accuracy of MDS'_{T1} due to the relatively lower number of duplicated records. Our intuition would also expect not only $H\_acc'_{T2} > H\_acc'_{T1}$, but also $H\_acc_{T1} > H\_acc'_{T1}$ and $H\_acc_{T2} > H\_acc'_{T2}$

The above intuitive expectations are confirmed by the measurement results, where $H\_acc'_{T1} = 2.7254$ and $H\_acc'_{T2} = 3.7254$.

The value of Acc measure is calculated using the formula (4):

Acc (MDS'_{T1}) = $H\_acc'_{T1}$ / $Hmax'_{T1}$, where $Hmax'_{T1} = 3.17$, thus

Acc (MDS'_{T1}) = 86.97%. Similarly, Acc (MDS'_{T2}) = 89.34%

Based on the analysis of the above measurement results we can conclude that both Tracking, and Consistency criteria hold for the Accuracy measures, thus we proved their theoretical validity ((Fenton, Bieman 2014),(IEEE Std 1061, 1998)).

4.4.4 Accuracy Profile for MDS

We propose to visualize the Accuracy indicator results by depicting the Acc values of MDS graphically; this will allow data engineers to easily trace the accuracy of individual DS and identify those MDS whose records need to be analyzed further and merged, where applicable. Figure 4.4 illustrates the Accuracy Profile graph for the four MDS (MDST1, MDST2, MDS'T1 and MDS'T2).

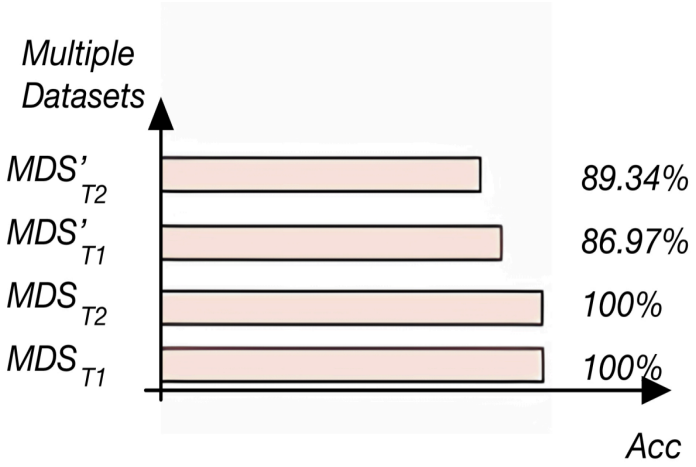

Figure 4.4: Illustration of the Accuracy Profile Graph for MDS

Hence, *Acc* indicator of Validity not only allows objectively to compare different MDS in terms of their accuracy, but likewise visualizes the Accuracy measurement results to facilitate the decision-making of the Big Data users.

4.5 The Credibility Indicator (Cre)

4.5.1 The Notion of Credibility

The notion of credibility in ISO/IEC DIS 25024 standard represents "the degree to which data has attributes that are true and accepted by users in a specific context of use" (ISO/IEC DIS 25024, 2015). In order to measure credibility, we assume the existence of up-to-date information on qualified sources.

4.5.2 Base Measures and Derived Measures

Base Measures. Let $cre_{source}: DS \rightarrow [0\ldots1]$ be a function that returns 1 if the source of a DS is qualified for use, or 0 otherwise. Two base measures are defined in the measurement model of credibility:

- *Number of DS in Big Data set (Nds).* Nds is a simple counting of the total number of DS in MDS.
- *Number of credible DS in Big Data set (Nds_cr)*, where the measurement method is counting of DS with qualified sources:

$$Nds\_cr(MDS) = \sum_{\forall\ DS \in MDS} cre_{source}(DS)$$

Derived measures. We define Credibility measure as a ratio of the total number of credible DS and all DS. The measurement function for *Cre* is specified as follows:

$$Cre\ (MDS) = \frac{Nds_{cr}(MDS)}{Nds\ (MDS)}$$

Regular collection of Cre measurement data would allow practitioners to gain timely valuable control over the credibility of the data sources in their MDS and eventually trace the MDS credibility over time. When calculating creadability its important to note that because we meaure Cre_source and Nds as our base measures the time complexity of this measure remains linear.

4.5.3 Illustration of the Credibility Indicator

We illustrate the Cre indicator through a simple extract of MDS at two-time frames. The first base measure we need to calculate is Nds, which is defined as the number of DS present in MDS: Nds = 3.

Next, we assess the credibility of the DS sources, where *cre_source (DS*) is set to 0 or 1 value, depending on whether or not a specific DS is credible. In this example we assume that DS_1 and DS_3 are credible:

cre_source (DS_1) = cre_source (DS_3) = 1 and,

cre_source (DS_2) = 0.

We also assume that the credibility of the DS at times T1 and T2 remain the same. Next, we normalize the credibility of MDS by *Nds*. The measurement value of *Cre (MDS)* is ⅔ (or 66%), which indicates the proportion of credible DS.

	Dataset 1			Dataset 2			Dataset 3		
	Name	Salary	Debt	Name	Salary	Debt	Name	Salary	Debt
T1	Jill	50,000	10,000	Jessica	55,000	10,000	Chris	50,000	90,000
	Jill	50,000	10,000	Jessica	55,000	10,000	Beth	40,000	100
	Adam	75,000	5,000	Melvin	80,000	400	Beth	40,000	100
T2	Jacky	55,000	10,000	Robin	15,000	3,000	Ace	70,000	30,000
	Brook	45,000	1,500	Luffy	50,000	90,000	Ace	70,000	30,000
	Cathy	4,000	90	Zoro	400	10,000	Jenella	6,000	0

Figure 4.5: Illustration of Credibility with duplicated records (Time T1 and T2)

4.5.4 Theoretical Validation of Credibility

In this section we assess the Tracking and Consistency criteria of the Cre measures. Intuitively, the more credible sources that exist in MDS, the higher the value of credibility in MDS. In order the validate the Cre measurement values against this intuitive expectation, we fix the value of Nds through time and track the changes of cre_source (DS) data.

In the previous example (see section 5.3) the value of *cre_source (DS)* doesn't change from T1 to T2, neither does *cre (MDS*), as expected. If we, however, assume that at time T2 the credibility of

DS'₃ changes (*cre_source (DS'₃) = 1* at time T2), then we expect the credibility of the MDS_{T2} to increase. The measurement value of *cre* (MDS_{T2}) proves that the intuitive expectation is preserved by the measurement value of Credibility indicators, which increased from 0.66 to 1 (meaning 100% credible MDS). These calculations establish the theoretical validity of the Credibility measures, as required by the representational theory of measurement.

4.6 The Compliance Indicator

In this section we introduce a hierarchy of measures for compliance, which evaluate compliance at the record (REC), dataset (DS) and multiple datasets (MDS) levels reflecting the corresponding entities in the NIST hierarchy.

4.6.1 Notion of Compliance

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, according to ISO/IEC 25024 definition. This means that whether or not a data element is deemed as compliant depends on the judgment of the data scientists, the organization, standards and local laws and regulations.

4.6.2 Base Measures and Derived Measures

We define a function $Comp_{Source}$: *Rec* → *[0...1]* that returns 1 if the source record is compliant with the set of standards that have been set by the researchers; otherwise, the returned value is 0.

Base Measure. The base measure rec_comp counts the number of compliant records in a DS, as defined below:

$$rec\_comp(DS) = \sum_{\forall\, rec \in DS} Comp_{Source}(rec)$$

Derived Measures. The proposed derived measure for Compliance is defined as a ratio of the Big Data entities (records, DS, or MDS) that have values and/or format that conform to standards,

conventions or regulations, divided by the total number of data entities. We propose two derived measures for the Compliance indicator measuring the above ratio at the level of a DS and the level of MDS as defined below:

- ***DataSet Compliance*** DS_{comp}. The measurement function for Compliance along a DS is defined by *DS_comp (DS)* as follows,

$$DS\_comp(DS) = \frac{rec\_comp(DS)}{Ldst\ (DS)}$$

where *Ldst (DS)* is the number of records in a specific DS.

- ***Multiple DataSets Compliance*** MDS_Comp ***(MDS).*** Finally, we define a measurement function for quantifying objectively Compliance across all DS in MDS as follows:

$$MDS\_Comp\ (MDS) = \frac{\sum_{\forall\ DS\ \in MDS}\ Nrec_{Comp}(DS)}{Nds\ (MDS)}$$

Regular collection of Compliance measurement data would be necessary to flag data that does not comply with local laws and regulations. For instance, in the Healthcare industry, HIPPA in the USA or PIPEDA in Canada require compliance with their privacy and data security regulations by law, thus the measurement of compliance for such sensitive data will become imperative. When calculating Compliancce its important to note that because we meaure rec_comp, Ldst and Nds as our base measures the time complexity of this measure remains linear.

4.6.3 Illustration of the Compliance Indicator

Figure 4.6 shows the same data as in Figure 4.7, where the data that is non-compliant is highlighted. We assume that for this example, all data elements in Salary or Debt columns must be numerical (commas are allowed). In this example, rec_Comp (DS1) = 2, rec_Comp (DS2) = 1 and rec_Comp (DS3) = 3 at time T1. The value of the Compliance Indicator at the DS level at time T1 is as follows: DS_Comp (DS1) = 0.66, DS_Comp (DS2) = 0.33 and DS_Comp (DS3) = 1. The result of the Compliance at time T1 in terms of MDS is defined to be the average compliance of all DS; MDS_Comp (MDST1) = 0.66. We perform the same steps to measure Compliance at record, DS and MDS levels in Time T2. Finally, MDS_Comp (MDST2) = 0.72. As expected, MDS_Comp (MDST2) > MDS_Comp (MDST1) = 0.66

	Dataset 1			Dataset 2			Dataset 3		
	Name	Salary	Debt	Name	Salary	Debt	Name	Salary	Debt
T1	Jill	50,000	10,000	Jessica	55,000	10000$$	Chris	50,000	90,000
	Jill	50,000	10,000	Jessica	55,000	10000$$	Beth	40,000	100
	Adam	75,000	5000$$	Melvin	80,000	400	Beth	40,000	100
T2	Jacky	55,000	10,000	Robin	15,000	3,000	Ace	70,000	30,000
	Brook	45,000	1,500	Luffy	50,000	90000$$	Ace	70,000	30,000
	Cathy	4,000	90	Zoro	400	10,000	Jenella	6,000	0

Figure 4.6: Illustration of non-compliant data in MDS

4.6.4 Theoretical Validation of Compliance

Based on the meaning of Compliance, the more credible datasets the Big Data contains, the larger the Cre indicator value. The perception of 'more' should be preserved in the mathematics of the measure: the more compliant records that exist in a particular dataset, the higher the rate of compliance as defined earlier. We validate theoretically Compliance by fixing the value of Nds (MDS) to 3 and tracking the change in (MDS) through time: We see from the example above that, as the value of (DS_1) and (DS_2) at the DS level increases from T1 to T2, Compliance at the MDS level increases from 0.66 to 0.72. which represents a 9% increase. This is to be expected: the percent change of records compliance in the DS increases respectively by 8.5% (DS_1) and 9% (DS_2). The above calculations establish the theoretical validity of the Compliance measures by demonstrating both the Tracking and Consistency criteria, as required by the representational theory of measurement.

4.7 Hierarchy of the Validity Measures

The objective of this section is to define the Validity indicator Mval that would allow objectively to compare different Big Data sets in terms of the indicators Accuracy, Credibility and Compliance, and to present graphically this new hierarchical measurement model tailored specifically to the Validity of Big Data.

4.7.1 Validity Indicator Mval

Validity Indicator proposed in this research is defined as a vector ***Mval = (Acc (MDS), Cre (MDS), DS_comp(DS), MDS_comp(MDS))*** that reflects correspondingly the accuracy, credibility and compliance of the underlying data at the level of DS or MDS.

4.7.2 The Measurement Hierarchy

The measurement information model proposed in this work is a hierarchical structure linking the goal of Big Data Validity to the relevant entities and attributes of concern, such as an entropy of a MDS, number of records, number of DS, etc. In our approach, the Validity characteristics were decomposed through three layers as depicted in Figure 4.7. The measurement information model defines how the relevant attributes are quantified and converted to indicators that provide a basis for decision-making.

4.8 Conclusion

In this chapter, we proposed a new theoretically valid measurement information model to evaluate Validity of Big Data in the context of the MEGA framework, applicable to a variety of existing Big Data Pipelines (Bhardwaj, Ormandjieva, 2021). Four levels of entities have been considered in the definitions of the measures, as derived from the NIST hierarchy: data element, record, DS, and MDS. The model elements are compliant with ISO/IEC/IEEE Std. 15939 guidelines for their definitions, where five base measures are first defined, assembled into four derived measures, evolving into four indicators. Theoretical validation of the Validity measures has been demonstrated.

The model is suitable for Big Data in any forms of structured, unstructured, and semi-structured data.

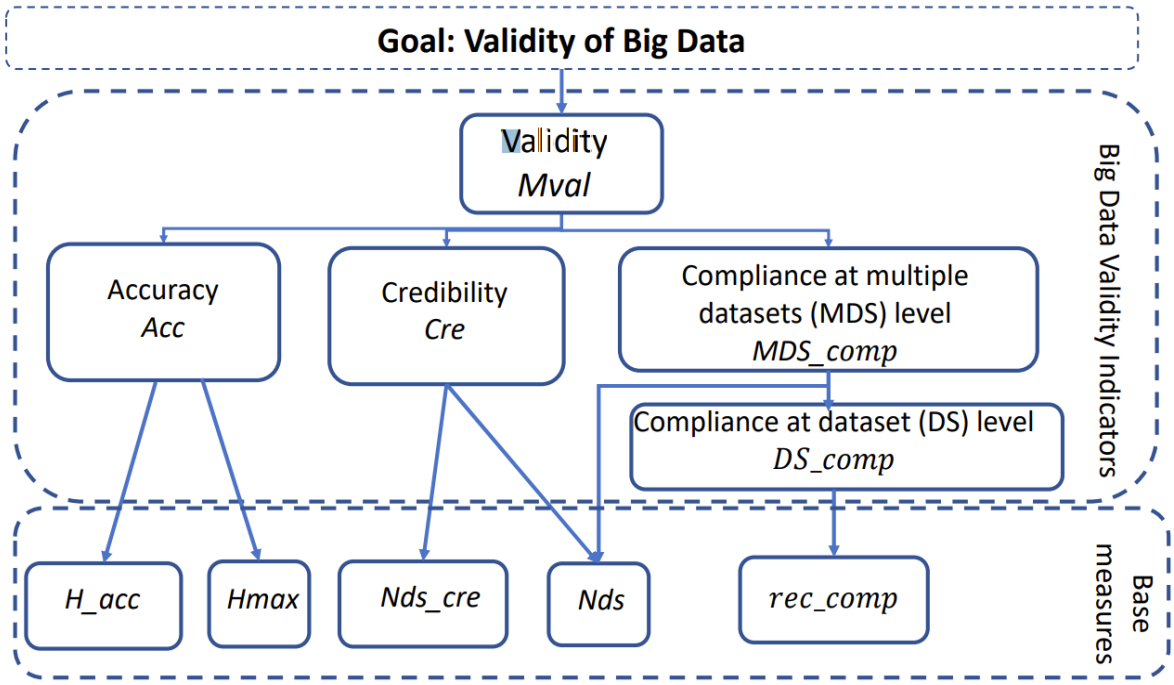

Figure 4.7: Hierarchical Measurement Model of Validity

We illustrated the Validity measurement model by collecting measurement data on small examples and showed how the Validity indicators Accuracy, Credibility and Compliance can be used to monitor data quality issues that may arise. The relevance of such a model for the industry can be illustrated with simple examples of usage of these measures and indicators:

- **Accuracy (Acc)** and its profile: Accuracy is useful to objectively compare MDS in terms of their information content, as well as to oversee variations of Big Data Accuracy over time. A decrease in Accuracy might trigger investigation as actions might be needed to merge duplicated (common) information.
- **Credibility (Cre)** and its trend allows easy and objective comparisons of MDS in terms of their credibility. A Credibility trend showing a decrease might trigger investigation as a source of data could be damaged or unavailable.
- **Compliance (Comp)** at DS and MDS levels, and the corresponding trends: Compliance allows to track an important characteristic of Validity at the level of record, DS and MDS; decrease in the measurement results over time might trigger investigation of the potential legal issues with the usage of the Big Data.

Our future research will enhance the theoretical findings presented in this chapter with empirical evidence through evaluation of these measures with open-access data and industry data.

Chapter 5 Measurement Information Model for Vincularity of Big Data

5.1 Introduction

In this chapter we propose a new hierarchical measurement of the Big Data quality characteristic referred to as Vincularity. The proposed measurement model is built upon; i) the NIST (National Institute of Standards and Technology) taxonomy towards to the standardization of big data technology (NIST 2018), ii) measurement principles described in ISO/IEC/IEEE Std. 15939 (ISO/IEC/IEEE 15939, 2017), and iii) the hierarchical measurement models discussed in (Taleb, Dssouli & Serhani 2015) and (Ramaswamy, Lawson & Gogineni, 2013). The newly proposed Vincularity measurements are validated theoretically using the representational theory of measurement (Fenton, Bieman 2014).

5.2 Measurement Information Model for the Vincularity of Big Data

Vincularity refers to the connectivity and linkage of data (Mahshewari, R., 2015). In essence, As Figure 10 shows below, Vincularity is directly mapped to the Traceability of Big Data. However, few studies have been done on the evaluation of data vincularity.

5.2.1 Mapping of Vincularity to the ISO/IEC DIC 25024 Data Quality Characteristics

Big Data Vincularity is measured in this thesis from the perspective of traceability which has been adapted and refined in order to provide an evaluation of the Big Data Vincularitywith respect to defined information needs of its measurement model. Figure 5.1 shows the relationship of our hierarchical measurement model and the Big Data characteristic of Vincularity.

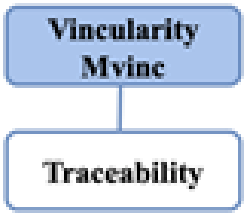

Fig. 5.1. Big Data Validity Mapping to ISO25024 data quality characteristics

The measurement information model for the Vincularity of Big Data defines a single indicator for measuring traceability.

5.3 Traceability Indicator (Trace)

Big data traceability is essential for the Vincularity of Big Data. The users of big data sets may require that their data can be traced back to its original source in order to validate the trustworthiness of the data.

5.3.1 Notion of Traceability

In this thesis we define traceability by the ISO/IEC DIC 25024 definition. Traceability represents the degree to which data has attributes that provide an audit trail of access to the data and any modification in a specific context of use. (ISO/IEC DIS 25024, 2015.)

5.3.2 Base Measures and Derived Measures

In this thesis we introduce our evaluation of Vincularity measured at the level of multiple datasets (MDS) while Traceability is evaluated at the level of the dataset (DS). We measure Traceability at the level of the dataset by determining the amount of data records in a dataset that contain data elements that can be audited. For instance, if a data element exists within a record that cannot be traced, then it is said that the record isn't traceable. For a record to be traceable, this thesis requires

that all data elements contained with a record be traceable. This method of evaluation has been adopted from ISO/IEC 25024 and we can describe it mathematically as follows:

Let $Trace_{Source}$: $rec \rightarrow [0...1]$ be a function that returns 1 if the source data record has metadata (or another form of audit) to track changes that have impacted *all* data elements contained within the record. We define rec_trace(DS) to be the number of traceable records in the dataset. More specifically it represents the number of record elements that have implemented some sort of traceability standard.

$$Rec_{Trace}(DS) = \sum_{\forall\, rec \in DS} trace_{Source}(rec)$$

The measurement function for Traceability along an entire dataset is defined by Trace (DS) and is as follows:

$$Trace(DS) = \frac{rec_{trace}(DS)}{Ldst(DS)}$$

Ldst(DS) represents the total number of records in the particular dataset, for a more detailed explanation on Ldst refer to (Ormandjieva, Omidbakhsh & Trudel 2021). Finally, we can average all the Traceability values across DS in MDS in order to derive Vincularity (Mvin), as follows:

$$Mvin(MDS) = \frac{\sum_{\forall DS \in MDS} Trace(DS)}{Nds(MDS)}$$

Regular collection of Vincularity would allow practitioners to gain timely and valuable control over the traceability of the data sources in their MDS and gain greater confidence over their dataset's quality. When calculating Traceability and Vincularity its important to note that because we meaure trace_source, Ldst and Nds as our base measures the time complexity of this measure remains linear.

5.3.3 Illustration of the Traceability Metric and Vincularity

We can illustrate the measure of Vincularity through a simple extract of multiple datasets at two different time frames T1 and T2, where T2 > T1.

	Dataset 1			Dataset 2			Dataset 3		
	Name	Salary	Debt	Name	Salary	Debt	Name	Salary	Debt
T1	Jill	50,000	10,000	Jessica	55,000	10,000	Chris	50,000	90,000
	Jill	50,000	10,000	Jessica	55,000	10,000	Beth	40,000	100
	Adam	75,000	5,000	Melvin	80,000	400	Beth	40,000	100
T2	Jacky	55,000	10,000	Robin	15,000	3,000	Ace	70,000	30,000
	Brook	45,000	1,500	Luffy	50,000	90,000	Ace	70,000	30,000
	Cathy	4,000	90	Zoro	400	10,000	Jenella	6,000	0

Fig. 5.2. Illustration of Accuracy measurement with duplicated records (time T1 and time T2). Table Title: MDS_1

We first begin by assigning unique identifiers to the distinct data elements that need to be distinguished. In this case we have NameJill, NameAdam, NameJessica, NameMelvin, NameChris, NameBeth, NameJacky, NameBrook, NameCathy, NameRobin, NameLuffy, NameZoro, NameAce, NameJanella all represent the different data elements associated with the each of the 3 datasets at Time 1 and 2. This is done for all elements and so we end up having an additional list of values representing salary and debt: Salary: S_50000, S_75000, D_55000, S_55000, S_45000, S_4000, S_80000, S_15000, S_400, S_40000, S_70000, S_6000. Debt: D_10000, D_5000, D_1500, D_90, D_400, D_3000, D_90000, D_100, D_30000, D_0.

We can also record a few more base measures that are necessary. Ldst is defined as the number of records in a dataset, therefore, Ldst(DS1_T1) = 3, Ldst(DS2_T1) = 3, Ldst(DS3_T1) = 6 and at T2 we get Ldst(DS1_T2) = 6, Ldst(DS2_T2) = 6, Ldst(DS3_T2) = 6. Finally, we can measure Nds, the number of datasets in MDS. In this case Nds(MDS) = 3, giving us 3 datasets in total.

Fig. 5.3 shows the same data as shown in Fig. 5.2 but here the data highlighted in red is data that isn't traceable. Because traceability is found at the level of the record, we can assume that if a data element is untraceable then the record itself can be labelled as untraceable.

	Dataset 1			Dataset 2			Dataset 3		
	Name	Salary	Debt	Name	Salary	Debt	Name	Salary	Debt
T1	Jill	50,000	10,000	Jessica	55,000	10000$$	Chris	50,000	90,000
	Jill	50,000	10,000	Jessica	55,000	10000$$	Beth	40,000	100
	Adam	75,000	5000$$	Melvin	80,000	400	Beth	40,000	100
T2	Jacky	55,000	10,000	Robin	15,000	3,000	Ace	70,000	30,000
	Brook	45,000	1,500	Luffy	50,000	90000$$	Ace	70,000	30,000
	Cathy	4,000	90	Zoro	400	10,000	Jenella	6,000	0

Fig. 5.3. Illustration of non-traceable data in MDS

From this table we can record rec_trace (DS1) to be 2, rec_trace(DS2) to be 1 and rec_trace(DS3) to be 3 at T1. With this we can calculate Trace (DS1_T1) = 0.66. We simply repeat the same steps for each dataset, and we see that Trace (DS2_T)1 is 0.33 and Trace (DS3_T1) to be 1.00. We repeat the same steps again when T2 can eventually be read. Here, Trace (DS1_T2) =0.66 Trace (DS2_T2) =0.50 and Trace (DS3_T2) to be 1.00. Traceability in terms of MDS is defined to be Vincularity (Mvin). We can calculate Vincularity by averaging the Traceability of datasets across MDS using Nds (MDS). We find that Mvin (T1) = 0.66 and that Mvin (T2) = 0.72. This shows us that as the ratio of untraceable records decreases, Vincularity increases.

We propose to compare the Vincularity of MDS graphically to the Traceability of each of the dataset like in the example shown below in Figure 5.4.

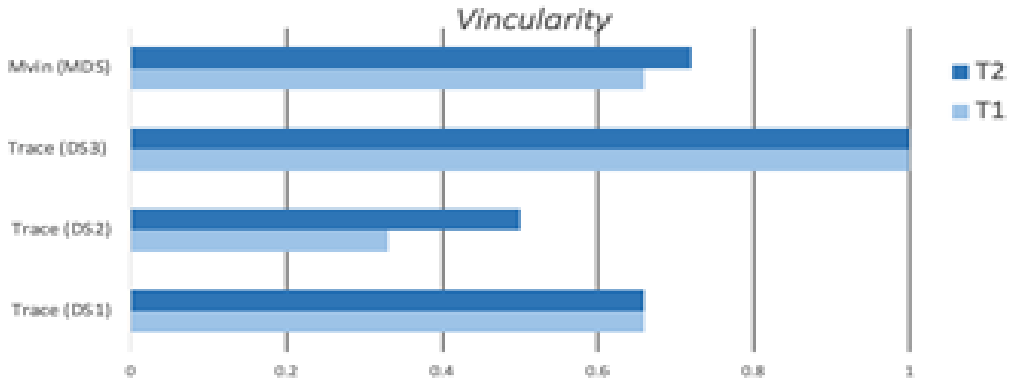

Figure 5.4. Illustration of Vinculairty and Traceability using clustered bar graphs for MDS

Graphical representations like the one shown in Figure 5.4 makes it significantly easier for data engineers to quickly find issues and monitor individual datasets and in this case, we see the

difference between Trace (DS1), Trace (DS2) and Trace (DS3) and can compare it to the Vincularity of MDS. The case here shows us how a significant increase in the Trace (DS1) from T1 to T2 gave us a moderate increase in out Vincularity at T2. We can also see visually how the measure of Traceability and Vincularity can objectively compare multiple datasets in terms of their traceability.

5.3.4 Theoretical Validation of Traceability

Vincularity is mapped directly to Traceability. We can theoretically validate the base measures of traceability to validate the measures of Traceability and Vincularity using the Tracking and Consitency Criterions.

Traceability and Vincularity. The more traceable records that exist in a particular dataset, the higher the rate of traceability and Vincularity as defined earlier.

We first validate the measure rec_trace (DS). Rec_trace (DS) is defined as the number of traceable records in DS. We know that because the base measure trace_source (DS) is a value that is either 0 or 1 and that rec_trace is the summation of this value, we find that rec_trace acts as a count for traceability in DS. We see this in the example for section IV when we recorded that rec_trace (DS1) is 1 and at T2 it goes to 2, doubling in size. This means that the measure follows both the tracking criteria as it's able to count traceability over time (T2>T1 as expected) and the Consistency Criterion (T2 increases by the expected value of 1). Knowing this, we can validate Traceability Trace (DS). Assuming Ldst (DS) stays the same, like in the example from Section IV we find that as the traceability of a dataset grows, so does Trace (DS), for example with 1 traceable item Trace (DS2_T1) is 33% (Ldst (DS2) = 3) and with 3 traceable items Trace (DS3_T1) is 100% (Ldst (DS3) = 3). Trace (DS) stays both consistent and allows for tracking, assuming that Ldst is static. This is the same case for Vincularity except we keep Nds (MDS) to be static. For example, Mvin (MDS_T1) = 0.66 and Mvin (MDS_T2) =0.72. This increase is to be expected because of the increase in Trace (DS2) changing from 33% to 50% between T1 and T2.

5.4 Hierarchy of the Vincularity Measure

The objective of this section is to define the Vincularity indicator (Mvinc) that would allow objectively to compare different Big Data sets in terms of the Traceability indicator, and to present graphically this new hierarchical measurement model tailored specifically to the Vincularity of Big Data.

5.4.1 Vincularity Indicator Mvin

Vincularity Indicator proposed in this research is defined as

$$Mvin(MDS) = \frac{\sum_{\forall DS \in MDS} Trace(DS)}{Nds(MDS)}$$

that reflects correspondingly the Traceability of MDS.

5.4.2 The Measurement Hierarchy

The measurement information model proposed in this work is a hierarchical structure linking the goal of Big Data Vincularity to the relevant entities and attributes of concern, such as the MDS, number of records, number of DS, etc. In our approach, the Vincularity characteristic was decomposed through a few layers as depicted in Figure 5.5. The measurement information model defines how the relevant attributes are quantified and converted to indicators that provide a basis for decision-making.

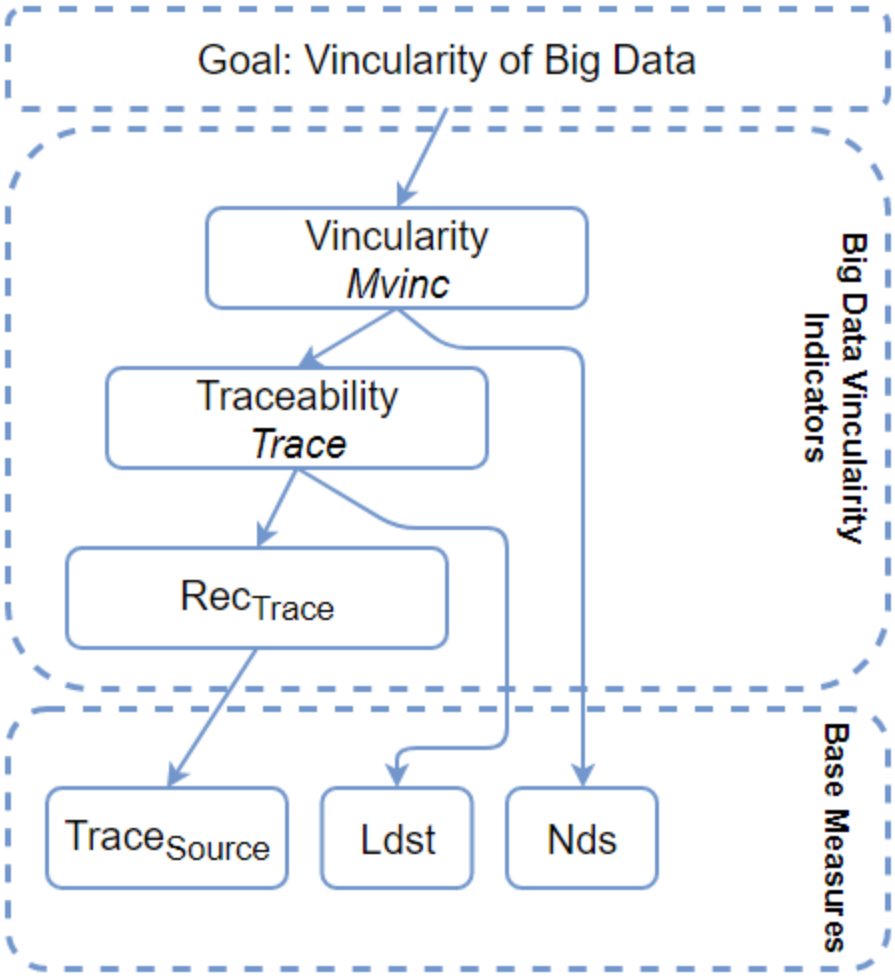

Figure 5.5: Hierarchical Measurement Model of Vincularity

5.5 Conclusion

In this chapter, we proposed a new theoretically valid measurement information model to evaluate Vincularity of Big Data in the context of the MEGA framework, applicable to a variety of existing Big Data Pipelines (Bhardwaj, Ormandjieva, 2021). Four levels of entities have been considered in the definitions of the measures, as derived from the NIST hierarchy: data element, record, DS, and MDS. The model elements are compliant with ISO/IEC/IEEE Std. 15939 guidelines for their definitions, where two base measures are first defined, assembled into two derived measures,

evolving into one indicator. Theoretical validation of the Vincularity measure has been demonstrated.

The model is suitable for Big Data in any forms of structured, unstructured, and semi-structured data.

We illustrated the Vincularity measurement model by collecting measurement data on small examples and showed how the Vincularity indicator (Traceability) can be used to monitor the linkage of data. The relevance of such a model for the industry can be illustrated with simple examples of usage of the traceability measure:

- **Traceability (Trace)** at DS and MDS levels, and the corresponding trends: Traceability allows to track an important characteristic of Vincularity at the level of record, DS; decrease in the measurement results over time might trigger investigation of the potential misuse of the metadata or that of an unwanted source in the Big Data Pipeline.

Our future research will enhance the theoretical findings presented in this chapter with empirical evidence through evaluation of these measures with open-access data and industry data.

Chapter 6 MEGA Architecture

6.1 Introduction

Having objective measures for the V's of Big Data like Volume, Velocity, Veracity, Vincularity, Variety and Veracity is an important first step to understanding the quality of data. The next step is to have a framework that can be applied to Big Data systems and allow practitioners of Big Data to easily apply these measurements to their systems. This is where we propose a novel quality measurement framework called (MEGA), that can be used to apply these measures. The aim of the framework, beyond using the measures defined, is to continuously collect, analyze and report objective data and information on the underlying data. This can allow practitioners to potentially identify and flag issues that may arise throughout the Big Data process and allow practitioners to make any required changes before the data propagates further into the system.

In this chapter, we introduce the motivations of this work, provide an overview of the framework, and present a case study to show its usefulness in a simulated scenario.

6.2 Background and Related Work

6.2.1 Comparison with Related Work

The authors in (Pääkkönen & Ovaska 2015) propose a framework for managing Big Data quality by adding an additional layer to the Big Data Architecture proposed in (Pääkkönen & Pakkala 2015). This framework establishes points at which quality policies can be used to evaluate Big Data Quality. It uses two primary policies, the organizational policy and decision-making policy. The organizational policy defines the selected data sources, quality attributes and metrics that can be evaluated during phases like data extraction and data pre-processing. The decision-making policy handles the rules associated with the decision-making phase of the pipeline. Our approach differs from the architecture described in (Pääkkönen & Ovaska 2015) in that we accommodate a

variety of existing Big Data Pipelines, while the focus of the authors in (Pääkkönen & Ovaska 2015) is limited to a much narrower field of social media data. In our MEGA architecture, the quality policies are flexible enough to target a variety of Big Data applications, including IoT, health, agriculture, etc.

A quality evaluation framework for a big data pre-processing service is proposed in (Taleb, Dssouli & Serhani 2015). This framework lacks the ability to define quality measurements in every stage of the Big Data Pipeline. In contrast, MEGA architecture allows data engineers and users to select V's according to their needs and provides them with much more flexibility in how data can move through the pipeline, while data is being evaluated by the Big Data Quality layer.

A conceptual framework for quality assessment and management of Biodiversity Information Standards (TDWG) is proposed in (Veiga, Saraiva, Chapman, et al. 2017). It serves as a common ground for a collaborative development of solutions in biodiversity informatics. The authors suggest a basic architecture for a computational platform based on this conceptual framework, which consists of three main parts: (1) registering and retrieving biodiversity data quality status (2) registering and retrieving methods and tools for meeting biodiversity informatics requirements, and (3) registering and retrieving biodiversity data quality needs. The conceptual framework described in (Veiga, Saraiva, Chapman, et al. 2017) differs significantly from our contribution in that it tackles the specific needs of the TDWG community. In addition, it is to be noted that the authors' measurement terminology does not comply with the established by ISO/IEC/IEEE Std. 15939 software measurement terminology and guidelines (ISO/IEC/IEEE 15939, 2017).

The 3As Quality-in-Use model proposed in (Merino, Caballero, Rivas, Serrano & Piattini 2017) considerably differs from our approach in that it tackles Quality-in-Use of Big Data solutions, but not the quality characteristics of Big Data, which are the targeted in our work. Moreover, the implementation of the Quality-in-Use characteristics proposed in (Merino, Caballero, Rivas, Serrano & Piattini 2017) targets data at the record level only, while MEGA framework assesses a hierarchy of quality characteristics not only at data elements and record level, but also at a dataset and multiple datasets levels, which is more suitable for the quality characteristics (the V's) of Big Data.

Characteristics	MEGA Approach	Evaluating the Quality of Social Media Data in Big Data Architecture	Big Data Pre-Processing: A Quality Franework	Conceptual Framework for Quality assessment and mangement of biodiversity data	Data quality in use model for big data
Defined Measurements for each step in the Data Pipeline	✓	✓	✓	NA	NA
Heuristics For automatically discovering rules	NA	NA	✓	NA	NA
Big Data Quality Vizulations	✓	✓	✓	✓	NA
Validation of Measurements based on the Representational Theory of Measurement	✓	NA	NA	NA	NA
Measurement of Big Data 10V's	✓	NA	NA	NA	NA
User-configurable policies	✓	✓	✓	✓	NA
Metadata Repository	✓	✓	NA	✓	NA
User-configurable policies for each step in the data pipeline	✓	NA	NA	NA	NA
User Configurable Measurements and visualization techniques	✓	NA	NA	✓	NA

Figure 6.1.: Comparison of MEGA against other related works

The proposed here measurement framework of Big Data MEGA is built upon the standard measurement principles described in ISO/IEC/IEEE Std. 15939 and the hierarchical measurement models discussed in (Ormandjieva, Omidbakhsh, & Trudel, 2020) and (Ormandjieva, Omidbakhsh, & Trudel, 2021). It automates the 3V's measurement information model aimed at evaluating three aspects of Big Data – Volume, Velocity and Variety (Ormandjieva, Omidbakhsh, & Trudel, 2020), and the Veracity measurement information model (Ormandjieva, Omidbakhsh, & Trudel, 2021). The hierarchical measurement models are briefly reviewed next.

6.2.2 Comparison Against Similar Patents

United States Patent No.: US10572456B2: (Staeben, Maier, et.al, 2020)

Targets big data pipelines and extracts metrics directly from the data in the pipeline. The major difference seems to be in the level of measurement: their understanding of a metric is equivalent to "base measure" in the standard terminology followed in our work (such as the number of records processed successfully). There is no validation of the metrics in their work.

United States Patent No.: US20190286617A1: (Hisham, Xiuzhan, et al., 2019)

The goal of this invention is to link data elements within datasets to establish relationships between the data records in order to extract linkage data structures and consequently simplify the datasets or data structures.

United States Patent No.: US 10,838,921 B2 (Mattelli et al., 2020)

The Goal of this invention is to assess the complexity of data cleansing and governance, and criticality of data attributes to the to one or more enterprise dimensions in order to prioritize efforts and meet schedule in Information integration projects. US10572456B2 patent also targets big data pipelines and extracts metrics directly from the data in the pipeline. One major difference seems to be in the level of measurement: their notion of metrics is equivalent to "base measure" in the standard terminology followed in our work (i.e., the number of records processed successfully). There is no validation of the metrics in their work.

In contrast, MEGA approach targets the quality characteristics specific to the "Big" aspect of data, the V's. the goal of our proposal is to extract measurement data on big Data's quality characteristics Volume, Variety, Validity, etc. (V's), analyze the data and generate indicators of V's at every step on the big data pipelines. The measurement process will be used to flag low quality data that is flowing through the pipelines, before it affects the decision-making. The proposed V's are modeled as a hierarchy of measurements (base measures, derived measures, indicators – see the standard terminology in ISO 15939). The measures are validated theoretically, which eliminates systematic errors of the measurement results. Measurement methods for

collecting measurement data are objective thus do not depend on a particular set of data. MEGA Quality Indicators are constructed and reported automatically from the collected objective measurement data thus random error is eliminated, which increases the reliability and trust in the MEGA measurement results.

United States Patent No.: US 9,984,235 B2 (Madera et al., 2018)

The authors claim that there are only 4 V's of Big Data. They propose to track changes of Veracity only. They claim that Veracity and Trustworthiness are the same notions. They quantify Veracity Key Performance Indicator as a trustworthy index score only. In contrast, Veracity of big data commonly refers to the degree of data accuracy, trustfulness and precision (ref: Lukoianova, T., & Rubin, V. (2014). Veracity Roadmap: Is Big Data Objective, Truthful and Credible? Advances In Classification Research Online, 24(1). doi:10.7152/acro. v24i1.14671.). Therefore, Veracity is a multidimensional model. For instance, in MEGA approach we proposed a new 5-dimensional veracity model built upon NIST and ISO25024 standards. We do not measure trustworthiness of data as it is outside ISO25024's scope.
Their architecture is different from Big Data Pipelines thus is unrelated to our approach.

United States Patent No.: US 10,191,962 B2 (Shkapenyuk et al., 2019)

The main goal of this Patent is to identify outliers of data file content. In principle, this patent aims at solving a similar problem, that is, detecting data quality anomalies. The approach here is pure statistical analysis: the authors propose to capture random data errors (outliers) by generating models from historical similar data and analyzing the new data vs so named base models. Their approach will not detect systematic errors due to lack of goal-driven approach to measurement.
In our approach, we also track measurements over time and detect anomalies. In contrast to their work, we first derive measurement models for Big Data specific quality characteristics – the V's, defining procedures for collecting the corresponding base measures and generating indicators of V's. This assures that the generated indicators and their interpretation is tailor-made for the Big Data quality needs. Therefore, it will flag both random and systematic errors in the data flowing through the Big Data pipelines.

Their architecture is different from the Big Data Pipelines. Our Framework also differs in organization and uses more components for the User to specify their conditions.
In addition, they do not use the standard Big Data terminology defined in NIST or standard measurement methodology defined in ISO15939

6.3 System Architecture

The objective of the proposed novel quality measurement framework for Big Data (MEGA) is to provide a flexible and scalable architecture that can be used on most Big Data Pipelines for the purpose of assessing Big Data quality in terms of the widely used 10V's. The V's measurement methods are designed to be collected and analyzed at any step of the process or throughout the entire pipeline, based on the users' individual context of use. The MEGA solution proposed here is built around the Big Data Pipeline architecture described in (Pääkkönen & Pakkala 2015) and shown in Fig. 17, as well as the Big Data Quality Architecture described in (Pääkkönen & Ovaska 2015).

The goal is to allow for the MEGA architecture to run in parallel to the Big Data Pipeline and to be used to halt the pipeline when data issues are flagged. This will permit the data engineers to

assess the quality of the data before certain steps in the process so as to avoid costly mistakes along the Pipeline.

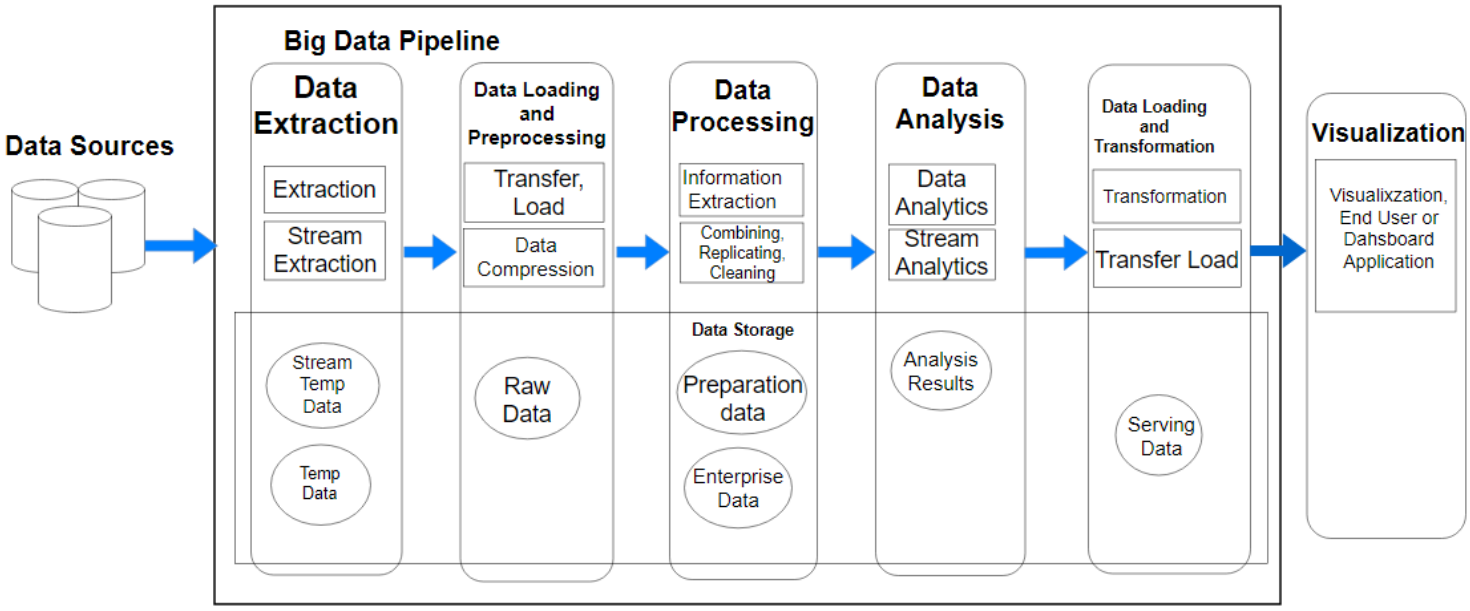

Fig 6.2. Big Data Architecture Diagram from

In our approach, the Big Data Quality Architecture has been adopted and improved to allow for the use of additional quality characteristics and for a large variety of Big Data Systems. It's also been improved to work with more flexibility at each step of the Big Data Pipeline. Fig. 6.3 depicts the overview of the proposed MEGA architecture; it also illustrates the flow of data.

Comparing it to Fig. 6.2, data no longer travels from just left to right through each phase of the Pipeline. The Big Data Quality Framework now requests data from each phase based on a scheduling protocol, so that Big Data quality measurements of the V's can be collected and analyzed at each step. This scheduling protocol is defined by the data engineer in the Quality Policy Manager (see the example shown in Fig. 6.4).

The MEGA architectural solution allows for data to travel from the left to the right of the Big Data Pipeline unimpeded according to the scheduling protocol. This permits the Big Data Quality layer to measure attributes in parallel, while the pipeline continues processing data.

The Big Data Quality layer permits the user to halt the pipeline process until quality validation is completed, by specifying this in the Quality Policy Manager. The goal is to allow the user the most amount of freedom in terms of adapting the Big Data quality assessment to their specific needs at each phase of the Pipeline.

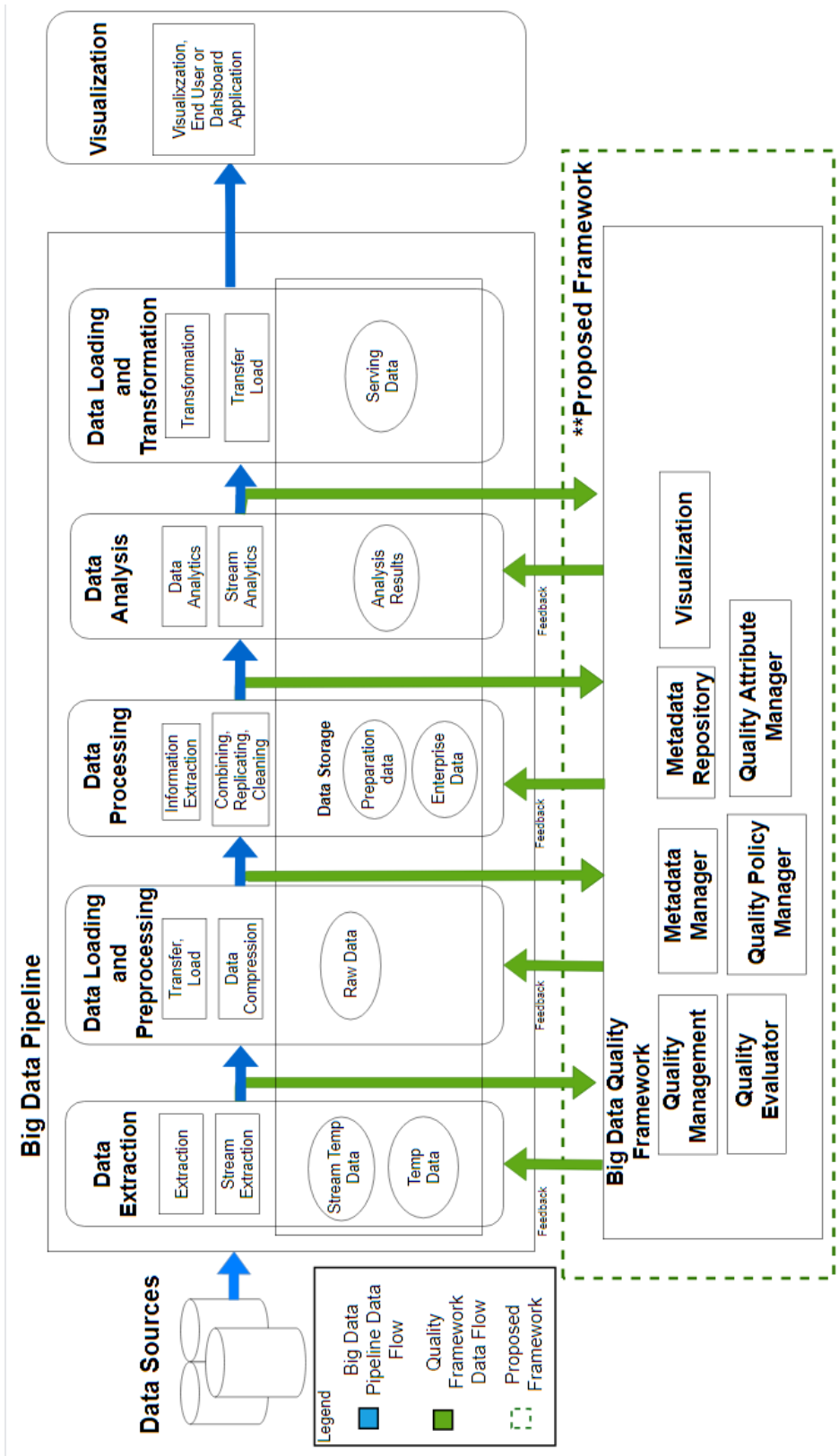

Figure 6.3: Big Data Quality Architecture Diagram

6.4 Architecture Components

6.4.1 Quality Policy Manager

The Quality Policy Manager is, in essence, a configuration file for users to control the behaviour of the system. This includes which quality attributes should be measured and which data stores should be validated. Beyond that the Quality policy Manager also defines the scheduling process of the Big Data Quality Architecture on a step-by-step case. This means that, for example, the data extraction phase can be run for quality metrics on a daily basis while the data processing phase can be verified hourly. Additionally, verification of data quality can be defined by setting thresholds and defining what occurs if a threshold is to be passed.

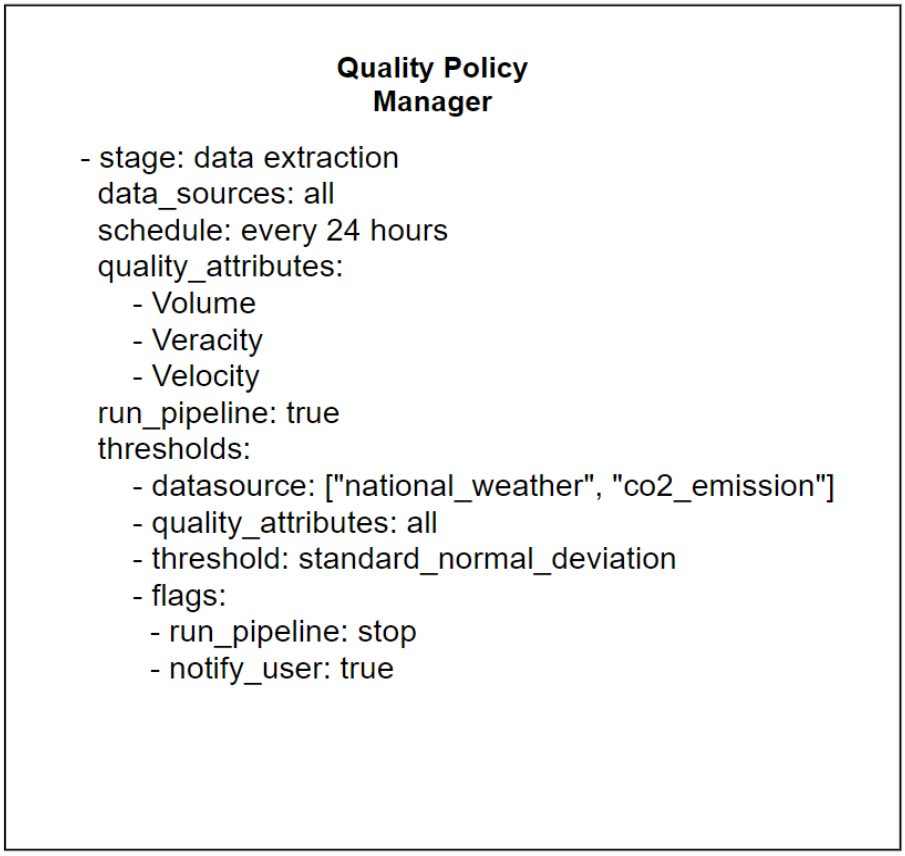

Figure 6.4: Example Quality Policy Manager Diagram

Figure 6.4 describes an example of how a Quality Policy Manager could be set up. In this example, we set the stage (or phase) that we want to validate, here it's the Data Extraction phase. We define the data sources that we want to check, in this case, because we want to check all data sources we can just define 'all' as the value. We also set up our schedule to validate data every 24 hours, this tells our Quality Manager when to ping the pipeline and ask for data. Depending on the context, developers and users may want to adjust this to validate data quality every second or every week. Perhaps the most important section of this file will be defining the quality attributes to measure. These measures can be pre-defined in the Quality Attribute Manager and customized there. Other things that can be done include allowing the pipeline to continue running or stopping until the data has been validated. Additionally, we can add thresholds to validate the data and notify the users of potential anomalies that have been identified. It is important to note however that these thresholds are to be determined by the data engineers themselves since these can be very context specific.

6.4.2 Quality Manager

The Quality Manager handles the framework as defined by the Quality Policy manager. This includes retrieving data from the Pipeline, making requests to the Metadata Manager, or providing data for the Attribute Evaluator. It also handles feedback to the Big Data System. In cases where the Big Data Pipeline needs to be paused or the user needs to be notified the Quality Manager will be responsible for handling these requests.

6.4.3 Metadata Manager and Metadata Repository

The Metadata Manager as described by (Pääkkönen & Ovaska 2015.) enables the extraction of metadata and access to metadata. It acts as the doorkeeper to the Metadata Repository, validating that the structure of the metadata is proper before saving or accessing information from the Metadata Repository.

The Metadata Repository is a data store used to store the Big Data Quality measurements taken by the Quality Attribute Evaluator. Additionally, other metadata information can be stored depending

on how the user defines the Metadata Manager. Information like timestamps and visualization data can be saved to the metadata store for easy access later.

6.4.4 Quality Attribute Evaluator

The Attribute Evaluator measures base measures and calculates the necessary derived measures expressed by the Quality Policy Manager. To evaluate these base measures and derived measures the evaluator looks at how it has been defined in the Quality Attribute Manager and evaluates them accordingly.

6.4.5 Quality Attribute Manager and Visualization Dashboard

Like the Quality Policy Manager, the Quality Attribute Manager allows the user to define base measures and the derived measures based on those defined base measures. This gives the user the greatest amount of flexibility in what attributes and metrics are needed for their specific use case. Furthermore, the best visualization techniques can be defined from the Quality Attribute Manager. This will be useful later for the end-user to have an overview of what the metrics calculated.

6.5 Multiphase Measurements

The proposed architecture adds much more granularity to how Big Data Quality Measures can be taken in the context of an actual Big Data Pipeline. Specific data stores can be validated while others can be left untouched. Additionally, anomalies can be detected automatically by applying thresholds with statistical methods such as a standard normal deviation where the Quality Manager can halt the pipeline and notify the user should it be necessary.

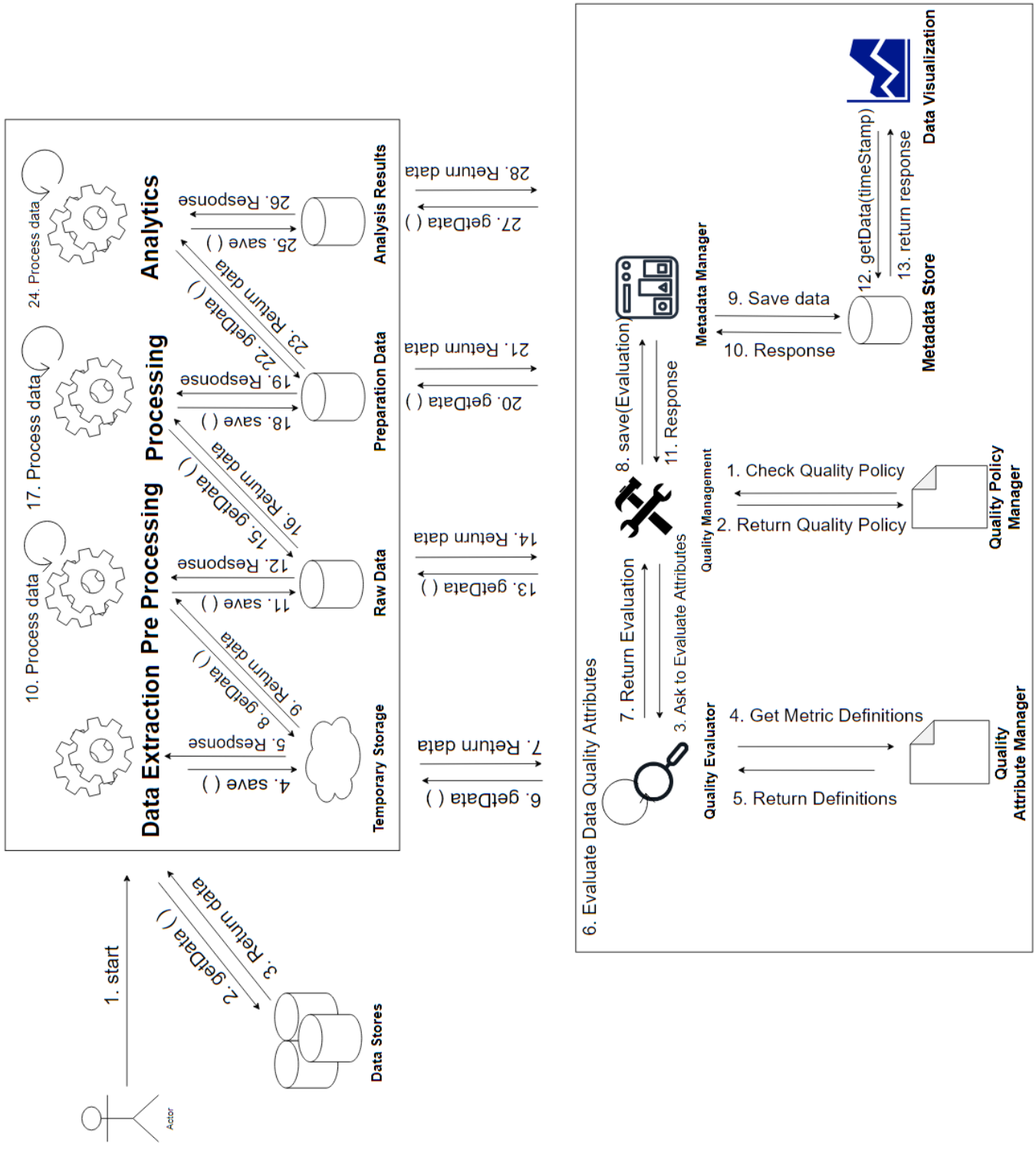

Figure 6.5: Detailed Quality Diagram

Fig. 6.5 describes in more detail how the Big Data Quality Architecture functions. The Quality Manager will periodically get data from each of the available data stores based on the scheduling policy described in the Quality Policy Manager. Once data has been retrieved the Quality Manager will check with the Quality Policy Manager to determine which quality attributes need to be evaluated. Once done the data is sent to the Quality Evaluator and the metrics are measured then returned back to the Quality Manager to be saved. The Quality Manager passes the data onto the Metadata Manager where it is structured and stored in the Metadata Store. This all happens parallel to the Big Data Architecture pipeline.

6.6 Case Study

A suitable example to illustrate the flexibility of the MEGA framework would require us to use data that is both real-time, time-stamped, freely available and traceable to a source for the purposes of reproducibility. We therefore opted to use stock market data from Yahoo's Finance API to build out a simple Big Data Pipeline architecture that was used to insert our MEGA Framework onto it.

6.6.1 Measuring Data Through the Pipeline

Our MEGA Framework was used to collect data quality measurements each time a data frame was created or modified at some stage in the pipeline.

The first step of the MEGA workflow is to select an appropriate measurement information model and to schedule the measurement data collection and analysis. In this example, we selected the 6V's measurement information model and implemented the measurement methods of its base measures Ndde, Lbd, Nrec, Lrec, Nds (please, refer to section II.A). These base measures were collected at each step of the Big Data Pipeline, including Data Extraction, Preprocessing, Processing, and Analysis.

Additionally, we automated the calculations of the 3V's derived measures and the corresponding indicators Mvol (Volume), Mvel (Velocity), Mvar (Variety), Mver (Veracity), Mval (Validity), & Mvinc (Vincularity). The above measurement procedures were implemented in the Quality Attribute Manager and described in the Quality Policy Manager. We scheduled the Quality Policy Manager to collect the measurement data at the end of each phase in the pipeline, except for the Data Analysis phase. Typically, we would recommend that measurement data to be collected at each stage and even at the end of the pipeline. However, in this case study we aim to show the flexibility of the MEGA framework in allowing the data engineers to choose where to collect and analyze the measurement data, and where they may feel it isn't necessary in the context of their customized workflow.

6.6.2 Measuring Data Through the Pipeline

The steps of the MEGA framework are designed to be easily learned and applied in different contexts of usage.

The process starts with ***Data Extraction.*** In this step, we extracted pricing data for Apple (AAPL), Tesla (TSLA) and the S&P 500 from Yahoo's Finance API using a ***datareader*** library provided by Pandas. The data we collected initially from each stock included the date, high, low, open, close, volume and adjusted close. This step is part of the ***data extraction phase***. In our case study, the data extraction started with the time frame T1. T1 is defined as the time period between January 1, 2021, and January 6, 2021. There were three-time frames in total, with each time frame lasting 5 consecutive days and each new time frame beginning when the previous one ended.

Data Preprocessing. In this step the pipeline is set to remove any unnecessary columns from the extracted dataframes; this meant removing 'Adj Close' for the Tesla and Apple data frames respectively. For the S&P 500 dataframe, all features except 'Close' were removed for the reason that in this pipeline our goal was to predict the next Close of the S&P 500 based on financial data from Tesla and Apple.

Data Processing. This step involved using pandas' in-built merge library for T1 to T3 to merge three data frames corresponding to each of the tickers (TSLA, APPL, S&P 500) into one larger dataframe for further analysis. This new dataframe was also evaluated by the MEGA Framework and its measurement data, as well as the analysis results, were stored in the Metadata Repository.

Data Analysis. This step involved splitting the data into test and training sets. The training set was used to train a KNN model *where n_neighbors* was set to 3, *p* to 1, and the metric used was '*minkowski*'. As users, we decided that our Quality Policy Manager will not collect any measurement data in this step thus our Quality Manager did not request the pipeline for more data.

Data Extraction, ***Data Preprocessing***, ***Data Processing*** and ***Data Analysis*** steps are repeated sequentially for each time frame (T1 to T3 in this example). Once each phase completes its process, the MEGA framework collects base measures described in the *Quality Policy Manager*, by using the Quality Manager to request and retrieve the measurement data automatically from the pipeline. Once collected, the Metadata Manager records the measurement data acquired. Finally, we used

the Metadata Repository, along with the Quality Attribute Manager, to build visualizations of the 6V's indicators after collecting and analyzing the measurement data. These visualizations were defined by us for what we believe to be most relevant to the context of our work. In our case study we displayed the updated visualizations at the end of the pipeline; however, data engineers may choose to update a dashboard of important indicators in real time as new data is constantly being added. Either approach will work depending on the context and the needs of the data engineers and users.

6.6.3 MEGA Results of Stock Analysis

As mentioned previously, the Quality Attribute Manager is responsible for not only implementing the measurement methods for the base measures, but also for defining procedures for the measurement data's statistical analysis and the visualization of the quality indicators. The case study that was implemented here depicts the visualizations of the 6V's indicators to rapidly turn the collected measurement data into easy to interpret graphs, as shown below.

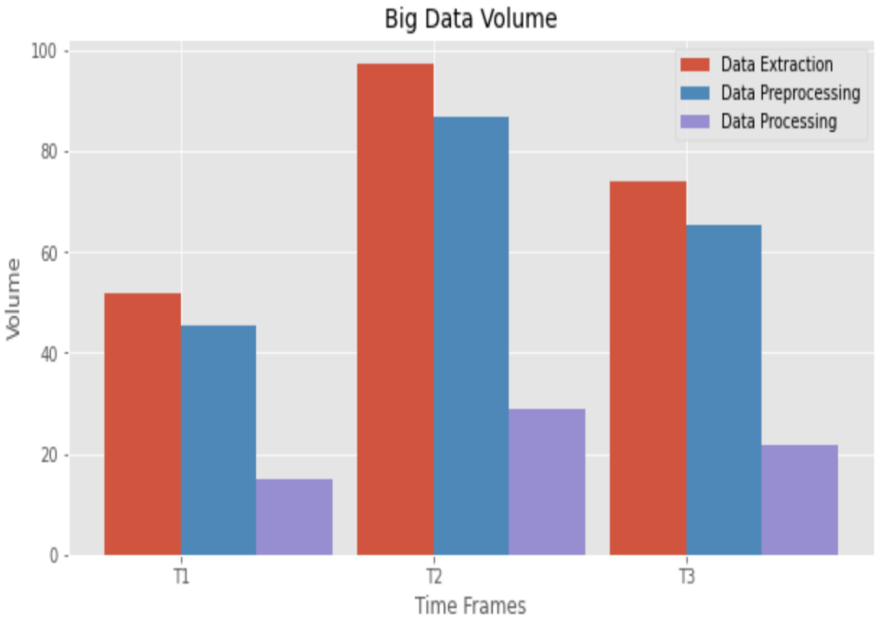

Figure 6.6: Big Data Volume for Stock Analysis

Fig. 6.6 visualizes the indicator Mvol gathered through T1 to T3. We see clearly here that, as data flows through the sequence of steps involved in our process, the amount of data decreases between the extraction, pre-processing, and processing phases. Data Volume Mvol may seem

rather trivial, however, if this were a real-time system and we noticed a lower than average or completely empty time frame, it may be a cause for serious concern, for instance, a server going down. In this case, we notice that pre-processing and processing are both doing their jobs at reducing our overall volume.

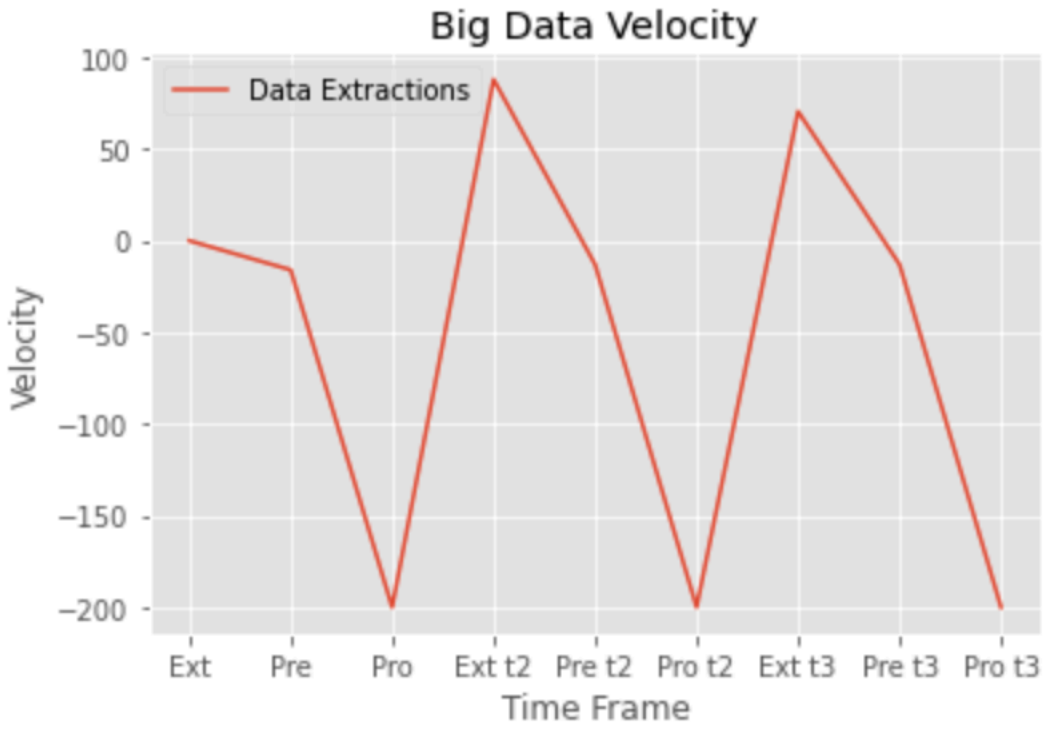

Figure 6.7: Big Data Velocity for Stock Analysis

Fig. 6.7 illustrates the Velocity (*Mvel*) measurement results. As we saw before, our pre-processing and processing cause significant decreases in Volume with only the addition of new data from times T2 and T3 creating an increase. We also notice very quickly from our graph that the rate at which we lose data from each time frame is roughly the same. This is very good news as it's to be expected since we filter the data in the same way during each phase.

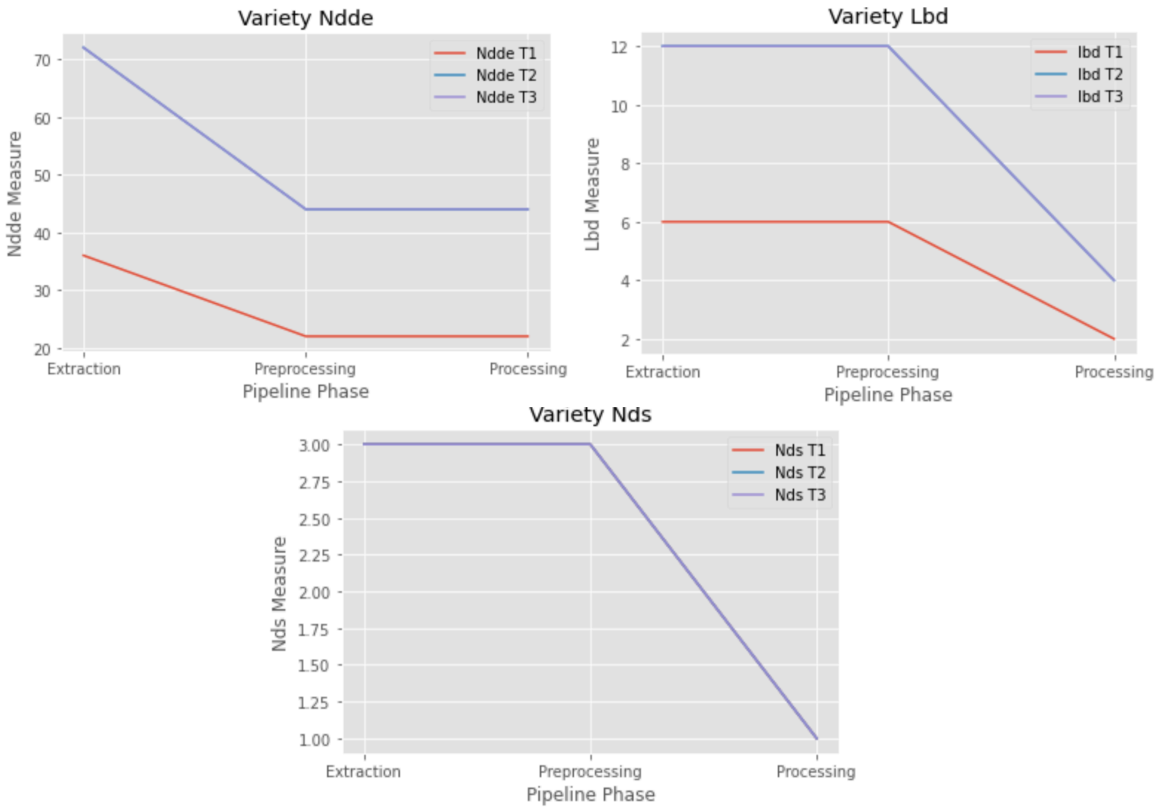

Figure 6.8: Big Data Variety for Stock Analysis

Fig. 6.8 illustrates the Variety (*Mvar*) measurement results. Because Variety is a multidimensional object, visualizations of the Mvar indicator are spread over multiple graphs, each illustrating a different measure. *Mvar*, as described in section II.B, is a tuple composed of *Ndde, Lbd* and *Nds*. These metrics serve as indicators that allow data engineers to measure and visualize the effect their automation has on the data. For instance, in Fig. 6.8 *Ndde* consistently has a drop occurring at the pre-processing phase. This is to be expected since we remove features (or columns) in the dataset. This explains why during the same phase *Lbd* remains constant. We see that the opposite occurred in the Processing phase. Merging datasets together reduces the number of records but keeps *Ndde* the same. Visualizing data quality characteristics is very beneficial for data engineers. For example, *Mvol* indicated that there was a significant reduction in our data volume, but the reason wasn't obvious. The *Mvar* indicator showed that *Nds* was reduced from 3 to 1, as expected, because we merged these datasets. However, the consequence of that was that *Lbd* was cut by a third. From this analysis we know that we would need 3 times the amount of information during data extraction in order to maintain the number of records we thought we would have for our ML model to work with.

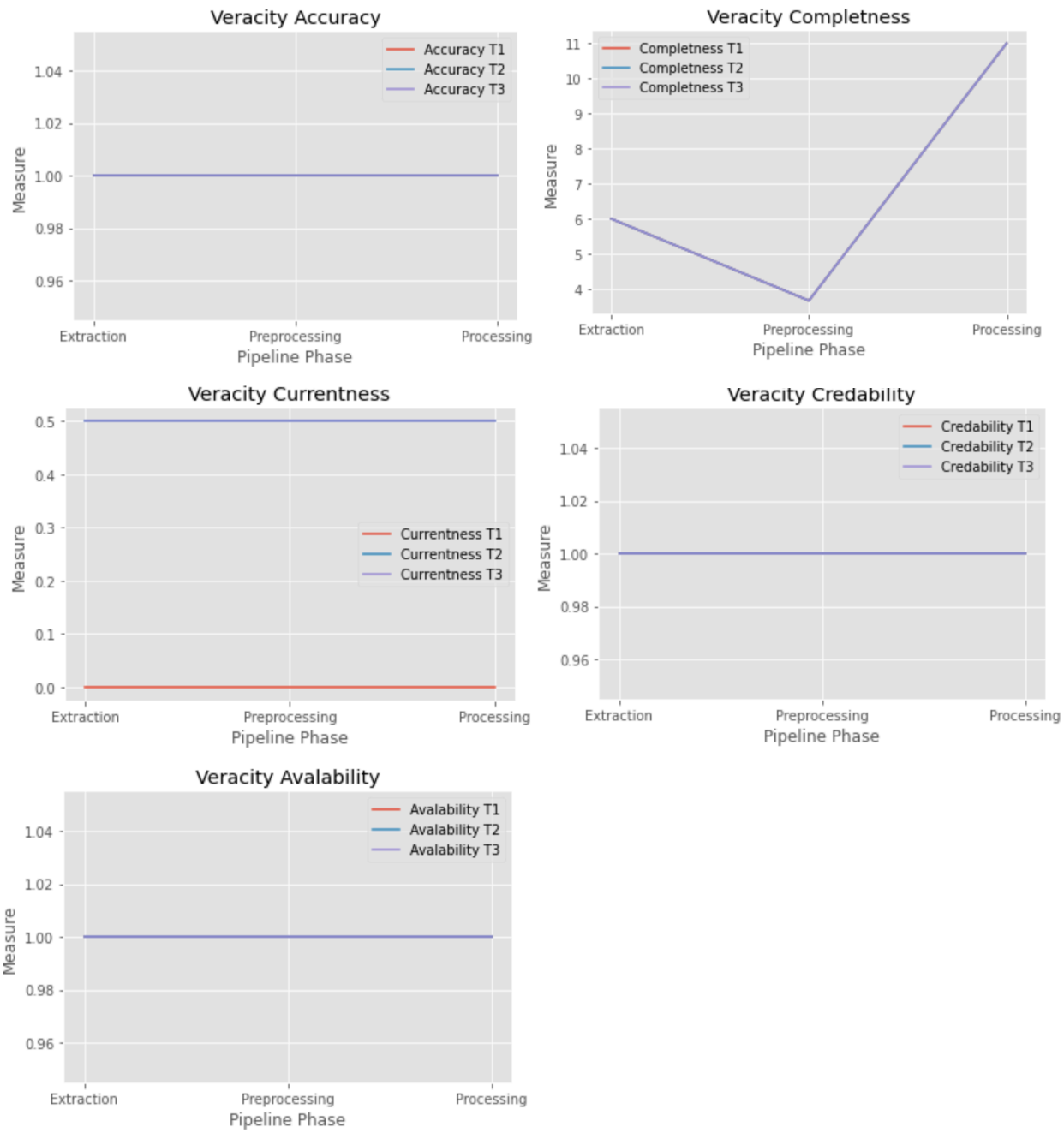

Figure 6.9: Big Data Veracity For Stock Analysis

As with Variety, Veracity is similarly a multidimensional object composed of five indicators. These include Accuracy, Completeness, Currentness, Credibility and Availability. The way in which these indicators are measured are specified in the paper (Ormandjieva, Omidbakhsh & Trudel, 2021). The first is indicator that is interesting to look at is Accuracy. As Accuracy is defined in section 4.4.2, we know that if it's the case that Accuracy is 100%, then there are no

duplicate data rows. This is highly likely because in the case of stock data, the time stamp will always be unique in the time column, in that specific dataset. However, having the same time stamp across multiple datasets will occur, since the stock exchange opens and closes on the same days and times for every stock listed on the exchange. However, it's highly unlikely for the Open, High, Low, Close, Adj Close and Volume values to all be the same across separate stocks. Additionally, none of our processing would change any of those results. Therefore, we can logically expect a high or perfect accuracy score.

Completeness is another indicator related to Variety. In terms of Completeness at the level of MDS, Completeness checks for the percentage of rows with null values across MDS. Here, Figure 6.8 shows us that even from the data source (the Yahoo Fainace API), we receive complete data and that our processing does not disturb that.

Currentness, as implemented in (Ormandjieva, Omidbakhsh & Trudel, 2021) uses a box plot method to determine a range of acceptable data. We find T1 has a score of 0, throughout all processing phases. The fact the score does not change at any phase makes sense. Obviously, currentness is based on time stamps and those come directly from the moment the stock data was collected by the API. However, it is interesting that T1 is still 0 given that T2 and T3 are only a few days away. Upon further investigation we found that T1 only contains 2 datapoints and that their ages were 222 days old and 221 days old. Because of the methodology used the boxplot gave a range of 221.25 days old – 221.75 days old as the acceptable range. This is interesting as it acts as an example of where our measurements may not always be perfect. In this case, neither 221 nor 222 fall within the acceptable range at T1. Of course, with more data points the range expands and we find that more can fit those acceptable ranges.

Credibility is the next indicator to make up Veracity. Here credibility is defined as the ratio of "credible" data to all data, according to the formula from (Ormandjieva, Omidbakhsh & Trudel, 2021). Because all data comes from a reputable source (Yahoo's API), we can assume that their data is credible. But in this instance, it's important to note that we chose to believe that the data provided by the API is credible but determining credibility is difficult and researchers will have to come up with better more robust solutions at determining whether something is truly credible or not.

Finally, we have the last indicator for Veracity, Availability. Availability as defined here (Ormandjieva, Omidbakhsh & Trudel, 2021), in the context of a measurement is the ratio of successful requests to total requests made to a datastore. In this case, we find that there are never any situations in which the data was unavailable. This indicator, while simple is very important as it has the ability to show researchers times at which their data or the data of others that they rely upon may be down or unavailable all together. This can help quickly diagnose problems related to missing data from a specific datastore.

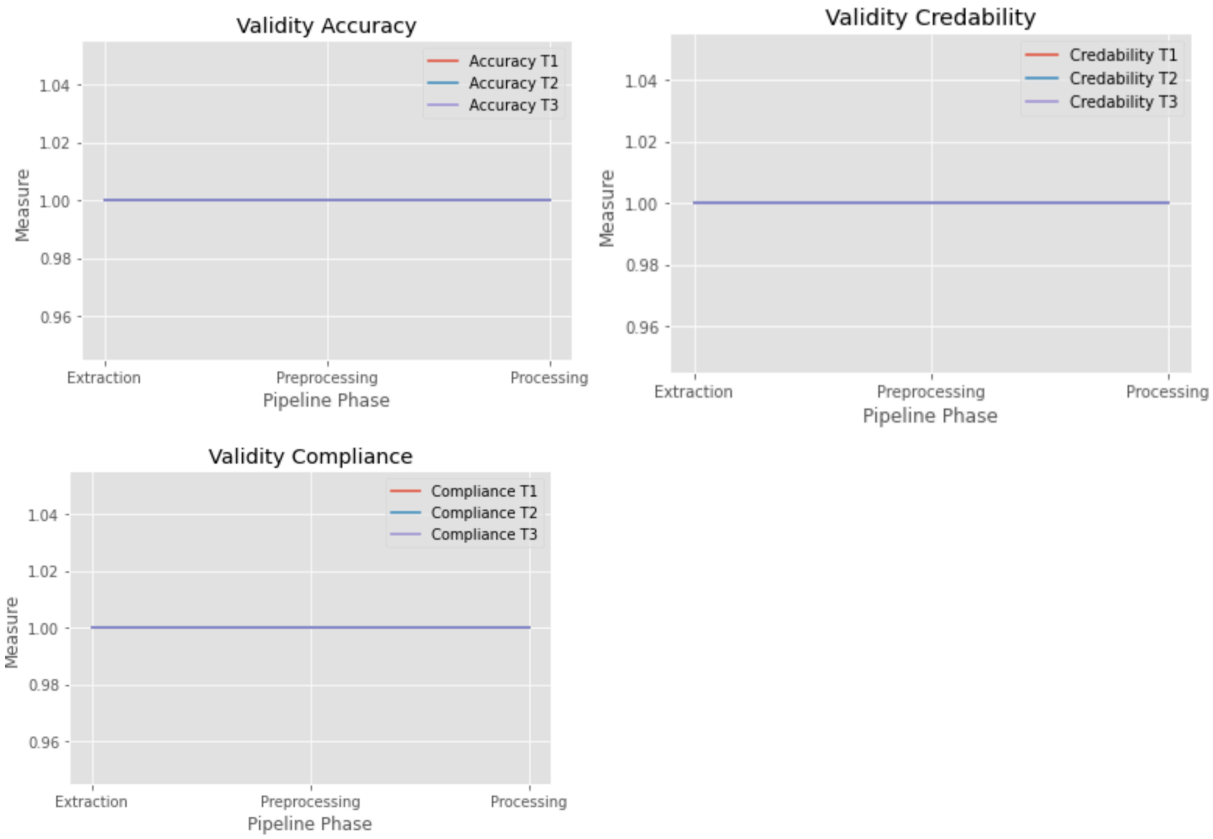

Figure 6.10: Big Data Validity For Stock Analysis

Figure 6.10 shows us the indicators used to determine Big Data Validity. Interestingly Validity shares some of its indicators with Veracity, these include Accuracy and Credibility and because we analyze the same data these visuals are the same, so they follow the same logic described earlier. Compliance, on the other hand, is new to Validity. Compliance is defined as how much the data itself adheres to any standards we place on it. In the case of stock data, we want to make sure that time stamps are formatted correctly and that the values are in USD. In this case, because we obtain our values from the Yahoo API, we find that our data is formatted correctly. In some other

situation it may be necessary for developers to have to manually write code using tools like RegEx to ensure that their data meets their specific needs.

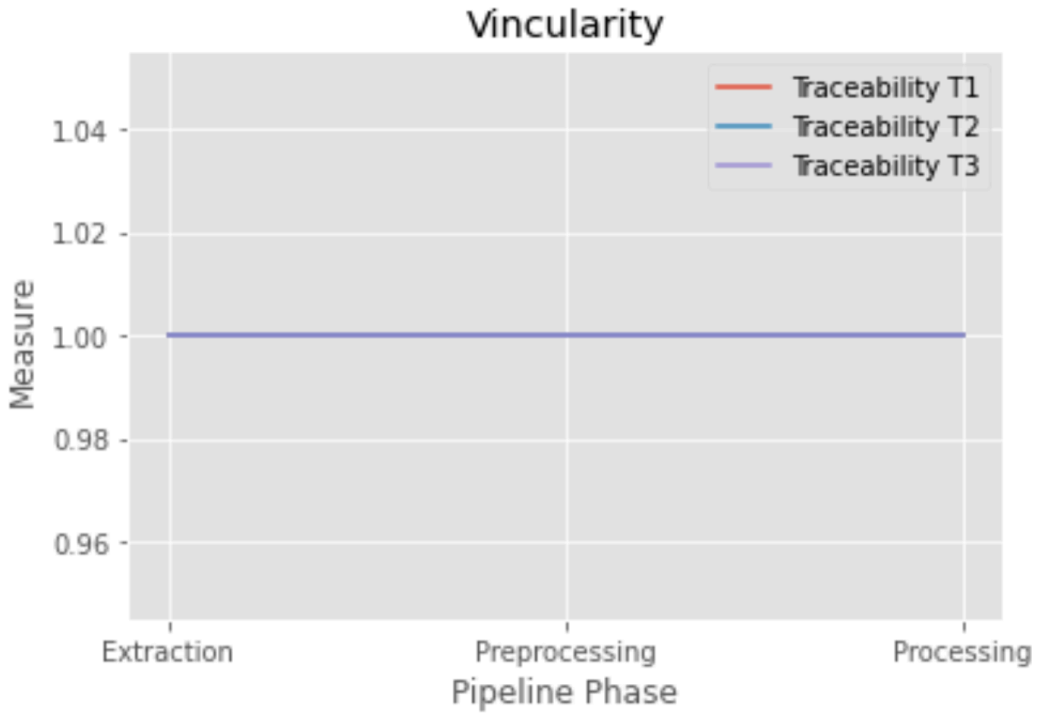

Figure 6.11: Big Data Vincularity For Stock Analysis

Figure 6.11 shows the Big Data Vincularity. Since Vincularity is composed of only one indicator Traceability (at the level of MDS). Here, Traceability is defined as how much of the data we change is stored in metadata, so that we can see how data changes from each phase. Here, because of the way the code is written we made sure to store versions of the data before making changes on to the datasets. As such, we have perfectly traceable data. But it's important to note that this is from the perspective of our own system. Our data is originally taking from Yahoo's API. How do we know what changes Yahoo's systems have made before we received the final version of the data? This is difficult to ask since we do not have internal information of how Yahoo Finance processes their data and in a way, this leaves an important part of traceability still on the table. In the future we'd like to solve this issue or a least provide more guidelines in dealing with visualizing traceability of data when its source is unknown.

6.7 Conclusion

In this chapter we've described the MEGA architecture in detail and have demonstrated how it may be used to measure and analyze the underling data of a Big Data Pipeline that will be built or a pre-existing one. The architecture allows for the V's of big Data to be used in real world systems and provides can sub-systems, such as the Quality Policy Manager, that can be customized to fit a variety of needs of users. This makes the architecture very powerful as it's able to be flexible for a variety of problems that exist in the industry.

The framework currently has been implemented and tested on 6 V's (Volume, Velocity, Variety, Veracity, Validity and Vincularity). This is a good start but, in the future, we'd like to complete our development of objective measures and have metrics for the remaining V's. We'd also like to implement these into our framework to hopefully allow users to have much greater insight into their systems and provide a greater overall metric for the "quality" of their Big Data systems. Additionally, in the future, our goal is to benchmark the performance cost of having the framework running while the pipeline runs in parallel. We believe that researchers understanding the performance cost of the system is important for how they may want to design their approach given their requirements.

Chapter 7 Conclusions and Future Work

In this thesis, we presented a novel measurement framework MEGA for monitoring the quality characteristics of Big Data (the V's). The newly proposed MEGA framework can be used for assessing the alignment of Big Data solutions to the needs of their users, guided by users' best interests, before committing to any alternative. We proposed two new theoretically valid measurement information models to evaluate Validity and Vincularity of Big Data in the context of the MEGA framework. The models' elements that we've developed are compliant with ISO/IEC/IEEE Std. 15939 guidelines for their definitions. We've demonstrated their theoretical validation and have shown that they can be used along side the four other indicators (Ormandjieva, Omidbakhsh, & Trudel, 2020) (Ormandjieva, Omidbakhsh, & Trudel, 2021). These models are suitable for Big Data in any forms of structured, unstructured, and semi-structured data.

We described the proposed MEGA architecture applicable to a variety of existing Big Data Pipelines. Our MEGA solution takes into consideration the flexibility that would be required by the data engineers when measuring specific data and specific points in the Big Data pipeline. We illustrated the MEGA framework by collecting measurement data on V's of Big Data from Yahoo's Finance API stock market data and showed how the framework can be used to monitor data quality issues that may arise.

We will also focus more heavily on testing the framework on a variety of different pipelines including handling real-time data streaming as well as trying to understand the overhead cost involved with using the Framework. We plan on building out the libraries for the framework in a two-month projected period, with testing taking 6 to 8 months.

The proposed novel measurement framework MEGA is also forward-looking — considering not only the understanding of the Big Data quality characteristics that are generally accepted today (the V's), but also exploring emerging new quality characteristics rooted in innovative usages of Big Data. These include having the pre-built libraries that can be used by developers and easily integrated instead of having a framework that is abstract.

References

Allan Koch Veiga, A. M. (2017). A conceptual framework for quality assessment and management of biodiversity data. *PLOS ONE, 12*(6), 178731. From https://doi.org/10.1371/journal.pone.0178731

Andreu-Perez, J., Poon, C., Merrifield, R., Wong, S., & Yang, G. (2015). Big Data for Health. *IEEE Journal of Biomedical and Health Informatics, 19*(4), 1193-1208.

Bates, D., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big data in healthcare: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs, 33*, 1123-1131.

Bhardwaj, D., & Ormandjieva, O. (2021). Rigorous Measurement Model for Validity of Big Data: MEGA Approach. *Proc. 25th International Database Engineering & Applications Symposium (IDEAS'21)*.

Bhardwaj, D., & Ormandjieva, O. (2021). Toward a Novel Measurement Framework for Big Data (MEGA). 3rd IEEE International Workshop on Big Data Computation, Analysis & Applications (BDCAA 2021). *Proc. IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC'21), 1580-15887.*

Cedrine, M., & Alain, M. (2017). *US Patent No. US9858420B2.*

Emden, M. H. (1975). An analysis of complexity. *Mathematisch Centrum Amsterdam*, 86.

Fenton, N., & Bieman, J. (n.d.). *Software Metrics: A Rigorous and Practical Approach* (3rd edn ed.). CRC Press. From https://doi.org/10.1201/b1746

Gudivada, V., Apon, A., & Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software, 10*(1), 1-20.

(1998). *IEEE Standard for Software Quality Metrics Methodology.* IEEE Std 1061-1998.

Immonen, A., Pääkkönen, P., & Ovaska, E. (2015). Evaluating the quality of social media data in big data architecture. *IEEE Access, 3*, 2028-2043.

(2015). *ISO/IEC DIS 25024 Systems and Software Engineering- Systems and Software Quality Requirements and Evaluation (SQuaRE)-Measurement of data quality.*

(2017). *ISO/IEC/IEEE International Standard - Systems and software engineering--Measurement process.* ISO/IEC/IEEE 15939:2017.

Jorge Merino, I. C. (2016). A data quality in use model for big data. *Future Generation Computer Systems, 63*, 123-130.

Kelly, B., & Knezevic, I. (2016). Big Data in food and agriculture. *Big Data & Society, 3*(1), 205.

Lee, J. (2013). Recent advances and trends in predictive manufacturing systems in big data environment. *Manufacturing letters, 1*(1), 38-41.

(2018). *NIST Big Data Interoperability Framework: Volume 1, Definitions. Volume2, Big Data Taxonomies.* Commerce, NIST U.S. Department of Commerce.

Omidbakhsh, M., & Ormandjieva, O. (202-). Toward A New Quality Measurement Model for Big Data. *Proc. 9th International Conference on Data Science, Technology and Applications (DATA)*.

Ormandjieva, O., Omidbakhsh, M., & Trudel, S. (2020). Measuring the 3V's of Big Data: A Rigorous Approach. *Proc. IWSM-MENSURA 2020: In Proc. Joint conference of the 30th International Workshop on Software Measurement (IWSM) and the 15th International Conference on Software Process and Product Measurement (MENSURA)*.

Ormandjieva, O., Omidbakhsh, M., & Trudel, S. (2021). Measurement Model for Veracity of Big Data. *The 7th International Symposium on Big Data Principles, Architectures & Applications (BDAA 2020). As part of 18th International Conference on High Performance Computing & Simulation (HPCS 2020). In Proc. HPCS 2020.*

Pääkkönen, P., & Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems. *Big Data Res*.

Pääkkönen, P., & Pakkala, D. (2015). *Reference architecture and classification of technologies, products and services for big data systems,'.* chicago: Big Data Res.

Ramaswamy, L., Lawson, V., & Gogineni, S. (n.d.). Towards a quality-centric big data architecture for federated sensor services. *Proc. IEEE Int. Congr. Big Data*, 86-93.

Rashidi, C. M. (2014). *US Patent No. US20160147798A1.*

Shkapenyuk, V., Dasu, T., Srivastava, D., & Swayne, D. (2021). *US Patent No. 16257936.*

Staeben, C., Maier, C., Savard, B., Wilbur., A., & BV, H. G. (2020). *US Patent No. US20200159702A1.*

Taleb, I., Dssouli, R., & Serhani, M. (2015). Big data pre-processing: A quality framework. *Proc. IEEE Int. Congr. Big Data*, 191-198.

Walker, R. (2015). *From big data to big profits: Success with data and analytics.* Oxford University Press.

Yu, S., & Song, G. (2016). *Big data concepts, theories, and applications.* (Y. Shui, & S. Guo, Eds.) Springer.

Zhou, N., Huang, G., & Zhong, S. (2018). Big data validity evaluation based on MMTD. *Mathematical Problems in Engineering, 6*, 1-6.