FRAMEWORK TO ENHANCE VERACITY AND QUALITY OF BIG DATA

FAKHITAH BINTI RIDZUAN

UNIVERSITI SAINS MALAYSIA

2021

FRAMEWORK TO ENHANCE VERACITY AND QUALITY OF BIG DATA

by

FAKHITAH BINTI RIDZUAN

**Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy**

October 2021

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful

I would like to extend my deepest praise to Allah S.W.T for the blessings and guidance in completing this research and thesis. Also, special thanks and gratitude to; Public Service Department Malaysia for the financial support throughout my studies, and also Universiti Sains Malaysia for the resources and facilities provided. Alhamdulillah, praise to Allah, finally I can complete this research.

To my parents; Salma Ismail and Dr Ridzuan Ahmad. I simply cannot thank you enough; I would have never been able to succeed in my PhD without your unconditional love, encouragement, and faith. I will forever be indebted for your efforts in raising me and the education you provided me.

To my supportive husband, Nik Mohamad Sharif, I am grateful for sharing my PhD journey; it has been a unique experience for both of us. I am eternally grateful for your endless love and support during this rough period. Being my husband is the best thing that ever happened to me during my PhD. To my beloved son, Nik Aidan Syamil, thank you for lightens up my PhD journey, and this is for you.

I would like to express my deep and sincere gratitude to my supervisor, Dr Wan Mohd Nazmee Wan Zainon. Thank you for giving me the opportunity and provide me invaluable guidance throughout my PhD. I am extremely grateful for what he has afforded me.

I would like to take this opportunity to say thank you to Dr Mohd Khairul Nizam, Dr Nur Ashikin Marzuki, Dr Mazidah Rejab, Dr Nur Liyana Sulaiman, Dr Nurul Asyikin Ibharim, Tasnim Ismail, Azratul Hizayu Taib, Nur Hidayah Ahmad and Noor Ropidah Bujal. Thank you for being with through thick and thin.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF APPENDICES

KERANGKA UNTUK MENINGKATKAN KESAHIHAN DAN KUALITI DATA RAYA

ABSTRAK

Sejumlah besar data tersedia untuk organisasi bagi memacu perniagaan mereka dan menandingi pesaing. Data dikumpulkan dari pelbagai sumber adalah kotor dan ini akan mempengaruhi keputusan perniagaan mereka jika tidak dikendalikan dengan betul. Pelbagai kajian telah dijalankan untuk membersihkan data tersedia dan mengatasi masalah data kotor dengan menawarkan kualiti data yang lebih baik. Ini akan membantu organisasi memastikan data mereka siap untuk dianalisis. Namun begitu, timbul beberapa permasalahan yang dibangkitkan mengenai kepercayaan hasilnya walaupun kualiti data yang tinggi. Kesahihan adalah salah satu ciri data raya, yang merujuk kepada tahap kebolehpercayaan terhadap data. Ia berkait rapat dengan kualiti data, tetapi tidak banyak kajian yang dijalankan yang khusus membincangkan standard bagi menentukan kualiti data khusus untuk data raya. Selain itu, kajian-kajian lepas yang dijalankan menunjukkan bahawa perlunya peraturan kualiti data untuk memastikan data yang dibersihkan adalah berkualiti. Namun, proses ini memerlukan pakar domain untuk memastikan peraturan yang ditetapkan adalah betul dan tepat. Oleh itu, penyelidikan ini mencadangkan kaedah baru untuk menukarkan kaedah sediakala kepada penghasilan peraturan kualiti data secara automatik dan kerangka penilaian kesahihan yang dipertingkatkan. Pendekatan utama adalah untuk mengekstrak peraturan kualiti data secara automatik daripada sumber data, yang akan mengurangkan interaksi sistem dengan pakar domain. Pendekatan yang dicadangkan akan dinilai menggunakan Kerangka Penambahbaikan Kesahihan (VEF) untuk memastikan data telah memenuhi dimensi kualiti data dan dapat memberikan hasil yang dapat dipercayai. Hasil eksperimen menunjukkan bahawa teknik automatik yang dicadangkan untuk mengekstrak peraturan kualiti data, mampu mengelaskan 9487 data dengan betul dengan peratusan kesalahan sebanyak 4.6% sahaja. Manakala untuk pelanggaran corak, VEF mencatatkan 1 untuk nilai kejituan dan pengingatan kembali. Selanjutnya, kerangka kerja kebenaran yang dipertingkatkan memberikan hasil yang berbeza dari ukuran kualiti data sedia ada. Sumbangan utama tesis ini adalah VEF yang mampu menilai tahap kualiti data bagi memastikan tahap kebenaran dan kualiti data mencapai titik ambang. Selain itu, teknik automatik untuk mengetahui peraturan kualiti juga mampu mengurangkan interaksi dengan pakar domain. Oleh kerana penyelidikan ini hanya tertumpu pada data berstruktur, maka penelitian ini dapat diperluas untuk memenuhi jenis data tidak berstruktur dan separa berstruktur.

FRAMEWORK TO ENHANCE VERACITY AND QUALITY OF BIG DATA

ABSTRACT

Massive amount of data are available for organisations to drive their business ahead of the competitors. Data collected from a variety of resources are dirty, and this will affect their business decisions. Various data cleansing tools are available to cater to the issue of dirty data. They offer better data quality, which will be a great help for the organisation to make sure their data is ready for the analysis. However, there has been an issue raised regarding the trustworthiness of the result, even though the quality of the data is high. Veracity is one of the characteristics of Big Data, which refers to the trustworthiness of the data. It always relates to data quality, but there has been less work on a standard that defines data quality, specifically for Big Data. Besides, most of the studies also show the need for data quality rule to satisfy a variety of errors present in the data. However, this process requires a domain expert that is expensive to employ. Consequently, this research proposes a method to automate data quality rules and an enhanced veracity assessment framework. The proposed method will automate the process of extracting data quality rules from the data source, which will reduce the interaction with the domain expert, and at the same time correctly verifying and validating the rules. The proposed method will be evaluated using the Veracity Enhancement Framework (VEF), to make sure the data has met the data quality dimension and able to deliver trustworthy result. The experimental result shows that the proposed automatic technique to extract data quality rules is able to correctly classify 9487 data with 4.6% error percentage. Meanwhile, for pattern violation, the value of precision and recall for VEF is 1. Furthermore, the enhanced veracity framework provides distinct results from the data quality measure. The main contribution of this thesis is the VEF, which is able to compute the quality of the data to make sure the data is in high veracity and high quality. Since this research is focused on structured data only, thus the research may be broadened to cater for unstructured and semi-structured data types.

CHAPTER 1

INTRODUCTION

1.1 Overview

In recent years, there has been an increasing interest in Big Data analytics, a complex process consist of collecting, organising, and analysing a large dataset to uncover the hidden pattern, market trends, and other useful information. The benefits of Big Data analytics are not restricted to a specific type of industry since analytics has proven to be critical for organisations to stay on top their competitors. The growing enthusiasm for making a decision based on data creates the importance of accurate and precise prediction.

The volume of data created, captured, and copied by the organisations has been increasing every year, and the process of analysing the data quickly becomes essential. Unfortunately, the process to analyse the vast amount of data is creating new challenges (Khan et al., 2018). Big Data is a term that relates to massive, heterogeneous, and often unstructured digital data that are difficult to handle with traditional data management tools and techniques (Rodríguez-Mazahua et al., 2016).

In general, Big Data refers to large and complex datasets that are often used for predictive analytics. Big Data is designed to gather, store, and manage the data through advanced analytic techniques and application. The rapid growth of the data drives new business opportunities if the data collected are handled properly (Sivarajah et al., 2017). However, data preparation has become more challenging and time-consuming, as the volume and variety of data has been increasing in recent years (Tian, 2017). The collected data need to be processed first to ensure its reliability before analytics can occur. This crucial phase will affect the result if unreliable data are used for the analysis. Thus, data cleansing is a critical process to make sure that the data is free from any type of errors.

Big Data can generally be characterised into five main dimensions, called 5V's. Figure 1.1 depicts a diagram for the 5V's of Big Data.

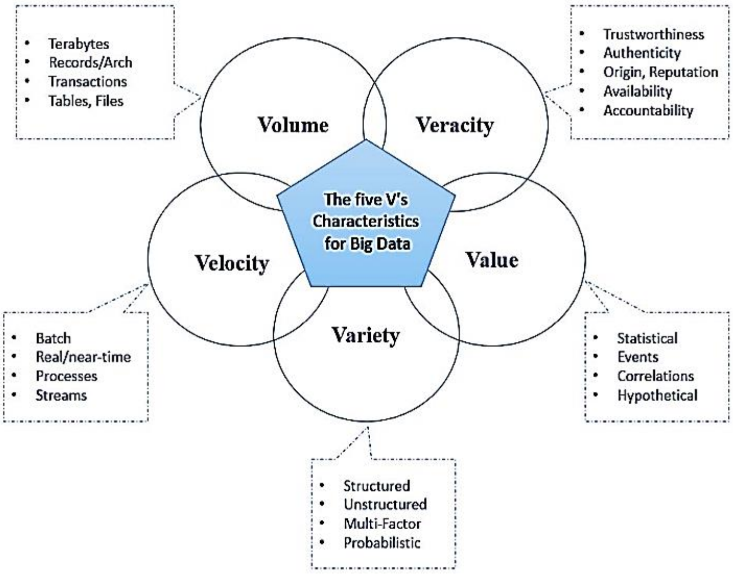

Figure 1.1 Big Data Characteristics (Hadi et al., 2015)

1. Volume

Big Data volume refers to the size of the data obtained, which is the most essential and distinctive feature of Big Data. Considering volume as the most effortless characteristic of Big Data to be defined, it also causes various challenges (L'heureux et al., 2017). There is no standard volume of how big the datasets should be considered as Big Data, and it may vary depending on the industries and application (Sonka, 2017). Enterprises are cramped with enormous amounts of information every day, and it is being collected at an unprecedented scale. This situation decreases the value of data collected and indirectly affects data quality and data analysis.

2. Velocity

Data velocity intends to measure how quickly the data are processed from the creation until the visualisation of the data. It can be created in real-time or near real-time. Sonka (2016) stated, '*Velocity refers to the ability to understand and respond to the events as they occur*'. How the data is being analysed, depends on the activities, whether it requires an immediate response. Velocity is important as time-sensitive activities, such as fraud detection requires Big Data to analyse the pattern of real-time transactional data and find the correlation between multiple sources (Oracle Corporation, 2015).

3. Variety

Variety refers to the different types of data that can be processed, which can consist of structured, semi-structured, or unstructured data. Text, images, videos, and audios are the examples of the data stored in the database. The variety of data is expanding wildly as the handwriting of a person might be captured and stored as data. Data variety might be the biggest obstacle to effectively use the large volume of data for the analysis (Swapnil et al., 2016), as accuracy and veracity become more challenging to establish in Big Data analytics (Hariri et al., 2019). Challenges such as incompatible data formats, incomplete data, non-aligned data structures, and inconsistent data can affect the analysis result.

4. Value

Data value refers to a measure of the usefulness of the collected data in decision-making (Swapnil et al., 2016). Having more data does not mean having better data. Big Data analytics is not about knowing the data, but to discover the predictive power behind the data collected. Value is closely related to the volume and variety because it depends on the events or processes, such as stochastic, probabilistic, regular, or random (Hadi et al., 2015).

5. Veracity

Veracity often relates to the trustworthiness and reliability of the data. According to Lozano et al., (2020), veracity is necessary when data is analysed and used for decision making because it deals with uncertain and imprecise data. In order to ensure the veracity of the data, the polluted data need to be sanitised by eliminating all the errors found inside the dataset (Rodríguez-Mazahua et al., 2016). Veracity is often associated with data quality, as it defines how data can be trusted and used for decision-making (Khan et al., 2018).

Volume, velocity, variety and value are indeed important in Big Data, but ensuring veracity is what makes Big Data useful. The first and essential step in order to get an accurate result is to have the correct data source. Therefore, this research aims to improve the veracity of the data source by proposing a Veracity Enhancement Framework (VEF) that will be able to automate the discovery of data quality rules and evaluate the veracity of the result.

1.2 Research Background

Recently, Big Data and data analytics have become exceptionally valuable in many areas of computer science, medicine, finance, and retail. Each day, Google processes 3.5 billion searches, and Facebook users upload 300 million photos, 510,000 comments, and 293,000 status updates (Hariri et al., 2019). This phenomenon has generated petabytes of data in various formats, such as structured data and unstructured data, which adds to the complication of data cleansing. The process of getting data is becoming easier; however, the quality of data remains to be a significant concern, as dirty data can lead to incorrect decisions and unreliable analysis (Chu et al., 2016). There is no longer an issue with the shortage of data; instead, a new problem arises to get good training data. Besides, new knowledge discovery techniques and methods are required to analyse these data (Corrales et al., 2018).

In knowledge discovery, the preliminary step is pre-processing, which includes cleansing of raw data. Data cleansing processes can take up to 50% of the effort and take up 80% of the time in a data mining project (Ismail et al., 2019). Data mining and cleansing are the most critical stages of processing to extract insights from Big Data. Data cleansing is defined as the continuous process that consists of detecting, identifying, and, imputing anomalous data (Khayyat et al., 2015). It deals with the errors, missing values, and inconsistencies within the data in order to improve the quality of the data. It is important to have a detailed understanding of the data sources in order to select, clean, construct, and format data (Ismail et al., 2019).

Data cleansing techniques address data quality and uncertainty problems resulting from the variety of Big Data (Hariri et al., 2019). Wang et al. (2017) stated that variety is one of the possible reasons for data quality problems as data comes from various sources that may contain inconsistencies, conflicts, and incompleteness. The tremendous growth of data volumes has complicated the task in validating the quality of Big Data. It is the degree to which the data is suitable for use in the required business processes and has a high impact on a process, such as data mining, knowledge discovery, and trend analysis executed in the existing data warehouse (Alotaibi, 2017). Data quality issues happen due to improper maintenance and will indirectly generate inconsistency in the database (Cohen et al., 2015). It is one of the obstacles to use the data effectively as dirty data may lead to a false decision. Moreover, the velocity of data contributes to frequent changes and it is difficult to maintain it manually.

Dirty data is inaccurate, inconsistent, and incomplete due to the errors found within the dataset. In the era of Big Data, the existence of dirty data is common, and it is becoming an obstacle in providing accurate results and may result in misguided decisions (Feng, 2018). Improving data quality is very important for data mining and data analysis to avoid losses, problems, and additional costs due to the poor quality of data. Data quality has become one of the most important data management issues as data from various sources in various formats are widely available (Côrte-Real et al., 2020).

In 2017, four crew members of an Irish Coast Guard were killed when their helicopter was crashed into Black Rock (Nagle et al., 2020). According to the accident report, the island was not in the database of a key safety system onboard their helicopter. Apparently, if the pilots get too close to terrain or a known obstacle, the warning system will alert them to take corrective action to avoid the collision. The crew was alerted to the island 13 seconds before the impact, which raises the question on the level of data quality of the official map data used by Irish coast guard helicopters. What is worse is the fact that this data quality concern was raised four years prior to the tragedy.

On the other hand, Experian reports that most of the companies lose about 12% of their sales due to wrong records, subsequently adding to reduced productivity, loss of resources, and significantly misused chances for the marketing of cross-channel (Jesmeen et al., 2018). Based on the survey conducted by Experian, approximately one-third of respondents think that they waste nearly 10% or more of their marketing budget because of the result from inaccurate data. 83% of the respondents stated that poor data quality has hurt their business objectives, while 66% reported that poor data quality has had a negative impact on their organisation in the last twelve months

(Heinrich et al., 2018). The organisation needs to have an excellent quality of data in order to obtain more accurate and valuable results, since they depend on the data like customer relationship management and supply chain management

Due to the 5V's of Big Data, the complexity of the data quality algorithm becomes more complex (Cappiello et al., 2018). Besides, the way to assess the quality of data in the context of Big Data requires a different approach. Although there are different standards for evaluating data quality in a traditional context, none of these models are appropriate for Big Data environments (Merino et al., 2016). In the era of Big Data, data quality has been receiving a lot of attention because of its complexity and how it requires well-defined, lightweight measurement processes that can run simultaneously with each phase (Taleb et al., 2015). Besides, it relies not only on its features, but also on the data-driven business environment, including business processes and business users (Cai & Zhu, 2015), which causes data quality to face several issues.

Veracity is one of the critical V's in Big Data. It is referred to the trustworthiness of the processed data (Mittal, 2013). Even though the use of veracity not widely practised, but it also includes the measure of data quality. Poor data veracity may affect organisational performance and is considered as one of the obstacles for the successful use of Big Data (Accenture, 2018; Crone, 2016). Although veracity is getting more attention, very few research has been conducted to address the issues and challenges in veracity (Lukoianova & Rubin, 2014).

According to Grover et al. (2018), veracity is the biggest challenge in Big Data analytics as compared to volume, velocity, and variety, as it deals with data validity and trustworthy. What is more worrying when the report released by O'Reilly in early

2020, stated that data quality might get worse before it gets better (Magoulas & Swoyer, 2020). This shows the importance of handling and managing the issues of data quality to ensure that organisations can move forward using data analytics.

Considering the vast amount of data available today, assessing and evaluating information reliability is even more critical. According to Beretta (2018), this process is time-consuming, and it is impossible for a human to identify reliable data from a large amount of information manually. There are few tools available to tackle the process of analysing data quality of complex datasets with simple command-line tools and scripting libraries. However, it is difficult for people from non-IT background to understand the result provided. Moreover, most of the researches focus on the common data quality issue, models, and metrics in the traditional data warehouse (Gao et al., 2016). Even though there are already researches on data quality dimension focusing on both structured (Batini & Scannapieco, 2016; Firmani et al., 2016) and unstructured data (Cappiello et al., 2018); however, there is no standard guideline for the data quality dimension (Cai & Zhu, 2015).

Veracity is one of the severe challenges faced by data scientists as they need to distinguish meaningful data and dirty data (El-Ghafar et al., 2018). Accuracy and veracity are essential in data cleansing as these two characteristics complement each other. Most data cleansing framework designed for Big Data still use data quality dimension as their baseline to confirm the accuracy of the proposed frameworks. The main idea in evaluating veracity is to determine the fitness for the use of the data. There is no denying the importance of data quality dimension in order to measure the quality of the data. However, due to the volume and variety, data quality assessment framework alone is not enough to address the issue (Crone, 2016).

Moreover, the term ‘veracity’ or referred to trustworthiness (Al-Salim et al., 2018; Amini & Chang, 2017; De Tré et al., 2018), is often left out when dealing with Big Data. Since the credibility and quality of data directly influence the performance of the business decision, veracity helps to make accurate and precise decisions from data and select relevant information from given data with noise removal. Achieving high data quality is an important and challenging task, and simple cleansing methods do not eliminate unpredictability from data. It is also observed that volume is inversely proportional to veracity (Hariri et al., 2019), which means as the size of data increases, uncertainty and truthfulness decreases.

1.3 Background of Problems in Veracity and Data Quality

In oxford dictionary, veracity is conformity to facts, accuracy and habitual truthfulness (Amer, 2019). Veracity is sometimes thought to be uncertain or imprecise data but may be more precisely defined as false or incorrect data. The information may be falsified intentionally, negligently, or mistakenly. It can be distinguished from data quality, generally defined as data reliability and application efficiency, and sometimes used to describe information that is incomplete, uncertain or imprecise (Walker, 2015).

Veracity focuses on the trustworthiness of the data, but data quality focuses on the usability of the data. When it comes to the accuracy of data, it is not about quality of the data only, but how much the data source, type and processing can be trusted (GutCheck, 2019). If the data is objectively false, then any analytical results are irrelevant and unreliable regardless of any data quality issues. In addition, the lack of data creates an illusion of reality that can lead to bad decisions and fraud: sometimes with civil liability, or even criminal consequences.

Even though there have been some researches (Arolfo & Vaisman, 2018; Talha et al., 2019; Xiaojiang et al., 2017) that include trust measure as part of the data quality dimension, but to measure the veracity requires more sophisticated measures, instead of relying only on the trust. Veracity highlights the importance of trust in the data itself, not on the quality of the data. Based on the report by Accenture, 79% of executives agree that organisations are relying their most critical systems and strategies on data, yet many have not invested in the capabilities to verify the truth within it (Accenture, 2018).

1.4 Problem Statement

The presence of the Big Data wave is a turning point for all organisations including health, financial, retail, and other industries. The worldwide data revolution is causing most organisations to rely more on data in decision-making. Therefore, to address this issue, various technologies have been developed to adapt to the evolving situation. However, the biggest issue in using the data is the quality of the data itself.

Data quality is commonly used as the baseline to measure the quality of the dataset. However, the traditional data quality dimension which are accuracy, completeness, consistency and completeness are no longer suitable in Big Data environment. Moreover, most of data cleansing framework designed for Big Data still use data quality dimension as their baseline to confirm the accuracy of the cleansed dataset. Based on the above considerations, the basic question is how to evaluate the quality of the data to ensure that the data provided has high-level data veracity.

Having corrected and completed data quality rules are important in order to make sure all the errors in the dataset are cleansed. Traditionally, the process of obtaining data quality is done manually and it is defined by the expert (Abdellaoui et al., 2017). This process is costly, time-consuming, and requires domain experts' interaction (Juddoo, 2015). Therefore, an automatic technique to formulate the data quality rules for the cleansing process are needed.

Veracity is the primary concern, as it relates to the presence of imprecise and uncertain data that could affect the accuracy and relevancy of the data. However, the veracity of the data is still a wrangling issue, since most of the research done were only on social media dataset (Ba et al., 2016; Beretta, 2018; Paryani, 2017). Thus, a framework to measure the veracity of the structured data are needed.

1.5 Research Objectives

The main aim of this research is to develop a Big Data Veracity Enhancement Framework (VEF), where this framework will detect the data quality rules and measure the veracity of the dataset. The overall objectives of the research are as follows:

1. To identify and develop a taxonomy of data quality dimension specifically for Big Data

2. To propose a method that can automatically discover data quality rules to feed into the cleansing system

3. To adapt Veracity Assessment Framework, based on the proposed data quality taxonomy

4. To evaluate the proposed VEF in improving the veracity of the dataset

1.6 Expected Contribution

The expected contributions for this research are as follows:

1. Taxonomy of data quality criteria

In this research, the taxonomy criteria for the data quality will be proposed. The taxonomy not only provides a framework for understanding the quality dimensions, but it also includes how to compute the data quality measure. Compared to existing work, most data quality dimensions only focus on accuracy, completeness, consistency, and timeliness. These four criteria are indeed important, but no longer enough in measuring data quality in Big Data environment. The proposed taxonomy will be used as the baseline for the veracity assessment framework.

2. An automatic method to discover data quality rules

Obtaining data quality rules from the domain expert is usually expensive and time-consuming. Thus, an automatic method to discover the data quality rule will be proposed to deal with the data and help to improve the data quality. This method should be able to eliminate the need of domain expert in defining the rules and able to detect dirty data in the dataset.

3. An enhanced veracity assessment framework

This research will propose a new veracity assessment framework that can be applied within the organisation to formalise the analysis of the cleansed data source. This assessment will make it easier for the decision maker to evaluate the data source. This framework will be applied to the cleansed dataset, where it would check the data quality and map it to the real world.

1.7 Scope and Limitation

The data extracted from Big Data may be structured, semi-structured, and unstructured. However, this research focuses on structured data only, which include text and numerical. Currently, an automatic method to discover the data quality rule can be applied to the structured data only, as it will generate the data quality rules based on the table analysis.

Since the scope of study focuses on structured data only, thus the quality dimension proposed only covered for the structured data type. Besides, this research is focused on data quality rule extraction and data quality validation only. The cleansing process is done using third party resources.

1.8 Thesis Structure

This thesis is organised into six chapters. Chapter 1 until Chapter 6 present the introduction, literature review, research methodology, Veracity Enhancement Framework, results and discussion, and conclusion. The organisation of the chapters in this thesis are as follows:

Chapter 2 reviews previous research and discusses preliminary knowledge related to the thesis. It begins with the data quality challenges, issues, categories, and dimensions. Next, the systematic literature review on the taxonomy of data quality will be presented in brief. The following sections include explanation of the data cleansing process and the existing research done to formulate the data quality rule in the dataset. Besides, the following section will focus on the veracity challenges and the veracity evaluation framework available.

Chapter 3 presents the research methodology. It presents the method followed in carrying out this research. The first section describes the research methodology framework. Next, the development of automatic data quality rule discovery method is elaborated in detail, followed by the description of the veracity assessment framework. The following section explains the evaluation of the proposed VEF. Finally, discussion on data selection as well as the modelling approach for the experiment is included.

In **Chapter 4**, details on the proposed VEF is explained, followed by the taxonomy of data quality dimension. Next, the development of the automatic data quality rule discovery method is presented together with the result tested on the benchmark dataset. Finally, the phase for veracity assessment framework is explained. It provides a systematic explanation of the procedures conducted to evaluate the veracity of the dataset.

Chapter 5 describes the findings and discussion of the proposed work. First, the proposed data quality rule discovery method will be applied to the big dataset. Next, the evaluation of the proposed data quality rule discovery method is explained. Then, the analysis and results of the veracity assessment are revealed.

Finally, **Chapter 6** consolidates the conclusion from the study, followed by the main findings derived from the analysis of the results produced. It includes the limitations of this research, as well as improvements for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter is divided into five subtopics: data quality, taxonomy of data quality dimension, data cleansing, data quality rules, and veracity. First, the overview of data quality will be explained in brief. Section 2.3 explains the Systematic Literature Review (SLR) for the taxonomy of data quality dimension. Next, section 2.4 explains the data cleansing and the process in detail. Section 2.5 discusses on the data quality rules discovery, and section 2.6 focuses on veracity, challenges, and available assessment. Finally, section 2.7 will summarise the chapter.

2.2 Data Quality

‘Garbage in garbage out’ always happens in an organisation. Dirty data is collected, stored, processed, and analysed, but still provide the wrong decision, which lead to losses of a business organisation. Data quality is the primary concern faced by most organisations. This issue rises due to improper maintenance and will indirectly generate inconsistency in the database (Cohen et al., 2015). Data collected is valuable for the organisation to catapult their business ahead. It can provide various services for the organisation, given that they have high-quality data which will help them to achieve the top service in the organisation (Sidi et al., 2012).

In 2016, IBM reported that the company’s low-quality data cost $3.1 trillion (Redman, 2016). This is happening because the decision makers, managers, knowledge workers, data scientists, and others must accommodate it in their everyday work. The organisations are paying extra (Gao et al., 2016; Noraini et al., 2015) because they do not understand their data quality. On the other hand, high-quality data may contribute to wise decisions that help companies succeed. The organisations that decided to enhance their data quality have shown an improvement in sales of 15% to 20% (GlobalTranz, 2017).

Data quality can be defined as the fitness of data to meet the business requirement. It is based on the organisation itself, requirements, nature, and scale of the business. Without proper data quality management, even minor errors might cause revenue loss, process inefficiency, and failure to comply with the industry and government regulations (Saha & Srivastava, 2014). Thus, data quality and data cleansing are always linked together as ensuring data quality is critical and necessary before the sharpening of analytic focus can occur (Shneiderman & Plaisant, 2015). Data quality represents the standard to which the information is fit for usage in the required business processes. It can be defined, measured, and managed through several data quality metrics, such as completeness, timeliness, consistency, and accuracy (Al-Mughni et al., 2020; Taleb et al., 2015; Tatian et al., 2019).

Data quality is not about how cleansed the data is, but it has become a critical issue because it involves operational processes and is perceived as the greatest challenge in data management (Ehrlinger et al., 2019). According to Taleb et al. (2015), data quality in Big Data requires well-defined and precise measurement processes, which includes data quality management, monitoring, and control. These processes are to keep track of any changes that would improve or degrade data quality. Data quality can help in problem identification, process improvement, productivity increase, customisation support, intelligent decisions, and optimised solution. However, due to the enormous volume of generated data, the fast velocity of arriving data, and the wide variety of heterogeneous data, the quality of data is far from perfect (Gao et al., 2016).

Data quality has received greater attention in recent years due to the emergence of various V's criteria of Big Data (Juddoo et al., 2018; Saha & Srivastava, 2014). The characteristics of Big Data itself give the challenge towards the data quality assessment despite the volume is high. Data quality assessment is an iterative process, as the quality of the data must be checked every time the cleansing process is done until it reaches the defined standard. Besides, the assessment is differs based on each application context. The evaluation of the data quality is based on the intended use and suitability for that purpose (Tatian et al., 2019). Furthermore, data quality is a combination of data content and format (Jesmeen et al., 2018) where data content must contain accurate information and data must be collected and visualised in an appropriate form. Content and form are important components to minimise data errors, as they reflect the challenge of data cleansing.

In general, data quality can be expressed as cleansing up a set of dirty data. There are five essential components for data quality, which are dimensions, metrics, data profiling, data cleansing, and discovery of data quality rules (Juddoo, 2015). Figure 2.1 depicts the components of data quality.

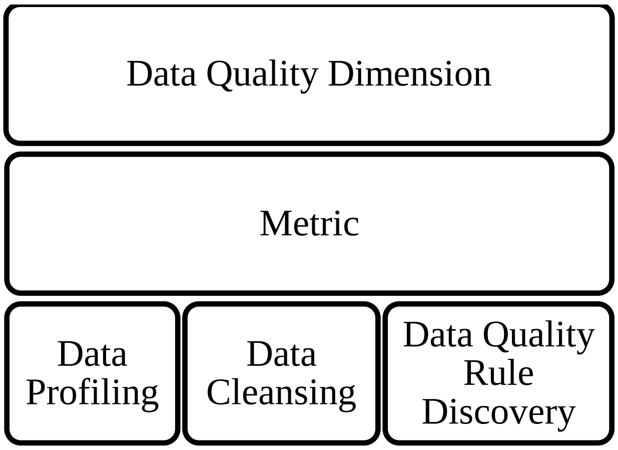

Figure 2.1 Data Quality Components (Juddoo, 2015)

A data quality dimension is a method of conveying the idea of data quality. The dimension is necessary to know what to measure in order to improve data quality (Heinrich et al., 2018; Sidi et al., 2012; Taleb et al., 2019). Metrics are needed to measure the dimensions in order to establish a particular notion of measurability and comparison of dimensions (Heinrich et al., 2018). It is a benchmark to distinguish between high-quality and low-quality data.

In order to improve the quality of data, several steps need to be conducted which are data profiling, data cleansing and data quality rule discovery. Data profiling is used to gain insight about the existing data quality (Tatian et al., 2019) and it focuses on the instance analysis of individual attributes. Meanwhile, data cleansing is important in the data quality components to solve the issues of data quality (Jesmeen et al., 2018) to make sure the data that will be used for analysis purposes is in high quality. The discovery of data quality rules will discover rules that would be used to detect dirty data and for cleansing purposes (Juddoo, 2015).

These five components of data quality are important to ensure that the data processed achieve high quality. However, challenges and issues that arise in data quality also need to be highlighted to facilitate data quality management. Therefore, the next section will discuss the challenges and issues that arise especially in Big Data.

2.2.1 Data Quality Challenge

Achieving and maintaining data quality is not easy, as many challenges affect data quality. Chen et al. (2018) have identified three characteristics associated with data quality: persistent, diverse, and severe. It is considered persistent because these issues have risen for more than 20 years. Besides, various types of the issue have risen, including duplication, inconsistency, inaccuracy, and incompleteness. Due to quality issues, precautions should be taken when performing associated data analysis. Data quality issues can also lead to business losses, as most organisations are depending on the data to make decisions and predictions.

According to Abdellaoui et al. (2017), there are three challenges related to data quality during data flow design: the rules defined ignore the concept of data quality rules, new errors may occur, and limited repair methods. Usually, the error detection is done based on the data quality rules defined, but if the rules are not designed properly, then the error detection process is difficult. Besides, a new error may arise after the cleansing process is done. Even though extensive studies have been conducted on repair methods, but they are inappropriate for flexible data structures such as graphs (Abdellaoui et al., 2017).

To evaluate the quality of the data, data security which allows easy and flexible access to all the data is required. However, it can make the process of managing data quality slower and more complex (Talha et al., 2019). Moreover, data cleansing is required to achieve high quality data, but the cleansing work has to be done manually which makes data quality assessment an expensive and time-consuming process (Hermans et al., 2017). Cai and Zhu (2015) added that common and approved data quality standards have not been developed, and research on the data quality of Big Data has just begun. The summary of data quality challenges is illustrated in Table 2.1.

Table 2.1 Data Quality Challenge

Authors	Data Quality Challenge	Description
Chen et al. (2018)	Persistent	Have been raised more than 20 years
	Diverse	Multiple data quality issues inside the dataset
	Severe	Issues that arise can affect the state of the existing dataset
Abdellaoui et al. (2017)	Ignore the concept of data quality rule	Data quality rule defined based on the detection of error in the dataset
	New errors may occur	Iterative data cleansing process may cause new error appeared
	Limited repair methods	The repair methods available do not suitable for flexible data structures
Talha et al. (2019)	Data security	Data security can make the process of managing data quality slower and more complex
Hermans et al. (2017)	Expensive and time-consuming	Cleansing work has to be done manually and the data are collected from various source
Cai and Zhu (2015)	No data quality standard	No unified and approved data quality standards are available

Guo et al. (2018) added that when the original data quality problem is solved, a new data quality issue arises. This is due to the variety and volume of Big Data, where it increases the difficulty of quality evaluation. Besides, other Big Data characteristics, such as variability, velocity, and volatility introduce challenges in managing and assessing Big Data quality. Five V's of Big Data need to be considered when dealing with and discussing the quality of Big Data. Gao et al., (2016) agreed that a massive volume of data, the fast velocity of arriving data, a variety of heterogeneous data is affecting the data quality of the dataset.

Data volume is enormous, and it is difficult to assess data quality within a reasonable period (Cai & Zhu, 2015; Talha et al., 2019). Besides, the data is collected from various resources which is making it is hard to achieve high data quality (Hermans et al., 2017). Inconsistent, incomplete, and noisy data have a detrimental effect on a variety of data types which need to be addressed when dealing with data quality (Abdallah, 2019; Hariri et al., 2019). The data comes in various types and formats, which makes it challenging to identify processes and filter the low quality of data. Cai and Zhu (2015) added that the variety of data sources increases the difficulty of data integration. Besides, the data consist of semi-structured or unstructured data, which makes a correlation between unstructured data a difficult task (Talha et al., 2019).

One of the major problems in data quality is rapid degeneration over time (Amini & Chang, 2017; Noraini et al., 2015). Experts say two percent of records in a customer file become obsolete in one month because customers die, divorce, marry, and move (Eckerson, 2002; Noraini et al., 2015). Timeliness is extremely important for applications where real-time responses are required. Data changes rapidly and the ‘timeliness’ of data is minimal, requiring higher requirements for processing technology (Cai & Zhu, 2015; Talha et al., 2019). Table 2.2 shows the data quality challenges in Big Data context.

Based on Table 2.2, the common criteria being discussed are volume, variety and velocity. Regardless of the challenge of data quality being associated with the V’s, these criteria are often left out when proposing data quality for Big Data. In this research, challenges related to volume and variety of the data will be taken into consideration when designing the VEF.

Table 2.2 Big Data Quality Challenge

Authors	Big Data Quality Challenge		
	Volume	Variety	Velocity
Talha et al. (2019)	√	√	√
Cai and Zhu (2015)	√	√	√
Amini and Chang (2017)			√
Guo et al. (2018)	√	√	√
Abdallah (2019)		√	
Taleb et al. (2018)		√	√
Hermans et al. (2017)		√	

2.2.2 Data Quality Issues

Data quality issues arise due to poor data management in the organisation. According to Taleb et al. (2018), data quality issues arise from a variety of causes or processes happening at different levels; data source, generation level, and application level. Thus, data quality must be controlled, particularly at the data source and generation level. In the data source, data quality is maintained through the data cleansing process. On the other hand, the generation level requires quality control when creating the dataset.

Outlier, missing value, repetitive value, redundant and irrelevant value, class imbalance, data shift problem, and high dimensionality of data were discovered as seven data quality issues associated with the software fault dataset (Rathore & Kumar, 2019). These issues occurred because of too much unnecessary information inside the dataset. Based on the study conducted by the authors, the major issue in data quality is high data dimensionality where data are filled with unnecessary features. Class imbalance problems and outliers are the second and third highly investigated data quality issues.

Serhani et al. (2016) have divided data quality issues into three groups: (1) error correction, (2) data conversion, and (3) data integration from multiple sources. Besides, the authors reported that volume, speed, and schema-less data are also related to the data quality issue. Meanwhile, Abdellaoui et al. (2017) listed data quality issues including (1) defining a set of quality rules, (2) implementing them on the data flow pipeline to detect violations, and (3) delivering accurate repairs for the detected violations.

Pei Li et al., (2019) classified data quality issues into four categories: redundancy, canonicalisation, strong logic, and weak logic error. In the process of data merging or multi-source data fusion, canonicalisation errors signify inconsistencies caused by different recording methods for the same attributes. The strong logic error refers to the truth-value error. A weak logic error is similar to strong logic, but there is a weak logic correlation among attributes.

A review conducted by Corrales et al. (2018) has shown that data quality issues for regression tasks, such as missing values, outliers, and redundancy have received greater attention from the research community. Meanwhile, since noise is characterised as a general consequence of data measurement errors, it has received less attention. Data quality issues cannot be solved in one cycle of data cleansing only. It requires numerous iterations to verify that all the errors are being corrected, as some errors are only visible after the transformation (Guo et al., 2018). Due to the abnormality of repaired data, the correctness of the data may not be achievable at the end of data cleansing and requires another cycle of the cleansing process (Cheng et al., 2018).

Various issues that arise in the dataset are due to the evolution of Big Data itself. Data is collected actively from numerous sources caused diverse issues to arise in the dataset which caused it to be dirty. Nevertheless, all of these issues can be grouped into single-source and multi-source problems (Rahman et al., 2019). Missing values and deduplication are considered as a single-source problem where it can happen in the schema and instance level. The important part in data cleansing is the instance level problem, as it reflects the errors and inconsistencies happening in the actual data contents, but the errors are not reflected in the schema level (Rahm & Do, 2000). Figure 2.2 shows the classification for the data quality issue in the data source.

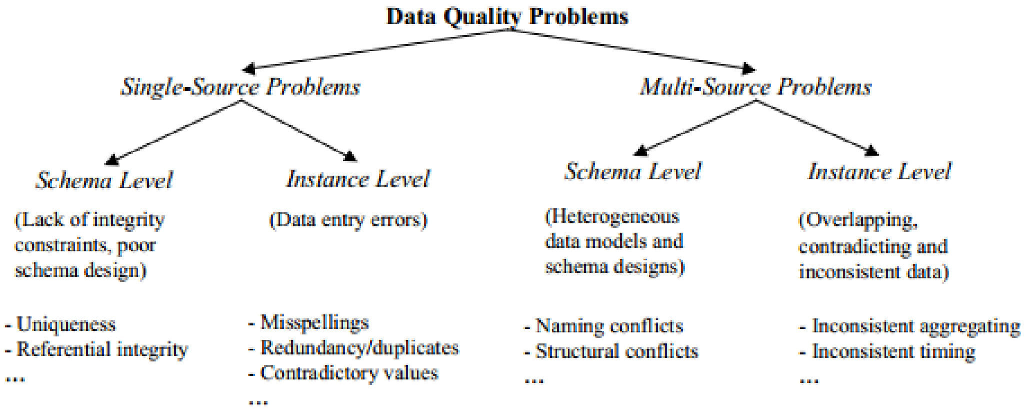

Figure 2.2 Classification of Data Quality Problems