# DATA TRANSFORMATION MODEL FOR ADDRESSING INCOMPLETE AND INCONSISTENT QUALITY ISSUES OF BIG DATA

**ONYEABOR GRACE AMINA**



**DOCTOR OF PHILOSOPHY**

**UNIVERSITI UTARA MALAYSIA**

**2024**

# DATA TRANSFORMATION MODEL FOR ADDRESSING INCOMPLETE AND INCONSISTENT QUALITY ISSUES OF BIG DATA

**ONYEABOR GRACE AMINA**



**A thesis submitted to the Awang Had Salleh Graduate School of Arts and Sciences in fulfilment of the requirements for the Doctor of Philosophy**
**Universiti Utara Malaysia**

Awang Had Salleh
Graduate School
of Arts And Sciences

Universiti Utara Malaysia

## PERAKUAN KERJA TESIS / DISERTASI
*(Certification of thesis / dissertation)*

Kami, yang bertandatangan, memperakukan bahawa
*(We, the undersigned, certify that)*

GRACE AMINA ONYEABOR

calon untuk Ijazah
*(candidate for the degree of)* PhD

telah mengemukakan tesis / disertasi yang bertajuk:
*(has presented his/her thesis / dissertation of the following title):*

"DATA TRANSFORMATION MODEL FOR ADDRESSING INCOMPLETE AND INCONSISTENT QUALITY ISSUES OF BIG DATA"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi.
*(as it appears on the title page and front cover of the thesis / dissertation).*



Bahawa tesis/disertasi tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : ***22 Mac 2023.***
*That the said thesis/dissertation is acceptable in form and content and displays a satisfactory knowledge of the field of study as demonstrated by the candidate through an oral examination held on:*
***22 March 2023.***

| | | |
|---|---|---|
| Pengerusi Viva: *(Chairman for VIVA)* | **Assoc. Prof. Dr. Fauziah Baharom** | Tandatangan *(Signature)* |
| Pemeriksa Luar: *(External Examiner)* | **Assoc. Prof. Dr. Umi Kalsom Yusof** | Tandatangan *(Signature)* |
| Pemeriksa Dalam: *(Internal Examiner)* | **Assoc. Prof. Dr. Azizah Hj. Ahmad** | Tandatangan *(Signature)* |
| Nama Penyelia/Penyelia-penyelia: *(Name of Supervisor/Supervisors)* | **Assoc. Prof. Ts. Dr. Azman Ta'a** | Tandatangan *(Signature)* |

Tarikh:
*(Date)* ***22 March 2023***

# PERMISSION TO USE

In presenting this thesis in fulfillment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for copying this thesis in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication, or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use that may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences

UUM College of Arts and Sciences

Universiti Utara Malaysia

06010 UUM Sintok

# ABSTRACT

Data Quality (DQ) assessment remains one of the major challenges for Big Data (BD) due to the complexity of handling large volumes of data. Traditional data transformation methods such as Extract-Transform-Load (ETL) use data sources from a diverse range of devices and locations resulting in incomplete and inconsistent DQ that may lead to wrong insights and decisions. Therefore, DQ is vital for the effective operation and management of BD. Recognizing many DQ features from its definition to the various dimensions is essential for equipping techniques and procedures to improve DQ. This research focuses on two aspects of DQ: completeness, and consistency. Firstly, an enhanced data transformation model (2CsDQT) is proposed to assess and improve big data quality. A new algorithm using ontology and clustering methods is used to identify and correct incomplete and inconsistent data, which resolves the availability and comprehensiveness of data, similarity between data items, and missing specific attributes of data. Secondly, using a clustering technique to analyse DQ, and improve employing results from the 2CsDQT model. The complete and consistent data are put into clusters, and the designed algorithm predicts the position of any incomplete and inconsistent data, based on its value to be added to the specific cluster. The study was evaluated using the developed model and benchmarked with existing data transformation techniques in the literature. This research shows that the 2CsDQT model successfully improves BD quality and outperforms previously proposed methods. Data completeness and consistency results outperform related articles and benchmark studies in the literature on the datasets of two different test cases. The theoretical contribution of this research work is to provide insight into the importance of DQ issues in BD and the effect of inconsistency and incompleteness on BD application. The practical contribution is the provision of enhanced data transformation models for DQ leading to better data analysis and strategic planning.



**Keywords:** Data quality, Data transformation, Data completeness, Data consistency, Big data.

# ABSTRAK

Penilaian Kualiti Data (KD) kekal sebagai salah satu cabaran utama dalam Data Raya (DR) disebabkan oleh kerumitan pengendalian kuantiti data yang besar. Kaedah transformasi data secara tradisional seperti Ekstrak-Tukar-Muat (ETM) menggunakan sumber data daripada pelbagai peranti dan lokasi mengakibatkan KD tidak lengkap dan tidak konsisten yang boleh membawa kepada penilaian dan keputusan yang salah. Oleh itu, KD adalah penting untuk operasi dan pengurusan DR yang berkesan. Menyedari banyak ciri definisi KD membawa kepada pelbagai dimensi, adalah penting untuk melengkapkan teknik dan prosedur untuk menambah baik KD. Penyelidikan ini memfokuskan kepada dua aspek KD: kesempurnaan, dan konsistensi. Pertama, model transformasi data yang dipertingkatkan (2CsDQT) dicadangkan untuk menilai dan meningkatkan kualiti data raya. Algoritma baharu menggunakan kaedah ontologi dan pengelompokan digunakan untuk mengenal pasti dan membetulkan data yang tidak lengkap dan tidak konsisten, yang boleh menyelesaikan ketersediaan dan kelengkapan data, persamaan antara item data dan atribut data tertentu yang hilang. Kedua, menggunakan teknik pengelompokan untuk menganalisis KD, dan menambah baik hasil keputusan daripada model 2CsDQT. Data yang lengkap dan konsisten dimasukkan ke dalam kelompok, dan algoritma yang direka bentuk meramalkan kedudukan mana-mana data yang tidak lengkap dan tidak konsisten, berdasarkan nilai yang ditambah kepada kelompok tertentu. Kajian ini dinilai menggunakan model yang dibangunkan dan ditanda aras dengan menggunakan teknik transformasi data yang sedia ada dalam literatur. Penyelidikan ini menunjukkan bahawa model 2CsDQT berjaya meningkatkan kualiti DR dan mengatasi kaedah yang dicadangkan sebelum ini. Hasil kesempurnaan dan konsistensi data mengatasi penanda aras dalam artikel dan literatur yang dibuat ke atas dua set data dalam kes ujian yang berbeza. Sumbangan teori kerja penyelidikan ini adalah untuk memberi gambaran tentang kepentingan isu KD dalam DR dan kesan data yang tidak konsisten, dan data yang tidak lengkap dalam aplikasi DR. Sumbangan praktikal ialah penyediaan model transformasi data yang dipertingkatkan untuk KD yang membawa kepada analisis data dan perancangan strategik yang lebih baik.

**Kata Kunci:** Kualiti data, Transformasi data, Kesempurnaa data, Konsistensi data,Data raya.

# ACKNOWLEDGEMENT

I am grateful to the Almighty God, the Father of our Lord Jesus Christ, for His mercies, love, and provision in helping me complete my PhD. To God be the praise and glory. I am grateful to my supervisor, Associate Professor Dr. Azman Ta'a, who provided invaluable mentoring, advice, and feedback. He's been invaluable in offering painstaking guidance and thorough direction throughout the study process. I pray to God to reward his efforts. I also appreciate the contributions of my proposal defence panel, which helped to improve the study.

I am grateful for my institution's assistance, the Federal University of Technology, Minna that provided the platform for the course and made it possible for me to complete it. I acknowledge the Federal Government of Nigeria's enormous assistance for the PhD path through the Tertiary Education Trust Fund. These organizations have helped me tremendously.

I appreciate my family's sacrifices in bearing the brunt of my extended absence from home. I appreciate my lovely husband, Tobias, and our children, Alice and Arnold, for their patience, prayers, support, and inspiration for the program's success. God will also help you achieve your goals in life. Another family I appreciate so much in this journey is members of Data Science Research Lab (DSRL), School of Computing (SOC), UUM especially Professor Dr. Ku Ruhana Ku Mahamud (who was like a mother to us all) and Professor Dr. Yuhanis Yusof for the beautiful times shared in research, presentations, partying and encouraging one another during the fight. I will never forget any of you.

I appreciate the prayers and support of my pastors, Pastor Leonard Lim and members of Full Gospel Church, Changloon, Pastor (Dr) Ikenna and Pastor Peace and members of the Redeemed Christian Church of God, Changloon, Rev. Jacob Amadu, Pastor Lanre Faboyede and members of the Apostolic Faith Church in Nigeria for their prayers and unwavering support throughout the voyage. My gratitude also extends to all of my colleagues at Federal University of Technology, Minna especially those in the defunct Information and Media Technology department and Information Technology department. Finally, I want to express my gratitude to all of my friends like Nurul, Thillaga, Dr. Sumbo, Dr. Aisha, Dr. Kay, Dr. Taiye, Dr. Abdullah, Dr Sunny, Dr. Sidra (etc) too numerous to mention who have supported me during the programme. May God continue to bless them all.

## Contents

**List of Figures**

**List of Tables**

## List of Abbreviations

2CsDQTM: Completeness and Consistency Data Quality Transformation Model

3Vs: 3 Volume, Velocity, Variety

BD: Big Data

BI: Business Intelligence

DI: Data Integration

DQ: Data Quality

DQD: Data Quality Dimensions

DQO: Data Quality Ontology

DL: Data Lake

DW: Data Warehouse

DWQ: Data Warehouse Quality

ETL: Extract Transform Load

ELT: Extract Load Transform

IID: Incomplete and Inconsistent Data

IT: Information Technology

IoT: Internet of Things

KBV: Knowledge-Based View

OBDM: Ontology-Based Data Management

OBDT: Ontology-Based Data Transformation

OIPV: Organizational Information Processing View

OWL-DL: Web Ontology Language-Description Logic

TDQM: Total Data Quality Management

TIQM: Total Data Quality Management

# CHAPTER ONE: INTRODUCTION

This chapter underscores the background and motivation of this research study. The chapter also states the research problems and the research gaps, including the research questions and objectives. The research scope and the significance and contribution of the study are also discussed. This chapter ends with an overview of the thesis organization and summary.

## 1.1 Background

Today businesses and organisations generate vast amounts of data. At the same time, we receive and collect huge volumes of data from different resources and store them, which leads us into the Big Data (BD) era. BD is a name for massive varying format datasets produced at rapid speed and the usage of standard database management systems makes it utterly difficult to handle (Kanchi et al., 2015). Several businesses are constantly accumulating massive datasets that record consumer interactions, product sales, and other types of information due to online marketing initiatives. For example, Facebook shared daily 1.7 millions contents, and gathers 300 petabytes of data (Domo Inc., 2022).

Consequently, there is an urgent need to guarantee BD quality. Data Quality (DQ) allows data to meet the requirements of a user and it accurately represent the real-world entity in structure and meaning. Moreover, it's suitable for intentional usage when applied in tasks, planning and decision making in organizations or by users. In other words, data must not lose its value as it goes from one process of transformation to the other; instead the value should be improved to enable valuable operations,

planning and decision making. The benefits of high quality data by Côrte-Real et al. (2020) are presented in Table 1.1. The key factors for the growth of any business are efficiency, performance, decision making, cost savings and organizational reputation, which also positively correlate to better DQ.

*Table 1.1: Benefits of high quality data (Côrte-Real et al., 2020)*

| **Benefits** | **Percentage** |
|---|---|
| Increased efficiency | 58% |
| Mobile website performance | 55% |
| Improved decision-making | 51% |
| Cost savings | 47% |
| Protection of organization reputation | 46% |

Data in BD storage systems is dispersed which in result allows for distributed computation and huge datasets can be processed with no storage constraints. Nevertheless, this may cause DQ concerns like consistency across several data centres (Fürber & Hepp, 2010a). Due to the diversity of the sources, the data obtained might have varying noise levels, redundancies and consistency etc. Moreover, the transfer and storage of raw data would cost as well. Russom (2013) suggested that consistency is the primary dimension used in measuring DQ in the BD. Degraded DQ leads to incorrect assumptions about these processes which ultimately results in the waste of all sorts of resources and assets (Ali & Warraich, 2020). Heterogeneous data sources and various data formats can lead to DQ issues like inconsistency and incomplete data. These issues can be summarized as data conflicts structurally and semantically (Bala et al., 2017; Franklin et al., 2005; Kettouch, 2017). Furthermore, no information about the data origin may lead to receive corrupt data from the same source which was used previously.

In order to encourage a broad range of company's decision making, reports, analysis and planning, a company's of Data Warehouse (DW) consolidates data from many

sources. DW procedures are susceptible to DQ for the mentioned tasks. They are dependent on data completeness, consistency and accuracy, which are defined as DQ dimensions. Therefore, data transformation is an important factor to increase accuracy of the overall system by verifying the origin of the data.

On the consumer side, some techniques and applications of data analysis might have severe DQ requirements. Data transformation methods that improve the DQ should thus be employed in BD systems (Taleb & Serhani, 2017). Data transformation is a process in DW framework called Extract Transform Load (ETL) used to transform and integrate data from various sources. It makes the seamless integration of many data sources by using a particular tool. It works as an integrator, extracts data from multiple sources to perfectly adapt them according to the principles of business rules and loads it into cohesive databases or DWs (Tesfagiorgish & JunYi, 2015). These processes always lead to the creation of data inconsistency and incompleteness (Elouataoui et al., 2022).



Inconsistent data contain differences in codes or names (synonyms and nicknames, prefix and suffix variations, abbreviations, truncations or initials). There might be discrepancies between duplicate records; different data sources might carry different column names for the same values due to non-uniform naming principles or data codes, functional dependency and/or referential integrity violations. This arises from various data sources. Incomplete data lack of attribute values or is deficient in specific attributes of interest or comprises only of cumulative data. They come from non-available data values when accumulated because of various reasons such as data gathering, hardware or software problems. They distinguish between two types of data

conflict: (i) attribute value that uncertainty produced by missing data, and (ii) attribute value contradictions that generated by differing attribute values.

### 1.2 Research Motivation

One of the largest misunderstandings enterprises have about their existing data is that it needs to be cleaner, complete, and consistent. They consequently need to pay more attention to the potentially catastrophic impact that inadequate data is already having within their corporate systems. The fact is, as more information enters the company from diverse sources, the risk of mistakes occurring naturally grows. Poor data loses enterprises an average of $15 million each year (Elliott et al., 2021). As organizations grow more operationally dependent on technology and algorithms, high-quality data will become increasingly more crucial. From the researcher's perspective, duplicated data, which is because of inconsistent data, for example, is one of the primary sources of incorrect insights. A corporation might believe it has 100 active users. Still, owing to duplicate data that happens throughout numerous data sources, it's entirely feasible that the company only has 63 active users while the remaining 37 are duplicates as a consequence of discrepancies. Consider this observation at an exponentially huge level with millions of rows of data, and one is very likely to make false inferences from the data.

Also, companies lose income by allowing their sales staff to squander time on faulty leads created by low-quality data. They risk losing contact with their consumers through incomplete contact information and customer details. Marketing departments undertake costly, ineffective marketing initiatives targeting the incorrect demographics owing to insufficient and inconsistent data. Companies are taken by

surprise by their revenue estimates, which is wrong if obtained with insufficient financial data sets. For example, many organizations allow clients 30 days to complete their payment after being invoiced. Nonetheless, incomplete client contact information slow down the accounting process and push back payments.

The foregoing is a small-scale example, but bad DQ also lead to more major difficulties which leads to financially disastrous judgments and becomes more typical with firms that handle with BD sets. Many social media firms, for example, rely on data about their customers to effectively sell to them. Marketing studies demonstrate that personalisation is crucial to reaching new customers for your product or service. But incomplete and inconsistent data delivers foggy or partial perspective of client demographic and bog down marketing initiatives. Facebook, for example, derives an estimated 97.6% of its revenue via ads. The social media powerhouse had bragged to its customers that it reached an audience of 41 million Americans between the ages of 18-24, a lucrative market for many enterprises. However, the US Census Bureau stated that the overall population of this youthful cohort is just 31 million. As a result, Facebook had to face with a class-action lawsuit from its advertisers arguing that bad data allowed the social media corporation to inflate its effect (Ghasemaghaei, 2020).

Furthermore, more subtle, and harder to measure, but nonetheless destructive, are the wasted opportunities that emerge from inconsistent data. Identifying, making judgments and acting on the wrong leads, being unable to pre-qualify leads, and the inability to recognize potential prospects all hinder a developing organization. For example, sales teams spend significant time by reaching out to the same lead twice owing to inconsistent data that is duplicated. This makes the firm appear less

professional to the potential consumer. It also implies the sales team is spending time and money on the wrong leads when time is of the importance when prospecting. Poor DQ allows organizations to wrongly assess product feedback, hampering their capacity to proactively enhance their services and solve concerns. It also causes marketing efforts to be exposed to the wrong sort of customers. Personalized marketing for client demography may not be as successful with faulty data. Poor data influences social media data analysis and result in less-than-stellar results for the marketing initiatives. Limited or poor-quality data can stop a firm in its tracks, slowing growth by failing to identify and swiftly act upon future prospects .

In addition, improper decisions and wasted time caused by poor-quality data affects an organisation's reputation. Often, A damaged reputation can be harder to fix and have a more lasting impact than financial loss. Inaccurate product and consumer information undermines business integrity and reliability. Many of these inaccuracies are linked to poor-quality data. The failure to accurately organize, identify, and secure data contributed to challenges with legal standards concerning data handling. Companies must secure customer data. Companies that break data compliance regulations lose their reputation with local authorities and customers, putting them under heightened scrutiny. For example, many organizations in the finance sector must comply with anti-money laundering legislation, such as the Know Your Customer law. This law requires financial organizations to obtain client data before fulfilling their transactions. This is aimed to verify that they are not fraudsters, money launderers, terrorists, criminals, or hostile foreign governments. Inconsistent, incomplete, or unfiled client data easily leads to incompliance in this circumstance.

Data has produced a very significant resource for social and business life. Humans use data daily for decision making and transactional processes (Fürber & Hepp, 2010b). Besides, the fact that information is derived from data, it is assumed that high-quality data can only provide high-quality information. Hence, if the data consumed is incorrect, the information derived will be incorrect as well. As a result wrong processes or decisions that depends on the erroneous data may fail. The presence of inconsistent and incomplete data in a dataset might lead to incorrect conclusions or assumptions. This is also evident in Figure 1.1, showing primary DQ issues faced by various organizations as of 2020 (Roger & Steve, 2020).

It can been seen that DQ is degraded by 58% due to the inconsistency in the data which is the highest percentage in all the other DQ issues. Bloomberg Trading Solutions (undated) notes that the entire cost of inconsistencies does not quantify. The costs of accounting errors, numerous feeds costs and all work necessary to deal with inconsistencies can be included.

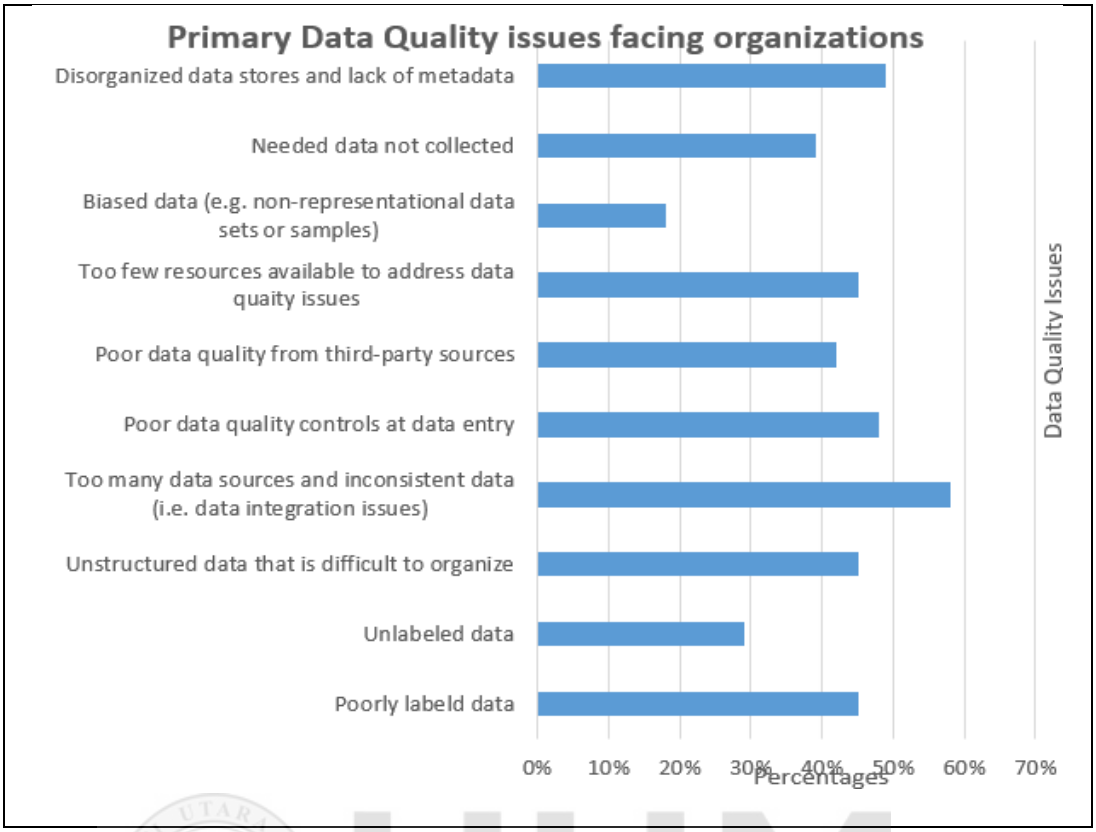



*Figure 1.1: Data Quality Issues*

The incomplete data is presented in first two issues of the DQ in Figure 1.1, i.e, "disorganized data" (i.e., lack of metadata) and "needed data not collected". Naumann and Herschel (2010) identified data incompleteness as a challenge in a unified dataset. A high number of records found in noisy data can nonetheless be misanalyzed through an analyzer of data. For instance, a data analyst might assign incomplete data as complete if the data incompleteness has not been recognised. The analyser must recognise the incomplete data inside a retrieved record to correctly analyse the data. But it might be challenging to physically recognise them in BD because a record set in a table style does not allow the incomplete data to be easily viewed even in traditional data.

## 1.3 Problem Statement

Inconsistent and incomplete data (IID) is a threat to both BD and conventional databases (Chika, 2015). Traditional DBs are queried to retrive the records in tabular form (i.e., each row contains objects and columns). BD is more vulnerable to IID because of significant data transformation in a schema-less and complex BD environment. Therefore, the problem of IID is crucial for the success of BD and also is of interest among the research community. Unlike the Data Warehouse (DW) approach, BD utilizes a data lake approach. This leads to a more scalable way of storing and retrieving data.

However, the absence of a predefined schema on data may lead to IID issues more than often because of data being integrated and transformed from many diverse sources (Impact et al., 2022). In such an integration, various values from different sources might be linked with an attribute of an object. These attributes may not only have inconsistent values, but may not even contain the needed values (i.e., incompleteness). Analysis on significantly a large dataset can be diluted by the existence of IID in that dataset leading to lead to incorrect assumptions and conclusions while improving DQ is vital for policy and planning. Data availability does not automatically translate into availability of the quality data needed. Sound decisions are based on sound data; therefore, it is essential to ensure that the data are of good quality (Daneshkohan et al., 2022). Otherwise, dependence consequently will be placed on the data analyst to recognize and deal with the IID in the dataset which may further worsen the situation.

The conventional ETL models' standard means of overseeing DQ are generally meant to discover and correct errors in data from known sources based on a restricted set of

rules. Instead, the number of rules may be enormous in a BD environment, and correcting flaws would not be practical or suitable. Therefore, it is vital to rethink how to supervise IID and integrate them into the BD framework in these cases, there hasn't been much study done on DQ in BD beyond data cleansing to the best of our knowledge. In many respects, the issue of value extraction from BD resembles the long-standing distillation of business intelligence (BI) in transactional data (Mikalef et al., 2020) . The process for extracting data from different sources, transforming it into a form which is fit for analysis and loading it into DW for further analysis which is a process of standard ETL for DW is the heart of this problem. This research focuses on the two crucial dimensions of DQ identified in existing literature and discussed above (i.e., completeness and consistency of data).

Poor-quality data affect different levels of various sectors and systems in different ways. For example, health-care providers at the facility level, which is the patient treatment, can be compromised if the information about the patient is missing or inconsistent. For program managers, poor-quality data can lead to wrong judgments that can damage the overall running of the program and, ultimately, the population's health. At the planning level, poor-quality data can weaken evidence of progress towards health sector objectives and may hamper annual planning processes by delivering deceptive findings. Furthermore, when selecting investments in the health sector, poor-quality data can lead to inefficient targeting of resources (World Health Organization, 2021).

An informed decision-making in Primary Health Care (PHC) is the foundation of universal health coverage and depends on high-quality data. This issue plays a vital

part in the healthcare industry, and sometimes, decision-making based on these data might imply a difference between life and death (Daneshkohan et al., 2022). In the United States, an average of 71,000 deaths are recorded annually owing to medical errors connected to poor data quality of patient treatment. In general, informed judgments, efficient patient' treatment, and equity in resource allocation in PHC, as a basis of health systems, depend on high-quality data to operate successfully (Gonete et al., 2021).

Current applications in data management, such as corporate data management, requires integrating existing data sources and a standard user interface for access from diverse sources. Data integration was thoroughly explored in the previous ten years using integrity constraints with exciting results (Barbosa et al., 2018). However, there are often inconsistencies between different sources that are always difficult to recognise and assess (Quadir et al., 2020). In recent years, there has been a rising interest in querying instance data using description logic ontologies, often known as ontology-based data access utilising database-style queries (Alizadeh et al., 2019). An ontology provides a way for users to consistently and accurately employ uniform terminology about the same entity in certain domains and enables automated auditing of DQ for requests. Automated auditing of DQ for analysis can be utilized as DQ assurance in the data entry sites where data that do not fit the ontology automatically cannot get accepted by the system (Liu et al., 2023).

Most of the previous DQ researches in BD have conducted research using ontology to only access DQ within BD (Jarwar et al., 2019; Wahyudi et al., 2018). However, there is not much work done in attempt to improve DQ in BD situations, especially with

inconsistency and incomplete data problems. Also, previous researchers did not consider integrative view but as a partial approach of just the assessment of the DQ issues without improving it. The ontology should be further expanded and formalised to support more use-cases, also, the DQ can be extended with additional rules. Ontologies should scale on both complexity and size (Engineering, 2022). Meanwhile detecting errors in BD can be achieved using clustering (Juddoo & George, 2020).

### 1.4 Research Questions

Due to DQ's problems in line with quality issues in BD, as explained in the previous sections, a model for transformation process of DQ in BD environment need to be enhanced. As this research focused on the dimensions completeness and consistency of DQ, these bring in the following research questions:

i) How does incompleteness and inconsistency affect the quality of BD?

ii) How to develop a data transformation model with these DQ dimensions?

iii) How to produce quality data from the proposed data transformation model?

### 1.5 Research Objectives

Based on the aforementioned problem statement and research questions, the main objectives of this research are as follows:

i) To measure the effect of incompleteness and inconsistency of quality data in BD.

ii) To develop the data transformation model with the completeness and consistency data quality dimensions.

iii) To evaluate the effectiveness of the proposed model in order to produce quality data of BD

## 1.6 Scope

This research explores the problem of DQ that occurred in the data transformation of BD. The focus is on transformation model development and evaluation which is applicable to various domains. In order to design the transformation model, this research needs to understand all the component functions and the requirements of the model. It covered textual BD datasets, but audio, video and streaming datasets were not covered. Furthermore, the study used semantic technologies in the form of ontologies and clustering algorithm in Machine Learning (ML) to identify and assess inconsistency and incompleteness of data to enhance DQ in BD. This study does not cover all the DQ dimensions, and is only limited to consistency and completeness dimensions.

## 1.7 Significance of the Study

The significance of this study is focused in two areas: theoretical contribution and practical contribution. It is crucial to have implications for researchers and practitioners interested in increasing DQ by incorporating the contributions made in this study. The model proposed in this study is domain-independent. In other words, it can be applied in other domains aside from the domains used in this research. The major theoretical contribution is extending DW framework by improving data transformation process using ontology approach and clustering method in BD environment. Some major steps to follow as a documentation for any organization to implement this study model is as follows:

1.2 Purpose of the Documentation

**Data Quality Model Overview**

2.1 Gathering Big Dataset

2.2 Data Preprocessing

2.3 Developing Data Quality Ontology

2.4 Data Improvement through Clustering

2.5 Data Quality Monitoring Tool with Alerts and Notifications

**Implementation Steps**

3.1 Gathering Big Dataset

3.2 Data Preprocessing

3.3 Developing Data Quality Ontology

3.4 Data Improvement through Clustering

3.5 Data Quality Monitoring Tool



## 1.8 Thesis Organization

**i) Chapter 1 – Introduction**

The first chapter gives an overview of the thesis by explaining the research background and the motivation for carrying out the research, the problem was also stated, and the research gap was not left out. Furthermore, the research questions, objectives, research scope, and research contributions were highlighted in this chapter.

**ii) Chapter 2 – Literature Review**

Chapter two discusses related literature to DQ and BD. DQ dimensions relevant for BD, that is, completeness and consistency, are also reviewed with the existing data transformation models and their effects. DQ with ontology and clustering are discussed extensively.

**iii) Chapter 3 – Research Methodology**

The methods utilised to fulfil the transformation model's aims for resolving quality concerns in BD are addressed in this chapter. The chapter explains the different phases the research underwent in order to achieve its purpose. These are:

a) Identification of the DQ dimensions that are mostly used and how the violoations of them can affect the results of data transformation of BD and identification of the inconsistency and incompleteness effect on data transformation in BD.
b) Development of the Data Transformation Model with the DQ Dimensions.
c) Evaluation of DQ of data processed using the Data Transformation process model.

**iv) Chapter 4 – Completeness and Consistency Data Quality Transformation Model**

Chapter four discussed the steps to develop the data transformation model for DQ based on the literature findings. This will enable the identification of DQ violations. That is incompleteness and inconsistencies within the integrated dataset.

**v) Chapter 5 – Evaluation of Completeness and Consistency Data Quality** Transformation Model. This chapter presents the evaluation process of the transformation model. The efficiency of the 2CsDQT model is evaluated using truth dataset to validate the model and compared with benchmark paper articles based on the papers dataset. The chapter also compares the proposed study and the benchmark studies to assess the effectiveness of the proposed model.

**vi) Chapter 6 – Conclusions and Recommendations**

The research objectives and findings are reviewed in this chapter. The summary of this research work was given toward the contributions of the study. The core contributions are allied with general contributions, as presented in chapter one. Furthermore, the limitations and recommendations for future studies are given and finally, a concluding remark.

## 1.9 Summary

The background and problem statement were presented in this chapter leading to the sustaining research motivation. The identified gap from the previous studies propelled the requirement for the research questions and the corresponding objectives as sta. This study aims to develop and evaluate a model for improving DQ in the BD environment's through the transformation process. BD is more vulnerable to inconsistent and incomplete data because BD is schema-less, and meaningful data transformation in a schema-less and complex BD world is a big open research challenges. The reliable data needs to be added to the transformation model, which give rise to the importance of data transformation and integration in BD. The next chapter discusses the literature related to DQ in BD in detail.

# CHAPTER TWO: LITERATURE REVIEW

This chapter explain about literature of the research. The aim of this chapter is to expound on different concepts highlighted in the previous chapter and critically review literatures related to this research study. Further into the chapter also comes selected related studies reflecting the author's achievements, strenght and limitations which helps in revealing the gap of where this study steps in.

## 2.1 Introduction

Review of essential literature remains an integral ingredient of sound research. Given this, previous related studies are reviewed to consolidate understanding of the aim of this study. The review is grouped under several subheadings of related topics such as DQ, DQ dimensions, and DQ dimension metrics. A review on BD followed the above; its characteristics; DQ in BD; DQ dimensions of BD emphasizing on completeness and consistency DQ dimensions for the DQ of BD. In addition, antecedents of DQ, transformation and effects are equally reviewed. It is followed by a review on data DW, and extract transformation load-process. Also, data transformation, DQ defects with the ETL process, and DQ approaches in the ETL process are equally reviewed.

Similarly, a review is conducted on process-oriented approaches, data-oriented approaches; syntactic oriented approaches; sematic oriented approaches and context-oriented approaches. Accordingly, issues such as common DQ issues in DW; ontology-based data management, its paradigm; and uses for DQ are discussed. Second to the last include reviews on ontology-based data transformation; DQ with clustering algorithm; DQ methodologies; total information quality management; and

total data quality management. Finally, DW quality; summary of the three chosen methodologies; DQ tools; BD quality rules; review of related studies; benchmark study and review summary.

## 2.2 Data Quality

The importance of data for businesses has been increasing for decades, and high-quality data is commonly regarded as an antecedent of organizational performance and competitive advantage (Altendeitering & Guggenberger, 2021). Thus, "Data is the new oil" became a common saying, which refers to how data has become an essential operating and strategic resource (Cappa et al., 2021). In detail, data is a crucial driver of practically any innovation in our digital business world. Either directly for building new data-driven products or services e.g Altendeitering & Guggenberger (2021) or indirectly through business intelligence, e.g., for process optimization or decision support (Akter et al., 2016). Even the development of new physical items needs data to understand client needs and derive design requirements (Burley et al., 2019). Moreover, the development of digital platforms and their associated ecosystems rely on the use and interchange of data (Altendeitering & Guggenberger, 2021). To achieve these business objectives, data needs a specified “fitness for use”, which relates to the data's quality (Wang & Strong, 1996).

Before considering BD quality, it is necessary to define DQ in general. The majority of previous DQ research has come from the database management field. According to research published in Oliveira and Belo (2017), the concept of DQ has proved challenging to define. Fürber (2015) defined DQ as "fitness in use" or "meeting the needs of users". As Nikiforova, (2020) states, “You can’t do anything important in

your company without high-quality data". DQ is context-specific and hence determined by the user. The value acquired from data processing relies on DQ, which remains one of the major concerns addressed by modern data management (Legner et al., 2020). Following the Data Management Association (DAMA), DQ management is the "application of Total Quality Management (TQM) concepts and practices to improve data and information quality, including setting DQ policies and guidelines, DQ measurement and validation, DQ analysis, data cleansing and correction, DQ process improvement, and DQ education" (Altendeitering & Guggenberger, 2021). Thus, at least two basic approaches can boost DQ (Byabazaire et al., 2020). First, optimizing data acquisition, which can be seen as the optimal solution for sustainable quality optimization. Second, quantitative and qualitative analysis and curation of existing data sets (Altendeitering & Guggenberger, 2021). The latter is a labor-intensive and complex procedure that can be supported by data transformation (Liu et al., 2023; Altendeitering & Guggenberger, 2021).



Several academics have described DQ in the literature as data that meets users' expectations or data appropriate for use by data consumers (Sebastian-Coleman, 2012). The International Organization for Standardization/International Electrotechnical Commission 25012 standard (Hassenstein & Vanella, 2022) defines DQ as the degree to which a collection of data characteristics (also called DQ dimensions) meets requirements. Completeness, accuracy, timeliness, and consistency are examples of DQ dimensions (Batini & Scannapieco, 2016b). While requirements reveal the needs and limits that contribute to solving a problem, data must be processed before any analysis can be performed. Nonetheless, at this critical stage in the BD

value chain, several problems have been identified. This includes DQ, which must be carefully addressed in the context of BD.

Before the advent of BD into the field, there were DQ issues. The researchers classified DQ difficulties and obstacles into three categories, according to Daraio et al. (2016): (i) error correction, (ii) unstructured data to structured conversion, and (iii) integrating data from multiple sources of data. Aside from the difficulties described above, there are other unique BD challenges, such as the massive volume of data created by web 3.0 travelling at a breakneck pace and being confined within schema-less structures. Due to these general problems, BD cleaning and sifting processes must be applied before analysing uncertain data of poor quality. The quality of these data has become so attractive due to the magnitude of the data generated, the speed at which the data travels and the wide spectrum of heterogeneous data.



### 2.2.1 Data Quality Dimensions

DQ is commonly considered a multi-dimensional definition, which defines specific attributes by dimensions of DQ (e.g., completeness, consistency, timeliness). A set of DQ attributes that reflect any element of DQ are termed as dimensions of DQ. Generally, DQ dimensions are seen as a canopy by which a particular feature of DQ is captured in each dimension (Elouataoui et al., 2022). Diversity in data use and data purposes may result in different requirements for DQ (Wang & Strong, 1996). Here, universal and objective measures for establishing DQ play a significant role and add valuable criteria to rather subjective criteria, such as the relevancy of the info to the user. The concept of the DQ dimensions seeks to grasp the basic ideas underlying the data quality dimension DQ, in which each dimension reflects a specific trait (Wang &

Strong, 1996) were early advocates and established the dimensions from a fit-for-use data user standpoint. They identified DQ attributes that consolidate in 20 dimensions categorized into four overarching DQ categories: Intrinsic, contextual, representational and accessible. Here, intrinsic defines that "data have quality in their own right" (N & Mahadevan, 2019), contextual stresses the dependence of data on its purpose. Simultaneously, representational and accessible describe, in a broader sense, the accessibility and usability of the data (Krishna et al., 2023)

The concept of DQ dimensions is still employed today—albeit in varied ways and perspectives. For instance, in a recent literature review, Haug, (2021) acknowledged the presence of over 300 DQ dimensions with distinct denotation. However, some were related either to the same or similar dimensions, whereas some equal dimensions were labeled divergently. This variance illustrates that no uniform definitions have been applied for DQ dimensions. Therefore, users adopting DQ dimensions should rely on the most common definition and additionally provide a description. The International Organization for Standardization also defines a rather technical set of DQ characteristics as part of standards for software quality within their DQ model Haug, (2021), Hassenstein & Vanella (2022), distinguishing between two overarching DQ categories: inherent and system-dependent DQ. Here, inherent features represent those having an intrinsic capability to meet the requirements, whereas the system-dependent characteristics propose those conditional on the information technology (IT) system employed to manage the data. Here, we presume, for example, that relevancy and timeliness of the data take on a different significance to the user (i.e., user-centric view) than to the software developer (i.e., generalist view), for whom availability and understandability may be more significant. Therefore, the views offer varied

perspectives on DQ, as the requirements among stakeholders may vary. Regardless, from our point of view, the definitions of the dimensions themselves do not diverge from the view or perspective backed by the given definitions below. Focusing on the most employed dimensions discovered from the literature to cover the bigger picture.

Completeness is one of the most commonly applied DQ dimensions and generally measures the presence of data. The dimension maps the extent to which all expected and necessary data (records, attributes) are present or not missing, respectively (Hassenstein & Vanella, 2022). Furthermore, it depicts the degree of sufficient resolution for the intended use (Wang & Strong, 1996). The clue to this dimension is to know what the data should contain and analyse whether the data is complete (Taleb et al., 2016). Completeness may be computed by dividing the available items or records by the expected total number resulting in a percentage if multiplied by 100 (Ramasamy & Chowdhury, 2020a). For example, a local population register contains all 932,904 inhabitants, but the date of birth is only available for 930,611 persons. This results in 99.75% (930,611/932,904 * 100) completeness.

Consistency is also known as consistent representation and is the degree to which data are free of contradiction with themselves (Hassenstein & Vanella, 2022) or follow an established rule (Dictionary of Data Quality, n.d.), and are provided in the same format by Pipino et al. (2002) that compatible with previous data. Consistency may be represented as the proportion of items or records found to be consistent (Ramasamy & Chowdhury, 2020a). For instance, we presume the date of birth within a population register should be stored in the "YYYY-MM-DD" (year-month-day) format. In 61,196 of 930,611 total instances, date of birth was stored inverted as "DD-MM-YYYY",

resulting in 93.42% ((930,611 − 61,196)/930,611 * 100) consistency (Hassenstein & Vanella, 2022).

Consistency DQ dimension is linked to semantic rules defined over data objects. This DQ dimension is also related to integrity constraints since the accuracy of the data can be checked using this. An example of an integrity constraint can be a rule or constraint whereby “Loan” is a relational schema with an attribute margin on that loan. The constraint could be anywhere from 0 to 20% of the margin. The constraints must not be only related to one attribute, depending on the relational system, though they can have multiple attributes. This type of constraint is called intra-relation integrity constraint. Another type is the inter-relation integrity constraints with an example: the credit sum for the application will be equal to the value of the loan drawn. This is to say, the application for the loan and the loan amount are their relations in this case. Consistency may also be known as internal or external consistency (Adane et al., 2021).



Furthermore, It is expected that DQ should be quantitatively measured using metrics. DQ dimensions and DQ metrics have to be defined by the DQ organization. This way, DQ can be termed as in agreement with each dimension. Closely 200 such terms have been recognized with little agreement in their definitions, nature, and the way they are measured (Gärtner & Hiebl, 2018).

### 2.2.2 Data Quality Dimension Metrics

According to Günther et al., (2019), DQ is known as multidimensional in literature to represent its aspects and impacts to classify quality-related issues accurately. It is a

multidimensional term since multiple factors must be taken into consideration. These factors are modelled by DQ dimensions measured using metrics. Each dimension is linked to a sequence of metrics that can be evaluated and assessed. A single or multiple DQ metrics may be used to measure DQ dimensions. These metrics typically lead to numerical values. A metric is a function which links a DQ dimension to a value that enables an understanding of the fulfillment of a dimension. This DQ metric can be calculated at various aggregation levels: at a value-level, column-level, tuple-level or a record-level, table-level or relationship-level, and database-level.

For example, the aggregation might be done using the weighted arithmetic mean of the calculated metric results from the preceding level (e.g., the record-level results to generate the table-level metric). A weighted arithmetical average of the computed metric values from the introductory level (e.g. values of the record level in calculating the table level metric) may be used due to the aggregation. Ehrlinger et al., (2018) suggested the five DQ-metrics criteria for accurate decision making: "the minimum and maximum metric values (R1), the metrics value interval scaling (R2), the parameters configuration quality and metrics values determination (R3), the sound aggregation of the metric values (R4) and the economic efficiency of the metric (R5)".



### 2.3 Big Data

Big Data (BD) as a critical challenge could be defined as "a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the process of capturing, storing, distributing, managing, and analyzing the information" (Al-Sai & Abualigah, 2017; Alashhab et al., 2021). Based on the previous research Al-Sai & Abualigah (2017), Al-Sai et al. (2019a), Song

et al. (2022), Hajjaji et al. (2021), Al-Sai et al. (2020), there is no consistent definition for BD between industry and academics. The Statistical Analysis System Institute (SAS) defines BD as a "Popular term used to describe the exponential growth, availability, and use of information, both structured and unstructured" (Al Nuaimi et al., 2015).

IBM also contributed a definition for BD, "Data is coming from everywhere; sensors that gather climate information, social media posts, digital videos and pictures, purchase transaction record, and GPS signal of mobile phone to name a few" (Al-Sai et al., 2019b; Al Nuaimi et al., 2015). Therefore, BD can be seen as both an entity and a process. BD as an entity encompasses a volume of data acquired from multiple resources (i.e., internal and external) and consists of structured, semi-structured, and unstructured data that cannot be analyzed using traditional databases and software methodologies. BD as a process refers to the organization's architecture and technologies utilized to acquire, store and analyze various forms of data (Al Nuaimi et al., 2015; Al-Sai et al., 2020; Alfadda & Mahdi, 2021).

Moreover, BD is pointed out as a technology that facilitates the processing of unstructured data; BD technologies are the systems and tools used to process BD, such as NoSQL databases, the Hadoop Distributed File System, and MapReduce (Benfeldt et al., 2020; Novita & Husna, 2020). BD provides fresh insights on finding new values, helping organizations to benefit from a thorough understanding of the hidden values (Hariri et al., 2019). Big Data analytics (BDA), as technologies (data management and data mining tools) and techniques (analytical methods and techniques), can be employed to transform and analyze large-scale and complex data for a variety of

applications prepared to increase the performance and effectiveness of the organization (Al-Sai & Abualigah, 2017; Al-Sai et al., 2020; Al-Sai et al., 2019b; Al-sai et al., 2023; Al-sai et al., 2023).

### 2.3.1 Big Data Characteristics

The 3Vs — volume, velocity, and variety – are commonly used to characterise the significant characteristics of BD. The concept of many "Vs" is commonly used by academics and practitioners to describe it. The three “Vs” of BD have developed from the traditional three-volume, variety, and velocity (Caulkins et al., 2018; Palareti et al., 2016). However, the relevance of another V of BD, veracity, is progressively becoming apparent to extract value (another V characteristic) to make BD successful. Inconsistency and DQ problems are directly denoted by the term "veracity". One of the primary difficulties with BD is the tendency for mistakes to rise. Even tiny data errors can accumulate without appropriate DQ processing, resulting in revenue loss, inefficient processing, and inability to comply with government and industry requirements (Fosso Wamba et al., 2018). Volume and variety relate to the vast amount of data and many sources and types of data, whilst velocity refers to the speedrate at which data is created (Dubey et al., 2018). On the other hand, Veracity denotes the unreliability of some data sources, which impacts DQ and necessitates DQ management to make accurate predictions (Haider et al., n.d.). Finally, value refers to how meaningful insights and benefits are derived from data analysis via BD (Fosso Wamba et al., 2018).

If the data is wrong, having large amounts of data at rapid speeds is useless. This erroneous data can result in a slew of serious issues for both businesses and

individuals. Analytical results based on insufficient quality data might be inaccurate. Compared to other factors like volume and velocity, veracity is today's most significant hurdle in data analysis. Veracity aids in invalidating data gathered, and believing the data gained goes a long way towards adopting automated decision-making system choices. As a result, there is a pressing need to verify the authenticity of BD. Data veracity investigates data anomalies, noise, and biases. Before it can be used, this information must be cleansed and validated. Since dirty data serves no purpose and prevents business insight, it must be avoided at all costs. While techniques to cleanse and prepare this BD existence is still in their infancy, a solid and effective solution to solving the data veracity problem is urgently needed today (Reimer & Madigan, 2019).

### 2.4 Data Quality in Big Data

BD altered the method in which we collect and analyze data. The quantity of knowledge available is gradually rising, and businesses are increasingly dependent on data collection to gain their competitive edge (Park and Kim, 2019; Pratima Verma and Vimal Kumar, 2021). However, such data volumes only can generate meaningful value when paired with quality: the outcomes of accurate, consistent and complete data are good decisions and behaviour (Tozzi, 2020; Pratima Verma & Vimal Kumar, 2021). In today's BD management, as DQ plays a significant role, it is met with many issues with the explosive growth in data from various data sources. They are usually incomplete, inaccurate, inconsistent, and vague or ambiguous, which could lead to false decisions (Côrte-Real et al., 2020; Janssen et al., 2017; Hariri et al., 2019; Tang & Liao, 2021; Muslihahet al., 2021). Relevant data can be selected in such a scenario by methods and techniques for evaluating the DQ. However, various evaluation approaches are proposed for only conventional databases (but the old DW systems

were not sensitive to the latency presented by this workflow (Berkani et al., 2013; Meehan & Zdonik, 2017; Hamid & Djamel, 2019) as is seen in Section 2.2. It is crucial to extend conventional DQ principles to BD directly. New algorithms need to be built within the BD scenario to resolve current specifications in terms of volume, variety and velocity, value and veracity issues (Hamid & Djamel, 2019).

The data volume grows gradually in the BD era. According to Anosh et al., (2017), this is primarily because of data revision relating to our desire to convert multiple facets of our life into and use digital data. In reality, the authors further say BD's big five challenges (the 5Vs), are to turn the related data into effective choices to reach the optimum benefit. Compared to classical data sources, BD has important peculiarities and is typically distinguished by errors and missing values. The concern is that not all data are relevant. One of the basic issues is that the information collected is noisy, biased, outdated, misleading, incorrect, and thus unreliable (Pavol et al., 2017). The trustworthiness of the data analytics depends on the reliability of the data evaluated for the specific job. This reliability can be measured with DQ metrics, giving data customers an indicator of the quality of their results.

Due to the scale of the data sources, it is often helpful in a DQ study to pick from all available data a more accurate component for research that discards data that have low quality characteristics, because it is not possible to apply the traditional data cleaning approaches in a BD scenario. The assessment of DQ can be a firm basis on which irrelevant details can be identified. DQ literature provides considerable contributions (Fürber, 2015; Chika, 2015; Geisler et al., 2016; Vaziri et al., 2016; Du & Zhou, 2012; Hamdy & Shaalan, 2018; Cure, 2012) suggesting measurement

algorithms for conpleteness and consistency DQ dimensions for structured data, but BD faces new problems in its essential features (i.e., 5Vs). DQ assessment and enhancement processes need to be revamped to fix variety and veracity problems in particular.

The DQ research field is concerned with developing methods for identifying, analyzing and optimizing the adequacy and usability of data in the sense it is meant to be used. Compared to classical data management that concentrated on data approaches and techniques for structured data, BD poses a new challenge to be solved by modern methods (Muslihahet al., 2021). To mine BD, it needs the ability to extract valuable and quality information. We derive data from large data sets and many structured, semi-structured and non-structured forms, making it more challenging. Data consistency is somewhat uncertain since vast volumes of data are produced, and a wide range of heterogeneous data are generated (Muslihahet al., 2021). Previous research showed massive datasets and the internet, creating loss of money, poor service reliability and enormous expense of system repairs, are running low-quality large-scale BD (Hajli et al., 2020; Xu et al., 2016; Verma & Kumar, 2021). Organizations have thus suffered significant casualties. Therefore, the relevance of veracity and value of BD is gradually being understood. According to Hamid and Djamel (2019), studies report that US corporations lose about $600 billion annually in erroneous information in some organizations, the prevalence of data errors is between 1% to 5%, while it is more than 30% in some other big situations. DQ involves tedious data cleaning processes, including discovering rules, identifying/checking for inconsistencies, and data repairing. All these corrections will cost companies approximately 30% to 80% of construction time and budget. It is also essential to control DQ. Data transformation

is a complex operation, requiring several different cleaning activities as a final transformation solution. Conclusively, while techniques exist for performing common data cleaning activities, traditional data cleaning systems design poses many challenges and opportunities. New strategies for transformation have to respond to the rising BD age demands. DQ questions remain largely unexplored.

### 2.4.1 Data Quality Dimensions of Big Data

The importance of obtaining and maintaining a high standard of DQ is significantly essential in every firm. Traditional DQ dimensions have garnered considerable interest in the past few years. According to Ibrahim (2021), due to its simple nature, an extensive range of DQ dimensions can be simply implemented on traditional structured data. However, the research for BD quality dimensions are ongoing initiatives. Only a limited number of the present works explored how to boost the BD quality, and this enhancement procedure was applied on fewer than the structured data quality dimensions. In contrast, the rest of the dimensions are still buried (Juddoo & George, 2018).

It is vital to review the existing quality measures used widely to assess the DQ to determine whether they also apply to BD. For structured data, DQ literature includes various contributions that propose assessment methods for these consolidated dimensions, but BD brings new issues associated with their main properties (Ramasamy & Chowdhury, 2020b). While considering BD, the most significant features, the 3 Vs - Velocity, Variety and Volume play a crucial role. BD indicates a collection of features that directly impact DQ (Taleb et al., 2016; Taleb & Serhani, 2017). Therefore, more emphasis should be placed on applicable DQ dimensions and

matching BD characteristics (Ibrahim, 2021). The ever-increasing studies on BD quality dimensions are scattered, so it is vital to review DQ dimensions suitable for the BD age. In particular, to address these characteristics, it is required to adapt assessment methodologies for utilizing parallel computing situations and minimizing the computation space (Ardagna et al., 2018; Ramasamy & Chowdhury, 2020b).

The earliest research on BD quality with emphasis on DQ dimensions was conducted at the end of 2013 (Ramasamy & Chowdhury, 2020b). This shows that BD systems are still in the early phases of research on DQ. According to Ardagna et al., (2018), most BD quality research recognises the traditional dimensions' importance in the DQ assessment of BD. The readability and structure of data is particularly relevant for large-scale data systems. At the same time, dimensions like as completeness, consistency, accuracy and timeliness that all relate to whether the information is dependable are the most often expressed aspects (Ardagna et al., 2018). On the track of traditional DQ dimensions, most of the researchers in BD area follow the same traditional quality dimensions such as accuracy, consistency, and completeness some of which is the sources of the review in this session.

Taleb and Serhani (2017) proposed a BD quality rules generation and discovery methodology from the quality evaluation findings. The quality is estimated before any pre-processing work. This step gives well-constructed DQ rules to be used in the pre-processing phase. These target data attribute for specific DQ dimensions based on quality evaluation ratings. This information provides a mechanism to develop quality rules for qualities with poor quality scores. The rule set can be adjusted by a user expert and applied to increase quality dimension. DQ dimensions that were measured

were Accuracy (Acc), Completeness (Comp) and Consistency (Cons). Employing accuracy, consistency and completeness of DQ dimensions, suggested a Quality of BD evaluation scheme to provide a set of actions to raise the DQ of BD set using a BD quality evaluation method based on BLB, a bootstrap sample for BD (Taleb et al., 2016). The BLB sampling helps in obtaining an efficient DQ evaluation by decreasing computation time and resources.

Furthermore, Caballero et al. assert that the significant DQ dimension to be considered for BD is consistency, which they explain as the competence of information systems to ensure uniformity of datasets. At the same time, data are being transmitted between networks and systems (Caballero et al., 2014). Their essential claim is that the business value of a dataset can only be measured in its context of use. They further partition consistency into three successive sections, as explained below. However, they have coupled several of the classic DQ dimensions with the three consistency subdomains. Caballero et al. (2014) mapped how the 3v's of BD affect the 3Cs of DQ as seen in Table 2.1.

*Table 2.1: Matrix of 3Cs relative to the 3Vs (Caballero et al., 2014).*

| **Subdomains** | **Velocity** | **Volume** | **Variety** |
|---|---|---|---|
| Contextual | Consistency, Credibility, Confidentiality | Completeness, Credibility | Accuracy, Consistency, understandability |
| Temporal | Consistency, Credibility, Currentness, Availability | Availability | Consistency, Currentness, Compliance |
| Operational | Completeness, Accessibility, Efficiency, Traceability, Availability, Recoverability | Completeness Accessibility, Efficiency, Availability, Recoverability | Accuracy, Compliance, Accessibility, Efficiency, Traceability, Availability, Recoverability, Precision |

Certain standard DQ dimensions still apply to BD. Precision, integrity, consistency, originality and punctuality are among them. Some researchers have identified new aspects important to large data, such as accuracy, consistency, completeness and timeliness (Firmani et al., 2019; Benkhaled & Berrabah, 2019) among which completeness and consistency DQ dimensions were chosen for this study. More so, there are also relationships between these DQ dimensions for example, data set can only be accurate when it is consistent and complete Gökalp, Kayabay, Gökalp, Koçyiğit, & Eren, 2020). Furthermore, timeliness DQ dimension is most suitable for streaming data, but our study data is textual.

### 2.4.2 Data Quality Dimension Metrics of Big Data

Many of BD studies utilize the same standard metrics for assessing the level of DQ, which is (1 – (Nic/Tn), where Nic indicate the number of incorrect values and Tn represents the total number of records (Serhani et al., 2016; Taleb & Serhani, 2017; Taleb et al., 2016). However, few articles discovered that these employed formulas are more connected to traditional data but didn't specify any BD-related measures (Juddoo, n.d.). Differently from the typical utilized metrics, ISO/IEC 25024 was employed as a metric for the DQ dimensions (Merino et al., 2016a). ISO/IEC 25024 contains several ideas and interactions between these concepts to measure DQ dimensions. Like quality measure, quality measure elements, property, etc. Ardagna et al., (2018) employed standard metrics to quantify accuracy, completeness, and consistency.

However, alternative metrics are utilized, such as distinction to quantify the percentage of unique values in a dataset; Precision is used to measure the degree to

which the values of an attribute are near to each other. In particular, accuracy is calculated by examining the mean and the standard deviation of all the values of the studied property. The Timeliness is evaluated by the timestamp that is related to the last update (currency) and the average validity of the data (volatility) (Bovee et al., 2003). The Volume measures the percentage of values present in the studied Data Object. The C4.5 technique (Abdallah, 2019) is used to quantify Noise by categorizing if data is noisy or not. While Commercial Sensitivity is determined by searching data item anonymization or transformation. Commercial sensitivity information could potentially be expressed as part of the metadata. Heterogeneity was difficult to determine. Thus, they draw this information from publications that had utilized these datasets earlier.



## 2.5 Antecedents of Data Quality, Transformation and Effects

DQ research has historically focused on organizational data such as business transactions, structured data and usually deposited in relational databases. Integrity constraints have been used as the primary mechanism to impose DQ. This method avoids unnecessary modifications to data in the database (Debattista et al., 2016). The method does not tackle problems that stem from the data source, such as inconsistent, incomplete, inaccurate, and outlier data. Operational data alone is insufficient to satisfy the demands of strategic decision-making around the enterprise. This void was filled by Data Warehouses (DWs). DW collects, cleanses, transforms and integrates data from numerous operational datasets in an all-inclusive database.

A series of ETL tools are used to promote DW development. Data in DW facilities are seldom modified but regularly revised and are configured for read-only service. DW

design needs extreme care to ensure DQ before the data is loaded in the warehouse. DW poses more obstacles to DQ relative to database settings. Because DWs integrate data from various sources, quality issues connected to data collections, cleansing, transformations, mapping, and integration becomes critical. Data incompleteness, inconsistency resolution, recognition of duplicates, record linkage and detection of an outlier (Cai & Zhu, 2015; V. Gudivada et al., 2017) have been dealt with through various types of rules and statistical algorithms (Cai & Zhu, 2015).

As a normal evolution, the real-world databases today are massive, and thus very likely to include unreliable and low-quality data that are unstable and thus can't be used for data mining, whereas clean-data driven decision-making on business intelligence is vital (Gudivada et al., 2017; Fatima et al., 2017). Online data sources were used in subsequent DQ research. The truthfulness of web-based data sources assessment considers the consistency of hyperlinks, browsing history and source factual data (Dong et al., 2015). Other studies have used the connections between online sources and the accurate information offered by the data sources (Yin et al., 2008). DQ issues were compounded by BD's rapid growth and ubiquity (Gudivada et al., 2015). New problems are the processing of data, the heterogeneity of data and cloud deployments. Moreover, the creation of machine learning models requires high-quality datasets. In the training dataset, for example, outliers may cause inconsistency or non-convergence in ensemble learning. Incomplete and inconsistent data can result in significant prediction deterioration (Gudivada et al., 2017). Decision-makers often conclude for many reasons, from data sources with access to massive quantities of data, which may have DQ issues. DQ is a significant problem in decision-making.

In reality, missing values of the dataset attributes is the most critical issue in knowledge discovery. Moreover, according to Gudivada et al., (2017), DQ problems emerge from unhandled missing data at the transformation stage. Incomplete or missing data, that is, fields for which data is unavailable or missing, is a vital problem since this will lead the analyst to draw erroneous conclusions. When data is retrieved from a DW or database (a typical case when aggregating data from many sources), a cleaning procedure is usually used to eliminate the occurrence of missing data or inconsistent data as much as possible. Missing data attributes problems are sometimes not fixed because the database manager may not have any way of knowing the missing value or incomplete data. At this stage, a database manager can either delete missing data records or supply the dataset with missing records when power losses have occurred for all other attributes in those records (Fatima et al., 2017).

According to Fatima et al. (2017), many data management frameworks have been implemented under the concept of NoSQL to solve different problems in various BD applications. NoSQL systems are fitted with a range of data models and query languages (Huan Liu et al., 2002; García et al., 2016), unlike the relational database model and the star schema-based databases for data management. In addition, data are at the centre of every national, social, and private enterprise in the BD era. Efforts have been made to obtain beneficial information which cannot exist if data are of low quality. DQ is thus treated as a crucial component in the processing stage of BD. Low-quality data is not inserted into the BD value chain in this process.

BD moves through all stages of its life cycle. This includes processing, analytics, and visualization. Nevertheless, these phases will not make the data clean, complete, and

consistent. However, the processing of data is highly vulnerable if data is inappropriately prepared for processing. This leads to unusable data analysis due to factors such as data origin, data format, nature of data and inadequate data preparation. In BD, data itself and its consistency are critical issues. BD shows a variety of functionality specifically influencing DQ (Karau et al., 2015). A data transformation is expected to focus on maintaining a certain DQ standard (García et al., 2016).

Incomplete and inconsistent data are typically present in a relational database and in BD where data from various sources have been combined. This is because multiple values from various sources can be integrated with an entity attribute in such centralized databases. Furthermore, without a proper check for data varification may also lead to incomplete and incosistant data addition to the DW. An object attribute can associate contradictory values (i.e., inconsistent data), while certain object attributes may not have values (i.e., incomplete data). Data is distributed in BD storage networks, allowing distributed processing and handling of vast volumes of data without storage limitations. Furthermore, fault tolerance, geo-circulation, and data replication offer high availability. This may however, lead to DQ problems such as inconsistencies in many data centres. Because of the variety of sources, the data collected may have varying quality levels such as redundancy inconsistencies, integrity, etc., types of quality. It would have expensive transmission and retrieval of raw data. On the user side, particular approaches and frameworks for data processing may have stringent DQ requirements.

Addressing DQ involves coping with specific concerns such as knowledge evaluation in different DQ stages, error detection, missing data discovery, and improving DQ in

general. Thus, the DQ challenges fit the creation of data processing approaches and models. The lower the DQ, the less valuable the data will be.

### 2.6 Data Warehouse

Business Intelligence and Analytics (BI&A) refers to a wide variety of applications, technologies, and approaches for selecting, storing, distributing, and analysing data to allow better decision-making in business organizations. DW are the cornerstone of BI & A systems as "a subject-oriented, integrated, time-variant, non-volatile collection of data supporting the management decision-making process". The function, subject-oriented, refers to a DW that includes only the data used to support system processes within an enterprise. Data for urgent data analysis that does not assist decision-making was included in the operational database systems compared to DW. The subject aspect of the definition above refers to the data characteristics used in the company's Decision Support Framework (Ricamato & Chicago, 2015). As the second feature of the definition of a DW, data is obtained from various heterogeneous and self-sufficient data sources in data warehousing. However, the data in the warehouse needs to be stored inside a schema that satisfies consumer requirements analysis.

Data from the source is typically integrated and transformed until the data is loaded into the DW, meaning that DW users can use the merged data rather than think about the completeness and consistency of the source data. The non-volatile attribute of the DW definition suggests that the data is normally long-term, is not modified or re-established in real-time. In operational database systems, data is typically modified, and upgrading procedures, such as addition, data shift and deletion, are generally performed. The data is typically used for the care of decision-making support

mechanisms in data warehousing. Concentration will now be on querying the data instead of modifying or deleting it after the data has been loaded into the DW. However, a DW has to be refreshed from time to time so that updates are revealed inside the major data sources. On the other hand, bulky storage is usually used to store the previous data in the DW and delete redundant data throughout the periods. By using this time-variant knowledge or information, consumers would evaluate and forecast the company's success and future developments. Defiantly, operational database applications mostly consider only the current dataset.

In summary, a DW is configured for a Decision Support System analyst or an organization that does not have to be technically oriented to access the prevalent business-related information through the organization effectively. This entails integrating data derived from some data sources that can be segregated, dispersed, autonomous and heterogeneous such that data can be taken advantage of. DW is single, reliable and whole. The essential duties of a DW are to gather, cleanse, filter, integrate, transform, and reorganize source data to be loaded into a single storage facility with a schema that satisfies the user's analytical requirement. And this method is known as the Extract-Transform-Load (ETL). The technique is used for all the above tasks.

Figure 2.1 demonstrates the BI & A architecture in which data is obtained from an external and operational database (Yulianto, 2019) . The collected data has various forms and formats, is heterogeneous. These data are transformed and integrated by ETL processes before loading into the DW. This is carried out by three fundamental functions: (i) retrieve from data sources; (ii) transform data to the proper format or framework for querying and review purposes (i.e., data cleaning, reformatting,

comparing, aggregating, etc); and (iii) load the resultant data collection into a target device, usually the DW. Datamart (Data Cube) is the DW environment's access layer used to relay data to decision-makers.

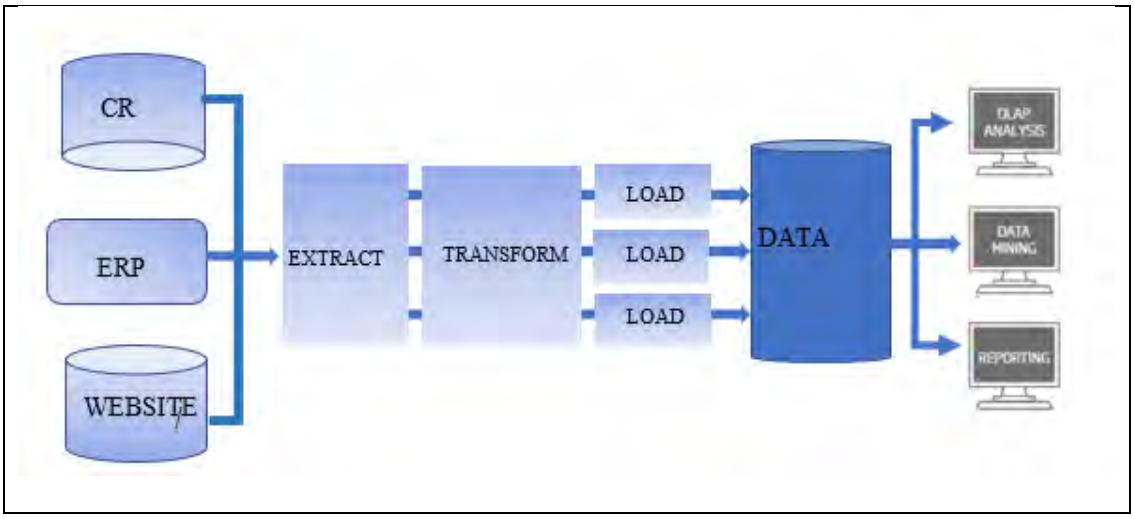



*Figure 2.1: Data Warehouse ETL Process (Yulianto, 2019)*

### 2.6.1 Extract-Transform-Load Process

As shown in Figure 2.1, ETL is main process for transforming and integrating data for various sources into a single format structure of data store. The main processes of ETL are described as follow:

**i) Extract**

Data extraction is the first step in every ETL scenario. The Extract step is responsible for extracting data from data sources and making it available for processing. The primary goal of this ETL phase is to get all necessary data from the data source using as few resources as feasible. The data extraction step should be built to have no detrimental impact on the response time and performance of the data source system.

**ii) Transform**

The second stage in the ETL process is Data transformation. The transformation step cleans and adapts the incoming data to provide accurate data that is consistent, comprehensive, correct, and unambiguous. DQ comes in because these extracted data may be stale or noisy, resulting in inconsistencies, missing values, or other irregularities. This part consists of the processes such as data cleansing, transformation, and integration. It determines the level of granularity in fact tables, dimension tables, DW schema (star or snowflake), derived facts, slowly changing dimensions, and factless fact tables. The descriptions of all transformation rules as well as the schemas that arise are part of metadata repostory. This study is concentrating on the transformation step of the ETL.

The transform stage applies to a collection of rules for transforming data from a source to a destination. This involves the conversion of measured data to the same dimension, that is, conformed dimension using the same units to enable them to be later joined. Joining data from several sources, producing aggregates, generating surrogate keys, sorting, deriving new computed values, and applying complex authentication rules are all part of the transform process. Data retrieved from the source may require several calculations or a qualitative modification according to the rules specified in the mapping file before processing into the ETL loading phase. Transforms and maps from the source data system all essential data pieces to the desired target data system.

**iii) Load**

The last ETL step is to import data into the multidimensional structure that you want. The extraction and transformation of data into the dimentional structures which are

accessable to the end user is involved in this step. The loading of dimesion and fact tables are part of the loading procedures. It is critical to ensure that the load is completed successfully while utilising a few resources as feasible during the load stage. The objective of this procedure is to create a database regularly. Any restrictions and indexes must be disabled before the load and only enabled once the load is complete to make the load more efficient. The referential integrity requires to be maintained by ETL tool to guarantee consistency.

Basically, as can be seen from Figure 2.1, The Extract, Transform, and Load (ETL) procedures are used in DW. Data is taken (internally or externally) from defined operational information sources, transformed, and fed into the final target file. These extracted data may be stale or noisy, resulting in inconsistencies, missing values, or other irregularities. As a result, they will be transformed to comply with DW requirements, such as the DW's integrity constraints. Also, not all of the extracted data may be necessary for specific management decisions. Therefore just the data needed for that analysis will be put into the target file.

In ETL operations, issues with inconsistent and incomplete data are evident. Data cleaning, for example, is a critical activity in a DW's ETL operations, and it involves finding and eliminating mistakes, missing data, and inconsistencies in the data collection. In ETL procedures, data cleaning ensures that the data in the DW is accurate, consistent, and complete. They stress the need to recognise the same dimensional object across several source systems and resolve overlapping descriptions during the loading phase of ETL procedures. Works such as identified data inconsistency and data incompleteness as a challenge in DWs.

### 2.6.2 Data Transformation

Data transformation is a mechanism to manipulate data sources fitting the format of the DW structure. Usually, the mechanism is executed within a heterogeneous data sources environment. Data transformations are performed for complementing the ETL processes. Data must be transformed into a format compatible with the new system while still retaining the information stored in the old system. The majority of businesses are attempting to make themselves far more operational by utilising simple methods to work with. They're also making progress in building their DW to store and monitor real-time and historical data. The immaculate integration between various data sources is implied by the ETL tool. It functions as an integrator, taking data from many sources, converting it into the desired format using business transformation rules, and feeding it into a cohesive system, databases and/or DW. There may be data mismatch, miscalculation, and/or loss of important data during the transformation of a large amount of data, resulting in a failed data transformation. (Tesfagiorgish & JunYi, 2015).



Data is transformed in various ways, many times, and to a variety of formats, resulting in size reduction, schema-less to structured representation, and the elimination of noisy data, among other things. Transformations that collect data as input and produce a new dataset must be documented, stored, and queried. To accomplish DQ in BD, vital information such as the source, origin, types of data, whether structured or schema-less, date of data generation, creation, and transformation, among other things, must be specified and connected to the data from its genesis (Khalilijafarabad et al., 2016; Taleb et al., 2015).

Data transformation can be referred to as converting or transforming data from one format to another. This data transformation activities can in general be divided into two types i.e., syntactic and semantic transformations. Syntactic transformations are structured to convert a table from a syntactic format to another, usually without external information or reference details. Instead of merely sequencing characters, semantic transformations typically include an interpretation of the meaning or semantics of the data; instead of just as a character sequence; therefore, they typically include referencing external sources of evidence . For example, the table of data that include names, telephone numbers and fax numbers, initially organised in tabular format and with the telephone numbers and fax numbers in a complex data type column, is flaked into a regular relational table view in order to facilitate efficient data query demonstrating a syntactic transition. Additionally, by incorporating two dashes between nine digits, telephone and fax numbers are both transformed. These transformations do not need external expertise.



A semantical transformation in country codes transforms them into country names, which clearly involves additional information, such as an external knowledge base comprising complete and abbreviated country names. The syntactic data transformation method typically has three key components/dimensions: language, authoring, and execution. As a result, a dataset can be transformed in several ways; syntactical data transformation strategies generally embrace a transformation vocabulary that restricts the scope of potential transformations. The language might be a restricted range of operations permitted on the table or it might consist of a few features, for example, such as dividing a column and combining two columns, or it

might entail a set of functions for string manipulations (Gulwani, 2011), such as removing substring and concatenating two substrings.

The combination is important for a given language because it is sufficiently descriptive to cover several activities of transformation in the real world and, at the same time, sufficiently limited to allow the transformation to be written accurately and efficiently. Syntactic transformation platforms provide multiple styles of engagement with users to write transformation programs using the chosen language for a particular transformation mission. The models of interaction may usually be categorized as declarative, example-oriented or proactive. Users demonstrate the transformations explicitly in the declarative model, usually with a visual interface. Users must provide a few examples in the example-driven model, based on which the tool automatically deduces the possible transformations which fit the examples given. The transformation tool offers instructions or highlights in the proactive interaction model for data transformation and recommends potential transformations to make users easily accept or deny suggestions.

Finally, the specified transformation implementation applies the specific transformations to the dataset. As several transformations can be defined from the previous level, Transformation Tools generally assist users in choosing the appropriate transformation, for example by instantly presenting a sample data's effect of the defined transformation or presenting an interpretable description of the specified transformations. Disparate transformation tools provide diverse abilities along the three dimensions. Instead of merely as character sequences, semantic transformations

generally require an interpretation of significance/semantics or the regular use of the data.

In comparison with syntactic transformation, semantic transformations can be determined by the input values but typically involves additional external data sources to search for the values to be transformed. Example-driven techniques have been mainly used for semantic transformations (Malik & Singh, 2019; Singh & Dwivedi, 2020; Abedjan & Naumann, 2016). Input-output examples are usually requested of the users, and semantic transformation tools are then retrieved from the available external data sources to find a query that matches the examples. Then, it can be used the query to convert new invisible data (Chu et al., 2016).

### 2.6.3 Data Quality Defects within the ETL Process

Several reasons explain the need for transformation of data in the ETL: (i) heterogeneous formats; (ii) hard to read or unclear data formats; (iii) outdated structures using redundant databases; (iv) the structure of the data source is evolving over a while. DQ is unknown with all of these data source characteristics; and (v) data is recieved from unknown sources or is modified by any adversary. During the data integration process, some experiments have been undertaken to recognize various quality problems. Most of the studies accept that DQ faces multiple challenges. ETL is a vital part of the DWs in which most of the activities is cleaning and processing. Hence, the typical DQ issues in the transformation phase of the ETL. Each move is susceptible to various quality difficulties at both schematic and instance levels. DQ concerns in the DQ improvement method concentrate on the ETL process. The improvement process is also a required condition for the assessment of DQ in practice.

The incorporation of DQ in the ETL process demonstrates the difference between the outputs achieved and expected. In addition, it is also difficult to solve all DQ issues. Otherwise, the groundbreaking approaches to incorporating and enhancing DQ in the ETL method are classified.

### 2.6.4 Data Quality Approaches in the ETL Process

The literature demonstrates a range of methods to deal with DQ problems in the context of data transformation. Several efforts were made to surmount the problems of DQ. In short, however, two streams were defined for DQ management: (i) process-centred approaches that deal with DQ through the quality of ETL processes, and (ii) data-centred methods that consider DQ defaults.



**i) Process-Oriented Approaches**

It is understood that quality features are non-orthogonal. In addition, the understanding of quality and appropriate standards can vary between users for each feature. Thus, in view of the user's desires, it seems necessary to obtain the desired outcome in terms of priority quality functionality. Some authors have suggested a functional architecture of the user-centric ETL mechanism in which DQ characteristics are integrated based on the user-defined value of every characteristic feature. The user directly chooses quality attributes of concern to gain a recommendation on quality assessment. This stage helps the user decide how well they fit with the strategic targets that have already been established. The Quality Appraisal consequently allows users to evaluate which quality aspects can be changed to create a process version that suits the users' objectives.

The automated data generator Bijoux was proposed by Nakuçi et al., (2014). This tool extracts the semantics of the data operations and analyzes the shortcomings suggested by the input data based on an ETL process logic model. Bijoux produces approximately a collection of data to test the different execution scenarios. In order to cover all test scenarios, a separate path in the ETL method, models as a matrix, are enumerated using an algorithm. To produce test datasets, Bijoux analyzes for any direction the various limitations related to ETL operations. This method has proven helpful for evaluating the output of various implementations of the same ETL systems. On the other side, a tool named POIESIS was introduced that produces alternate flows by incorporating the original flow patterns. This is focused on various consistency characteristics on user expectations. DQ and output metrics are used in each generated flow. The user has to determine which is sufficient for his quality specifications. The authors have suggested component flux patterns among the DQ dimensions listed in the literature, including consistency, efficiency, reliability of accuracy, completeness, deductibility, and the filtering of missing values.



**ii) Data-Oriented Approaches**

The data level quality defaults are tackled using data-oriented approches which are further resolved. The current approaches are classifies into the following: (a) syntactic, (b) semantic, and (c) contextual points of view.

a) **Syntactic oriented approaches**

Data inconsistency is a significant quality default. Suppose an organization has separate division consumer databases and wants to merge them. In that case, the recognition of compatible information is a real difficulty because the same data are

represented in different formats. This problem is characterized as the mechanism by which documents are found and combined that reflect the same actual entity. Chu et al. (2016) suggested three algorithms to evaluate matching records based on the similarity of characteristic components. Similarly, the Word Frequency-Inverse Document Frequency (TF-IDF) metric has been used for the identification of duplicates.

Van Gennip et al. (2018) also proposed a new way of expanding TF-IDF to fix inconsistencies, missing data values and incompleteness. In particular, missing data may prevent inconsistency detection. In the following example, both documents are similar, but TF-IDF implementation does not suffice for a suitable outcome. Data from different sources loaded into DW has diverse coding and format (for example, data format). Since the user can reach different data formats, authors have proposed a Mass ETL open source tool that manages data translation and naming conflicts and the management of missing values. The default format is specified by a mapping rule of the data format.



b) **Semantic oriented approaches**

Knowledge about data semantics is important for data transformation and integration. The layer of Bergama projected an extraction-friendly tool that allows semi-automatic semantic mapping between the data source and DW features. This analysis describes the components of DW-linked data source systems and uses syntactic and semantic measures of similarities. The task is done in three separate phases: (i) the extraction from sources by using a wrapper for various forms of data sources (RDF, XML, etc.); (ii) WordNet annotation process, the process in which the most appropriate definition

of the database ontology is combined; (iii) the creation of a thesaurus of relations between schema components. Using the semantic enrichment results, the authors developed an algorithm for mapping the appropriate transformation function to get data types and the thesaurus produced for the transformation process.

There can be a wide range of anomaly types in the data, e.g. disputes with the names of modified objects, including homonymous attributes or synonyms. The lack of semantics triggers these issues. In addition, semantic data improvement enables increased cleansing. There is no formula for measuring the resemblance between Pekin and Beijing, for instance. The solution is to identify the data schema by suggesting a semantic name for each data source column. This is because the meaning of each column with regular expressions and the database is defined in the group and vocabulary. In addition, the syntactic and semantic authenticity of the data source values is provided with a semantic profiling algorithm. The algorithm returns a series of tables that include the profiling outcome given regular expressions, syntactic and semantic indicators, data dictionary, and the data source.



c) **Context oriented approaches**

As for the consistency of application functionality according to ISO / IEC2 50101 quality standard, context coverage has been introduced. Batini et al. (2009) have demonstrated that both qualitative and objective consistency characteristics must be tested concurrently. They suggested a theoretical model that stresses the importance of integrating processes of contextual DQ evaluation. van den Broeck et al. (2020) took the first step to formalize and explore the principle of DQ evaluation and DQ query response, using the data log and the first-order data logic as contexts related

operations. Another method of testing the usability of test cases is defined in where authors illustrate how a quality plan can systemically be built. In addition, a new set of background variables has been added, which may affect the DQ assessment. In the DQ assessment in the ETL phase, however, most prior studies fail to take into account context in DQ its meaning nor stress its primary function in influencing DQ steps.

### 2.7 Common Data Quality Issues in Data Warehouse

The quality of the ETL process, which is considered the prerequisite for good business intelligence (BI) and analytics method is used for DQs within the DWs and in data cubes. A literature review is performed to recognize the different methods in the ETL method to deal with DQ issues. It is observed in the literature that scholars are writing on ETL-related DQ issues from two major viewpoints: (i) method based and (ii) data based. In addition, the DQ drawbacks of ETL have been demonstrated.



The key point here is to consider the best way to improve quality when attempting to improve it without risking losing quality. The easiest way to solve a DQ problem is to find the principal root of the problem and ensure that it does not arise again. DQ problems can be produced at virtually every point in the data life cycle, distinguishing four significant classes of DW's problems: data source problems, data profiling issues, data staging issues or ETL, and data modelling issues (database schema design) level. DQ issues have been categorized into four major classes.

ETL is the most crucial step in the development of DQ problems. It is also the best place to conduct quality controls on the quality of data. This is because the data is formatted and transformed at the time of the ETL process. Execution such as data

cleansing, which denotes data cleaning to improve the consistency and completeness of DQ dimensions, are done herenot storing columns containing either missing or null values, for instance. The key or attribute level problem can be a common issue in the ETL method. This can happen because of errors such as logic deficiencies in the ETL process. The problems with the attributes relate to information that does not match various sets of data. This challenge impact negatively on DQ dimensions such as consistency. DQ problems such as lack of proper extraction logic and loss of data in the process also occur. This loss applies to the fact that data is omitted. Examples may be data that do not agree with any rules in the ETL phase or are denied because of quality concerns from a previous level.

For example, by producing business rules and checks for various sections of the ETL process, the bugs in the ETL process can be solved. The inspection is to ensure that the data are transformed following the rules required. The rules facilitate information technology (IT) roles to enable us to understand the whole ETL process. If the IT help is comprehending the procedure and rules, problems with the processing logic may be fixed whether they are violated or modified. Malik and Singh (2019) also recommended that no user-written code for the ETL method should be used, as this creates typically quality problems. Tools that are intended to be used for the ETL process ought to be used in the place of the user-written codes.

DQ Research suggests a range of consolidated approaches, which work well on relational models. Still, it is not sufficiently implemented in BD environments: BD needs new methods, models, and techniques for the description, assessment, and improvement of quality dimensions. Several articles in the literature contend that the

DQ dimensions must be redefined for BD. For example, it reflects on changing DQ dimensions and explains how their meanings vary due to the form of data and the sources and applications. In the BD case, DQ dimensions were also evaluated. The authors also describe a quality measurement framework for BD: the description of the quality dimensions here depends on the data collection targets, the organization context and the data sources involved. A model is introduced to evaluate the quality-in-use of data used in BD.

Several current ETL techniques process the static data generated by highly structured databases and complement DW in the format needed for further decision making. Some ETL techniques aim to manipulate data stream processing. Still, data streams are distinguished by the volume, velocity, and variety of data to be handled for data processing. From this point of view, the existing ETL methods are insufficient to solve all the BD functions. The ETL method, which can fix BD issues, is therefore required to be redefined. As provided by the conventional DW techniques, several emerging techniques transform structured relational data into fact tables and dimension tables. Yet BD is also heterogeneously characterized. Thus, the complexities resulting from the variability of the data must be addressed. Current approaches seek to deal with data heterogeneity by building a semantic representation of ontologies and related resource description framework (RDFs).

A technique for handling data streams in real-time ETL is given by the authors of. The proposed approach distinguishes ETL from historical ETL in real-time. It implements a dynamic storage area to solve the batch upgrade issue in sync with real-time data processing and new query results. In particular, the architecture does not resolve the

problems of processing high volume and speed of unstructured data. To handle semantic heterogeneity of data, authors and integrate semantic technology into the ETL method. The methodology generates a semantic data model by mapping ontological data and loading the RDFs to the DW. The methodology answers the challenge of integrating heterogeneous source data into a standard format, thus addressing variety, which is one of BD characteristics and offers an ability to coordinate the data through RDF connections to improve performance precision. This methodology does not consider veracity, velocity, and volume of the data in this methodology as ontologies are challenging to construct manually.

A semantic technique was proposed that uses the linked RDFs to describe the sets of data. Semantic filters are used to integrate vast quantities of heterogeneous data. Continuous queries filter semantic sources of data that summarization methods to remove some data once it crosses a limit to be handled. Consequently, if it arrives at a high volume and speed, the semantics of the original data will be lost.



### 2.8 Ontology-Based Data Management

Ontology-Based Data Management (OBDM) is a new framework for approaching data management based on the ontology conceptualization of the field of interest. A structure that is capable of applying the OBDM view contains three levels: the ontology providing a high level, structured, logical representation of this conceptualization; the data source layer, which signifies the data in the numerous system assets; the mapping between the two layers, an apparent description of the communication amid the ontology and the data sources. Though most of OBDM 's work focuses on data query via ontology, the latest papers argue that OBDM is an

excellent tool to assess DQ, mainly when various data sources are likely mutually incoherent (Croce et al., 2018).

The Resource Description Framework (RDF) also can be extended to define resources (meta-information) in a web context. It also aids in the representation of services. Semantic technologies for information systems may be focused on basic techniques, including glossaries (lists of words and their definitions), taxonomies (hierarchies of terms) and thesasauri-based methods (relationships of connections with synonyms) to evade syntactical and semantic glitches during the creation and interpretation of product data. Methods with additional fullness of semantic are topic maps and ontologies.

When generating and describing abstract knowledge and concepts relationship, semantic technology can help to understand knowledge by defining the meaning in question. The context, intent of the data (e.g., symbols, terms etc.) abstract concepts and sharing information for humans and machines can be better understood by semantic technologies. Metadata provides more information about other data and thus helps to locate knowledge and records more efficiently to promote semantic applications. Metadata in various data sources may also be related to other metadata. This requires, therefore, uniform guidelines (standardized rules), allowing metadata to be systematically defined. These rules are essential and allow for metadata sharing between applications, information systems, and workstations (De Giacomo et al., 2018).

As seen in Croce et al. (2018), semantic means developing information systems where data semantics are precisely defined and all the system functions are planned. This term has been increasingly relevant for a broader scope of data processing applications in the past two decades and has gained significant attention in the cultures of artificial intelligence, the databases, the web, and data mining. Here, the focus is predominantly on OBDM, implemented around ten years ago as a new way to model and communicate with data sources collection (Calvanese et al., 2017; Chen et al., 2009; Croce et al., 2018). This is the basis for the creation of data sources. According to this model of OBDM, information system clients are free from realizing that the data is structured in particular objects (databases, software applications, systems, etc.) and are able to communicate with the system in the form of accomplishing their queries and goals in terms of logical description of the interest area called ontology. Moreover, an OBDM system consists of an information processing system managed and operated by a single entity (or a group of users) whose design has the same framework as the traditionally data integration system. As with the following mechanisms: a set of data sources, an ontology, and the mapping amid the two.

a) Ontology is a formal philosophical description articulated in terms of the related concepts, concept attributes, relations between concepts and logical assertions which formally describe domain knowledge of the organisation's field of interest
b) Databases are archives where the organisation's access to store data relating to the domain. In general, these repositories are multiple, heterogeneous, each of which is individually controlled and preserved.
c) The correspondence between data comprising data sources and the ontological elements is described precisely by mapping. The element here means concept, attribute, or relationship.

The above three layers form an advanced structure for describing information, which can be controlled and considered by automatic reasoning techniques. For instance, practical algorithms respond to queries articulated in ontology by mapping the problem by automatically translating the data sources (Calvanese et al., 2017). While the issue with responding to queries over the ontology has been the key emphasis in recent years, an OBDM system possesses the ability to offer more services. A critical example of this is the assessment of DQ.

DQ is often one of the critical factors in providing value-added intelligence services (Ilyas & Chu, 2019). However, the heterogeneity of the data sources and the fact that their structure always depends on the applications they offer hinder the purpose of even inspecting the DQ, let alone attaining a reasonable degree of consistency in the delivery of information. How do we set standards for DQ if the semantics that data can be given are not well understood? The dilemma is compounded by linking to external data, e.g., clients, vendors, consumers, and even public sources. Again, it cannot be achieved without profound comprehension of the consistency of external data, even without determining whether to conciliate potential discrepancies or merely to incorporate those evidence as separate views. That is why extending the OBDM model to the issue of the DQ evaluation is essential and promising (Wand & Wang, 1996, Console & Lenzerini, 2014a; 2014b). Centered on a structured domain of interest conceptualization, this analysis helps us quickly remove all irrelevant data source information and concentrate on actual DQ problems. In addition, the multiple data sources may be analyzed using the same yardstick, i.e. ontology, and thus their content may be measured and compared in terms of quality. Finally, using conceptualization

shared by an organization's various properties allows for DQ evaluations to be conveniently expressed and applied in several different ways.

### 2.8.1 Ontology-Based Data Transformation

Ontology may be a crucial solution to integrate BD sources and traditional sources. Integration schemes depending on ontological frameworks can also allow for systematic quality assessment of inconsistency and incompleteness of integrated data (Al-sai et al., 2023). Apart from ontology-mediated inquiry and other data management activities, current studies contend that OBDA is a promising method for evaluating DQ, especially where numerous independent data sources are involved (Engineering, 2022). The following are some considerations: (i) based on a systematic conceptualization of the area of interest, and we can conveniently flutter out all of the irrelevant information of the single source of data and focus on the actual questions of the DQ; (ii) the same criterion, i.e. ontology can analyze various sources of data, and hence accessed/compared to quality; (iii) by conceptualization shared amongst the diverse assets of the organization permits for DQ assessments which enables the straightforward presentation and use in several diverse contexts. Quality tests are well known to be carried out in various ways, for example, consistency and completeness. The DQ dimension that deals with data coherence is consistency. Anything outside the example of consistency in data indicates that data suffers from credibility concerns and thus provides critical information on the assets which possess these data. In literature, consistency is also measured by confirming the material meets basic credibility rules, nevertheless, those rules are either tacit in conventional methods or defined according to the analyzes for the particular data source. OBDM, on the other hand, advocates a new approach, which would explicitly extract the regulations to be verified from the

ontology and verify the logical model of the domain in a step (Lenzerini, 2018; Engineering, 2022). Further, the inference capability of ODBA systems provides automatic techniques for the access to consistency, distinguishes the different inconsistencies contained in data, and rates them according to different pre-determined parameters instead of laborious consistency testing activities for the various sources. For instance, if ontology states that a physical variable might have a set of values and mapping states that specific sensors convey data about that variable, it may be liberated from applying a certain rule since it may rely on the OBDA mechanism to automatically validate the rule.

In recent years, we point out that rigorous research on OBDA has created optimized consistency checking algorithms, which scale when used in BD sources. Regarding DQ evaluation, the authors describe a general structure for OBDA data consistency and offer algorithms and complexity analysis in numerous diverse tasks associated with checking DQ under the consistency dimension.



Studies have been proposed using the ontology technique on BD to assess the quality of a dataset (Mendes et al., 2012; Zaveri et al., 2016). Even though these studies provide useful ways to assess the quality of a dataset, they often still do not address the improvement of the datasets (Rula & Zaveri, 2014), such as the case of a recent study in (Yuan et al., 2019). Most of them only suggest improving the DQ in their further studies (Cai & Zhu, 2015; Jarwar et al., 2019). The studies presented in the literature have shown that the DQ process problem still requires a lot of efforts, especially in BD, based on non-relational data sources, e.g., NoSQL databases (Abdelhedi et al., 2017; Croce et al., 2018). This gap drives the researcher to propose

an ontology-based DQ approach within the scope of quality-driven solutions by improving the quality of BD. The goal of this study is to improve BD quality using OBDM for DQ assessment. This research will show how ontology can have a significant role in addressing quality issues in data transformation from the perspective of the usual dimensions studied in the context of DQ, namely, consistency and completeness.

Ontologies, annotations, metadata, web service interfaces, and other components fall under this category. Frameworks such as the RDF are also helpful. Paganin et al., (2019) and Jafer et al., (2017) are creating web ontology language (OWL) that is a standardised way to represent data to make activities like finding, retrieving, and processing easier. Furthermore, Linked Data is a potential technique for simplifying data integration and retrieval in the IoT ecosystem (Zou, 2018). For example, the study presents web technologies and linked data principles as a semantic data integration framework.



### 2.8.2 Ontology for Addressing Data Incompleteness and Inconsistency

Besides ontology-mediated querying and other data management activities, recent research claims that OBDM is a viable method for analyzing data quality, especially in the presence of many independent data sources (Engineering, 2022; Lenzerini, 2018). Some of the causes are as follows: (i) basing DQ assessments on a formal conceptualization of the domain of interest allows us to easily blur out all the meaningless details of the single data source and focus on real DQ issues; (ii) different data sources can be analyzed using the same yardstick, that is, the ontology, and hence accessed and compared in terms of their quality; and (iii) the use of conceptualizations

shared among the different assets of an organization allows for DQ assessments that are easy to present and use in many different contexts.

Quality assessment is carried out through many dimensions: consistency, accuracy, completeness, and confidentiality. Regarding consistency, the quality dimension dealing with data coherence, counterexamples to consistency reveal that data suffers from integrity difficulties, thus offering critical information about the assets owning such data. The literature often proposes that consistency be examined by evaluating whether data follow certain principles for integrity (Engineering, 2022). However, in previous approaches, such rules are either implicit or defined based on the data source under analysis. On the contrary, OBDM proposes a new way, where the rules to be tested are drawn directly from the ontology and where they have also been validated by developing the conceptual model of the domain. In addition, instead of implementing laborious quality-checking tasks for the various sources, the inference capabilities inherent in ODBM systems provide automated techniques for accessing consistency, singling out the various inconsistencies present in the data, even ranking them according to various predetermined criteria (Xue & Zou, 2022).

One is not compelled to develop specific criteria for checking whether a client exists that is identified by the data sources as both an ordinary and a special customer. Indeed, reliance is placed on the automatic verification of the rule using the OBDM system as part of the consistency check of the complete OBDM system. Research carried out in the previous years has generated optimal methods for consistency testing, which scale nicely when applied to BD sources. Similar concerns exist regarding completeness DQ dimension (C. Liu et al., 2023)(Lenzerini, 2018).

## 2.9 Data Quality with Clustering Algorithm

Machine learning and clustering-based approaches are gaining an increasing interest in DQ assessment and improvement, such as in duplication detection in assessing data completeness (Ezzine & Benhlima, 2018; Huang et al., 2020; Zhao et al., 2018), in detecting patterns and data correlations (Homayouni et al., 2019; Rashighi & Harris, 2017), and in anomaly detection (Deja, 2019; Estiri et al., 2019; Lin & Su, 2019; He Liu et al., 2019; Nematzadeh et al., 2020). From a technical point of view, among the most recurrent clustering approaches, we can cite k-Means in anomaly detection and completeness assessment (Zhao et al., 2018). The local outlier factor in outlier detection (Xu et al., 2020) k-nearest neighbors (k-NN) in the correction process (Li & Gao, 2019).

Exploited character-based similarity measures were created by Levenshtein and Pradhan (1975) and Jaro-Winkler (1990). OpenRefine Ham, (2013) is commonly used in data profiling and quality enhancement (Noor et al., 2020). It takes advantage of a similar approach in the textual facets: based on the Levenshtein similarity metric and the k-NN clustering approach, it splits the tested column into bags of words, and users have the opportunity to correct errors (if any) by merging clusters and making the cluster content uniform. We take OpenRefine as a referring point in the evaluation phase, and we will show that without a dictionary of proper values, the likelihood of recognizing wrong values is reduced dramatically (Pellegrino et al., 2020; Pellegrino et al., 2020).

K-means is one of the most common and most accessible grouping algorithms. Data clustering aims to detect a set of patterns, points or objects as a natural grouping.

Concentration is to K-means in this study. K-means has an abundance and varied history as found separately in many sciences Steinhaus fields (1956), published by Lloyd (1957), Ball and Hall (1965), and MacQueen (1967). Although it is one of the most often used methods for clustering, K-means was initially introduced over 50 years ago. The principal grounds for its appeal are its easy implementation, simplicity, efficiency and empirical success. K-means is one of the unsupervised learning algorithms and can be employed to classify data through a number of clusters, for example, k-clusters, to reduce the squared error for optimised classification. In terms of simplification, robustness and speed, the method is beneficial. K-means algorithm uses Apriori specification number of clusters (Mittal et al., 2019).

K-means is one of the unsupervised learning algorithms and can be employed for classification of data through a number of clusters, for example k-clusters, in order to reduce the squared error for optimised classification. In terms of simplification, robustness and speed, the method is beneficial. It works relatively well with unique and linear datasets compared to other clustering Algorithms. K-means algorithm uses Apriori specification number of clusters (Mittal et al., 2019). The developments in K-means are summarised and then discussed the main techniques created to data clusters.

Detecting errors in text data can be carried out using machine learning clustering, such as k-means (Juddoo & George, 2020). K-means is a framework of clustering or a family of distance functions, which provides the basis for numerous forms of k-mean algorithms. K-means is commonly used in clustering BD due to its simplicity and fast convergence. K-means is based on a relatively simple calculation technique, yet it is

susceptible to outliers data shapes, and it presupposes that clusters have nearly equal amounts of observations (Hassan et al., 2021).

Related work includes Freudenberg et al. Zou (2018), Zou (2018) developed a graphical blending of knowledge-based functional catagories with cluster analysis of genomics data based on genes unique functional consistency. W. Hu et al. (2017) and Loureiro et al. (2004) offers a methodology for identifying erroneous international trade transactions using hierarchical clustering algorithms. The affinity propagation clustering algorithm was implemented in R programming by Bodenhofer et al. (2011) and Giunchiglia et al., (2007), which provided fundamental and optimised semantic matching algorithms and explored their implementation within the S-Match system for idea matching.



## 2.10 Data Quality Methodologies

There have been many methodologies developed to understand how DQ can be consistently improved. The DQ management operations of the two most common DQ management methodologies, Total Information Quality Managing (TIQM) and Total Data Quality Management (TDQM) are summarized in Table 2.2 while the discussion is provided in subsection below.

*Table 2.2: Summary of TIQM and TDQM*

| TIQM | TDQM |
|---|---|
| • Evaluate information management quality<br>• Data definitions as component requirements<br>• Evaluate for DQ | • Define (features, information product, and information quality)<br>• Measure (IQ metrics for quality control)<br>• Analyze (for errors)<br>• Improve (quality improvement) |

### 2.10.1 Total Information Quality Management (TIQM)

Total Information Quality Management (TIQM) methodology (formally known as Total Quality Data Management / TQDM) is an all-inclusive DQ management methodology with the goal to integrate DQ management and advantageous behavioural patterns into the principles of an organization. It is an integral component of a DQ management framework. The monitoring systems were initially developed for DW, but they also refer to other information systems. It also provides rules on developing a quality control culture for knowledge within an organization and operating procedures to increase understanding of the value of high-quality information to the corporate achievement. This study only focuses on the TIQM operational processes rather than the tools and methodologies to create an organisation's DQ culture. Total Information Quality Management (English, 1999) (cf. English, 1999, p. 70)

The TIQM methodology's operating processes begin by evaluating information management quality and data definitions, that is details concerning the data descriptions, the meanings, appropriate meaning settings, and the appropriate business rules. TIQM sees data descriptions as 'component requirements' of necessary data before the DQ can be evaluated. Therefore, the first phase category of TIQM is to determine the "efficiency of data description and architecture".

### 2.10.2 Total Data Quality Management (TDQM)

Total Data Quality Management (TDQM) is a DQ assessment methodology. It is the basis for many other DQ methodologies. The technique of the methodology can be used in any setting and not be constrained or developed for any reason. TDQM has four key components: define, measure, analyze and improve. The TDQM mechanism

was initially developed for DW projects to ensure that output is sustained to a specified or desirable level. Richard Wang presented TDQM in 1998 as a commodity viewpoint for the complete control of data content. The aim is to establish the information product (IP) life cycle through creation, improvement and comprehension. IP is the output of an information system in this sense. TDQM recommend that an IP map for the whole process be developed.

Information from data is known to have a life cycle in the TDQM technique. The definition of the features, information product, and information quality or DQ requirements are provided in the defining step. The IP features mean that one knows what information product one can concentrate on and guarantee high quality. The information quality or DQ and their study demonstrate Wang's IQ dimensions. To visually see what dimensions have potential problems, Wang proposes to use an IT tool or app. It is important to remember that you cannot specify the IQ requirements if you have not defined the IP characteristics and collected the data. Therefore, it is advised in the defining process to identify the information system from which the data comes to explain the various steps in the data flow.

IQ metrics are developed for quality control in the measurement phase. For example, the sum of missing data about customer addresses in the same customer information database may be designed depending on the DQ measurements. The measures are known as 'simple' metrics. In comparison, "advanced" metrics apply to business rules, which track items over time, for example. An explanation would be the missing values on customer gender. If an attribute model forecasts customer behaviour and one in the model is gender, it would also impact the long term if the sum of missing attributes

correlated with gender varies significantly. These metrics can be added to the current information system as an add-in or entrenched into the system itself if a new system has been installed. Based on the measurement findings, the analysis process takes place. Trying to locate the original cause of the problem at this point. When viewing past instances, we will verify it by generating a dumb customer and seeing if the data is flowing correctly. Another approach is to verify if the information system is functioning, but employees are faulted or are not meeting the instructions. The key is to identify and consider the root of the problem. By identifying the source, this step begins to enhance the results. The two last pieces are deemed to be test stages.

Improvement is the final process. The aim here is to solve the DQ problem and for the improvement of DQ. Improvement of DQ should be made to keep that from happening again. Thus, the analysis must also be performed attentively. The DQ improvement needs to meet the requirements of the organization. The proposed approaches are to build or use a technique of allocating resources for DQ enhancement in the knowledge manufacturing matrix analysis. The latter approach was developed because organizations typically have only small resources or resources for increasing efficiency so that resources are diverted to topics of more excellent quality value. The solution is based on a statistical model, which seeks to reduce the penalty and accuracy costs.

This method suggests that an organization can boost IQ or quality of knowledge by two methods, particularly at the defining point. The organization will either buy or create a new whole information system or use the results to develop and enhance the current information system as it has established the information commodity. It is also

noted that the purchase of a new device is enhanced in the long term when requirements can be added to the system itself, meaning that such IQs are tested before the final data are submitted. Whatever the establishment's direction, it is good to point out that changing IQ customer specifications could mean changing the IQ system.

### 2.11 Data Quality Tools

The quality of data influences its business value, and organizations must ensure high-quality data. To achieve this goal, several DQ tools evolved from literature and experience. We can distinguish them in data preparation tools, delivering data cleaning and validation services (Abedjan et al., 2016; Hameed & Naumann, 2020; Wang & He, 2019), measuring and monitoring tools (Ehrlinger et al., 2019), and general-purpose tools, which combine services of the two former types. Several survey studies evaluated existing solutions regarding their DQ capabilities and weaknesses. In their work, Abedjan et al. (2016) explored the cleaning aspect of DQ tools and evaluated approaches to data error detection. They listed four sorts of issues: outliers, duplicates, pattern violations and rule violations. They observed existing technologies are frequently specialized in a single type but cannot recognize many mistake types.

To overcome this constraint, a combination of tools and algorithms is necessary for acquiring a thorough picture of data flaws and generating superior outcomes. In their study, they further suggest that existing solutions lack a proper interactive dashboard. Since the work of DQ and domain experts is costly, a good user interface is crucial for the success of DQ tools. For instance, the program OpenRefine includes faceting and filtering functions for data exploration, which enables efficient work on big data sets (OpenRefine, 2020). The two commercial programs, Penthao and KNIME, provide

workflow-based user interfaces that make it easier to develop data science pipelines, including DQ aspects (Hitachi, 2020; KNIME, 2020). However, both technologies focus on applying existing DQ knowledge and do not assist in the discovery of new rules or errors.

Ehrlinger et al. (2019) analyzed DQ monitoring and measurement technologies. They offer a complete review of DQ capabilities common in commercial and scientific tools. They differentiate these capabilities into the ones that are more common in tools (e.g., cardinalities, value distributions) and the less common ones (e.g., dependencies, multi-column profiling). Specifically, for multi-column profiling, they show that no tool implements association rule learning, which, according to Abedjan et al. (2015), is a valuable functionality. Generally, they agree with the findings of Barateiro and Galhardas (2005) and argue that DQ problems involving numerous qualities are difficult to recognize and that further research and developments are necessary towards this end.



Furthermore, Ehrlinger et al. (2019) indicate that many DQ tools are not fully utilizing sophisticated machine learning approaches. Although many DQ tools advertise the application of such methods, they rarely give proper documentation, and it remains unclear how and what machine-learning algorithms are implemented. In the case of black-box machine-learning algorithms (e.g., neural networks), one reason for the lack of assistance is the lack of accessible and clearly interpretable results, which are essential "… to prevent a user from deriving wrong conclusions from the results" (Ehrlinger et al., 2019, p. 24). Further research is required on how DQ tools may utilize the advantages of machine-learning methodologies. Both Hameed and Naumann

(2020) and Ehrlinger et al. (2019) underline the necessity for a higher level of automation in DQ tools. Wang and He (2019, p. 812) highlight this issue and argue that automated data error detection is essential for the long tail of use cases and many less valuable data sets, where engaging DQ experts for per-table customization is not viable. This paper describes a data cleaning solution for master data sets, which offers services for mistake detection and DQ rule formulation.

Our approach addresses certain of the above-mentioned shortcomings of existing DQ tools. In particular, it enables more automation by employing machine-learning technologies and provides a more human-centric user interface (Altendeitering & Guggenberger, 2021).



### 2.12 Big Data Quality Rules

With the emergence of BD era, DQ Management has taken a new path. Because specific data, such as stock trades or machine/sensor produced events, is vast and fast, it is necessary to understand this data's properties and correct any errors as early as feasible. Furthermore, as data with varying structures, often from some sources, is combined, it determine data semantics and comprehends connections between characteristics. Unlike traditional DQ management, it is not feasible to specify all of the data's semantics in advance, and there is no global semantics that can suit all of the data. Context-aware DQ rules are required to detect semantic problems in our data and, much better, to fix such errors using DQ rules. From the filthy data itself, we may learn intriguing and informative rules (Taleb et al., 2015).

As a result, we go from the system database's near world hypothesis to an underdeveloped view of the world. Rules are derived from data and updated and verified as additional data is gathered, which is generally dependent on the new data to come. These principles may only apply to specific subsets of data due to the varying structure of data sources (Taleb & Serhani, 2017). Because rules are formed from the dirty data, proper measures must be developed to find stringent rules instead of outliers. Rule violations define data inconsistency. One may be compelled to use these inconsistencies without correcting them or discover methods to correct them based on the applications (Ardagna et al., 2018; Saha & Srivastava, 2014). Fixing must adjust between inconsistent data and incorrect limitations due to intrinsic noise in identifying rules (Ardagna et al., 2018).



### 2.13 Review of Related Studies

Data seldom generate value on its own. They must be connected and merged from a variety of sources, many of which have varying DQ. DQ improvement is a never-ending challenge. High-quality data are required for evaluating and utilising BD, as well as ensuring the data's usefulness. Quality standards and quality evaluation techniques for BD are still missing in extensive study and research. This section summarizes reviews of DQ research related to this study.

Some of the related studies are just reviews of DQ research. At the same time, some propose frameworks and models without implementing the proposed models and/or frameworks, such as in Hamdy and Shaalan (2018), where the authors proposed hybrid methodology for semantic integration fixed mapping issues and adaptation challenges as well as ontology conflicts. Other researchers also followed in this direction are

(Yuan et al., 2019). These authors also did not consider consistency and completeness in their DQ assessment. Others are (Tepandi et al., 2017; Cai & Zhu, 2015; Urrutia et al., 2017; Jarwar et al., 2019).

Furthermore, most of those that implemented their proposed models or frameworks made use of relational datasets. Such studies include Geisler et al. (2016), where the authors applied DQ management in a data stream management system using DQ Ontology. Chika (2015) dealt with inconsistent and incomplete data in a semantic technology setting, identifying and visualizing inconsistent and incomplete data in RDF data using FCA tools and techniques. Some other authors that experimented DQ using relational data are Vaziri et al., (2016), Fürber and Hepp, (2011), Cure, (2012), and Du and Zhou, (2012).

Summary of some of the past studies in the literature can be found in Table 2.3. Table 2.4 shows the benchmark work to this study are recorded. These studies were chosen based on the dimensions of completeness and consistancy. Furthermore, all three research work used different methods for the same data quality dimensions and data type. The benchmark articles include research works by Nath et al., (2017), Fürber, (2015) and Luján-Mora and Palomar, (2001). Detailed discussion on the research works by the benchmark studies are given in the next section.

*Table 2.3: Summary of past studies in the literature*

| S/No | Reference | Method | Data Quality Dimensions | Data Type | Strengths | Limitations |
|---|---|---|---|---|---|---|
| 1 | Wang et al., (2020) | Data Quality Assessment based on Rules | Completeness, Consistency, Timeliness & Duplication | Relational Data | DQ rules in a single-center case study to assess the ability of rules-based DQ assessment to identify data errors of importance to physicians and system owners | DQ was only assessed and not improved |
| 2 | Zheng, (2021) | Contextual Data Cleaning with Ontology Functional Dependencies | Consistency | Relational Data | OFDClean, an algorithm that computes a minimal number of updates to an ontology, and to the data based on dominance | The methodology was limited to only relational data |
| 3 | (Hamdy & Shaalan, 2018) | Hybrid Methodology for Semantic Integration | Completeness, consistency, accuracy and timeliness | Relational Data | Proposed Hybrid Methodology for Semantic Integration. The issues of | Study was mainly literature review of various DQ semantic |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | mapping, the conflicts of ontology and challenges related to adaptation were fixed. | integration. There was no experimental implementation of the proposed model |

Table 2.3 Continued.

| | | | | | | |
|---|---|---|---|---|---|---|
| 4 | (Woods et al., 2021) | Demonstration of the use of the reference ontology in an application-level ontology using an industrial use case | Consistency | Relational Data | An ontology for maintenance activities and its application to DQ | This work is confned to closed environments having systems with relational information. |
| 5 | (Nayak et al., 2021) | Linked DQ Assessment using Ontological Approach | Conciseness | RDF Data | An end-to-end system that uplifts non-RDF to RDF, assesses its quality, and identifies root causes of quality violated triples to improve the quality | No DQ improvement, only suggested |
| 6 | (Yuan et al., 2019) | DQ evaluation methods using Linked Data technology and ontology | Correctnes, rationality and accuracy | Big Data | The detection and evaluation of oil big data quality | DQ was only assessed, not improved. Also, other DQ dimensions such as consistency and completeness were not evaluated |

Table 2.3 Continued.

| | | | | | | |
|---|---|---|---|---|---|---|
| 7 | (Tepandi et al., 2017) | Description of a DQ framework is proposed for the public sector organizations | Completeness, non-redundancy consistency, up-to-datedness, accuracy, compliance, confidentiality, credibility, and once-only principle | relational Data | In order to improve data DQ in public sector, a handbook containing the guidelines is produced | Non-experimental solution |
| 8 | (Wahyudi et al., 2018) | Inductive derivation of process pattern model | Accuracy Objectivity, Interpretability, Consistent, Accessibility, Security and Value-added | Big Data | The literature is enlightned with the process patterns and strategies to enhance the repository of big data. | A single case study in a particular area is used for the derivation of patterns. |
| 10 | (Tute et al., 2021) | Interoperable knowledge-based DQ assessment | and completeness | Relational Data | Applicable concepts and a tested exemplary open source implementation for interoperable and knowledge-based DQ-assessment in healthcare | Describes a method for interoperable and knowledge-based DQ-assessment. The authors didn't go beyond the assessment |

Table 2.3 Continued

| | | | | | | |
|---|---|---|---|---|---|---|
| 11 | (Daneshkohan et al., 2022) | Hands-on experiences in dealing with DQ improvement for the CPCSSN database | Accuracy/precision, completeness/ comprehensiveness, consistency, timeliness, uniqueness and coherence | Relational Data | This article provided an overview of hands-on experiences of healthcare data issues and approaches to improve DQ | The dataset was limited to relational structured and hands-on experiences are just overviews, not implemented. |
| 12 | (Cai & Zhu, 2015) | Formulated a hierarchical data quality framework | Availability, usability, relevance, reliability & presentation quality | Big Data | The feedback mechnisms is introduced for the formulation of a dynamic BD quality evaluation process, which laid a good foundation for further study of the evaluation model. | The framework is limited to only DQ assessment, and no implementation was carried out |
| 13 | (Urrutia et al., 2017) | DQ assessment process using domain ontologies DQ metrics. | Uniqueness, existence & consistency | | Use of the ontology and the possible metrics proposed in a case study that applies to the health care domain | The framework is limited to only DQ assessment |
| 14 | Muhammad Aslam Jarwar, Sajjad Ali, (Jarwar et al., 2019) | Proposed microservices based linked data quality model to assess and improve the energy-related data with defined data quality dimensions and metrics | Accessibility, Consistency and Completeness | Big Linked Data | Design of a microservices based modular and layered architecture. Also, the extension of the W3C Linked Data Quality Vocabulary | There was no experimental implementation of the proposed model |

## 2.14 Benchmark Study

This section explains the studies selected to benchmark the proposed study to determine the performance of the proposed study. The selected studies were used to compare the proposed study and to check if the proposed studies address the weaknesses and limitations of the existing study. This thesis selected three studies for the benchmarking and comparative study. They share the attributes which are common to the work presented in this research such as in technique, methodology and the DQ dimensions. The selected information is extracted from the research articles and is summarized in Table 2.4.

*Table 2.4: Benchmark Study*

| No. | Reference | Technique | Data Quality Dimensions |
|---|---|---|---|
| 1 | (Elouataoui et al., 2022) | Weighted Metrics | Completeness, consistency, timeliness, volatiliy, reliability, readabilty |
| 2 | (Christina Latsou, Marta Garcia I Minguell, Ayse Nur Sonmez, Roger Orteu I Irurre, Martinmark Palmisano, 2022) | Data Quality Ontology | Timeliness, accuracy, completness, consistency, validity, uniqueness, |

## 2.15 Summary

Chapter two gives details of related literature on the essential issues that have to do with this research. The chapter provides a comprehensive survey of DQ definitions, dimensions, methodologies, frameworks and models in the literature, not leaving out the concepts of BD, including semantic technologies and clustering and data transformation models adapted to develop the desired transformation model for BD quality improvement. A summary table identifies the gaps and what is missing in existing DQ methods, models, and frameworks in the literature. Finally, the research technique and model used in this study are highlighted. The following chapter provides the research methodology in more detail to attain the research objectives.

# CHAPTER THREE: RESEARCH METHODOLOGY

This chapter explain about methodology used in the research. This begins from the research approach of the study to the where tasks of each phase are highlighted to enable the achievement of obejectives of the study. The chapter also discussed how some theories are associated with research studies and how the whole methods will be implemented for the benefit and aim of the research in general

## 3.1 Introduction

The methods utilised to fulfil the transformation model's objectives for resolving quality concerns in BD are described in this chapter. The research technique used in this study is based on the information systems domain, including problem awareness, suggestion, creation, and assessment. The details of the research approach in the next section is outlined in 3.2. Furthermore, the researcher's methodology starts with discussing phase I of the study methodology in Section 3.2.1. This, if dealt with, will achieve the first objective of the study. In Phase II of the methodology discussed in section 3.2.2 comes the development and validation of the proposed data transformation model towards achieving the second objective. This is followed by the implementation and the evaluation process in section 3.2.3, where the quality of the BD is to be accessed and improved. This will be achieved to measure the performance evaluation of the transformed data using DQ monitoring and clustering for consistency and completeness in the data. Finally, in section 3.6, the study's cycle is completed by presenting the findings, conveying them as contributions to knowledge, and offering recommendations for future research. Design Science Research Methodology (DSRM) is adopted for this study as discussed in the next session to achieve all of the above.

### 3.2 Research Approach

According to Hevner et al. (2019), behavioural science and design science are the dominant study paradigms in information systems. Behavioural science aims to create and test hypotheses that explain or predict human or organisational behaviour. The design science paradigm aims to push the frontiers of human and organisational capabilities by producing new and creative objects. The development of this study model follows the Design Science (DS) approach by Hevner et al., (2004), and Peffers et al. (2007) because it guides the development of artefacts that are both practice-inspired and theory-ingrained as stated in chapter two coupled with other reasons as stated in same chapter. As an approach, DS provides the right balance between research rigor and relevance in research (Rosemann & Vessey, 2008), which is important given the researcher's aim to develop a data transformation model that aims to help organizations for the improvement of DQ of BD datasets. The study is guided by DS guidelines as shown in Figure 3.1. The artefact is inspired by the lack of completeness and consistency DQ dimensions in BD even with different approaches in the literature.

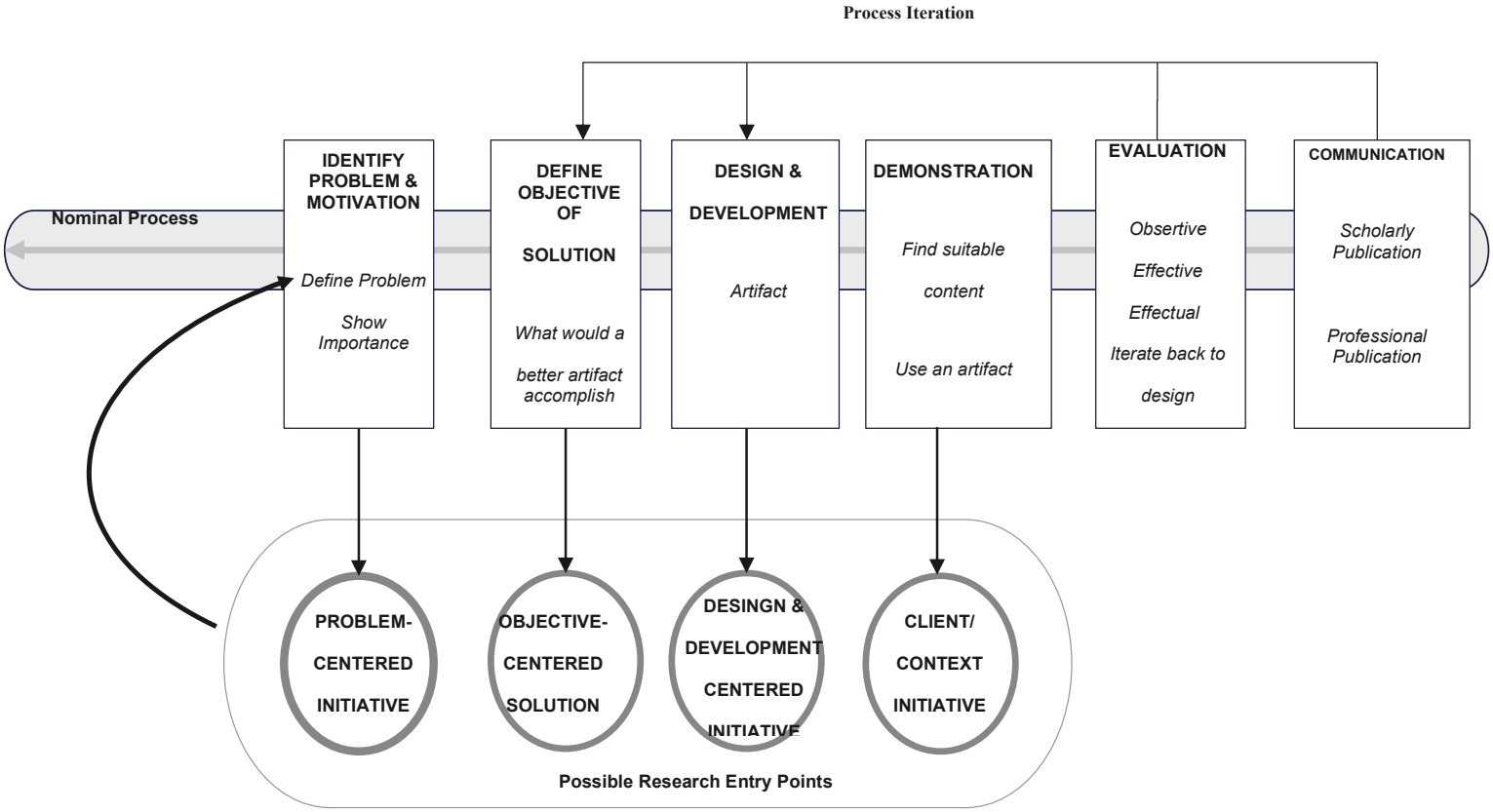



*Figure 3.1: Design Research Methodology. Source: Peffers et al. (2007)*

Figure 3.2 depicts the researcher's framework in accordance with the DSR approach. The phases of the research process are presented in the centre of the diagram, with the specific outputs from each process depicted on the right and the flow of information on the left. This DSR methodology was adopted to bring out the research methodology which is presnted in colored blocks which identifies the phases aligned with the tasks and the tasks are mapped with the deliverables. At each deliverable, the research objective is achieved. The details related to phase, tasks and deliverables are provided in Figure 3.2.

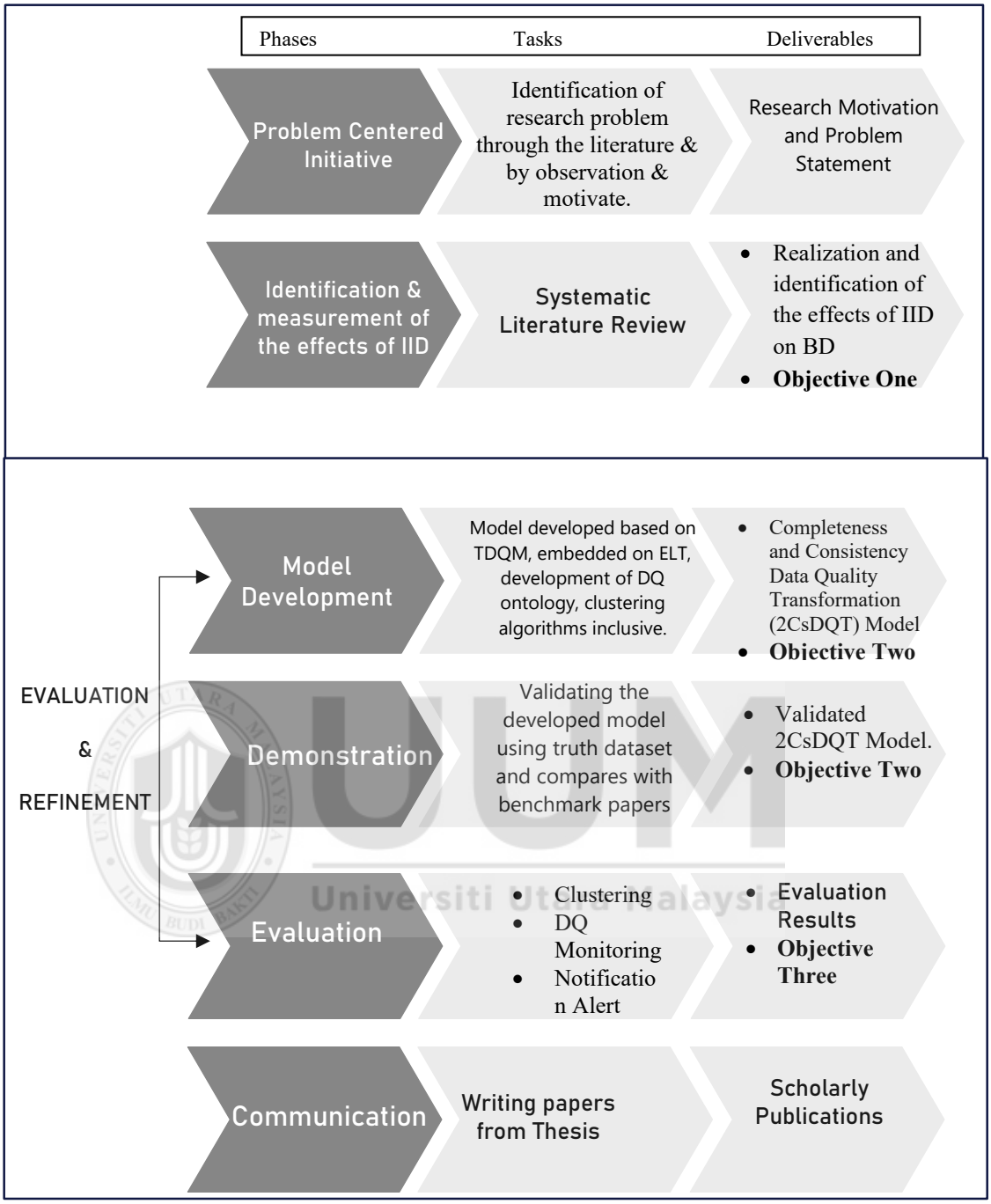



*Figure 3.2: Researcher's Framework integrated into Design Science Research (DSR) Methodology proposed by (Hevner et al. 2007)*

This research methodology is actually based on some established theories called Semiotic Theory. Semiotic Theory is used along with completeness and consistency DQ dimensions as core theories that inform our artefact and create a mapping between these to provide a foundation for the model development. Semiotics is the study of

signs and symbols used to convey meaning to various users. DQ researchers have also adopted the semiotic perspective of data; for instance, Zhang et al. (2019) identified three DQ levels: syntactic, semantic, and pragmatic. The relationship between semiotic levels and structure, data, information, and knowledge facilitate the identification of unique DQ issues.

In the case of this study, the identification of the effects of data incompleteness and inconsistency in BD are based on Knowledge-based View (KBV), and Organizational Information Processing View (OIPV) concepts. KBV views knowledge as associated intangibles – from individual to collective knowledge – to be valuable, unique, inimitable, and non-substitutable resources that are crucial to competitive advantage. The knowledge resource is also dependent on DQ, as it is only with high-quality that BD provide a competitive advantage to organizations (Dwivedi et al., 2011). An organizational Information Processing View (OIPV) relates to the concept of data inconsistency and incompleteness. According to OIPV, efficient data usage necessitates a context-specific composition of information processing processes (Kowalczyk & Buxmann, 2014).

In this regard, DQ management is regarded as one of the essential information processing techniques for businesses. They can aid in the reduction of uncertainty and ambiguity in many decision-making processes. The completeness of knowledge and the interpretation of information may be influenced by DQ levels (Hazen et al., 2014; Krishna et al., 2023).

### 3.2.1 Phase I: Problem Centered Initiative

This phase entails critical literature studies in DQ in BD to better understand the problem and the importance of its solution to gain a better picture of the present situation. This is where the awareness of the unique research problem that prompted this investigation is made known. It is the point at which the BD quality issue is conceptually atomized to convey its complexity. The DSR process usually begins with the awareness of the research problem and rationalising its significance (Peffers et al., 2018). In BD, with the large volume and variety of data, verification of every data input becomes a challenging task. The data becomes untrustworthy if the quality can not be assessed (Ramasamy & Chowdhury, 2020). The deliverables at the end of this phase are the motivation of the study and the statement of the problem.

### 3.2.2 Phase II: Identification and Measurement of the Effects of Incompleteness and Inconsistency on Big Data

This phase which is the first objective to be achieved came with the task of the employing method of a Systematic Literature Review (SLR) to identify the effects of BD quality dimensions. In the past few years, the significance of BD has grown, enabling organizations to gain insights and make data-driven decisions. However, ensuring the quality of BD is crucial as it directly impacts the reliability and usefulness of the derived information. Incompleteness and inconsistencies are two major challenges that can have a substantial effect on BD quality, potentially leading to inaccurate conclusions and misguided decision-making (Smith, 2018).

Incompleteness refers to the absence of certain data elements or attributes within a dataset, which can arise due to various factors such as data collection processes, system limitations, or human error. Incomplete data can hinder the accuracy and

comprehensiveness of analyses, resulting in biased outcomes and flawed conclusions. Thus, it is imperative to identify and measure the extent of incompleteness in big data to comprehend its influence on data quality (Jones & Brown, 2020). Inconsistencies, on the other hand, arise when there are discrepancies or conflicts within the data. These inconsistencies can manifest in different ways, including variations between data sources, contradictory values, or breaches of predefined rules or constraints. Inconsistent data can introduce ambiguity and uncertainty into analyses, compromising the dependability of the derived insights. Hence, it becomes vital to identify and evaluate inconsistencies in large datasets in order to gauge their influence on data quality and the decision-making process (Johnson et al., 2019).

Addressing these challenges requires a holistic approach that involves identifying and measuring how incompleteness and inconsistencies impact the quality of BD. This approach necessitates utilizing a combination of techniques such as data profiling, data cleansing, and statistical analysis to gain a comprehensive understanding of the extent and nature of these issues. By comprehending the effects of incompleteness and inconsistencies, organizations can make well-informed decisions regarding data utilization, implement appropriate data cleansing strategies, and allocate resources effectively to enhance data quality (Miller, 2021).

**i) Scope of the Review**

The aim of this review is to identify and measure how incompleteness and inconsistencies affect the quality of BD. The scope of this systematic literature review is limited to research addressing how to identify and measure how incompleteness and inconsistencies affect the quality of BD in BD systems.

**ii) Methodology**

Systematic literature review to identify and measure how incompleteness and inconsistencies affect the quality of BD according to the guidelines presented by Creswell and Creswell (2017), and Kitchenham (2004). Clearly defining the research objectives is the first stage. In this instance, the goal is to recognize and measure how incompleteness and consistency impact the quality of BD. Following the definition of the research objective, a thorough literature search was conducted to acquire pertinent research papers and publications. I start with keywords like "Big Data" and "Data Quality" before focusing on more targeted phrases like "Incompleteness," "Inconsistencies," and "Big Data Quality." To find pertinent material, academic databases like Scopus, IEEE, Google Scholar, and ResearchGate were used and the number of articles was 41.



Next relevant research papers were sorted, the following selection criteria were established. Include criteria like English-language publications, full-text access to journals, and an emphasis on data quality in BD platforms. 18 papers detail how incompleteness and inconsistencies affect the quality of BD while papers that don't go into great detail about how incompleteness and consistency affect the quality of BD was excluded. Selected research papers were read to find the ones that specifically spoke about how incompleteness and consistency affect the quality of BD. To make sure the papers are in line with the goals of the research by using the review's title as a guide. Detailed notes on the methodologies, findings, and key points discussed in each paper. After reading, the researcher move on to extract relevant information from the selected research articles, paying particular attention to methodology,

measurements, and approaches used to assess the effect of incompleteness and consistency on the quality of BD to synthesize the findings from the data analysis and interpret the results. A comprehensive review and synthesis of relevant literature sheds light on the effects of completeness and inconsistencies on the quality of big data. By examining pertinent studies and articles, we can gain valuable insights into how these factors impact big data quality.

The presence of all pertinent data elements in big data systems is of utmost importance for ensuring completeness. Insufficient data can lead to skewed or biased outcomes, compromised analysis, and hindered decision-making processes. Various metrics and techniques have been proposed to measure completeness in BD, including missing data rates, coverage ratios, and data density analysis. Inconsistencies within big data sets, characterized by contradictory or conflicting information, significantly affect data quality. Such inconsistencies can arise during data integration, merging, or extraction processes. To tackle this issue, researchers highlight the significance of data cleansing and validation techniques in addressing inconsistencies within BD. Approaches like rule-based consistency checks, anomaly detection, and data profiling are suggested for measuring and mitigating inconsistencies. Moreover, researchers have examined how completeness and inconsistencies impact specific domains within big data, such as healthcare, finance, or social media. For instance, in healthcare, the completeness and consistency of patient records directly influence clinical decision-making and patient outcomes. Domain-specific measures and frameworks have been proposed to assess the impact of completeness and inconsistencies within these contexts.

Quantitative and qualitative approaches are employed to measure the effects of completeness and inconsistencies on BD. Quantitative analysis involves statistical

techniques to calculate metrics such as error rates, completion percentages, or inconsistency scores. Qualitative analysis incorporates subjective evaluations, expert judgments, or user feedback to comprehend the practical implications of completeness and inconsistencies in big data.

### iii) The Impact of Inconsistencies and Incompleteness on Big Data Quality

The impact of knowledge inconsistencies and incompleteness in the decision-making process of rational agents (Zhang and Grégoire, 2011). It suggests that the process of generalizing knowledge becomes increasingly challenging as the dimensionality of input data increases (Domingos, 2012). The section also highlights that human knowledge, assumptions, and heuristics may not take into consideration the different probabilistic distributions and temporal and spatial dimensions of input data instances (Zhang, 2013a).



In feature engineering, large raw datasets that are not suitable for learning purposes may be used to construct features, resulting in knowledge inconsistency and incompleteness and reduced predictive accuracy in decision making (Zhang, 2013a). Actionable knowledge, which facilitates the decision-making process, may contain different levels of noise, data, and information (Zhang, 2013a). The section defines knowledge as the connection of influential events and links, encompassing rules and facts that lead to decision-making in specific scenarios. Big Data represents values drawn from structured or unstructured sources, while information provides meaning to data values in specific contexts. Knowledge inconsistency can occur at various granularities of knowledge content, including data, information, knowledge, meta-knowledge, and expertise knowledge (Zhang, 2013b).

Data inconsistency refers to patterns in data that do not conform to established ranges or interpretations (Zhang & Grégoire, 2011). Information inconsistency occurs when it is functionally dependent on other information (Zhang, 2013b). Knowledge inconsistency occurs when known knowledge leads to incorrect outcomes for the same set of conditions, potentially due to violations of consistency rules (Zhang, 2013b). Anokhin and Motro (2001) discuss data value level inconsistencies from a multiple data sources perspective, while Zhang (2013a) focuses on data level inconsistencies within a single large dataset.

*Table 3.1: Impact of incompleteness and inconsistencies in big data systems*

| Aspect | Impact | Reviewed Papers |
|---|---|---|
| Incompleteness | Skewed or biased results | 1. "An Advanced Big Data Quality Framework Based on Weighted Metrics" (2022) by Zhang et al. |
| | Compromised analysis | 2. "Big data quality framework: a holistic approach to continuous quality management" (2021) by Li et al. |
| | Hindered decision-making | 3. "Data Completeness in Healthcare: A Literature Survey" (2017) by Wang et al. |
| Inconsistencies | Contradictory or conflicting information | 4. "Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling" (2021) by Chen et al. |

**iv) Data Quality Issues and problems**

Data is always altered due to many factors. When it needs a quality evaluation and improvement, these factors which include completeness and consistencies. Several factors or processes generated bad data: human data entry, sensors devices readings, social media, unstructured data, and missing values. Laranjeiro, Soydemir, and

Bernardino, (2015) enumerate many reasons of poor data which affect its quality elements and its related dimensions. In Table 3.1, a shortlist of the well-known data issues with respect to consistency and completeness.

Completeness and consistency are widely recognized as important factors for measuring and managing DQ. Within the literature, several commonly cited DQ dimensions are often utilized. The first dimension, completeness, focuses on evaluating the presence of missing values within a dataset. It involves assessing the extent to which all expected data elements are available and accounted for. On the other hand, consistency pertains to ensuring that the data adheres to predefined constraints, rules, or standards. This emphasizes the internal coherence and logical integrity of the data.

*Table 3.2: How to identify and measure inconsistencies and incompleteness in BD*

| | | |
|---|---|---|
| Data Validation | **Checking data against predefined rules or constraints** | **Zhang, P., Chen, Y., Li, D., & Zhang, W. (2019). A survey on data validation in big data era. IEEE Access, 7, 103650-103666.** |
| Data Cleaning | Correcting inconsistent or incomplete data | Rahimian, F., Doan, A., Wang, Z., & Allen, D. (2019). Data cleaning: Overview and emerging challenges. Foundations and Trends® in Databases, 10(1-2), 1-154. |
| Data Integration | Combining data from multiple sources and resolving inconsistencies | Li, X., Zhang, H., Li, J., Zhang, H., & Wu, J. (2020). A survey of data integration: From big data to big integration. Big Data Research, 21, 1-19. |
| Entity Resolution | Identifying and merging duplicate or related entities | Christen, P. (2018). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection (2nd ed.). Springer International Publishing. |
| Data Completeness Analysis | Assessing the extent to which data is missing or incomplete | Khan, S., Ameen, A., Saeed, A., Ahmed, S., & Haider, N. (2020). Evaluating the completeness of big |

data: A systematic review. IEEE Access, 8, 97824-97845.

Based on the systematic review of relevant literature aiming to identify and assess the effects of incompleteness and inconsistencies on the quality of BD as shown in Table 3.2, several key findings can be derived:

1. **Impact on Data Quality**: The review demonstrates that both incompleteness and inconsistencies significantly influence the quality of big data. Incompleteness refers to missing values or attributes within the dataset, while inconsistencies pertain to conflicting or contradictory information. These issues can result in inaccurate or unreliable insights and analyses derived from big data.
2. **Data Cleaning and Preprocessing**: The literature highlights the significance of employing data cleaning and preprocessing techniques to address incompleteness and inconsistencies in big data. Various methods, such as data imputation, outlier detection, and deduplication, are commonly utilized to mitigate these issues and enhance data quality.
3. **Domain-specific Challenges**: The impact of incompleteness and inconsistencies on big data quality varies across different domains. Specific industries or applications may be more prone to particular types of data quality issues. For instance, healthcare datasets may encounter challenges related to missing patient information, while financial datasets may face inconsistencies in transaction records.
4. **Quantifying Data Quality Impact**: Researchers have proposed several metrics and approaches to quantitatively measure the influence of incompleteness and inconsistencies on big data quality. These metrics include measures of completeness (e.g., percentage of missing data) and inconsistency measures (e.g.,

rate of conflict resolution). Such metrics facilitate the evaluation of the severity of data quality problems and the comparison of the effectiveness of different data cleaning techniques.

5. **Data Integration Challenges**: Incompleteness and inconsistencies become more pronounced when integrating multiple heterogeneous data sources in big data environments. The review emphasizes the complexity associated with integrating data from various formats, structures, and quality levels. It underscores the necessity of robust data integration techniques capable of effectively handling and reconciling diverse data sources.
6. **Machine Learning and AI Techniques**: The literature suggests that machine learning and AI techniques can play a crucial role in addressing incompleteness and inconsistencies in big data. Automated approaches for data cleaning, imputation, and validation have shown promising outcomes in enhancing data quality and minimizing the impact of these issues.



In summary, the literature in general underscores the crucial importance of the effects of incompleteness and inconsistencies as there are not much articles dwelling on the effects of incompleteness and inconsistencies to ensure high-quality of BD. It provides insights into the challenges, techniques, and metrics associated with managing DQ issues in the context of BD.

### 3.2.3 Phase III: Development of Data Transformation Model

In this phase, the significant problems in DQ dimensions are further examined by concentrating on two primary dimensions for BD: Consistency DQ dimension, Completeness DQ dimension. The study requires a data transformation model using

the mentioned DQ dimensions. The development of this model is based on Total Data Quality Management (TDQM). TDQM is a DQ management methodology invented by Wang (1998). One of the core ideas of TDQM is the application of the Deming Cycle concept (Gartner & Naughton, 1988) and the Total Quality Management (Martínez-Lorente et al., 1998) approach to his task of DQ management. Like the Deming cycle, the TDQM cycle also consists of four phases: (1) definition phase, (2) measurement phase, (3) analysis phase, and (4) improvement phase. This is presented in Figure 3.3 which is mapped with the mechanism adopted in this research.

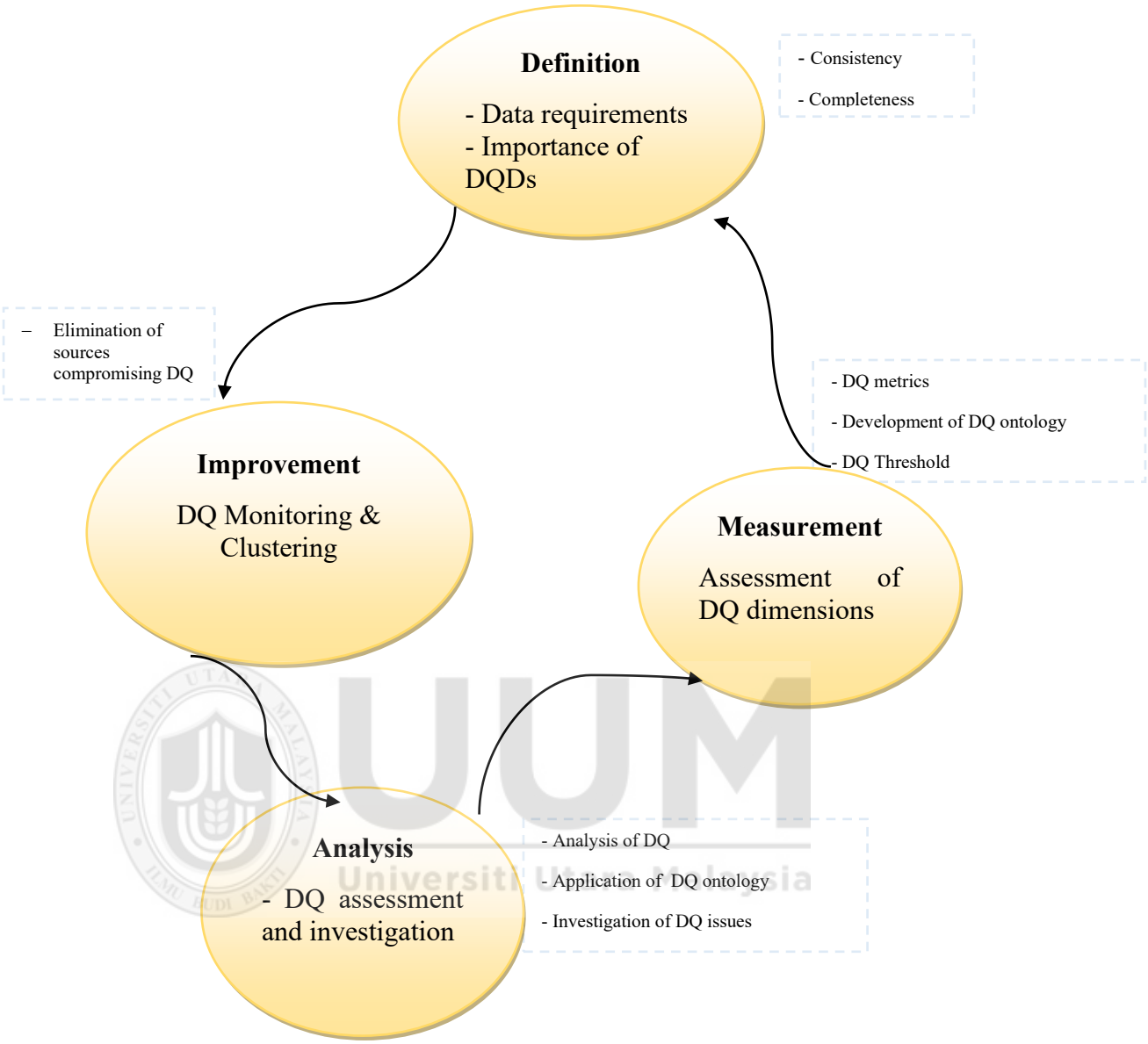



*Figure 3.3: TDQM invented by Richard Wang in 1998 (Wang, 1998)*

In the definition phase, characteristics such as data requirements or rules are captured, and the importance of DQ dimensions are identified (Wang, 1998) which in our case are consistency and completeness. DQ metrics are initially identified in the measurement phase. Remember, as stated in chapter two, DQ metric is a measurement method to assess the extent of the DQ within the dataset. Here, is to establish the initial damage of the lack of DQ. DQ ontology will assist in the assessment. In order to measure the DQ dimension, the metrics are and the developed DQ ontology

implemented in a system and applied to the data. During the analysis phase, the root triggers of the detected DQ issues are investigated based on the measurement result using ontology. Furthermore, from further examination comes the improvement phase where are modified, expanded, or enhanced or the identified sources of quality concerns must be eliminated.

The proposed model is embedded in BD's Extract Transform and Load (ETL) process since it's a transformation model, as shown in Figure 3.4. The initial model is presented in this Section, which is divided into three stages namely, Extract, Transform and Load. Historically ETL procedures and systems have been characterized as integrated into a central repository from several heterogeneous sources referred to DW (Laranjeiro et al., 2015; Debroy et al., 2018). Each ETL system typically consists of data extraction from a source or mapping or transformation of the data in a possible differing format (inclusive of filtering data into what is to be transferred), and the insertion of the transformed data into a separate system. Whether the sources are heterogeneous or the repository represents an actual DW (or some physically or conceptually based storage system) (Debroy et al., 2018). In this case, Extract, Load and Transform (ELT) is mainly applied in the BD environment.

The introduction of Hadoop Technology and decreasing hardware and storage facilities costs have made this more prominent. Data are retrieved from several sources, but the raw data is placed into a database instead of initially transformed. Often, the loading process requires no schema (i.e., schema independent), and in the repository, the data might be unpolished for an extended period (typically in archives). If necessary, specialists create a schema and transformation of the data takes place.

ELT can save raw data for an extended period inside the storage facility that allows experts, without previous practises by experts, to alter the stored data according to their respective tasks. This study will follow the improved ELT process as shown in Figure 3.4

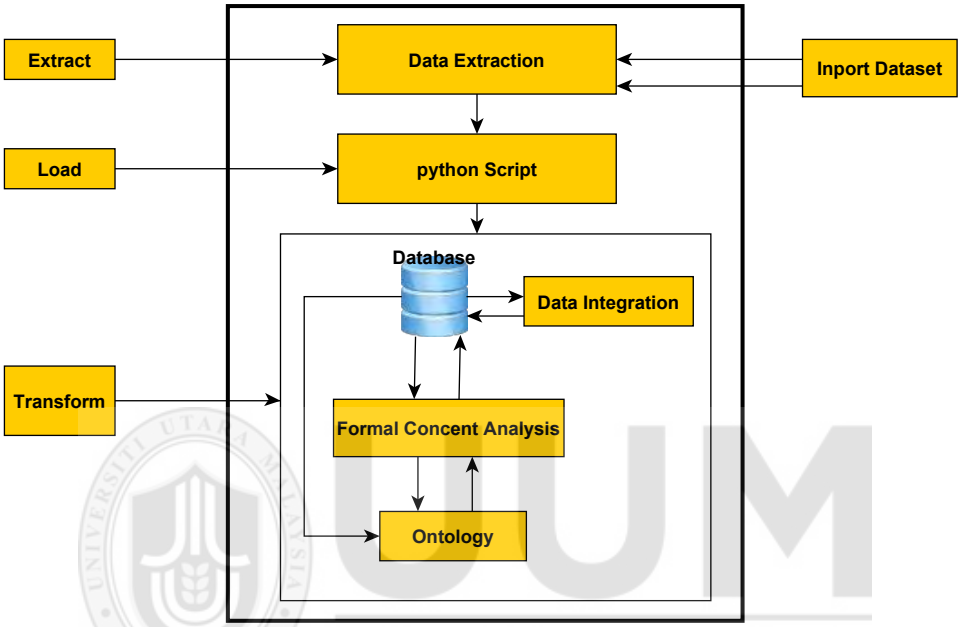



*Figure 3.4: Initial ELT Model Architecture*

### 3.2.4 Phase IV: Validation of the Model

According to Hevner et al. (2004), a developed model is effective when it satisfies the requirements and constraints of the problem it was meant to solve. Venable et al. (2012) further indicates that DS artifacts can be evaluated for (at least) three main reasons, viz: rigor, efficiency and ethics. In this study, we focus on evaluation of efficiency in terms of the model's effectiveness. Specifically, the researcher is to conduct demonstration and evaluation using a set of data to validate the efficiency of the model. The model is therefore to be validated by the said dataset to ensure its utility and applicability in other domains. The validation allows the researcher to reason

about the model's applicability in textual BD settings in general since our study is based on textual BD.

### 3.2.5 Phase V: Evaluation of the Model

After the development and validation of the model comes its evaluation using precision and recall method and benchmark studies. The concepts of recall and precision from the field of information retrieval are used in DQ management to evaluate the transformation model. They are used as indications for the performance of this approach (Batini & Scannapieco, 2016a). Details of the implementation of the precision and recall method and of the benchmark studies comparison are given in chapter five which is the chapter dedicated for the study evaluation which will be the achievement of the third objective of this study.

## 3.3 Summary

This chapter outlines the research methodology and techniques that was utilised and how it was applied in this study. The chosen research design is described and justified and mapping the design research approach to this study context. Furthermore, theories applicable to this research studies are discussed and finally, the chapter discusses different phases, task and deliveries for the achievement of the research objectives.

# CHAPTER FOUR: COMPLETENESS AND CONSISTENCY DATA QUALITY TRANSFORMATION MODEL

In this chapter, based on the results of the systematic literature reveiew, a DQ model is developed. A detailed study of current information in the realm of DQ dimensions and ontology design was carried out in the preceding chapters. This is utilised as the basis for the completeness and consistency DQ transformation (2CsDQT) model. The discussion about DQ transformation model ends with the summary of the chapter.

## 4.1 Introduction

Though substantial research has been done in the DQ domain, not much work has been done in the improvement stage, especially in the BD domain consisting of the semi-structured and unstructured data called the NoSQL data. After the proposed model development, it was validated using a truth data set as the processes and steps identifying inconsistency and incompleteness in integrated data are highlighted using metrics, threshold and DQ ontology design. This chapter provides insight about the implemention of methods, tools, and techniques used to achieve the purpose of the proposed model development and validation which is the core of this research. Nevertheless, before delving into model development proper, systematic literature review to achieve objective number one is done according explaination in section 3.2.2, chapter 3.

## 4.2 Proposed Model

Adapting the general architecture for extracting, loading, and transforming (ELT) as shown in Figure 3.3 chapter 3, the proposed data transformation model is an extension of the transformation stage of the existing ELT. Figure 3.4 shows a more

comprehensive approach in the transformation process of the ELT processes from data integration phase and identification of violation of the DQ dimensions (inconsistency and incompleteness) phase and then the development of the ontologies – both the integration and the DQ ontology to the clustering phase. The 2CsDQT data transformation model, in brief, contains the data extraction phase, data gathering phase, data preparation, DQ violation identification phase, and the DQ improvement phase.

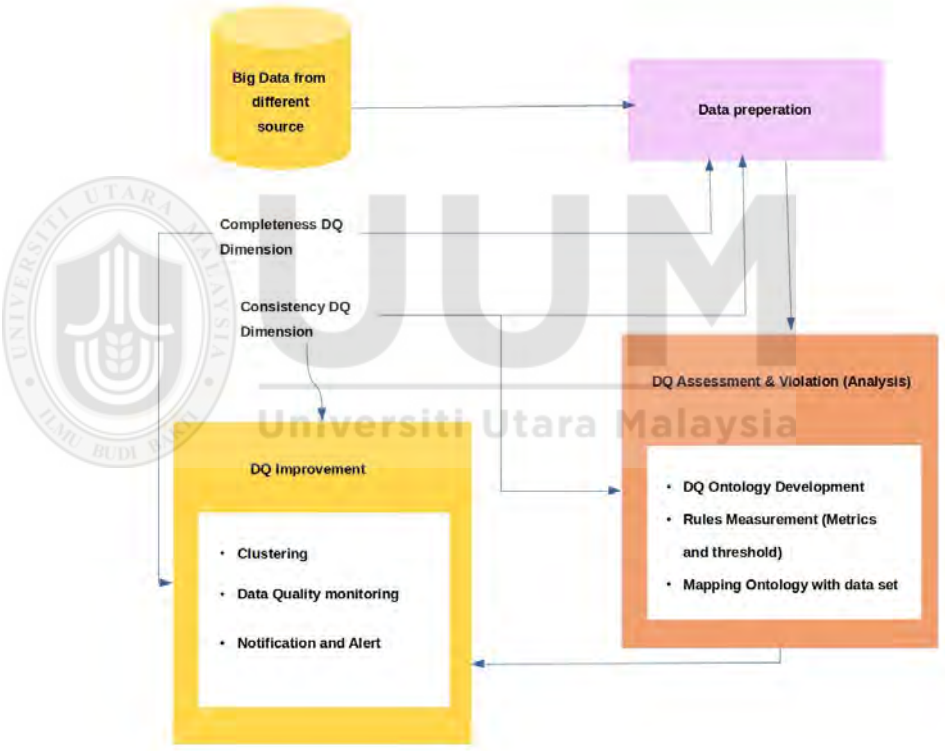



*Figure 4.1: Completeness and Consistency Data Quality Transformation Model*

Based on knowledge from the literature reviews in previous chapter concerning the effects of DQ dimension violations and with the previous technique not being able to meet up with the demand of BD and its quality. Therefore, the researcher came up with a proposed model called Completeness and Consistency Data Quality Transformation

Model (2CDQTM) as shown in Figure 4.1. The model is an enhancement of the "Transform" phase of the initial architecture in Figure 3.4 chapter 3.

From the Figure 4.1, data collected or extracted from different sources of BD are being integrated and preprocessed. The data sets are integrated and preprocessed ready for transformation. The core aim of this step is to ensure that all data set within same domain are centrally kept. This also helps to identify inconsistencies and incompleteness within the data. This further leads to the DQ assessment and violation phase where DQ ontology development, DQ requirements etc are showcased. The phase intensify the identification of the quality violation. Next is the DQ improvement phase where the clustering algorithm is to be applied applied to the result of the violations to ensure the completeness and consistency of the data set. The following subsections discuss the different components of the proposed model for this study and its implementation steps.



### 4.2.1 Component 1 – Data Gathering

Component 1 explains the data types and sources that are adopted and extracted for the proposed study. The data sources used in this study are XML, JSON and CSV. The model is designed to allow it to be easily extended to support other formats as well. A comma-separated values (CSV) file allows to save data in a table-structured format. Each row of the table is represented by a single file line, consisting of one or more fields separated by commas. Datasets are extracted from the different data sources after which their integration commences.

### 4.2.2 Component 2 - Data Preparation

Data preparation is the process of gathering, combining, structuring and organizing data so it can be used in business intelligence, analytics and data visualization applications. The components of data preparation include data preprocessing, profiling, cleansing, validation and transformation; it often also involves pulling together data from different internal systems and external sources merging the different BD set of same domain from different sources into only a source. It is easier to perform transformations on a single or centralized data set in a domain than when they are scattered in their different sources. To help achieve this purpose, data preparation can be done using pandas dataframe in python which is usually seamless. This can also be done using tools such as OpenRefine.

### 4.2.3 Component 3 – Data Quality Assessment and Violation

The component 3 deals with the process and analysis of the integrated dataset to identify DQ violations. This is where a DQ ontology is developed to enable the detection or identification of the DQ violations specifically based on completeness and consistency DQ dimensions. The developed DQ ontology is to be mapped with the integrated data set and violations are enabled base on the rules/requirements, that is, metrics and threshold set for the model and contained in the DQ ontology. The report of the DQ violated is reoprted for the implementation of the next phase, that is the improvement phase.

### 4.2.4 Component 4 – Data Quality Improvement

In the DQ improvemet phase is DQ monitor which monitors violations of stated rules and noifications are sent as alerts  and K means clustering is employed to improve the

DQ based on the violations. Clustering is an unsupervised machine learning task which involves automatically discovering natural grouping in data. Clustering technique is applied after the identification of incompleteness and inconsistencies within the data. These DQ violations also includes the findings of the previous phase.

### 4.3 Proposed Model Initial Requirements

The aim of any DQ management activity is to improve the quality of data in an organization. In order to achieve this goal, a method for the identification and removal of the causes of DQ issues is needed. Among the description of some DQ methodologies given in chapter two, the two most popular DQ dimensions to improve DQ are Completeness and Consistency. The common activitities identified for the two methodologies include (Batini & Scannapieco, 2016a):

- Definition and identification of quality-relevant requirements and metadata,
- Information quality assessment and measurement,
- Analysis of the root causes of identified DQ problems, and
- Resolution of the identified root causes

Organizations in need of DQ require the implementation of a DQ management methodology. Using the findings from the comparison of TDQM and TIQM to define a DQ management process that fits an organization, the DQ management contains the following subtasks:

**Data Requirements Formulation**: Formulating data requirements specific to completeness and consistency. Additionally, a common language and structure is used to formulate the requirements.

**Assessment of Data Sources' Quality State**: Transparency regarding the quality condition of a data source should be generated based on the data requirements. During the analysis phase, the detected requirement breaches are investigated to determine the root reasons for the requirement violations.

**Identifying and Eliminating the Root Causes of Requirements Violations**: The fundamental causes of requirement violations are discovered and eliminated using the requirement violation reports. Finally, the requirement breaches are rectified in the improvement phase to rebuild the state in accordance with the requirements.

Thus, data requirements management is one of the most critical and essential aspects of DQ management. Throughout the management cycle, it is used to formally define the intended condition of data. To put it another way, data requirements are the knowledge of the characteristics of high-quality data (Fürber, 2015). The description of the requirements of the 2CsDQT model is separated by functional and research requirements which are presented as follows:

### 4.3.1 Functional Requirements

Functional requirements are those that specify an artefact's desired functionality (Grande, 2011). Because the artefact's functions must enable the execution of the defined tasks, functional requirements of 2CsDQT may be inferred from the task requirements. The job may be used to produce the following functional requirements.

**Formulation of Data Requirements**: The requirements are formed based on completeness and consistency DQ dimension requirements. The follwoing functions are required for the identification of Requirement Violations:

- Use the formulated data requirements to identify requirement violations in the tested data
- Generate a report with violated instances indicating the type of violation specific to completeness and consistency DQ dimension.

**Evaluation of the Quality State of Data Sources**: Generate a report that shows the ratio between correct instances and instances with requirement violations of completeness and consistency DQ dimension.

### 4.3.2 Research Requirements

2CsDQT also covers research requirements that must be considered to fulfil the research aim or induce the research settings under which the artefact is created. Since this study uses ontology to identify violations and pre-processing with clustering to improve DQ, one or more ontologies are part of the model. Table 4.1 provides the summary of requirements to develop data transformation model.

*Table 4.1: Summary of the initial requirements for the creation of a data transformation model*

| Requirement ID | Requirement | Requirement Type |
|---|---|---|
| 1 | Data is captured in a machine-readable form | Functional |
| 2 | The form of data can be used to test data automatically, and it shows violations. | Functional |
| 3 | Create a report listing instances of violations and specifying the type of violation. (data incompleteness or inconsistency) | Functional |

| | | |
|---|---|---|
| 4 | Data extraction from BD sources | Technological |
| 5 | Integration of data from multiple sources | Technological |
| 6 | Performance and scalability | Technological |
| 7 | System constraints | Technological |
| 8 | Use ontologies | Research |
| 9 | Data clustering | Research |

### 4.4 Data Gathering

The dataset used for the research work was gotten from reputable data sources New York City agencies (https://opendata.cityofnewyork.us/), which include data.gov.us (USA), data.gov.uk (UK and social media data gotten by the use of API that allows extraction of data from their platforms). A ground truth dataset was used for evaluating/validating the DQ model. A ground truth dataset is a reference dataset that serves as a standard or authoritative source of information for validating and evaluating the performance of DQ models or natural language processing models (Arhin et al., 2021). Also, data set for comparism was gotten from the benchmark paper articles.

### 4.5 Data Preparation

Initially, this step starts early in any data transformation task. In this model, understanding BD requirements begins by gathering the user requirements through traditional and commonly used dataa preparation techniques are data cleaning, transformation, normalization, and others.

#### 4.5.1 Data Preprocessing

The data set was prepared using Python and OpenRefine. Preprocessing data using Python and OpenRefine involves a combination of programming and Graphical User Interface (GUI) tools. Python offers a wide range of libraries for data manipulation

and analysis, while OpenRefine is a powerful GUI tool for cleaning and transforming data. The cleaning process as shown in Figure 4.2 is done using OpenRefine tool to prepare the data for next step of data transformation.

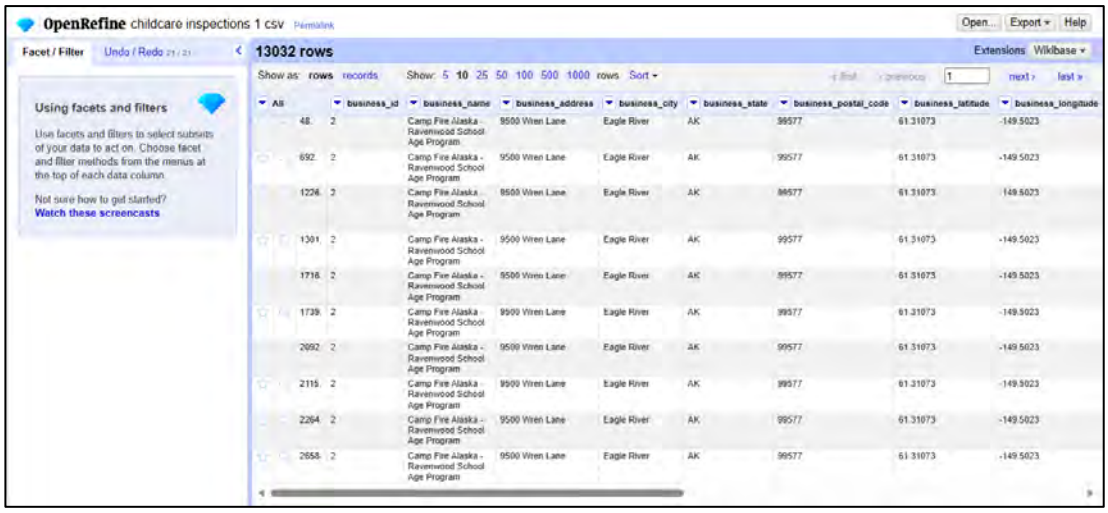



*Figure 4.2: Data Cleaning Process*

By combining Python for more advanced data manipulation and analysis with OpenRefine for interactive and visual data cleaning, this helps to effectively preprocess and clean data for various analytical and modeling tasks. Based on the complexity of the data and preprocessing requirement both of these tools were used for data preprocessing pipeline.



### 4.6 Data Quality Assessment and Violation

The aim of any DQ management activity is to improve the quality of data in an organization. Thus, data requirements management is one of the most critical and essential aspects of DQ management. Throughout the management cycle, it is used to formally define the intended condition of data. To put it another way, data requirements are the knowledge of the characteristics of high-quality data (Fürber, 2015). To achieve this, DQ rules or requirements ought to be defined or formed. This has to do with data requirements formulation. Formulating data requirements base on

the data domain and in this study specific to completeness and consistency DQ dimensions. The requirements are embedded in the DQ ontology and can also be formed during the implementation of the model specific to a data domain.

### 4.6.1 Ontology and Resource Description Framework

It is critical to capture the semantics of data at the conceptual level. Because of the following reasons, the proposed model is developed using ontological notions. Ontologies, firstly, enable the semantic integration of heterogeneous data sources. We may express clearly how two ontology concepts are structurally related and linked to distinct properties. Secondly, it's machine-readable, which means we can automate ELT modelling operations. Thirdly, the DW is simpler to develop than alternative representations. Some standard languages, such as RDFS or OWL, can describe an ontology. They both offer fundamental constructs for defining the ontology's formal semantics. OWL uses owl: Class and rdfs:Property to define the ontology's concepts and properties, respectively. Rdfs:subClassOf and rdfs:subPropertyOf are used to define hierarchical relationships between concepts and properties, while rdfs:domain and rdfs:range correlate properties with concepts.

A shared conceptual data model for collecting datasets, quality rules, and quality metrics is one of the fundamental prerequisites for the suggested solution. In the context of databases, such shared data schemas are referred to as global or mediated schemas (Nguyen et al., 2018) (Cipolla, 2019) or ontologies (Nguyen et al., 2018) (Dos Reis, 2018) in relation to the Semantic Web, intelligent systems, agents and knowledge representation, etc.

### 4.6.2 Data Quality Ontology Development and Requirement

The DQ vocabulary described here is an ontology that offers a uniform data structure for storing quality-related knowledge so that generic SPARQL queries may analyse it and discover gaps and inconsistencies in data instances. Some of the DQ ontologies in the literature are listed in Table 4.2 below. The Semantic Web Publishing Vocabulary (SWP51) and the Open Provenance Vocabulary are two ontologies realy used to express DQ information of data in Semantic Web systems (OPV52). Table 4.2 also illustrates the present Linked Open Vocabularies in terms of quality and trust Vandenbussche et al., (2017), a website that maintains a list of open vocabularies the Semantic Web. While most of the ontologies listed above may be expressive enough to describe certain quality information relevant to DQ dimensions such as timeliness, Ruiz-Benito et al., (2020), they lack expressiveness in representing various sorts of data requirements, such as legal property values or functional dependencies which expresses data consistency except for DQM vocabulary by (Fürber, 2015).

*Table 4.2: Ontologies in the DQ space of Linked Open Vocabularies*

| **Prefix** | **Namespace** | **Title** |
|---|---|---|
| Cert | http://www.w3.org/ns/auth/cert# | The Cert Ontology |
| Dqm | http://purl.org/dqm-vocabulary/v1.1/dqm# | The Data Quality Management Vocabulary |
| Irw | http://www.ontologydesignpatterns.org/ont/web/irw.owl# | The Identity of Resources on the Web ontology |
| Opmv | http://purl.org/net/opmv/ns# | Open Provenance Model Vocabulary |
| Pav | http://swan.mindinformatics.org/ontologies/1.2/pav/ | Provenance, Authoring and Versioning Ontology |
| Prov | http://purl.org/net/provenance/ns# | Provenance Vocabulary Core Ontology |
| Prvt | http://purl.org/net/provenance/types# | Provenance Vocabulary |
| Voag | http://voag.linkedmodel.org/schema/voag# | A vocabulary of Attribution and Governance |
| Wot | http://xmlns.com/wot/0.1/ | Web of Trust |

The Data Quality Ontology (DQO) in this study has been meticulously crafted with a strong focus on the DQ dimensions of completeness and consistency. It leverages standard metrics and well-defined thresholds to ensure that data requirements, DQ assessment results, data cleansing rules, and data requirement violations are represented comprehensively and in a manner that maintains data consistency. This ontology serves as a robust foundation for enabling DQ monitoring, assessment, and cleansing within Semantic Web architectures, all while upholding the rigorous standards of completeness and consistency through the use of established metrics and thresholds. The DQO was designed using Protege OWL. The formulated data requirements are used to identify requirement violations in the tested data and a report generated with violated instances indicating the type of violation specific to completeness and consistency DQ dimension. Finally, the requirement breaches are rectified in the improvement phase which is the final phase to rebuild the state in accordance with the requirements.



A shared conceptual data model for collecting datasets, quality rules, and quality metrics is one of the fundamental prerequisites for the suggested solution. In the context of databases, such shared data schemas are referred to as global or mediated schemas (Nguyen et al., 2018; Cipolla, 2019) or ontologies (Nguyen et al., 2018; Dos Reis, 2018) in relation to the Semantic Web, intelligent systems, agents and knowledge representation, etc. The DQ vocabulary described here is an ontology that offers a uniform data structure for storing quality-related knowledge so that generic SPARQL queries may analyse it and discover gaps and inconsistencies in data instances. Figure 4.5 shows the DQ ontology graph was created in protégé-OWL. Also, screenshots of the classes and object properties are displayed as shown in the Figure 4.3 and 4.4.

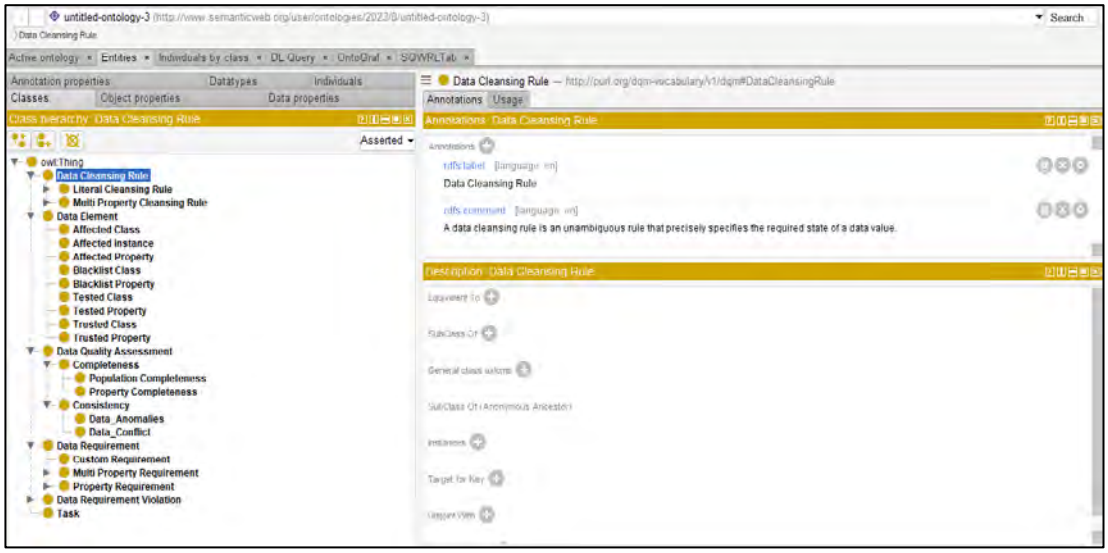



*Figure 4.3: Data Ontology classes*

A class is a collection of individuals or objects that describes concepts in the domain (e.g., data quality assessment). A class can be defined either by extension (specifying members) or by intension (specifying conditions), which are sets of collections of various objects are termed as classes (Ta'a et al., 2011). One of the classes are: Data Cleansing Rule which has subclasses as LiteralCleansingRule and MultiplePropertyCleansingRule, which also has their subclasse. The next is Data Element class with subclasses such as AffectedClasses, AffectedInstance, AffectedProperty, BlacklistClass, BlacklistProperty, TestedClass, TestedProperty, TrustedClass and TrustedProperty and so on and so forth.

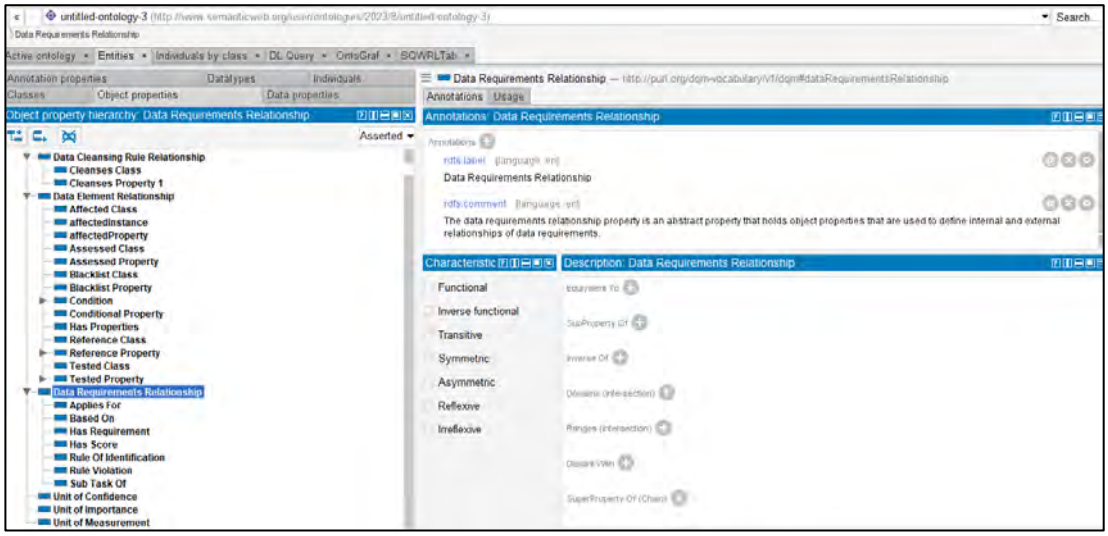



*Figure 4.4: Data Ontology Object Properties*

In OWL, object properties are used to assert relationships between individuals (or instances). Object properties in OWL can have characteristics such as being transitive *or* symmetric. We can assert additional information about properties such as their domain and range, along with defining inverse properties (Dimassi et al., 2021).

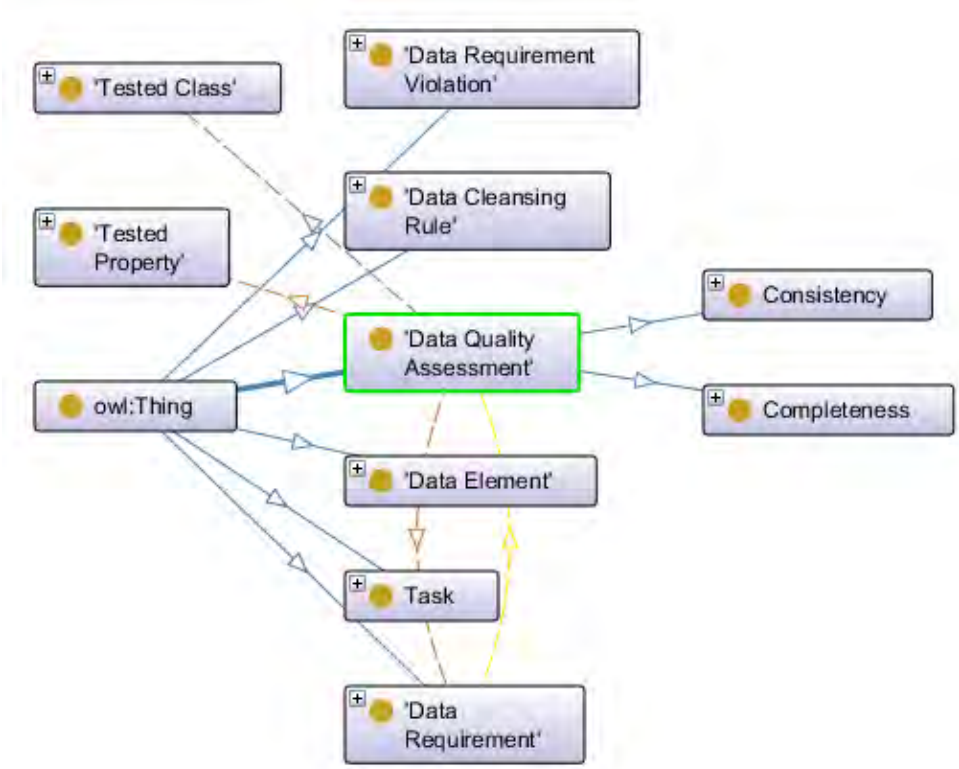



*Figure 4.5: Data Quality Ontology with Set Requirement*

This study DQ ontology is created in protégé-OWL. It comprise of classes, object properties, and datatype properties, and it's written in OWL DL to make it easier to use in knowledge bases that rely on defensible reasoning. The DQ ontology has the following primary functions:

(i) Machine-readable representation of data needs.

(ii) Data components are annotated with quality-relevant meta-information.

The class Data Requirement is the superclass of all data requirements and includes common properties shared by all data requirements. Data Requirement is a crucial class in ontology since it is the foundation of all DQ management. The object property applies because data requirements may vary depending on the job (task-dependent). There are keywords for each task. For example, “for” can be used to link a specific

requirement to a Task class. This allows task-dependent data requirements to be filtered based on particular tasks. It also aids in identifying the activities that may be impacted if the data requirement is not met. Among the major classes of the DQ ontology are:

1. **Data Requirement**: The class Data Requirement is a crucial class in ontology since it is the foundation of all DQ management. A data requirement is a written directive or contract that specifies the content and/or structure of high-quality data instances and values.
2. **Data Element**: This is a class, a property, an instance or a literal value. These elements are the building blocks of data, and their quality directly impacts overall DQ. Within the category of data elements, it's important to delineate specific aspects, including the affected class, impacted instances, affected properties, and the inclusion of blacklist subclasses
3. **Data Requirement Violation**: This arises when a data value or a data instance fails to satisfy its requirements. Within the realm of data elements, there are subclasses to consider, each with its own unique characteristics. These subclasses encompass duplicates instances, functional dependencies, illegal values, missing elements, out-of-range values, and syntax violations. Additionally, we must specify attributes such as the affected class, impacted instances, affected properties, and the inclusion of blacklist classes, as these all contribute to a comprehensive understanding of DQ.
4. **Data Quality Assessment:** Within this class, DQ assessments are conducted, with a particular focus on two vital dimensions: completeness and consistency. Completeness assessment evaluates the extent to which data meets defined requirements, while consistency assessment ensures that data is coherent and free

from contradictions. DQ dimensions scores that reflect the quality status of classes and/or properties may be structured using this class.

5. **Data Cleansing Rule:** This is a definite rule that stipulates a data value's required state. This class represents rules and guidelines employed to improve the quality of data by addressing issues such as errors, duplicates, and missing values and blank rows.
6. **Task:** The task class categorizes the task for which DQ requirement is applied.

*Table 4.3: Forms by 2CsDQT's Data Requirements*

| Serial No. | Form | Purpose |
|---|---|---|
| 1 | Property requirements | Identify data requirements related to single attributes. |
| 2 | Conditional requirements | Capture data requirements that are valid for a specific subset of instances of a class. |
| 3 | Duplicate rules | Identify the data requirements that apply to a particular subset of a class's instances. |
| 4 | Functional dependency reference rules | Capture data requirements that refer to a trusted data source to identify functional dependency violations. |
| 5 | Custom requirements | Capture data requirements that are not expressible with the above forms. |
| 6 | Tested Classes | Register classes with instances that shall be analyzed for DQ problems. |
| 7 | Tested Properties | Register properties that shall be analyzed for data quality problems. |
| 8 | Conditions | Define conditions that shall be used for conditional requirements to filter a relevant subset of a class. |
| 9 | Trusted Classes | Register classes of another data source as a trusted reference for legal value rules and functional dependency reference rules. |
| 10 | Trusted Properties | Register properties of another data source as a trusted reference for legal value rules and functional dependency reference rules. |

In order to discover requirement breaches and compute DQ scores, instances of the class Data Requirement and its subclasses are utilised. As a result, further examples for Data Requirement Violation and Data Quality Assessment can be derived from these instances. The class Data Requirement Violation is used to annotate instances that violate data requirements with information on when they were discovered, the classes and properties affected, and the data requirement that detected the violation. The results of DQ evaluations may be saved in the Data Quality Assessment class. The class then offers properties to specify when the assessment was performed, the measurement's requirement, the classes and properties examined, the actual score, and its unit. The class Data Element and its subclasses provide the range for the classes Data Requirement, Data Requirement Violation, and Data Quality Assessment. As a result, every class and property used in a Data Requirement instance must either be a direct instance of one of the Data Element's subclasses or be mapped to one of its instances through its properties. In the first option, the knowledge base is transformed and becomes OWL Full in the former. A visualization of part of the DQ ontology is shown in Figure 4.5.

Pseudocodes showing the mapping of the DQ ontology and datasets using python script. This pseudocode describes a methodical way to evaluate DQ using ontology-based approaches, which applies to all the datasets used in the study. It includes the crucial processes of loading data, establishing quality thresholds, mapping columns to DQ dimensions, calculating metrics, and reporting results. DQ ontology is developed to enable the detection or identification of the DQ violations specifically based on completeness and consistency DQ dimensions. The developed DQ ontology is used to be mapped with all the BD sets in the study and violations are enabled base on the

rules or requirements, that is, metrics and threshold set for the model and contained in the DQ ontology.

### 4.7 Data Transformation Model Validation

The model is validated based on use of an intelligent algorithm, set of rules in DQ ontology and ground truth BD set and data sets from benchmark paper articles from literature.

#### 4.7.1 Model Validation Criteria

The DQ ontology has defined what is consider to be valid or acceptable DQ for the dataset in terms of data completeness and consistency. The ontology may already specify some criteria, but other parameters was put to customize them to match the specific needs of the dataset and organization such as threshold and metrics. DQ ontology is applied to the integrated dataset. This process identifies the 2Cs problem in the dataset, which affects the DQ.



#### 4.7.2 Selection Validation Metrics

Appropriate metrics or measures for each DQ dimension defined in the ontology. These metrics are quantifiable and objective, allowing for assess to the DQ accurately. The chosen metrics are for completeness and consistency. DQ dimensions.

#### 4.7.3 Data Preparation

**Collection of Validation Data**: For validation purposes, a separate dataset (a ground truth dataset) that is considered to be of high quality was used. This dataset was used to compare and validate the results of the model.

**Description of Data Set for Model Validation**

Truth data set is said to be a data set that is free of DQ issues. In this study, a truth data set is a data set that is complete and consistent. This type of data set is suit for model validation to ensure the developed model is effective and efficient. The data set used for the model validation is COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths Data set gotten from (https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3mention sources). It is found to be complete and consistent.

**COVID-19 Daily Counts of Cases, Hospitalizations, and Deaths Data set**

This data set was Created on June 9, 2020 and last updated on December 2, 2020. This dataset represents citywide and borough-specific daily counts of COVID-19 confirmed cases and COVID-related hospitalizations and confirmed and probable deaths among New York City residents. Counts are aggregated separately by the following dates of interest:



- Cases are by date of diagnosis (i.e., date of specimen collection),
- Hospitalizations are by date of admission
- Deaths are by date of death

To address variation in cases diagnosed per day, a 7-day average (mean) of daily confirmed case counts is included. This is calculated as the average of number of confirmed cases diagnosed on the current day and the previous six days. Additionally, this dataset includes a 7-day average of COVID-related hospitalizations, calculated as the average number of COVID-19 patients who were hospitalized on the current day and the previous six days, and a 7-day average of confirmed deaths, calculated as the average of number of deaths occurring among confirmed COVID-19 cases on the

current day and the previous six days. The dataset includes citywide and borough-specific 7-day averages of cases, hospitalizations, and confirmed deaths.

This file includes data since February 29, 2020, which is the date that the Health Department classifies as the start of the COVID-19 outbreak in NYC as it was the date of the first laboratory-confirmed COVID-19 case.Data on confirmed cases were passively reported to the NYC Health Department by hospital, commercial, and public health laboratories. In March, April, and early May, the NYC Health Department had discouraged people with mild and moderate symptoms from being tested, so data from those months primarily represent people with more severe illness. Until mid-May, patients with more severe COVID-19 illness were more likely to have been tested and included in these data. Data on hospitalizations for confirmed COVID-19 cases were obtained from direct remote access to electronic health record systems, regional health information organization (RHIOs), and NYC Health and Hospital information, as well as matching to syndromic surveillance.



**Pseudocodes Showing the Mapping of the Data Quality Ontology and Datasets using Python Script**

This pseudocode describes a methodical way to evaluate DQ using ontology-based approaches which applies to all the data sets used in the study. It includes the crucial processes of loading data, establishing quality thresholds, mapping columns to DQ dimensions, calculating metrics, and reporting results. DQ ontology is developed to enable the detection or identification of the DQ violations specifically based on completeness and consistency DQ dimensions. The developed DQ ontology is to be mapped with all the BD sets used in the study and violations are enabled base on the

rules/requirements, metrics and threshold set for the model and contained in the DQ ontology.

**Step 1**: Load the Data Quality Ontology (e.g., in RDF format)

ontology = LoadDataQualityOntology("data_quality.owl")

**Step 2**: Load the Dataset (e.g., in CSV format)

dataset = LoadDataset("dataset_name.csv")

**Step 3**: Initialize a Mapping Dictionary to Store Mappings

mapping = {}

**Step 4**: Define Thresholds

completeness_threshold = 95  # Minimum completeness threshold

consistency_threshold = 95    # Minimum consistency threshold

**Step 5**: Iterate Through Data Quality Dimensions (Completeness and Consistency)

for quality_dimension in ["Completeness", "Consistency"]:

**Step 6**: For Each Quality Dimension, Identify Matching Columns in the Dataset

matching_columns = IdentifyMatchingColumns(dataset, quality_dimension)

**Step 7:** Store the Mapping in the Dictionary

mapping[quality_dimension] = matching_columns

**Step 8:** Analyze Data Quality Metrics for Mapped Columns

for quality_dimension, columns in mapping.items():

**Step 9**: Calculate Data Quality Metrics (Completeness and Consistency) for Each Mapped Column

metrics = CalculateDataQualityMetrics(dataset, columns, quality_dimension)

**Step 10**: Store the Metrics or Take Further Actions (e.g., visualization)

StoreMetricsForQualityDimension(quality_dimension, metrics)

**Step 11**: Report or Visualize Data Quality Metrics for the Entire Dataset

ReportDataQualityMetrics(mapping)

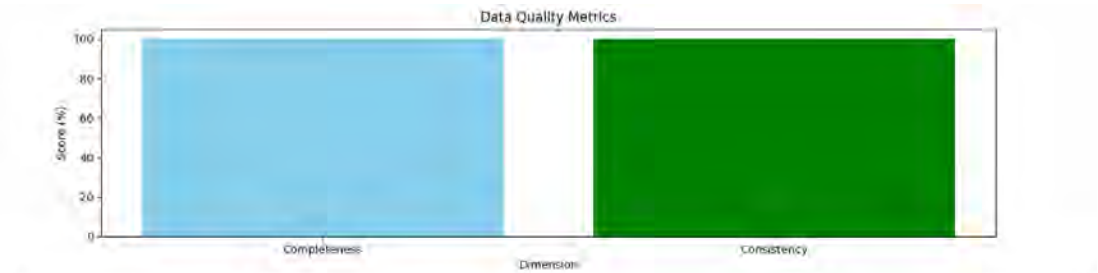
**Step 12**: End

**Assessment of Model Performance**: Comparing the results produced by the truth dataset and the DQ ontology model requirements/rules. The model's performance metrics, such as the score for completeness and completeness score using completeness and consistency metrics as defined below:

$$\text{Completeness (\%)} = \frac{\text{Number of Unique row}}{\text{Total Values}} X\ 100$$

$$\text{Consistency (\%)} = \frac{\text{Number of Values with Consistent types}}{\text{Total Rows}} X\ 100$$

Threshold for completeness and consistency was set as 95% as the acceptable result for evaluating the DQ. Results was gotten using python script to map the DQ ontology and dataset. The first approach was to use the truth dataset from COVID-19 Hospital daily record of death. (https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3).




| Dimension | Value |
|---|---|
| Completeness | 100.00% |
| Consistency | 100.00% |
| Columns Below Threshold | 0 |
| Rows Below Threshold | 0 |

*Figure 4.6: Result from the Truth Dataset*

The result from the Truth dataset shows 100% score for both completeness and consistency and hence our basis for evaluating the model over other papers.

### 4.7.4 Model Validation with Benchmark Papers

Furthermore, in continuation of the model validation, the results above were compared with two benchmark studies from the literature using two study data sets on our model. The two benchmark paper articles by Elouataoui et al., (2022) and Latsou et al., (2022).

**First Paper:** (Elouataoui et al., 2022)

**Dataset Description**

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

Result from the above author's paper article for completeness is 60.62% while the result for consistency is 45.5%.

**Data Pre-processing**

The first step of solving the IID problem after collecting the data according to our model and before applying the DQ ontology is to perform a pre-process on the dataset.

The data was preprocessed using OpenRefine and Python These can be a powerful combination to clean, transform, and prepare data for analysis or other downstream tasks. Here are the steps followed to preprocess data using these two tools:

1. **Data Exploration**: OpenRefine's interface is used to explore the dataset and understand its structure.
2. **Remove Duplicate**: This phase identifies and removes duplicate rows if necessary.
3. **Data Standardization**: Standardize text data by correcting typos and capitalization.
4. **Split and Merge Cells**: Split or merge columns to organize the data better.
5. **Remove or Fill Missing Values**: Handle missing data by removing rows or filling in missing values.
6. **Transform Data**: Create new columns, apply mathematical operations, or perform other transformations.



During the first step of the pre-processing (**Data Exploration)**, the result of data exploration showed the following chart in terms of completeness and consistency:

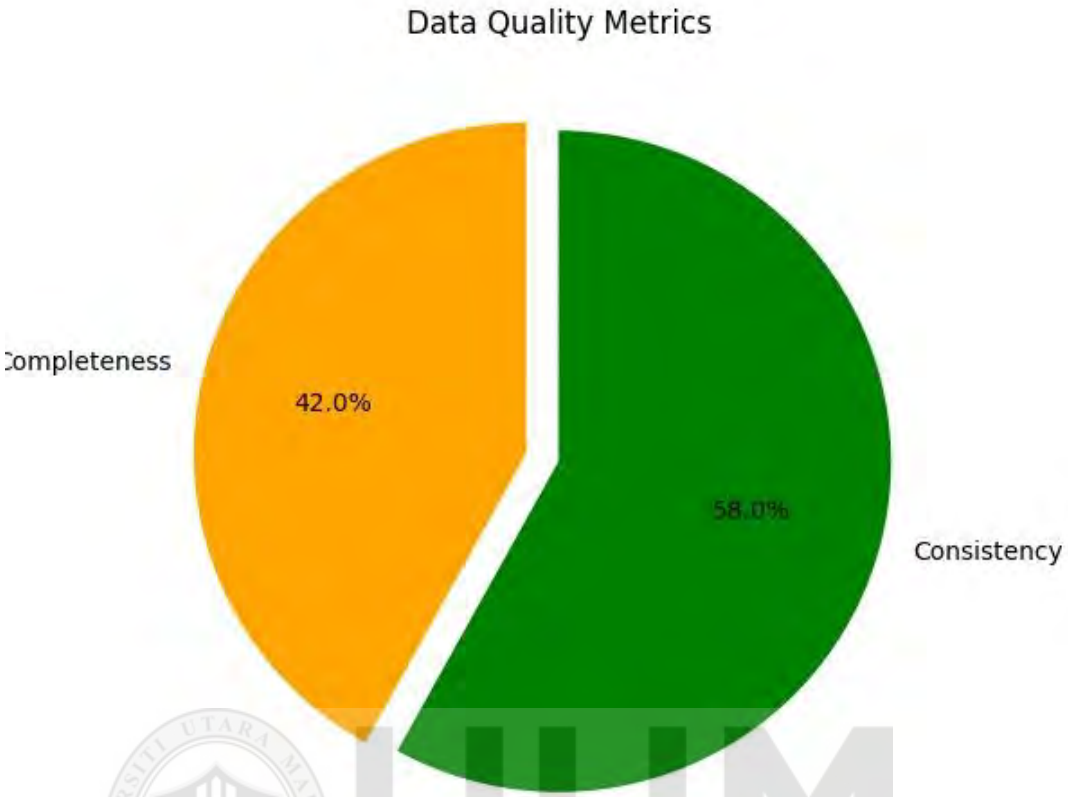
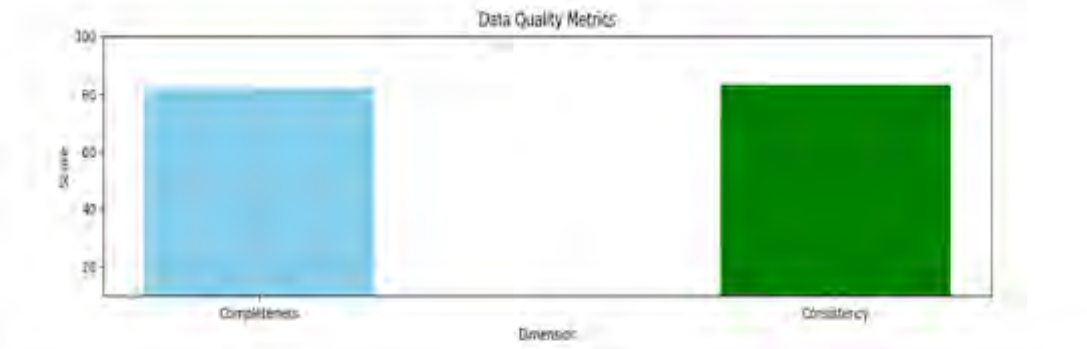



*Figure 4.7: Result from first paper Data Exploration*

Figure 4.7 shows the result from this research model using the same dataset from the stated paper article, that is, after the data set was preprocessed and fed into our study model. The following result shown underneath was obtained:




| Dimension | Value |
|---|---|
| Completeness | 82.01% |
| Consistency | 83.00% |
| Columns Below Threshold | 0 |
| Rows Below Threshold | 0 |

*Figure 4.8: Result from the First Paper Dataset*

As shown in Figure 4.8, the completeness DQ dimension scored 82.01% and consistency DQ dimension 83%. The results for both DQ dimensions are below the set threshold of 100%. The above result shows the model used in this research performs better than the model used in the paper mentioned above. Therefore, our model is more efficient.



**Second Paper:** (Christina Latsou, Marta Garcia I Minguell, Ayse Nur Sonmez, Roger Orteu I Irurre, Martinmark Palmisano, 2022)

**Dataset Description**

The data set used in the above article provided relates to the management and development of a number of critical assets involved in global operations. It holds information for thirty four assets and five model types that operate in various locations. The dataset consists of 23 columns and 39569 rows. It contains information for the usage of assets over time, recoding data in terms of land and nautical mileage, average speed of land, and flight hours, average altitude, and minimum and maximum water temperatures. The challenge arises with the 'asset usage' dataset is that the data collected over the years has been merged into one file to have a single point of

reference. This can affect the quality of data due to its heterogeneity, including structural and lexical differences across the merged datasets.

*Table 4.4: Paper Two (2) Result*

| **Dimensions** | **DQ Level** | **Results** | **State** |
|---|---|---|---|
| Completeness | 38876 | 98.25% | Accepted |
| Consistency | 38267 | 96.71% | Not Accepted |

Table 4.4 shows from the second paper's result using their model. Same as in the first, at the first step of the pre-processing (**Data Exploration)**, the result of data exploration showed the following chart in terms of completeness and consistency: Completeness DQ dimension is 40.1% which is acceptable based on their metrics score and Consistency is 59.9% which is not acceptable. And the same dataset was used on this study research model after preprocessing and the result shown in Figure 4.9 was obtained.

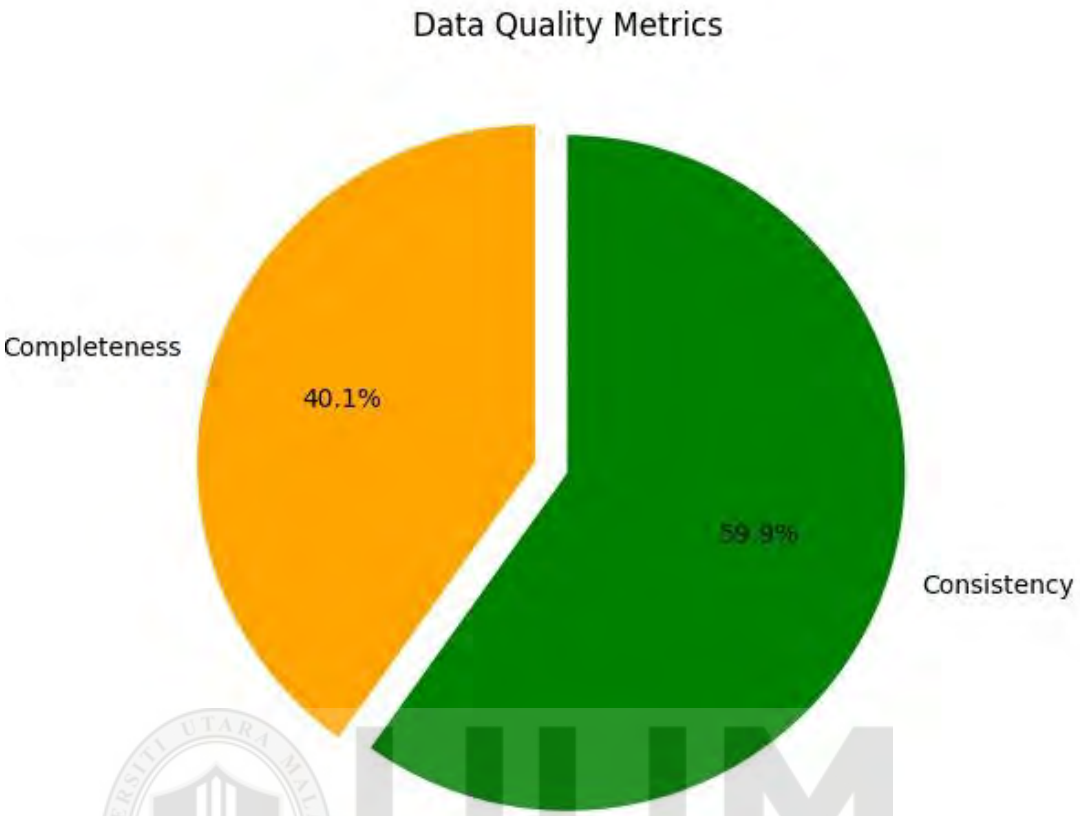



*Figure 4.9: Result from second Data Exploration*

Figure 4.9 is the result from this research model using the same dataset from the stated paper article, that is, after the data set was preprocessed and fed into our study model. The following result shown in Figure 4.10 was obtained.


| Dimension | Value |
|---|---|
| Completeness | 97.50% |
| Consistency | 98.00% |
| Columns Below Threshold | 0 |
| Rows Below Threshold | 0 |

*Figure 4.10: Result from the Second Paper Dataset*

As shown in Figure 4.10, the completeness DQ dimension scored 97.50% and consistency DQ dimension 98.00%. The results for completeness DQ dimensions are below the set threshold. The above result shows the model used in this research performs better than the model used in the paper mentioned above. Therefore, our model is more efficient.



### 4.7.5 Benchmark Paper Articles

The selected two papers were used to compare the proposed study and to check if the proposed studies address the weaknesses and limitations of the existing study. Although, the reason for the selected articles as bechmark for this research study is based on the problem domain that is, DQ issues especially in completeness and consistency dimensions and the benchmark study data sets were used on our model, they share same attributes which are common to the work presented in this research such as in technique, methodology and the DQ dimensions. The selected information is extracted from the research articles and is summarized in Table 4.5

*Table 4.5: The Selected Papers for Benchmarks*

| No. | Reference | Technique | Data Quality Dimensions |
|---|---|---|---|
| 1 | (Elouataoui et al., 2022) | Weighted Metrics | Completeness, consistency, timeliness, volatiliy, reliability, readabilty |
| 2 | (Christina Latsou, Marta Garcia I Minguell, Ayse Nur Sonmez, Roger Orteu I Irurre, Martinmark Palmisano, 2022) | Data Quality Ontology | Timeliness, accuracy, completness, consistency, validity, uniqueness, |

### 4.7.6 Clustering

Clustering is applied to check the DQ violations. Based on the findings that are gathered at the DQ ontology implementation stage, the results are used to apply a clustering technique. Cluster analysis or clustering is a method of grouping sets of objects so that objects in the same cluster or group are more like each other by first applying a pre-processing method to the data. To solve the DQ violation, the result achieved from the DQ violation identification stage is applied on clustering, which performs preprocess steps to filter the inconsistent and incomplete features in the dataset and cluster applied to group based on requirement needs using the K-Means clustering steps as shown in Figure 4.11.

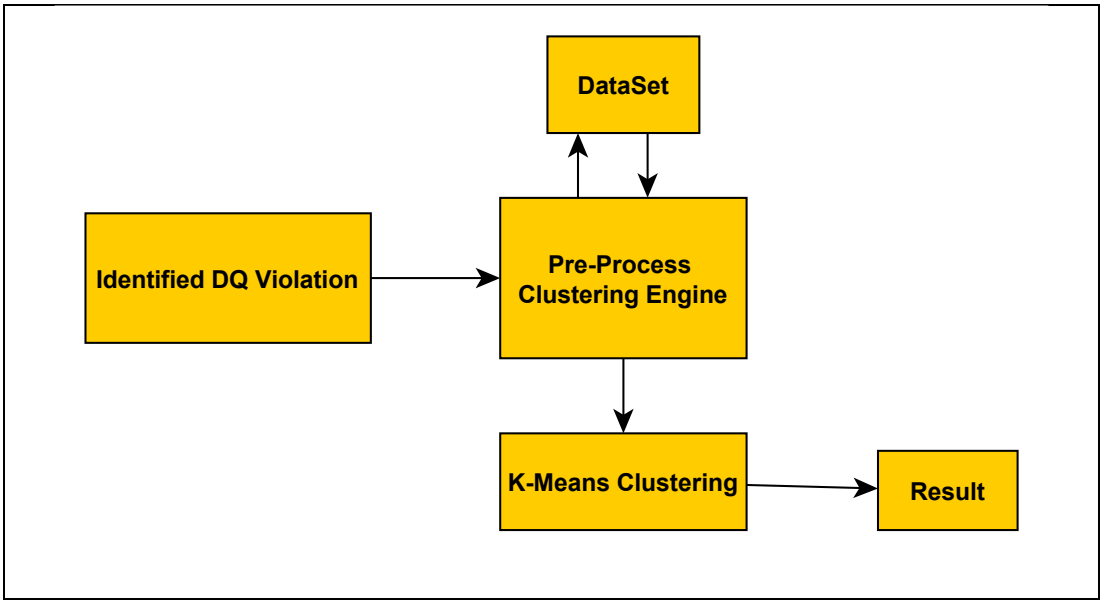



*Figure 4.11: Clustering Steps*

K-Means algorithm is used in this work because it is used to develop clusters for unlabeled data. In BD, the data is unlabeled and the results are developed based on the clusters that it generated. Details of the DQ improvement is given in the next chapter, 5, where a use case is applied to the model and results of the DQ is improved by clustering and DQ monitoring.



### 4.8 Summary

This chapter started by accomplishing the first objective of the study which is a systematic literature review. Furthermore, the formulation of the different aspects of the proposed model is explained in this chapter. The entire model is divided into four stages, namely**:** data gathering from different data sources, data preparation, DQ violation and DQ improvement. The first stage is where the dataset are collected from their differnt sources, the second is the collected data set must be structured or transformed and preprocessed. The main focus of this study is on stages three and four, where a DQ ontology designed in this chapter and presented to detect DQ violation based on DQ dimensions chosen for this study. The DQ ontology is applied to the to the data set to perform data search to identify the violation. After which comes stage

three, the final stage, a clustering technique is applied to solve the DQ violation. Based on the findings achieved in stage two, requirement gathered at the ontology stage is used to apply a clustering technique that focuses on data linkage to solve the DQ violation problem. The achievement in this chapter brings us to the accomplishment of objective number two which is to develop and validate a DQ transformation model which is the core of this study.

# CHAPTER FIVE: EVALUATION AND DISCUSSION

This chapter applies the 2CsDQT model on a use case. This is done by applying the model to several use-case to demonstrate the model's practical application and for continous monitoring and assessment of DQ specically on completeness and consistency DQ dimensions. The discussion about general findings on evaluation process ends with the conclusion of the chapter.

## 5.1 Application of 2CsDQT Model

**Error! Reference source not found.** shows the steps followed for the evaluation of t his study. The first step is data collection in which data from different sources but a similar domain is collected. The second step is data integration and transformation. The collected data is integrated and transformed. The integration is carried out in which the data is integrated from different sources to form a single dataset. The third step is the transofrmation process i.e., to apply 2CsDQ, and this step involves the study and analysis of the transformed data to identify the DQ violation requirements and problem. It also involves steps to solve such identified problem. The last step is the evaluation of the study and a comparative study of the benchmark studies.

## 5.2 Application Procedure of 2CsDQT Model

The 2CsDQT model is applied to the case study after the transformation to (i) identify requirement violations, and (ii) solve the requirement violations problem and monitor the quality state of data. Therefore, the 2CsDQT model was set up and configured as explained in Chapter 4. These steps aid as the best practice to help determine and

monitor the performance of the model. The model laid a step by step guide for the successful execution of the study.

### 5.3 Applicability of 2CsDQT

The model is a complete method containing the tools, requirements, and steps to arrive at and conclude on the research aim. The model of this study is applied to a case study to help draw a meaningful conclusion of the study and present a better way to measure the effectiveness of the study. The section further shows steps on how the 2CsDQT model is applied on the datasets and case studies to achieve the expectation of the study. Case study is chosen from GoodRead Books becuase the volume of books which is 10000 books characterize a BD set and every book has features which needs classification and transformation.

### 5.4 Case Study: Evaluation of GoodRead Books Data Set

This case study deals with over 10,000 good read books integrated into one dataset. This dataset is regarded as a BD set since it carries two important characteristics of a BD set i.e., volume and variety. With a volume of over 10,000 Goodreads, data was collected and integrated to form one dataset. A study and analysis are carried out on the datasets to determine the required type of integrated data. During this process, it is identified that the dataset includes ratings for ten thousand trendy books. In general, each book receives 100 reviews, while some receive fewer or more. The ratings run from one to five. Both book and user IDs are linked together. They range from 1 to 10000 books and 1 to 53424 users. Each user has given at least two ratings. Each user has an average of eight ratings. There are also books marked to read by the users, book metadata (e.g., author, year, etc.), and tags.

### 5.4.1 Data Pre-processing

The first step of solving the IID problem after collecting the data according to our model and before applying the DQ ontology is to perform a pre-process on the dataset. The data was preprocessed using OpenRefine and Python These can be a powerful combination to clean, transform, and prepare data for analysis or other downstream tasks. Here are the steps followed to preprocess data using these two tools:

7. **Data Exploration**: OpenRefine's interface is used to explore the dataset and understand its structure.
8. **Remove Duplicate**: This phase identifies and removes duplicate rows if necessary.
9. **Data Standardization**: Standardize text data by correcting typos and capitalization.
10. **Split and Merge Cells**: Split or merge columns to organize the data better.
11. **Remove or Fill Missing Values**: Handle missing data by removing rows or filling in missing values.
12. **Transform Data**: Create new columns, apply mathematical operations, or perform other transformations.



During the first step of the pre-processing (**Data Exploration)**, the result of data exploration showed the following chart in terms of completeness and consistency:


*Figure 5.1: Goodreads Data Exploration*

Figure 5.1 shows the visibility of the identified DQ problem in the dataset as it shows the incomplete value and properties of the explored data, which identified 21.1% data completeness and 78.9% consistency.

This data exploration led to the following assumptions made on the data domain during the data preprocessing:

i) Exclusion of published year that is more than four-digit and less than four-digit or not equal to four digits.

ii) Exclusion of any Author Name field that is without any value.

iii) Combine all Editor and Publisher field

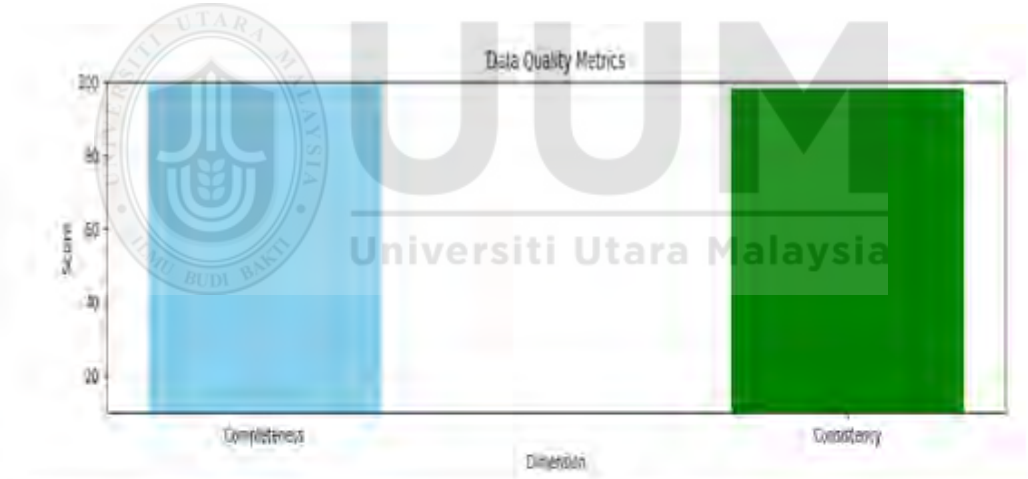
iv) Exclusion of any book title that cannot be used to identify Author Name.

### 5.4.2 Identify Requirement Violations

Furthermore in the process, the developed DQ ontology is applied to the same dataset to identify the requirement violations based on the DQ requirements, metrics and set threshold. Python uses an intelligent algorithm and set of rules that also deals with missing values and inconsistencies, together with the DQ ontology imported into the software's device. To achieve that, the DQ ontology is imported and stored in the same

directory with the executable file of the local storage drive in the computer, and python dependency is installed.

The terminal initiates the binding to use the DQ ontology, python automatically fetches the ontology into its algorithm and uses it as a dictionary search of DQ features. In essence, DQ ontology requirements helps to identify violations while Python contribute by helping correcting incompleteness and inconsistencies in the dataset. This process helps to reduce as much as possible DQ issues with great improvement as can be seen in Figure 5.2. The data set has improved in both the consistency 98% and completeness at 99.37% in the presentation of the data values.




| Dimension | Value |
|---|---|
| Completeness | 99.37% |
| Consistency | 98.00% |
| Columns Below Threshold | 0 |
| Rows Below Threshold | 0 |

*Figure 5.2: Incomplete and Inconsistencies in Goodreads Data*

### 5.5 Data Quality Improvement

To ensure the best DQ, this research adopts the k-means clustering technique, applied to the result presented in **Error! Reference source not found.**2 and Figure 5.3. The c lustering algorithm is applied to solve the identified problem and produce the 2Cs datatype based on the DQ violation. By doing this, the clustering algorithm focused its technique on the consistent and complete datatypes as identified. Clustering's objective is to discover similarities between strings and cluster them together based on a similarity criterion and removes the incomplete and inconsistent dataset.

#### 5.5.1 Clustering (K-Means)

Using k-means clustering to improve the data quality involves identifying and handling outliers or noise in the dataset. By grouping similar data points together into clusters in order to identify and potentially address data quality issues. The following Procedure was carried out.



#### 5.5.2 Procedure for Clustering

**Feature Selection:** Relevant features (columns) for clustering was selected. Not all features contribute equally to data quality.

**Scaling/Normalization:** There are different scale and hence the need to normalize them so that they have similar ranges. K-means is sensitive to the scale of features.

**Choosing the Number of Clusters (K):** Hierarchical clustering dendrogram was chosen to determine the optimal number of clusters for the data. This is crucial for effective clustering.

**Applying K-means Clustering:** The K-means clustering algorithm was ran on the chosen dataset with the chosen number of clusters (K). This group similar data points together.

**Identify Outliers and Noise:** After the clustering, the clusters and data points that are assigned to them was examined. Outliers and noise can often be identified as data points that do not belong to any cluster or are in very small clusters.

**Handling Outliers and Noise:** Depending on the nature of the outliers and noise, you can take various actions:

**Data Removal**: Removing extreme outliers due to data entry errors or anomalies.

**Data Imputation**: missing values or outliers are due to DQ issues; they were imputed using statistical methods or machine learning models.

**Cluster Deletion**: In the case were cluster consists mostly of noise that cluster will be removed

**Cluster Splitting**: when cluster contains both relevant data and noise they will be split it into sub-clusters or investigate to separate the noise.

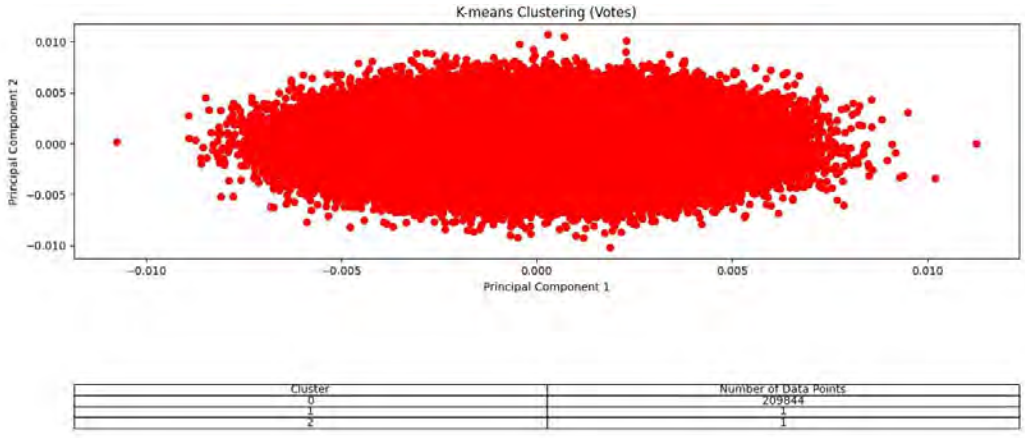



*Figure 5.3: Cluster Result from Goodreads Refine Test Dataset*

**Cluster 0**: This cluster contains 209,844 data points. It is the largest cluster by far and represents a group of data points that share similar characteristics or content. The fact that it has a significantly larger number of data points suggests that this cluster captures consistency and completeness

**Cluster 1**: This cluster contains only 1 data point. It is a very small cluster, and its presence means outlier that doesn't fit and hence represent the presence of completeness and insignicant inconsistency.

**Cluster 2**: This cluster also contains only 1 data point, similar to Cluster 1. It is another small cluster, and like Cluster 1, its presence means outlier that doesn't fit and hence represent the presence of completeness and insignificant inconsistency. This is to show our model has performed efficiently.

### 5.6 Data Quality Monitoring

Talend stands out as a popular, open-source platform designed for integrating and managing data comprehensively. It offers a wide array of tools for crafting, developing, testing, and deploying solutions related to data integration, data transformation, and DQ. Talend simplifies the complex task of connecting to diverse data sources, transforming data as needed, and efficiently loading it into target systems. This versatility positions Talend as a valuable asset for organizations seeking to enhance data integration, facilitate ETL processes (Extraction, Transformation, Loading), and maintain DQ standards.

Talend Data Quality was used for effective data quality management and monitoring, including data profiling, cleansing, and alerting. The alerting mechanism ensures that you and relevant stakeholders are notified promptly when data quality violations occur,

enabling swift corrective actions. The following procedure will be utilized when carrying these operations.

### 5.6.1 Set Up Data Monitoring

Configure data monitoring tasks into talend. Specify the frequency of monitoring and conditions under which monitoring tasks should run based on the data quality model as described above.

### 5.6.2 Implementation of Data Quality Alerts

Alerting mechanisms was configured within Talend to notify relevant stakeholders when data quality violations are detected. Talend provides email notifications, which you can customize based on specific conditions.

### 5.6.3 Monitor Data Quality Dashboards

Utilize Talend's data quality dashboards and reports to visualize data quality metrics and trends. The dashboards were customized to suit the monitoring needs.



## 5.7 Summary

This chapter applied a use case to the research proposed model, and the implementation of the model developed and validated in chapter four carried out was tested. It shows the study's outcome and presented the result presenting the effectiveness and performance of the model.

# CHAPTER SIX: CONCLUSION AND FUTURE WORK

It is crucial to examine the core areas and objectives of this research for the enablement to measure the degree of achievement of the objectives before the conclusion of the thesis. Therefore, the next section will reviews the objectives of the research and presents the research contributions. The limitations in developing 2CsDQT are discussed and overviews of the solutions are highlighted. This chapter ends by proposing the future work.

## 6.1 Review of Research Objectives.

The primary goal of this study is to create and test a model for creating DQ throughout the transformation process in a BD environment. This testing was done as BD usually contains inconsistencies and incompleteness, leading to poor data analysis quality. Thus, the study proposed a transformation model for BD, which was implemented and experimented. Therefore, this section articulates the accomplishments logged at various phases of the model developed, which resulted in the fulfilment of the objectives as stated in chapter one. Summary of the approaches of the objectives stated in chapter one is outlined below:



### 6.1.1 Objective One - To identify and measure how incompleteness and inconsistency of data affect the quality of BD.

One of the problems stated in the first chapter of this research work is that it is challenging to identify DQ problems in BD due to its volume, velocity, and varied nature, which further compound the identification of DQ dimensions relevant to BD. This leads to the purpose of the research objective: to identify the effects of the DQ dimension in the data transformation of BD. The objective was achieved with the assistance of relevant related literature reviewed on BD and DQ dimensions and majorly on conducting a systematic literature review. The key to

DQ dimensions to be assessed for BD for measuring the level of quality in using heterogeneous datasets for BD projects are completeness and consistency (Merino et al., 2016b). If datasets are transported through networks and shared by various applications, consistency is defined as the capacity of data and systems to retain the uniformity of certain properties, and completeness is determined to include all expected data values in a given data collection (Batini et al., 2016). Moreover, an integrated DW of heterogeneous multiple data sources in BD is an example of a critical case in which inconsistent and incomplete data are most likely to exist. Works such as SETL in Batini et al. (2016), Fürber, (2015) and Chika, (2015) assess the integrated database as susceptible to inconsistent and incomplete data.

The study went further by identifying the effects of the DQ dimensions identified or rather the effects of poor DQ on the identified DQ dimensions on data transformation in BD. This came from the fact that data transformation techniques in traditional ETL cannot detect the DQ dimension violations in BD. DQ is unknown with especially with the said BD characteristics. Most of the studies accept that DQ faces multiple challenges, especially in BD. This poses new challenges that new methods have to address. It is complicated because we're mining data from extensive databases and dealing with a mix of structured, semi-structured, and unstructured data types. Thus, DQ must be managed properly (Ijab et al., 2019).

### 6.1.2 Objective Two - To develop and evaluate a 2CsDQT data transformation model with the completeness and consistency DQ dimensions.

The second objective is to develop an enhanced data transformation model that will improve the DQ. New algorithms have to be designed to deal with novel requirements related to variety, volume, velocity, value and veracity issues since the literature discovered that applying the traditional database numerous assessment methods of DQ concepts directly to BD encounters

some difficulties in the BD scenario. The goal of the DQ research field is to provide techniques for defining, evaluating, and enhancing the appropriateness and usefulness of data for the context in which it is intended to be utilised. Many of the contributions were on methods and strategies for dealing with structured data, but BD introduces new difficulties that new approaches must address. It is necessary to be able to extract essential and high-quality data in order to exploit BD. This is complicated because we're working with data from extensive databases and a variety of organised, semi-structured, and unstructured data (Hafez, 2016)(Ijab et al., 2019). Though this research identified the DQ dimensions suitable for BD and the effects of their violations, from the few studies that can work on DQ improvement on BD, especially on consistency and completeness, DQ dimensions only emphasize DQ assessment leaving out the improvement aspect. The studies that made efforts to improve did not emphasize on consistency and completeness DQ dimensions.

In achieving this objective, the data transformation model was developed with the DQ Dimensions. This section briefly explain some activities involved in the model development. The central task in this objective is to develop a data transformation model, as shown in Figure 4.1, Chapter 4. The model is based on an initial model in Figure 3.3, Chapter 3, which highlight the ELT. The emphasis of the proposed model here is on the transformation stage of the ELT. The major phases in the transformation stage are the DQ violation identification phase. This is where the inconsistencies and incompleteness in the dataset are identified with the aid of a DQ ontology is developed as seen in figure 4.2. Furthermore, the model was validated by a ground truth BD set and with some benchmark studies in the literature. The next phase in the model is where clustering is applied to the results of the identified inconsistencies and incompleteness within the dataset to improve the data in order to obtain consistent and complete dataset.

### 6.1.3 Objective Three - To evaluate the effectiveness of the proposed model in order to produce quality data of BD using a case study

To achieve the above objective, we evaluate the proposed 2CsDQ model approach. Firstly, there is the practical applicability of 2CsDQ by applying the model to a use case and clustering applied after the results of DQ violation. Then, the continous DQ monitory evaluation of DQ whereby notification alerts are sent to the DQ monitor for action thereby solving incompleteness and inconsistencies in BD.

## 6.2 Research Contributions

Significant theoretical and practical contributions to the body of knowledge of DQ can be observed from the findings of this study. This is specifically for the improvement of DQ by resolving inconsistencies and incompleteness in BD. This study demonstrated an innovative approach to understanding and identifying inconsistencies and incompleteness in BD and insight into the resolution of the deficiency in these DQ dimensions. Likewise, the study results contribute theoretically and practically in the direction of the overall goal of data management, as explained in the following sections.



### 6.2.1 Theoretical Contribution

The major theoretical contribution is extending DW framework by improving data transformation process using ontology approach and clustering method in BD environment. Theoretically, this study provides perceptions to the comprehension of the importance of data management, especially in the aspect of DQ. It also provided deeper understanding of the importance of resolving inconsistencies and incompleteness in dataset which will in turn benefit the quality in data analysis. Specifically, this work has increased the available documentations on identifying and exposing DQ dimensions relevant for BD, contributing to

the literature of the various DQ dimensions for quality data transformation for BD. As explained earlier in chapter one, one of the rationales for doing this research is to increase the documentation on the effects of DQ dimensions violations in BD. It is shown that there are ample documentation on how inconsistent and incomplete data can be dealt with in traditional databases. This is unlike in BD settings. In this work, the ways BD quality issues can be dealt with using semantic web technologies and clustering were explored.

This work, in identifying and exposing existing data transformation approaches for BD quality in the literature which do not fully address the quality issues has contributed to the body of knowledge. It is also significant in contributing to the growing literature on the development of an effective data transformation model involving semantic technologies, ontology in particular, to identify and assess DQ measuring the consistency and completeness DQ dimensions of BD and the application of clustering algorithm to improve the quality. The evaluation has also offered guidance to researchers that are interested in this field of study. It will also help them to comprehend what has been accomplished and what remains to be accomplished. Consequently, this study has contributed to the growing body of knowledge by offering a novel approach in addressing DQ issues in BD by identifying DQ violations using DQ ontology and clustering algorithms and DQ monitoring for the DQ improvement majoring in consistency and completeness DQ dimensions.



### 6.2.2 Practical Contribution

This study offers practical suggestions as following:

i) To the authorities in data management by enhancing the DQ transformation model for BD is significant for organizations to adopt it to improve their business and take better strategic decisions to advance their enterprises. Furthermore, the quality of service

provided by firms that apply the model will be improved, and a standard for qualitative data processing for commercial enterprises or organisations will be enforced, thereby enhancing the DQ by the users. Bad DQ (e.g., incomplete, inconsistent data and without data provenance) is inefficient in terms of consumption and may directly affect the decrease in company decisions' efficiency, as harmful effects have been observed on businesses.

ii) To the software developer for DQ system or application. Since various applications have been developed to produce the DQ system, this research will develop and produce quality data, especially for incomplete and inconsistent data from the big data environment and provides a roadmap to load only verifiable data to the transformation model. The various techniques suggested in this research can strengthen the current method used in the current system or applications.

In general, since businesses are expected to provide high-quality data, corporations consider data as more analytical during their decision making. It's more helpful when businesses have access to complete and consistent data, more valuable knowledge and insight. Though advanced technology and analytical tools help you recognise data patterns and valuable information, data insight ultimately depends on DQ. In the absence of high DQ, it will be hard for businesses to diagnose and have profound and useful information adequately.

### 6.3 Research Challenges

In respect to the modeling and designing the 2CsDQT, this research suggests the following challenges to be tackled accordingly. The suggestion can be classified into either short (i.e., immediately implement after this research) and long term (i.e., possibility involving new trends in DQ in BD). Nevertheless, future work can complete the application prototype for

implementing DQ in the case study. However, several research issues are still left open on the grounds of this research work and discussed as following:

**Data privacy:** To ensure that data collection and processing abide with applicable data privacy laws (such as GDPR and HIPAA).

**Resources Needed**: It may use big labour cost, and take a lot of money to create and maintain a data quality model.

**Continuous Improvement**: Monitoring and adjusting the quality of the data is a continuous activity, and always changing according to new requirements of user applications.

### 6.4 Research Limitations

This research is overtly centred on achieving DQ in BD. Therefore, the study only covers the relevant efforts necessary within the scope as stated in chapter one. Nevertheless, more efforts should have been put in DQ areas, but this work cannot go further due to resources needed and time constraints. Moreover, the current behaviour of data produced from unpredictable resources was a great challenge to address DQ issues, especially in data incompleteness and inconsistency. Companies looking to retain high-quality data, a strong data quality model that includes data collection, preprocessing, ontology construction, data improvement through clustering, and a monitoring tool with warnings and notifications is essential. Companies and stakeholders can assure data accuracy, dependability, and relevance by following the instructions provided in this material, which will ultimately result in greater decision-making and operational efficiency.

## 6.5 Future Study

In this research, the data transformation model was developed encompassing a data integration with ontology to integrate the multiple dataset from different data sources into one dataset and a DQ ontology used to identify DQ violations using the python scripts. Moreover, a clustering algorithm to resolve the identified 2Cs DQ violations. Currently, the 2CsDQT model focuses on only two DQ dimensions: consistency and completeness. Future work could address other DQ dimensions such as accuracy, timeliness, integrity and so on, especially to cover specific DQ dimensions relevant to data types other than a textual dataset. The developed model can be further extended to use more than one clustering algorithm as this study only used one clustering algorithm, namely the k-means algorithm. Furthermore, this research is currently focused on just textual dataset. The researcher will recommend using other BD types such as streaming data or social media data for case study and evaluation of the model.

## 6.6 Concluding Remarks

DQ management has been an ongoing research area for decades, and researchers have been trying to proffer solutions in the best possible ways. Nevertheless, the issues in DQ will continue as long as data is being produced in the world. This is especially true in this BD system, where data are produced in milliseconds, and data inconsistencies and incompleteness always happen. This thesis shows how the data transformation process using ontologies, and clustering algorithms can be employed to resolve DQ, data assessment, and data improvement. The evaluation results documented reveal that the developed transformation model (2CsDQT) is an improvement over the previous study.

# REFERENCE


Abdelhedi, F., Brahim, A. A., Atigui, F., & Zurfluh, G. (2017). MDA-based approach for NoSQL databases modelling. *International Conference on Big Data Analytics and Knowledge Discovery*, 88–102.

Abedjan, Z., & Naumann, F. (2016). CONFERENCE: Data Profiling. *Icde*, 1432–1435.

Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. (2016). How to improve firm performance using big data analytics capability and business strategy alignment? *International Journal of Production Economics*, *182*, 113–131. https://doi.org/10.1016/j.ijpe.2016.08.018

Al-sai, Z. A., Husin, M. H., Syed-mohamad, S. M., Abdullah, R., Zitar, R. A., Abualigah, L., & Gandomi, A. H. (2023). *Big Data Maturity Assessment Models : A Systematic Literature Review*. *Idc*.

Alizadeh, M., Shahrezaei, M. H., & Tahernezhad-Javazm, F. (2019). Ontology based information integration: a survey. *ArXiv Preprint ArXiv:1909.13762*.

Altendeitering, M., & Guggenberger, T. M. (2021). Designing Data Quality Tools: Findings from an Action Design Research Project at Boehringer Ingelheim. *Twenty-Ninth European Conference on Information Systems*, *Ecis*, 1–16. https://aisel.aisnet.org/ecis2021_rp/95

Ardagna, D., Cappiello, C., Samá, W., & Vitali, M. (2018). Context-aware data quality assessment for big data. *Future Generation Computer Systems*, *89*, 548–562.

Arhin, K., Baldini, I., Wei, D., Ramamurthy, K. N., & Singh, M. (2021). *Ground-Truth, Whose Truth? -- Examining the Challenges with Annotating Toxic Text Datasets*. *1*(1), 1–15. http://arxiv.org/abs/2112.03529

Barbosa, L., Crescenzi, V., Dong, X. L., Merialdo, P., Piai, F., Qiu, D., Shen, Y., & Srivastava, D. (2018). Big Data Integration for Product Specifications. *IEEE Data Eng. Bull.*, *41*(2), 71–81.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52. https://doi.org/10.1145/1541880.1541883

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2016). From data quality to big data quality. In *Big Data: Concepts, Methodologies, Tools, and Applications* (pp. 1934–1956). IGI Global.

Batini, C., & Scannapieco, M. (2016a). *Data and Information Quality: Concepts, Methodologies and Techniques*. Springer Cham.

Batini, C., & Scannapieco, M. (2016b). Data and information quality. *Cham, Switzerland: Springer International Publishing. Google Scholar*, 43.

Benkhaled, H. N., & Berrabah, D. (2019). Data quality management for data warehouse

systems: State of the art. *CEUR Workshop Proceedings*, *2351*, 1–10.

Berkani, N., Bellatreche, L., & Khouri, S. (2013). Towards a conceptualization of ETL and physical storage of semantic data warehouses as a service. *Cluster Computing*, *16*(4), 915–931. https://doi.org/10.1007/s10586-013-0266-7

Bodenhofer, U., Kothmeier, A., & Hochreiter, S. (2011). APCluster: an R package for affinity propagation clustering. *Bioinformatics*, *27*(17), 2463–2464.

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Kalro, T., Liang, Y., … Zardecki, C. (2019). RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, *47*(D1), D464–D474. https://doi.org/10.1093/nar/gky1004

Byabazaire, J., O'hare, G., & Delaney, D. (2020). Data quality and trust: Review of challenges and opportunities for data sharing in IoT. *Electronics (Switzerland)*, *9*(12), 1–22. https://doi.org/10.3390/electronics9122083

Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*(0). https://doi.org/10.5334/dsj-2015-002

Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M., & Xiao, G. (2017). Ontop: Answering SPARQL queries over relational databases. *Semantic Web*, *8*(3), 471–487.

Cappa, F., Oriani, R., Peruffo, E., & McCarthy, I. (2021). Big Data for Creating and Capturing Value in the Digitalized Environment: Unpacking the Effects of Volume, Variety, and Veracity on Firm Performance*. *Journal of Product Innovation Management*, *38*(1), 49–67. https://doi.org/10.1111/jpim.12545

Caulkins, J. P., Bao, Y., Davenport, S., Fahli, I., Guo, Y., Kinnard, K., Najewicz, M., Renaud, L., & Kilmer, B. (2018). Big data on a big new market: Insights from Washington State's legal cannabis market. *International Journal of Drug Policy*, *57*(March), 86–94. https://doi.org/10.1016/j.drugpo.2018.03.031

Chen, H., Yang, C. C., Chau, M., & Li, S. H. (2009). Lecture Notes in Computer Science: Preface. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 5477).

Chika, N. H. (2015). *Dealing with inconsistent and incomplete data in a semantic technology setting*. Sheffield Hallam University.

Christina Latsou, Marta Garcia I Minguell, Ayse Nur Sonmez, Roger Orteu I Irurre, Martinmark Palmisano, S. W. L.-V. (2022). *Developing an Ontological Framework for Effective Data Quality Assessment and Knowledge Modelling*. *November*.

Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. *Proceedings of the 2016 International Conference on Management of Data*, 2201–2206.

Cipolla, C. N. (2019). Taming the Ontological Wolves: Learning from Iroquoian Effigy Objects. *American Anthropologist*, *121*(3), 613–627. https://doi.org/10.1111/aman.13275

Côrte-Real, N., Ruivo, P., & Oliveira, T. (2020). Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value? *Information and Management*, *57*(1). https://doi.org/10.1016/j.im.2019.01.003

Croce, F., Savo, D. F., & Lenzerini, M. (2018). *On the experimental usage of ontology-based data management for the italian integrated system of statistical registers: quality issues*.

Cure, O. (2012). Improving the data quality of drug databases using conditional dependencies and ontologies. *Journal of Data and Information Quality*, *4*(1), 1–21. https://doi.org/10.1145/2378016.2378019

Daneshkohan, A., Alimoradi, M., Ahmadi, M., & Alipour, J. (2022). Data quality and data use in primary health care: A case study from Iran. *Informatics in Medicine Unlocked*, *28*, 100855. https://doi.org/10.1016/j.imu.2022.100855

Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). The advantages of an Ontology-Based Data Management approach: openness, interoperability and data quality. *Scientometrics*, *108*(1), 441–455. https://doi.org/10.1007/s11192-016-1913-6

De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., & Rosati, R. (2018). Using ontologies for semantic data integration. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years* (pp. 187–202). Springer.

Debattista, J., Auer, S., & Lange, C. (2016). Luzzu--A Framework for Linked Data Quality Assessment. *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, 124–131.

Debroy, V., Brimble, L., & Yost, M. (2018). NewTL: engineering an extract, transform, load (ETL) software system for business on a very large scale. *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 1568–1575.

Deja, K. (2019). Using machine learning techniques for Data Quality Monitoring in CMS and ALICE. *Proceedings of Science*, *350*, 0–10.

Dictionary of Data Quality. (n.d.). *Dictionary of dimensions of data quality (3DQ) Dictionary of 60 Standardized Definitions*. http://www.dama-nl.org/wp-content/uploads/2020/11/3DQ-Dictionary-of-Dimensions-of-Data-Quality-version-1.2-d.d.-14-Nov-2020.pdf

Dimassi, S., Demoly, F., Cruz, C., Qi, H. J., Kim, K. Y., André, J. C., & Gomes, S. (2021). An ontology-based framework to formalize and represent 4D printing knowledge in design. *Computers in Industry*, *126*(0). https://doi.org/10.1016/j.compind.2020.103374

Dong, X. L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., Sun, S., & Zhang, W. (2015). Knowledge-based trust: Estimating the trustworthiness of web sources. *ArXiv Preprint ArXiv:1502.03519*.

Dos Reis, J. C. (2018). Mapping refinement based on ontology evolution: A context-matching approach. *CEUR Workshop Proceedings*, *2228*, 251–256.

Du, J., & Zhou, L. (2012). Improving financial data quality using ontologies. *Decision Support Systems*, *54*(1), 76–86.

Dubey, R., Gunasekaran, A., Childe, S. J., Luo, Z., Wamba, S. F., Roubaud, D., & Foropon, C. (2018). Examining the role of big data and predictive analytics on collaborative performance in context to sustainable consumption and production behaviour. *Journal of Cleaner Production*, *196*, 1508–1521. https://doi.org/10.1016/j.jclepro.2018.06.097

Dwivedi, Y. K., Venkitachalam, K., Sharif, A. M., Al-Karaghouli, W., & Weerakkody, V. (2011). Research trends in knowledge management: Analyzing the past and predicting the future. *Information Systems Management*, *28*(1), 43–56. https://doi.org/10.1080/10580530.2011.536112

Elouataoui, W., Alaoui, I. El, & Mendili, S. El. (2022). *An Advanced Big Data Quality Framework Based on Weighted Metrics*.

Engineering, S. (2022). *Software Engineering and AI for Data Quality in Cyber-Physical Systems / Internet of Things*.

English, L. P. (1999). *Improving data warehouse and business information quality: methods for reducing costs and increasing profits* (Vol. 1). Wiley New York.

Estiri, H., Klann, J. G., & Murphy, S. N. (2019). A clustering approach for detecting implausible observation values in electronic health records data. *BMC Medical Informatics and Decision Making*, *19*(1), 1–16. https://doi.org/10.1186/s12911-019-0852-6

Ezzine, I., & Benhlima, L. (2018). A Study of Handling Missing Data Methods for Big Data. *Colloquium in Information Science and Technology, CIST*, *2018-Octob*, 498–501. https://doi.org/10.1109/CIST.2018.8596389

Fatima, A., Nazir, N., & Khan, M. G. (2017). Data cleaning in data warehouse: a survey of data pre-processing techniques and tools. *IJ Information Technology and Computer Science*, *3*, 50–61.

Firmani, D., Tanca, L., & Torlone, R. (2019). Ethical dimensions for data quality. *Journal of Data and Information Quality*, *12*(1), 10–14. https://doi.org/10.1145/3362121

Fosso Wamba, S., Gunasekaran, A., Dubey, R., & Ngai, E. W. T. (2018). Big data analytics in operations and supply chain management. *Annals of Operations Research*, *270*(1–2). https://doi.org/10.1007/s10479-018-3024-7

Fürber, C. (2015). *Data quality management with semantic technologies*. Springer.

Fürber, C., & Hepp, M. (2011). Towards a vocabulary for data quality management in semantic web architectures. *ACM International Conference Proceeding Series*, 1–8. https://doi.org/10.1145/1966901.1966903

Fürber, C., & Hepp, M. (2010a). Using semantic web resources for data quality management.

*International Conference on Knowledge Engineering and Knowledge Management*, 211–225.

Fürber, C., & Hepp, M. (2010b). Using SPARQL and SPIN for data quality management on the semantic web. *International Conference on Business Information Systems*, 35–46.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., & Herrera, F. (2016). Big data preprocessing: methods and prospects. *Big Data Analytics*, *1*(1), 9.

Gärtner, B., & Hiebl, M. R. (2018). Issues with Big Data. *The Routledge Companion to Accounting Information Systems (S. 161-172). New York: Routledge*.

Geisler, S., Quix, C., Weber, S., & Jarke, M. (2016). Ontology-Based Data Quality Management for Data Streams. *Journal of Data and Information Quality*, *7*(4), 1–34. https://doi.org/10.1145/2968332

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.

Gudivada, V. N., Baeza-Yates, R., & Raghavan, V. V. (2015). Big data: Promises and problems. *Computer*, *3*, 20–23.

Hafez, H. A. A. (2016). Mining big data in telecommunications industry: challenges, techniques, and revenue opportunity. *Dubai UAE Jan*, 28–29.

Haider, M., Gandomi, A., & Rogers, T. (n.d.). When Big Data made the headlines: Mining the text of Big Data coverage in the news media *. * *Accepted in the International Journal of Services Technology and Management. † Corresponding Author*, *1*(516), 1–43. https://ssrn.com/abstract=3426256

Hamdy, M. E., & Shaalan, K. (2018). A hybrid framework for applying semantic integration technologies to improve data quality. *International Journal of Information Technology and Language Studies*, *2*(1), 27–44. https://pdfs.semanticscholar.org/5daa/1e0197a403f193b7e530579ba53d988f7790.pdf

Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in big data analytics: survey, opportunities, and challenges. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0206-3

Hassenstein, M. J., & Vanella, P. (2022). Data Quality—Concepts and Problems. *Encyclopedia*, *2*(1), 498–510. https://doi.org/10.3390/encyclopedia2010032

Haug, A. (2021). Understanding the differences across data quality classifications: a literature review and guidelines for future research. *Industrial Management and Data Systems*, *121*(12), 2651–2671. https://doi.org/10.1108/IMDS-12-2020-0756

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, *154*, 72–80. https://doi.org/10.1016/j.ijpe.2014.04.018

Hevner, A., vom Brocke, J., & Maedche, A. (2019). Roles of Digital Innovation in Design Science Research. *Business and Information Systems Engineering*, *61*(1), 3–8. https://doi.org/10.1007/s12599-018-0571-z

Homayouni, H., Ghosh, S., Ray, I., & Kahn, M. G. (2019). An Interactive Data Quality Test Approach for Constraint Discovery and Fault Detection. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, 200–205. https://doi.org/10.1109/BigData47090.2019.9006446

Hu, W., Zaveri, A., Qiu, H., & Dumontier, M. (2017). Cleaning by clustering: methodology for addressing data quality issues in biomedical metadata. *BMC Bioinformatics*, *18*(1), 415.

Huang, J., Mao, B., Bai, Y., Zhang, T., & Miao, C. (2020). An integrated fuzzy C-means method for missing data imputation using taxi GPS data. *Sensors (Switzerland)*, *20*(7), 1–19. https://doi.org/10.3390/s20071992

Ijab, M. T., Surin, E. S. M., & Nayan, N. M. (2019). CONCEPTUALIZING BIG DATA QUALITY FRAMEWORK FROM A SYSTEMATIC LITERATURE REVIEW PERSPECTIVE. *Malaysian Journal of Computer Science*, 25–37.

Ilyas, I. F., & Chu, X. (2019). *Data cleaning*. Morgan & Claypool.

Impact, T., Data, B., Making, D., Sector, H., By, J. P., & Submitted, T. (2022). *Salahaldeen Yousef Student No " 201820629 " Prof . Faisal Aburub*.

Jafer, S., Chhaya, B., & Durak, U. (2017). OWL ontology to ecore metamodel transformation for designing a domain specific language to develop aviation scenarios. *Simulation Series*, *49*(7), 23–33. https://doi.org/10.22360/springsim.2017.mod4sim.003

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, *70*, 338–345. https://doi.org/10.1016/j.jbusres.2016.08.007

Jarwar, M. A., Ali, S., & Chong, I. (2019). Microservices based Linked Data Quality Model for Buildings Energy Management Services. *ArXiv Preprint ArXiv:1910.06115*.

Juddoo, S., & George, C. (2020). A Qualitative Assessment of Machine Learning Support for Detecting Data Completeness and Accuracy Issues to Improve Data Analytics in Big Data for the Healthcare Industry. *2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering, ELECOM 2020 - Proceedings*, 58–66. https://doi.org/10.1109/ELECOM49001.2020.9297009

Kanchi, S., Sandilya, S., Ramkrishna, S., Manjrekar, S., & Vhadgar, A. (2015). Challenges and Solutions in Big Data Management--An Overview. *2015 3rd International Conference on Future Internet of Things and Cloud*, 418–426.

Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). *Learning spark: lightning-fast big data analysis*. " O'Reilly Media, Inc."

Khalilijafarabad, A., Helfert, M., & Ge, M. (2016). Developing a Data Quality Research Taxonomy-an organizational perspective. *ICIQ*, 176–186.

Gökalp, M. O., Kayabay, K., Gökalp, E., Koçyiğit, A., & Eren, P. E. (2020). Towards a model based process assessment for data analytics: an exploratory case study. In Systems, Software and Services Process Improvement: 27th European Conference, EuroSPI 2020, Düsseldorf, Germany, September 9–11, 2020, Proceedings 27 (pp. 617-628). Springer International Publishing. Kowalczyk, M., & Buxmann, P. (2014). Big data and information processing in organizational decision processes: A multiple case study. *Business and Information Systems Engineering*, *6*(5), 267–278. https://doi.org/10.1007/s12599-014-0341-5

Krishna, C. M., Ruikar, K., & Jha, K. N. (2023). *Determinants of Data Quality Dimensions for Assessing Highway Infrastructure Data Using Semiotic Framework*.

Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015). A survey on data quality: classifying poor data. *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)*, 179–188.

Legner, C., Pentek, T., & Otto, B. (2020). Accumulating design knowledge with reference models: Insights from 12 years' research into data management. *Journal of the Association for Information Systems*, *21*(3), 735–770. https://doi.org/10.17705/1jais.00618

Lenzerini, M. (2018). *the Lens of an Ontology*. 65–75.

Li, B., & Gao, Q. (2019). Improving data quality with label noise correction. *Intelligent Data Analysis*, *23*(4), 737–757. https://doi.org/10.3233/IDA-184024

Lin, L., & Su, J. (2019). Anomaly detection method for sensor network data streams based on sliding window sampling and optimized clustering. *Safety Science*, *118*(February), 70–75. https://doi.org/10.1016/j.ssci.2019.04.047

Liu, C., Talaei-Khoei, A., Storey, V. C., & Peng, G. (2023). A Review of the State of the Art of Data Quality in Healthcare. *Journal of Global Information Management*, *31*(1), 1–18. https://doi.org/10.4018/JGIM.316236

Liu, He, Wang, X., Lei, S., Zhang, X., Liu, W., & Qin, M. (2019). A rule based data quality assessment architecture and application for electrical data. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3371425.3371435

Liu, Huan, Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, *6*(4), 393–423.

Loureiro, A., Torgo, L., & Soares, C. (2004). Outlier detection using clustering methods: a data cleaning application. *Proceedings of KDNet Symposium on Knowledge-Based Systems for the Public Sector*.

Luján-Mora, S., & Palomar, M. (2001). Reducing inconsistency in integrating data from different sources. *Proceedings of the International Database Engineering and Applications Symposium, IDEAS*, *February*, 209–218. https://doi.org/10.1109/ideas.2001.938087

Malik, V., & Singh, S. (2019). Big Data: Risk Management & Software Testing. *International Journal of Computational Intelligence & IoT*, *2*(1).

Meehan, J., & Zdonik, S. (2017). (jk-idea)(jk-related)Data Ingestion for the Connected World. *Cidr*. https://cs.brown.edu/courses/cs227/papers/brown-data-ingest.pdf

Mendes, P. N., Mühleisen, H., & Bizer, C. (2012). Sieve: linked data quality assessment and fusion. *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, 116–123.

Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016a). A Data Quality in Use model for Big Data. *Future Generation Computer Systems*, *63*, 123–130. https://doi.org/10.1016/j.future.2015.11.024

Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016b). A Data Quality in Use model for Big Data. *Future Generation Computer Systems*, *63*, 123–130. https://doi.org/https://doi.org/10.1016/j.future.2015.11.024

Mikalef, P., Pappas, I. O., Krogstie, J., & Pavlou, P. A. (2020). Big data and business analytics: A research agenda for realizing business value. *Information and Management*, *57*(1). https://doi.org/10.1016/j.im.2019.103237

Mittal, K., Aggarwal, G., & Mahajan, P. (2019). Performance study of K-nearest neighbor classifier and K-means clustering for predicting the diagnostic accuracy. *International Journal of Information Technology (Singapore)*, *11*(3), 535–540. https://doi.org/10.1007/s41870-018-0233-x

N, C. H., & Mahadevan, G. (2019). *Data Quality: An Assessment of Information Quality Dimensions*. 19–25. https://doi.org/10.17758/ERPUB5.ER09191004

Nakuçi, E., Theodorou, V., Jovanovic, P., & Abelló, A. (2014). Bijoux: Data generator for evaluating ETL process quality. *DOLAP 2014 - Proceedings of the ACM 17th International Workshop on Data Warehousing and OLAP, Co-Located with CIKM 2014*, 23–32. https://doi.org/10.1145/2666158.2666183

Nath, R. P. D., Hose, K., Pedersen, T. B., & Romero, O. (2017). SETL: A programmable semantic extract-transform-load framework for semantic data warehouses. *Information Systems*, *68*, 17–43.

Naumann, F., & Herschel, M. (2010). An introduction to duplicate detection. *Synthesis Lectures on Data Management*, *2*(1), 1–87.

Nayak, A., Božić, B., & Longo, L. (2021). (Linked) data quality assessment: An ontological approach. *CEUR Workshop Proceedings*, *2956*.

Nematzadeh, Z., Ibrahim, R., Selamat, A., & Nazerian, V. (2020). The synergistic combination of fuzzy C-means and ensemble filtering for class noise detection. *Engineering Computations (Swansea, Wales)*, *37*(7), 2337–2355. https://doi.org/10.1108/EC-05-2019-0242

Nguyen, A., Gardner, L., & Sheridan, D. (2018). Building an ontology of learning analytics. *Proceedings of the 22nd Pacific Asia Conference on Information Systems - Opportunities and Challenges for the Digitized Society: Are We Ready?, PACIS 2018*.

Nikiforova, A. (2020). Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. In *Baltic Journal of Modern Computing* (Vol.

8, Issue 3). https://doi.org/10.22364/BJMC.2020.8.3.02

Oliveira, B., & Belo, O. (2017). Approaching ETL Processes Specification Using a Pattern-Based Ontology. In *Data Management Technologies and Applications* (pp. 65–78). https://doi.org/10.1007/978-3-319-62911-7_4

Paganin, L. B. Z., Borsato, M., Schmidt, J., Domingos, A., & Sato, R. S. (2019). Test planning based on ontological models constructed from product usage profiles. *Product Management & Development*, *17*(2), 110–122. https://doi.org/10.4322/pmd.2019.015

Palareti, G., Legnani, C., Cosmi, B., Antonucci, E., Erba, N., Poli, D., Testa, S., & Tosetto, A. (2016). Comparison between different D-Dimer cutoff values to assess the individual risk of recurrent venous thromboembolism: Analysis of results obtained in the DULCIS study. *International Journal of Laboratory Hematology*, *38*(1), 42–49. https://doi.org/10.1111/ijlh.12426

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). *Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research*. Taylor & Francis.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218.

Quadir, B., Chen, N.-S., & Isaias, P. (2020). Analyzing the educational goals, problems and techniques used in educational big data research from 2010 to 2018. *Interactive Learning Environments*, 1–17.

Ramasamy, A., & Chowdhury, S. (2020a). Big Data Quality Dimensions: A Systematic Literature Review. *Journal of Information Systems and Technology Management*, *May*. https://doi.org/10.4301/s1807-1775202017003

Ramasamy, A., & Chowdhury, S. (2020b). Big Data Quality Dimensions: A Systematic Literature Review. *Journal of Information Systems and Technology Management*, *17*. https://doi.org/10.4301/s1807-1775202017003

Rashighi, M., & Harris, J. E. (2017). 乳鼠心肌提取 HHS Public Access. *Physiology & Behavior*, *176*(3), 139–148. https://doi.org/10.1053/j.gastro.2016.08.014.CagY

Ricamato, A., & Chicago, W. (2015). *Hyperients*. *2*(12).

Ruiz-Benito, P., Vacchiano, G., Lines, E. R., Reyer, C. P. O., Ratcliffe, S., Morin, X., Hartig, F., Mäkelä, A., Yousefpour, R., Chaves, J. E., Palacios-Orueta, A., Benito-Garzón, M., Morales-Molino, C., Camarero, J. J., Jump, A. S., Kattge, J., Lehtonen, A., Ibrom, A., Owen, H. J. F., & Zavala, M. A. (2020). Available and missing data to model impact of climate change on European forests. *Ecological Modelling*, *416*(March 2019), 108870. https://doi.org/10.1016/j.ecolmodel.2019.108870

Rula, A., & Zaveri, A. (2014). Methodology for Assessment of Linked Data Quality. *LDQ@ SEMANTICS*, 34.

Saha, B., & Srivastava, D. (2014). Data quality: The other face of big data. *2014 IEEE 30th International Conference on Data Engineering*, 1294–1297.

Sebastian-Coleman, L. (2012). *Measuring data quality for ongoing improvement: a data quality assessment framework*. Newnes.

Singh, S. K., & Dwivedi, D. R. K. (2020). Data Mining: Dirty Data and Data Cleaning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3610772

Ta'a, A., Abdullah, M. S., & Norwawi, N. M. (2011). *ETL Processes Specifications Generation Through Goal-Ontology Approach*. *061*, 8–9.

Taleb, I., Dssouli, R., & Serhani, M. A. (2015). Big Data Pre-processing: A Quality Framework. In *2015 IEEE International Congress on Big Data* (pp. 191–198). https://doi.org/10.1109/BigDataCongress.2015.35

Taleb, I., & Serhani, M. A. (2017). Big Data Pre-Processing: Closing the Data Quality Enforcement Loop. *2017 IEEE International Congress on Big Data (BigData Congress)*, 498–501.

Tang, M., & Liao, H. (2021). From conventional group decision making to large-scale group decision making: What are the challenges and how to meet them in big data era? A state-of-the-art survey. *Omega (United Kingdom)*, *100*, 102141. https://doi.org/10.1016/j.omega.2019.102141

Tepandi, J., Lauk, M., Linros, J., Raspel, P., Piho, G., Pappel, I., & Draheim, D. (2017). The data quality framework for the Estonian public sector and its evaluation: Establishing a systematic process-oriented viewpoint on cross-organizational data quality. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10680 LNCS*, 1–26. https://doi.org/10.1007/978-3-662-56121-8_1

Tesfagiorgish, D. G., & JunYi, L. (2015). Big Data Transformation Testing Based on Data Reverse Engineering. In *2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)* (pp. 649–652). https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.129

Tute, E., Scheffner, I., & Marschollek, M. (2021). A method for interoperable knowledge - based data quality assessment. *BMC Medical Informatics and Decision Making*, 1–15. https://doi.org/10.1186/s12911-021-01458-1

Urrutia, A., Chavez, E., Motz, R., & Gajardo, R. (2017). An Ontology to Assess Data Quality Domains. A Case Study Applied to a Health Care Entity. *IEEE Latin America Transactions*, *15*(8), 1506–1512. https://doi.org/10.1109/TLA.2017.7994799

van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2020). On the tractability of SHAP explanations. *ArXiv*.

van Gennip, Y., Hunter, B., Ma, A., Moyer, D., de Vera, R., & Bertozzi, A. L. (2018). Unsupervised record matching with noisy and incomplete data. *International Journal of Data Science and Analytics*, *6*(2), 109–129.

Vandenbussche, P. Y., Atemezing, G. A., Poveda-Villalón, M., & Vatant, B. (2017). Linked

Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. *Semantic Web*, *8*(3), 437–452. https://doi.org/10.3233/SW-160213

Vaziri, R., Mohsenzadeh, M., & Habibi, J. (2016). TBDQ: A Pragmatic Task-Based Method to Data Quality Assessment and Improvement. *PLoS One*, *11*(5), e0154508. https://doi.org/10.1371/journal.pone.0154508

Wahyudi, A., Kuk, G., & Janssen, M. (2018). A Process Pattern Model for Tackling and Improving Big Data Quality. *Information Systems Frontiers*, *20*(3), 457–469. https://doi.org/10.1007/s10796-017-9822-7

Wang, Richard YWang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5–33., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM*, *41*(2), 58–65.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Wang, Z., Talburt, J. R., & Wu, N. (2020). *A Rule-Based Data Quality Assessment System for Electronic Health Record Data*. 622–634.

Woods, C., Selway, M., Bikaun, T., Stumptner, M., & Hodkiewicz, M. (2021). *An ontology for maintenance activities and its application to data quality* (Vol. 0, Issue 0).

Xu, X., Lei, Y., & Li, Z. (2020). An Incorrect Data Detection Method for Big Data Cleaning of Machinery Condition Monitoring. *IEEE Transactions on Industrial Electronics*, *67*(3), 2326–2336. https://doi.org/10.1109/TIE.2019.2903774

Yin, X., Han, J., & Philip, S. Y. (2008). Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, *20*(6), 796–808.

Yuan, M., Hu, C., Liu, Y., & Mu, Y. (2019). Research on Quality Assessment Model of Big Data in Petroleum Field Based Ontology and Linked Data. *International Field Exploration and Development Conference*, 1274–1282.

Yulianto, A. A. (2019). Extract Transform Load (ETL) Process in Distributed Database Academic Data Warehouse. *APTIKOM Journal on Computer Science and Information Technologies*, *4*(2), 61–68. https://doi.org/10.11591/aptikom.J.Csit.36

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., & Auer, S. (2016). Quality assessment for linked data: A survey. *Semantic Web*, *7*(1), 63–93.

Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering Data Quality Problems: The Case of Repurposed Data. *Business and Information Systems Engineering*, *61*(5), 575–593. https://doi.org/10.1007/s12599-019-00608-0

Zhao, L., Chen, Z., Yang, Z., Hu, Y., & Obaidat, M. S. (2018). Local Similarity Imputation

Based on Fast Clustering for Incomplete Data in Cyber-Physical Systems. *IEEE Systems Journal*, *12*(2), 1610–1620. https://doi.org/10.1109/JSYST.2016.2576026

Zheng, Z. (2021). Contextual Data Cleaning with Ontology FDs. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21), June 20â•fi25, 2021, Virtual Event, China* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3448016.3450583

Zou, Q. (2018). Represent Changes of Knowledge Organization Systems on the Semantic Web. *International Journal of Librarianship*, *3*(1), 67–77.

# APPENDIX A

**Appendix A:** Mapping Codes for DQ ontology and the Study Datasets

```
import pandas as pd

import matplotlib.pyplot as plt

from matplotlib.gridspec import GridSpec

from rdflib import Graph, URIRef, RDF

import rdflib.plugins.sparql as sparql

import rdflib.plugins.sparql.parser as sparql_parser


dataset = pd.read_csv('goodreads_refine_test.csv')

ontology = Graph()

ontology.parse('data_quality_ontology.owl', format='xml')

query = """

   PREFIX dqo: <http://semantics.org/data-quality-ontology#>

   SELECT ?attribute ?dimension

   WHERE {

      ?attribute a dqo:DataAttribute .

      ?attribute dqo:hasQualityDimension ?dimension .

   }

"""
```

```
results = ontology.query(query)

attribute_dimension_mapping = {}

for row in results:

    attribute_uri, dimension_uri = row

    attribute_name = str(attribute_uri).split('#')[-1]

    dimension_name = str(dimension_uri).split('#')[-1]

    attribute_dimension_mapping[attribute_name] = dimension_name

def calculate_column_completeness(data):

    total_rows = len(data)

    completeness = 10 + 90 * (1 - (data.isnull().sum() / total_rows))

    return completeness

def calculate_row_completeness(data):

    total_columns = len(data.columns)

    row_completeness = 10 + 90 * (1 - (data.isnull().sum(axis=1) / total_columns))

    return row_completeness

def check_data_consistency(data):

    data_types = data.infer_objects().dtypes

    inconsistent_attributes = []

    for column in data.columns:

        inferred_dtype = data[column].infer_objects().dtype
```

```
        if inferred_dtype != data.dtypes[column]:
            inconsistent_attributes.append(f"Inconsistent Data Type: {column}")
    return inconsistent_attributes
def check_duplicates(data):
    duplicate_attributes = []
    for column in data.columns:
        if data[column].duplicated().any():
            duplicate_attributes.append(column)
    return duplicate_attributes


column_completeness = calculate_column_completeness(dataset)
row_completeness = calculate_row_completeness(dataset)
inconsistent_attributes = check_data_consistency(dataset)
duplicate_attributes = check_duplicates(dataset)


column_completeness_threshold = 10
row_completeness_threshold = 10
consistency_threshold = 10


overall_completeness = column_completeness.mean()
```

```
columns_below_threshold = column_completeness[column_completeness < column_completeness_threshold]

rows_below_threshold = row_completeness[row_completeness < row_completeness_threshold]

overall_consistency = 100 if not inconsistent_attributes else 100 - (len(inconsistent_attributes) / len(dataset.columns) * 100)

if duplicate_attributes:
    overall_consistency -= 10

fig = plt.figure(figsize=(8, 6))

gs = GridSpec(3, 1, height_ratios=[2, 1, 1])

ax1 = plt.subplot(gs[0])

ax1.bar(["Completeness", "Consistency"], [overall_completeness, overall_consistency], width=0.4, color=['skyblue', 'green' if overall_consistency >= consistency_threshold else 'red'])

ax1.set_xlabel('Dimension')

ax1.set_ylabel('Score')

ax1.set_title('Data Quality Metrics')

ax1.set_ylim(10, 100)  # Set the y-axis scale from 10 to 100


# Subplot for the Table
```

```
ax2 = plt.subplot(gs[1])
ax2.axis('off')

table_data = [["Completeness", f"{overall_completeness:.2f}%"],
              ["Consistency", f"{overall_consistency:.2f}%"]]
table_data.extend([["Columns Below Threshold", f"{len(columns_below_threshold)}"],
                   ["Rows Below Threshold", f"{len(rows_below_threshold)}"]])
table_data.extend([[attribute, "Inconsistent"] for attribute in inconsistent_attributes])

table = ax2.table(cellText=table_data, cellLoc='center', colLabels=["Dimension", "Value"],
loc='center')
table.auto_set_font_size(False)
table.set_fontsize(12)
table.scale(1.2, 1.2)

plt.tight_layout()
plt.show()
```

# APPENDIX B

**Appendix B:** Codes for Clustering

```
import pandas as pd

import matplotlib.pyplot as plt

from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn.cluster import KMeans

from sklearn.decomposition import TruncatedSVD

import numpy as np

import rdflib

ontology_path = 'data_quality.owl'

graph = rdflib.Graph()

graph.parse(ontology_path, format='xml')

data = pd.read_csv('good_reads.csv')

text_data = data['Authors']

text_data = text_data.dropna()  # Drop rows with missing values

text_data = text_data[text_data != ''']  # Drop rows with empty strings

text_data.reset_index(drop=True, inplace=True)

if text_data.empty:

    print("No valid text data to cluster.")
```

```
else:
    tfidf_vectorizer = TfidfVectorizer()
    tfidf_matrix = tfidf_vectorizer.fit_transform(text_data)
    kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
        cluster_labels = kmeans.fit_predict(tfidf_matrix)
    data.loc[text_data.index, 'Cluster'] = cluster_labels
    svd = TruncatedSVD(n_components=2)
    tfidf_2d = svd.fit_transform(tfidf_matrix)
    cluster_colors = np.array(['red', 'green', 'blue'])
    point_colors = cluster_colors[cluster_labels]
    fig, axs = plt.subplots(2, 1, figsize=(8, 10))
    axs[0].scatter(tfidf_2d[:, 0], tfidf_2d[:, 1], c=point_colors)
    axs[0].set_title('K-means Clustering (Authors)')
    axs[0].set_xlabel('Principal Component 1')
    axs[0].set_ylabel('Principal Component 2')
    cluster_summary = pd.DataFrame({'Cluster': range(3)})
    cluster_summary['Number of Data Points'] = [sum(cluster_labels == i) for i in range(3)]
    axs[1].axis('off')  # Turn off axis for the second subplot
    axs[1].table(cellText=cluster_summary.values,
                 colLabels=['Cluster', 'Number of Data Points'],
```

```
            cellLoc='center', loc='center')

plt.tight_layout()

 plt.show()
```