Journal of Theoretical and Applied Information Technology
15th May 2023. Vol.101. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DATA QUALITY ASSESSMENT USING TDQM FRAMEWORK: A CASE STUDY OF PT AID

TEDI WAHYUDI[1], SANI MUHAMMAD ISA[2]

[1]Student, Department of Computer Science, Bina Nusantara University, Jakarta, Indonesia

[2]Lecture, Department of Computer Science, Bina Nusantara University, Jakarta, Indonesia

E-mail: [1]tedi.wahyudi@binus.ac.id, [2]sani.m.isa@binus.ac.id

ABSTRACT

Data has a significant meaning for a business and is used to guide objectives and decision-making. In order to produce high-quality data, a corporation must be able to analyze and manage the data properly. This is a challenge for companies, especially those with a wide variety of data sources, because of the risk of increasing the inaccuracy of the data they have, which can result in making inappropriate decisions. Data processing and acquisition activities are carried out by AID, a company that specializes in "Data as A Service," spanning twelve business units with different business lines that are managed by a conglomerate group that mostly serves the financial services industry. The current state of the organization presents many challenges, as data sources still lack standardization and control, or monitoring of data completeness and accuracy, and the organization has never measured the quality of existing data. In order to obtain quality of data, it is necessary to apply specific methods, processes and techniques, to measure the data. The approach taken in this study to evaluate the quality of the data are Total Data Quality Management (TDQM) and the six dimensions from the DAMA white paper. The results of this evaluation procedure can be used to examine the company's existing data quality and to provide recommendations for changes that need be made internally. The results showed that the quality of data owned by the company was at the threshold of a very high-quality level. Additionally, it is envisaged that this data quality assessment can be applied to all business units and conducted on a regular basis.

Keywords: *Data Quality, Total Data Quality Management, TDQM, data quality dimension, DAMA*

1. INTRODUCTION

In this all-digital era, data collections can come from various sources, both internal and external to the company, having a very large volume, various types, and data generated at high speed requires different techniques from ordinary data transactions. This data set, known as “Big Data”, is expected to provide the required information as well as add value to the company. In searching for this information, traditional databases will not be able to solve all aspects of Big Data, namely "3V" - Volume, Velocity, and Variety [1]. Miloslavskaya, & Tolstoy added four other “Vs” to the Big Data criteria including Veracity, Variability, Value, and Visibility [2].

The Data Lake concept is emerging as a popular way to organize and build next-generation systems to address Big Data challenges. Data lakes can manage and use data with increased volume, variety, and speed rarely seen before, which makes companies strive to implement them [3]. Organizations are adopting the Data Lake model because it provides raw data that can be used to perform data experiments, advanced analytics, and scalable storage to handle growing amounts of data and provide the agility to provide insights more quickly [4].

One of a company's most precious assets is its data. If the data is handled properly, it will result in information that the company can utilize to make decisions. [5]. A company is expected to have the ability to analyze and manage data properly so as to produce quality data. Data quality plays a crucial part in all business and government applications [6]. This is a challenge for companies, especially for companies that have many varied data sources because they risk increasing the inaccuracy of the data they have and can lead to inappropriate decision-making.

AID is a company engaged in the Data as a Services (DaaS) sector, which is responsible for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

carrying out data acquisition and management activities from business units (BU) under the auspices of a conglomerate group that focuses on the financial services sector. Data from business units is stored in a Data Lake company for further use for analysis, reporting, prediction, and modeling purposes. As illustrated in figure 1, there are fifteen business units from various types of businesses such as leasing, financing (cars, motorcycles, heavy equipment, electronics, etc.), insurance, venture funds, pension funds, transportation and logistics, peer-to-peer lending, e-wallet, and digital applications. The results of the analysis, prediction, or modeling will be reused by the business units for their various business needs.

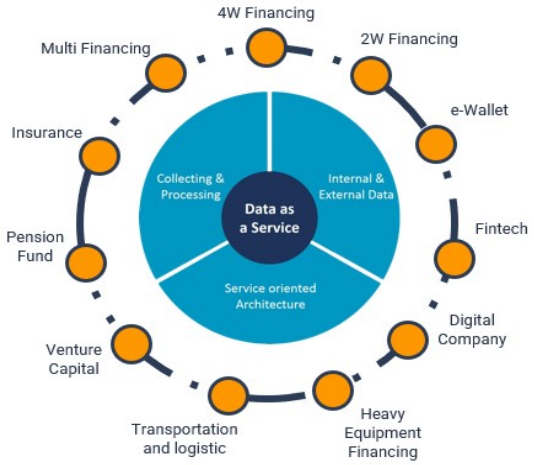

Figure 1: AID company overview

The company is currently facing many difficulties because the company has never measured quality of the data in the Data Lake and data sources from various business units still lack standardization, control, or monitoring of the completeness and accuracy of the data, which causes many inaccuracies in the results of analysis, predictions, reports or dashboards. Therefore, a method is needed to measure the quality of these data. Poor data quality greatly affects decision-making by companies that are less than optimal. Poor data quality can also result in compliance risk, namely when data quality standards do not match the expectations of supervisory authorities [7]. Poor data quality can have a negative impact on data utilization efficiency, leading to serious decision-making mistakes [8]. In data integration scenarios, data quality is a major challenge. Data integration and quality have a mutually beneficial relationship [9].

"Garbage in – garbage out" (GIGO) is a general term that is often used in the field of computer science or information and communication technology, referring to incoming data that is of poor quality which will lead to unreliable data output results. The information collected must be highly accurate otherwise data analysis, applications, or business processes will not be reliable [10]. The consequences of poor data can range from significant to catastrophic. Data quality problems can cause projects to fail, resulting in lost revenue, reduced customer relationships, and even lost customers [11].

This case study was conducted to evaluate the data quality in the AID Data Lake, build up data quality rules and metrics, and recognize where data quality is substandard and must be rectified to enhance the aspects of quality, accuracy, availability, and integrity.

2. LITERATURE REVIEW

2.1 Data Quality

The term Data Quality (DQ) refers to the characteristics associated with high-quality data and the processes used to measure or improve data quality [12]. Data quality is defined as "fitness for use" and its assurance is recognized as a valid and important activity, but in practice few people list it as the highest priority [13]. Data quality is defined as the degree to which the data is relevant, timely, accurate, full, and up to date in accordance with all business regulatory requirements [14]. Data Quality encompasses the extent to which data is reliable, accurate, applicable, applicable to the given context, easily understandable, and timely [15]. In the early development stage, how to determine data and data quality is an important aspect that varies in the literature. This includes context, nature, and data type [16].

2.2 TDQM

Total Data Quality Management (TDQM) can be considered as the first approach related to data quality proposed by Wang [17]. TDQM is inspired by Total Quality Management (TQM) which is used for product quality. The methodology views data (or information) as product entities, as there is an analogy between product manufacturing and information manufacturing, i.e., products produced from raw materials by assembly lines, and in the same way, information generated from raw data by Information Systems. Therefore, quality problems can be solved in the same way [18]. TDQM is a comprehensive and structured approach

to organizational management in improving data quality [19].

In business and manufacturing the Deming product cycle (or PDCA cycle) has proposed, through the Plan, Do, Check and Act cycle, the manufacturing process is continuously improved [18]. By the same method, the TDQM cycle includes four main phases as shown in figure 2.

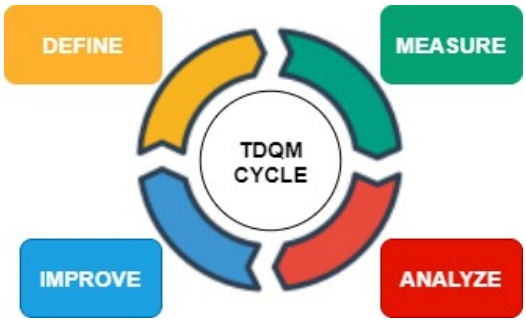

Figure 2: TDQM Framework.

2.3 Data Quality Dimension

The data quality dimension is a collection of data quality attributes that represent one aspect of data quality [20]. DAMA defines a data quality dimension as a feature or characteristic of data that can be measured [12]. The term dimension is defined as "a measurable level of a certain type, such as length, width, depth, or height". Dimension deals with measurement or, in other words, is the quantification of the characteristics of an object or phenomenon [21]. If quality is "a distinctive attribute or characteristic possessed by someone or something" then the data quality dimension is a general category that can be measured for the typical characteristics (quality) possessed by the data [22].

Many researchers proposed various types of dimensions to define and assess data quality. In Plotkin's book describes eight dimensions of data quality such as completeness, uniqueness, validity, reasonableness, integrity, timeliness, coverage, and accuracy [23]. DAMA issued a white paper describing the six core dimensions of data quality consist of completeness, uniqueness, consistency, validity, accuracy, and timeliness [12]. Meanwhile Loshin describes the definitions of various types of data rules which are reflected in the following seven dimensions [24]. The most common dimensions are completeness, timeliness, and accuracy, followed by consistency and accessibility [16]. The data quality of a system is evaluated using several dimensions in the literature on data quality. Accuracy, completeness, validity, uniqueness, consistency, etc. are some of the most often used metrics to assess the data quality of systems. [25]

To measure the level of data quality, it is necessary to select dimensions that are relevant to certain business processes [26]. In determining the dimensions of data quality, there are many other aspects that may be specific to an industry, such as internal information policies, line of business levels, data source conditions, and data needs in an organization. Thus, the determination of dimensions needs to be tailored to the organization as it will be incorporated into metrics and protocols to assess and monitor key data quality performance factors [7]. Dimensions must match the actual state of the organization in order to be able to measure data quality [8]. The use of data quality dimensions varies depending on the business needs, the context in which the data is used, and the industry involved [27]. Business requirements, database design analysis results, and characteristic of data are taken into consideration while determining measurement dimensions [28].

2.4 Data Quality Metrics

Generally, metrics used to evaluate data quality tend to span a range of 0 to 1, with 0 indicating a value that is incorrect and 1 indicating a value that is accurate. The following formula is used for precise evaluation of various dimension, such as completeness, uniqueness, accuracy, and consistency:

$$D = 1 - (Ni/Nt) \quad (1)$$

While for the dimension of timeliness, it is evaluated according to the subsequent formula:

$$D = \text{current Time} - \text{update Time} \quad (2)$$

Where D is the metric for a particular dimension, Ni is the number of noncompliant values, and Nt is the total number of values for the relevant dimension [12][29][30].

3. RESEARCH METHODOLGY

The research begins by identifying and exploring the problems. The analysis is carried out on the existing problems and their impact on the organization. The literature study stage is carried out by conducting a review of previous research, including methods that can be used to measure and improve data quality [6], data quality assessment in university [28], data quality assessment in higher

education [31][32], data quality assessment in telecommunications company [33], data quality assessment in National Scientific Repository [34], data quality assessment in sharia companies [35], data quality assessment on open government data [36], data quality assessment using MDM [37], and data quality assessment on national remote sensing data bank [38].

The method used in this research is a mixed method of qualitative and quantitative. Qualitative data collection methods were carried out through interviews with relevant teams, and document observations. As for the quantitative method, the researcher will conduct data profiling against the target table in organization Data Lake. The goal is to enable researchers to analyze the data patterns and values contained in the target table and help researchers discover data quality rules and requirements to support further assessment of the overall data quality.

Currently, the company uses Hadoop platform technology running on a private cloud as its Data Lake. The data that is evaluated for quality focuses on three major tables: customer, contract, and financial data. Apache Hue is leveraged as an interface to analyze data quality through HiveQL scripts. The steps followed to conduct this study covers the first three steps of the TDQM method. Subsequently, the stages undertaken in this study have been collated and outlined demonstrated as follows:

Stage 1 : Analyze the primary table to ascertain which attributes will be gauged for data quality.

Stage 2 : Comprehend and expound upon the characteristics of each attribute that will be evaluated to formulate the data quality rules.

Stage 3 : Identify the criteria for assessing data quality, taking into consideration the dimensions of the data.

Stage 4 : Establish thresholds for each rule and dimension used to measure the quality of the data.

Stage 5 : Specify the quality metrics for each of the applied dimensions.

Stage 6 : Measuring data quality by performing direct queries according to predefined criteria.

Stage 7 : Investigate the results of data quality assessments to identify the underlying cause of rules with low data quality levels.

In the final stage, recommendations are given for solutions that will be implemented by the organization so that in the future it can improve the quality of its data. The steps of the research method used can be seen in Figure 3.

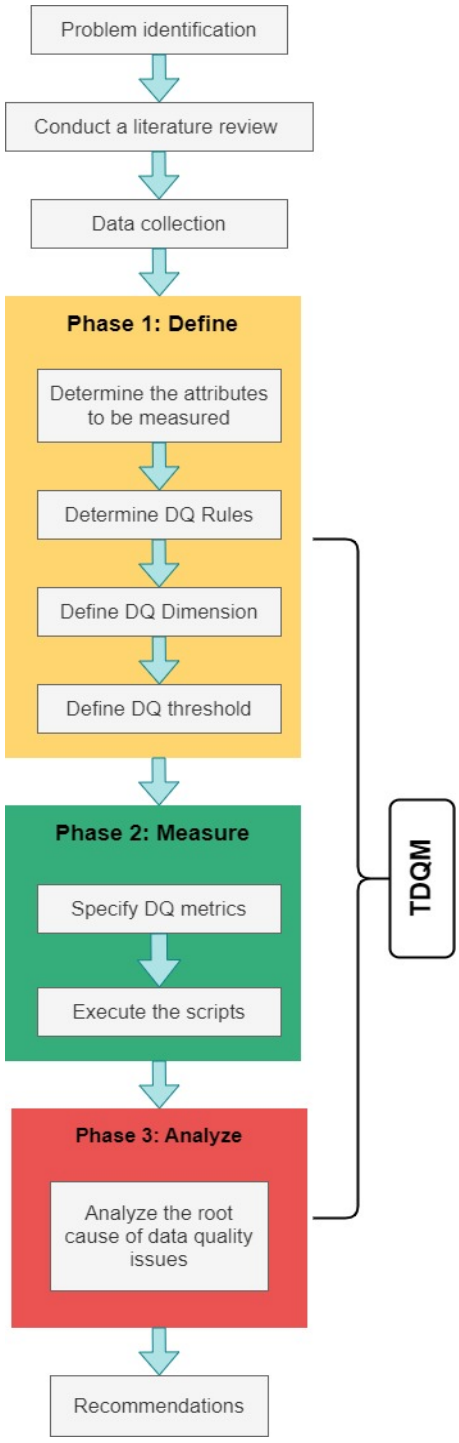

Figure 3: The steps of research method.

4. RESULT AND DISCUSSION

4.1 Define Phase

The total attributes used in this measurement are 80 out of 87 total attributes derived from three main tables. The AID Data Governance team is accountable for identifying and defining these attributes. The team's consideration in determining the attributes to be measured is to see how crucial and often they are used in analysis, dashboard, or machine learning process. The distribution of attributes can be seen more detail in Figure 4 below.

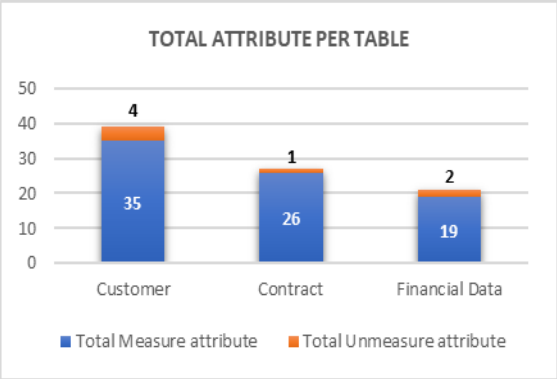

Figure 4: Total attributes to be measured.

By utilizing the 80 pre-defined attributes, 173 data quality rules were generated based on the details obtained from data profiles, literature reviews, and interviews. This implies that a single attribute can generate one or more data quality rules. Refer to Figure 5 for a broader summary of the allotment of data quality rules based on their origin information.

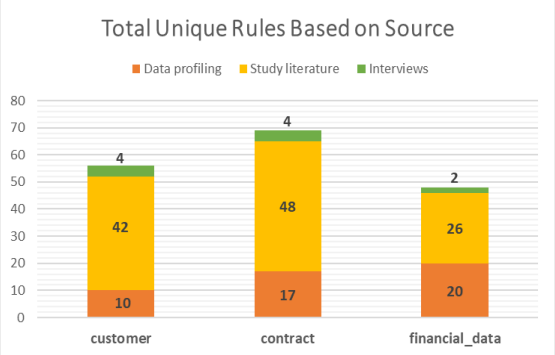

Figure 5: Total unique rules generated.

In addition, the status of data ownership is taken into consideration when deciding which rules should apply to each business unit. Five out of fifteen business units that are not represented in the customer, contract, and financial data tables. Three among those five business units are no longer integrated with AID. Another factor taken into account is to determine whether the columns in the main table owned by AID have a mapping with the columns in the table on the business unit side. The determination of this column can be obtained either from the results of interviews or by observing the Master Mapping document. Thus, the total number of measured data quality rules reached 1,460 rules.

In this study, the DQ dimensions used adhere to the six dimensions outlined in the 2013 DAMA white paper, which are completeness, uniqueness, accuracy, consistency, validity, and timeliness [12]. Using these six dimensions can enable a comprehensive assessment of the data that can be evaluated for both quality and suitability for AID's current objectives. Following, the corresponding DQ dimensions were aligned to the applicable DQ rules.

The AID Data Governance team also plays a key role in establishing and demarcating the threshold parameters of each data quality rule and its dimensions. The threshold implementation is divided into 3 levels: High (green), Medium (yellow), and Low (red). Due to the variety of sources and data conditions of the business units that integrate with AID, it is essential to have adjustable threshold levels for each data quality rule and dimension to meet the changing needs of the business. The consideration used to determine the threshold for each data quality rule is looking at the attribute's critical level. Additionally, as outlined in Table 1, the team assigned the same thresholds for each dimension except for the timeliness dimension.

Table 1: Mapping between DQ Dimension and Threshold

Dimension	Threshold		
	Low	Medium	High
Completeness	x < 75%	75% <= x < 85%	x >= 85%
Uniqueness			
Consistency			
Accuracy			
Validity			
Timeliness	x < 100%	-	x = 100%

Table 2 in the appendix presents the metrics utilized to calculate the score of each data quality rule based on its corresponding dimension category. In contrast to the other dimensions, timeliness dimension has a distinct formula in determining data quality values.

Table 2: Data Quality Metrics

4.2 Measurement Phase

4.2.1 Completeness Dimension

The quality rule that fall under this dimension generally verifies whether the measured column contains incomplete values such as partial, null, or empty values. The following columns are deemed significant in the evaluation process:

1. primary key table.
2. partition key table.
3. Personal Identifying Information (PII) columns for instance customer ID number, customer name, address, date of birth, mobile phone number, email address, etc.
4. Amount columns such as amount finance, outstanding principal amount, outstanding interest amount, installment amount, tenor, etc.
5. Date columns such due date, maturity date, go live date, create date, last update, etc.
6. Columns that has a reference to the Master Attribute table. The Master Attribute table contain lists all attributes utilized in the main tables. For example customer status, customer type, gender, contract status, line of business, installment type, etc.

The information in Table 3 is presented that the completeness dimension holds the most significant number of data quality rules, comprising 530 in total. The average value of this dimension is depicted in Figure 6 as 85.31%, which is at an adequate level, but there is still potential for improvement, especially in business units L, O, and R.

Table 3: Total rules on each BU in the completeness dimension

4.2.2 Uniqueness Dimension

In this dimension, the data quality rules usually constitute the attributes that serve as primary keys in the table. However, the financial data table is an exception to that rule, as it does not posses one. The data quality rules were established through observation and analysis of the metadata documents:

1. Each business unit in the customer table must possess a uniquely assigned customer ID value.
2. The contract id value in the contract table must be unique for each business unit.

The overall compliance of this dimension is demonstrated through the assessment of its 18 rules, which can be seen in Table 4. All rules are at an excellent level of data quality and achieve a score of 100% as shown in Figure 6, indicating that the ingestion process is functioning optimally without any duplication or inconsistency.

Table 4: Total rules on each BU in the uniqueness dimension

4.2.3 Consistency Dimension

Data profiling is an instrumental process in uncovering the data quality rules in this dimension. As detailed in the Table 5, this makes the consistency dimension the second highest in terms of the number of data quality rules, behind the completeness dimension with 447 rules. This dimension yielded an average value of 80.99%, as per Figure 6.

Table 5: Total ruleson each BU in the consistency dimension

The following are a few things to consider measuring in this dimension:

1. Comparing the minimum, maximum, and average values of data in AID with data in business units. The data compared is data with the types number and date.
2. Cross Table Validation Rules, these rules looks at the relationships between column sets in various tables using foreign key analysis, which identifies orphaned data and determines semantic and syntactic

differences. In addition to reducing redundancy, this can reveal data value sets that can be mapped together. For example in this case, the columns which has reference with Master Attribute Table.

4.2.4 Accuracy Dimension

Table 6 reveals that 12 rules were generated in this dimension, deriving from the contract table. The level of this dimension is classified as medium, as evidenced by an average value of 75.35%. All business units, except for business unit K, have a rule that is exhibiting a low standard of data quality.

Table 6: Total rules on each BU in the accuracy dimension

The criteria of the data quality rules that determine this scope include:

1. If there is contract status is in early repayment state then the close date value equal to maturity date.
2. The next due date value should be bigger than due date value.
3. The next ins sequence number value should be bigger than the ins sequence number value.

4.2.5 Validity Dimension

Table 7 demonstrates that 357 rules have been defined and are being used to measure quality. The average score for this dimension was 89.31% at completion as display at Figure 6. Master Metadata document are integral to establishing data quality rules within this domain.

Table 7: Total rules on each BU in the validity dimension

By comparing the current data with established rules, it is possible to determine whether the system's data quality meets this criterion. Among the guidelines established in this dimension are:

1. PII columns should be in hash format.
2. The length of hashed value should be 32 digit.
3. The length of zipcode should be 5 digit.
4. All date columns must be in YYYY-MM-DD format.
5. Amount columns should be in decimal format and cannot be minus (-).
6. The tenor value cannot be less than 1.

4.2.6 Timeliness Dimension

As listed in table 8, the number of data quality rules contained in the timeliness dimension is 61 rules. The measurement results show very satisfactory results where all rules get a score of 100% ass displayed in Figure 6. This means that the integration process in AID is running very well, as expected.

Table 8: Total rules on each BU in the timeliness dimension

The parameters for the data quality rules applied in this dimension are:

1. The values in the create date and last update columns of the customer table must have an updated date based on the execution date.
2. The values in the go live date, create date and last update columns of the contract table must have an updated date based on the execution date.
3. The values in the create date and last update columns of the financial data table must have an updated date based on the execution date.

The amount of gap time is determined by the type of customer and whether the integration process is daily or monthly. If a business unit with an individual customer type, the gap time used is 2 days. As for business units with corporate customer types, the gap time used is 1 month. This rules are obtained from the results of interviews with the Data Engineer team. If calculated value from the formula is higher than the gap value, then the DQ Score will be 0, otherwise the DQ Score will be 100.

4.3 Analyze Phase

A number of measurements representing different dimensions of data quality have been identified; however, there are underlying factors contributing to problematic data. These issues can be identified via the examination of integration processes and interviews with Data Engineering teams as the team who holds the responsibility for designing, constructing, and sustaining the infrastructure and data flow. A variety of factors can impact the quality of data, including:

1. No standardization of values is observed, particularly for the critical columns that contain null or empty values.
2. It was discovered that certain values in the columns of the primary table did not correspond to the values in the Master Attribute table.
3. The transformation process is encountering a configuration error, leading to data that is either lost or recorded incorrectly.
4. Processes within the Business Unit that allow for the acceptance of null or empty values are in place. One example of this finding is that one of the business units involved in the digital application sector does not require date of birth to be a field that must be completed, whereas AID uses this attribute to be part of its process in constructing customer golden records.
5. Migrating from the prior system to the new one on the business unit side is a potential cause of diminished data quality.
6. One of the reasons for low data quality is that the dataset or table does not contain any data that satisfies the criteria of the data quality rules. This situation typically occurs in the accuracy, consistency, and validity dimensions and is brought on by the measured attribute having null or empty values. One could say that the those dimensions are dependent on completeness dimension. Kaiser pointed out that when attribute values are incomplete, they are automatically inconsistent and inaccurate, thus leaving consistency and accuracy metrics inapplicable in the case of incomplete information [39].

4.4 Recommendations

The results derived from conducting data quality measurements on the three central tables of the AID Data Lake in the context of this case study suggest that there are a few strategies which can be applied to optimize the data quality:

1. The findings of this research should be implemented through the establishment of comprehensive policies which prioritize the standardization of any column containing null or empty values, as well as mandating data owners/business units to consistently fill in crucial details such as ID card number, customer name, and date of birth.
2. Periodically maintain and update the mapping in master reference tables.
3. The AID Data Governance team should coordinate with the corresponding Data Governance team of the business unit to discuss and implement an improvement plan for this finding, as the data belongs to the business unit. Especially for the critical attributes that are not mandatory to fill on the business unit side.
4. It is highly recommended that all currently owned documents, such as the Master Metadata and Master Mapping documents, are kept up to date. Currently, these documents are tracked manually in excel files, therefore, the company should explore automated alternatives and implement a Data Catalog tool that aligns with the company's needs and conditions.
5. It is critical to evaluate the accuracy of data from business unit sources, as this data is gathered from multiple core systems within the business unit, modified to meet AID's requirements.

5. CONCLUSION

Based on measurements of data quality that were made on three main tables at AID Data Lake using the TDQM method and six core dimensions from DAMA (completeness, uniqueness, accuracy, consistency, validity, and timeliness), indicates that the quality of AID data is at a satisfactory level with a DQ score of 88.49 percent. The average percentage for the dimension completeness of the 530 data quality rules measured was 85.31 percent. The average score for 17 rules in

the accuracy dimension is 75.35 percent. While in the consistency dimension, 80.99 percent is the average value obtained from 447 eligible rules. In the validity dimension, 387 data quality rules show that the average valid data that meets the specified criteria is 89.31 percent. The average percentage for the timeliness dimension was 85.31 percent from the 61 data quality rules that were measured. As for the dimensions of uniqueness and timeliness, both are at the excellent level of 100 percent compliant with the rules.

The result of this study shows that TDQM can be used as a data quality management strategy by measuring the quality. The scope of measurement is limited to three main tables that are in a structured format and the process is still in batch form. Going forward, data quality assessment will not be limited solely to the customer, contract, and financial data tables but will also apply to all primary tables in the AID Data Lake. Moreover, this will extend to all business units under this group.

REFERENCES:

[1] Madden, Samuel. (2012). From Databases to Big Data. IEEE Internet Computing - INTERNET. 16. 4-6. 10.1109/MIC.2012.50

[2] Miloslavskaya, N., & Tolstoy, A. (2016). Big Data, Fast Data and Data Lake Concepts. Procedia Computer Science, 88, 300–305. doi: 10.1016/j.procs.2016.07.439

[3] Fang, H. (2015). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. 2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER). doi:10.1109/cyber.2015.7288049.

[4] Singh, A., & Ahmad, S. (2019). Architecture of data lake. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 5(2).

[5] Hikmawati, Sanny & Santosa, Paulus & Hidayah, Indriana. (2021). Improving Data Quality and Data Governance Using Master Data Management: A Review. IJITEE (International Journal of Information Technology and Electrical Engineering). 5. 90. 10.22146/ijitee.66307.

[6] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41, 3, Article 16 (July 2009), 52 pages. DOI = 10.1145/1541880.1541883 http://doi.acm.org/10.1145/1541880.1541883.

[7] Loshin, D. (2011). Evaluating the business impacts of poor data quality. Knowledge. Integrity Incorporated Bus. Intell. Solutions, Silver Spring, MD, USA, Tech. Rep.

[8] Cai, L., & Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, 2. DOI: http://doi.org/10.5334/dsj-2015-002.

[9] Endler, G., Schwab, P. K., Wahl, A. M., Tenschert, J., & Lenz, R. (2015). An architecture for continuous data quality monitoring in medical centers. In MEDINFO 2015: eHealth-enabled Health (pp. 852-856). IOS Press.

[10] Kilkenny, M. F., & Robinson, K. M. (2018). Data quality: "Garbage in – garbage out.". Health Information Management Journal, 47(3), 103 – 105. https://doi.org/10.1177/1833358318774357.

[11] Gudivada, V., Apon, A., & Ding, J. (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. In: International Journal on Advances in Software 10.1, pp. 1 - 20.

[12] Gabr, M.I., Helmy, Y.M., & Elzanfaly, D.S. (2021). Data Quality Dimensions, Metrics, and Improvement Techniques. Future Computing and Informatics Journal.

[13] Data Management Association., Henderson, D., Earley, S., Sebastian-Coleman, L., Sykora, E., & Smith, E. (2017). DAMA-DMBOK: Data management body of knowledge. 2nd Edition. Bradley Beach, NJ: Technics Publications, LLC.

[14] Tayi, G.K. & Ballou, D.P. (1998). Examining data quality, Communications of the ACM, Vol. 41 No. 2, pp. 54-7.

[15] Mosley, M., & Data Administration Management Association. (2008). The DAMA dictionary of data management. 1st Edition. Bradley Beach, NJ: Technics Publications, LLC.

[16] Cichy, C., & Rass, S. (2019). An Overview of Data Quality Frameworks. IEEE Access, vol. 7, pp. 24634 - 24648. doi: http://doi.org/10.1109/ACCESS.2019.2899751.

[17] Wang, R. Y. (1998). A product perspective on total data quality management. Commun. ACM, vol. 41, no. 2, pp. 58-66.

[18] Vaziri, R. (2012). A Questionnaire-Based Data Quality Methodology. International Journal of Database Management Systems, 4(2), 55–68. doi:10.5121/ijdms.2012.4204.

[19] Veiga, A. K., Saraiva, A. M., Chapman, A. D., Morris, P. J., Gendreau, C., Schigel, D., & Robertson, T. J. (2017). A conceptual framework for quality assessment and management of biodiversity data. PLOS ONE, 12(6), e0178731. doi: 10.1371/journal.pone.0178731.

[20] Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. J. Manage. Inf. Syst., vol. 12, no. 4, pp. 5-33.

[21] Jayawardene, V., Sadiq, S., & Indulska, M. (2015). An analysis of data quality dimensions.

[22] Sebastian-Coleman, L. (2013). Measuring data quality for ongoing improvement: A data quality assessment framework. Chapter 4: Data Quality and Measurement (pp 39-53). Amsterdam: Elsevier.

[23] Plotkin, David. (2014). Data stewardship: an actionable guide to effective data management and data governance. Waltham.

[24] Loshin, D. (2011). The Practitioner's Guide to Data Quality Improvement. Burlington, MA:: Morgan Kaufmann.

[25] Ramasamy, Anandhi & Chowdhury, Soumitra. (2020). Big Data Quality Dimensions: A Systematic Literature Review. 10.4301/S1807-1775202017003.

[26] Jugulum R. (2016) Importance of Data Quality for Analytics. In: Sampaio P., Saraiva P. (eds) Quality in the 21st Century. Springer, Cham. https://doi.org/10.1007/978-3-319-21332-3_2.

[27] Mahanti, Rupa. (2019). Data quality: Dimensions, measurement, strategy, management and governance. ASQ Quality Press, Milwaukee.

[28] Cahyono, S. H., & Sucahyo, Y. G. (2020). Pengukuran Kualitas Data Menggunakan Framework Total Data Quality Management (TDQM): Studi Kasus Sistem Informasi Beasiswa Universitas Indonesia (Data Quality Assessment Using the TDQM Framework: A Case Study of University of Indonesia (UI) Scholarship Information System). JURNAL IPTEKKOM (Jurnal Ilmu Pengetahuan & Teknologi Informasi), 22(2), 193-206. doi: http://dx.doi.org/10.33164/iptekkom.22.2.2020.193-206.

[29] Juddoo, S. (2015). Overview of data quality challenges in the context of Big Data. In 2015 International Conference on Computing, Communication and Security (ICCCS) (pp. 1-9). IEEE.

[30] Serhani, M. A., El Kassabi, H. T., Taleb, I., & Nujum, A. (2016). An Hybrid Approach to Quality Evaluation across Big Data Value Chain. 2016 IEEE International Congress on Big Data (BigData Congress). doi:10.1109/bigdatacongress.2016.

[31] Prasetyo, R. T., Ruldeviyani, Y., Purnamasari, E. D., & Wibowo, A. F. (2021). Data Quality Assessment on Lecturer Primary Data: A Case Study on Higher Education Database at Ministry of Education and Culture Republic of Indonesia. IOP Conference Series: Materials Science and Engineering, 1077(1), 012036. doi:10.1088/1757-899x/1077/1/012036.

[32] Wijayanti, W., Hidayanto, A. N., Wilantika, N., Adawati, I. R., & Yudhoatmojo, S. B. (2018). Data Quality Assessment on Higher Education: A Case Study of Institute of Statistics. 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). doi:10.1109/isriti.2018.8864476.

[33] Andini, I. D., Ruldeviyani, Y., Maulana, A. H., & Hidayat, A. (2020). Penilaian Kualitas Data Broadband Customer Profiling (BCP) Pelanggan Fixed Broadband PT Telekomunikasi Indonesia Tbk. (Data Quality Assessment of Broadband Customer Profiling (BCP) of Fixed Broadband Customer of PT Telekomunikasi Indonesia Tbk.). JURNAL IPTEKKOM (Jurnal Ilmu Pengetahuan & Teknologi Informasi), 22(1), 31-43.

[34] Riyanto, S., Marlina, E., Subagyo, H., Triasih, H., & Yaman, A. (2020). METODE PENILAIAN KUALITAS DATA SEBAGAI REKOMENDASI SISTEM REPOSITORI ILMIAH NASIONAL. Baca: Jurnal Dokumentasi dan Informasi, 41, 11-22. doi: 10.14203/J.BACA.V41I1.544.

[35] Bowo, W.A., Suhanto, A., Naisuty, M., Ma'mun, S., Hidayanto, A.N., & Habsari, I.C. (2019). Data Quality Assessment: A Case Study of PT JAS Usi ng TDQM Framework. 2019 Fourth International

Conference on Informatics and Computing (ICIC), 1-6.

[36] Li, X.-T., Zhai, J., Zheng, G.-F., & Yuan, C.-F. (2018). Quality Assessment for Open Government Data in China. Proceedings of the 2018 10th International Conference on Information Management and Engineering - ICIME 2018. doi:10.1145/3285957.3285962.

[37] Hikmawati, S., Santosa, P., & Hidayah, I. (2021). Improving Data Quality and Data Governance Using Master Data Management: A Review. IJITEE (International Journal of Information Technology and Electrical Engineering). 5. 90. 10.22146/ijitee.66307.

[38] Payani, A. S., Mayanda, A. M., Adiresta, A., & Ruldeviyani, Y. (2022). Data Quality Management Strategy To Improve Remote Sensing Data Quality: A Case Study On National Remote Sensing Data Bank. IPTEK The Journal for Technology and Science, 33(3), 162-174.

[39] Kaiser, M. (2010). A conceptional approach to unify completeness, consistency, and accuracy as quality dimensions of data values. In European and Mediterranean Conference on Information Systems (EMCIS 2010). Abu Dhabi, UEA: Academic Press.

Journal of Theoretical and Applied Information Technology
15th May 2023. Vol.101. No 9

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FIGURES:

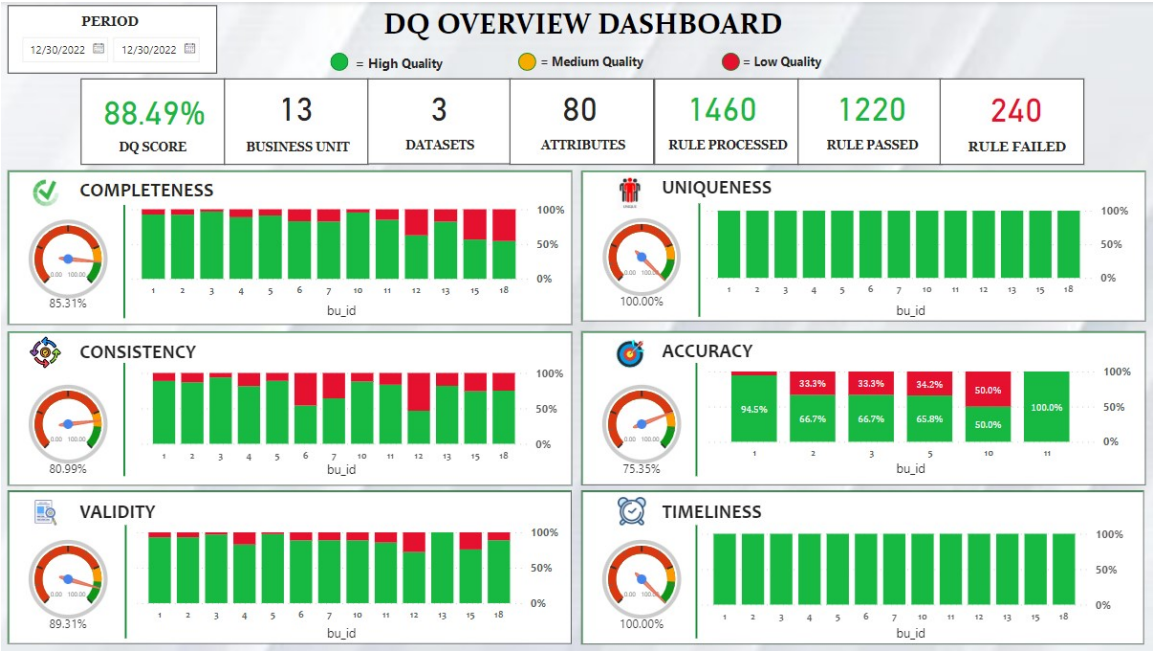

Figure 6: Summary result of Data Quality Measurement

TABLES:

Table 2: Data Quality Metrics

Dimension	Metrics	Definition
Completeness	(1 - (Number of incomplete rows / total of rows)) * 100%	The proportion of null/empty/incomplete records to the overall number of records in the examined characteristic.
Uniqueness	(1 - (Number of non-unique rows / total of rows)) * 100%	The proportion of non-unique values to the overall quantity of values in the recorded attribute.
Consistency	(1 - (Number of inconsistent rows / total of rows)) * 100%	The proportion of records that do not match the original source to the total number of records in the examined feature.
Accuracy	(1 - (Number of incorrect rows / total of rows)) * 100%	The proportion of inaccurate records in relation to the aggregate amount of records present in the evaluated attribute.
Validity	(1 - (Number of invalid rows / total of rows)) * 100%	The proportion of records that do not conform to the prescribed format compared with the entirety of records associated with the attribute being observed.
Timeliness	current Time – update Time	Calculates the difference between the expected time and the current data.

Table 3: Total rules on each BU in the Completeness Dimension

TABLE	QUALITY LEVEL	BU_ID																	
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
CUSTOMER	HIGH	18	16	18	N/A	17	14	14	N/A	N/A	N/A	17	10	14	N/A	13	N/A	N/A	9
	MEDIUM	0	0	0		0	1	1				0	0	0		0			0
	LOW	2	0	1		3	3	1				1	7	3		4			8
CONTRACT	HIGH	22	21	25	16	21	N/A	N/A	N/A	N/A	23	22	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MEDIUM	0	2	0	0	0					0	0							
	LOW	4	3	1	3	4					1	3							
FINANCIAL DATA	HIGH	19	17	18	14	18	N/A	N/A	N/A	N/A	18	10	11	N/A	N/A	7	N/A	N/A	N/A
	MEDIUM	0	0	0	0	0					0	0	0			0			
	LOW	0	2	1	3	1					1	7	8			10			

Table 4: Total rules on each BU in the Uniqueness Dimension

TABLE	QUALITY LEVEL	BU_ID																	
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
CUSTOMER	HIGH	1	1	1	N/A	1	1	1	N/A	N/A	N/A	1	1	1	N/A	N/A	1	N/A	1
	MEDIUM	0	0	0		0	0	0				0	0	0			0		0
	LOW	0	0	0		0	0	0				0	0	0			0		0
CONTRACT	HIGH	1	1	1	1	1	N/A	N/A	N/A	N/A	1	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MEDIUM	0	0	0	0	0					0	0							
	LOW	0	0	0	0	0					0	0							
FINANCIAL DATA	HIGH	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MEDIUM																		
	LOW																		

Table 5: Total rules on each BU in the Consistency Dimension

TABLE	QUALITY LEVEL	BU_ID																	
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
CUSTOMER	HIGH	15	13	15	N/A	13	9	9	N/A	N/A	N/A	14	5	9	N/A	N/A	14	N/A	12
	MEDIUM	0	1	0		0	0	2				1	0	0			0		0
	LOW	2	3	2		4	8	6				2	12	2			3		4
CONTRACT	HIGH	17	18	19	11	17	N/A	N/A	N/A	N/A	14	16	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MEDIUM	0	0	0	0	0					0	0							
	LOW	4	2	2	2	1					3	3							
FINANCIAL DATA	HIGH	15	14	14	11	15	N/A	N/A	N/A	15	11	9	N/A	N/A	N/A	10	N/A	N/A	N/A
	MEDIUM	0	0	1	0	0				0	0	0				0			
	LOW	1	2	1	3	1				1	3	7				4			

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 6: Total rules on each BU in the Accuracy Dimension

TABLE	QUALITY LEVEL	BU_ID																	
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
CONTRACT	HIGH	2	2	2	N/A	2	N/A	N/A	N/A	N/A	1	3	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MEDIUM	1	0	0		0					0	0							
	LOW	0	1	1		1					1	0							

Table 7: Total rules on each BU in the Validity Dimension

TABLE	QUALITY LEVEL	BU_ID																	
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
CUSTOMER	HIGH	18	18	18	N/A	17	14	14	N/A	N/A	N/A	16	13	14	N/A	14	N/A	N/A	14
	MEDIUM	0	0	0		0	0	0				0	0	0		0			0
	LOW	0	0	0		1	4	4				2	5	2		4			4
CONTRACT	HIGH	12	13	14	8	13	N/A	N/A	N/A	N/A	12	12	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MEDIUM	0	0	0	0	0					0	0							
	LOW	3	2	1	0	1					1	2							
FINANCIAL DATA	HIGH	11	10	11	7	11	N/A	N/A	N/A	N/A	9	7	6	N/A	N/A	7	N/A	N/A	N/A
	MEDIUM	0	0	0	0	0					0	0	0			0			
	LOW	0	1	0	3	0					2	3	5			4			

Table 8: Total rules on each BU in the Timeliness Dimension

TABLE	QUALITY LEVEL	BU_ID																	
		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
CUSTOMER	HIGH	2	2	2	N/A	2	2	2	N/A	N/A	N/A	2	2	2	N/A	2	N/A	N/A	2
	MEDIUM	0	0	0		0	0	0				0	0	0		0			0
	LOW	0	0	0		0	0	0				0	0	0		0			0
CONTRACT	HIGH	3	3	3	3	3	N/A	N/A	N/A	N/A	3	3	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	MEDIUM	0	0	0	0	0					0	0							
	LOW	0	0	0	0	0					0	0							
FINANCIAL DATA	HIGH	2	2	2	2	2	N/A	N/A	N/A	N/A	2	2	2	N/A	N/A	2	N/A	N/A	N/A
	MEDIUM	0	0	0	0	0					0	0	0			0			
	LOW	0	0	0	0	0					0	0	0			0			