JOIV : Int. J. Inform. Visualization, 8(4) - December 2024 2195-2202

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Improving Data Reliability Assessment in ETL Processes through Quality Scoring Technique in Data Analytics

Nor Famiera Atika Razali [a], Salmi Baharom [a,*], Salfarina Abdullah [a], Novia Indriaty Admodisastro [a]

[a] *Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia*

Corresponding author: [*]salmi@upm.edu.my

***Abstract*— The foundation of a relevant and accurate data analysis is reliable data. Technique and measurement are essential to evaluate current data quality regarding reliability and establish a baseline for ongoing improvement initiatives. Without tools or visualizations, data engineers may find it challenging to monitor and maintain the reliability of the massive data from the extraction, transformation, and loading (ETL) data load process. Data reliability assessment is a helpful technique in analyzing the quality of data reliability and information on the present state of data before commencing any analytics. The proposed technique hinges on the metric and measurement defining data reliability and the dashboard platform where the integration with the user in dictating the weight of data and the final output, which is the final data reliability score, will be projected. The score obtained affirms whether improvements are needed on the data or if an organization can proceed with data analytics. The technique considers the data extraction, transformation, and loading (ETL) procedures used to gather datasets. Data significance or weight was determined according to the analytics needs and preferences, indicating an acceptable score for generating insights. Ultimately, when utilizing the data reliability assessment metrics technique, we are credited with an overall picture of our data's reliability aspect, as only one look is offered based on the intended analysis. This new approach boosts the confidence among data practitioners and stakeholders, especially those relying on findings generated from data analysis. Furthermore, the overview assists in enhancing the current state of data, where the derived score helps identify possible areas of improvement in the ETL process. Accuracy and efficiency assessment of the proposed technique also showed positive feedback in measuring the method in measuring the reliability of data.**

***Keywords*— Data reliability; extraction; transformation; data reliability metrics; data weight.**

Manuscript received 11 Jan. 2024; revised 9 Aug. 2024; accepted 22 Oct. 2024. Date of publication 31 Dec. 2024.

I. INTRODUCTION

Data analytics involves using various techniques and tools to analyze large datasets to extract valuable insights meant to be leveraged in making data-driven decisions. This can enhance processes and procedures in a variety of domains and sectors. It is a useful approach for businesses trying to make the most of the abundance of data at their disposal to obtain a competitive advantage and make wise decisions. Inadequate data quality can diminish the effectiveness of data utilization and potentially result in substantial errors during decision-making processes [1]. Having trustworthy data at the very beginning is crucial in producing insightful analysis. Hence, data reliability is one of the critical aspects of data analytics. A data reliability quality dashboard is a tool or interface that provides an at-a-glance view of the quality, specifically data reliability within an organization, to benefit all data practitioners. The data's reliability will be assessed based on the importance of the selected data fields in the analytics, hence using the Data Reliability Assessment Metrics technique. The outcome and score produced will serve as a baseline or indicator for the data practitioners in their primary area of focus.

Managing data quality is notably challenging when it is poor since it complicates the decision-making processes in organizations and their day-to-day operations. Data entered is often inaccurate, incomplete, or contains inconsistencies when captured across different systems. Therefore, it is unsuitable for making sound analyses and thus prevents our collection of valuable information. It is, therefore, essential to fix this issue so that trust can be rebuffed regarding our data, decisions, and overall organizational performance. While it may be relatively easy to define the data collection process, most collected data often lack consistency and are unreliable [1]. As the amount of data generated from various sources grows and becomes 'big data,' data profiling and quality management are gaining currency.

The previous methods have not been effectively used for data quality management. The introduction highlights that data quality management has been an area of concern for research and development since the 19th century, and different frameworks and methodologies have been developed to address data quality problems. However, it has also been noticed that many commercial data quality management tools are costly, come with a complicated interface, and can be less easy to integrate [2]. The quality of Big Data continues to pose a substantial challenge, given the presence of errors, inconsistencies, and inaccuracies within numerous data sources, potentially affecting the precision of the derived insights [3]. In addition, even the latest studies fail to pay much attention to data weight as a parameter essential for reliability measurements [4]. They strongly support judging the data fields by their importance, not by equal estimations of all fields. On the other hand, data visualization is helpful when analyzing data elements as it provides ease of use, while data quality control tools play a vital role in oversight [5]. A basic dashboard for monitoring current conditions is even more appropriate for managing current problems, and addressing them effectively is encouraged [6]. On the other hand, it was mentioned that there were issues with processing large volumes of data from various sources and ensuring this was done quickly and efficiently because of inadequate insight [7].

In the opinion of data analytics in industrial use, the prior work on defining and assessing data quality concerns either encompassed large numbers of data quality dimensions or considered only dimensions irrelevant to data usage in industrial analysis and data management. Thus, in addition to the work done for the calls of the research work, what was shown in the practicality part for the actual industrial test cases has narrowed the gap [8]. The proposed solution will be used as a benchmark or reference point for the data practitioners. But again, it should be noted that to improve the data quality, data engineers might need to modify and make some changes to the extract, transform, and load (ETL) if the score requires the standard level set by the business entity or the project.

It can also identify to data engineers which aspect of achieving quality data needs further enhancement from the total ETL process. For data analysts, it offers an early look at the state and condition of the data before proceeding with data processing, as it is a first step and a prerequisite to creating timely and meaningful analyses. From the business analyst's standpoint, such scores will assist them in coming to certainty that such information and conclusions are based on such reliable data.

This paper is organized as follows: Section II reviews related material and the proposed solution's method, Section III demonstrates the implementation and discusses the result, and Section IV states the conclusion.

II. Materials and Method

A. Extract, Transform, Loading (ETL) Processes

Data analysis involves massive amounts of data originating from various sources, so the insight can be descriptive, explorative, diagnostic, predictive, or prescriptive analysis that will suit the organization [9]. Using traditional ways to process such massive data and different resources to deal with such issues is very challenging. The actual or prescribed activities of the old ways of data processing were found to be less efficient and sometimes erroneous. ETL is quite an essential component of a data warehouse, which starts with extracting data from multiple sources, transforming it if necessary, and loading it into the data warehouse [10]. To facilitate data management and preprocessing, ETL is a single tool combining three distinct yet crucial processes [11].

a. Extract: This stage focuses on collecting information from different source systems. Data extraction can be done on big data sources, unstructured data, and structured databases.
b. Transform: After extraction, the data is refined and normalized to the chosen grading for evaluation and analysis. The transformation step may also encompass other activities such as data aggregation, computation, and enrichment.
c. Load: The converted data is placed in a data lake, warehouse, or other storage context. This stage ensures that the data is compiled and cleaned up for reporting and analysis.

In this way, a good ETL process should involve extracting data from various sources, converting the data to ensure reliability and conformity, and loading the data into the intended database [12]. The amount of work required to gather the data is enormous; besides, the data comes in different formats, which analyzes big data a challenging endeavor. It is essential to clean the collected data to increase its quality. Still, it cannot guarantee that the enhanced significant data quality will address the needs of big data applications [13]. Therefore, the assessment of data quality is much needed.

B. Other Data Loading Process vs ETL Process

Of all the data process methods, the ETL (Extract, Transform, Loading) commonly appears in data analytics because it is effective in managing numerous data types and processing raw data to make it ready for analysis. It is beneficial for joining potentially unrelated datasets and translating data to improve its quality. Some papers focusing on the benefits of ETL also prove that it can help resolve issues of making the data available in a consistent format and appropriateness for all systems or big data or streaming data [14]. ETL goes beyond basic data loading processes, such as file-based loading, where raw data are loaded directly into data storage systems with minimum transformation activities. If the primary need is the mere uploading of data into a system, then the file-based loading is satisfactory; however, for analytic purposes, ETL has the added advantage of cleansing the information and putting it into a standardized format.

On the other hand, API-based loading might load data much faster for quick data retrieval and enable communications between systems but does not cater to optimization for the analytical domains required with history. The involved ETL process in data science extends beyond simply data extraction and loading. Instead, it involves data transformation, which ensures accurate insights from quality data for strategic decision-making. It is not just about moving vast masses of data or enabling rapid access; it focuses on

establishing confidence in decision-making based on filtered data sets. Lastly, ETL ensures that analysts get credible and determinative data to work with, extracting rich information to be used by decision-makers.

C. Data Quality

The main problem with data quality that the researchers have pointed out is the problem of dirty data, which can be due to various reasons, including human error and technical malfunctions. Such poor-quality data can negatively affect the decision's reliability and result in incorrect decisions. Thus, there is a need to utilize strategies that may enhance data quality. Assessing the data quality is an important and starting point for improving this quality [15]. Some of the concerns were even about the quality of the data. It was said that the data was incompletely biased, thus causing some concern about the reliability of the data. Nevertheless, digitization entails challenges that include data inaccuracies and the absence of information because of a standard validation database [16]. Data quality includes aspects such as the accuracy of the data, the extent to which data is comprehensive and free from inconsistencies, and how updated and credible it is. Assessing data quality involves determining individual parameters for these aspects. It enables accurate, reliable decision-making and does not require additional manual input and manipulation, thus cutting time and other resources [7].

Furthermore, a lack of good-quality data will negatively impact the quality of predictive analytics [17]. The main challenge with data quality that they identified is the consequences of data quality error or misinterpretation on the effectiveness and excellence of the big data application. Big data must be collected and analyzed to help increase its efficiency and produce the practical results anticipated. First, the amount of data that needs to be collected is immense, and the nature of data can be varied, making analyzing big data even more challenging. Therefore, it is crucial to enhance the quality of the available data for effective and efficient implementation of big data solutions [18]. Data quality issues, according to findings made by Springboard.com [19], revealed that such costs are estimated to be around $9 for an organization. About 7 million annually reduces the worker's productivity by 20%, making 40% of business goals and objectives unachieved. Essentially, it is clear that data quality is a critical success factor in the organization, influencing decision-making, organizational operations, customer relations, legal requirements, and organizational performance. However, data quality's negative aspects include reduced productivity, wastage of marketing budgets, high storage expenses, and a lack of coordinated customer visions [6]. Accurate data is essential for providing sound guidance for obtaining organizational goals and aims. The internal implementation of data quality into ETL processes ensures that data accurate to business requirements is acquired and preserved in consistency, accuracy, completeness, and clarity via actions such as data cleansing and rectification [10].

D. Data Reliability

A solid infrastructure and procedures that ensure consistency, accuracy, and trustworthiness are necessary to produce reliable data. Reliability is characterized as the measure of trustworthiness in data, encompassing accuracy, consistency, and integrity [20]. Reliability is also one of the terms used to describe the quality dimension of trust, believability, and reputation and identify the extent to which data originates from authoritative sources [21]. Furthermore, reliability refers to whether we can trust the data [1]. Moreover, data integrity and volatility are the elements that define data reliability [7]. Data integrity refers to the correctness of the data, and data volatility refers to how long the retention data will still be considered valid. They also shared how they calculate the score for these two metrics. Uniqueness is one of the metrics in the data reliability dimension in research conducted by Kristyanti et al. [6]. They calculate the percentage of the duplicated data, and the only tolerable percentage is less than 0%. The completeness of data is also one of the elements considered when discussing data reliability [22].

Data reliability is sustained and enhanced over time by adherence to best practices within ETL workflows, process improvement, and ongoing monitoring. ETL might deliver data reliability accurately and consistently by managing processes of extracting, transforming, and loading data with heightened concern. ETL ensures that only clean data gets to the transformation stage by cleaning it before transforming it by three subprocesses, which include validation checks and deleting inconsistencies. In addition, loading validation and verification methods employed during the loading process ensure that the alterations are loaded accurately into the group, thus making it more reliable and freer from errors. Moreover, poor design of the applications that utilize the data warehouse or even the lousy quality of the source data might not affect the overall effectiveness of ETL; nonetheless, any mistakes in this process might lead to skewed data being written into a data warehouse [23]. This could threaten the quality of the data acquired, including the reliability and credibility of the result.

E. Existing Dashboard

A data quality dashboard, a specialized type of dashboard, displays information about an organization's data quality. It presents key metrics and insights regarding timeliness, accuracy, completeness, and consistency. These dashboards provide a quick overview of data health, including quality scores, error tracking, trend analysis, and potential areas for improvement. Dashboards help interpret data quality findings [24]. During the data visualization stage, the results of big data analysis should be provided in the form of charts and formulas. This allows the senior decision-makers in the user unit to appreciate and review the analysis better, thereby accelerating the progression of the following tasks [18]. Furthermore, using the dashboard will help people understand the quality of the published datasets, reducing the time taken to assess the data quality [25].

Talend and Ataccama are two tools that offer a good way of showcasing data quality on dashboards. ETL tool – Talend has the characteristic of providing data-quality dashboards. However, Ataccama is not only an ETL tool but a toolbox with many functionalities related to the ETL process, improving the quality of data in use. A data quality tool called Looker Block for Talend Studio has been created in collaboration with Talend and Looker [26]. The article "What

is Data Health" by Talend [27] explains ways of measuring these dimensions of data health. They include the data quality sum, the number of rows passed through the ETL pipeline, the number of passed and failed rows, and data quality trends over time and by departments. Unfortunately, it is unclear whether data weight determines their scores. The dashboard determines the data quality by summing the results. It also does not specify which columns it uses to make such a determination in the dataset.

Ataccama, as a data quality platform, mainly deals with data formatting, normalization, filtering, identification, and sorting. Their solution, Ataccama ONE, is a state-of-the-art, fully autonomous data management and governance solution. It allows users to start the application with essential features and add more per company requirements. Data quality's default dimensions are validity, completeness, uniqueness, and accuracy [28]. As is the case with Talend, they do not consider data weight, and the final score is computed for all the columns in the datasheet.

F. Evaluation Assessment Strategy

Widad et al. [7] conducted an assessment evaluation to gauge the precision and scalability of the methodology. Empirical testing yielded an accuracy score of 99.91% and an F-score of 98.07%. Conversely, Byabazaire et al. [29] research uses Pearson Correlation coefficients to investigate the relationship between Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of the findings. This was considered when incorporating their newly proposed measurements and verifying the high link between the indicators. However, the comparative analysis was used as the evaluation method in the investigations by Widad et al. [7]. A questionnaire is one of the tools used in the research conducted by Vaidyambath et al. [24] to evaluate the method's usability and effectiveness. McCarthy et al. [30] research is assessed based on the speed and accuracy of the procedure using both human and automated methods

G. Analysis and Problem Definition

Based on the implication of the literature review conducted, we can see that there is no specific approach to assess the reliability of data and ETL for business intelligence and analytics. Another aspect is that despite numerous suggested measurements for evaluating data quality, including dashboard options, attention is not given to the significance of data weight in generating an accurate score that users can define based on selected columns. Answering stakeholders' questions regarding data reliability and the insights' trustworthiness is challenging. Finally, the lack of a specific technique makes it difficult to identify the appropriate measurements and calculate the final score. The case studies we have created for this purpose have helped arrive at the data reliability scores using proposed metrics and the weight of data.

Regarding the proposed solution, we selected the metrics most often cited to examine reliability. We chose the five most important qualities based on their definition and described a way to combine these measures. We developed the procedures to capture user preferences for arriving at the final score. Finally, all the metrics and their measurements, the weight of the data, and the platform were integrated into one single process.

H. Design and Development

Calculation of scale measures for data reliability culminated in formulating the data reliability assessment metrics. These encompassed accuracy, integrity, completeness, uniqueness, and timeliness. There was also a scale to measure the weight of the data to be collected. It provided the final reliability score as the average of all the reliability metric scores combined with the data weight of the reliability analysis.

Components for our prototype are the set of chosen datasets obtained from the ETL process automatically, finalized metrics measuring data reliability, and the implementation of the data weight scale for the final rating. These steps are crucial before designing our solution prototype. For our prototype, we extract data from APIs using scripts in Python, and after pre-processing, the collected data is inserted into the database automatically. These results can then be exported and displayed as a dashboard for further interaction with the data. Software development entails understanding the software's requirements, creating the interface layout, coding using PHP, CSS, and Javascript, and testing.

In terms of designing the metrics, each metric would have a dedicated measurement considering the Extract Transform Loading (ETL) data processing base. Below is the measurement for each metric, referring to the definition from existing literature and the situation reflected in the ETL pipeline design.

1) Completeness: This metric calculates non-empty rows or NULL, which indicate incomplete data or missing records. The calculation proceeds by dividing the count of non-empty rows in the column by the total count of records in the table and then multiplying the outcome by 100 [4].

$$Completeness\ (\%) = \frac{Number\ rows\ with\ no\ NULL}{Total\ number\ of\ rows}\ X\ 100 \qquad (1)$$

2) Uniqueness: The uniqueness metric will oversee the table's data duplication or redundancy level for whole columns. The same data with more than one occurrence will be counted, impacting its scoring [4].

$$Uniqueness\ (\%) = \frac{Number\ of\ unique\ rows\ of\ data}{Total\ values}\ X\ 100 \qquad (2)$$

3) Timeliness: Timeliness metrics will measure the number of rows that are not current according to ETL schedule updates.

$$Timeliness\ (\%) = 100\ X\ (1 - \frac{Number\ of\ Rows\ Not\ Updates}{Total\ Number\ of\ Rows}) \qquad (3)$$

4) Accuracy: Accuracy metrics measure the data between two schemas, staging and the user layer. It may not be 100% accurate if some data from the source needs cleansing. However, we can overlook this as we understand and know where the score comes from upon checking. The main objective of this metric is to evaluate the precision of how different data is from the source itself and the data that has

been transformed. The same attribute within the table column will be compared in both schemas.

$$Accuracy\ (\%) = 100\ X\ \frac{Number\ of\ Matching\ Rows}{NULLIF(Number\ of\ rows\ in\ source\ ,0)} \quad (4)$$

5) Integrity: Integrity will be checked based on the validity of columns in the table and the referential integrity, that is, the relationship between tables. This will consider the primary or foreign key for invalid reference and the data out of validity for each column. Mathematically, this can be defined as below:

- VPK is the count of valid primary keys.
- VNN is the count of valid non-null columns.
- VRC is the count of valid columns.
- TR is the total number of rows.
- TNV is the total number of validity check types.
- CT2 is the count of invalid references in the second table.
- TT2 is the total number of rows in the second table.
- TNC is the total number of validity checks and invalid reference checks.

$$Integrity\ (\%) = \left(\frac{1}{TNC}\left(\frac{100\times(VPK+VNN+VRC)}{TR\times TNV}+ 100\times\left(1-\left(\frac{CT2}{TT2}\right)\right)\right)\right) \quad (5)$$

A. Evaluation

The proposed technique will be tested empirically by comparing the results obtained with the existing method in terms of accuracy and efficiency. The following evaluation is based on case studies detected in problem analysis and an additional test item. Precise and meticulous calculations will be conducted to mitigate potential validity concerns. The acquired evaluation results were assessed, analyzed, and interpreted. The accuracy analysis evaluates the existing method and the newly proposed solution using the Pearson Correlation Coefficient to determine the correlation. In the context of efficiency, the time required to compute the reliability score between the new and existing techniques was compared. Using the left-tailed test, whether statistically significant differences in the time taken by the latest tools from the existing technique were tested.

III. Results and Discussion

A. Database Design

Regarding the database design, we have implemented two separate schema structures that serve as the separation layer, facilitating the seamless data flow from its source to its destination. Although each table, column, connection, and constraint would have its distinct entity in this schema separation, for this project, the schemas are nearly identical, mirroring each other except for a few additional rows in the main schema that hold transformed data. The staging schema is where the first incoming data from sources is initially stored. Its primary purpose is to load and lodge the source's original raw data efficiently. Since this data was obtained straight from the source with slight modifications, we can state that it is the most recent information available. In addition to serving as the temporary repository, this staging schema is prepared for modification for future user usage.

On the other hand, the user layer, or Production schema, is the main repository that the user can access, and it's known as the Production schema. The data in this schema is typically in the final, usable state, allowing the user to query and make inferences. The data in the production schema generally has undergone much processing, several transformations, or even some cleaning that may have been overlooked when scripts were running to retrieve the data from the API.

B. Data Acquisition

In gathering the data using ETL, the pipeline starts by recording the package's start time of the operation. Not only is this needed for audit purposes, but it also helps us to know when, where, and how long the process will take. Next, the staging area directly connected to the source table is truncated to create a clean environment for data ingestion, making the table always have the latest and fresh data upon the data acquisition process. A Python script enclosed in a batch file is executed to retrieve and clean the data. These are the scripts that are being mentioned previously where it directly has a connection with the source data and destination database. After completing data retrieval and cleansing in the staging layer, the staged data seamlessly combines with the main database table, or, in our development term, the production schema, the user layer. This integration process combines the cleaned data into the main layer, ensuring the database remains updated with the latest information. Once the data is loaded in staging and production, the pipeline ends with recording the package end time for the same reason as recording the start time. An example of these processes is briefly illustrated in Fig 1.

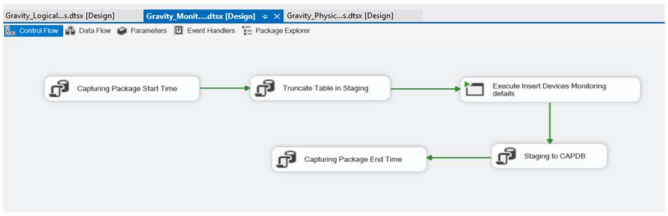

Fig 1 Snippets of SSIS packages in automating ETL processes

C. Data Reliability Metrics Assessment

The proposed data reliability assessment metrics solutions are divided into the following categories. We are focusing on metrics frequently mentioned in the literature as a beginning. Table 1 lists all the top 5 metrics in the literature review, mentioning data reliability.

TABLE I
PROPOSED DATA RELIABILITY ASSESSMENT METRICS

Metrics	Occurrence	Author/Tool
Accuracy	9	[1], [7], [10], [15], [20], [22], [25], Ataccama, Talend
Completeness	6	[1],[7],[20], [25], Ataccama, Talend
Uniqueness	6	[6], [7], [15], [25], Ataccama, Talend
Integrity	4	[1], [4], [15], [21]
Timeliness	4	[4], [10], [25], Talend

D. Data Weight

The Likert scale, initiated in 1832 by Rensis Likert, is famously known as one of the psychometric tools used to

evaluate human attitudes, opinions, and perceptions [31]. Since dictating the preference of data or considering which data has a higher weight than others is based more on the data practitioner's view or opinion regarding what data they need to use, the Likert scale is used in developing the data reliability assessment metrics tool. Data weighting makes it possible to weight data based on the importance of each data dimension and how it affects the data quality. This is made possible by allocating proper weights to these dimensions so that the assessment exercise better reflects the actual effects of each dimension on the general quality of the data [15].

In this development context, the range of scale can be referred to as below:

a. Strongly Disagree (1): The information is not essential, but it is not wrong to include it in the analysis, maybe just for a description or supporting details. The inexistence of this data would not interfere with any decision or conclusion that needs to be drawn.
b. Disagree (2): Data that are "good to have" in producing insight and lacking quality are the slightest concerns.
c. Neutral (3): "Good to have" data and if it will be good if it is in good shape.
d. Agree (4): the kind of data supporting the decision's main details extracted.
e. Strongly Agree (5): This data with this weight of data is significant in integrating each table to conclude. Primary and foreign key IDs and the most prominent data or tables are examples of dictating with a higher scale range. It is like the attribute that makes up the whole purpose of having reliable insight. For example, having a 5 rating on timeliness ensures data use is updated.

Existing studies also integrated the calculation with weightage; however, the ranking ranges from 1 – 10, 1 from less essential fields from the whole dataset in the table [4]. For this study, the weight is limited to the range 1-5 due to the lack of much of a difference in scoring for both range scales with the data we selected to make the deduction. This was captured in our little experiment to determine the scale. On the other hand, the 1-5 scale is much easier and simpler to define compared to the existing research. The approximate value of the differences in using the existing scale and the newly proposed is 0.65. This can be considered when benchmarking other situations unless there are specific guidelines within the respective industries or the context of the situation that would require otherwise.

E. Metric and Weight Calculations

Every metric score we measure and cumulate will have specific data weight, and the final score should comprise both equation parameters. The following procedure determines the overall metric score from a given row of metrics and their weight. The first operation is to obtain the metrics and weights of every row. Then, add all the individual weights to get the total weight. Apply this total to estimate the proportion of each row's overall weight out of the whole weight; this indicates the weight's impact. Multiply each row's metric value by weight proportion to get the row score. Finally, the sum of all the values present in rows would decide the overall for one specific metric score. The last score is an average of all the weights in the rows with additional weighting based on their level of importance. The same applies to our final reliability calculation of all the metrics, including accuracy, completeness, uniqueness, integrity, and timeliness.

F. Dashboard

The output of this project, visualized as the Data Reliability Metrics Assessment Dashboard, showcases the calculated metrics. Each metric calculation is unique and varies from one another. By clicking on the "Data Reliability Metrics" tab, users will see a series of buttons representing completeness, uniqueness, timeliness, accuracy, integrity, and overall reliability. Begin by choosing the row from the column that will be used to create data-driven decision-making visualizations or the information needed for analysis. Each row already precalculated its percentage scoring for each metric, saving the user time not to calculate them manually. Once the row is selected, as needed by the user for making the visualization, it will be appended to the selected row table. The user must then input the weightage according to each column's importance in creating the insight to obtain the final score for each metric. The snippet of the dashboard for this user action can be referred to in Fig 2. Repeat the whole process for each remaining metric. Bear in mind that each metric would have a different column to select as it has a different way on the measurement methods. All the metric scores will be displayed on the overall page, where the user must once again input the weight. This allows the user to customize the assessment by prioritizing specific metrics over others. Fig 3 illustrates the final score of data reliability to be referred to in making the visualization or analysis.

G. Empirical Evaluation Result Analysis

Elouataoui et al. [4] technique in the year 2022 is the basis for this experiment's existing method. The measurement of the method nearly encompassed all the new tool's finalized metrics. However, the measures for the integrity and accuracy criteria and the data weight scale varied slightly. The current approach combines integrity and accuracy under a single metric, whereas the new tool expands each with a separate measurement. The new tool sets the rank of preferences only to the needed selected data column. However, the existing ones have the scale on all columns. This experiment tested both approaches on the same dataset with the same data significance preferences in creating data analysis.

The Pearson correlation coefficient result of $r = 0.996$ shows a strong positive linear relationship between the reliability scores generated by the new tool and the existing method. This can be categorized under strong correlation, which means the newly proposed technique effectively assesses reliability, just as the existing technique does. They generate a score that closely matches the existing technique, proving their high accuracy and correlation. Moreover, in terms of efficiency, the experiment conducted was able to reject the null hypothesis as the calculated t-value is less than the critical value. This supports the conclusion that there is a statistically significant difference in the execution time between tools and existing manual techniques.

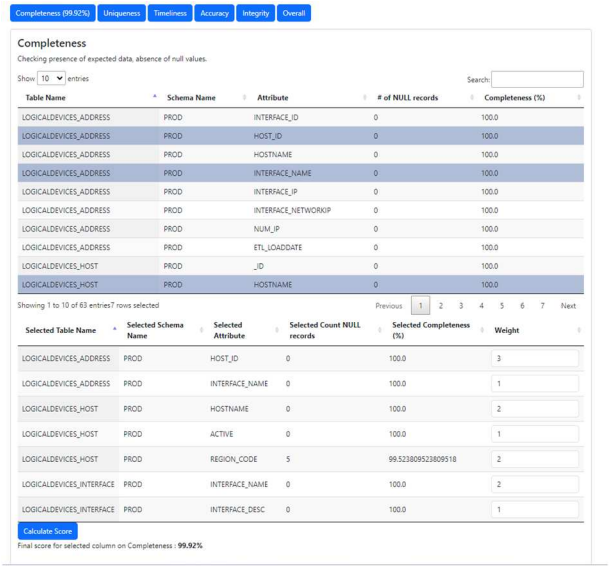

Fig 2 Table and column selection, weight inclusion, and final score for the dedicated metric

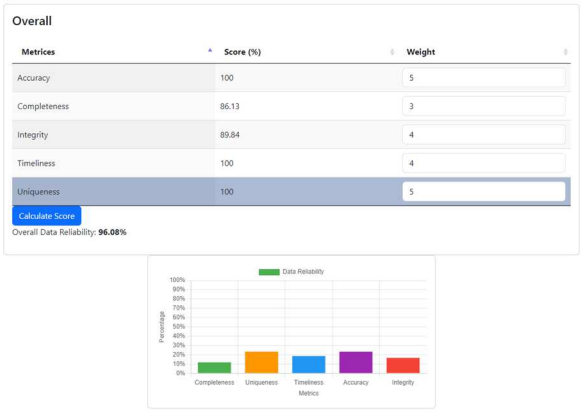

Fig 3 Overall score of all metrics measured from Data Reliability Metric Assessment Technique

H. Discussion

The comparison of the proposed and existing techniques in the experiment states their correlation so that the new procedure might resolve the problem in data quality assessment within the ETL (Extract, Transform, Load) process. More general than the current one, the new technique focuses on reliability and adds new metrics for accuracy and integrity. This change drastically imposes on data reliability scoring, considering only the data columns relevant to each analysis within the ETL process. Concerning effectiveness, the identified experiment reveals that the new technique offers results in less time than the previous compilation of Excel spreadsheets, decreasing the element of human fallibility inherent in other manual approaches. The proposed method allows automating such part of ETL, which, in turn, would improve the data handling quality within the process, making the analysis more correct, consistent, and efficient.

I. Threat to Validity

For internal validity, we avoid measurement bias by being very cautious and ensuring the new and the old are working on the same data set and reporting the same measures simultaneously. All the test cases are designed in advance, and we take turns using the new tool and employing the previous technique to eliminate any influence.

Another concern about external validity is whether the measurement of the reliability of data captured in those scenarios used is realistic. Relying on case studies effectively clarified the problem and determined whether the conditions and scenarios required for testing met real-world environments. The goal is to obtain an understanding that must include practical and realistic results that will reflect actual requirements and test cases.

IV. Conclusion

The data reliability assessment metrics technique assesses the reliability of measurements through data completeness, uniqueness, time, accuracy, and data integrity while employing the Likert scale to account for data weight. It captures data through ETL processes for data acquisition, user-friendly input through a web interface, and the final reliability score. The technique has been used successfully to evaluate data quality, thus allowing data practitioners to operate with reliable data and enhance ETL steps if needed. This, in turn, results in improved data analysis and more accurate decision-making information being generated.

However, the technique may be a little cumbersome, particularly for individuals who may not be familiar with data analysis and datasets themselves. This means that users require a clear understanding of the whole data structure and definition and the relationship between them. However, the same issue persists even with the existing method. Data analysts need to be proficient in the data they are about to work with.

Integrating new data quality measures to extend the tool's usage to various domains and increase sample sizes in the future is possible. In this big data age, advanced tools of cloud computing, artificial intelligence, and machine learning are also being integrated into different fields of industrial applications to improve large-scale big data facilities and intelligence [13]. Utilizing machine learning concepts and elements of artificial intelligence could further identify potential problems with data quality in advance, explain data dependencies to new users, and lessen the dependency on seasoned data analysts. Another possible change is to expand the array of information shown on the dashboard according to users' requests.

References

[1] L. Cai and Y. Zhu, "The Challenges of Data Quality and Data Quality Assessment in the Big Data Era," Data Science Journal, vol. 14, no. 0, p. 2, May 2015, doi: 10.5334/dsj-2015-002.

[2] S. Loetpipatwanich and P. Vichitthamaros, "Sakdas: A Python Package for Data Profiling and Data Quality Auditing," 2020 1st International Conference on Big Data Analytics and Practices (IBDAP), pp. 1–4, Sep. 2020, doi:10.1109/ibdap50342.2020.9245455.

[3] I. El Alaoui, Y. Gahi, and R. Messoussi, "Big Data Quality Metrics for Sentiment Analysis Approaches," Proceedings of the 2019 International Conference on Big Data Engineering, Jun. 2019, doi:10.1145/3341620.3341629.

[4] W. Elouataoui, I. El Alaoui, S. El Mendili, and Y. Gahi, "An Advanced Big Data Quality Framework Based on Weighted Metrics," Big Data and Cognitive Computing, vol. 6, no. 4, p. 153, Dec. 2022, doi:10.3390/bdcc6040153.

[5] V. Azzolini et al., "The Data Quality Monitoring Software for the CMS experiment at the LHC: past, present and future," EPJ Web of

Conferences, vol. 214, p. 02003, 2019, doi:10.1051/epjconf/201921402003.
[6] S. F. Kristyanti, T. F. Kusumasari, and E. N. Alam, “Operational Dashboard Development as A Data Quality Monitoring Tools Using Data Deduplication Profiling Result,” 2020 6th International Conference on Science and Technology (ICST), pp. 1–6, Sep. 2020, doi: 10.1109/icst50505.2020.9732870.
[7] E. Widad, E. Saida, and Y. Gahi, “Quality Anomaly Detection Using Predictive Techniques: An Extensive Big Data Quality Framework for Reliable Data Analysis,” IEEE Access, vol. 11, pp. 103306–103318, 2023, doi: 10.1109/access.2023.3317354.
[8] N. West, J. Gries, C. Brockmeier, J. C. Gobel, and J. Deuse, “Towards integrated Data Analysis Quality: Criteria for the application of Industrial Data Science,” 2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI), pp. 131–138, Aug. 2021, doi: 10.1109/iri51335.2021.00024.
[9] A. Kohli and N. Gupta, “Big Data Analytics: An Overview,” 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1–5, Sep. 2021, doi: 10.1109/icrito51393.2021.9596417.
[10] Munawar, “Extract Transform Loading (ETL) Based Data Quality for Data Warehouse Development,” 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), pp. 373–378, Oct. 2021, doi: 10.1109/iccsai53272.2021.9609770.
[11] B. Singhal and A. Aggarwal, “ETL, ELT and Reverse ETL: A business case Study,” 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), pp. 1–4, Dec. 2022, doi:10.1109/icatiece56365.2022.10046997.
[12] A. P. Pereira, B. P. Cardoso, and R. M. S. Laureano, “Business intelligence: Performance and sustainability measures in an ETL process,” 2018 13th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–7, Jun. 2018, doi:10.23919/cisti.2018.8399473.
[13] W. Han and M. Jochum, “A Machine Learning Approach for Data Quality Control of Earth Observation Data Management System,” IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, pp. 3101–3103, Sep. 2020, doi:10.1109/igarss39084.2020.9323615.
[14] A. Qaiser, M. U. Farooq, S. M. Nabeel Mustafa, and N. Abrar, “Comparative Analysis of ETL Tools in Big Data Analytics,” Pakistan Journal of Engineering and Technology, vol. 6, no. 1, pp. 7–12, Jan. 2023, doi: 10.51846/vol6iss1pp7-12.
[15] R. Ji, H. Hou, G. Sheng, and X. Jiang, “Data Quality Assessment for Electrical Equipment Condition Monitoring,” 2022 9th International Conference on Condition Monitoring and Diagnosis (CMD), pp. 1–4, Nov. 2022, doi: 10.23919/cmd54214.2022.9991385.
[16] M. Al Amin, MD. Jawad-Al-Mursalin Hoque, Z. Nazzum, M. A. Sayed, S. Tanveer Ahmed Rumee, and M. I. Zaber, “Data Quality Assessment of Substation Data in Bangladesh: Insights from Handwritten Data Digitization,” 2023 10th IEEE International Conference on Power Systems (ICPS), pp. 1–6, Dec. 2023, doi:10.1109/icps60393.2023.10428984.
[17] V. Pattana-Anake, F. J. J. Joseph, and P. Pachaivannan, “Data Wrangling for IoT Based Aquarium Water Quality Management System,” 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI), pp. 1–5, Dec. 2022, doi:10.1109/icdsaai55433.2022.10028891.
[18] X. Zuo, “Research on Data Quality Improvement Program Based on Big Data Application,” 2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), pp. 1742–1745, May 2023, doi: 10.1109/iciba56860.2023.10165495.
[19] L. Davidson, "What is data quality and why does it matter?" Springboard, 2019. [Online]. Available: https://www.springboard.com/blog/data-analytics/data-quality/.
[20] P. Zhang, F. Xiong, J. Gao, and J. Wang, “Data quality in big data processing: Issues, solutions and open problems,” 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 1–7, Aug. 2017, doi: 10.1109/uic-atc.2017.8397554.
[21] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi, “From Data Quality to Big Data Quality,” Journal of Database Management, vol. 26, no. 1, pp. 60–82, Jan. 2015, doi: 10.4018/jdm.2015010103.
[22] H. A. Sulistyo, T. F. Kusumasari, and E. N. Alam, “Implementation of Data Cleansing Null Method for Data Quality Management Dashboard using Pentaho Data Integration,” 2020 3rd International Conference on Information and Communications Technology (ICOIACT), pp. 12–16, Nov. 2020, doi: 10.1109/icoiact50329.2020.9332030.
[23] H. Homayouni, “Testing Extract-Transform-Load Process in Data Warehouse Systems,” 2018 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW), pp. 158–161, Oct. 2018, doi: 10.1109/issrew.2018.000-6.
[24] R. Vaidyambath, J. Debattista, N. Srivatsa, and R. Brennan, "An intelligent linked data quality dashboard," in AICS 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, Galway, Ireland, 2019, pp. 5-6.
[25] T. Samakit, C. Anutariya, and M. Buranarach, “QUALYST: Data Quality Assessment System for Thailand Open Government Data,” 2023 20th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 196–201, Jun. 2023, doi:10.1109/jcsse58229.2023.10202060.
[26] Talend, "Data quality Looker Block released for Talend Studio," Talend Community, Nov. 6, 2020. [Online]. Available: https://community.talend.com/s/article/Data-quality-Looker-Block-released-for-Talend-Studio?language=en_US.
[27] Talend, "What is data health? Definition and how to measure," Talend, [Online]. Available: https://www.talend.com/resources/what-is-data-health/.
[28] Ataccama, "Data quality management," Ataccama, [Online]. Available: https://www.ataccama.com/dictionary/data-quality-management.
[29] J. Byabazaire, G. M. P. O’Hare, and D. T. Delaney, “End-to-End Data Quality Assessment Using Trust for Data Shared IoT Deployments,” IEEE Sensors Journal, vol. 22, no. 20, pp. 19995–20009, Oct. 2022, doi: 10.1109/jsen.2022.3203853.
[30] S. McCarthy, A. McCarren, and M. Roantree, “A Method for Automated Transformation and Validation of Online Datasets,” 2019 IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC), pp. 183–189, Oct. 2019, doi:10.1109/edoc.2019.00030.
[31] R. Likert, "A technique for the measurement of attitudes," Archives of Psychology, vol. 22, no. 140, pp. 55, 1932.