A Fuzzy Logic-Based Data Integrity Assessment for Data Warehouses

Millina Achary[1] Pranali Mahadik[2]
[1,2]Department of Information Technology
[1,2]Keraleeya Samajam's Model College, Dombivli East, Mumbai, Maharashtra, India

Abstract — Data warehouses integrate data from multiple sources to support analytical and decision-making processes, making data integrity an important factor in ensuring the reliability of analytical results. Conventional data integrity assessment methods often rely on predefined thresholds or binary classifications, such as valid or invalid, which may not adequately represent the gradual and uncertain nature of data quality. This study proposes a Fuzzy Logic-Based Data Integrity Assessment for Data Warehouses that evaluates data integrity using multiple quality indicators, including accuracy, completeness, consistency, timeliness, and validity. Fuzzy logic is used to transform these indicators into linguistic variables such as Low, Medium, and High and combine them to generate an overall data integrity score ranging from 0 to 1. The proposed approach can be applied during or after the ETL/ELT process to identify potentially unreliable datasets and prioritize them for further investigation or corrective action. By representing data quality as degrees rather than binary classification, the approach aims to provide a more flexible, granular, and interpretable assessment of data integrity in data warehouse environments.

Keywords: Fuzzy Logic, Data warehouse, Data Integrity, Data Quality, Fuzzy Membership, ETL, ELT, Uncertainty, Data Analytics

I. Introduction

Organizations increasingly use data warehouses to integrate information from multiple operational and external sources and support reporting, analytics, and decision-making. The usefulness of a data warehouse depends not only on the volume of data it contains but also on the quality and integrity of that data. Data may contain missing values, inaccurate values, inconsistent records, outdated information, or values that do not conform to required rules. Such problems can affect analytical results and lead to unreliable decisions. Traditional data-quality assessment commonly uses predefined rules and thresholds. These methods are useful, but they can force data into rigid categories such as acceptable/unacceptable or valid/invalid. In practice, data quality may exist at different levels. Fuzzy logic provides a mathematical framework for representing such gradual conditions using membership values between 0 and 1 linguistic categories such as Low, Medium, and High. This research proposes a fuzzy logic-based approach for assessing data integrity in data warehouse environments. The approach considers multiple quality indicators- accuracy, completeness, consistency, timeliness, and validity and combines them through fuzzy rules to produce an overall integrity score. The study also uses a public survey and statistical analysis to examine perceptions related to conventional and fuzzy based data quality assessment.

II. Literature Review

Data warehouses (DWs) support organizational decision-making by consolidating data from heterogeneous sources, but the reliability of this consolidated data depends heavily on its quality and integrity. Traditional data quality checks rely on crisp, binary rules (a record either passes or fails a validation rule), which struggle to capture the vagueness, imprecision, and gradation inherent in real-world data. Fuzzy logic, since it allows partial membership and linguistic reasoning ("mostly accurate", "moderately consistent"), has increasingly been proposed as a more natural framework for assessing data quality. Cichy and Rass (2019) demonstrate this directly, using fuzzy approximation and if-then rules to build an explainable data quality indicator, arguing that fuzzy models better reflect the subjective, expert-driven judgments involved in quality assessment than rigid thresholds do. Similar reasoning underlies fuzzy quality models proposed for linked open data, where fuzzy ontologies are used to represent and reason about source quality across multiple, often conflicting, quality dimensions [1]. A second stream addresses fuzzy data warehouse architecture itself-embedding fuzziness into the DW structure rather than using it only as an external evaluation tool. Fasel and Zumstein's series of papers (on wed analytics, customer performance measurement, and meta-table structures) introduce fuzzy dimensions and fuzzy facts, allowing values to belong to multiple linguistic classes simultaneously (e.g., a sales figure being partly "high" and partly "medium"). They show that such fuzzy classification produces more accurate and human-interpretable analysis than sharp classification, and that both qualitative and quantitative attributes can be handled within the same model [2]. A third, more directly relevant stream applies fuzzy logic during the ETL (Extract-Transform-Load) process. Where data integrity is actively enforced before data enters the warehouses. The fuzzy-based ETL filter approach defines linguistic variables and a fuzzy rule base to interpret and filter incoming data for business intelligence systems, offering a more nuanced alternative to conventional crisp ETL data-cleansing rules (which typically only flag data as strictly valid/invalid). This contrasts with non-fuzzy ETL quality frameworks- such as automated ETL testing approaches and ETL- based cleaning frameworks- which rely on fixed thresholds for dimensions like completeness, consistency, and accuracy. The comparison suggests fuzzy ETL filtering cloud meaningfully improve integrity assessment over these crisp baselines, though the existing fuzzy ETL literature remains sparse and largely conceptual [3]. A fourth stream uses fuzzy logic operationally, for monitoring or querying warehouse data after it has been loaded. The fuzzy expert system for server-failure data warehouse analysis, for example, applies fuzzy rule-based inference so decision-makers can query the warehouse using natural, vague terms rather than precise thresholds. This shows fuzzy logic being used not just to assess quality abstractly but to support real decision-making directly on warehouse data- closer in spirit to an "integrity assessment" that a user or system might actually rely on [4].

III. DATA INTEGRITY

Data integrity refers to maintaining the correctness, completeness, and trustworthiness of data throughout its lifecycle. In the information security context, the National Institute of Standards and Technology (NIST) defines data integrity as the property that data has not been changed or altered in an unauthorized manner after it has been created, transmitted, or stored [5]. Therefore, data integrity is important because decisions made from a data warehouse depend on the reliability of the stored data. In a data warehouse, data is collected from different operational systems and transformed before being stored for analysis. During this process, data may be affected by missing values, incorrect values, inconsistent formats, duplicate records, or outdated information. Such problems can reduce the reliability of analytical results. Therefore, data integrity assessment should consider different characteristics of data quality rather than checking only whether data exists in the database.

A. Importance of Data Integrity in Data Warehouses

Data warehouses integrate information from multiple sources and maintain historical data for reporting and decision-making. Because data passes through several stages such as extraction, transformation, cleaning, and loading, maintaining data integrity is essential throughout the data warehouse process [6]. Poor data integrity can lead to incorrect reports, unreliable analysis, and inappropriate business decisions. For example, if customer information is incomplete or sales values are incorrect during the ETL process, the final analytical results may also be inaccurate. In this research, data integrity is assessed using five important data quality indicators: accuracy, completeness, consistency, timeliness, and validity. These dimensions are widely discussed in data quality research and provide different views of the reliability and usefulness of data [7], [8].

B. Data Integrity and Data Quality

Data integrity and data quality are closely related, but they are not exactly the same concept. Data integrity mainly focuses on whether data has been protected from improper or unauthorized modification. Data quality is broader and evaluates whether data is suitable, correct, complete, consistent, timely, and valid for its intended purpose [5], [7]. For example, a customer's record may remain unchanged after being stored in a database, so it can maintain integrity from a security perspective. However, if the customer's address is outdated, the data may still have poor quality for a particular application. Therefore, this study uses selected data quality dimensions as measurable indicators for assessing the overall integrity of data within a data warehouse. This approach allows the study to represent data condition using different degrees rather than treating the data only as completely valid or completely invalid.

C. Data Integrity and Data Security

Data integrity is one of the major objectives of information security. NIST explains integrity as protection against improper modification or destruction of information, along with related properties such as authenticity and non-repudiation [5]. However, data integrity should not be considered the same as overall data security. Data security includes broader protection mechanisms such as confidentiality, integrity, and availability. Data integrity specifically focuses on maintaining data in a correct and protected state. For this research, the focus is not on security mechanisms such as encryption or access control. Instead, the study focuses on measuring the condition of data using quality indicators such as accuracy, completeness, consistency, timeliness, and validity.

D. Data Integrity Dimensions Used in This Study

The five dimensions selected for this study are described below:

Dimension	Meaning
Accuracy	Measures whether the data correctly represents the real-world object, event, or value that it describes.
Completeness	Measures whether all required or relevant data is available without important missing values.
Consistency	Measures whether data remains compatible and does not contain conflicting values across records or systems.
Timeliness	Measures whether the data is sufficiently up-to-date and available at the required time.
Validity	Measures whether data follows the required format, type, range, or predefined rules

These dimensions have been selected because they are commonly used in data quality research and are relevant to evaluating the condition of data for analytical purposes [7], [8], [9].

E. Role of Data Integrity in the proposed Study

The proposed study develops a fuzzy logic-based data integrity assessment for data warehouses. Instead of using only a binary decision such as valid or invalid, the proposed approach represents the condition of data using degrees. The five input dimensions- accuracy, completeness, consistency, timeliness, and validity are measured and provided to the fuzzy logic system. The system then combines these inputs and produces an overall data integrity score between 0 and 1. This approach is intended to represent gradual differences in data condition. For example, two datasets may both contain some quality problems, but one may have considerably better quality than the other. A degree-based assessment can represent this difference more clearly than a simple binary classification.

IV. DATA WAREHOUSE

A data warehouse is a centralized data environment designed to integrate information from different sources and support analysis and decision-making. Data warehouses generally organize historical and integrated data so that users can perform analytical queries and generate reports. A typical data warehouse process involves collecting data from multiple source systems, transforming and cleaning the data, and loading the processed data into the warehouse. Research on data warehouse ETL processes describes the general flow as data extraction from different sources, transformation and cleansing in a staging area, and finally loading the processed data into the data warehouse [6]. The data warehouse is therefore an important part of this research because the proposed data integrity assessment is designed to evaluate the quality of data that is stored and processed within such an analytical environment.

A. *Characteristics of a Data Warehouse*

A data warehouse commonly has the following characteristics:

1) Subject-Oriented: Data is organized around important subjects such as customers, products, sales, or transactions rather than individual operational processes.
2) Integrated: Data from different sources is combined into a common structure and format.
3) Time-Variant: A data warehouse generally maintains historical information so that data can be analyzed over different periods.
4) Non-Volatile: Data Stored in a warehouse is mainly used for analysis and reporting. Once loaded, it is generally not changed as frequently as data in operational systems.

These characteristics distinguish data warehouses from conventional operational databases and make them suitable for analytical processing [6].

B. *Data Warehouse Architecture*

A basic data warehouse architecture consists of several components that work together to collect, process, store, and analyze data.

1) The major components are:

- Data Source: Operational databases, applications, files, and other systems that provide data.
- ETL/ELT: Process used to extract, transform, clean, integrity, and load data.
- Staging Area: Temporary environment where extracted data can be stored and processed before loading.
- Data Warehouse: Central repository containing integrated and historical data.
- OLAP/Reporting: Analytical tools and applications used to query and analyze warehouse data.

This architecture is consistent with research describing data warehouse systems as having source systems, a data staging area, and the main data warehouse, with ETL processes connecting these components [6].

C. *ETL and ELT*

ETL stands for Extract, Transform, and Load. It is a process used to move data from source systems into a data warehouse. The three main steps are:

1) Extract: Data is collected from one or more source systems.
2) Transform: The extracted data is cleaned, converted, standardized, and integrated into a suitable format.
3) Load: The transformed data is loaded into the target data warehouse.

During transformation, data quality problems such as missing values, inconsistent formats, and incorrect structures can be identified and handled [2]. In an ELT approach, data is first extracted and loaded into the target environment, and transformation is preformed after loading. Both ETL and ELT can be used for integrating data into modern analytical environments.

D. *Staging Area*

A staging area is an intermediate storage environment used during the data integration process. Extracted data can be temporarily stored in the staging area before transforming and loading into the data warehouse. The staging area allows data to be cleaned, transformed, standardized, and checked before it becomes part of the warehouse. Academic research on ETL processes describes the staging area as a physical container for temporary data produced during extraction and transformation [6]. For the proposed research, this stage is important because data quality checks can be performed before the data is stored in the final warehouse.

E. *Fact Table*

A fact table is a central table in a dimensional data warehouse that stores measurable or quantitative information. It generally contains numerical measures along with keys that connect the facts to related dimension tables. For the proposed research, a fact table can be designed to store the quality measurements and the calculated overall integrity score. A possible structure is:

Field	Description
Dataset_ID	Unique Identifier of the dataset.
Date_ID	Reference to the data dimension.
Source_ID	Reference to the source system.
Accuracy_Score	Measured accuracy value.
Completeness_Score	Measured completeness value.
Consistency_score	Measured consistency values.
Timeliness_score	Measured timeliness value.
Validity_score	Measured validity value.
Overall_Integrity_Score	Final score generated by the proposed fuzzy model.

The exact fields and scheme should be finalized according to the dataset and implementation used in the study.

F. *Dimension Table*

A dimension table contains descriptive information that provides context for the measurements stored in fact tables. Common dimensions include data, source, customer, product, location, and other descriptive categories. For this research, possible dimensions include:

- Data Dimension: Data, month, quarter, and year.
- Source Dimension: Source system or data origin.
- Dataset Dimension: Dataset name, type or category.

Fact and dimension tables are commonly used in dimensional data warehouse designs such as the star schema [10].

G. OLAP

OLAP (Online Analytical Processing) is used to analyze data stored in a data warehouse; it allows users to examine information from different perspectives and perform analytical operations on historical data. For example, an organization may analyze data integrity scores by- data sources, Month or year, Dataset type, individual quality dimension, or overall integrity level.

OLAP systems commonly use dimensional models containing fact and dimension tables to support analytical queries and reporting [10]

H. Role of the Data Warehouse in the Proposed Study

The data warehouse provides an environment in which data from different sources can be integrated, stored, and analyzed. In the proposed study, data quality measurements are generated for the five selected dimensions: accuracy, completeness, consistency, timeliness, and validity. These measurements are provided as inputs to the fuzzy logic model. The model processes the input values and produces an overall integrity score. The score can then be stored along with the individual quality measurement for further analysis and reporting. Therefore, the data warehouse serves as the central analytical environment for the proposed fuzzy logic-based data integrity over time.

V. DATA QUALITY DIMENSIONS

Data quality refers to the extent to which data is suitable, reliable, and useful for its intended purpose. Data quality is a multi-dimensional concept because the quality of a dataset cannot normally be described by only one characteristic. Different studies have identified several dimensions of data quality, including accuracy, completeness, consistency, timeliness, and validity. These dimensions are also commonly used in modern data-quality frameworks. [7], [11]. In this research, five dimensions are selected to assess the quality and integrity of data stored in a data warehouse: accuracy, completeness, consistency, timeliness, and validity. These dimensions provide different views of the condition of a dataset and can be combined to obtain an overall assessment.

A. Accuracy

Accuracy refers to the degree to which data correctly represents the real-world value or situation that it is intended to describe. Accurate data should contain correct values and should not contain unnecessary errors. For example, if a customer's actual age is 25 but the database stores the age as 52, the data is inaccurate. In a data warehouse, inaccurate data can lead to incorrect analysis and unreliable decisions.

B. Completeness

Completeness refers to the degree to which all required or expected data is present. Missing values, incomplete records, or missing attributes can reduce the completeness of a dataset. For example, if a customer record requires name, address, email, and phone number but the phone number is missing, the record is incomplete. Completeness is important because missing information may affect subsequent analysis and reporting.

C. Consistency

Consistency refers to the degree to which data agrees with itself across records, systems, or sources. Data should not contain conflicting values or representations for the same entity. For example, if a customer's address is recorded as "Mumbai" in one system and "Pune" in another system for the same customer without a valid reason, the data is inconsistent.

D. Timeliness

Timeliness refers to the degree to which data is sufficiently current and available when it is required. Data that is correct but outdated may not be useful for time-sensitive decisions. For example, an inventory dataset showing stock information from several months ago may not be suitable for making a current inventory decision. Therefore, the age of data and the time at time at which it becomes available are important aspects of timeliness.

E. Validity

Validity refers to the degree to which data conforms to predefined rules, formats, types, ranges, or other requirements. Valid data follows the expected structure and constraints. For example, if an age field is defined to accept values between 0 and 120, an age value of 250 would be considered invalid. Similarly, an incorrectly formatted email address may fail a validity check.

F. Importance of These Dimensions in the Proposed Study

The five dimensions provide complementary information about data quality. A dataset may be accurate but incomplete, complete but inconsistent, or valid but outdated. Therefore, considering multiple dimensions gives a more detailed representation of the condition of warehouse data. The proposed study uses these five dimensions as input indicators for the fuzzy logic-based assessment. Each dimension can be represented using a numerical quality value and then converted into fuzzy linguistic categories such as low, medium, and high. The fuzzy system can subsequently combine these inputs to generate an overall data integrity score. Thus, the selected dimensions provide the foundation for the proposed fuzzy based data integrity assessment model.

VI. RESEARCH GAP

Although existing studies demonstrate the usefulness of fuzzy logic in data quality assessment, data warehouse schema evaluation, fuzzy data warehouse modelling, ETL filtering, and decision-support applications, these approaches address different aspects of the data warehouse lifecycle. Existing data warehouse studies largely focus on schema-level quality, analytical modelling, or data cleansing, while fuzzy data-quality studies often operate outside the specific context of data warehouses. Moreover, existing approaches generally evaluate individual quality dimensions or perform specific validation and filtering tasks rather than integrating multiple dimensions into a unified measure of overall data integrity. Therefore, a research gap exists in developing a unified fuzzy logic-based framework that evaluates multiple data-quality dimensions of data warehouse data, such as accuracy, completeness, consistency, timeliness, and validity, and combines their fuzzy assessments into an overall integrity score ranging from 0 to 1. Such a framework could provide a more gradual and interpretable assessment of interpretable assessment of data integrity than conventional binary or fixed-threshold validation approaches.

VII. PROBLEM STATEMENT

Data warehouses integrate data from multiple heterogeneous sources and are widely used to support organizational analysis and decision-making. However, the reliability of decisions made from a data warehouse depends on the integrity and quality of the data stored within it. Existing data-quality and integrity assessment approaches commonly rely on predefined rules, fixed thresholds, or binary classifications in which data is considered valid or invalid. Such approaches may not adequately represent the gradual and uncertain nature of real-world data quality, where a dataset may be partially accurate, moderately complete, somewhat consistent, or relatively outdated. Furthermore, data warehouse integrity is influenced by multiple quality dimensions, including accuracy, completeness, consistency, timeliness, and validity. Assessing these dimensions independently does not provide a unified view of the overall integrity of the warehouse data. Therefore, there is a need for an approach that can accommodate degrees of quality and combine multiple quality dimensions into a single, interpretable assessment. This study addresses this problem by proposing a fuzzy logic-based framework for post-load data integrity assessment in data warehouses. The proposed framework measures selected data-quality dimensions, converts their numerical values into fuzzy linguistic values, applies fuzzy inference rules, and produces an overall data integrity score ranging from 0 to 1. The resulting score can be interpreted as low, moderate, or high integrity, thereby providing a more flexible and interpretable assessment than conventional crisp validation approaches.

VIII. PUBLIC SURVEY

A public survey was conducted as part of this research to understand people's perceptions regarding data quality and data integrity assessment. The main objective of the survey was to examine respondents' opinions about different data-quality factors and to understand whether a degree-based approach could provide more information than conventional binary data-quality assessment.

A total of 38 responses were collected through a structured questionnaire. The respondents were asked about their familiarity with computers, data, and technology, as well as the importance of different data-quality dimensions such as accuracy, completeness, consistency, timeliness, and validity.

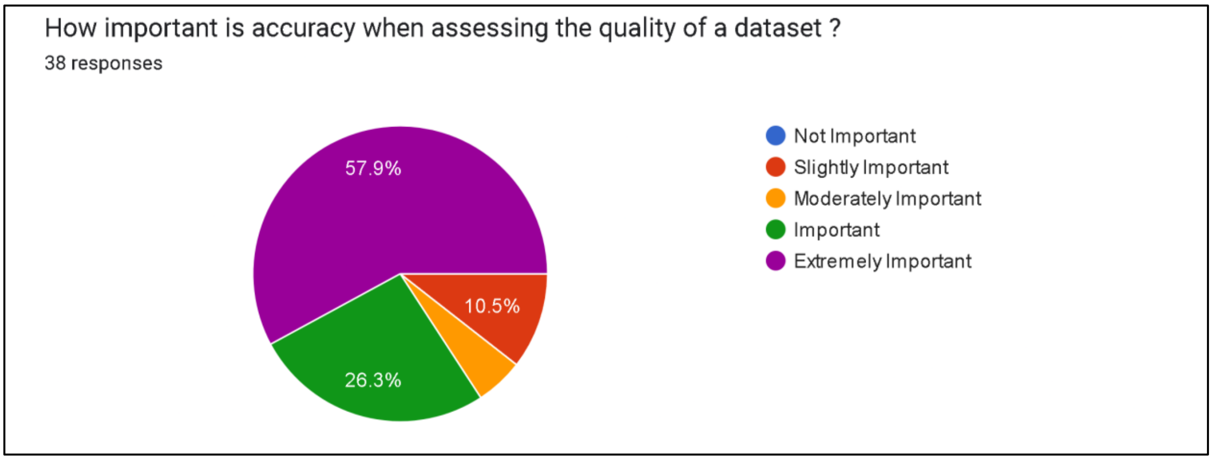
The survey also included questions about whether data quality can always be accurately represented using only "Valid" or "Invalid" classifications. Respondents were asked whether a dataset can have different degrees of quality rather than simply being classified as "Good" or "Bad." They were also asked which approach would provide more information about the quality of a dataset: the conventional Valid/Invalid approach or the proposed Low/Medium/High with a Numerical Score approach.

In addition, the questionnaire examined respondents' opinions about the usefulness of an overall data-integrity score between 0 and 1, combining multiple factors for data-quality assessment, and classifying datasets as Low, Medium, or High integrity. The survey also considered whether such a system could help organizations identify datasets that require correction or further investigation.

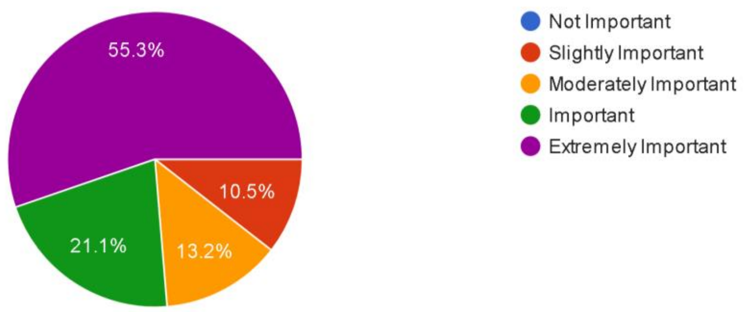
The collected responses were organized and analyzed to identify general trends in public perception. The survey results were subsequently used for hypothesis testing using the Pearson Chi-Square test. The hypothesis test focuses on the respondents' preference between the conventional Valid/Invalid approach and the proposed fuzzy logic-based approach.

The survey was conducted to provide supporting empirical evidence for the proposed fuzzy logic-based data integrity assessment framework, which represents data quality using degrees of membership and produces an overall integrity score rather than relying only on binary classifications

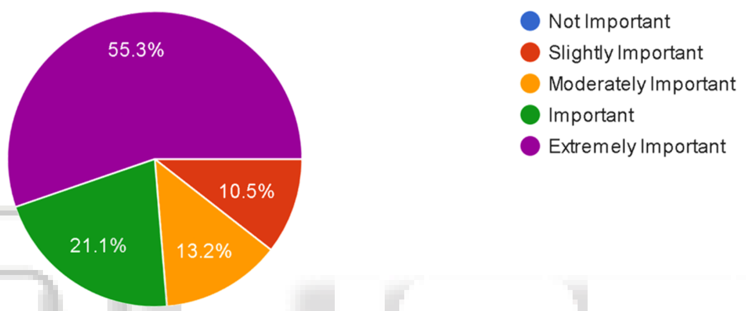
IX. QUESTIONNAIRE:



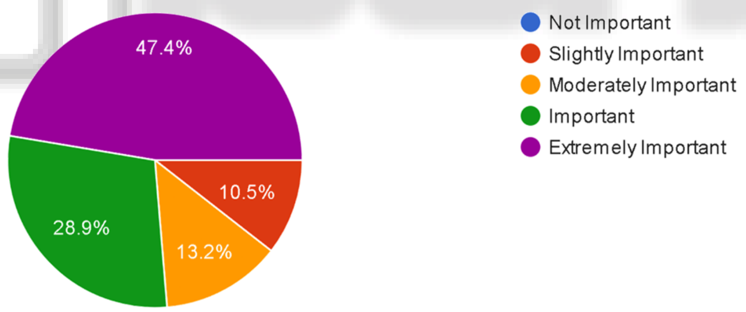
How important is completeness when assessing the quality of a dataset ?
38 responses



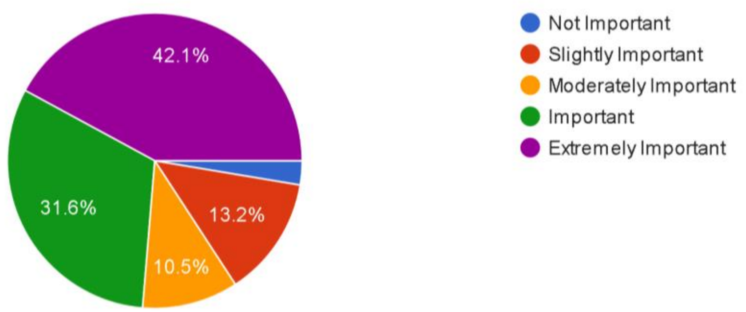
How important is completeness when assessing the quality of a dataset ?
38 responses

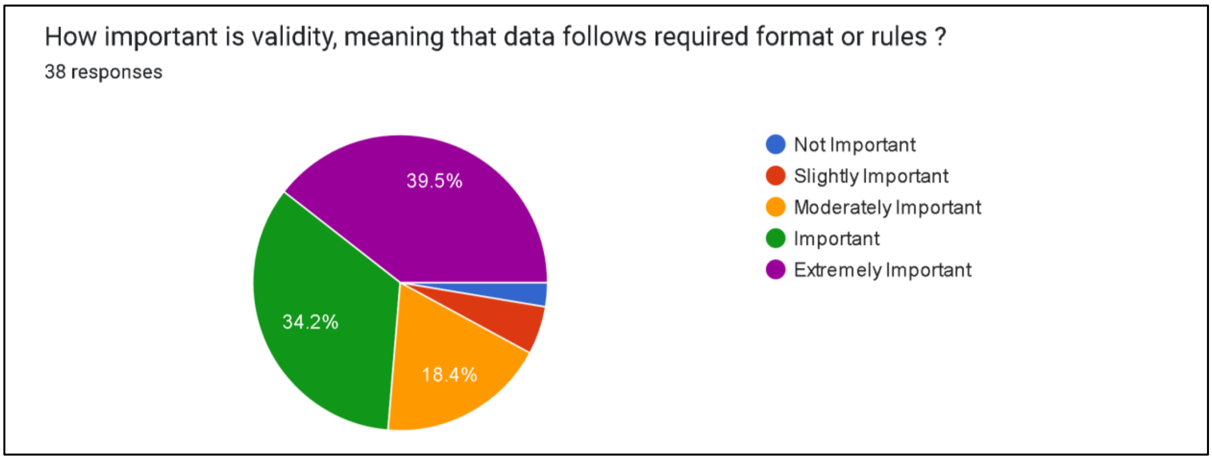


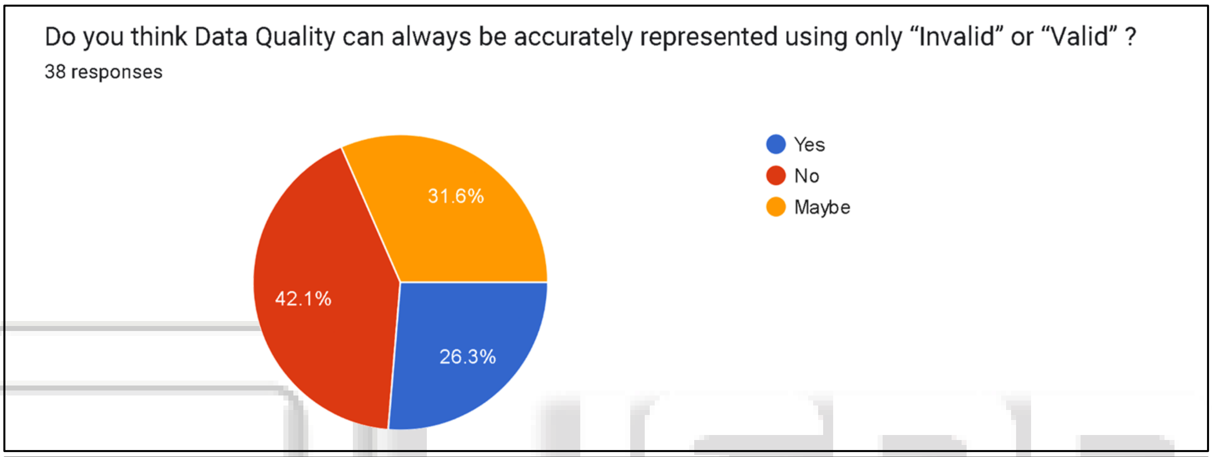
How important is consistency when assessing the quality of a dataset ?
38 responses

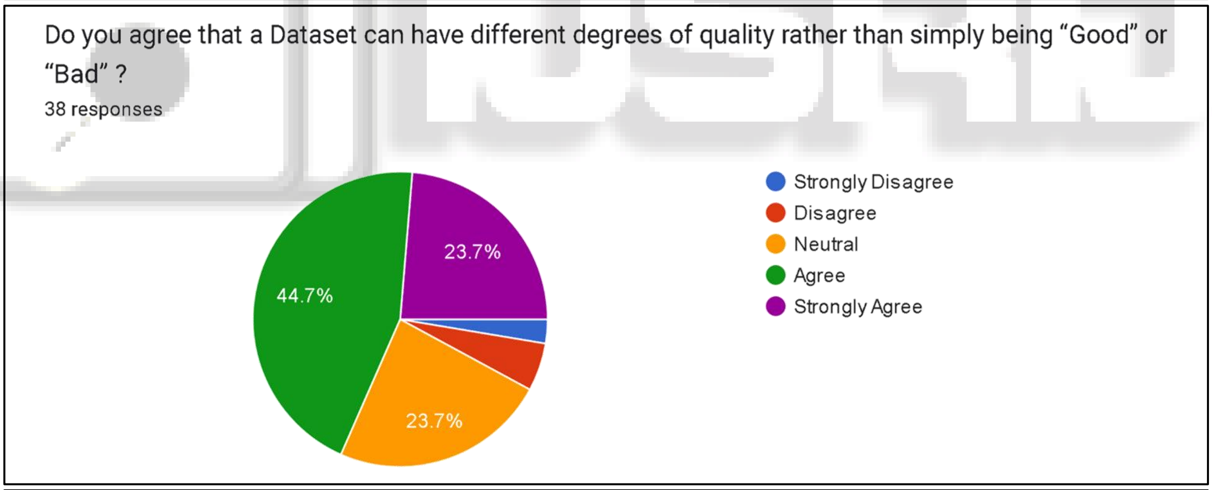


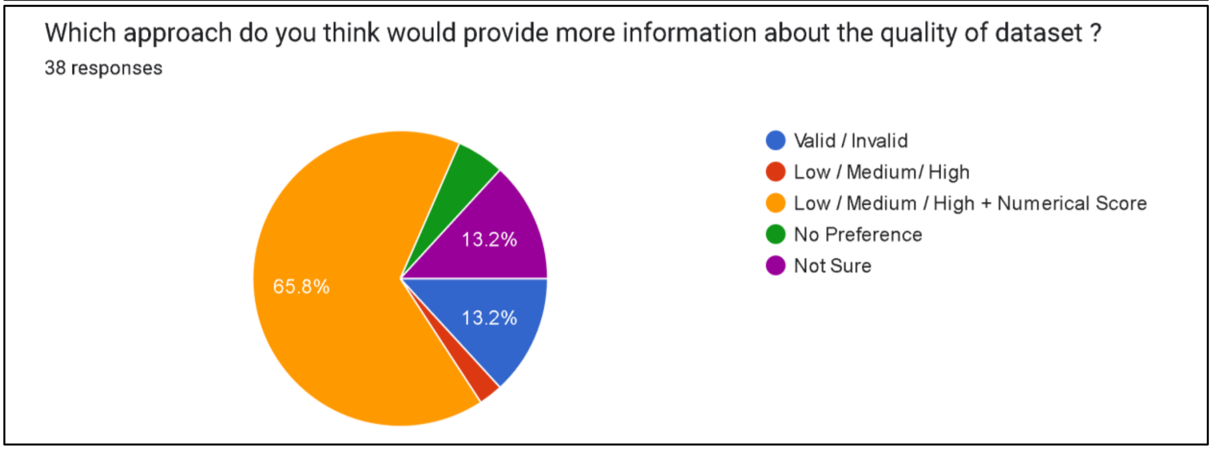
How important is timelines/up-to-date information when assessing the quality of a dataset ?
38 responses

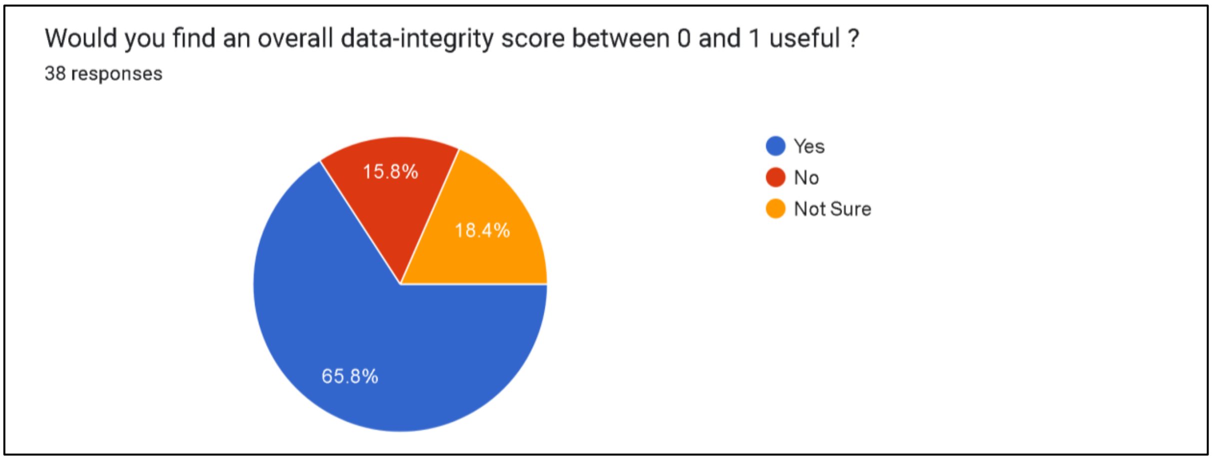


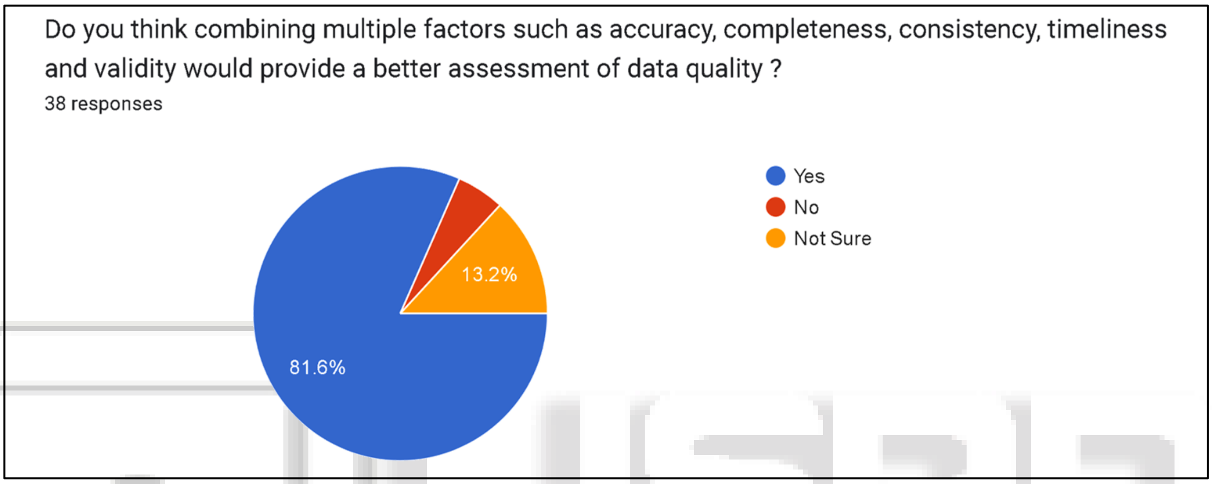

How important is validity, meaning that data follows required format or rules ?
38 responses
39.5%
34.2%
18.4%
Not Important
Slightly Important
Moderately Important
Important
Extremely Important

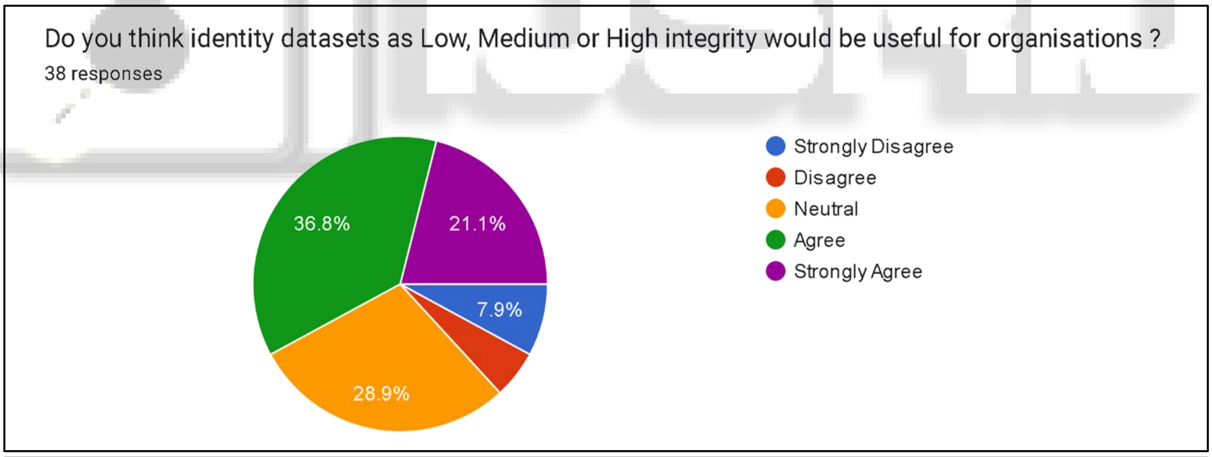

Do you think Data Quality can always be accurately represented using only “Invalid” or “Valid” ?
38 responses
31.6%
42.1%
26.3%
Yes
No
Maybe

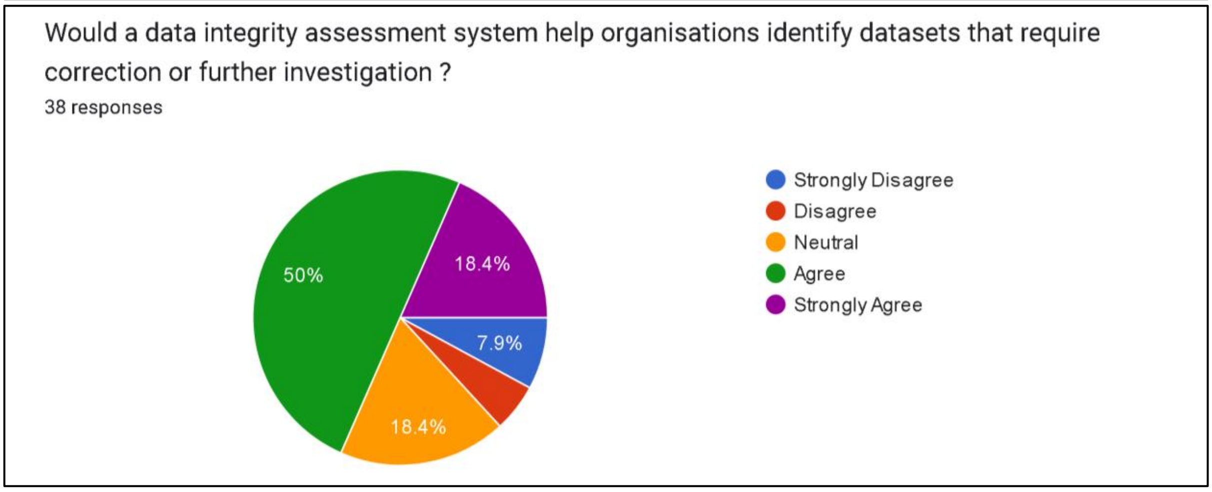

Do you agree that a Dataset can have different degrees of quality rather than simply being “Good” or “Bad” ?
38 responses
23.7%
44.7%
23.7%
Strongly Disagree
Disagree
Neutral
Agree
Strongly Agree

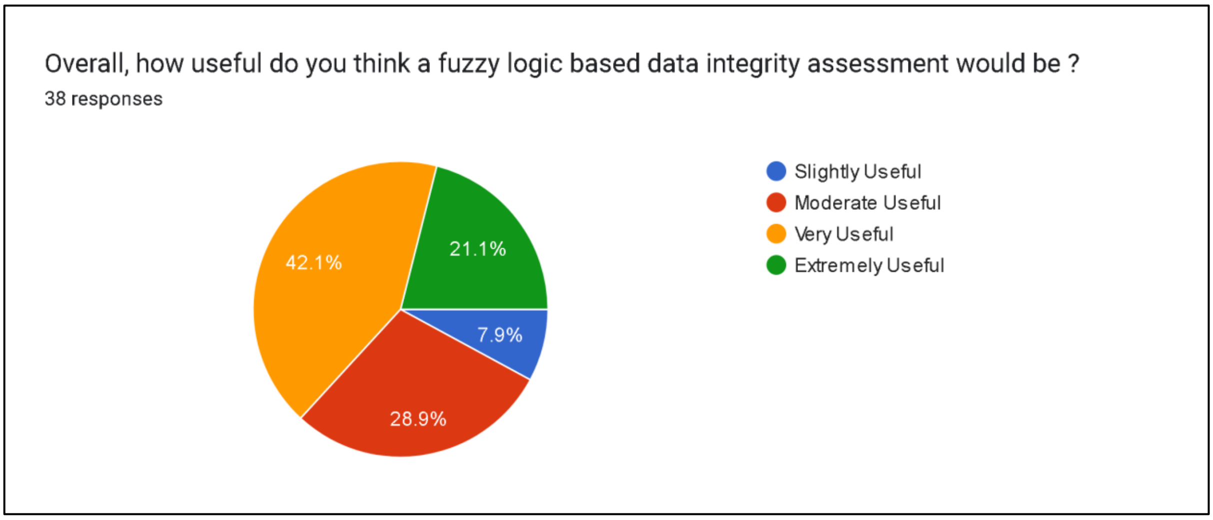

Which approach do you think would provide more information about the quality of dataset ?
38 responses
13.2%
65.8%
13.2%
Valid / Invalid
Low / Medium/ High
Low / Medium / High + Numerical Score
No Preference
Not Sure


Would you find an overall data-integrity score between 0 and 1 useful ?
38 responses
15.8%
18.4%
65.8%
Yes
No
Not Sure


Do you think combining multiple factors such as accuracy, completeness, consistency, timeliness and validity would provide a better assessment of data quality ?
38 responses
13.2%
81.6%
Yes
No
Not Sure


Do you think identity datasets as Low, Medium or High integrity would be useful for organisations ?
38 responses
36.8%
21.1%
7.9%
28.9%
Strongly Disagree
Disagree
Neutral
Agree
Strongly Agree


Would a data integrity assessment system help organisations identify datasets that require correction or further investigation ?
38 responses
50%
18.4%
7.9%
18.4%
Strongly Disagree
Disagree
Neutral
Agree
Strongly Agree



X. Hypothesis Testing

Hypothesis testing is a statistical method used to analyze sample data and draw conclusions about a population. It involves formulating two statements: the null hypothesis (H_0) and the alternative hypothesis (H_a). The null hypothesis represents the assumption that there is no significant difference or relationship between the variables being studied. The alternative hypothesis represents the presence of a significant difference or relationship.

The hypothesis-testing procedure uses sample data to determine whether the null hypothesis should be rejected or not. If the null hypothesis is rejected, the result provides statistical evidence in support of the alternative hypothesis.

For this paper, the hypotheses are formulated based on the survey responses regarding the conventional and fuzzy logic-based approaches to data-quality assessment.

- Null Hypothesis (H_0): There is no significant difference between respondents' preference for the conventional Valid/Invalid approach and the fuzzy logic-based Low/Medium/High + Numerical Score approach for assessing data quality.
- Alternative Hypothesis (H_a): There is a significant difference between respondents' preference for the conventional Valid/Invalid approach and the fuzzy logic-based Low/Medium/High + Numerical Score approach for assessing data quality.

XI. Test Statistics

There are several statistical tests available for hypothesis testing. Some commonly used tests are:

1) Chi-Square Test
2) Student's t-test (T-test)
3) Fisher's Z-test

For this paper, the Chi-Square Test is used because the survey responses being analyzed are categorical. The Pearson Chi-Square test is used to determine whether the observed frequencies of the responses differ significantly from the frequencies expected under the null hypothesis.

The level of significance used for this study is:

$$\alpha = 0.05$$

The Chi-Square statistic is calculated using:

χ2=∑(O−E)2E\chi^2 = \sum \frac{(O-E)^2}{E}

where O represents the observed frequency and E represents the expected frequency.

Approach	Observed Frequency	Expected Frequency
Fuzzy Approach	26	15.5
Valid/Invalid Approach	5	15.5
Total	31	31

The calculated value is $\chi^2 = 14.226$, with df = 1 and p = 0.00016.

Since $p < 0.05$, the null hypothesis is rejected.

A. *Conclusion of Hypothesis Testing*

The Chi-Square test indicates that there is a statistically significant difference between respondents' preferences for the conventional Valid/Invalid approach and the fuzzy logic-based approach. The survey therefore provides statistical support for investigating the use of fuzzy logic and degree-based classifications for data-integrity assessment.

XII. Overall Conclusion

This study proposed a fuzzy logic-based framework for assessing data integrity in data warehouses. Conventional assessment methods classify data as valid or invalid using predefined rules or fixed thresholds, which cannot represent the gradual and uncertain nature of real-world data quality. The proposed framework instead evaluates five data-quality indicators – accuracy, completeness, consistency, timeliness and validity – converts them into linguistic values (Low, Medium and High), and combines them through fuzzy inference rules into an overall data integrity score between 0 and 1. This gives a more granular and interpretable view of data condition, allows datasets of different quality levels to be distinguished more clearly, and helps identify datasets that need correction or further investigation.

The framework is designed to be applied during or after the ETL/ELT process, for example in the staging area or after loading. The individual dimension scores and the overall integrity score can be stored in a fact table linked to date, source and dataset dimensions, so that integrity can be analysed through OLAP by source, time period, dataset type and quality dimension. By addressing multiple dimensions within a single fuzzy model, the study addresses the gap identified in the literature, where fuzzy approaches have mostly been applied to individual dimensions or to isolated tasks such as schema evaluation, ETL filtering or fuzzy warehouse modelling.

The public survey supports this direction. Among the 31 respondents who expressed a preference, 26 preferred the degree-based fuzzy approach over the conventional Valid/Invalid approach, and the Chi-Square test ($\chi^2 = 14.226$, df = 1, p = 0.00016) showed this preference to be statistically significant at $\alpha = 0.05$. This indicates that a graded, score-based assessment is perceived as more informative than binary classification and justifies further investigation of fuzzy logic for data integrity assessment.

The survey, however, measures respondents' perception and not the performance of the model. The present work is therefore a proposed framework, and its effectiveness still needs to be validated. Future work should define the membership functions and fuzzy rule base in detail, implement the model on real data warehouse datasets, and compare its results with crisp threshold-based methods. Further extensions may include weighted or expert-defined rules, automated computation of the quality indicators, and integration of the integrity score into ETL/ELT monitoring and reporting.

XIII. References:

[1] Cichy, C., & Rass, S. (2019). A Fuzzy-Approximation Approach to Explainable Information Quality Assessment. In Proceedings of the 34th IBIMA Conference, pp. 3919-3931.-https://researchgate.net/publication/337567351
[2] Ali, K.B., & Gosain, A. (2012). Predicting quality of data warehouse using fuzzy logic. International Journal of Business and Systems Research. 6 (3), 255-268. DOI: 10.1504/IJBSR.2012.047925-https://inderscience.com/info/inarticle.php?artid=47925
[3] Fasel, D., & Zumstein, D. (2009). A Fuzzy Data Warehouse Approach for Web Analytics. In M.D Lytras et al. (Eds.), WSKS 2009, LNAI 5736, pp. 276-285. Springe-https://www.researchgate.net/publication/225155489
[4] Fuzzy-based filter for ETL process-https://www.researchgate.net/publication/289618068
[5] National Institute of Standards and Technology(NIST), Data Integrity, NIST Computer Security Resource Center Glossary-https://www.nccoe.nist.gov/publication/1800-26/VolA/index.html
[6] S. H. Ali El-Sappagh, A. M. Ahmed Hendawi, and A. H. El Bastawissy, "A proposed model for data warehouse ETL processes," *Journal of King Saud University - Computer and Information Sciences*, vol. 23, no. 2, pp. 91–104, 2011-https://www.sciencedirect.com/science/article/pii/S131915781100019X?via%3Dihub
[7] C. Fox, A. Levitin, and T. Redman, "The nation of data and its quality dimension," Information Processing & Management, vol.30, no.1, pp. 9-19 -https://www.sciencedirect.com/science/article/abs/pii/0306457394900205?
[8] Overview of Data Quality-https://pmc.ncbi.nlm.nih.gov/articles/PMC9912223/?
[9] Big Data Quality Dimensions-https://www.scielo.br/j/jistm/a/dz8CSmJT3MXC7J5dw8qPfss/?format=html&lang=en&ilang=en
[10] IBM, "What is a Data Warehouse?"-https://www.ibm.com/think/topics/data-warehouse?
[11] Data Quality-Driven Improvement in Health Care-https://www.sciencedirect.com/org/science/article/pii/S1438887124005090?