© DEC 2023 | IRE Journals | Volume 7 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I6-1716725

Automated Data Quality Scoring for Analytics Readiness Using Integrated Profiling and Validation Frameworks

SAI LALITESH POTHUKUCHI

Abstract- ***The quality of datasets is a critical factor in identifying the stability and accuracy of the present-day analytics systems. In most data-intensive contexts, data quality, including missing values, data duplication, inconsistency in the schema, and drift in distribution, can cause a substantial impact on the results of analytical processes and result in unreliable insights. Initial data quality assessment has pointed to the necessity of systematic mechanisms in measuring the reliability, completeness, and consistency of datasets prior to their application in analytical decision-making and the significance of structured data quality assessment mechanisms in complex information systems (Christopher S. Carson, 2001). In spite of these developments, a number of existing analytics pipelines do not have automated processes that can measure the reliability of datasets in a single and scalable way. The current paper will present an automated framework of dataset trust scoring aimed at assessing analytics readiness by fusing messages of data profiling with rule-based validation procedures. The suggested solution consists of the combination of indicators of profiling, which include missingness, duplicate records, distribution drift, and invalid categorical values, and validation rules implemented by automated data integrity checks. The emergence of recent automated data profiling and quality scoring tools has confirmed the efficacy of the algorithmic assessment techniques in identifying data anomalies and enhancing the predictive analytics integrity (Hugo Moura et al., 2024). Based on such improvements, the suggested framework consists of a structured scoring model that combines various data quality indicators into a single dataset trust score. The framework also assesses the possibility of using dataset trust scores as predictors of downstream analytics stability and errors in analytical processes. The adaptive data quality scoring models have demonstrated their potential in industrial contexts in which drift-sensitive monitoring systems are utilized to ensure the stability of data-driven systems in the long term (Fatih Bayram et al., 2024). The proposed framework builds a scalable and automated data readiness evaluation system before the analytical processing by extrapolating these concepts. The research also adds to the expanding area of automated data governance by introducing an effective model of continuous quality assessment of the dataset, which would help organizations increase the reliability of analytics, minimize the spread of errors, and increase the confidence in the decision-making systems it is based on.***

Keywords: Automated Data Quality Scoring; Dataset Trust Score; Data Profiling and Validation; Analytics Readiness Assessment; Data Drift and Integrity Monitoring; Data Quality Automation; AI-Driven Data Governance.

I. INTRODUCTION

The swift growth of data-driven technologies has essentially reshaped organizational analytics, predictive modeling, and strategic decision-making. Analytics systems are becoming increasingly dependent upon in the context of finance, manufacturing, healthcare, as well as cloud computing to produce insights that affect operational and strategic decisions. The success and trustworthiness of these systems, however, are limited by the quality of the underlying data on a fundamental level. Unstable models, biased predictions, and unreliable analytical results may be the result of datasets that contain missing values, duplication, schema inconsistencies, invalid categorical entries, or distributional drift.

The contemporary analytics landscape is defined by the volumes, speed, and diversity of data, and the data sources are also heterogeneous and decentralized, and may include enterprise databases, cloud data warehouses, real-time data streams, and sensor-based platforms. Although these architectures can offer scalability and flexibility, they also increase data quality risks that the traditional manual inspection approach or ad hoc validation technique can hardly deal with. The larger and more dynamic the datasets, the more the traditional methods of quality assurance will prove ineffective, and the necessity to develop automated, systematic, and scalable mechanisms to determine the reliability of datasets before they are processed in analysis.

The aforementioned research has provided a definitive status that the quality of data dimensions, including completeness, consistency, uniqueness, validity, and

stability, is a crucial determinant of the credibility of the end product of analytics and machine learning. It has been demonstrated that poor-quality data can affect the statistical inference, decrease the predictive accuracy, augment model bias, and decrease the reproducibility of the analytical outcomes. Such failures can spread through the operational and decision-sensitive pipelines in environments where they are not allowed to happen, causing expensive mistakes and low confidence in the data-driven decision-making. Therefore, viewing data quality is not as a peripheral technical operation as it used to be, but rather as a prerequisite to analytics preparedness.

The latest developments in automated data profiling, rule-based validation structures, and AI-based data governance have enhanced the process of identifying the problem of data quality in large-scale and distributed data systems. Profiling methods make it possible to detect missing values, duplicates, anomalies, and distributional drift, and validation frameworks impose schema constraints, domain rules, and integrity policies. In spite of these advances, the majority of analytics pipelines continue to be based upon fragmented quality checks, which respond to individual problems but offer no integrated and comprehensible measure of the overall reliability of a dataset. Consequently, organizations do not have a common process of deciding whether a dataset is trustworthy enough to be analytically deployed or whether data quality problems can affect the stability of downstream analytics.

Although previous research has considered data profiling, validation models, and quality measurements separately, the current research paper provides an integrated trust scoring architecture that directly associates automated data quality testing with analytics preparedness and stability of downstream models. As opposed to the current methods, where the sole objective is attained by detecting or correcting data, the suggested framework models data quality as a forecasting tool of analytical consistency. This placement of data quality testing is an expansion of data quality testing beyond a diagnostic undertaking to a decision-making path in the enterprise analytics pipelines.

In order to close this gap, the present paper suggests an automated dataset trust scoring model that will be used to measure analytics readiness by combining automated data profiling cues with rule-based validation results. The framework brings together several dimensions of data quality into one composite score of trust, which provides a clear and scalable data set reliability measure. Through its operationalization of data quality as a measurable trust indicator, the suggested solution broadens the current quality assessment techniques to exceed the diagnostic mechanism of detecting quality towards a prospective technique of evaluating analytics consistency and error propagation.

The suggested framework also puts the dataset's trust scoring in the perspective of AI-based data governance and constant monitoring. The framework facilitates real-time quality analysis on cloud analytics and distributed systems as it can automatically assess, normalize, weight, and aggregate quality signals. By so doing, it offers organizations a viable way of ranking datasets, optimizing remediation with respect to these datasets, and minimizing analytical risk throughout the data life cycle.

The rest of this paper follows the following structure. Section 2 explores the notion of analytics preparedness and its correlation with the quality of data in contemporary data ecosystems. Section 3 discusses basic methods of data quality measurement and scoring. Section 4 presents the automatic data profiling signals that are applied into evaluating the trustworthiness of data sets, and Section 5 addresses the rule-based validation framework that implements data integrity. The design of the dataset trust score framework, scoring architecture, normalization, and weighting strategies are given in Section 6. Section 7 examines how dataset trust scores, stability of analytics, and propagation of errors relate. Section 8 makes the framework a part of the practices of AI-driven data governance and continuous monitoring. Section 9 addresses implications for scalable analytics systems, followed by conclusions and directions of future research.

II. DATA QUALITY AND ANALYTICS READINESS IN MODERN DATA ECOSYSTEMS

Analytics preparedness: The degree to which data resources, policies, and technologies are all accessible to credible analytical procedures. Most importantly,

data availability does not qualify readiness, but it is rather the level to which data meets the basic quality dimensions, including completeness, consistency, validity, and structural integrity (Elragal and Elgendy, 2024; Helbig et al., 2019).

In the event that datasets do not uphold these quality standards, systems used to carry out analysis become vulnerable to instability, low reproducibility, and high error rates. The empirical research has proved that data quality inefficiency may lead to biased statistical inference, reduced predictive accuracy, and poor reliability of machine learning models, especially in decision-intensive settings (Buchanan and Scofield, 2018; Razali, 2024). As a result, the need to critically assess the quality of datasets before using them to conduct advanced analytics and machine learning efforts has become a mandatory requirement in these endeavors (Sundararaj, 2023).

The other critical dimension of analytics preparation is that between trustworthy data infrastructure and effective organizational decision-making processes. Decision-driven environments that are data-driven demand that the analytical results should be an accurate reflection of data patterns. In case datasets have latent anomalies or structural errors, analytical findings are prone to make erroneous strategic decisions. The frameworks that focus on measuring the readiness of systems to analytical processes highlight the significance of the diagnostic assessment systems that could pinpoint the flaws of data systems prior to the deployment of analytical processes (Michael E. Ezerins et al., 2022). Such diagnostic models offer entities with instruments of assessing the capability of their data systems to facilitate sound analytical works.

The predictive analytics and machine learning models also rely on reliable datasets as a core factor to perform their role successfully. Predictive models are highly dependent on training groups to determine patterns and produce an accurate forecast. In the case of missing values, duplicate records in the training datasets or varying data structure, the performance of the model can be significantly impaired. Such problems may cause the model bias, higher predictive error, and lower validity of analytical information. A study of quality scoring methods in Extract-Transform-Load (ETL) pipelines indicates that the systematic data quality scoring systems may play an essential role in enhancing the quality of analytical results obtained by exposing and addressing data integrity violations in the course of the data processing phases (Azman Razali, 2024).

With the continued expansion of analytics capabilities of organizations, the necessity of automated processes of measuring the readiness of the dataset has become relevant. Conventional manual methods of assessing data quality tend to be inadequate to contemporary data ecosystems of big statistics of information combination, decentralized computers, and live analytics lines. Automated data quality scoring systems offer a flexible data quality scoring system where the integrity of a dataset is constantly measured and possible quality problems are recognized ahead of time before interfering with the analytical processes. Using automated assessment systems in analytics pipelines, organizations will be able to enhance their decision-making processes and make sure predictive models are trained using robust and credible data.

Figure 1: Data quality and analytics readiness in modern data ecosystems

III. FOUNDATIONS OF DATA QUALITY MEASUREMENT AND SCORING

The research synthesizes an extensive collection of traditional and current studies on the issues of data quality measurement, automated profiling, and the assessment of the trustworthiness of the analytical system. Early efforts like Christopher S. Carson (2001) give the conceptual base of the structured data quality assessment, with dimensions of completeness, consistency, and validity being key factors of the reliability of the dataset. In the same vein, the harmonized framework of data quality assessment put forward by Michael G. Kahn et al. (2021) emphasizes the role of standardized quality evaluation mechanisms in large observational data, which justifies the necessity of systematic quality measurement in the contemporary analytics setting.

Recent works build upon these basic ideas and present automated and adaptable methods of dataset analysis. As an example, Fatih Bayram et al. (2024) prove that adaptive data quality scoring frameworks that have drift-sensitive mechanisms enhance reliability and monitoring in industrial analytics pipelines to a large degree. Their results are aligned with the fact that evolving data problems can be identified with the help of a dynamic quality scoring system that traditional tools like static validation could fail to identify.

In general, the literature surveyed in this paper has proven that there is a continuum between traditional data quality assessment systems and modern automated scoring systems, with the latter gaining importance as a scalable and automated tool to assess the preparedness of a dataset to be analysed in a large-scale analytical system. Through a combination of these articles, the current paper places the suggested data quality evaluation framework as a trust score into a new research area of automated, scalable, and interpretable data quality assessment.

The trend of choosing larger datasets to perform analytics and make decisions has garnered a lot of interest in formal methods to quantify and evaluate the quality of data. Data quality measurement frameworks offer methodical ways of determining whether the datasets are up to standard specifications to be subject to reliable analytical processing. These systems generally consider several aspects of data integrity such as completeness, consistency, accuracy and structural validity. The need of such multidimensional evaluation methods is explained by the fact that the issues of data quality are seldom singular; on the contrary, they are usually the results of interactions between various characteristics of the datasets. Studies of data quality assessment models have put forth that any effective evaluation model must include various quality dimensions to make sure that datasets are interpretable to analytical and operational use (Christian Helbig et al., 2019).

The major difficulty with data quality measurement is that it requires the creation of standardized frameworks that would be able to support varied data environments. Heterogeneous data structures are frequently found in observational research data sets, enterprise databases, and in large scale analytical platforms and demand harmonized methods of evaluation. To deal with this issue, various works have put forward elaborate frameworks that incorporate standard metrics and evaluation methodologies which are used to assess the reliability of a dataset in various fields. These models emphasize the significance of well-organized quality assessment systems that allow companies to detect and address problems of data integrity prior to datasets being processed through analytical processes (Michael G. Kahn et al., 2021).

In addition to the conventional quality evaluation models, the notion of trust measurement has gained more and more popularity in digital systems. Measuring trust is concerned with establishing reliability of data sources, system or services using performance criteria and error trend. Trust scores are determined with statistical measurements and model evaluation, which is used to gauge the reliability of the system and the integrity of the data that is stored in the digital environment in many digital environments. The studies focusing on the way of predicting trust in the context of digital service ecosystems prove that quantitative scoring frameworks, typically assisted by confusion matrix assessment metrics, can be successfully used to determine the reliability of the system and help to prioritize the mechanisms in data-driven settings (Muhammad Hasnain et al., 2020). Such methods of measuring trust can be useful in understanding how reliability scoring mechanisms can be modified to measure the integrity of data sets in the analytical systems.

Data quality scoring is now even broader with the latest innovations in machine learning and automated analytics. The application of data-driven learning methods is becoming more widespread in order to evaluate the reliability of data sets by automated anomalies, inconsistencies and structural error detection. Such approaches allow the analytical systems to dynamically detect the quality problems and to respond to the varying data environment. Research investigating data-driven learning models in data quality evaluation prove that data quality issues in large datasets that continuously change can be detected using algorithmic models in large quantities (Elisa Tute et al., 2021).

The other new approach to data quality measurement is the incorporation ofFAIR data principles that take

into account that datasets are findable, accessible, interoperable, and reusable. According to FAIR, high-quality data is vital because it is necessary to ensure that not only high-quality data is used in the short term, but also in the long term to conduct data reuse and interoperability across digital ecosystems. Research on the FAIR-based data quality models highlights that effective data quality evaluation strategies should be transparent and have quality indicators that would endorse sustainable data governance practices (Nicholas Nicholson et al., 2025). These frameworks add to building more efficient and scalable data quality assessment models through the introduction of the tenets of transparency, interoperability, and controlled metadata management.

Together, these theoretical models and scoring systems define the conceptual basis of automated dataset trust scoring systems. The modern data quality measurement framework is characterized by the availability of tools that help organizations to assess the reliability of the datasets in the complex analytical environment due to the combination of multidimensional quality indicators, statistical evaluation methods, and automated detection mechanisms. Such foundations help to develop automated scoring architectures that can be used to measure the trustworthiness of a dataset and to make sure that the operating analytical systems deal with reliable and high-quality data.

Table 1: Core Dimensions of Data Quality Measurement

Dimension	Description	Common Causes of Issues	Impact on Analytics
Completeness	Degree to which required data values are present in a dataset	Data entry omissions, sensor failures, incomplete records	Leads to biased models, inaccurate statistical inference
Consistency	Extent to which data values follow predefined formats, rules, or categories	Inconsistent formatting, integration of heterogeneous data sources	Reduces feature reliability and causes model instability
Uniqueness	Degree to which records appear only once without duplication	Repeated ingestion, system integration errors, duplicate transactions	Distorts statistical distributions and aggregation results
Validity	Compliance of dataset values with defined schemas, ranges, or domain constraints	Schema violations, incorrect data types, invalid categorical values	Causes processing failures and analytical inaccuracies

The fundamental dimensions that are employed to analyze the quality of datasets in the analytical environments are expressed in Table 1. All the dimensions are indicative of a serious concern of data integrity: Completeness is the rate of lost or null values, Consistency is the rate of pre-determined formats and valid entities, Uniqueness is the rate of duplicate records that may alter the statistical analysis, and finally, Validity is the rate of compliance with schema regulations and domain constraints. The new column, Common Causes of Issues, shines light on the common causes of problems in quality, whereas the Impact on Analytics elucidates the impact of each dimension on the reliability of analytics. Combined, these dimensions will give an overall picture regarding the reliability of the dataset, which will be the basis of automated scoring.

© DEC 2023 | IRE Journals | Volume 7 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I6-1716725

IV. AUTOMATED DATA PROFILING SIGNALS FOR DATASET TRUST ASSESSMENT

Automation of data profiling is important for checking data reliability before it enters analytic workflows. 'Data profiling' refers to the systematic analysis of data to summarize statistics and spot any unusual patterns, outliers, or errors that could impact results. As organizations handle larger and more complex data, automated profiling is necessary to catch data quality issues that may otherwise go unnoticed. Research on automated profiling shows that the measures identified during profiling help determine if data is suitable for prediction or machine learning (Hugo Moura et al., 2024).

Missingness detection refers to the process of identifying and quantifying missing or null values within a dataset. Missing data are values that are absent from specific fields, while null values refer to entries explicitly marked as empty. The existence of missing data can significantly affect the reliability of analytical models, particularly when these gaps occur in essential variables for predictive modeling. In most profiling tools, null percentages are calculated for each column to determine which fields have incomplete data. Recognizing patterns of missingness helps analysts decide whether to impute, remove, or correct values based on the nature and distribution of the missing data. In ongoing data quality management systems, monitoring missing data patterns is vital for maintaining data integrity, especially in large-scale systems (Taleb et al., 2021).

Duplicate detection is another significant profiling signal that is concerned with detecting repeated records in a dataset. The cause of duplicate entries can be system integration errors, or because there is a repeated processes in the ingestion of data, or because of inconsistency in data entry mechanisms. These replicas may cause distortion of the statistical distributions, prejudice aggregation findings, and errors in analysis models. Duplicate detection algorithms are thus used in automated profiling systems to take the ratio of duplicates in datasets by comparing record identifiers or key attributes. Benchmark research that assesses data cleaning and preprocessing systems emphasizes that working duplicate detection systems really enhance information credibility and general analytics execution in significant real-life datasets (Pedro Martins et al., 2025).

Another profiling dimension is the data drift detection, which is the change in the data statistical distribution with time. Data drift may arise when the properties of actual data are not similar to those of previous data that was used to train predictive models. These changes will reduce the accuracy of the model and cause inaccurate analytical results. The statistical monitoring techniques, e.g., the distribution divergence measures, are thus generally utilized in detecting the drift trends in the changing datasets. To improve the automated quality of the system, drift detection is becoming a common feature of quality improvement systems, whereby analytical models are kept consistent with the current data trends, especially in dynamic data systems where datasets are constantly being updated (Sarr, 2024).

Lastly, anomaly detection and category error detection are significant profiling indicators to trust the dataset. Categorical variables in most datasets would need to fit the predetermined sets of values, or domain constraints. There is a possibility of invalid categories or unanticipated label patterns that may mean that there is an error in recording data, a discrepancy in the schema, or even integration issues between data systems. Such anomalies can be identified with the help of automated profiling tools that examine the distribution of categorical values and reveal such entries that do not correspond to expected distributions. With the addition of anomaly detection to data profiling processes, analytical systems can detect possible quality problems in the early data pipeline stages and minimize the likelihood of errors being carried forward into the latter analytics.

Collectively, these profiling messages can be used as the necessary signals to determine the integrity and reliability of the data sets. When coupled together in automated evaluation processes, organizations can have scalable data quality monitoring systems that will be able to continuously perform evaluations on the trustworthiness of datasets in complex analytical environments.

© DEC 2023 | IRE Journals | Volume 7 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I6-1716725

Table 2: Automated Profiling Signals for Dataset Trust Scoring

Profiling Signal	Measurement Metric	Detection Technique	Quality Dimension Affected
Missingness	Percentage of null or missing values per column	Automated column profiling and completeness checks	Completeness
Duplicates	Duplicate record ratio within dataset	Record comparison and key-based matching algorithms	Uniqueness
Data Drift	Statistical divergence between historical and current distributions	Kolmogorov–Smirnov test, PSI monitoring, distribution comparison	Stability
Category Errors	Frequency of invalid or unexpected categorical values	Pattern detection and rule-based validation	Consistency

Table 2 presents the main indicator automated profiling signals to evaluate the level of dataset trustworthiness. Every signal is connected to a quantifiable score, a detection method, and the quality dimension to which the signal influences. Missingness checks the existence of the null values, Duplicates checks the repeated records, Data Drift checks the changing of the statistics with time, and Category Errors checks the invalid categorical values. This table connects every signal with the corresponding quality dimension to give us a structured perspective of how automated profiling can inform dataset trust scoring so that organizations can be informed about analytics readiness and enhance it in a systematic manner.

V. RULE-BASED VALIDATION FRAMEWORKS FOR DATA INTEGRITY

A data integrity step is important in the preparation of datasets to be used to make reliable analytics. Mechanicalized rule-based validation systems have now become imperative in applying data quality norms in a structured manner to extensive and multifaceted collections of information. Such structures enable organizations to establish, instantiate, and maintain rules of validation that guarantee a dataset to satisfy structural, semantic, and domain specifications prior to being subjected to analytical channels. The research has revealed the increasing role of automated validation tools in ensuring data soundness, especially in large-scale and distributed data systems (Fariha A. et al., 2021; C. Devi et al., 2025).

Pandera and Great Expectations are two popular models used in this sphere. Pandera is a Python package that can be used to verify the schema through the use of column types, nullable constraints, and ranges of values that can be imposed directly on dataframes. This will be used to make sure that the dataset structure is in line with the predefined analytical expectations and thus avoids future analytics workflow errors. In Great Expectations, in contrast, a flexible, declarative method of checking data is provided against a very diverse range of rules, such as range checks, unique key constraints, and categorical constraints. These two frameworks can be used to promote automation, constant monitoring, and connection with the data pipelines, allowing real-time awareness of any data quality breach (O. Ozonze et al., 2023; AI Augmented Frameworks for Data Quality Validation, 2025).

Mechanisms of Enforcement of Rules.

- Schema Validation: Assures that every column is validated to a specified type of data, whether it is null or not, and structure. Schema validation eliminates data errors during processing and ensures that the data sets are of a baseline quality (Fariha et al., 2021).
- Constraint Checking Constraint checking will compare the data to a set of rules, e.g., numeric ranges, permitted categories, or relational constraints. The given step assists in detecting oddities, which do not conform to logical or

domain-specific assumptions, including invalid categorical choices or out-of-range values (Devi et al., 2025).

- Data Integrity Policies: Imposes more generic policies in the organization, such as key identifiers being unique, referential integrity between tables, and meeting governance policies. The application of these policies in automated structures would also guarantee the data sets to be reliable in the long run and across systems (AI Augmented Frameworks for Data Quality Validation, 2025).

With a combination of profiling signals in the preceding sections and rule-based validation, organizations will be able to gain a robust dataset trust score in such a way that both statistical and rule-based measurements are combined into a single measure.

Table 3: Example Validation Rules for Automated Data Integrity

Validation Rule	Description	Tool Support
Schema validation	Enforces column types, nullability, and structure	Pandera
Range constraints	Ensures numeric values fall within defined limits	Great Expectations
Allowed categories	Validates that categorical values adhere to controlled vocabularies	Both
Unique keys	Ensures records are unique to prevent duplicates	Both

An example of rules-based validation mechanisms that are employed to enforce the icons of dataset integrity is presented in Table 3. Schema validation can be used to ensure that datasets match structural requirements, and range constraints and legal categories can be used to validate the validity of numerical and categorical data. Distinct key enforcement eliminates replication and ensures that the identity is unique. The frameworks, such as Pandera and Great Expectations, use these automated rules to conduct continuous monitoring and verification of dataset quality, which is part of the automated dataset trust scoring framework.

VI. DESIGNING THE DATASET TRUST SCORE FRAMEWORK

The methodological value of this paper can be characterized as the creation of a systematic dataset trust scoring system that combines automated data profiling indicators with rule-based validation systems. As opposed to the conventional methods of assessing the quality of individual measures in isolation, the suggested framework combines various quality measures, such as completeness, consistency, uniqueness, and compliance with validation within a single scoring model, which would consider the overall trustworthiness of a dataset.

The framework is mainly conceptual, though it is accompanied by a demonstrative case scenario explaining the ways in which the metrics of profiling and validation rules may be aggregated to compute a dataset trust score. This case draws attention to the fact that automated scoring systems can be used in real-world analytical pipelines to determine the readiness of datasets prior to analytical processing.

The existing conceptual framework can be applied in various analytical settings since the current research is founded on a generalizable architecture, although the use of large-scale real-world datasets would further bolster the framework. The frameworks of this type are regularly applied in new domains of the data governance research to frame underlying models prior to the large-scale validation tariff being established. Further research can build upon this by applying the proposed system of scoring as an operational analytics pipeline and assessing its predictive capabilities on a variety of datasets and machine learning problems.

The dataset trust score framework gives an organized method of determining the reliability and analytical preparedness of datasets. Through the combination of profiling indicators and rule-based validation outcomes, the framework produces one composite rating that demonstrates the general credibility of a

dataset. This score may be utilized to rank datasets to be analyzed, track the data quality across time periods, and intervene in order to enhance reliable datasets. Recent studies show the power of scoring models within the industrial and organizational analytics setting by showing that a structured trust score would be useful to forecast downstream model stability and error rates (Fatih Bayram et al., 2024; S. M. R. Nalla, 2025).

6.1 Simulated Case Study: Dataset Trust Scoring in Practice.

In order to demonstrate the practical implementation of the proposed framework, a simulated dataset trust scoring situation is described. Suppose that there is a structured dataset with 1 million transactional records of a cloud-based enterprise system.

Step 1: Profiling Results

Quality Signal	Measured Value
Missing values	8% of records
Duplicate records	3%
Category errors	5%
Data drift	Moderate drift detected

Step 2: Normalized Scores (0–1 Scale)

Component	Score
Completeness	0.92
Consistency	0.95
Uniqueness	0.97
Validation compliance	0.90

Step 3: Weighted Trust Score Calculation

Trust Score=(0.92×0.30)+(0.95×0.25)+(0.97×0.20)+(0.90×0.25)=0.93

Interpretation

The score of 0.93 is high analytics readiness. This data would be allowed to be used in predictive modeling and downstream analytics based on predefined thresholds. On the contrary, a dataset with a score of less than 0.80 would automatically initiate remediation processes before analytical.

Scoring Architecture

Scoring architecture is developed in a way that it integrates several quality dimensions into a single indicator. All the data sets are assessed in terms of profiling signals like missingness, duplicates, data drift, and category errors, and rule-based compliance with validation. These measures are then added up individually and then summed up to create one composite score. The framework promotes transparency by distinguishing between measurement and aggregation, and this is why it allows an analyst to track which quality dimensions matter the most to the trust score (Razali, 2024).

Signal Aggregation

Signal aggregation entails a weighted combination of multiple quality indicators in a bid to represent their relative weight. As an example, missing data can be a stronger predictive analytics issue than small-scale categorical inconsistencies, so it is given a higher weight in the composite score. The aggregation makes sure that the score of trust is a more moderate outlook of the entire quality of the dataset, and not some particular measure predominating the trust score. Depending on the context of application, the aggregation step may make use of linear combinations, geometric means, or other statistical methods.

Normalization and Weighting.

The quality signals are brought to a standard scale (e.g., 0 to 1) in order to make them comparable between different metrics. Normalization helps in avoiding the occurrence of signals that have bigger numeric values and hence may affect the trust score out of proportion. A predetermined weight is then used to multiply each of the normalized signals, representing the relative significance of the signal to the reliability of the analysis. Lastly, the weighted metrics are added together to generate the final dataset

trust score, which gives one interpretable score of dataset reliability that can be fed into downstream analytics.

This framework offers a framework through which automated profiling metrics can be combined with rule-based validation compliance, and thus offer a scalable, systematic, and interpretable system of assessing the readiness of datasets in large and complex data ecosystems.

Table 4:Dataset Trust Score Components

Component	Metric	Weight
Completeness	Missing values	0.30
Consistency	Valid categories	0.25
Uniqueness	Duplicate ratio	0.20
Validation compliance	Rule satisfaction	0.25

Table 4 identifies the elements with which the dataset trust score is calculated. The measures of completeness measure the percentage of data that are missing. Consistency is the measure that determines whether data meet the correct categorical values. Uniqueness measures whether there are duplicate records. Lastly, the measure of validation compliance measures the extent to which data meets the rule-based validation checks. Every component is given a weight in terms of its relative significance in establishing the overall data reliability. A combination of these weighted parts gives one dataset trust score, which organizations can use to have a clear and practical view of their analytics preparedness.

VII. PREDICTING ANALYTICS STABILITY AND ERROR PROPAGATION

The dataset trust scores not only act as a diagnostic of the quality of data, but also as a predictor of the stability of downstream analytics. Data that has a high trust are always linked to a reduction in prediction error, enhanced model reproducibility, and less sensitivity to data changes, whereas data with low trust has the propensity to increase the error propagation throughout analytical pipelines (Bayram et al., 2024; Luo et al., 2023).

Trust scoring frameworks enable organizations to offer an active approach to reducing the risk of analytics before model creation, rank quality data assets high, and limit instability in predictive and machine learning frameworks by quantifying data reliability before model development (Saini et al., 2024; Doris and Potter, 2024).

Dataset Trust Score and Analytics Stability

The concept of analytics stability is defined as the stability of model results or analytical results in the event of slight variations or updates in dataset. Poorly quality datasets, such as missing values, duplicate records or drift in distribution can exaggerate the diversity of model predictions, lower the reproducibility and both diminish trust in analytic knowledge. Research on industrial and cloud-based analytics settings indicates that trust-based scoring systems can be trusted as reliable proxies to analytics stability by advising data collections, which have a high likelihood of exhibiting stable performance with time passing (Luo et al., 2023).

Prediction Error Rates

Correlation with the prediction error rates in machine learning and statistical models can also be performed with the trust score. The increase in the level of trust correlates with the decrease in the mean squared errors, better classification outcomes, and decreased bias of regression and forecasting analytic tasks. With automated scoring systems, organizations can determine the possible error proliferation by labeling datasets with lower scores prior to processing and, therefore, lowering retraining expenses and enhancing reliability overall (Doris and Potter, 2024).

Reliability of Machine Learning Outputs

The sensitivity of machine learning models to data quality problems is extra, since any error in input data can grow and multiply as a result of a sophisticated model architecture. Datalysts that have high ratings of trust lead to more stable feature representations, stable model training, and strong evaluation metrics. The inclusion of dataset trust scores into the model development lifecycle allows organizations to focus on better data to train and validate their models and

eventually improve machine learning results in terms of reliability and interpretability (Saini et al., 2024).

Practically, the dataset trust score is both diagnostic and predictive, and it correlates the data quality indicators that can be measured with the anticipated stability and error properties in the downstream analytics. This connection offers a quantitative basis for the proactive data control and evidence-based ruling of the dataset usage in predictive modeling and analytical chains.

VIII. AI-DRIVEN DATA GOVERNANCE AND CONTINUOUS MONITORING

With the continued deployment of cloud-based and large-scale analytics environments by organizations, AI-based data governance has become crucial in guaranteeing the reliability, compliance, and usability of datasets. Automated governance systems are policy enforcement, constant monitoring, and smart quality analysis mechanisms used to ensure the integrity of data across distributed systems. Using AI technologies, organizations will be able to decrease the number of people working on the manual level, identify irregularities in real-time, and employ proactive measures to enhance the quality of data (N. Prasad & L. K. Paripati, 2025).

Automated Monitoring and Governance Policies

Data quality metrics are constantly monitored through automated monitoring systems, and governance policies are implemented so that the data sets meet the organizational and regulatory requirements. As an illustration, the cloud analytics environment is likely to enforce a rule of schema verification, range enforcement, and duplicate detection to ensure reliability in data. Such systems, along with policy-based control of AI-based detection algorithms, allow detecting quality violations in real-time and sending alerts or remedial measures. Research on AI-assisted governance points to the fact that the meaning of automated monitoring is that the spread of errors decreases, regulatory compliance is facilitated, and the efficiency of making decisions in the organization is improved (K. B. Tenneti et al., 2024).

Automated Quality Improvement

In addition to the observation, AI-based models can proactively enhance the quality of data by automated remediation and grading systems. Automated quality improvement involves the activities of imputation of missing values, the correction of inconsistent categorical items, the reconciliation of duplicate records, and the identification of abnormalities to intervene in correcting the anomalies. The implementation of automated scoring systems into these workflows allows assigning the so-called trust scores, which not only evaluate the quality of raw data but also the efficiency of corrections introduced to the latter. In recent studies, it has been proven that AI-based quality assessment can be faster, more consistent, and more accurate than manual assessment (especially in dynamic or high-volume data), and it is more efficient, particularly in costly or labor-intensive settings (Marco Bevilacqua et al., 2025; N. M. Bui and J. S. Barrot, 2025; Nakamoto et al., 2023).

Integration to Dataset Trust Scoring.

Automated monitoring, rule enforcement, and quality improvement with the help of AI help to maintain a data quality cycle. Signals, rule-driven validations, and AI-assisted corrections are used to feed the dataset trust scoring system to make sure that the calculated trust score is a reflection of the quality of the current dataset and of the effects of automated remediation actions. This integration helps organizations to have sustainable analytics preparedness, minimize chances of error propagation, and build trust in machine learning products and forecast models.

Organizational management can establish a scalable and dynamic data monitoring, scoring, and data improvement mechanism in real-time through the integration of AI-driven governance in cloud analytics processes, which can form the basis of trustworthy, data-centric decision-making.

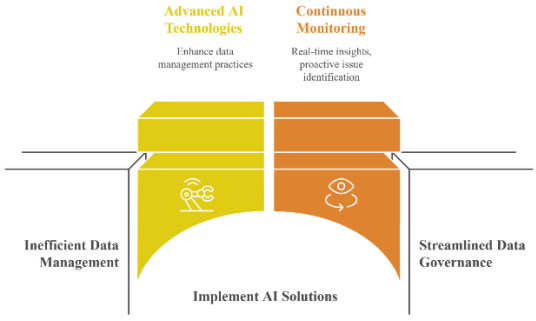

Figure 2: AI-driven data governance and continuous monitoring

© DEC 2023 | IRE Journals | Volume 7 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV7I6-1716725

IX. IMPLICATIONS FOR SCALABLE DATA ANALYTICS SYSTEMS

The introduction of the dataset trust score model has strong implication to organizations that want to scale their analytics operations. The trust score allows enterprises to place emphasis on high-quality datasets, perform continuous monitoring of data quality, and integrate datasets into the process of analytical work with certainty by offering a quantitative metric of the reliability of datasets. This method eliminates the possibility of error diffusion, further improves the reproducibility of analysis findings, and makes predictive models more stable overall (Abdelrahman Elragal and Nour Elgendy, 2024).

Benefits of Trust Scoring

Trust scoring offers understandable and transparent measures that determine the quality of the data at various dimensions. The scores can be used by organizations to determine which datasets need to be remedied and made available to analytical processes, allocate resources effectively, and make wise decisions regarding data integration. Trust scoring, a formal method of evaluating the reliability of datasets, also enables a conversation between technical personnel, data stewards, and business decision-makers, so that data quality implications are introduced into the business analytics plans accordingly.

Impact on Enterprise Analytics

The datasets in an enterprise analytics environment tend to be fed by several systems and updated in real time. A uniform trust scoring system allows organizations to test and compare datasets in an objective way and enhances a consistency level among the analytical pipelines. Studies focusing on data quality frameworks, which comply with FAIR, have highlighted the advantages of including structured indicators of quality in improving the interoperability of an analytical system and the reliability of running on reliable data when working with large-scale analytics systems (Nicholas Nicholson et al., 2025). As a result, cloud and distributed analytics setups are capable of delivering greater efficiency, decreasing operational risk, and more congruence with the governance standards in organizations.

Improved Decision Reliability

The trust score framework is also directly related to the reliability of enterprise decision-making. Organizations can minimize uncertainty in predictions, prevent the causes of errors that are costly, and facilitate evidence-based strategic planning by ensuring that the models and results of the analytical processes are based on high-quality data. Research has shown that datasets having a high trust score are always associated with lower error rates, more predictable machine learning, and better predictive accuracy, and this has instilled more confidence in decisions made by using data (Elragal and Elgendy, 2024).

In general, automated data quality scoring and trust-based assessment as a part of enterprise analytics pipelines enforces the creation of a scalable, reliable, and governance-oriented system of contemporary data management. By enabling the connection between raw data quality measurement and actionable analytics, the dataset trust score framework can allow organizations to take advantage of high-volume distributed datasets and allow organizations to maintain the integrity of decisions.

Figure 3: Implications for scalable data analytics systems.

X. CONCLUSION

This paper presents a broad automated dataset trust scoring system that would assess analytics preparedness in current data ecosystems. The framework makes use of profiling signals, such as missingness, duplicates, data drift, and category errors, and rule-based validation frameworks, such as Pandera and Great Expectations, to offer a single metric, which measures the reliability of the data. The

scoring architecture of the framework includes normalization, weighting, and aggregation so that separate quality dimensions have a contribution to the overall dataset trust score.

The suggested framework facilitates a higher quality of analytics readiness assessment regarding encouraging organizations to gain information on both high- and low-quality datasets, anticipate possible error propagation, and enhance machine learning results reliability. The system can facilitate scalable, real-time data quality analysis, with the help of AI-based automated monitoring and continuous governance policies, and minimize the threat of mistakes in cloud-based analytics and distributed data environments. The conceptual and operational elements of this framework are represented in tables 1 to 4 that establish the relationship between quality dimensions, profiling signal, validation rules, and the calculation of trust score in a systematic and publication-presentable way.

Future Work

This study can be extended in the future through:

- Adaptive weighting models: This is where the contribution of individual quality dimensions is dynamically adjusted depending on the dataset context, the demand for the analysis, or even the sensitivity of the model.
- Real-time scoring systems: adopt streaming data evaluation systems to constantly update trust scores in response to the changing datasets.
- Integration with machine learning pipelines: make trust scoring an explicit part of model training, validation, and deployment processes, to take control of input data quality and enhance predictive accuracy.

The dataset trust scoring framework offers a strong basis of scalable, trustworthy, and governance-relevant analytics by supporting informed decision-making in contemporary enterprise and research settings, and automated and AI-assisted data quality measurement.

REFERENCES

[1] (2025). *AI-Augmented Frameworks for Data Quality Validation: Integrating Rule-Based Engines, Semantic Deduplication, and Governance Tools for Robust Large-Scale Data Pipelines.* International Journal of Advanced Artificial Intelligence Research, 2(08), 9–15. https://aimjournals.com/index.php/ijaair/article/view/382

[2] Bauer, J. C., John, E., Wood, C. L., Plass, D., & Richardson, D. (2020). Data Entry Automation Improves Cost, Quality, Performance, and Job Satisfaction in a Hospital Nursing Unit. *Journal of Nursing Administration*, *50*(1), 34–39. https://doi.org/10.1097/NNA.0000000000000836

[3] Bayram, F., Ahmed, B. S., & Hallin, E. (2024). *Adaptive Data Quality Scoring Operations Framework Using Drift-Aware Mechanism for Industrial Applications. Journal of Systems and Software, 217.* https://doi.org/10.1016/j.jss.2024.112184

[4] Bevilacqua, M., Oketch, K., Qin, R., Stamey, W., Zhang, X., Gan, Y., … Abbasi, A. (2025). When Automated Assessment Meets Automated Content Generation: Examining Text Quality in the Era of GPTs. *ACM Transactions on Information Systems*, *43*(2). https://doi.org/10.1145/3702639

[5] Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, *50*(6), 2586–2596. https://doi.org/10.3758/s13428-018-1035-6

[6] Bui, N. M., & Barrot, J. S. (2025). ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. *Education and Information Technologies*, *30*(2), 2041–2058. https://doi.org/10.1007/s10639-024-12891-w

[7] Carson, C. S. (2001). Toward a framework for assessing data quality. *IMF Working Paper.* https://doi.org/10.5089/9781451844269.001

[8] Devi, C., Inampudi, R. K., & Vijayaboopathy, V. (2025). *Federated Data-Mesh Quality Scoring*

with Great Expectations and Apache Atlas Lineage. Journal of Knowledge Learning and Science Technology, 4(2), 92–101. https://doi.org/10.60087/jklst.v4.n2.008

[9] Doris, L., & Potter, K. (2024). Continuous Monitoring and Improvement: Implement continuous monitoring of AI models to detect and correct issues in real-time. I-Manager s Journal on Artificial Intelligence & Machine Learning.

[10] Elragal, A., & Elgendy, N. (2024). A data-driven decision-making readiness assessment model: The case of a Swedish food manufacturer. *Decision Analytics Journal*, *10*. https://doi.org/10.1016/j.dajour.2024.100405

[11] Ezerins, M. E., Ludwig, T. D., O'Neil, T., Foreman, A. M., & Açıkgöz, Y. (2022). Advancing safety analytics: A diagnostic framework for assessing system readiness within occupational safety and health. *Safety Science*, *146*. https://doi.org/10.1016/j.ssci.2021.105569

[12] Fariha, A., Tiwari, A., Radhakrishna, A., Gulwani, S., & Meliou, A. (2021). *Conformance Constraint Discovery: Measuring Trust in Data-Driven Systems. Proceedings of the ACM SIGMOD International Conference on Management of Data.* https://doi.org/10.1145/3448016.3452795

[13] Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking. *IEEE Access*, *8*, 90847–90861. https://doi.org/10.1109/ACCESS.2020.2994222

[14] Helbig, C., et al. (2019). Data quality assessment framework for critical raw materials. *Resources, Conservation and Recycling.* https://doi.org/10.1016/j.resconrec.2019.104564

[15] Kahn, M. G., et al. (2021). Facilitating harmonized data quality assessments: A data quality framework for observational research data collections. https://doi.org/10.1186/s12874-021-01252-7

[16] Luo, F., Ge, N., & Xu, J. (2023). Power Supply Reliability Analysis of Distribution Systems Considering Data Transmission Quality of Distribution Automation Terminals. *Energies*, *16*(23). https://doi.org/10.3390/en16237826

[17] Martins, P., Cardoso, F., Váz, P., Silva, J., & Abbasi, M. (2025). *Performance and Scalability of Data Cleaning and Preprocessing Tools: A Benchmark on Large Real-World Datasets.* Data, 10(5), 68. https://doi.org/10.3390/data10050068

[18] Moura, H., et al. (2024). *Automated Data Profiling and Scoring Methods for Predictive Analytics. Information Processing & Management.* https://doi.org/10.1016/j.ipm.2024.103903

[19] Nakamoto, R., Flanagan, B., Yamauchi, T., Dai, Y., Takami, K., & Ogata, H. (2023). Enhancing Automated Scoring of Math Self-Explanation Quality Using LLM-Generated Datasets: A Semi-Supervised Approach. *Computers*, *12*(11). https://doi.org/10.3390/computers12110217

[20] Nalla, S. M. R. (2025). *Building an AI Trust Score: A Data-Driven Framework to Evaluate Dataset Fitness.* International Journal of Computing and Engineering, 7(20), 54–63. https://doi.org/10.47941/ijce.3091

[21] Nicholson, N., Carvalho, R. N., & Štotl, I. (2025). A FAIR perspective on data quality frameworks. *Data.* https://doi.org/10.3390/data10090136

[22] Ozonze, O., Scott, P. J., & Hopgood, A. A. (2023, December 1). Automating Electronic Health Record Data Quality Assessment. *Journal of Medical Systems*. Springer. https://doi.org/10.1007/s10916-022-01892-2

[23] Prasad, N., & Paripati, L. K. (2025). AI-Driven Data Governance Framework For Cloud-Based Data Analytics. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5052472

[24] Razali, A. (2024). *Improving Data Reliability Assessment in ETL Processes Through Quality Scoring Techniques in Data Analytics. International Journal on Informatics Visualization.* https://doi.org/10.1109/icst50505.2020.9732870

[25] Saini, H., Singh, G., Dalal, S., Moorthi, I., Aldossary, S. M., Nuristani, N., & Hashmi, A. (2024). A hybrid machine learning model with self-improved optimization algorithm for trust and privacy preservation in cloud environment. *Journal of Cloud Computing*, *13*(1). https://doi.org/10.1186/s13677-024-00717-6

[26] Sarr, D. (2024). Towards explainable automated data quality enhancement without domain knowledge. https://doi.org/10.48550/arXiv.2409.10139Taleb, I., Serhani, M. A., & Bouhaddioui, C. (2021). *Big Data Quality Framework: A Holistic Approach to Continuous Quality Management. Journal of Big Data, 8(76).* https://doi.org/10.1186/s40537-021-00468-0

[27] Tariq, A., et al. (2025). A survey of data quality measurement and monitoring tools. *Frontiers in Artificial Intelligence.* https://doi.org/10.3389/frai.2025.1621514

[28] Tenneti, K. B., Pandula, S., & Pandula, S. (2024). Comparative Analysis of Traditional and AI-Driven Data Governance: A Systematic Review and Future Directions in IT. *International Journal of Computer Trends and Technology*, *72*(11), 150–158. https://doi.org/10.14445/22312803/ijctt-v72i11p116

[29] Tute, E., Ganapathy, N., & Wulff, A. (2021). *A Data-Driven Learning Approach for the Assessment of Data Quality.* BMC Medical Informatics and Decision Making, 21, 302. https://doi.org/10.1186/s12911-021-01656-x

[30] Venkatraman, S., & Sundarraj, R. (2023). Assessing organizational health-analytics readiness: artifacts based on elaborated action design method. *Journal of Enterprise Information Management*, *36*(1), 123–150. https://doi.org/10.1108/JEIM-10-2020-0422