

Research on Enterprise-level Data Quality Measurement and Standards for Closed-Loop Governance of Lakewarehouse Integrated Data Platform

Yun Lou

Michigan State University, East Lansing, MI, 48824, US

Keywords: Data Platform; Data Quality; Data Standards; Lake Warehouse Integration; Data Governance; Closed-Loop Management

Abstract: Enterprise Data Platforms are evolving from centralized storage facilities to all-purpose foundations for connecting business semantics, data products and intelligent applications. The old "post-event cleanup" fails to address cross-source heterogeneity, data-quality drift and accountability gaps. This paper proposes a governance mechanism for lakewarehouse integrated platforms that includes standard objects, quality indicators, rule enforcement, lineage identification and work order closure. Based on literature from the past three years and two publicly available enterprise surveys, the study establishes indicators such as completeness, accuracy and consistency, as well as risk priority models, designs hierarchical standards, data contracts and tiered service agreements. Based on the above results, the main way to enhance the quality of enterprises is not to add standalone verification rules, but to strengthen standards, responsibilities, evidence and improvement actions throughout the operating chain of the platform. This solution will help the group company improve the quality of data, reduce fault handling costs, and enhance the reliability of data for analysis and artificial intelligence.

1 Introduction

With the simultaneous construction of transaction systems, Internet of Things services, customer relationship management systems, and generative artificial intelligence applications, enterprise data quality has shifted from report accuracy to a system-level capability spanning collection, processing, sharing, and consumption. Fu et al. argue that data quality should be understood in relation to provenance, user interaction, and the value created in a specific industrial context rather than as a one-time cleaning result.[1]

Current research examines both organizational governance and technical quality management. Bernardo et al. show that data governance and quality management increasingly support innovation and value creation rather than serving only as compliance mechanisms.[2]

Different industries still use inconsistent names and scopes for concepts such as accuracy, completeness, timeliness, and semantic consistency. Miller et al. map heterogeneous terminology to ISO 25012 and introduce governance, usefulness, quantity, and semantic dimensions, reinforcing

the need for cross-domain standardization.[3]

Reliable process data require more than technical validation; authority, access control, accountability, and lifecycle management also matter. Goel et al. develop a data-governance framework for process mining through a Delphi study and show that governance practices need to be embedded in process-data management.[4]

Based on this, this paper will investigate "how data platforms can transform enterprise standards into an executable, measurable and accountable closed-loop mechanism". It will use a combination of literature review, secondary statistical analysis of public surveys and mechanism design methods to follow the following steps: current situation analysis, problem identification, strategy construction and conclusion. The contributions of this paper are as follows: Use standard objects as the main line of governance, unify five core quality indicators, connect producers and consumers through data contracts, and drive work orders and debriefing via quality risk prioritization; thus, shift governance from project-based rectification to the platform's routine operation.

2 Current Status Analysis of the Research Topic

2.1 Platform-based governance shifts from centralized processing to end-to-end collaboration

Enterprise architectures are increasingly moving from centrally managed datasets toward data products delivered through data-mesh and data-fabric approaches. Blohm et al. emphasize that reusable, discoverable, understandable, and fit-for-purpose data products require explicit management mechanisms and domain-aware quality expectations, making the platform a key carrier of enterprise data governance.[5]

Industry evidence from data-mesh implementations shows that decentralized data ownership and federated governance create practical challenges in responsibility allocation, product maintenance, and cross-domain coordination. These findings support an end-to-end governance model in which platform capabilities and organizational accountability evolve together instead of relying on isolated quality checks.[6]

2.2 Quality standards shift from field-based rules to contextualized data product contracts

Uniform coding, data types, and value ranges remain the foundation of standardization, but field-level constraints alone cannot guarantee fitness for use. Research on service-based data pipelines shows that quality requirements must be evaluated together with metadata, processing stages, and protection policies across the pipeline lifecycle, making context and consumption requirements part of quality management.[7]

Recent big-data quality-management research further emphasizes standardized quality dimensions, centralized governance, continuous monitoring, and AI-assisted improvement across distributed environments. Therefore, mature enterprise practices should connect enterprise standards, subject-domain standards, data-product contracts, and executable operating rules through a unified platform mechanism.[8]

2.3 Governance pressures revealed by public surveys

Recent multidimensional data-quality assessment research argues that static rule sets are insufficient for heterogeneous modern data environments and calls for automated, scalable evaluation across accuracy, completeness, timeliness, consistency, relevance, and credibility.[9] The following industry survey is used only as descriptive evidence of the operational pressure created by data-quality problems.

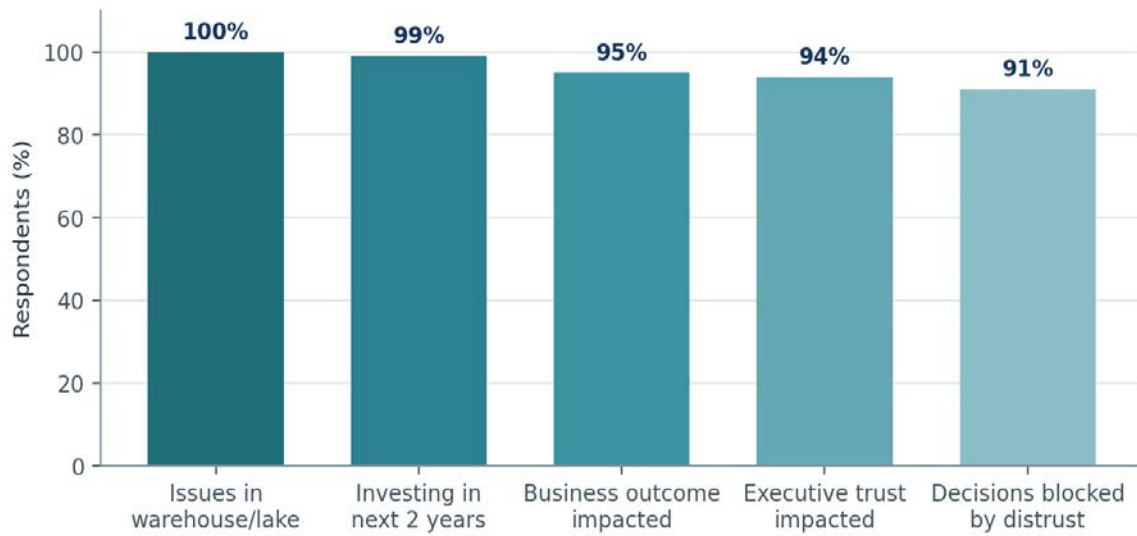


Figure 1. Public survey statistics on the impact of enterprise data quality problems on business (unit: %).

Data source: publicly released Anomalo State of Enterprise Data Quality 2024 survey data (industry survey; used only for descriptive statistics); more than 300 data managers and quality-related professionals were surveyed.

Chart analysis: As shown in Figure 1, all five indicators exceed 90%, indicating that data-quality problems are not confined to individual systems but can become enterprise-wide operational risks. The reported figures also associate data-quality problems with business impact, management trust, and reduced willingness to rely on data-driven decisions; therefore, governance should address technical detection, business impact, and trust restoration together.

Recent review work further emphasizes that data quality is a socio-organizational challenge requiring governance and continuous assurance across the full data lifecycle, especially as artificial-intelligence use expands.[10] The second industry survey below provides a complementary descriptive view of changing governance and data-integrity pressure.

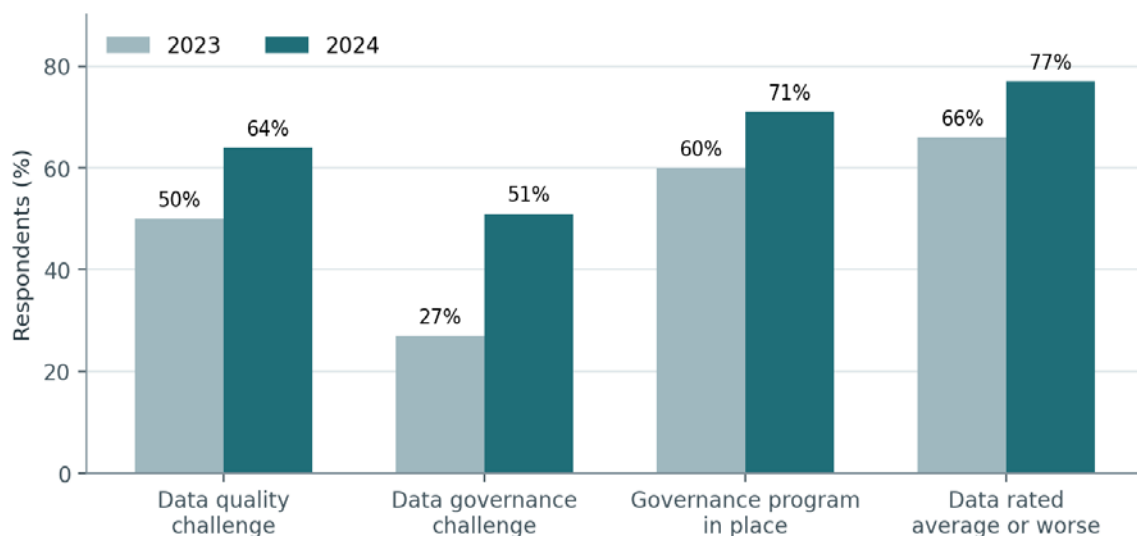


Figure 2. Changes in Key Data Integrity Indicators from 2023 to 2024 (unit: %).

Data source: publicly released Precisely and Drexel University 2025 Outlook: Data Integrity Trends and Insights survey data (industry survey; used only for descriptive statistics); sample size exceeds 550 data and analytics professionals.

Chart Analysis: As shown in Figure 2, the reported proportion of organisations viewing data quality as a major challenge increased from 50% in 2023 to 64% in 2024, while the data-governance challenge increased from 27% to 51%. Although governance-program coverage also increased, a large share of respondents continued to rate their data as average or worse. This contrast indicates that expanding policy coverage alone does not guarantee better quality; governance must be translated into executable platform rules, accountability, and verifiable outcomes.

3 Problem Statement

3.1 Standard objects are scattered, and the integrity criteria cannot be reused.

Large enterprises generally use the same data standards for policy documents, Excel ledgers and other system configurations. Fields with the same name may have different meanings, and the same code may have multiple value domains. Without a common identifier, version, effective date and scope of application, the platform cannot automatically determine which data product a rule should be applied to. Completeness should first be shown by the absence of null values in required fields:

$$C_d = \frac{N_{total} - N_{null}}{N_{total}} \quad (1)$$

C_d is the integrity rate of data object d , N_{total} is the total number of records in the inspection, and N_{null} is the number of records with missing key fields. This indicator should be related to the importance of the field and the business scenario, and the average value of the entire table should not be used to hide missing key attributes.

3.2 The rules are redundantly constructed, and there is a lack of reliable references for accuracy.

Business units, data development and audit teams often independently write the same rules without sharing reference data, sample evidence and threshold justifications. The number of rules continues to rise, but they have not addressed the problems of false positives, false negatives and business losses. For data with authoritative references, accuracy can be calculated by comparing with verified correct records.

$$A_d = \frac{N_{correct}}{N_{verified}} \quad (2)$$

A_d is the accuracy of data object d ; $N_{correct}$ is the number of records that are consistent with authoritative sources or human verification results; and $N_{verified}$ is the number of records that have been verified. The accuracy rate should also record the reference source, sampling method and verification date to avoid treating unknown records as errors directly.

Table Analysis: As shown in Table 1, the diagnosis matrix connects surface anomalies with platform symptoms, business consequences, and governance control points. Semantic drift and schema drift require vocabularies, lineage, and data contracts; rule duplication, ambiguous responsibility, and evidence gaps correspond to rule registration, RACI, and evidence stores. The matrix indicates that enterprise data quality cannot be solved by a single cleaning tool; technical metadata, business metadata, and responsibility metadata must be linked so that anomalies can be traced to rules, data products, and accountable parties.

Table 1. Enterprise Data Platform Quality Problem Diagnosis Matrix.

Issue	Platform Symptom	Business Consequence	Control Focus
Semantic drift	Same field, different meanings	Conflicting KPI results	Glossary and lineage
Schema drift	Unannounced type or column change	Pipeline failure	Data contract
Rule duplication	Repeated checks across tools	High maintenance cost	Rule registry
Weak ownership	Unassigned incidents	Long resolution cycle	RACI and SLA
Evidence gap	No samples or root cause record	Low auditability	Quality evidence store

3.3 The detection and rectification process is disconnected, making it difficult to prove governance performance.

Many platforms can show the null value rate or the number of anomalies, but they fail to convert alarms into work orders with an impact scope, responsible party, time limit and review record. Without lineage propagation analysis, a small upstream change can lead to a cascading failure of multiple reports and models; without a baseline and versioning, it cannot be determined whether this is due to actual quality improvement or a relaxed threshold, etc. Therefore, the root of the problem is that standards, indicators, implementation and responsibility have not formed an integrated control loop.

4 Problem-solving mechanisms and strategies

4.1 Establish a "Four-Tier Standard Object + Unified Identifier" System

The first level is the general standard for the entire company, and it contains general terms, master data codes, etc. The second level is the standard for the subject area, such as customers, products and finance. The third level is a set of data product contracts for consumer scenarios that specify the schema, refresh frequency, quality threshold and compatibility strategy. The fourth layer is for rule instances that the platform can execute. Each object has an ID, owner, version, status, effective time, upstream and downstream links, etc. Standard changes need impact analysis, review, canary release and decommissioning.

Table 2. Standardised Quality Objects and Platform Implementation Elements.

Object	Required Metadata	Platform Service	Output
Business term	Definition, owner, domain, status	Business glossary	Shared semantics
Data element	Type, format, code set, sensitivity	Metadata catalog	Technical standard
Data contract	Schema, SLA, threshold, compatibility	Contract registry	Producer-consumer agreement
Quality rule	Metric, logic, scope, severity, version	Rule engine	Executable control
Quality incident	Impact, lineage, assignee, deadline	Workflow service	Closed-loop evidence

Table Analysis: Table 2 links business semantics and platform execution through five categories of objects. Business terms and data elements define "what the data is"; data contracts specify "what producers promise consumers"; quality rules determine "what the platform verifies"; and quality event records anomalies in "how they are handled". Once these objects are connected by a common identifier, standards are no longer just documents but can be retrieved from directories, invoked by

rule engines, propagated through lineage services, and audited by workflows, thus forming a continuously evolving standard asset.

4.2 Develop reusable indicators, rules, and baselines.

Build an all-encompassing, high-quality, accurate, timely, valid and unique database of platform metrics, and enable business domains to set thresholds via a single definition. Attribute conflicts of the same business key in different sources can be quantified as the consistency rate.

$$S_d = 1 - \frac{N_{\text{conflict}}}{N_{\text{compare}}} \quad (3)$$

S_d is the cross-source consistency rate of data object d ; N_{conflict} is the number of records where the same business key conflicts between the authoritative source and the target source; and N_{compare} is the total number of comparable records. Standardise the time range, code mapping and primary key rules before the calculation.

Rule execution uses a "template-instance" model; templates define general logic, such as null values, ranges, regular expressions, referential integrity, cross-table balancing and distribution drift, and instances bind data objects, scheduling cycles, thresholds and severity. The platform saves the samples, baselines, rule versions and computation environments of each run to ensure reproducibility of the metrics. Sliding window monitoring is used for high-frequency streaming data, and batch verification reduces the computation of stable master data.

4.3 Establish a quality scoring and grading service agreement

Based on the business objectives, combine the individual indices and avoid a high mean that conceals serious defects. The total quality score of the data object d is as follows:

$$Q_d = \sum_{k=1}^m w_k q_{dk}, \quad \sum_{k=1}^m w_k = 1 \quad (4)$$

Q_d is the overall quality score in the formula, q_{dk} is the k -th standardised quality indicator, w_k is the corresponding weight, m is the number of indicators, and the sum of the weights is 1. The weight should be according to the level of the data product and consumption risk, and a "veto" threshold can be set for key regulatory fields.

Therefore, the data products are classified into three tiers: essential, necessary and general. Critical data products need real-time or hourly monitoring, have strict change compatibility and a short repair cycle; general products can be processed in daily batches. In addition to the target score, the service agreement should also specify the maximum data latency, alarm confirmation time, repair time and allowable exemption conditions, and display the latest trend instead of a single instantaneous score on the directory page.

4.4 Closed-loop disposal driven by kinship and risk priorities

When a rule fails, the platform identifies the upstream source table, processing task, report and model based on field-level lineage, automatically merges alarms with the same root cause, and generates a work order. To address the problem of too many high-risk events, a risk priority for event j can be set as follows:

$$R_j = \frac{(1-Q_j)I_jF_j}{1+T_j} \quad (5)$$

R_j is the risk priority of event j ; Q_j is the quality score of the relevant data object; I_j is the

business impact coefficient; F_j is the impact frequency; and T_j is the tolerable time in the formula. A high score indicates a more severe problem, and this will be dealt with first. Scale all the coefficients according to the risk policy of the company.

The seven links in the loop are: discovery, verification, location, allocation, repair, validation, and examination. Producers are responsible for the source repair; Data administrators verify adherence to standards; the platform team ensures the normal operation of rules and lineage; consumers confirm the business results. Recurring events will automatically trigger root cause analysis and rule optimisation; serious events that exceed the service time limit will be reported to the governance committee. Closing a work order needs repair records, a before-and-after comparison, a list of affected parties, and the verifier's signature.

Table 3. Quality Closed-Loop Roles and Service Level Configurations.

Role	Primary Accountability	Key Evidence	Target SLA
Data owner	Approve standard and risk level	Approval record	2 business days
Data steward	Define rule and validate result	Rule version and sample	1 business day
Data producer	Fix source or pipeline defect	Change and test record	4–24 hours
Platform team	Operate engine, lineage and alerting	Run log and lineage path	99.9% availability
Data consumer	Confirm fitness for use	Acceptance record	1 business day

Analysis of Tables: Table 3 divides the general "data department responsible" into owners, administrators, producers, platform teams and consumers of governance responsibility. Each role has an auditable evidence and a target timeframe: Owners approve standards and risks; Administrators manage rules; Producers correct root causes; The platform team ensures normal operation; Consumers confirm applicability. Thus, the platform team will not be responsible for the meaning of the data on behalf of the business, and the business department will not shift technical failures to the users; thus, a stable operating organization for closed-loop governance has been established.

4.5 Phased Implementation and Effect Verification

The three steps of the implementation are as follows: First, select high-value subject areas, such as customers, products and finance; Second, review standard objects; Third, set a quality baseline. Phase 2 will introduce contract registration, a rule engine, lineage alerts and a work order interface to build an all-encompassing closed-loop system. Phase 3 will add more data products and introduce anomaly detection and rule recommendations. The performance index should include the proportion of complete essential fields, cross-source consistency ratio, mean discovery duration, mean repair time, recurrence frequency and consumer trust before and after optimization. If the quality score rises but the repair cycle and business complaints do not fall, then the weights, thresholds or sampling methods may have been altered to gamify the metrics.

5 Conclusion

The issue of enterprise-level data quality management is not a lack of rules; rather, there are inconsistent standards, an inability to reproduce quality evidence, and a disconnect between platform alerts and accountability. This paper focuses on lakewarehouse integrated data platforms to build a closed-loop mechanism of "layered standards—data contracts—unified metrics—lineage positioning—risk work orders—retrospective optimisation" and uses five editable formulas to

clarify the calculation logic for completeness, accuracy, consistency, overall quality and event priority. According to the above statistics, even after the expansion of governance projects, enterprises have been subjected to higher data-quality requirements; thus, the results of governance should be reflected in both operating indicators and business performance. In the future, multiple real-world enterprises will be selected as research subjects to conduct longitudinal studies on how different weights and thresholds, as well as organisational maturity, affect remediation efficiency, data product reusability, and the reliability of artificial intelligence applications.

REFERENCES

- [1] Fu Q, Nicholson GL, Easton JM. *Understanding data quality in a data-driven industry context: Insights from the fundamentals*. *Journal of Industrial Information Integration*, 2024, 42: 100729. DOI: 10.1016/j.jii.2024.100729.
- [2] Bernardo BMV, São Mamede H, Barroso JMP, dos Santos VMPD. *Data governance & quality management—Innovation and breakthroughs across different fields*. *Journal of Innovation & Knowledge*, 2024, 9(4): 100598. DOI: 10.1016/j.jik.2024.100598.
- [3] Miller R, Whelan H, Chrubasik M, Whittaker D, Duncan P, Gregório J. *A Framework for Current and New Data Quality Dimensions: An Overview*. *Data*, 2024, 9(12): 151. DOI: 10.3390/data9120151.
- [4] Goel K, Martin N, ter Hofstede AHM. *Demystifying data governance for process mining: Insights from a Delphi study*. *Information & Management*, 2024, 61(5): 103973. DOI: 10.1016/j.im.2024.103973.
- [5] Blohm I, Wortmann F, Legner C, Köbler F. *Data products, data mesh, and data fabric*. *Business & Information Systems Engineering*, 2024, 66: 643-652. DOI: 10.1007/s12599-024-00876-5.
- [6] Bode J, Kühl N, Kreuzberger D, Holtmann C. *Toward Avoiding the Data Mess: Industry Insights From Data Mesh Implementations*. *IEEE Access*, 2024, 12: 95402-95416. DOI: 10.1109/ACCESS.2024.3417291.
- [7] Polimeno A, Braghin C, Anisetti M, Ardagna CA. *Maximizing data quality while ensuring data protection in service-based data pipelines*. *Journal of Big Data*, 2025, 12: 62. DOI: 10.1186/s40537-025-01118-5.
- [8] Nguyen T, Nguyen HT, Nguyen-Hoang TA. *Data quality management in big data: Strategies, tools, and educational implications*. *Journal of Parallel and Distributed Computing*, 2025, 200: 105067.
- [9] Yalaoui M, Boukhedouma S. *A structured machine learning based framework for multidimensional data quality assessment*. *Journal of Big Data*, 2026. DOI: 10.1186/s40537-026-01520-7.
- [10] Zhang, Z. (2026). *Research on Service-oriented Packaging of AI Systems and Data-driven Decision Support for Real-world Business Processes*. *Engineering Advances*, 6(3).
- [11] Wang, B. (2025). *Methods of Load Optimization for Computer Systems Based on Physical Principles*.
- [12] Wang, Z. (2026). *Research on Operation Data Analysis and Efficiency Improvement of the New Energy Industry Based on Machine Learning*. *Procedia Computer Science*, 285, 605-612.
- [13] Han, W. (2025, November). *An Empirical Study on Building a High-precision Forecasting System for Financial Market Trends Based on the PyTorch Deep Learning Framework*. In *International Conference on Computational Technologies for Research in Data Analytics* (pp. 518-528). Cham: Springer Nature Switzerland.

- [14] Wu, L. (2026). *Construction and Evolutionary Analysis of a Game Model for Supply Chain Finance Funding Based on Blockchain Technology*. *Procedia Computer Science*, 282, 2004-2012.
- [15] Ding, J. (2025, December). *Research on Logistics Cost Control and Optimization in Automotive Manufacturing Supply Chain Based on DMAIC Model*. In *2025 IEEE 1st International Conference on Recent Trends in Computing and Smart Mobility (RCSM)* (pp. 1-7). IEEE.
- [16] Yu, X. (2026). *Exploration of Multi-Channel Conversion Path Optimization Methods Based on A/B Testing*.
- [17] Xiaoyu Gu. (2025) *Research on Generative AI Psychological Intervention Models Based on Multimodal Emotion Recognition*. *Advances in Computer and Communication*, 6(5), 304-309.