JURNAL RESPONSIVE
Teknik Informatika

Desember 2025 | Vol.9 | No.02
E-ISSN : 2614-7602
DOI: 10.36352/jr.v3i2

DATA QUALITY MANAGEMENT IN DATA WAREHOUSES: A SYSTEMATIC LITERATURE REVIEW

Romiko Afriantoni*[1], Naisya Nindy Pangestuti[2]
[1,2]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi – Universitas Ibnu Sina
e-mail: *[1]romiko@uis.ac.id,

Abstract

This study presents a systematic literature review of Data Quality Management (DQM) in data warehouse environments, aiming to map key dimensions, processes, and architectural/technological enablers, and to identify research gaps. Searches were conducted across Scopus, ScienceDirect, IEEE Xplore, ACM Digital Library, SpringerLink, and Google Scholar (as a complement) for the period 2009–2025, following PRISMA 2020. Of 200 initial records, 133 were excluded during the first screening, 67 underwent further assessment, and 6 studies met the inclusion criteria for in-depth analysis. Thematic synthesis indicates that effective DQM rests on four integrated pillars: (1) standardized quality dimensions and metrics (accuracy, completeness, consistency, timeliness, and traceability), (2) prevention–detection–correction processes embedded along the ETL/ELT pipeline (including consistent SCD policies and handling of late-arriving data), (3) architectural/technological support (automated data tests within CI/CD, catalogs/metadata, data lineage, observability, and data contracts), and (4) governance that clarifies roles and accountability (data owners/stewards) with incident-response procedures. Practically, organizations should start from critical data elements and high-priority consumption paths, translating SLA/SLI into executable rules. Limitations include the small number of included studies and contextual heterogeneity, motivating further work on cross-domain metric standardization, open DQM benchmarks, cost–benefit evaluations of observability/contract enforcement, and the impact of data quality on analytic/AI performance in near real-time settings..

Keywords— Data quality management; data warehouse; ETL/ELT; data governance; data lineage; observability; data contracts; PRISMA; systematic literature review.

INTRODUCTION

The explosion in data volume, variety, and velocity makes Data Warehouses (DWs) the backbone of organizational analytics due to their ability to integrate diverse sources into a structured repository for decision-making. However, low data quality (DQ)—for example, inaccurate, incomplete, inconsistent, or untimely—will propagate throughout the ETL process to dashboards and analytical models, reducing the reliability of insights and decisions. This situation underscores the urgency of systematic Data Quality Management (DQM) in the DW context (Ehrlinger & Wöß, 2022; Liu et al., 2020).

DQM in DWs is challenging due to heterogeneous integration, pipeline complexity, and dynamic changes in schemas/business rules. Recent findings indicate that dozens of factors influence data pipeline quality—from data aspects, infrastructure, lifecycle, development & deployment, to processing—and root causes often emerge during the cleansing and integration stages. This demands a proactive and continuous DQM approach throughout the data lifecycle in DWs (not just a final inspection). (Foidl et al., 2024; Taleb et al., 2021).

At the organizational level, data governance plays a central role as an umbrella of policies, roles, and accountability mechanisms to ensure the quality of data shared and consumed across units. Recent systematic reviews confirm that good governance maturity correlates with improved data quality and value, and encourages standardization of DQM processes. Furthermore, case-based studies propose a design theory for DQ tools in data ecosystems to prescriptively enforce quality checks on shared data (Bernardo et al., 2024; Altendeitering et al., 2024).

On the methodological side, recent literature points to the direction of continuous DQM (continuous quality management) through quality profiles, rule verification, and continuous monitoring, including in the context of big data and AI. Similarly, some disciplines are developing domain-specific frameworks (e.g., for medical training data) that extend traditional quality dimensions to address the reliability of machine learning-based applications. These findings indicate the need for a conceptual synthesis linking generic DQM principles to the needs of modern DW (Taleb et al., 2021; Schwabe et al., 2024).

A significant gap that remains is the fragmentation of quality terms and dimensions across domains and the limited standardization of measurements that can be directly operationalized in DW. Recent studies highlight the need for coherence in DQ terminology and dimensions to enable organizations to design consistent and measurable quality metrics, rules, and controls across DW pipelines (Miller, 2024; Ehrlinger & Wöß, 2022).

RESEARCH METHODS

This study uses a systematic literature review (SLR) with a qualitative-descriptive approach to evaluate and synthesize studies on Data Quality Management (DQM) in data warehouses. The reporting protocol follows PRISMA 2020 to ensure transparency and traceability of the search, filtering, and reporting processes.

Literature Sources and Criteria

The search was conducted in reputable databases: Scopus, Web of Science (where available), ScienceDirect, IEEE Xplore, ACM Digital Library, SpringerLink, and Google Scholar (as a supplementary source). The timeframe 2009–2025 was chosen to capture the evolution of DQM practices in DW from the 2010s to the latest developments. Inclusion criteria:

(a) published primary documents (reputable journal articles and conference proceedings),
(b) written in Indonesian or English,
(c) focused on DQM (quality dimensions/matrices, process control, governance, testing/monitoring, lineage, metadata, data contracts, observability) in the context of connected data warehouses/ETL/ELT/lakehouses,
(d) presenting replicable/adoptable methods, frameworks, tools, or evaluations.

Exclusion criteria: non-scholarly articles, duplications, studies addressing data quality outside the DW context (e.g., sensors/IoT with no relation to DW), or not providing sufficient methodological details. Multi-database selection was performed to mitigate coverage bias and refer to comparative evidence of search system performance for SLR.

Search Strategy and Reference Management

Queries are derived from three sets of concepts: (i) data quality domain, (ii) DW/pipeline architecture, (iii) DQM practices/processes. Example strings (customized per database):

- (“data quality” OR “data-quality”) AND (“data warehouse” OR “data warehousing” OR ETL OR ELT OR lakehouse)
- (“data governance” OR “metadata management” OR “data lineage” OR “data contract*” OR “observability*”) AND (“warehouse” OR “ETL” OR “BI”)

Boolean operators, exact phrases, and year filters are used according to each database's policies. Publish or Perish (PoP) is used to aggregate Google Scholar results for export to CSV and consolidation in Microsoft Excel for metadata deduplication and normalization. Reporting strategies and search syntax follow PRISMA-S recommendations.

Prosedur Seleksi & Diagram PRISMA

The selection process involved two stages: (1) title–abstract screening, and (2) full-text review. Each record was checked against inclusion/exclusion criteria. Disagreements were resolved through criteria-based discussions. An initial target of ±100 articles was set to ensure breadth of mapping; the final number was determined based on eligibility after evaluation. The results of each stage (number of records found, screened, excluded with reasons, and included) are reported in the PRISMA Flowchart.

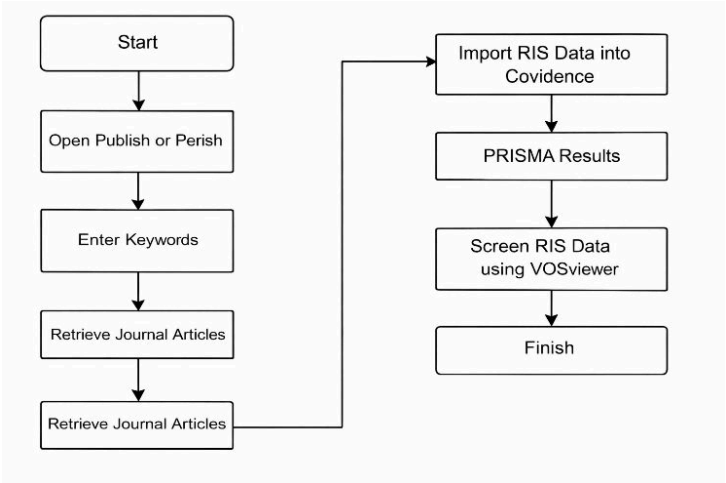

Figure 1. Flowchart

SLR process flow diagram for Data Quality Management topic in Data Warehouse. The flow starts from Start → opening Publish or Perish → entering keywords related to DQM & data warehouse → obtaining a list of journal articles. The list is exported (RIS/CSV format) and then imported into Covidence (or similar tool) for reference management and appropriateness assessment. The selection results are summarized in a PRISMA diagram. Next, screening is carried out using VOSviewer (co-occurrence mapping/citation network) to support relevance assessment and identify topic clusters. The final stage is Finish, which is a collection of studies that pass the selection for further analysis.

RESULTS AND DISCUSSION

Results

The research "Data Quality Management in Data Warehouses: A Literature Review" was conducted through three main stages following a Systematic Literature Review (SLR) approach. First, the study plan was formulated, focusing on DQM in the context of data warehouses—including the research question statement, search strategy, and inclusion-exclusion criteria related to quality dimensions (accuracy, completeness, consistency, timeliness) and DW/ETL-ELT practices. Second, the literature search and selection process included retrieving records from reputable databases, deduplication, title-abstract screening, full-text review, and core data extraction (quality control mechanisms, architectural support such as lineages, metadata catalogs, and data contracts). Third, the results were synthesized and reported, where findings were mapped thematically to demonstrate trends, contributions, and gaps in DQM research in data warehouses. The entire process was reported systematically, transparently, and replicabably.

1. Planning Phase

The planning phase begins by defining the focus of the study on Data Quality Management (DQM) in the context of a data warehouse, encompassing quality dimensions (accuracy, completeness, consistency, timeliness, traceability) and supporting practices and architectures (ETL/ELT, lineage, metadata catalog, data contracts, and observability). From this focus, research questions are derived that guide the entire process: (i) what is the landscape of DQM concepts, dimensions, and metrics relevant to data warehouses; (ii) how is the prevention–detection–correction approach operationalized throughout the data warehouse lifecycle; (iii) what technologies/architectures effectively support DQM; and (iv) what governance factors influence implementation success.

To ensure transparency and repeatability, an SLR protocol was developed detailing the timeframe (2009–2025), publication type (reputable journal articles and conference proceedings), language (Indonesian/English), and domain boundaries (studies explicitly addressing DQM in data warehouses, including relevant lakehouse integrations). The protocol also includes a cross-database search strategy (Scopus, ScienceDirect, IEEE Xplore, ACM Digital Library, SpringerLink, and the Google Scholar supplement), a plan for using Publish or Perish to retrieve Scholar results, and the definition of keywords and equivalents: “data quality,” “data-quality,” “data warehouse/warehousing,” ETL, ELT, lakehouse, “data governance,” “metadata management,” “data lineage,” “data contract*,” “observability*,” and “SLA/SLI.” Variations of Boolean operators and exact phrases are set up per database to maximize coverage and precision.

Next, inclusion-exclusion criteria and methodological quality indicators to be used during full study screening and assessment (clarity of objectives, transparency of methods, adequacy of data/evaluation, and reporting of limitations) were established. A data extraction form was also developed that captured publication identity, DW context (on-prem/cloud/hybrid), DQ dimensions addressed, pipeline stages discussed (acquisition, transformation, storage, ingestion), control mechanisms (validation rules, automated data testing, SLA/SLI, data contracts, lineage/metadata), tools/architecture, evaluation metrics, and key findings. To support reference management and pipeline documentation, the use of Microsoft Excel (metadata deduplication/normalization) and VOSviewer (co-occurrence mapping and citation networks) was planned. All these planning decisions were linked to the PRISMA 2020 reporting plan to ensure that every step—from identification to synthesis—is recorded systematically, transparently, and replicable.

2. Implementation Phase

The implementation phase implements the SLR protocol developed for Data Quality Management (DQM) in the data warehouse. The process begins with a cross-database literature search (Scopus, ScienceDirect, IEEE Xplore, ACM Digital Library, SpringerLink, and Google Scholar via Publish or Perish). The search string is customized per database using a combination of data quality domain terms and DW/pipeline architecture, for example: "data quality" AND ("data warehouse" OR ETL OR ELT OR lakehouse) AND (governance OR lineage OR "metadata management" OR "data contract" OR observability* OR SLA OR SLI)*, and a publication year filter of 2009–2025. All queries, search dates, and the number of initial hits are recorded as part of the audit trail.

Search results from each source are exported in RIS/CSV format and then consolidated in Microsoft Excel for metadata normalization (title, author, year, venue, DOI) and deduplication. Deduplication is performed in layers: DOI match; If unavailable, a combination of title-year-author; and manual verification for ambiguities. Each track record is assigned a unique ID and process status (discovered, duplicated, filtered, included/excluded) to facilitate PRISMA reporting.

The screening phase is two-stage. First, the title-abstract is screened against inclusion-exclusion criteria: an explicit focus on DQM (dimensions/metrics, control, monitoring) in the context of data warehouses/ETL-ELT/lakehouses, published scientific documents, and

Indonesian/English language. Articles that refer to "data quality" but are not related to DW (e.g., pure IoT without DW integration) are flagged for exclusion. Second, a full-text review is conducted to assess thematic fit and methodological feasibility; reasons for exclusion (e.g., outside the DW context, inadequate method description, duplication) are documented in detail. If there is more than one reviewer, disagreements are resolved through protocol-based discussions; inter-reviewer agreement can be quantified (e.g., using κ) for reporting purposes, where relevant. Articles that pass the quality assessment stage are assessed using a checklist that includes clarity of purpose, transparency of methods, adequacy of data/evaluation, reporting of limitations, and direct relevance to DQM–DW. The assessment results are not solely for eliminating studies, but rather for weighting them during synthesis.

Data extraction is then performed based on a predetermined form: study identity; DW context (industry/academic; on-prem/cloud/hybrid); quality dimensions addressed (accuracy, completeness, consistency, timeliness, traceability/lineage); pipeline stages (acquisition, transformation, storage, consumption); control mechanisms (validation rules, automated data testing in CI/CD, quality SLAs/SLIs, data contracts, catalog/metadata, lineage/observability); tools/architectures used; evaluation metrics; and key findings and limitations of the study.
To support this, a bibliometric analysis using VOSviewer is performed to map keyword co-occurrence and citation networks. Parameters (e.g., minimum occurrences, synonymous fusion—lineage/provenance, observability/monitoring) are documented. The mapping results were used to confirm thematic clusters—for example, DQ dimension clusters, data control & testing, governance & metadata/lineage, and DW/ETL–ELT/lakehouse architecture—and to track topic evolution.

The final implementation phase was the initial PRISMA synthesis and reporting. The number of records in each flow box (discovered, after deduplication, filtered, fully reviewed, excluded with reasons, and included) was entered into the PRISMA diagram. All decisions—from queries and exclusion lists with reasons, to extraction forms—were archived to ensure a systematic, transparent, and replicable process consistent with the study's objectives.

3. Reporting Stage

The reporting stage is the final part of the systematic literature review process, which aims to compile and convey research findings in a coherent, objective, and scientific manner. At this stage, all analytical results obtained from selected articles are presented in narrative form, tables, or data visualizations to facilitate interpretation within the context of Data Quality Management (DQM) in a data warehouse.

Researchers compile the report by grouping key findings based on specific themes, such as data quality dimensions (accuracy, completeness, consistency, timeliness, traceability/lineage), control and monitoring mechanisms (validation rules, automated data testing, quality SLAs/SLIs, data contracts), and architectural support (ETL/ELT, metadata catalog, lineage, observability) and factors that support/inhibit DQM implementation in a data warehouse. Each result is critically analyzed to identify trends, research gaps, and potential future research directions in the DQM–DW domain.

Furthermore, the researchers outlined the study's limitations and provided relevant recommendations for academics and practitioners—for example, strengthening governance and the role of data stewards, standardizing the definition of quality dimensions, and integrating quality control as code in the pipeline. The entire reporting process and results were compiled with the principles of transparency and scientific accountability in mind, ensuring they can serve as valid references for further research and application within organizational data warehouse environments.

The next stage involved using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework to assess the suitability of articles to the DQM research topic in data warehouses. This process involved screening based on inclusion and exclusion criteria to ensure that only articles or journals truly relevant to the study's focus were considered. The results of this selection phase narrowed down to six journals deemed appropriate for the

research title and worthy of further in-depth analysis. This procedure aimed to ensure that the data used had strong validity and supported the achievement of the overall research objectives.

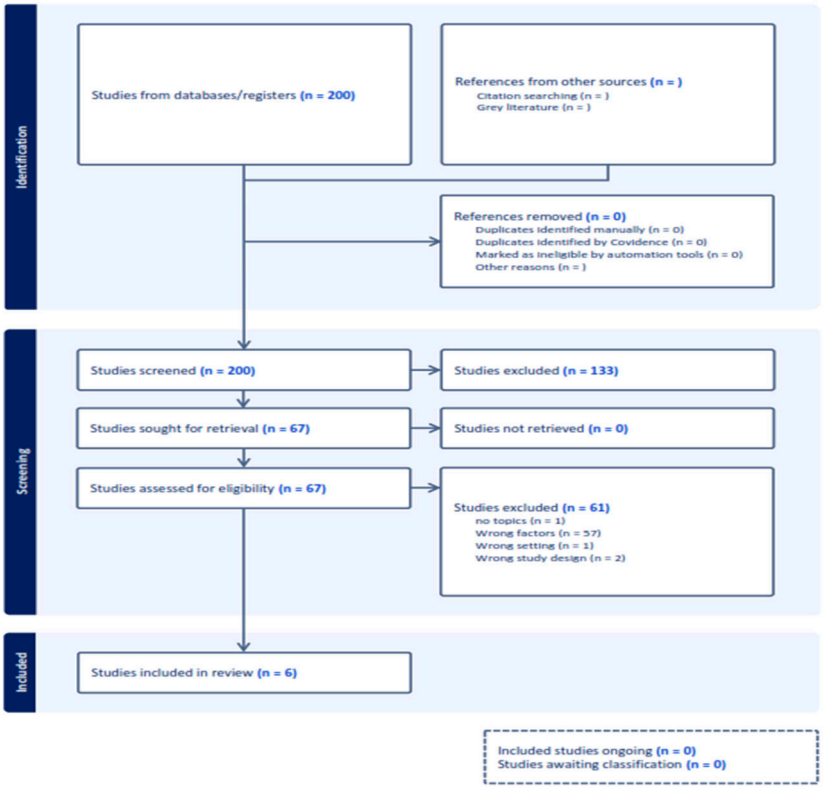

Figure 2. Covidence Prism

The PRISMA diagram displays the study screening process for a systematic review of the role of social media in supporting digital business sustainability. Of the 200 papers collected through various databases and additional sources, 133 were excluded during the initial screening stage due to not meeting relevance criteria. Sixty-seven articles were then reviewed further, but 61 were excluded from the final selection—due, among other things, to inappropriate topic focus, irrelevant variables, or inadequate research design. Ultimately, six studies met all inclusion criteria and served as the basis for the study synthesis. This flowchart demonstrates a systematic and transparent literature identification process, strengthening the traceability and validity of the research findings.

Table 1. Systematic Literature Review (SLR) References

No	Referensi	Goal/ Focus	Method/ Design	Context (DW/ETL/ Lakehouse)	DQ Dimensions/Aspects	Key Findings	Implications for DQM–DW
1	Ehrlinger, L., & Wöß, W. (2022). A survey of data quality measurement and monitoring tools. Frontiers in Big	Mapping of measuring instruments & monitor	Survey/comparative tools	DW & pipeline (measurement/monitoring)	Measurement, monitoring, completeness,	DQ tool features and limitation	A guide to selecting a DQ toolchain for DW and

	Data, 5, 850611. https://doi.org/10.3389/fdata.2022.850611	ing of data quality			timelines s	s are mapped; automation support varies	integration into CI/CD
2	Foidl, H., Golendukhina, V., Ramler, R., & Felderer, M. (2024). Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. Journal of Systems and Software, 207, 111855. https://doi.org/10.1016/j.jss.2023.111855	Factors affecting data pipeline quality and root causes	Empirical/ analytical studies	ETL/ELT (pipeline DW)	Pipeline quality, root cause, process	Identifying problematic factors and areas (cleaning, integration)	DQ control assignment at critical points of DW pipeline
3	Taleb, I., Serhani, M. A., Bouhaddioui, C., & Dssouli, R. (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	Holistic DQ framework & sustainable management	Framework/conceptual development	Big data ↔ can be adopted DW/lakehouse	Continuous DQ, quality profile, rule verification	Proposing a continuous DQ cycle across data stages	The basis for implementing continuous DQ monitoring in DW/lakehouse
4	Liu, Q., Feng, G., Tayi, G. K., & Tian, J. (2020). Minimizing the data quality problem of information systems: A process-based method. Decision Support Systems, 137,	Process approach to minimize DQ problems	Process method/operational framework	IS/BI ↔ relevant to DW/ETL	Prevention, detection, correction	DQ is more effective when embedded in the	"Quality by design" in ETL/ELT and DW governance

	113381. https://doi.org/10.1016/j.dss.2020.113381					proc ess, not final inspe ction .	
5	Altendeitering, M., Guggenberger, T. M., & Möller, F. (2024). A design theory for data quality tools in data ecosystems: Findings from three industry cases. Data & Knowledge Engineering, 153, 102333. https://doi.org/10.1016/j.datak.2024.102333	Design theory for DQ tools in data ecosyst ems	Multi-case/theor etical design	Data ecosystem ↔ integration to DW	Governa nce, tool capabilit y, prescripti ve rules	DQ tool desig n princ iples that can be enfor ced acros s team s	Criteria for selecting /designi ng DQ tools compati ble with DW
6	Miller, R. (2024). A framework for current and new data quality dimensions: An overview. Data, 9(12), 151. https://doi.org/10.3390/data9120151	Data quality dimensi on update/ alignm ent	Conceptu al/synthes is	General ↔ standardizat ion for DW	DQ dimensio ns (accuracy , complete ness, consisten cy, etc.)	Prop osed more unifo rm dime nsion s and defin ition s	Standard ization of DQ DW metrics and quality scorecar ds

Source, processed data

Discussion

The results of a systematic review of Data Quality Management (DQM) in the context of data warehousing reveal a consistent picture: data quality is not simply a function of a tool, but rather the result of a close interaction between the definition of quality dimensions, operational processes throughout the DW pipeline, architectural/technology support, and organizational governance. The six selected studies demonstrate that the effectiveness of DQM depends on how these four elements are designed as a unified, auditable and automated whole, rather than handled as downstream inspection activities.

First, in terms of dimensions and metrics, the majority of articles emphasize the importance of aligning the definitions of accuracy, completeness, consistency, and timeliness with DW use cases (reporting, descriptive analytics, or advanced). A recurring finding is the persistence of variation in terminology and indicators across organizations, which makes it difficult to conduct cross-study comparisons or establish meaningful quality SLAs/SLIs. The synthesis highlights the need for a more uniform conceptual dictionary and a clear distinction between outcome metrics (e.g., error rates on facts/dimensions) and process metrics (e.g., rule pass rates at each ETL/ELT stage).

Second, in the end-to-end DQM process, the most consistently effective practice is the prevention-detection-correction pattern embedded upstream. In data sources and staging areas, prevention takes the form of schema validation, domain constraints, and the establishment of critical data elements. In transformation, detection is realized through data tests (e.g., referential integrity testing, row count reconciliation, range checks, and business rules) that run automatically during each build/deployment pipeline. Corrections are then operationalized as integrated jobs (deduplication, record linkage, and safe type casting) with audit logs. For slowly changing dimensions (SCD), the study emphasized the importance of consistent historiography policies (e.g., Type 2) to prevent historical accuracy from being sacrificed for current value.

Third, in terms of architecture and technology, there has been a shift to cloud-based ELT approaches for data warehouses/lakehouses, along with increased adoption of metadata/lineage and data catalogs. Automated data testing combined with CI/CD and data observability (monitoring for volume, schema, and value anomalies) emerged as key levers for reducing mean time to detect (MTTD) and mean time to recover (MTTR) of quality incidents. For near-real-time streams, the most prominent challenges are schema drift and late-arriving dimensions; a reported effective solution combines schema registry, contract testing, and graceful degradation at the consumption layer.

Fourth, in the governance and organization domain, the study positions data owners and data stewards as key actors bridging business needs and technical controls. Data contracts are used to establish accountability between data producers and consumers: quality rules, schemas, and SLAs are agreed upon upfront, tracked via lineage, and automatically enforced in the pipeline. Investments in training, data incident runbooks, and post-incident review mechanisms have been shown to reduce defect recurrence and improve documentation discipline.

Based on these cross-theme findings, this discussion develops an integrative framework for DQM-DW, encompassing four pillars. The Standardization Pillar defines the dimension vocabulary, the scope of critical data elements, and the basis for establishing SLAs/SLIs. The Process Pillar maps prevention–detection–correction controls at each pipeline node (source, staging, transformation, loading, and consumption). The Technology Pillar consolidates tests-as-code data, catalog/lineage, observability, and contract enforcement within the CI/CD cycle. The Governance & Roles Pillar establishes cross-team accountability, incident procedures, and change management for schemas and business rules. These four pillars reinforce each other: without standardization, SLAs are meaningless; without process, tools are ineffective; without technology, processes are unscalable; without governance, accountability is weakened.

Theoretically, this study enriches the literature by linking classic quality dimensions with modern practices such as observability and data contracts, and affirms DQM as a socio-technical discipline, not simply a tool. A notable gap is the limited availability of open benchmarks and replication packages that allow objective cross-context evaluation—this gap limits generalizability and slows metric convergence.

Practically, several immediate implications can be drawn. Organizations operating DW are advised to start by establishing decision-oriented quality SLAs (e.g., the timeliness of fact sales before the reporting cutoff), then scale them down to automated rule execution in the pipeline. Observability instruments should be installed at critical control points—schema changes, key fact-dimension joins, and key generators—and linked to threshold-based alerting. SCD and late arrival handling should have explicit policies that are periodically tested through regression tests based on historical data.

The discussion also highlights the cost-benefit trade-off. Full automation across all controls is not always optimal; studies show that the highest returns typically come from securing critical data elements and high-priority consumption paths (e.g., regulatory reports or key performance metrics). Therefore, an effective DQM roadmap moves from minimally viable DQ (CDE + rule prioritization + alerting) to broader scope as teams and infrastructure mature.

Meanwhile, significant open challenges remain: unifying semantic layers across domains to ensure consistent conformed dimensions; handling semantic drift when business definitions change; managing quality in hybrid batch-streaming environments; and measuring the impact of

quality on the accuracy of analytical/AI models consuming DW data. Another methodological gap is the rarely published reporting of negative outcomes (e.g., rules that fail to add value), which are crucial as a guidepost for future practice.

Relevantly, a realistic further research agenda includes: developing a unified metrics framework that combines outcome and process indicators; designing synthetic/annotated data-based DQM benchmarks to compare toolchains; quantitative cost-benefit studies of observability and contracts; and research on the relationship between data quality and AI model performance in near-real-time scenarios.

Finally, this discussion acknowledges the limitations of the review: the relatively small number of included studies (six), heterogeneity in context (industry, platform, and maturity level), and potential publication bias. However, transparent reporting through PRISMA and the use of explicit selection criteria strengthen the traceability and validity of the synthesis. Overall, the evidence gathered supports the conclusion that standardized, process-integrated, technology-driven, and governance-enforced DQM is a key prerequisite for data warehouses to maintain analytical reliability amidst the complexity of modern architectures.

CONCLUSION

This study confirms that the success of Data Quality Management (DQM) in data warehouses is supported by four integrated pillars: dimension-metric standardization, prevention-detection-correction processes throughout the pipeline, architectural/technology support (automated data testing, catalog/metadata, lineage, observability, and data contracts in CI/CD), and governance with clear roles and accountabilities. Practically, organizations need to start from critical data elements and prioritized consumption paths, reduce SLAs/SLIs to automatically executed rules, and enforce data contracts to accelerate incident detection/recovery. Theoretically, this study bridges classic quality dimensions with modern practices (observability and contracts-as-code), but gaps remain in metric standardization and the availability of open benchmarks. The limited number and heterogeneity of studies limit generalizability, so further research is recommended on a unified metrics framework, DQM benchmarks, automation cost-benefit evaluation, and the data quality-AI model performance relationship in near-real-time scenarios..

BIBLIOGRAPHY

Altendeitering, M., Guggenberger, T. M., & Möller, F. (2024). A design theory for data quality tools in data ecosystems: Findings from three industry cases. Data & Knowledge Engineering, 153, 102333. https://doi.org/10.1016/j.datak.2024.102333

Bernardo, B. M. V., Mamede, H. S., Barroso, J., & Santos, V. (2024). Data governance & quality management—Innovation and breakthroughs across different fields. Journal of Innovation and Knowledge, 9(4), 100598. https://doi.org/10.1016/j.jik.2024.100598

Ehrlinger, L., & Wöß, W. (2022). A survey of data quality measurement and monitoring tools. Frontiers in Big Data, 5, 850611. https://doi.org/10.3389/fdata.2022.850611

Foidl, H., Golendukhina, V., Ramler, R., & Felderer, M. (2024). Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. Journal of Systems and Software, 207, 111855. https://doi.org/10.1016/j.jss.2023.111855

Miller, R. (2024). A framework for current and new data quality dimensions: An overview. Data, 9(12), 151. https://doi.org/10.3390/data9120151

Taleb, I., Serhani, M. A., Bouhaddioui, C., & Dssouli, R. (2021). Big data quality framework: A holistic approach to continuous quality management. Journal of Big Data, 8, Article 76. https://doi.org/10.1186/s40537-021-00468-0

Liu, Q., Feng, G., Tayi, G. K., & Tian, J. (2020). Minimizing the data quality problem of information systems: A process-based method. Decision Support Systems, 137, 113381. https://doi.org/10.1016/j.dss.2020.113381

Schwabe, D., et al. (2024). The METRIC-framework for assessing data quality of medical training data. npj Digital Medicine, 7, 116. https://doi.org/10.1038/s41746-024-01196-4

Ehrlinger, L., & Wöß, W. (2022). A survey of data quality measurement and monitoring tools. Frontiers in Big Data, 5, 850611. https://doi.org/10.3389/fdata.2022.850611

Foidl, H., Golendukhina, V., Ramler, R., & Felderer, M. (2024). Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers. Journal of Systems and Software, 207, 111855. https://doi.org/10.1016/j.jss.2023.111855

Taleb, I., Serhani, M. A., Bouhaddioui, C., & Dssouli, R. (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

Liu, Q., Feng, G., Tayi, G. K., & Tian, J. (2020). Minimizing the data quality problem of information systems: A process-based method. Decision Support Systems, 137, 113381. https://doi.org/10.1016/j.dss.2020.113381

Altendeitering, M., Guggenberger, T. M., & Möller, F. (2024). A design theory for data quality tools in data ecosystems: Findings from three industry cases. Data & Knowledge Engineering, 153, 102333. https://doi.org/10.1016/j.datak.2024.102333

Miller, R. (2024). A framework for current and new data quality dimensions: An overview. Data, 9(12), 151. https://doi.org/10.3390/data9120151