Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Metadata-Driven Data Quality Frameworks for AI-Ready Healthcare Data Lakes: Design Patterns and Implementation on Google Cloud Platform

A Practitioner Framework for HIPAA-Compliant AI-Readiness Scoring in Regulated Healthcare Environments

Sulthanul Arif Ibrahim Liakath Ali
Data Engineering & AI Architecture Practice
Accenture LLP, Philadelphia, Pennsylvania, USA

Abstract - Healthcare data lakes in payer organizations accumulate vast volumes of clinical, administrative, and pharmaceutical data. However, volume alone does not guarantee fitness for artificial intelligence and machine learning (AI/ML) consumption. Traditional rule-based data quality frameworks are static, schema-bound, and insufficient for the dynamic, multi-dimensional quality requirements of AI training pipelines. This paper proposes a Metadata-Driven Data Quality Framework (MDDQF) for AI-ready healthcare data lakes, implemented on Google Cloud Platform (GCP). The framework defines six AI-readiness quality dimensions — Completeness, Consistency, Freshness, Lineage, Schema Stability, and Semantic Fidelity — and formalizes them into a composite AI-Readiness (AIR) Score (Equations 1-4). A three-layer architecture (Metadata Capture, Quality Scoring, Remediation and Routing) is implemented using GCP Dataplex, Dataform, BigQuery, Cloud Composer, and Cloud DLP. An anonymized HEDIS measure pipeline case study illustrates the framework applied to a realistic healthcare payer ETL topology. Simulated evaluation across five healthcare datasets projects mean AIR Score improvement from 0.578 to 0.860 following framework application, with a corresponding reduction in ML model feature rejection rate from 34% to 8%. All results are presented as exploratory simulated projections, not empirical production validation.

Keywords - data quality; AI-readiness; healthcare data lake; metadata-driven framework; GCP; Dataplex; BigQuery; HEDIS; HIPAA; ETL pipeline; Dataform; feature engineering

I. INTRODUCTION

Note to Reader: The framework, architecture, and performance projections presented in this paper are derived from design pattern analysis, architectural modeling, and simulated benchmarking. No proprietary production system or identifiable patient data was used. All results are exploratory projections intended as a practitioner reference, not empirical validation from a deployed production system.

Healthcare payer organizations are among the most data-intensive enterprises in the global economy. A typical large payer processes tens of millions of medical and pharmacy claims annually, maintains eligibility records for millions of members, and integrates data from hundreds of provider systems across multiple interoperability standards including HL7 FHIR, EDI 837, and NCPDP [1]. The resulting data lakes are voluminous, heterogeneous, and continuously evolving — characteristics that create significant challenges for AI and machine learning consumption.

The promise of AI in healthcare — predictive risk stratification, automated prior authorization, population health management, care gap identification — is contingent on the quality of the data ingested into ML training pipelines [2]. Low-quality data does not merely reduce model accuracy; it introduces systematic bias, generates clinically dangerous predictions, and creates regulatory exposure under HIPAA [3]. Garbage in, garbage out is not a metaphor in healthcare AI: it is a patient safety issue.

Traditional data quality frameworks — built around rule-based validation engines, static schema contracts, and threshold-based profiling — were designed for operational reporting workloads, not AI consumption. They measure quality in isolation at the point of ingestion, without awareness of downstream ML requirements such as feature completeness, temporal freshness, lineage traceability, or semantic consistency across datasets [4][5].

This paper addresses the gap between operational data quality and AI-readiness quality through five contributions:

- Six formally defined AI-Readiness quality dimensions specific to healthcare data lake environments, each with a mathematical scoring formulation.
- A composite AI-Readiness Score (AIR Score) equation aggregating dimension scores into a single ML-pipeline gating metric.
- A three-layer Metadata-Driven Data Quality Framework (MDDQF) architecture implemented on GCP using Dataplex, Dataform, BigQuery, Cloud Composer, and Cloud DLP.
- An anonymized HEDIS measure pipeline case study demonstrating framework application across three source datasets with AIR Score gating before ML feature engineering.
- Simulated evaluation benchmarks projecting AIR Score improvement and ML feature rejection rate reduction following framework implementation.

II. BACKGROUND AND RELATED WORK

A. Traditional Data Quality Dimensions

The foundational data quality literature defines quality as fitness for use [6]. Wang and Strong [7] established the influential four-category, fifteen-dimension framework — Intrinsic, Contextual, Representational, and Accessibility — that remains the most widely cited taxonomy in data quality

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

research. Batini and Scannapieco [8] extended this to include temporal dimensions including currency, timeliness, and volatility, which are particularly relevant to healthcare claims data where processing lag is common. Redman [9] emphasizes that data quality is ultimately measured by its impact on downstream decisions — a framing that directly supports the AI-readiness orientation of this paper.

Abedjan et al. [10] survey data profiling techniques for big data environments, establishing that automated metadata extraction is a prerequisite for scalable data quality assessment. These classical frameworks, however, were not designed for the multi-modal, streaming, and schema-on-read characteristics of modern data lake architectures, and none specifically addresses AI training pipeline readiness.

B. Healthcare Data Quality Challenges

Healthcare data quality presents domain-specific challenges beyond general enterprise data. HEDIS (Healthcare Effectiveness Data and Information Set) measure calculations, maintained by NCQA [11], require high precision data completeness: a single missing pharmacy dispense record can incorrectly classify a member as non-compliant with a chronic disease management measure, affecting quality bonus payments and regulatory reporting. HL7 FHIR [1] standardizes clinical data exchange but does not enforce semantic consistency across payer implementations — two systems may use different coding systems for the same clinical concept, creating silent inconsistency that standard rule-based validators cannot detect.

Polyzotis et al. [12] document systematic data quality failures in production ML systems at Google, including distribution shift, schema staleness, and silent feature corruption — all of which are prevalent in healthcare payer ETL pipelines. Sculley et al. [13] characterize hidden technical debt in ML systems, with data pipeline debt identified as the dominant contributor to production model degradation. Breck et al. [14] propose data validation for ML at scale, introducing the concept of schema-based expectation testing as a ML readiness gate — a concept this paper extends with a full six-dimension scoring model.

C. GCP Data Quality Tooling

Google Cloud Platform provides a native suite of data quality and governance tools. Dataplex [15] provides a unified data governance layer with automated metadata discovery, data scanning, data quality rules, and a centralized metadata catalog across BigQuery, Cloud Storage, and Cloud Bigtable. Dataform enables SQL-based transformation pipelines with built-in assertion testing for data quality validation within BigQuery. Cloud DLP (Data Loss Prevention) provides automated PII detection and classification, critical for HIPAA-compliant data lake governance. Cloud Composer (managed Apache Airflow) orchestrates multi-step quality workflows with dependency management and retry logic.

To the authors' knowledge, no prior published work formalizes a six-dimension AI-readiness scoring framework combining GCP-native tooling with a composite AIR Score for HIPAA-regulated healthcare data lake environments — the primary contribution of this paper.

III. AI-READINESS QUALITY DIMENSIONS

Six quality dimensions are defined as necessary and sufficient for AI-readiness assessment in healthcare data lake environments. Each dimension addresses a distinct failure mode observed in ML training pipelines consuming healthcare payer data. Table I provides formal definitions and healthcare-specific examples.

Dimension	Definition	Healthcare Example
Completeness (C)	Ratio of non-null values to total records L-required fields	Member date of birth present in eligibility record
Consistency (K)	Agreement of values across related datasets and coding systems	ICD-10 code matches diagnosis in claims and clinical notes
Freshness (F)	Recency of data relative to ML pipeline SLA window	Claims adjudicated within 72-hour ingestion SLA
Lineage (L)	Traceability of data from source system to ML feature	Pharmacy dispense traceable to NCPDP source and transformation
Schema Stability (S)	Rate of non-breaking vs. breaking schema changes over time	New optional column added vs. required column renamed
Semantic Fidelity (M)	Correctness of clinical concept encoding relative to reference terminology	SNOMED CT code maps accurately to intended clinical concept

TABLE I. SIX AI-READINESS QUALITY DIMENSIONS WITH HEALTHCARE DEFINITIONS

Completeness is formalized as the ratio of non-null values across all ML-required fields $f_1 \ldots f^r$ for a dataset D_i with N records:

$$C_i = (1/|F|) \times \sum_{t \in F} (count(D_i[ff] \neq NULL) / N) \quad (1)$$

Freshness decays linearly from 1.0 as data age Δt exceeds the ML pipeline SLA window τ:

$$F_i = max(0,\ 1 - \Delta t_i / \tau_i) \quad (2)$$

where Δt_i is the elapsed time since last successful refresh and τ_i is the dataset-specific SLA threshold (e.g., $\tau = 72$ hours for claims, $\tau = 24$ hours for eligibility).

Schema Stability measures the proportion of non-breaking schema changes over a rolling 90-day window:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

$S_i = 1 - (breaking\_changes_i / total\_schema\_changes_i)$ (3)
Breaking changes include column type modifications, required column removals, and partition key changes. Non-breaking changes include optional column additions and index additions. A schema stability score below 0.70 triggers a schema review gate before ML pipeline activation.

Consistency (K_i), Lineage (L_i), and Semantic Fidelity (M_i) are scored on [0,1] via automated rule evaluation in Dataplex data quality rules and Cloud DLP classification confidence scores, respectively. Full scoring heuristics for these dimensions are defined in the MDDQF configuration specification and are beyond the scope of this paper.

IV. MDDQF THREE-LAYER ARCHITECTURE

The MDDQF organizes data quality activities across three sequential architectural layers. Fig. 1 illustrates the layer structure and data flow. Each layer is independently deployable, allowing organizations to adopt the framework incrementally.

Fig. 1 — Three-Layer Metadata-Driven Data Quality Framework Architecture

A. Layer 1: Metadata Capture

Layer 1 is responsible for continuous, automated metadata extraction from all registered data lake assets. Two complementary mechanisms operate in parallel:

- Dataplex Auto-Discovery: The Dataplex data scanning engine profiles registered BigQuery tables and Cloud Storage objects on a configurable schedule (default: daily). It extracts statistical metadata including null rates, value distributions, cardinality, and data type inference for each column, storing results in the Dataplex metadata catalog.
- Custom Tag Templates: Organization-specific metadata not captured by auto-discovery — including ML sensitivity classification (training-safe vs. PHI-adjacent), HEDIS measure association, source system lineage identifiers, and clinical coding system version— is captured via custom Dataplex tag templates applied programmatically through the Dataplex API during ETL pipeline execution.

All metadata produced by Layer 1 is written to a centralized BigQuery metadata store, partitioned by dataset identifier and scan timestamp. This partitioning scheme supports temporal quality trend analysis and dimension score history tracking.

B. Layer 2: Quality Scoring Engine

Layer 2 reads dimension-level metadata from the BigQuery metadata store and computes a composite AI-Readiness Score (AIR Score) for each registered dataset. The AIR Score is a weighted linear combination of the six dimension scores:

$AIR_i = \sum_j w_j \times D_{ij}, \quad \sum_j w_j = 1.0$ (4)

where $D_{ij} \in [0, 1]$ is the score for dataset i on dimension j, and w_j is the dimension weight. Table II defines the default weight assignments for healthcare payer ML pipelines. These weights are theoretically motivated by the relative frequency and severity of each dimension's failure modes in healthcare ETL environments, and are presented as a practitioner heuristic pending empirical calibration across multiple organization deployments.

Dimension (D_{ij})	Default Weight (w_j) — Healthcare Payer
Completeness (C)	$w_1 = 0.25$ — highest impact on ML feature null rejection rate
Consistency (K)	$w_2 = 0.20$ — cross-dataset code alignment drives model bias risk
Freshness (F)	$w_3 = 0.20$ — stale eligibility/claims causes training distribution shift
Lineage (L)	$w_4 = 0.15$ — untraceable features blocked by ML governance review
Schema Stability (S)	$w_5 = 0.10$ — breaking changes trigger pipeline rebuilds, not training failures
Semantic Fidelity (M)	$w_6 = 0.10$ — clinical coding errors affect prediction target correctness

TABLE II. DEFAULT AIR SCORE DIMENSION WEIGHTS (healthcare payer context; theoretically motivated)

An AIR Score threshold of 0.80 is defined as the AI-readiness gate for ML pipeline activation. Datasets scoring $AIR_i \geq 0.80$ are routed to ML feature engineering. Datasets scoring $AIR_i < 0.80$ are routed to Layer 3 for remediation. The 0.80 threshold is a practitioner-calibrated default; organizations should tune this value based on their specific ML model sensitivity profiles and acceptable feature rejection rates.

C. Layer 3: Remediation and Routing

Layer 3 executes targeted remediation workflows for datasets failing the AIR Score gate, orchestrated by Cloud Composer DAGs. Remediation is dimension-specific:

- Completeness failures: Dataform backfill assertions identify null fields and trigger source-system re-extraction or imputation (mean/mode for non-PHI numerical fields; flag-based imputation for categorical clinical codes per clinical domain expert rules).
- Consistency failures: Cross-dataset reconciliation SQL assertions in Dataform identify coding mismatches; SNOMED/ICD-10 mapping tables are applied to normalize terminologies to a canonical coding system.
- Freshness failures: Cloud Composer DAG retry logic re-triggers source extractions; if source system latency is confirmed, a staleness flag is written to the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

metadata store and the downstream ML pipeline is notified to apply temporal discount weighting.
- Lineage failures: Dataplex lineage graph API is queried to identify broken transformation hops; missing lineage segments are reconstructed from Cloud Composer DAG execution logs.

Following remediation, datasets are re-scored by the Layer 2 engine. A maximum of three remediation attempts per pipeline execution cycle is enforced; datasets failing to achieve $AIR_i \geq 0.80$ after three cycles are quarantined and escalated to the data engineering team via Cloud Monitoring alerting.

V. GCP IMPLEMENTATION

Fig. 2 illustrates the full GCP component integration for the MDDQF. Table III maps each framework component to its GCP service, primary function, and HIPAA compliance role.

Fig. 2 — GCP Component Integration for Metadata-Driven Data Quality

MDDQF Component	GCP Service	HIPAA Role
Metadata Catalog	Cloud Dataplex	PHI asset classification and tagging
PII Classification	Cloud DLP	Automated PHI detection, de-identification audit
Quality Assertions	Dataform (BigQuery)	SQL-level validation with audit logging
Metrics Storage	BigQuery	Encrypted at rest (CMEK); access-controlled via IAM
Orchestration	Cloud Composer	DAG execution audit trail for compliance reporting
ML Feature Store	Vertex AI Feature Store	AIR Score-gated feature registration; lineage tracking

TABLE III. GCP COMPONENT MAPPING — MDDQF IMPLEMENTATION

HIPAA Compliance Note: The MDDQF implementation as described involves processing of datasets that may contain Protected Health Information (PHI). Cloud DLP classification, Dataplex PHI tagging, and BigQuery CMEK encryption are components designed to support HIPAA-compliant data governance. However, HIPAA compliance is an organizational, legal, and technical determination that requires formal risk assessment, Business Associate Agreements with all GCP services, and review by qualified compliance counsel. This paper does not constitute compliance guidance.

The MDDQF implementation follows GCP's VPC Service Controls perimeter model, ensuring all data quality metadata and PHI-adjacent scoring operations occur within a controlled network boundary. Cloud Composer DAGs authenticate via Workload Identity Federation, eliminating service account key management risk. All inter-service communication uses Private Google Access with no data traversal over public internet.

VI. ANONYMIZED CASE STUDY: HEDIS MEASURE PIPELINE

To ground the MDDQF in operational reality, this section presents an anonymized case study based on a composite of healthcare payer HEDIS measure calculation ETL topologies. All system identifiers, member counts, and vendor-specific details have been removed or generalized.

A. Pipeline Topology and Data Sources

The HEDIS measure pipeline draws from three source datasets, each with distinct quality profiles and AI-readiness challenges:

- Member Eligibility (daily refresh from enrollment system): High completeness for demographic fields; frequent freshness failures during month-end enrollment processing windows; moderate schema stability risk due to benefit plan restructuring.
- Medical Claims (72-hour adjudication SLA): High volume (multi-million records per month); significant consistency challenges due to ICD-10 coding variation across 200+ provider organizations; lineage complexity due to claims adjustment and resubmission chains.
- Pharmacy Dispense (NCPDP daily feed): High completeness and semantic fidelity; freshness risk during pharmacy system outages; schema stability risk from NCPDP standard version transitions.

Fig. 3 illustrates the HEDIS pipeline AIR Score quality gate workflow. Datasets passing the 0.80 threshold are routed to HEDIS measure feature engineering (numerator/denominator extraction for measures such as CDC-HE Diabetes Care and CHL Cholesterol Management). Datasets failing the gate enter the Layer 3 remediation cycle before re-scoring.

Fig. 3 — HEDIS Measure Pipeline: AIR Score Quality Gate Workflow

B. Quality Dimension Scores by Dataset

Table IV reports the AIR Score and dimension breakdown for each source dataset before and after MDDQF implementation, based on simulated profiling using synthetic data distributions calibrated to documented healthcare payer data quality benchmarks [12][14].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Dataset	Before AIR	After AIR	Δ AIR	Primary Dimension Improved
Member Eligibility	0.61	0.88	+0.27	Freshness, Lineage
Medical Claims	0.54	0.83	+0.29	Consistency, Semantic
Pharmacy Dispense	0.68	0.91	+0.23	Schema Stability
Lab Results	0.49	0.82	+0.33	Completeness, Lineage
Prior Authorization	0.57	0.86	+0.29	Consistency, Freshness

TABLE IV. AIR SCORE BEFORE/AFTER MDDQF — HEDIS PIPELINE DATASETS (Simulated Projections)

VII. EVALUATION AND RESULTS

Simulation Scope: All results in this section are derived from simulated benchmarking using synthetic healthcare data distributions. No production ML model was trained or deployed. Results represent projected improvements based on documented relationships between data quality dimensions and ML pipeline outcomes in the literature [12][13][14]. Physical production validation on real patient data in a deployed GCP environment is identified as future work.

A. AIR Score Distribution Improvement

Fig. 4 presents the before/after AIR Score comparison across all five HEDIS pipeline datasets. Prior to MDDQF implementation, four of five datasets fall below the 0.80 AI-readiness threshold. Following framework application, all five datasets exceed the threshold, with a mean improvement from 0.578 to 0.860.

Fig. 4 — AIR Score Before/After MDDQF Implementation (Simulated Projections)

The Lab Results dataset shows the largest improvement (+0.33), primarily driven by Completeness remediation: synthetic profiling identified a 22% null rate in the lab result value field, attributable to a source system interface defect that was remediated through Dataform backfill assertion and source re-extraction. The Medical Claims dataset shows the largest absolute Consistency improvement, driven by ICD-10 normalization across provider coding variations using NCQA-specified crosswalk tables.

B. Quality Dimension Heatmap

Fig. 5 presents the post-framework quality dimension score heatmap across all five datasets, revealing remaining quality risk concentrations. Schema Stability scores remain the lowest dimension across all datasets (range: 0.58-0.72), reflecting the high rate of HEDIS measure specification changes (annual NCQA updates) that trigger schema modifications in the measure calculation pipeline. This finding suggests that Schema Stability weight ($w_5 = 0.10$) may need upward revision in organizations with annual HEDIS technical specification changes.

Fig. 5 — Quality Dimension Score Heatmap by Dataset (Post-Framework, Simulated)

C. ML Feature Rejection Rate

A key downstream metric for the MDDQF is the ML feature rejection rate: the proportion of candidate features excluded from Vertex AI Feature Store registration due to quality gate failure. Table V reports simulated feature rejection rates before and after framework application across three ML model types consuming HEDIS pipeline data.

ML Model Type	Feature Rejection Before (%)	Feature Rejection After (%)
Risk Stratification Model	31%	7%
Care Gap Identification Model	38%	9%
Prior Auth Prediction Model	33%	8%
Mean Across Models	**34%**	**8%**

TABLE V. ML FEATURE REJECTION RATE BEFORE/AFTER MDDQF (Simulated Projections)

The 34% to 8% mean reduction in feature rejection rate represents the primary operational value proposition of the MDDQF for ML engineering teams: fewer pipeline runs blocked by quality gate failures, reduced manual data remediation effort, and faster iteration cycles for model retraining. These projections are consistent with documented feature rejection rates reported in production ML systems consuming healthcare claims data [12].

D. Limitations

This work carries limitations that bound its generalizability. First, all quantitative results are simulation-derived using synthetic data distributions; physical production validation is required. Second, AIR Score dimension weights are theoretically motivated and require empirical calibration via cross-organization regression analysis. Third, the MDDQF is designed for the GCP toolchain specifically; organizations using AWS (Glue/Athena) or Azure (Purview/Synapse) would require platform-specific implementation adaptations. Fourth, the 0.80 AIR Score gate threshold requires tuning per ML model sensitivity profile; a risk

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

stratification model may tolerate AIR = 0.75 while a clinical decision support model may require AIR = 0.90.

VIII. FUTURE WORK

Five directions extend and validate the MDDQF contributions:

- Production Deployment Validation: Deploy the MDDQF in a live GCP healthcare payer environment and measure actual AIR Score distributions, remediation cycle times, and ML feature rejection rates against the simulated projections in Tables IV and V. This is the critical empirical validation step identified by this paper.
- Empirical AIR Score Weight Calibration: Conduct a multi-organization study across 5-10 healthcare payer GCP environments to calibrate dimension weights ($w_1 ... w_6$) using multivariate regression against observed ML model performance degradation as the ground truth quality impact signal.
- Streaming Quality Assessment: Extend the MDDQF from batch-oriented Dataplex scanning to streaming quality assessment for real-time data ingestion pipelines (Pub/Sub, Dataflow), enabling continuous AIR Score monitoring rather than scheduled evaluation.
- Adaptive Threshold Management: Develop a Vertex AI AutoML-based threshold optimizer that dynamically adjusts dimension weights and the AIR Score gate threshold based on observed model retraining outcomes, creating a feedback loop between ML performance metrics and data quality scoring.
- Cross-Cloud Framework Portability: Design platform-agnostic MDDQF specification layers (dimension definitions, scoring equations, remediation logic) that can be implemented on AWS Lake Formation + Glue Data Quality and Azure Purview + Synapse Analytics, enabling the framework to be adopted across multi-cloud healthcare organizations.

IX. CONCLUSION

This paper has presented the Metadata-Driven Data Quality Framework (MDDQF) for AI-ready healthcare data lakes, addressing the gap between operational data quality management and the dynamic, multi-dimensional quality requirements of ML training pipelines in regulated healthcare environments.

The framework's five contributions — six formally defined AI-readiness quality dimensions (Equations 1-3), a composite AIR Score (Equation 4), a three-layer GCP-native architecture, an anonymized HEDIS measure pipeline case study, and simulated benchmark projections — provide data engineering practitioners with a structured, implementable approach to ML feature quality governance on Google Cloud Platform.

Simulated evaluation projects mean AIR Score improvement from 0.578 to 0.860 across five HEDIS pipeline datasets, with ML feature rejection rates declining from 34% to 8% following framework application. These projections are explicitly presented as exploratory estimates pending production validation, and the AIR Score dimension weights are identified as theoretically motivated heuristics requiring empirical calibration.

The MDDQF's most immediate practical value is the operationalization of AI-readiness as a first-class data engineering concern. In regulated healthcare environments where low-quality training data can produce clinically dangerous model predictions, systematic metadata-driven quality scoring before ML pipeline activation is not merely an engineering best practice — it is a patient safety imperative. As healthcare organizations accelerate AI adoption across care gap identification, risk stratification, and prior authorization automation, frameworks that encode quality governance into the data pipeline architecture — rather than delegating it to ad hoc data scientist interventions — will be essential infrastructure for responsible AI deployment.

ACKNOWLEDGMENT

The author acknowledges the data engineering and AI architecture teams at Accenture LLP for collaborative development of healthcare data governance frameworks that informed the design patterns presented in this paper. The author thanks the Google Cloud Dataplex, Dataform, and Vertex AI product teams for open documentation and technical reference materials that supported the GCP implementation architecture described herein.

REFERENCES

[1] HL7 International, "HL7 FHIR Release 4: Fast Healthcare Interoperability Resources," HL7 FHIR Foundation, 2019. [Standards reference.]

[2] R. Kohavi and F. Provost, "Glossary of terms pecial issue on applications achine learning and the knowledge discovery process," Machine Learning, vol. 30, pp. 271-274, 1998.

[3] U.S. Department of Health and Human Services, "HIPAA Security Rule," R Parts nd 164, 2003. [Regulatory reference.]

[4] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," Journal anagement Information Systems, vol. 12, no. 4, pp. 5-33, 1996.

[5] T. C. Redman, Data Quality for the Information Age. Artech House, 1996.

[6] J. M. Juran, Juran on Quality by Design. Free Press, 1992.

[7] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," Journal anagement Information Systems, vol. 12, no. 4, pp. 5-33, 1996.

[8] C. Batini and M. Scannapieco, Data and Information Quality. Springer, 2016.

[9] T. C. Redman, "Data quality: The field guide," Digital Press, 2001.

[10] Z. Abedjan, L. Golab, and F. Naumann, "Profiling relational data: A survey," VLDB Journal, vol. 24, no. 4, pp. 557-581, 2015.

[11] National Committee for Quality Assurance (NCQA), "HEDIS 2024 Technical Specifications ealth Plans," NCQA, 2024. [Standards reference.]

[12] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data lifecycle challenges in production machine learning: A survey," SIGMOD Record, vol. 47, no. 2, pp. 17-28, 2018.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

[13] D. Sculley et al., "Hidden technical debt in machine learning systems," Advances in Neural Information Processing Systems (NeurIPS), vol. 28, 2015.

[14] E. Breck, N. Polyzotis, S. Roy, S. Whang, and M. Zinkevich, "Data validation for machine learning," in Proc. SysML Conference, 2019.

[15] Google Cloud, "Cloud Dataplex: Intelligent data fabric for data management," Google Cloud Documentation, 2024. [Practitioner reference; not peer-reviewed.]

[16] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs," in Proc. SEMANTiCS, 2016.

[17] A. Halevy, A. Rajaraman, and J. Ordille, "Data integration: The teenage years," in Proc. VLDB, 2006.

[18] P. Christen, Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, 2012.

[19] M. Stonebraker et al., "Data curation at scale: The data tamer system," in Proc. CIDR Conference, 2013.

[20] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, "Profiler: Integrated statistical analysis and visualization for data quality assessment," in Proc. AVI, 2012.