

(RESEARCH ARTICLE)



INTELLIGENT HEALTHCARE DATA LAKEHOUSES: INTEGRATING MEDALLION ARCHITECTURE, DATA QUALITY, AND AI-READY ANALYTICS

Anil Kumar Kolla ^{1, *} and Ameya Kokate ²

¹ Department of Information Technology, HonorHealth, USA.

² Department of Data Science and Analytics, HonorHealth, USA.

* Corresponding Author

ORCID Details

Anil Kumar Kolla: <https://orcid.org/0009-0001-8541-0855>

Ameya Kokate: <https://orcid.org/0009-0005-9777-3273>

Global Journal of Engineering and Technology Advances, 2026, 28(03), 280–287

Publication history: Received on 14 June 2026; revised on 20 September 2026; accepted on 22 September 2026

Article DOI: <https://doi.org/10.30574/gjeta.2026.28.3.0261>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Healthcare organisations hold large volumes of clinical, administrative, and claims data, yet routinely find that data unusable for analytics or machine learning without substantial remediation. The obstacles are structural: heterogeneous source semantics, incomplete and temporally inconsistent records, weak provenance, and privacy constraints that limit the free movement of data across environments. This paper presents a layered lakehouse architecture for healthcare analytics that combines a bronze–silver–gold (medallion) refinement model with an embedded, declarative data-quality regime and a terminology standardisation stage aligned to established interoperability and common-data-model standards. The contribution is threefold. First, we specify what each medallion layer must guarantee in a healthcare context, replacing the informal "raw / cleaned / curated" description with explicit per-layer contracts covering immutability, identity resolution, temporal semantics, terminology binding, and de-identification state. Second, we define a data-quality framework mapping the harmonised electronic health record quality dimensions of conformance, completeness, and plausibility onto executable assertions positioned at specific layer boundaries, with quarantine rather than silent drop as the default failure action. Third, we specify AI-readiness criteria for the gold layer - point-in-time correctness, label provenance, cohort reproducibility, and documented population coverage - that distinguish a table that is merely queryable from one that can responsibly support model development. We describe implementation on a transactional lakehouse, propose an evaluation protocol, and analyse limitations including the persistence of unstructured clinical narrative and the risk that quality filtering itself introduces population bias.

Keywords: Healthcare Informatics, Data Lakehouse, Medallion Architecture, Data Quality, Interoperability, FHIR, Common Data Model, AI Readiness, Clinical Analytics.

1 INTRODUCTION

Analytic ambition in healthcare consistently outruns data readiness. Organisations commission predictive models for readmission, deterioration, or utilisation, and then discover that the underlying data cannot support them: encounter records lack reliable timestamps, patient identity is fragmented across source systems, laboratory results are recorded in inconsistent units, and the meaning of a field has silently changed at some point in the historical record. The result is projects that stall in data preparation, or worse, models trained on data whose defects are invisible in aggregate but decisive in effect.

The medallion pattern - progressive refinement through bronze, silver, and gold layers on a lakehouse substrate [1], [2] - has become common as an organising principle. Its weakness is that it is usually adopted as a naming convention rather than as a contract. Teams create three schemas and place tables in them by intuition, with no stated guarantee about what a consumer of a silver table may assume. In a healthcare setting, where a misinterpreted table can propagate into clinical decision support, this ambiguity is not benign.

This paper argues that the value of a layered architecture lies entirely in the guarantees each layer makes, and that those guarantees must be domain-specific and machine-enforced. We contribute:

- Per-layer contracts for healthcare data specifying, for each medallion layer, what is guaranteed about immutability, identity, time, terminology, and identifiability.
- A placement model for data-quality assertions that maps the conformance/completeness/plausibility taxonomy for electronic health record data [3], [4] to specific layer boundaries, with defined failure semantics.
- AI-readiness criteria for gold-layer assets, addressing the specific defects that make otherwise-clean healthcare tables unsuitable for model training.
- An evaluation protocol for assessing the architecture's effect on data usability and analytic delivery.

Contribution type and scope. This is an architecture and methodology contribution. It specifies layer contracts, assertion placement, and AI-readiness criteria, together with the protocol by which their effect should be measured. It does not report an empirical evaluation on clinical data. the Evaluation Protocol section states the measurements and Table 2 gives the pre-specified reporting template; the values are to be supplied by an implementer on their own estate.

2 BACKGROUND AND RELATED WORK

2.1 Lakehouse Foundations

The lakehouse model [1] reconciles the low-cost, open-format storage of data lakes with the transactional guarantees and schema enforcement of warehouses. Transactional table formats [2] supply ACID commits, schema evolution control, and time travel over object storage. Three of these properties are load-bearing for the architecture proposed here: schema enforcement makes layer contracts machine-checkable; time travel makes historical reconstruction possible, which is essential for auditability and for point-in-time correct feature construction; and atomic commits allow quality gates to fail a batch without leaving partial state visible to consumers.

2.2 Healthcare Interoperability and Common Data Models

Two standardisation traditions are relevant. HL7 FHIR [5] defines resource-oriented representations for clinical concepts and is oriented toward exchange between systems. The OMOP Common Data Model, developed within the OHDSI community [6], defines a relational representation with standardised vocabularies oriented toward observational research and cross-institution comparability. The two are complementary rather than competing: FHIR is typically the shape in which data arrives, while OMOP or a similar CDM is a useful target shape for the analytic layer. The architecture proposed here treats terminology binding - the mapping of local codes to standard vocabularies - as an explicit, versioned transformation with its own quality assertions, rather than as an incidental step inside a transformation script.

2.3 Data Quality in Healthcare

General data quality frameworks [7], [8] define dimensions such as accuracy, completeness, consistency, and timeliness. Healthcare-specific work is more directly applicable. Weiskopf and Weng [3] reviewed methods for assessing EHR data quality for secondary use and documented the inconsistency of both terminology and method across studies. Kahn *et al.* [4] proposed a harmonised terminology organising quality assessment into conformance (does the value conform to structural, computational, and relational constraints), completeness (is the value present when it should be), and plausibility (is the value believable given clinical knowledge, internal consistency, and temporal patterns). This taxonomy is the basis for the assertion placement model in the Proposed Architecture section.

From the machine learning side, Breck *et al.* [9] demonstrated the value of continuous, automated validation of production data, and Polyzotis *et al.* [10] and Sculley *et al.* [11] documented the maintenance costs incurred when data dependencies are unmanaged.

2.4 Gap

The literature contains mature treatments of lakehouse architecture, of healthcare data models, and of data quality assessment, but these are largely developed in isolation. Practitioner accounts of medallion architecture rarely specify enforceable layer contracts, and healthcare data quality research rarely addresses where in a modern pipeline its assessments should execute or what should happen when they fail. Recent work has begun to automate the bronze-to-silver boundary itself: Kolla *et al.* [16] drive column mapping, type conversion, declarative quality rules, and quarantine routing from a central metadata repository on a Databricks lakehouse, showing that validation at a layer boundary can be specified as governed configuration rather than as per-table code. That framework is domain-neutral, however: it does not address the healthcare-specific guarantees — identity resolution, temporal semantics, terminology binding, and de-identification state — that this paper treats as the substance of a layer contract. This paper connects the three.

3 PROPOSED ARCHITECTURE

3.1 Layer Contracts

Bronze — faithful capture. The bronze layer guarantees that the platform holds an immutable, append-only record of what each source system transmitted, with ingestion metadata (source system identifier, extraction timestamp, batch identifier, file or message checksum) attached to every record. Payloads are retained in their arrival structure, including malformed records. Nothing is corrected, deduplicated, or dropped. The rationale is evidentiary: when an analytic result is later disputed, the organisation must be able to demonstrate what it received and when. Bronze is by definition identifiable data and inherits the strictest access controls.

Silver — conformed and resolved. The silver layer guarantees five properties:

- *Structural conformance*: records satisfy declared schemas and data types; non-conforming records are quarantined, never coerced silently.
- *Identity resolution*: a stable enterprise patient identifier is assigned, with the matching method and confidence recorded per linkage, and probabilistic matches below threshold flagged rather than merged.
- *Temporal semantics*: event, recording, and effective times are distinguished explicitly. Conflating them is a common and consequential defect, because a model trained on recording time will silently leak information unavailable at prediction time.
- *Terminology binding*: local codes are mapped to standard vocabularies using a versioned mapping table; the vocabulary version is recorded on each row, and unmapped codes are retained with an explicit unmapped marker rather than discarded.
- *Unit and value normalisation*: measurements are converted to canonical units, with the original value and unit preserved alongside the normalised pair.

Silver is deduplicated but not aggregated, and remains at the grain of the clinical event.

Gold — analytic and AI-ready. The gold layer contains purpose-built assets — cohort tables, measure tables, feature tables, reporting marts. Its guarantees are described in the AI-Readiness Criteria for Gold Assets subsection.

De-identification. De-identification is modelled as a transformation producing a parallel asset with its own lineage and its own access policy, not as an in-place mutation. The identifiability state of every asset is an explicit, queryable attribute, so that an access decision never depends on convention or on a table's name.

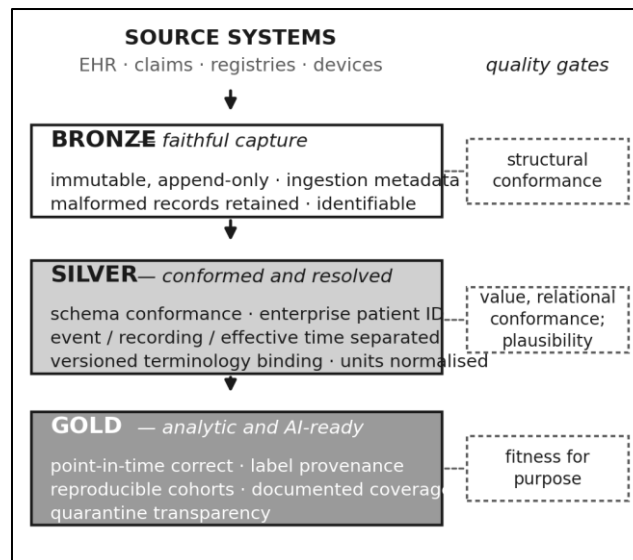


Figure 1: Medallion layers expressed as enforceable contracts, with quality gates positioned at each layer boundary.

3.2 Quality Assertion Placement

Assertions are declarative, versioned alongside the transformation code, and executed as part of the pipeline rather than as an out-of-band audit. Placement follows the taxonomy of Kahn *et al.* [4]:

Table 1: Data-quality assertion placement at layer boundaries, with assertion class, representative checks, and failure action

Boundary	Assertion class	Examples	Failure action
Ingest → Bronze	Structural conformance	File parseable, expected columns present, checksum valid, batch not duplicated	Fail batch; alert; no partial commit
Bronze → Silver	Value conformance	Data types, code set membership, permitted value ranges, referential integrity to reference tables	Quarantine record to dead-letter table with reason code
Bronze → Silver	Relational conformance	Every observation resolves to a valid patient and encounter	Quarantine; escalate if rate exceeds threshold
Silver	Completeness	Expected-field population rates by source and by period; absence of unexplained gaps in time series	Alert; annotate asset; do not block
Silver	Plausibility (atemporal)	Physiologically implausible values; internally inconsistent combinations such as procedure–sex incompatibility	Flag record; retain with quality annotation
Silver	Plausibility (temporal)	Distribution stability across periods; detection of abrupt shifts indicating upstream semantic change	Alert; hold promotion to gold pending review
Silver → Gold	Fitness for purpose	Cohort size within expected bounds; feature null rates within tolerance; point-in-time constraints satisfied	Block promotion

Two design decisions warrant emphasis. First, quarantine rather than drop: records failing conformance are written to a dead-letter table with a machine-readable reason, preserving both the record and the reason for exclusion. Silent dropping destroys the evidence needed to diagnose upstream problems and biases downstream populations invisibly. Metadata-driven implementations of this pattern on a Databricks lakehouse [16] show that the valid/quarantined split, the per-record reason code, and the run-level audit counts can all be generated from configuration rather than written by hand for each table. Second, completeness and plausibility annotate rather than block. A physiologically implausible laboratory value may be an error or may be a genuine extreme finding; automated deletion of extremes in a clinical

dataset removes precisely the observations that matter most. The pipeline's job is to attach a quality annotation and let the analytic consumer decide.

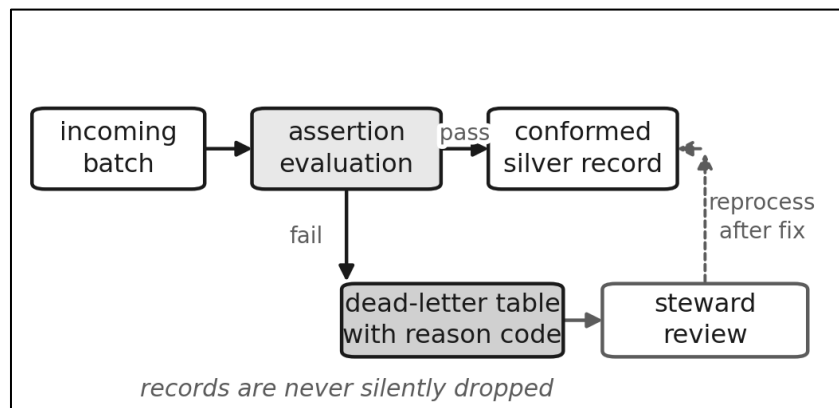


Figure 2: Quarantine semantics. Non-conforming records are diverted with a machine-readable reason code rather than dropped, preserving the evidence needed for upstream diagnosis.

3.3 AI-Readiness Criteria for Gold Assets

A gold asset intended to support model development must additionally satisfy:

Point-in-time correctness. Every feature value must be reconstructible as of the prediction timestamp using only information recorded at or before that time. Retrospective corrections and back-dated entries — common in clinical documentation — are the principal source of leakage in healthcare models and must be handled explicitly using recording-time filters rather than event-time filters.

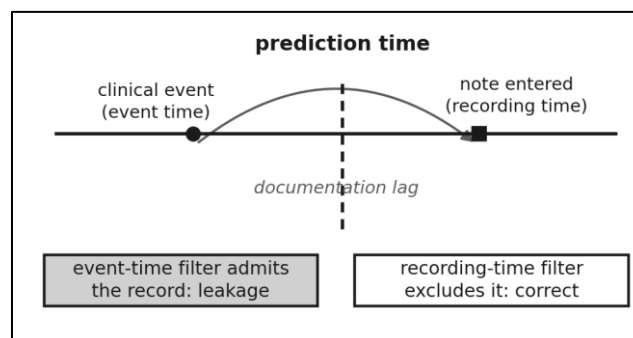


Figure 3: Point-in-time correctness. Back-dated clinical documentation leaks into features when the feature window is filtered on event time rather than recording time.

Label provenance. The definition of the outcome variable must be expressed as versioned code with a stated ascertainment window, not as an ad hoc query. Ambiguity in outcome definition is a frequent cause of irreproducible results.

Cohort reproducibility. Inclusion and exclusion criteria are executable and versioned, and any cohort can be regenerated exactly using the table version identifiers recorded in its metadata.

Documented population coverage. Each asset carries a statement of which source systems, care settings, and time periods it covers, and which populations are systematically under-represented. This is the dataset-documentation practice advocated by Gebru *et al.* [12], applied to internal assets.

Quarantine transparency. The proportion of source records excluded by quality gates is reported per cohort, stratified by available demographic attributes, so that quality-induced selection bias is visible.

3.4 Governance and Privacy

Access control is attribute-based, driven by the identifiability state, data domain, and purpose-of-use recorded on each asset. Lineage is captured at column level where the platform supports it, enabling both impact analysis and the

reconstruction of any published figure. Retention policies are enforced per layer, with bronze typically subject to the longest retention for evidentiary reasons and derived gold assets subject to shorter, purpose-bound retention.

4 IMPLEMENTATION CONSIDERATIONS

Ingestion is incremental and idempotent, keyed on source batch identifiers, so that replayed deliveries do not duplicate. Distributed execution uses a unified analytics engine [13]. Transformations between layers are expressed declaratively where possible and are idempotent with respect to reprocessing, since reprocessing after a mapping correction is routine rather than exceptional.

Terminology mappings are held as versioned reference tables rather than embedded in code. When a mapping is corrected, the correction is applied as a new version, and affected silver and gold partitions are rebuilt with the new version recorded — never patched in place. This preserves the ability to explain why an earlier report differed.

Slowly changing dimensions are required for provider, facility, and patient-attribute tables; type-2 history in the dimensional-modelling sense [14] is the default, because analyses frequently need the attribute value as of the encounter rather than as of today.

Performance considerations include partitioning by event date rather than ingestion date for query-aligned pruning, and clustering on patient identifier for cohort extraction workloads, which are dominated by selective per-patient access.

5 EVALUATION PROTOCOL

Implementers should report the following, measured on their own environment; where an institutional estate cannot be shared, a public critical-care corpus such as MIMIC-III [15] provides a reproducible substrate for the conformance, completeness, and point-in-time checks below. *This paper specifies the measurements; the values are not asserted here.*

Data quality outcomes. Conformance failure rate by source and layer; quarantine rate with reason-code distribution; completeness by expected-field and period; count of temporal plausibility alerts and the proportion confirmed as genuine upstream changes. The last is the key indicator of whether the plausibility layer is producing signal or noise.

Analytic delivery. Time from new source onboarding to a validated silver asset; time from analytic request to gold asset delivery; proportion of analytic requests satisfied from existing gold assets without new engineering.

Reproducibility. Proportion of published analytic results regenerable to exact agreement from recorded table versions.

AI readiness. Proportion of gold feature tables passing point-in-time correctness verification, tested by an adversarial leakage check that compares model performance on properly and improperly time-filtered feature sets; a large unexplained gap indicates leakage.

Table 2: Pre-specified reporting template. Values are to be measured on the implementer's own estate and are not asserted in this paper

Measure	Pre-existing architecture	Unlayered lake	Proposed architecture
Conformance failure rate	—	—	—
Quarantine rate	—	—	—
Expected-field completeness	—	—	—
Unmapped terminology proportion	—	—	—
Temporal plausibility alerts confirmed	—	—	—
Results regenerable to exact agreement	—	—	—
Gold tables passing leakage check	—	—	—
Time from source onboarding to silver	—	—	—

Baselines. Comparison against the pre-existing architecture on the same workloads, and against an unlayered lake configuration, is necessary for the results to be interpretable.

6 DISCUSSION AND LIMITATIONS

Quality filtering can introduce bias. Every quarantine decision removes records non-randomly. Records from smaller source systems, from settings with less structured documentation, or from populations with fragmented care are disproportionately likely to fail conformance checks. Reporting quarantine rates stratified by demographic attributes is therefore not an optional nicety; without it, the architecture may improve measured quality while degrading representativeness.

Unstructured narrative remains largely outside the model. A substantial proportion of clinically meaningful information resides in free-text notes. The architecture accommodates extracted structured representations as derived silver assets, but the extraction step carries its own error characteristics, and those errors are not addressable by the conformance and plausibility assertions described here.

Terminology mapping is not a solved problem. Local-to-standard code mapping requires clinical domain judgement, is incompletely automatable, and degrades as source systems change. Recording the unmapped proportion per source and period is essential, since a rising unmapped rate is often the earliest visible symptom of an upstream change.

Cost. Immutable bronze retention with full history is expensive at scale. Organisations should expect to make explicit, documented retention trade-offs rather than assuming indefinite retention.

Organisational dependency. Layer contracts are only meaningful if enforced. Where teams retain the ability to publish directly to gold, the guarantees degrade to conventions. The architecture presupposes governance authority that many organisations do not have, and this should be treated as a prerequisite rather than an implementation detail.

7 CONCLUSION

This paper argued that the medallion pattern delivers value only when its layers carry enforceable, domain-specific contracts, and it specified such contracts for healthcare data covering immutability, identity, temporal semantics, terminology binding, and identifiability. It positioned the conformance/completeness/plausibility quality taxonomy at defined pipeline boundaries with explicit failure semantics, defaulting to quarantine and annotation rather than silent exclusion. Finally, it defined AI-readiness criteria - point-in-time correctness, label provenance, cohort reproducibility, documented coverage, and quarantine transparency - that separate assets that can be queried from assets that can responsibly train models.

Future work should address automated detection of upstream semantic change without schema change, which remains the least tractable failure class; the integration of note-derived structured data under the same quality regime; and federated evaluation across institutions, where the architecture's assumption of centralised curation does not hold.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The authors thank the editorial team and the anonymous reviewers for their comments on the manuscript.

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding this manuscript.

REFERENCES

- [1] M. Armbrust, A. Ghodsi, R. Xin, and M. Zaharia, "Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics," in *Proc. Conf. Innovative Data Systems Research (CIDR)*, 2021.
- [2] M. Armbrust *et al.*, "Delta Lake: High-performance ACID table storage over cloud object stores," *Proc. VLDB Endowment*, vol. 13, no. 12, pp. 3411–3424, 2020.
- [3] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.

- [4] M. G. Kahn *et al.*, "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *eGEMs (Generating Evidence & Methods to Improve Patient Outcomes)*, vol. 4, no. 1, 2016.
- [5] Health Level Seven International, *HL7 Fast Healthcare Interoperability Resources (FHIR) Specification*, Ann Arbor, MI, USA. [Online]. Available: <https://www.hl7.org/fhir/>
- [6] G. Hripcsak *et al.*, "Observational Health Data Sciences and Informatics (OHDSI): Opportunities for observational researchers," *Studies in Health Technology and Informatics*, vol. 216, pp. 574–578, 2015.
- [7] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.
- [8] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–52, 2009.
- [9] E. Breck, N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data validation for machine learning," in *Proc. Conf. Systems and Machine Learning (SysML/MLSys)*, 2019.
- [10] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data management challenges in production machine learning," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2017, pp. 1723–1726.
- [11] D. Sculley *et al.*, "Hidden technical debt in machine learning systems," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2015, pp. 2503–2511.
- [12] T. Gebru *et al.*, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.
- [13] M. Zaharia *et al.*, "Apache Spark: A unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, 2016.
- [14] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed. Indianapolis, IN, USA: Wiley, 2013.
- [15] A. E. W. Johnson *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, art. 160035, 2016.
- [16] A. K. Kolla, A. Kokate, R. Vellanki, R. Lolam, and P. Bandi, "A metadata-driven framework for automated bronze-to-silver data ingestion, validation and loading on Databricks," *European Journal of Information Technologies and Computer Science*, vol. 6, no. 3, pp. 1–7, 2026, doi: 10.24018/compute.2026.6.3.70578.