Original Article

e-ISSN: 2774-2016 - https://journal.itera.ac.id/index.php/indojam/
p-ISSN: 2774-2067

Received 5th October 2026
Accepted 26th October 2026
Published 31st October 2026

Open Access

DOI
https://doi.org/10.35472/indojam.v6i2.3029

Evidence-First Data Management for Trustworthy AI: A Database Architecture for Ecological Tourism Intelligence

Ardika Satria*[a], Mika Alvionita Sitinjak [c], Ira Safitri[c], Febri Dwi Irawati[d]

[a,b,c,d] *Data Science Study Program, Faculty of Science, Institut Teknologi Sumatera, Lampung, Indonesia*

* *Corresponding E-mail: ardika.satria@sd.itera.ac.id*

Abstract: Multi-source operational data is rarely stored with enough structure to support trustworthy artificial-intelligence consumption: values are overwritten, provenance is recorded per record rather than per attribute, and unverified data reaches downstream consumers indistinguishably from verified data. This failure mode is shared by healthcare, disaster response, and crowdsourced supply chains; this paper studies it on tourism-village records from Indonesia's national JADESTA portal, a 21,461-record, single-source corpus in which only one record in five carries coordinates. The paper proposes P1, a database-driven architecture that stores every attribute value as a *claim* carrying source, time, and confidence; accumulates corrections as non-destructive *evidence events*; computes a confidence score per claim through a fully disclosed clipped linear function; and enforces a three-outcome *policy gate* before any value reaches an AI-facing *semantic view*. P1 is compared against an overwrite ETL pipeline (B0) and a record-level lineage pipeline (B1) under five injected failure scenarios, against a 304-pair manually labelled gold sample (Cohen's $\kappa = 0.742$), and through a four-variant component ablation. Three results are analytical consequences of the design and are reported as such: no single-source, unverified claim can exceed a confidence of 0.5375, so the measured zero leakage through the gate is a theorem rather than an observation; the gate separates corroborated from uncorroborated claims only for thresholds in the narrow window (0.540, 0.605], which a synthetic second-evidence-channel experiment confirms; and the entity-resolution score is a convex combination whose non-geospatial weights sum exactly to the merge threshold, so coordinate-less pairs merge only when name, district, and category agree perfectly, which explains why 121 of the 304 gold pairs score exactly 0.75 and why the merge F1 of 0.849 (precision 0.900, recall 0.804) is flat for thresholds between 0.55 and 0.75 and collapses to 0.265 above. Empirically, entity resolution raises F1 by 0.060 over the exact-key matching the baselines perform, all fifteen false merges are same-name re-registrations of one village under distinct JADESTA identifiers, and 78.6 per cent of the pairs P1 defers are genuinely distinct entities. A uniformly defined provenance-precision metric shows that record-level lineage and attribute-level claims re-stamp unchanged values at similar rates (0.471 and 0.478); the architectural difference is the 4.9 per cent of B1's stamps that attribute values to a record that never reported them, which the claim model excludes by construction. After correcting an indexing defect in the as-submitted benchmark harness, P1 is 10.7 times slower than B0 at 500 records and 86.6 times slower on the full corpus, with 87 per cent of that cost attributable to entity resolution.

Keywords: data provenance; entity resolution; evidence-first architecture; data quality management; tourism data infrastructure

Abstrak: Data operasional dari banyak sumber jarang disimpan dengan struktur yang cukup untuk dikonsumsi oleh sistem kecerdasan buatan (AI) secara dapat dipercaya: nilai ditimpa, *provenance* dicatat per *record* dan bukan per

Copyright © 2026 – Indonesian Journal of Applied Mathematics.
Content from this work may be used under the terms of the Creative Commons Attribution-NonCommercial 4.0 International Licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI. Published under licence by Indonesian Journal of Applied Mathematics.

Original Article

atribut, serta data yang belum terverifikasi sampai ke konsumen hilir tanpa dapat dibedakan dari data yang telah terverifikasi. Pola kegagalan ini juga dijumpai pada layanan kesehatan, penanggulangan bencana, dan rantai pasok berbasis *crowdsourcing*. Pada penelitian ini, persoalan tersebut dikaji menggunakan data desa wisata dari portal nasional JADESTA, yaitu korpus satu sumber berisi 21.461 *record* dengan hanya satu dari lima *record* yang memiliki koordinat. Penelitian ini mengusulkan P1, arsitektur berbasis basis data yang menyimpan setiap nilai atribut sebagai klaim yang membawa informasi sumber, waktu, dan tingkat kepercayaan; mengakumulasikan koreksi sebagai *evidence event* yang tidak bersifat destruktif; menghitung *confidence score* untuk setiap klaim melalui fungsi linear terpotong (*clipped linear function*) yang diungkapkan sepenuhnya; serta menerapkan *policy gate* dengan tiga keluaran sebelum suatu nilai diteruskan ke *semantic view* yang dikonsumsi AI. P1 dibandingkan dengan *pipeline* ETL berbasis *overwrite* (B0) dan *pipeline* dengan *lineage* tingkat *record* (B1) pada lima skenario kegagalan yang diinjeksikan, diuji terhadap *gold sample* berisi 304 pasangan yang dilabeli secara manual (Cohen's $\kappa = 0{,}742$), dan dianalisis melalui ablasi komponen dengan empat varian. Tiga hasil merupakan konsekuensi analitis dari rancangan arsitektur dan dilaporkan sebagai demikian. Pertama, klaim satu sumber yang belum terverifikasi tidak dapat melampaui *confidence* 0,5375 sehingga kebocoran nol yang terukur pada *policy gate* merupakan teorema, bukan hasil observasi. Kedua, *gate* hanya memisahkan klaim yang terkonfirmasi silang (*corroborated*) dari klaim yang tidak terkonfirmasi apabila ambang berada pada rentang sempit (0,540; 0,605], dan hal ini dikonfirmasi oleh eksperimen sintetis dengan kanal bukti kedua. Ketiga, skor *entity resolution* merupakan kombinasi konveks dengan jumlah bobot komponen non-geospasial yang tepat sama dengan ambang penggabungan (*merge threshold*). Akibatnya, pasangan tanpa koordinat hanya digabungkan apabila nama, kecamatan, dan kategori identik sepenuhnya. Kondisi ini menjelaskan mengapa 121 dari 304 pasangan *gold* memiliki skor tepat 0,75 serta mengapa *merge* F1 sebesar 0,849 (*precision* 0,900; *recall* 0,804) tetap datar pada ambang 0,55 hingga 0,75 dan turun tajam menjadi 0,265 di atas rentang tersebut. Secara empiris, *entity resolution* meningkatkan F1 sebesar 0,060 dibandingkan *exact-key matching* yang digunakan oleh *baseline*. Seluruh lima belas *false merge* merupakan registrasi ulang satu desa dengan nama yang sama di bawah *identifier* JADESTA yang berbeda, sedangkan 78,6 persen pasangan yang ditunda (*deferred*) oleh P1 memang merupakan entitas yang berbeda. Metrik *provenance precision* yang didefinisikan secara seragam menunjukkan bahwa *lineage* tingkat *record* dan klaim tingkat atribut menandai ulang (*re-stamp*) nilai yang tidak berubah dengan laju yang serupa, yaitu 0,471 dan 0,478. Perbedaan arsitekturalnya terletak pada 4,9 persen *stamp* B1 yang mengatribusikan nilai kepada *record* yang tidak pernah melaporkan nilai tersebut, sedangkan model klaim mengecualikan kondisi ini secara konstruktif (*by construction*). Setelah cacat pengindeksan pada *benchmark harness* versi naskah yang diajukan diperbaiki, P1 tercatat 10,7 kali lebih lambat dibandingkan B0 pada 500 *record* dan 86,6 kali lebih lambat pada keseluruhan korpus, dengan 87 persen dari biaya tersebut berasal dari *entity resolution*.

Kata Kunci: provenance data; resolusi entitas; arsitektur berbasis bukti; manajemen kualitas data; infrastruktur data pariwisata

Introduction

Artificial intelligence systems built on operational data are only as trustworthy as the data management layer beneath them. Model architecture and training procedure have absorbed the majority of research attention, in tourism informatics as elsewhere; substantially less attention has gone to how the data a model consumes was acquired, reconciled, corrected, and published. Yet errors introduced at this layer

propagate silently into every downstream recommendation, classification, or generated answer. Sculley et al. [1] described this asymmetry as hidden technical debt, observing that the code implementing a model is a small fraction of a production machine-learning system and that most of the risk accumulates in the surrounding data plumbing. Sambasivan et al. [2] documented the same asymmetry empirically across high-stakes deployments and named its consequence *data cascades*: compounding downstream failures whose root cause is an upstream data decision nobody recorded and nobody can now reconstruct. In a conventional database the concrete form of such a decision is an overwrite: an `UPDATE` that replaces a value in place leaves no record of which source asserted the previous value, when, or why it was replaced, so the cascade cannot be traced back to its origin even in principle. Polyzotis et al. [3] surveyed the resulting research agenda from the database-systems side and concluded that data validation, understanding, and lineage in production machine learning remain comparatively underdeveloped relative to model tooling.

This failure mode is not specific to tourism. Community-contributed data in healthcare registries, situational reports in disaster response, and supplier-reported attributes in crowdsourced supply chains exhibit the same overwrite-and-trust pattern, and the same cascades [2]. Tourism informatics is chosen here as the testbed for two reasons. First, the domain has moved, conceptually, from measuring visitor satisfaction toward measuring whether tourism restores the social-ecological system that hosts it, a shift Bellato et al. [4] formalised in the regenerative tourism framework and one that places community actors, not platform operators, at the centre of what counts as a valid observation. Operationally this means data arrives from heterogeneous sources of uneven quality: government tourism portals, community contributions, field surveys, and, in principle, IoT sensors and unstructured text. Second, a national-scale, openly scraped government corpus exists on which the reconciliation problem can be evaluated at realistic scale and sparsity. The smart-tourism literature has long anticipated this heterogeneity, and empirical comparisons of tourism data platforms have shown that the same destination is described inconsistently across sources in ways that materially change downstream analysis [5], [6], [7]. What that literature has generally not done is treat the reconciliation layer itself as a research object with its own baselines and metrics.

Two properties of this setting make naive data management strategies actively harmful rather than merely inefficient. The first is that silent overwrite destroys the ability to audit a claim's origin. A destination's operating hours, coordinates, or verification status may be asserted by a government portal at one time and corrected by a community member or field steward at another; if the correction overwrites the prior value, the system loses the ability to explain why the current value is what it is, or to recover from an incorrect correction. Provenance research in databases has established both the vocabulary and the machinery to prevent this: why-, how-, and where-provenance, interoperable provenance interchange, and a broad survey of what provenance is captured for and at what cost [8], [9], [10], [11]. The granularity actually deployed in applied platforms, however, is usually the record or the pipeline run, not the attribute. The reasons are practical rather than conceptual. Attribute-level provenance multiplies the number of stored rows by the number of attributes, so a record with ten fields becomes ten claim rows plus their events; every read of a current value becomes a query over a history rather than a lookup of a column; and a schema in which each attribute is an independent, append-only row is harder to evolve with the ordinary tooling of migrations and foreign keys than a flat current-value table [10]. These costs are real, and Section 3 measures them; the argument of this paper is that for data consumed by AI systems they are the price of a value being explainable at all.

The second property is that AI systems consume data as unqualified fact. Recommendation and reasoning components typically read a bare value from an operational table with no attached information about its source, recency, or verification status, so a single

Copyright © 2026 – Indonesian Journal of Applied Mathematics

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

unverified, low-confidence, or stale claim is indistinguishable, at the point of consumption, from a field-verified one. Retrieval-augmented generation sharpens this: the retriever places operational records directly in a language model's context window, the model repeats an unqualified value with the same fluency as a verified one, and the resulting fabrications are difficult for an end user to detect [12], [13]. The policy gate proposed here is the technical bridge between the two. It sits at exactly the boundary the retriever reads from, and it changes what the retriever can see: a value that has not cleared the gate is absent from the semantic view, not present with a warning, so the model cannot repeat it; a value that has cleared the gate arrives with its source, confidence, and verification status attached, so the prompt can carry them. Documentation conventions such as datasheets for datasets and model cards address qualification at the level of an entire dataset or model [14], [15]; the problem addressed here is one level finer, a per-value qualification that travels with the value into the query result.

Left unmanaged, tourism-village operational data exhibits six recurring failure modes. This study treats each as a first-class, separately measurable phenomenon rather than an incidental data-cleaning nuisance: duplicate records describing the same real-world entity under slightly different names or addresses; conflict, where two sources disagree on the value of the same attribute over overlapping validity periods; missing attributes that should exist but are absent from a given snapshot; stale claims whose values are never refreshed even as real-world time continues to pass; late updates, where a correction is processed after a temporally later update from another source has already been applied, risking silent regression; and schema drift, where a source begins reporting a field the ingestion pipeline does not yet recognise. These six are not an arbitrary list. They instantiate the data-quality dimensions established by Wang and Strong [16] and refined in subsequent assessment methodologies [17], [18]: duplicate maps to uniqueness, conflict to consistency, missing to completeness, stale and late update to timeliness, and schema drift to representational consistency. A data management layer that cannot detect, or at minimum survive, all six without losing information cannot be trusted to feed an AI system whose behaviour depends on that data being accurate, current, and traceable to its origin.

Existing work addresses parts of this problem convincingly but, to the best of our knowledge, no integrated data management architecture combines within one evaluable system all four of the following: attribute-level provenance, meaning a source, time, and confidence trail per attribute rather than per record; non-destructive temporal correction, in which corrections are recorded as new evidence rather than overwrites and any past state can be reconstructed; a confidence function propagated to an explicit policy gate rather than an implicit assumption that all stored data is equally trustworthy; and a semantic publication boundary safe for AI consumption, in which "not enough evidence" is a distinct, first-class outcome from both "true" and "false" and is never silently defaulted. Systems typically stop at ETL, a data warehouse, a data lake or lakehouse, or a knowledge graph without this specific combination evaluated together, end-to-end, against realistic baselines.

This study addresses that gap by proposing and evaluating P1, a database-driven, evidence-first data management architecture, following design science research as formulated and operationalised by Hevner et al. [19] and Peffers et al. [20]: an artefact is designed against an identified problem, instantiated, and then evaluated empirically against alternatives. Four concerns guide the evaluation. The first is whether an evidence-first architecture actually improves traceability and reproducibility relative to overwrite-based ETL and to record-level lineage, rather than merely appearing to. The second is whether integrating entity resolution with an attribute-level evidence ledger improves the quality of the resulting canonical entities relative to the naive matching the baselines perform. The third is whether a confidence function feeding an explicit policy gate can measurably withhold unverified or low-confidence data from an AI-facing view, and over what range of thresholds it does so correctly. The fourth is what this costs, and which component bears the cost.

The contribution is a database-driven data management architecture with six evaluated components (a canonical entity data model, attribute-level provenance management, an entity resolution mechanism, a confidence and policy-gate engine, a temporal evidence repository, and a semantic publication layer), compared against two realistic baselines and four ablated variants under controlled synthetic failure scenarios and a manually labelled gold sample on a real-world, national-scale tourism dataset. Relative to the version of this work first submitted, this revision discloses the confidence function in full, derives closed-form admissibility bounds and threshold windows for the gate, characterises the decision geometry of the entity-resolution score analytically, adds threshold and weight sensitivity analyses and an inspection of every error case, adds a component ablation, distinguishes definitional from empirical results throughout, and corrects a performance attribution that an indexing defect in the original benchmark harness had distorted.

Method

Data model. Let $S = \{s_1, s_2, \dots, s_n\}$ denote the set of data sources. Each source produces records, and a canonical mapping $M: R \to C$ associates a raw record with a canonical entity $e \in C$, a reconciled representation of one real-world tourism entity, identified independently of any single source record. The mapping is not required to be injective: several records may describe the same entity, which is exactly the duplicate failure mode, and entity resolution is the process of estimating M from record similarity [21], [22]. A claim is a tuple

$$c = \left(e, a, v, s, t_r, [t_{vf}, t_{vt}), \lambda, \sigma\right) \quad (1)$$

where e is the canonical entity, a the attribute, v the asserted value, s the source, t_r the retrieval or transaction time, $[t_{vf}, t_{vt})$ the real-world validity interval, $\lambda \in [0,1]$ a confidence score, and $\sigma \in \{\text{UNVERIFIED}, \text{VERIFIED}\}$ the verification status. The pairing of transaction time with validity time is the standard bitemporal construction [23], [24] applied here at attribute rather than row granularity. A claim is immutable once created: correcting it never mutates the original, but produces a new claim marked as superseding the prior one, which remains in the ledger permanently.

An evidence event is an immutable tuple $\varepsilon = (c, \tau, \alpha, \rho, t_o)$ recording that claim c was asserted, confirmed, corrected, or rejected ($\tau \in \{\text{ASSERTED}, \text{CONFIRMED}, \text{CORRECTED}, \text{REJECTED}\}$) by actor α holding authorisation role ρ at time t_o. Evidence events are never edited or deleted. The evidence graph for an entity-attribute pair has claims as vertices and a directed edge wherever one claim corroborates or contradicts another within overlapping validity; it is the structure over which corroboration and conflict terms are computed.

For a claim c, its provenance is $P(c) = (s, \varepsilon_c, \alpha)$: the originating source, the evidence events referencing it, and the asserting or verifying actors. The provenance function $\Pi(c)$ returns the complete, replayable lineage: the ordered chain of claims connected by their supersession links together with each one's evidence events. This is why-provenance in the sense of Buneman et al. [11], materialised rather than derived, and expressible in PROV terms without loss [8].

Confidence function. Confidence is a function $\lambda = F(P, Q, T, V)$ of provenance quality P (source reliability and actor authorisation), corroboration and conflict evidence Q, temporal freshness T, and verification status V. The choice of inputs follows the truth-discovery tradition, in which source reliability and corroboration are jointly informative about value correctness [25], [26], with freshness added because tourism attributes decay in currency even absent disagreement. The concrete form used throughout this study, previously withheld and now disclosed in full, is the clipped linear combination

$$\lambda = \text{clip}_{[0,1]}\left(w_s b_s r + w_f e^{-\Delta t / h_a} + w_k \min\left(1, \frac{k}{3}\right) + w_\rho \beta_\rho - \min(0.5, 0.15m)\right) \quad (2)$$

with weights $(w_s, w_f, w_k, w_\rho) = (0.35, 0.25, 0.20, 0.20)$. Here $b_s \in [0,1]$ is a base reliability

Copyright © 2026 – Indonesian Journal of Applied Mathematics

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

per source type (government portal 0.5, community contributor 0.3, field survey 0.6, sensor 0.6, extracted text 0.4, steward-verified 0.9), $r \in [0,1]$ is a contributor-reliability factor (0.5, the neutral default, for the portal scraper; 0.7 for field-survey submissions under the operational protocol; 1.0 for a steward's own verification), Δt is the claim's age in days and h_a an attribute-specific half-life (30 days for operational status, 60 for opening hours, 90 for price, 365 for name, address, and coordinates, 180 otherwise), k is the number of corroborating claims for the same value, $\beta_\rho \in \{0, 0.05, 0.20\}$ is an authorisation bonus for automated, community, and steward actors respectively, and m is the number of conflicting claims. Every parameter is a design hypothesis chosen by expert judgement against the operational protocol, not the result of tuning against labelled data; the study has no calibration set on which to tune them, and says so.

Two properties follow directly from (2) and are used in the analysis. First, λ is bounded in [0,1] and monotone: non-decreasing in b_s, r, k, and β_ρ, and non-increasing in Δt and m. Second, because the positive terms are bounded and the combination is linear, the supremum of λ under any evidence configuration is obtained by setting each free term to its maximum. For a claim whose only evidence is government-portal assertions by the automated scraper ($b_s = 0.5$, $r = 0.5$, $\beta_\rho = 0$, $m = 0$), the supremum over all ages and all corroboration counts is

$$\lambda^*_{\text{single}} = 0.35 \cdot 0.25 + 0.25 + 0.20 = 0.5375 < 0.60, \quad (3)$$

and without any corroboration it is 0.3375. Proposition 1 states the consequence: *no claim whose evidence consists solely of single-source, automated assertions can clear a policy threshold of 0.60, regardless of how many times the source re-asserts the value or how fresh it is.* The same substitution gives the supremum for every other configuration; Section 3 tabulates them and derives from them the window of thresholds over which the gate behaves correctly.

Policy gate. The policy gate is a function π: Claim $\rightarrow$ {allow, deny, insufficient-evidence} evaluating a claim's confidence, verification status, and conflict state against a domain-specific threshold θ: if $\lambda < \theta$ the outcome is insufficient-evidence; if $\lambda \geq \theta$ and an unresolved conflicting claim exists, deny; otherwise allow. The tourism domain uses $\theta = 0.60$ for general attributes. The three-outcome contract is fixed across every domain; only θ varies. The third outcome carries the design weight: "insufficient evidence" is not a soft form of "deny" but a distinct state, and preserving that distinction through to the query boundary is what prevents an AI consumer from reading absence as falsehood. One consequence of the evaluation order deserves stating now, because Section 3 exercises it: a conflict is reported as *deny* only when the post-penalty confidence still clears θ; a conflict that pushes confidence below θ is reported as insufficient evidence, which is the correct reading, since the claim is then both contested and under-evidenced.

The canonical state of attribute a of entity e at time t is the at-most-one currently allowed claim valid at t. The semantic view is the projection

$$V(e) = \{(a, e_t(a)) \mid a \in A(e),\ \pi(e_t(a)) = \text{allow}\} \quad (4)$$

in which every returned value carries its source, confidence, and verification status. An attribute with no allowed claim is simply absent from the view, never rendered as a null or a default, so that "no data", "unverified data", and "verified data" are never conflated at any query boundary.

Finally, given a locked source snapshot S_0 and its event log $E = (\varepsilon_1, \ldots, \varepsilon_k)$, each event induces an append-only state transition $D_{t+1} = \delta(D_t, \varepsilon_i)$, and the replayed database state is $D_t = \text{Replay}(S_0, E)$, obtained by applying each event to an initially empty state. Replay equivalence holds when re-running this replay on the same (S_0, E) produces a state that is field-by-field identical to the originally recorded one.

Entity resolution. Candidate pairs are blocked by regency (*kabupaten*) and scored by

$$s(x,y) = w_N J(x,y) + w_A A(x,y) + w_C C(x,y) + w_G G(x,y)\, \mathbf{1}[\text{both have coordinates}] \quad (5)$$

with $(w_N, w_A, w_C, w_G) = (0.40, 0.20, 0.15, 0.25)$, where J is token Jaccard similarity of the normalised names, A is 1 if both records report the same district (*kecamatan*), 0.5 if either district is unknown, and 0 if

both are known and differ, C is 1 if the category strings are identical and 0 otherwise, and $G = \max(0,1 - d/2\,\text{km})$ decays linearly with haversine distance d. A pair is merged when $s \geq \tau_{\text{merge}} = 0.75$, rejected when $s \leq \tau_{\text{reject}} = 0.35$, and deferred to a steward review queue otherwise. The weights and thresholds are expert heuristics set at design time; they were not tuned on the gold sample, and the sensitivity analysis in Section 3 is the first empirical examination of them. Because the weights sum to one, $s \in [0,1]$ is a convex combination of its signals, and one identity governs its behaviour on this corpus: the three non-geospatial weights sum to $w_N + w_A + w_C = 0.75 = \tau_{\text{merge}}$. Proposition 2 follows: *a pair in which either record lacks coordinates reaches the merge threshold if and only if* $J = A = C = 1$, that is, only when the normalised names are token-identical, both districts are known and equal, and the categories match. Section 3 shows how much of the corpus this governs.

Corpus. The evaluation uses a single real-world source: JADESTA (*Jaringan Desa Wisata*), the Indonesian Ministry of Tourism's national tourism-village portal. The snapshot comprises 21,461 records distributed across seven categories as shown in Table 1. The as-submitted version of this paper reported 6,568 nature attractions and a category total of 21,460 against a harness count of 21,461; the discrepancy was a line-counting artefact (the final record of the nature-attraction file carries no trailing newline, so a line count under-reports it by one), and the corrected count of 6,569 reconciles the table with the harness exactly. The corpus is sparse in precisely the fields entity resolution relies on: 4,429 records (20.6 per cent) carry coordinates, 5,131 (23.9 per cent) report a district, and only 493 (2.3 per cent) carry a non-empty category string, so the category term of (5) equals one for almost every pair.

Each record carries a fixed set of fields: name, category, description, administrative hierarchy, coordinates, facilities, photograph, price, contact details, and source URL. A published field allowlist determines which fields are ever normalised into a claim. Contact details are deliberately excluded from any claim or AI-facing surface and remain only in the immutable raw archive as an audit record of what was retrieved.

Table 1. Composition of the JADESTA snapshot used in this study.

Category	Records
Desa wisata rintisan (emerging village)	4,918
Desa wisata berkembang (developing village)	1,035
Desa wisata maju (advanced village)	338
Desa wisata mandiri (independent village)	38
Wisata alam (nature attraction)	6,569
Wisata budaya (cultural attraction)	5,469
Wisata buatan (man-made attraction)	3,094
Total	**21,461**

The proposed architecture is designed to be source-independent: any source that can produce (entity, attribute, value, timestamp) tuples can be reconciled through the same pipeline. The implementation and evaluation reported here use JADESTA as a single, representative operational dataset. This is a deliberate methodological choice. It is consistent with common practice in database-systems evaluation, where a proposed architecture is evaluated on one real-world dataset plus controlled synthetic scenarios rather than requiring simultaneous integration of multiple heterogeneous sources, and it allows the provenance, resolution, and confidence claims to be evaluated without conflating multi-source integration, which is an architectural capability, with data management quality, which is what is actually measured. Its cost is stated plainly in Section 3: with one

Copyright © 2026 – Indonesian Journal of Applied Mathematics
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

source and no verification channel, every claim is single-source and unverified, so the policy gate's behaviour in the mixed regime must be evaluated through a controlled synthetic second channel rather than through independent ground truth.

Pipeline. The pipeline processes a source through six stages, shown in Figure 1. A Bronze stage archives an immutable raw snapshot with a source hash and provenance tag. A Silver stage performs field-allowlist normalisation and entity resolution: candidate blocking by regency, weighted scoring by (5), and a merge, needs-review, or reject decision. An evidence ledger stage records each allowed field as one attribute-level, append-only, bitemporal claim. A confidence engine computes (2) per claim. A policy gate evaluates each claim against θ. A semantic view exposes only claims that cleared the gate, always carrying source, confidence, and verification status alongside the value. At the time of evaluation the Bronze stage is implemented as an object-store writer persisting hashed JSON rather than native open table formats; entity resolution is implemented for the tourism domain; the evidence ledger is implemented as a shared append-only pattern with concrete tables in two bounded contexts; the confidence engine is implemented as the rule-based function (2); and the policy gate and semantic view are implemented and exposed both per domain and through a single cross-domain endpoint sharing one filtering implementation, so that the two surfaces cannot disagree. The architecture is reported at its actual implementation status rather than at its long-run target status.

Compared systems and ablation.

To attribute any measured difference to architecture rather than to data or measurement differences, the proposed system is compared against two baselines that run on identical data and identical injected scenarios, and against four ablated variants of itself. Table 2 summarises them.



Figure 1. The six-stage evidence-first pipeline. New evidence producers (a community correction, a sensor reading, an extracted textual claim) enter through the same confidence engine and policy gate as every other source, so the gate cannot be bypassed by adding an evidence channel.

B1 exists specifically so that the statement "P1 is better than B0" can be decomposed: without B1, an improvement could be attributable merely to having any history, which B1 also has, rather than to attribute-level granularity, confidence, and gating specifically. The ablated variants decompose P1 itself. Attribute-level provenance cannot be switched off independently, because confidence and gating are defined over claims and have no meaning without them; the ablation therefore holds the claim ledger fixed and toggles the two mechanisms built on it.

Table 2. The compared systems. ER denotes blocking-and-scoring entity resolution; the alternative is the naive natural-key match (lower-cased name plus regency) that B0 and B1 perform.

System	History granularity	Entity resolution	Confidence and gate	Role in the study
B0	None (overwrite)	Naive key	None	Lower bound: no evidence discipline at all
B1	Per record	Naive key	None	Isolates the contribution of having history at all
P1-ER-G	Per attribute	Naive key	None	Claim ledger alone
P1-G	Per attribute	ER	None	Adds entity resolution to the ledger
P1-ER	Per attribute	Naive key	Yes	Adds confidence and gate to the ledger
P1	Per attribute	ER	Yes	Proposed system

Neither baseline is a strawman: B1 in particular reflects the audit-table and soft-versioning patterns genuinely used in production ETL systems and codified in dimensional warehouse practice [32]. Each system is materialised on its own branch of the same locked snapshot, so comparisons and the replay test both operate over data that is provably identical at the point of divergence.

Failure-scenario injection. The six failure modes are injected identically into the same locked snapshot before any system processes it, using deterministic, seeded injection functions, so that scenario identity is guaranteed across systems. Injecting controlled corruption into a real corpus rests on three converging methodological traditions. First, data-quality dimension theory selects *which* failure modes to inject: each of the six instantiates one of Wang and Strong's dimensions [16]. Second, controlled error generation over a real corpus is an established data-cleaning and record-linkage benchmarking method: BART formalises exactly this pattern, injecting known, parameterised errors into a cleaner real database so that ground truth is known by construction [27], and record-linkage methodology applies the same principle to duplicate generation [22]. The injection functions used here are direct implementations of this pattern against real records, never synthetic-from-scratch records, and always retain the original record identity for ground-truth comparison, inheriting the formalisation of record linkage as a precision-and-recall problem against known match status [21]. Third, mutation testing supplies the statistical logic for the injection *count*: a controlled number of precisely known faults is injected and the fraction caught is measured, which characterises detection capability rather than real-world fault frequency [28]. Twenty instances per scenario are injected, seeded for reproducibility; the number is not a claim about the prevalence of duplicates, conflicts, or missing values in JADESTA. Staleness is evaluated differently from the other five modes, and the difference is deliberate: because staleness is a continuous function of elapsed time rather than a discrete corruption of content, it is tested as a controlled parametric sweep of the freshness input of (2) with every other input held constant.

Gold sample. Entity-resolution quality is evaluated against a manually labelled gold sample. From the full corpus, 425 candidate record pairs were selected by stratified sampling: 150 scoring at or above

Copyright © 2026 – Indonesian Journal of Applied Mathematics

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

the merge threshold, 200 in the review zone, and 75 below the reject threshold. Of these, 304 pairs were labelled by two independent annotators as referring to the same or a different real-world entity, with a third annotator resolving disagreements; 49 pairs, or 16 per cent, required such resolution. Inter-annotator agreement was substantial in the conventional interpretation, with Cohen's $\kappa = 0.742$ over the 302 pairs carrying both annotators' labels [29], [30]. Labelling covered the merge-threshold and review-zone strata at 100 per cent and 77 per cent respectively; the reject-threshold stratum was not labelled, so entity-resolution quality is scoped to the merge decision, and the reject threshold is examined through sensitivity analysis and a census of the full candidate space rather than through labelled evidence.

Sensitivity analyses and synthetic second channel. Three analyses added in this revision use no data beyond the gold sample and the locked snapshot. The merge threshold is swept over [0.30,1.00] in steps of 0.025 and precision, recall, and F1 recomputed at each value; the reject threshold is swept over the same grid, counting the labelled SAME pairs that would be auto-rejected and the DIFFERENT pairs that would correctly leave the review queue; each weight of (5) is perturbed by ±0.10 and renormalised, each signal is removed in turn, and a coarse grid search reports the in-sample optimum, which is reported for comparison and explicitly not adopted. Every within-regency candidate pair in the corpus (1,055,971 pairs across 484 blocks) is scored and counted per decision zone and per structural case. For the policy gate, a seeded sample of 200 current claims per group is drawn from P1's output on the real corpus and each group receives a controlled second evidence channel: one, two, or three independent field-survey corroborations, a steward verification following the production confirmation path exactly, three corroborations plus one conflicting correction, or the maximal evidence configuration plus a conflict. Ground truth for "should be admitted" is fixed by construction (three independent corroborations or a steward verification should admit; fewer, or a conflict, should not), and the gate's leakage and wrongful-withholding rates are reported for every threshold from 0.40 to 0.80.

Metrics. Six metric families are used. Provenance coverage is the proportion of a canonical entity's attributes backed by at least one claim with source and retrieval time populated. Provenance precision is defined uniformly for every system with a history: among attribute-level stamps attached to a non-initial version (B1) or a superseding claim (P1 variants), the fraction whose value differs from the immediately preceding stamped value of the same entity-attribute; it is undefined for B0. The phantom-stamp rate, introduced in this revision, is the fraction of the same stamps that attribute a value to a source record which did not itself report that attribute; it is zero by construction for every claim-per-attribute variant and is measured for B1 by replaying records one at a time. Entity-resolution precision, recall, and F1 on the merge decision, together with pair-level false-merge and false-split rates, are computed against the gold sample for P1's scorer and for the naive key match. The unverified leakage rate is the proportion of currently published attributes that are unverified and should have been withheld; in the single-source corpus this coincides with the proportion of published unverified claims, and in the synthetic experiment it is computed against the constructed ground truth. Replay equivalence is field-by-field identity between a re-run and the originally recorded state. Throughput and latency are measured per system on the full corpus and on a fixed 500-record subset repeated twenty times for stable percentile estimates.

Proportion metrics are reported with bootstrap 95 per cent confidence intervals over 1,000 resamples except where noted, and two systems are described as differing only when their intervals do not overlap [31]. Throughout Section 3 each result is labelled as one of four kinds: *definitional* (a consequence of a system's schema that no data could change), *analytical* (derived from the equations of Section 2), *empirical* (measured on the real corpus or the gold sample), or *synthetic* (measured under constructed evidence or injected corruption). Table 7 collects the labels.

All measurements were produced on a single machine with an eight-core Apple silicon processor running Python 3.13, from a pinned commit of the experiment code, with the run timestamp and hardware

descriptor recorded in each result file. The as-submitted benchmark harness looked claims up by scanning the claim list linearly; profiling on a 3,000-record subset attributed 93 per cent of P1's wall time to those two scans, which grow quadratically in the number of claims and have no counterpart in a relational ledger that indexes claims by entity and attribute. The revised harness indexes claims by (entity, attribute); on the 3,000-record subset it produces a canonical state identical field by field to the unindexed harness in 0.83 seconds instead of 25.15. All performance figures in Section 3 are from the indexed harness, and the as-submitted figures are reported alongside them where the comparison matters.

Results and Discussion

Provenance. In the baseline scenario the three systems separate on provenance coverage exactly as their schemas dictate, and this part of the result is definitional: B0's coverage is 0.000 because no source column exists, and both B1 and P1 report 1.000 because every stored version or claim carries a source and a retrieval timestamp. Read alone, that pair of identical values would place B1 and P1 on equal footing, and the as-submitted version of this paper separated them with a provenance-precision metric computed for B1 only. A reviewer asked that the metric be defined across all systems, and doing so changes the finding. Under the uniform definition, B1's precision is 0.471 (3,417 of 7,249 stamps attached to non-initial versions carry a changed value) and the claim ledger without entity resolution, P1-ER, scores 0.478 (3,189 of 6,667), which is the same number within sampling noise. The full P1 scores 0.380 (615 of 1,620), lower rather than higher, because entity resolution merges fewer records than the naive key does (219 against 1,028) and the supersessions that remain are dominated by duplicate registrations re-asserting identical values. Value-change precision, then, measures how often this corpus re-asserts a value it already holds; it does not measure the architecture.

What does measure the architecture is the phantom-stamp rate. Of B1's 7,249 non-initial stamps, 354 (4.9 per cent) attribute a value to a source record that never reported the attribute, because B1 copies every attribute of the previous version forward and stamps all of them with the new record's source and time. No claim-per-attribute variant can produce such a stamp, because a claim is only ever created from an attribute the record actually reported; the rate is zero by construction, and that is the sense in which the as-submitted paper's phrase "phantom provenance" survives this revision, narrowed from half of B1's stamps to one in twenty. The remaining difference is semantic rather than numeric. When a duplicate registration re-asserts a village's name, B1 stores a second stamp indistinguishable from a change; P1 stores a second claim whose evidence event is a confirmation, counts it in the corroboration term of (2), and leaves the first claim's source and time intact. Figure 2 shows all three quantities for the four systems that have a history.

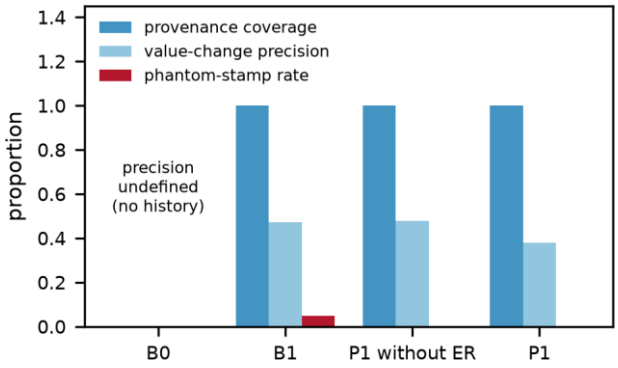


Figure 2. Provenance coverage (definitional), value-change precision (empirical, defined uniformly), and phantom-stamp rate for B1, the claim ledger without entity resolution, and P1. Precision is undefined for B0. Coverage confidence intervals over a 500-record subset with 200 resamples are degenerate for every system because the outcome is fixed by schema.

The lesson generalises beyond this corpus but is narrower than first stated. Data platforms routinely report lineage coverage as a governance indicator, and a pipeline that rewrites every row on every load reports perfect lineage. Value-change precision, which the as-submitted paper proposed as the corrective, turns out to be a property of the source's re-assertion behaviour and

Copyright © 2026 – Indonesian Journal of Applied Mathematics
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

is nearly identical for record-level and attribute-level designs. The quantity that discriminates is whether a stamp can name a source that never spoke, and any platform claiming lineage completeness can report it wherever the set of attributes each load actually carried is retained.

Behaviour under injected failures. The behaviour of the three systems under injected scenarios is shown in Figure 3 and Table 3; these results are synthetic by construction. B0 and B1 produce identical entity counts in every scenario, since both rely on the same exact-key match, and both are entirely unchanged by the missing and schema-drift scenarios, because neither detects nor reacts to those failure modes at all. P1's entity count rises by twenty under the conflict and late-update scenarios because the injected records introduce attribute values that fail to corroborate an existing entity strongly enough to auto-merge, and therefore contribute to the review queue rather than being silently absorbed.

Table 3. Canonical entities produced and pairs deferred to review, by scenario, on the full corpus.

Scenario	B0 entities	B1 entities	P1 entities	P1 pairs deferred
baseline	20,433	20,433	21,242	8,324
duplicate	20,453	20,453	21,254	8,336
conflict	20,433	20,433	21,262	8,344
missing	20,433	20,433	21,242	8,324
schema drift	20,433	20,433	21,242	8,324
late update	20,433	20,433	21,262	8,344

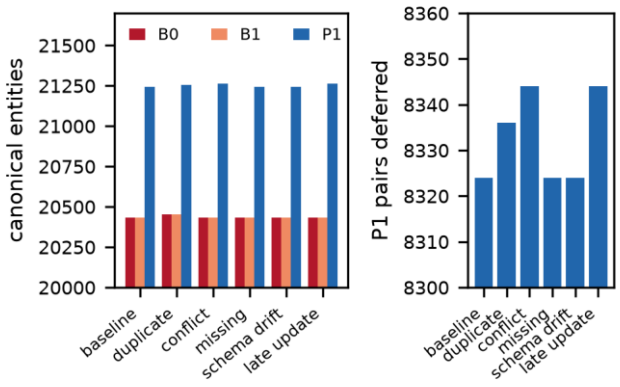


Figure 3. Canonical entity counts under each injected scenario (left) and the size of P1's review queue (right). The baselines are flat across scenarios they cannot detect; P1's response to conflict and late update is visible as a shift into the review queue rather than as a change in silently published state.

Reproducibility was tested by re-running the pipeline on an identical input. On the 500-record subset the two runs completed in 0.030 and 0.029 seconds and on the full corpus in 9.70 and 10.62 seconds, producing in both cases a canonical state identical field by field, over 149,926 attribute-level claims at full scale. This establishes in-process determinism, the operational form of replay equivalence available without a persisted snapshot store, and is distinguished below from full persisted-snapshot replay.

Entity resolution: the naive baseline and the merge decision. On the full corpus, exact-key matching collapsed the corpus to 20,433 entities, while P1's entity resolution produced 21,242, more rather than fewer. The reading, once the review mechanism is accounted for, is a precision-and-recall trade-off: P1 deferred 8,324 candidate pairs rather than merging them outright, most commonly because one or both sides lacked coordinates or a district. The gold sample makes the trade-off measurable, and the ablation makes the comparison direct, because the naive key can be scored on the same 304 pairs. Exact-key matching merges 136 of them and achieves a precision of 0.882, a recall of 0.714, and an F1 of 0.789; P1's scorer achieves a precision of 0.900, a recall of 0.804, and an F1 of 0.849 with a bootstrap 95 per cent interval of [0.804, 0.887]. Entity resolution therefore contributes 0.060 of F1, almost

Original Article

entirely through recall, by merging 15 true matches whose names differ in spacing, case, or a token that the exact key cannot see. The pair-level false-merge rate is 0.100 and the false-split rate 0.196. Figure 4 shows the decision breakdown and the comparison.

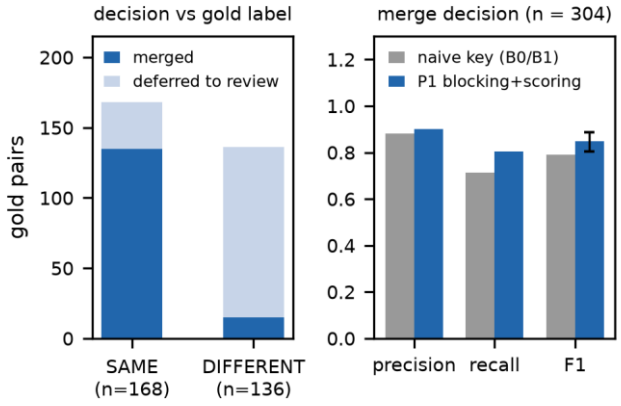


Figure 4. Left: P1's decision against each gold label; of 168 SAME pairs, 135 are merged and 33 deferred, and of 136 DIFFERENT pairs, 121 are deferred and 15 merged. No labelled pair reached the reject branch, consistent with the reject stratum being unlabelled. Right: precision, recall, and F1 on the merge decision for the naive key the baselines use and for P1's scorer, with the bootstrap interval on P1's F1.

The decision breakdown is more informative than the aggregate. Of 168 pairs labelled as the same entity, P1 correctly merged 135 and deferred 33 rather than rejecting any outright. Of 136 pairs labelled as different, P1 correctly withheld 121 and merged 15. Of the 154 pairs P1 deferred, 121, or 78.6 per cent, are genuinely distinct entities, so in roughly four cases out of five deferral protects the canonical entity set from a merge that would have been wrong; the remaining 21.4 per cent are the concrete source of the 0.804 recall.

Entity resolution: decision geometry and sensitivity. The reviewer's request for a sensitivity analysis of both thresholds is best answered by first stating what the score can take as values. Proposition 2 and the corpus's sparsity together quantise the score. Because the category string is empty in 97.7 per cent of records, $C = 1$ for almost every pair; because a district is reported in 23.9 per cent of records, $A = 0.5$ for most pairs; and because coordinates are present in 20.6 per cent of records, the geospatial term is absent for most pairs. The census of all 1,055,971 within-regency pairs shows the consequence: 980,163 pairs (92.8 per cent) have no coordinates on at least one side and an unknown district on at least one side, so their score is exactly $0.25 + 0.40\,J$. Such a pair can never merge, since that would require $J = 1.25$; it is rejected when $J \leq 0.25$ and deferred when $J > 0.25$. Across the whole candidate space 1,028,276 pairs (97.4 per cent) are rejected, 27,456 (2.6 per cent) are deferred, and 239 merge, of which 168 have no coordinates and therefore merge only by satisfying Proposition 2 exactly. The gold sample reflects the same quantisation: 121 of its 304 pairs score exactly 0.75, the coordinate-less ceiling, and the lowest score of any SAME pair is 0.359.

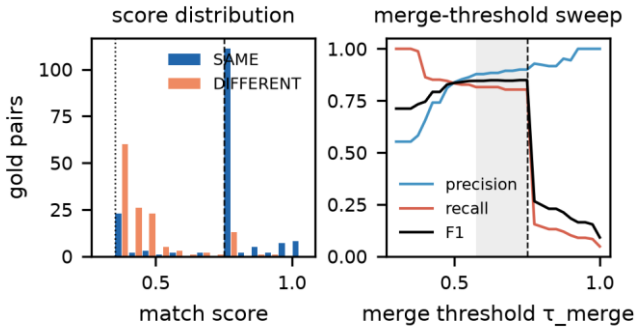


Figure 5. Left: score distribution of SAME and DIFFERENT gold pairs; the dotted and dashed lines mark the reject and merge thresholds, and the spike at 0.75 is the coordinate-less ceiling of Proposition 2. Right: precision, recall, and F1 on the merge decision as the merge threshold is swept; the shaded band marks the plateau on which F1 lies within 0.004 of its design-threshold value.

Figure 5 shows what this geometry does to the merge threshold. F1 is flat at 0.845–0.849 for every threshold between 0.55 and 0.75, rises slightly from 0.836 at 0.50, and collapses to 0.265 at 0.775 because the 0.75 atom, which holds 111 SAME pairs, moves wholesale from merge to review the moment the threshold exceeds it. The design threshold therefore sits at the upper edge of a wide plateau, which is the right place to be when the cost of a false merge exceeds the cost of a deferral, and the merge decision is robust to any threshold choice within the plateau. The weight

Copyright © 2026 – Indonesian Journal of Applied Mathematics

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

sensitivity shows the same structure from the other side (Table 4). Every perturbation that keeps $w_N + w_A + w_C \geq \tau_{\text{merge}}$ leaves F1 between 0.838 and 0.849; every perturbation that pushes the sum below the threshold, including removing name, district, or category, or shifting 0.10 of weight toward the geospatial term, collapses F1 to between 0.246 and 0.274, because the coordinate-less pairs can no longer reach the merge threshold at all. Removing the geospatial term entirely costs 0.005 of F1, which is what one expects when four pairs in five cannot use it. A coarse in-sample grid search finds a maximum F1 of 0.862 at weights (0.40, 0.20, 0.40, 0.00); the gain of 0.013 lies inside the bootstrap interval of the design weights and is reported only to show how far the heuristic sits from the empirical optimum, not adopted, since there is no held-out set.

Table 4. Weight sensitivity on the merge decision, 304 gold pairs. Perturbed weights are renormalised to sum to one. Σ is $w_N + w_A + w_C$ after renormalisation, the quantity Proposition 2 compares with the merge threshold; P and R are precision and recall.

Configuration	w_N	w_A	w_C	w_G	Σ	P	R	F1
Design weights	0.400	0.200	0.150	0.250	0.750	0.900	0.804	0.849
Name +0.10	0.455	0.182	0.136	0.227	0.773	0.899	0.798	0.845
Name −0.10	0.333	0.222	0.167	0.278	0.722	0.931	0.161	0.274
District +0.10	0.364	0.273	0.136	0.227	0.773	0.898	0.786	0.838
District −0.10	0.444	0.111	0.167	0.278	0.722	0.931	0.161	0.274
Category +0.10	0.364	0.182	0.227	0.227	0.773	0.900	0.804	0.849
Category −0.10	0.444	0.222	0.056	0.278	0.722	0.929	0.155	0.265
Geospatial +0.10	0.364	0.182	0.136	0.318	0.682	0.931	0.161	0.274
Geospatial −0.10	0.444	0.222	0.167	0.167	0.833	0.899	0.798	0.845
Without geospatial	0.533	0.267	0.200	0.000	1.000	0.888	0.804	0.844
Without name	0.000	0.333	0.250	0.417	0.583	0.889	0.143	0.246
In-sample grid optimum	0.400	0.200	0.400	0.000	1.000	0.892	0.833	0.862

The reject threshold cannot be validated by the gold sample, because no labelled pair scores below 0.35, and the as-submitted paper said so. The sweep in Figure 6 nevertheless bounds it from above. Raising the reject threshold to 0.375 would auto-reject 2 labelled SAME pairs, irrecoverably, in exchange for removing 17 DIFFERENT pairs from the review queue; raising it to 0.40 would lose 23 SAME pairs for 60. The SAME-pair density immediately above the design threshold is therefore substantial, 23 of the 168 SAME pairs lie in [0.35, 0.40), and nothing in the labelled data suggests that density falls to zero at 0.35 rather than continuing below it. The honest conclusion is twofold: the reject threshold should not be raised, and it may already be too high, which only labelling the 75 reject-stratum pairs already selected and blinded for this study can settle. Below 0.35 the census shows that 97.4 per cent of all candidate pairs are auto-rejected, so the review queue is already carrying the ambiguity the architecture can afford to carry.

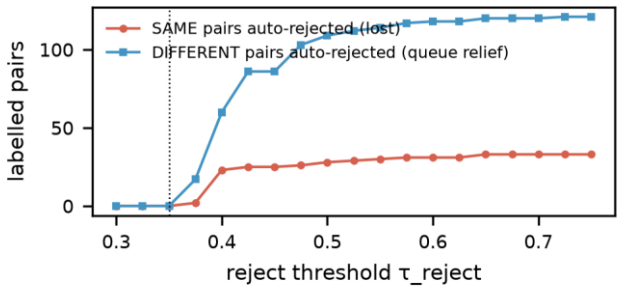


Figure 6. Reject-threshold sweep over the labelled pairs. Each point counts the SAME pairs that would be auto-rejected and lost, and the DIFFERENT pairs that would correctly leave the review queue, if the reject threshold were raised to that value. The dotted line marks the design value, below which no labelled pair lies.

Entity resolution: the error populations. Fifteen false merges and thirty-three deferred true matches are small enough populations to inspect individually, and the inspection reclassifies both. All fifteen false merges share the same structure (Table 5): fourteen have token-identical names, fourteen report the same known district, all fifteen have empty categories, and thirteen score exactly 0.75 because they have no coordinates and satisfy Proposition 2. Twelve of the fourteen same-name pairs are, by their JADESTA identifiers, re-registrations of one village name under distinct portal slugs that differ only by a numeric suffix, and in one of them the two street addresses are character-identical. The two remaining pairs carry coordinates at zero distance from each other. The annotators labelled all fifteen as different on the strength of differing street-address text, and for most pairs that judgement is defensible; for the character-identical pair it is not, and the merge decision's precision figure should be read as a lower bound that includes some label noise. More importantly, the class is homogeneous: a same-name, same-district, coordinate-less pair is exactly the case the current signals cannot separate, and the remedy is a signal the scorer does not yet have, such as address-text similarity or the portal identifier, not a different threshold.

Table 5. Structural inspection of the two error populations on the merge decision.

Property	False merges (n = 15)	Deferred true matches (n = 33)
Token-identical names	14	3
Same known district	14	0
District unknown on one side	1	32
Category empty on both sides	15	33
Identical street-address text	1	21
Same name, distinct JADESTA slug suffix	12	3
Both have coordinates	2	0
Identical coordinates (0 km)	2	0
Score exactly 0.75	13	0
Score in [0.35, 0.40)	0	23

The deferred true matches are the mirror image. Thirty-two of the thirty-three lack a district on one side and none carries coordinates, so by Proposition 2 none could have merged whatever the threshold; twenty of them score exactly 0.383, which is $0.25 + 0.40 \cdot 1/3$, a single shared name token. Twenty-one of the thirty-three have character-identical address text, and reading the records shows why: they are attraction-level entries listed under one village's address with distinct activity names, such as two traditional dances or two food products at the same cultural village. The annotators judged them the same place; the scorer, which sees only the names, judged them one shared token apart. These are not errors of calibration but of signal: the address string, which the scorer does not consume, carries the evidence the annotators used. The review queue is

Copyright © 2026 – Indonesian Journal of Applied Mathematics
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

therefore doing what it was designed to do with both populations, holding for a human the pairs on which the available signals are silent.

Policy gate: the single-source bound. Both baselines exhibit an unverified leakage rate of 1.000 by construction, since neither implements any policy-gate concept and every value is live the instant it is written. P1's measured leakage rate on the real corpus is 0.000: every one of its 148,306 current claims derives from a single government portal, is unverified, and receives the outcome insufficient-evidence. The as-submitted paper reported this as a measurement and a reviewer observed, correctly, that it mainly shows the gate working as designed. This revision goes further and states that it could not have been otherwise. By Proposition 1 the confidence of any claim whose evidence consists solely of automated single-source assertions is at most 0.5375, and at most 0.3375 without corroboration, against a threshold of 0.60; the leakage rate of zero is an analytical consequence of (2), and the bootstrap intervals of [0.000, 0.000] for P1 and [1.000, 1.000] for both baselines are degenerate for the same reason.

Table 6 tabulates the supremum for the evidence configurations that matter operationally, each obtained by substituting into (2) with freshness at its maximum. Two configurations clear the threshold: three independent field-survey corroborations, at 0.607, and a steward verification following the production path, at 0.605. Three observations follow. The steward's verification clears the gate by 0.005, so a threshold of 0.625 would withhold the one channel designed to admit a value on human authority. Two field surveys reach 0.540, so a threshold at or below that value would admit claims on two voices rather than three. And a community contributor corroborated three times reaches only 0.5125, so community corroboration alone can never publish a value, which is the intended division of authority between contributors and stewards. The admissible window for θ, the set of thresholds that admit exactly the configurations designed to be admitted, is therefore the interval (0.5403, 0.605]; the design threshold lies inside it, 0.06 above its lower edge and 0.005 below its upper edge. A reader who finds that margin uncomfortably small has understood the result: the gate is correct for the present parameters and has little room, and the ordering of the admissible configurations, not the absolute threshold, is what the parameters of (2) encode.

Table 6. Confidence supremum under each evidence configuration, from (2) with freshness at its maximum, and whether it clears the tourism threshold of 0.60. Field surveys use the operational reliability of 0.70 and the community authorisation bonus; the steward configuration follows the production verification path exactly.

Evidence configuration	λ^*	Clears 0.60
Government portal, single assertion	0.3375	no
Government portal, three same-source re-assertions	0.5375	no
Portal claim plus one field survey	0.4737	no
Portal claim plus two field surveys	0.5403	no
Portal claim plus three field surveys	0.6070	yes
Steward verification (reliability 1.0, level 2)	0.6050	yes
Community contributor corroborated three times	0.5125	no
Three field surveys plus one conflicting correction	0.4570	no (insufficient evidence)
Maximal evidence plus one conflicting correction	0.6550	no (deny)

Policy gate: the synthetic second channel. The reviewer asked for an evaluation using verified and unverified data from independent sources, or a clear statement of why that is not possible. It is not possible

Original Article

on this corpus, which has one source and no verification events, and the paper does not pretend otherwise. What it can do is show the gate operating in the mixed regime under evidence whose correctness is fixed by construction, which is the mutation-testing logic already used for the failure scenarios. Two hundred real current claims per group were drawn from P1's output and given a controlled second channel as described in Section 2. At the design threshold the gate admits exactly the two groups designed to be admitted and withholds the other five, so leakage and wrongful withholding are both zero; the group with three corroborations and one conflict is reported as insufficient evidence, because the conflict penalty lowers its confidence to 0.457, and the maximal-evidence group with a conflict is the only one reported as deny, at 0.655. Figure 7 shows the sweep. Below 0.540 the gate admits claims corroborated twice, and below 0.474 claims corroborated once, so leakage against the constructed ground truth rises to one group in seven and then two; above 0.605 it withholds the steward's verification and the three-survey corroboration, so wrongful withholding rises to two groups in seven. The window in which both rates are zero is exactly the analytical interval of Table 6, which is the sense in which the synthetic experiment confirms the analysis rather than adding to it.

What the result demonstrates is that the mechanism is wired correctly, separates evidence configurations in the intended order, and does so only within a narrow and now explicitly stated range of thresholds. What it does not demonstrate is that the confidence function is a calibrated probability estimator, or that three field surveys are the right number, or that the steward's margin of 0.005 is adequate. Establishing those requires independent field verification of individual claims, which is a primary data-collection effort rather than a labelling task, and which a single-source dataset cannot provide by construction; confirming a claim against the same source it was derived from would be circular.

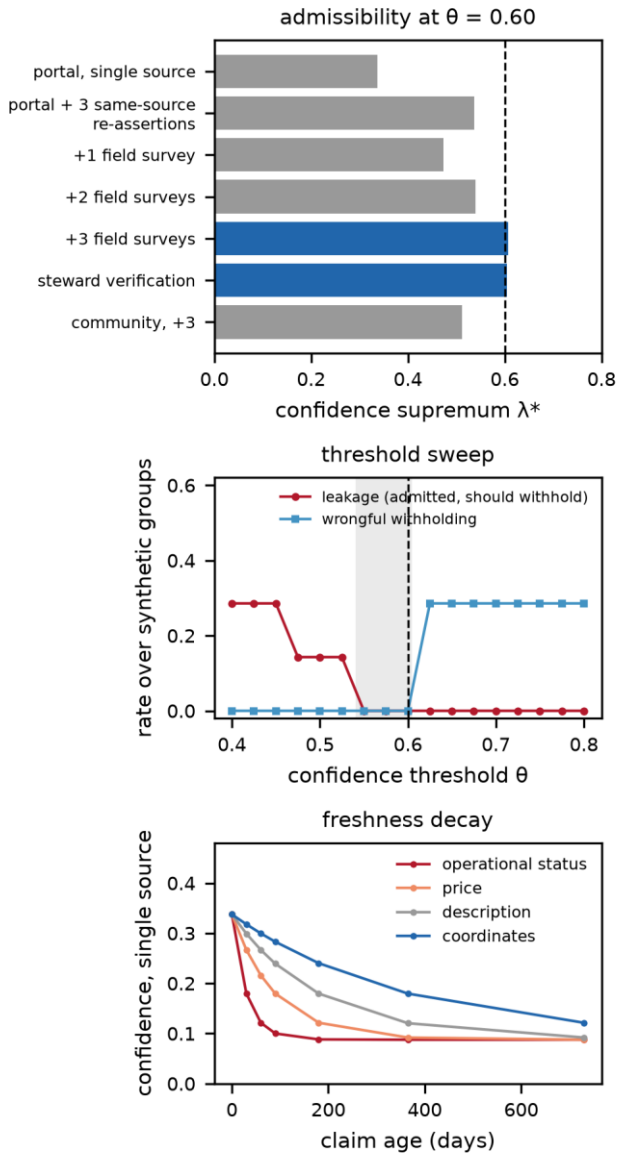


Figure 7. Top: confidence supremum per evidence configuration (Table 6) against the design threshold; filled bars clear it. Middle: leakage and wrongful-withholding rates over the seven synthetic evidence groups as the threshold is swept; the shaded band is the analytical admissible window (0.540, 0.605]. Bottom: freshness decay of a single-source claim's confidence for four attribute half-lives; the floor of 0.0875 is the source-reliability term alone.

Expected calibration error, Brier score, and reliability-diagram computation are implemented and unit-tested in the experiment code and remain available for a

Copyright © 2026 – Indonesian Journal of Applied Mathematics
Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

follow-up study with access to independent verification data. The freshness sweep in the bottom panel of Figure 7 shows the stale-claim behaviour the design intends: an operational-status claim loses half its freshness term in 30 days and reaches its floor by 180, while a coordinate claim retains most of it for a year.

Component ablation. Table 7 decomposes P1 by component on the full corpus and labels every result by kind. Three readings follow. First, the claim ledger alone (P1-ER-G) already differs from B1 in the only provenance quantity that discriminates, the phantom-stamp rate, and in nothing else: its entity count, leakage, and value-change precision are those of B1, which is what one expects of a storage model that changes what is recorded but not what is decided. Second, entity resolution and the gate act on disjoint metrics. Adding entity resolution (P1-G against P1-ER-G) changes the entity count from 20,433 to 21,242, creates the 8,324-pair review queue, and raises gold-sample F1 from 0.789 to 0.849, while leaving leakage at one; adding the gate (P1-ER against P1-ER-G) sets leakage to zero while leaving every entity-resolution metric untouched. Neither component substitutes for the other, and the full system is their sum. Third, the costs are not symmetric. On the full corpus the ledger alone runs in 1.08 seconds, the gate adds 0.43 seconds, and entity resolution adds 8.6 seconds, which is 87 per cent of P1's 9.9 seconds; the quadratic cost the as-submitted paper attributed to entity resolution is real, but it is an order of magnitude smaller than reported, for the reason given next.

Table 7. Component ablation on the full 21,461-record corpus, single run, with each result labelled by kind: D definitional, E empirical, A analytical, S synthetic. Gold-sample F1 is on the merge decision over 304 pairs; the naive-key figure applies to every variant without entity resolution.

Metric	Kind	B0	B1	P1-ER-G	P1-G	P1-ER	P1
Provenance coverage	D	0.000	1.000	1.000	1.000	1.000	1.000
Value-change precision	E	—	0.471	0.478	0.380	0.478	0.380
Phantom-stamp rate	D/E	—	0.049	0	0	0	0
Canonical entities	E	20,433	20,433	20,433	21,242	20,433	21,242
Pairs deferred to review	E	—	—	0	8,324	0	8,324
Gold-sample F1 (merge)	E	0.789	0.789	0.789	0.849	0.789	0.849
Unverified leakage	D/A	1.000	1.000	1.000	1.000	0.000	0.000
Wall time, full corpus (s)	E	0.13	0.52	1.08	9.72	1.51	9.89

Performance. The quality properties reported above are not free, and the as-submitted paper overstated by how much. It reported P1 at 559.3 seconds for the full corpus, approximately 5,447 times slower than B0, and attributed the cost to quadratic per-block entity resolution. Profiling the harness on a 3,000-record subset showed that 93 per cent of P1's time was spent in two linear scans over the claim list that located the current and historical claims of an entity-attribute pair, an $O(C^2)$ cost in the number of claims that exists only because the in-memory harness had no index, and that any relational ledger removes with an index on (entity, attribute). Entity resolution accounted for under 4 per cent. The revised harness indexes claims by (entity, attribute); it produces a canonical state identical field by field to the unindexed harness and is reported here. The correction does not change any quality result, since no decision changed, and it changes the performance narrative from one of infeasibility to one of a measurable, attributable cost.

On the fixed 500-record subset over twenty repeated runs, B0 processes approximately 183,606

records per second, B1 approximately 106,402, and P1 approximately 17,184. On the full corpus in a single clean run, B0 processes approximately 205,209 records per second, B1 approximately 54,231, and P1 approximately 2,368, a wall-clock time of 9.06 seconds for the complete national corpus. Table 8 and Figure 8 report both scales.

Table 8. Throughput (records per second) and wall time at two corpus scales with the indexed harness. Subset figures (500 records) are the median of twenty repeats; full-corpus figures (21,461 records) are a single clean run. The as-submitted, unindexed full-corpus P1 figure was 38.4 records per second (559.3 seconds).

System	Subset rec/s	Subset p50 (s)	Subset p95 (s)	Full rec/s	Full wall (s)
B0	183,605.5	0.002723	0.002915	205,209.4	0.105
B1	106,401.8	0.004699	0.004927	54,231.3	0.396
P1	17,183.5	0.029098	0.032280	2,368.3	9.062

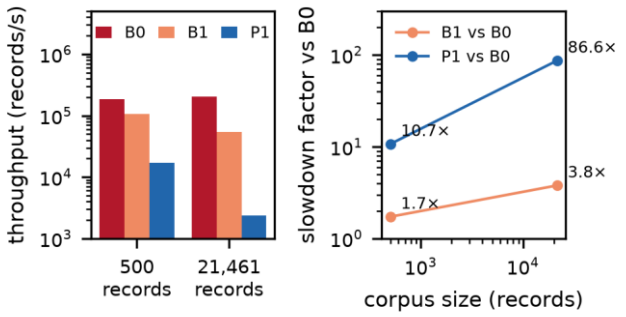


Figure 8. Left: throughput at both corpus scales on a logarithmic axis. Right: the slowdown factor relative to B0 as a function of corpus size; the gap between the record-level baseline and the evidence-first system widens with scale.

The headline ratio is still not the most important number in the table; the important observation is that the ratio is not constant. P1 is 10.7 times slower than B0 at 500 records and 86.6 times slower on the full corpus: a 43-fold increase in data volume corresponds to an 8-fold increase in the slowdown factor. B1 moves only from 1.7 to 3.8 times slower over the same range. The ablation attributes 87 per cent of P1's full-corpus time to entity resolution, whose candidate scan is linear in the number of existing entities per record and therefore quadratic overall, with blocking applied as a filter inside the scan rather than as an index over it. This remains the study's most actionable engineering finding, now at its correct magnitude: the current strategy will not scale gracefully to a materially larger or multi-source corpus without an indexed block structure, finer-grained blocking keys, or an approximate-nearest-neighbour geospatial join, which is precisely the cost the blocking and filtering literature exists to reduce [33]. Two secondary costs are reassuring (Figure 9).

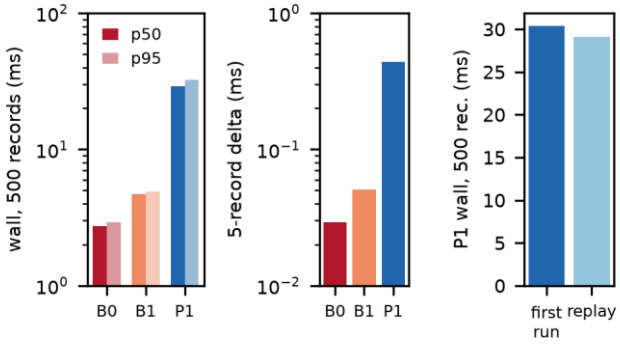


Figure 9. Left: median and ninety-fifth-percentile wall time over twenty repeated runs of the 500-record subset. Centre: incremental cost of a five-record delta. Right: cost of an exact replay on the subset.

Latency is stable rather than heavy-tailed: the ratio between the ninety-fifth and fiftieth percentile wall time is approximately 1.07 for B0, 1.05 for B1, and 1.11 for P1. And incremental work is proportionate: adding five records to an existing base of 495 costs P1 approximately 0.44 milliseconds against 0.03 and 0.05 milliseconds for B0 and B1, which matters more than any ratio for deployment planning, since community corrections arrive as small deltas rather than as full reloads. Replay is likewise inexpensive: a repeated run on the same input completed in 0.029 seconds against the original 0.030 on the subset, and in 10.6 seconds against 9.7 on the full corpus, with an identical resulting state, so exact

Copyright © 2026 – Indonesian Journal of Applied Mathematics

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

reproducibility costs approximately nothing beyond the cost of the original run.

Figure 10 places the quality and performance results on a single plane, now including the ablated variants. The baselines and the two ungated variants cluster at total leakage; the two gated variants sit at zero leakage, separated from each other by the cost of entity resolution. The figure is presented so that a reader can weigh the trade-off against their own deployment constraints: a national portal refreshed nightly and a real-time recommendation service have very different tolerances for a ten-second batch, and a reader who needs the gate but not the resolver can see what each costs alone.

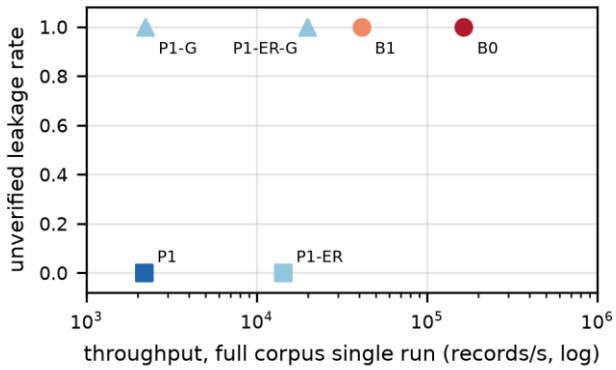


Figure 10. Quality against performance for the baselines, the proposed system, and its ablated variants. Throughput is a single full-corpus run on a logarithmic axis; the vertical axis is the unverified leakage rate on the baseline scenario.

Threats to validity. Several threats condition how far these results extend. The most consequential concerns construct validity around confidence. The confidence function is now disclosed, and its bounds and threshold window are derived, but it remains a design hypothesis: nothing here establishes that a value of 0.607 corresponds to any probability of correctness, or that three corroborations is the right requirement. That determination requires independent field verification and is deferred rather than claimed. A second construct-validity threat concerns the reject threshold, which the sweep bounds from above but which no labelled pair tests from below; the entity-resolution claim is stated as covering the merge decision, and the inspection of the false merges shows that the merge precision figure absorbs some label noise on same-name re-registrations. A third concerns the gold labels themselves: inter-annotator agreement of 0.742 is substantial but not perfect, and the two error populations are concentrated exactly where the annotators had information the scorer does not.

On internal validity, two safeguards were applied. Gold-sample annotators were not the individuals verifying claims in the production path, to avoid circular validation, and the entity-resolution and confidence logic never consumes gold-sample labels as pipeline input. The ablation holds the snapshot and the injection seeds fixed across variants, and the indexed harness was verified against the unindexed one before any revised figure was recorded. On external validity, the single-source constraint is the binding one. Results generalise to this architecture's behaviour on JADESTA-shaped data, in which four records in five lack coordinates and category is almost never populated; the decision geometry of Section 3 is a property of that sparsity as much as of the scorer, and a richer corpus would populate the geospatial term and loosen the quantisation. Multi-source integration is an architectural capability rather than an evaluated claim, and the synthetic second channel stands in for, rather than replaces, an independent one. Performance results were measured at the corpus's actual scale and on a 500-record subset, on a single machine, with no larger-scale stress test; the scale-dependent slowdown is itself evidence that these numbers must not be linearly extrapolated. Finally, the replay result establishes in-process determinism rather than full persisted-snapshot replay, in which a state is reconstructed from a durable event log after a process restart; the formal criterion admits both readings, and only the weaker one is evidenced here.

Conclusion

This study proposed and evaluated a database-driven, evidence-first data management architecture for tourism-village data, compared against two realistic baselines and four ablated variants on a national-scale, real-world corpus plus controlled synthetic failure scenarios and a manually labelled gold sample.

Original Article

Five findings emerge, and the revision has sharpened which kind of finding each is. Attribute-level claims and record-level lineage re-stamp unchanged values at the same rate, so value-change precision does not separate them; what separates them is the 4.9 per cent of record-level stamps that attribute a value to a source which never reported it, which the claim model excludes by construction. Entity resolution integrated with the evidence ledger raises merge F1 from 0.789 under exact-key matching to 0.849, almost entirely through recall, and its conservatism is justified in 78.6 per cent of the pairs it defers; its behaviour on this corpus is governed by one analytical identity, that the non-geospatial weights sum to the merge threshold, which makes the merge decision robust across a wide threshold plateau and brittle immediately above it, and which concentrates both error populations in pairs on which the available signals are silent. The policy gate withholds every single-source claim not as an observation but as a theorem, since no such claim can exceed a confidence of 0.5375 against a threshold of 0.60, and it separates corroborated from uncorroborated evidence in the intended order only for thresholds in the window (0.540, 0.605], which a synthetic second channel confirms and whose narrowness is reported as a design fact. The components act on disjoint metrics, so neither substitutes for the other. And the performance cost, after correcting an indexing defect in the original harness, is 10.7-fold at 500 records and 86.6-fold on the full corpus, with 87 per cent of it attributable to entity resolution, rather than the 5,447-fold the original submission reported.

The architecture's data model, confidence and policy contracts, and evaluation protocol are independent of any specific storage or processing technology and are intended to generalise beyond the single source evaluated here. Four directions follow most directly. The gold sample should be extended to the 75 already-selected reject-stratum pairs, so that the lower threshold rests on labelled evidence. The entity-resolution scorer should consume address text and portal identifiers, since both error populations show that this is where the discriminating evidence lies. An independent field-verification protocol should be established to calibrate the confidence function against ground truth and to widen, or justify, the gate's narrow admissible window. And the entity-resolution candidate scan should be indexed by block before the architecture is evaluated at a larger, multi-source scale.

References

[1] D. Sculley et al., "Hidden technical debt in machine learning systems," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 2503–2511.

[2] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, "'Everyone wants to do the model work, not the data work': Data cascades in high-stakes AI," in *Proc. CHI Conf. Human Factors in Computing Systems*, 2021, doi: 10.1145/3411764.3445518.

[3] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data lifecycle challenges in production machine learning: A survey," *ACM SIGMOD Record*, vol. 47, no. 2, pp. 17–28, 2018, doi: 10.1145/3299887.3299891.

[4] L. Bellato, N. Frantzeskaki, and C. A. Nygaard, "Regenerative tourism: a conceptual framework leveraging theory and practice," *Tourism Geographies*, vol. 25, no. 4, pp. 1026–1046, 2023, doi: 10.1080/14616688.2022.2044376.

[5] Z. Xiang, Q. Du, Y. Ma, and W. Fan, "A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism," *Tourism Management*, vol. 58, pp. 51–65, 2017, doi: 10.1016/j.tourman.2016.10.001.

[6] D. Buhalis and A. Amaranggana, "Smart tourism destinations enhancing tourism experience through personalisation of services," in *Information and Communication Technologies in Tourism 2015*, Springer, 2015, pp. 377–389, doi: 10.1007/978-3-319-14343-9_28.

[7] U. Gretzel, M. Sigala, Z. Xiang, and C. Koo, "Smart tourism: foundations and developments," *Electronic*

Copyright © 2026 – Indonesian Journal of Applied Mathematics

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia

Original Article

Markets, vol. 25, no. 3, pp. 179–188, 2015, doi: 10.1007/s12525-015-0196-8.

[8] K. Belhajjame, Y. Gil, and P. Groth, "PROV-DM: The PROV data model," W3C Recommendation, 2013. [Online]. Available: http://www.w3.org/TR/2013/REC-prov-dm-20130430/

[9] J. Cheney, L. Chiticariu, and W. C. Tan, "Provenance in databases: Why, how, and where," *Foundations and Trends in Databases*, vol. 1, no. 4, pp. 379–474, 2007, doi: 10.1561/1900000006.

[10] M. Herschel, R. Diestelkämper, and H. Ben Lahmar, "A survey on provenance: What for? What form? What from?," *VLDB Journal*, vol. 26, no. 6, pp. 881–906, 2017, doi: 10.1007/s00778-017-0486-1.

[11] P. Buneman, S. Khanna, and W. C. Tan, "Why and where: A characterization of data provenance," in *Database Theory — ICDT 2001*, LNCS 1973, Springer, 2001, pp. 316–330, doi: 10.1007/3-540-44503-X_20.

[12] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems 33*, 2020, pp. 9459–9474.

[13] Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, art. 248, 2023, doi: 10.1145/3571730.

[14] T. Gebru et al., "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021, doi: 10.1145/3458723.

[15] M. Mitchell et al., "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, and Transparency (FAT*)*, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[16] 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, doi: 10.1080/07421222.1996.11518099.

[17] T. C. Redman, "The impact of poor data quality on the typical enterprise," *Communications of the ACM*, vol. 41, no. 2, pp. 79–82, 1998, doi: 10.1145/269012.269025.

[18] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, art. 16, 2009, doi: 10.1145/1541880.1541883.

[19] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004, doi: 10.2307/25148625.

[20] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007, doi: 10.2753/MIS0742-1222240302.

[21] I. P. Fellegi and A. B. Sunter, "A theory for record linkage," *Journal of the American Statistical Association*, vol. 64, no. 328, pp. 1183–1210, 1969, doi: 10.1080/01621459.1969.10501049.

[22] P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin: Springer, 2012, doi: 10.1007/978-3-642-31164-2.

[23] C. S. Jensen and R. T. Snodgrass, "Temporal data management," *IEEE Transactions on Knowledge and Data Engineering*, vol. 11, no. 1, pp. 36–44, 1999, doi: 10.1109/69.755613.

[24] K. Kulkarni and J.-E. Michels, "Temporal features in SQL:2011," *ACM SIGMOD Record*, vol. 41, no. 3, pp. 34–43, 2012, doi: 10.1145/2380776.2380786.

[25] Y. Li et al., "A survey on truth discovery," *ACM SIGKDD Explorations Newsletter*, vol. 17, no. 2, pp. 1–16, 2016, doi: 10.1145/2897350.2897352.

[26] X. Dong et al., "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proc. 20th ACM SIGKDD Int. Conf. Knowledge Discovery and*

Data Mining, 2014, pp. 601–610, doi: 10.1145/2623330.2623623.

[27] P. C. Arocena, B. Glavic, G. Mecca, R. J. Miller, P. Papotti, and D. Santoro, "Messing up with BART: Error generation for evaluating data-cleaning algorithms," *Proceedings of the VLDB Endowment*, vol. 9, no. 2, pp. 36–47, 2015, doi: 10.14778/2850578.2850579.

[28] R. A. DeMillo, R. J. Lipton, and F. G. Sayward, "Hints on test data selection: Help for the practicing programmer," *Computer*, vol. 11, no. 4, pp. 34–41, 1978, doi: 10.1109/C-M.1978.218136.

[29] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977, doi: 10.2307/2529310.

[30] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960, doi: 10.1177/001316446002000104.

[31] B. Efron, "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979, doi: 10.1214/aos/1176344552.

[32] R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed. Indianapolis, IN: Wiley, 2013.

[33] G. Papadakis, D. Skoutas, E. Thanos, and T. Palpanas, "Blocking and filtering techniques for entity resolution: A survey," *ACM Computing Surveys*, vol. 53, no. 2, art. 31, 2020, doi: 10.1145/3377455.

Copyright © 2026 – Indonesian Journal of Applied Mathematics

Published by: Lembaga Penelitian dan Pengabdian Masyarakat (LPPM)
Institut Teknologi Sumatera, Lampung Selatan, Indonesia