

WCD-SBAS: An Explainable Welfare-Centric Data Warehousing Architecture for Automated Scheme–Beneficiary Alignment and Policy Decision Support

M. N. Nagendra Prasad¹, Dr. N. Geethanjali²

¹Research Scholar Department of Computer Science, Rayalaseema University Kurnool, Andhra Pradesh, India Email: mnnp999@gmail.com

²Professor, Department of Computer Science & Technology Sri Krishnadevaraya University, Anantapur, Andhra Pradesh, India Email: geethanjali.sku@gmail.com

Abstract: Social welfare delivery increasingly de-pends on information distributed across identity, civil-registration, income, education, health, employment, payment, grievance, and program-specific information systems. Such frag-mentation creates duplicate beneficiary records, stale or con-flicting attributes, heterogeneous schemas, limited provenance, opaque eligibility decisions, and weak cross-program analyti-cal capability. This paper proposes WCD-SBAS, an explain-able welfare-centric data warehousing architecture for auto-mated scheme–beneficiary alignment and policy decision support. The architecture integrates heterogeneous administrative sources through a governed interoperability and staging layer, performs privacy-aware beneficiary entity resolution, maintains a tempo-rally versioned beneficiary and household master, and organizes welfare processes using conformed dimensional models. Statutory eligibility is represented through versioned three-valued rules that distinguish satisfied, failed, and unknown conditions, allow-ing incomplete or stale evidence to be routed for verification instead of being interpreted as automatic ineligibility. Evidence-linked decision traces retain rule versions, source systems, times-tamps, and transformation provenance, while the OLAP layer supports coverage analysis, verification monitoring, grievance analysis, payment monitoring, and non-binding policy what-if evaluation. The proposed workflow is evaluated using a controlled synthetic validation environment designed to examine entity-resolution robustness, incomplete eligibility evidence, analytical-query behavior, and scalability. Under a stringent 40% field-corruption condition, the WCD-SBAS hybrid entity resolver achieves an F1 score of 0.817 ± 0.002 , compared with 0.810 ± 0.003 for deterministic multi-field matching, 0.808 ± 0.002 for a weighted probabilistic matcher, and 0.674 ± 0.002 for exact-ID matching. The evidence-aware rule engine automatically resolves 93.3% of evaluated beneficiary–scheme pairs and routes 6.7% for verification when mandatory evidence is unavailable or stale. In the modeled analytical workload, dimensional queries obtain median speedups of approximately $7.75\text{--}8.48\times$ over fragmented operational-query execution. Additional stress experiments show progressive degradation as linkage evidence becomes corrupted and increasing verification rather than automatic denial as eligibility evidence becomes incomplete. These findings demon-strate the functional behavior and robustness of the proposed architecture within the controlled validation environment; evalu- ation on authorized administrative data and production database platforms is required before operational performance or policy-effectiveness claims can be made.

Keywords: Data warehouse, social welfare, beneficiary matching, entity resolution, eligibility rules, explainable decision support, OLAP, data provenance, social registry

1. Introduction

Social protection and welfare programs are commonly de-livered through multiple institutions, each maintaining its own operational data. A single household may appear separately in identity systems, civil registries, education systems, em-ployment databases, disability registries, payment platforms, grievance systems, and individual program management infor-mation systems. The World Bank’s social-protection delivery-systems framework



emphasizes that fragmented delivery processes make programs difficult to navigate and that interoperable delivery platforms can improve coordination, inclusion, transparency, and efficiency [1]. Recent guidance similarly emphasizes dynamic social registries, secure identity, inter-operable data exchange, and digital payments as foundational components of modern social-protection delivery [2], [3].

The technical problem is broader than building a single beneficiary database. Cross-agency data may use incompatible identifiers, conflicting timestamps, different household definitions, inconsistent geographic codes, and varying data quality. A data warehouse must therefore support not only analytical integration but also identity resolution, temporal history, provenance, privacy controls, eligibility-rule execution, explanation, and multidimensional policy analysis. Traditional data warehousing provides mature concepts for ETL, conformed dimensions, fact tables, and OLAP [5], [6]; data-integration research formalizes the challenge of combining heterogeneous sources into a unified view [7]; and entity-resolution literature addresses record linkage, deduplication, scalability, and privacy [8]–[11]. These strands, however, are often treated separately from program-rule execution and welfare-policy analytics.

Automated eligibility also requires caution. A welfare system should not replace statutory criteria with an opaque predictor merely because a machine-learning model is convenient. Rule-based screening is attractive when program criteria are explicit, while learning-based methods can be useful for ranking, anomaly detection, or workload prioritization. Recent work such as ReliefAI illustrates the growing interest in hybrid rule-and-learning systems for public-resource eligibility and recommendation, but its evaluation is also simulation-based and focused on service discovery rather than a cross-agency data-warehouse backbone [18]. Explainability methods such as SHAP and LIME provide post-hoc explanations for complex models [14], [15]; for welfare eligibility, however, a stronger form of explanation is possible when the decision itself is derived from explicit predicates and data provenance.

This paper introduces WCD-SBAS (Welfare-Centric Data Warehousing–Scheme Beneficiary Alignment System), an architecture in which eligibility, provenance, data quality, and policy analytics are first-class warehouse concerns. The main contributions are:

- a welfare-centric canonical data model and dimensional warehouse that separates operational source data from analytical and decision-support layers;
- a privacy-aware entity-resolution stage that combines deterministic identifiers, field similarity, source reliability, household context, and manual-review thresholds;
- an evidence-aware eligibility engine based on three-valued predicates, preventing missing data from being silently interpreted as ineligibility;
- a native explanation object that records satisfied, failed, and unknown predicates together with values, source systems, timestamps, and rule versions;
- OLAP measures for coverage gaps, duplication, payment status, grievance aging, geographic disparities, verification backlog, and policy what-if analysis; and
- a conservative stress-test reporting strategy that emphasizes corrupted-record robustness and verification behavior rather than near-perfect clean-condition scores.

2. Related Work

A. Social-Protection Delivery Systems and Interoperability

The World Bank Sourcebook describes social-protection delivery as an end-to-end chain covering outreach, intake and registration, assessment, enrollment, provision, and beneficiary operations such as updates and grievances [1]. It notes that social registries increasingly interoperate with identity, civil-registration, tax, education, health, property, and other administrative systems. Such interoperability requires technical, data, and process standards, as well as protocols for conflicting or stale data. More recent World Bank guidance frames social registries, digital ID,

interoperable data exchange, case management, grievance mechanisms, and payments as reusable digital delivery infrastructure [2], [4]. These sources motivate

the system context of WCD-SBAS but do not prescribe a data-warehouse architecture for historical analytics, rule-level explainability, and scheme-to-beneficiary alignment.

B. Data Warehousing and OLAP

Data warehousing separates analytical workloads from operational transaction processing. Chaudhuri and Dayal describe the canonical warehouse pipeline of extraction, cleaning, loading, multidimensional models, metadata, and OLAP [5]. Kimball and Ross formalize dimensional modeling using business processes, facts, dimensions, grain, conformed dimensions, and star schemas [6]. These ideas are particularly suitable for welfare analytics because program participation, payments, grievances, eligibility evaluations, and case outcomes are event-like facts connected to beneficiary, household, scheme, geography, time, and agency dimensions.

C. Entity Resolution and Privacy

Entity resolution identifies records that refer to the same real-world entity despite inconsistent representations. Christen provides a comprehensive treatment of preprocessing, blocking, field comparison, classification, quality evaluation, and practical limitations [8]. Scalable blocking and indexing are necessary because all-pairs comparison becomes infeasible at administrative scale [9]. Privacy-preserving record linkage (PPRL) addresses the additional risk that linkage often relies on sensitive identifiers [10], [11]. WCD-SBAS therefore separates linkage keys from analytical surrogate keys, supports hashed/tokenized comparison where permitted, and retains manual review for uncertain links.

D. Rules, Explainability, and Provenance

Production-rule systems have long used efficient pattern matching such as the Rete algorithm [12]. In welfare administration, such rule execution is useful because many eligibility conditions are explicitly stated as predicates over age, income, household status, geography, disability, enrollment, or other attributes. For predictive components, explainable-AI techniques such as SHAP and LIME can expose feature contributions [14], [15]; however, statutory eligibility should ideally be explainable by construction. Data provenance provides the complementary ability to answer where data came from and how it reached the current representation [13]. NIST’s AI Risk Management Framework further emphasizes governance, transparency, and risk management for AI-enabled systems [16]. WCD-SBAS combines rule traces with provenance so that explanations are evidence-linked rather than merely post-hoc.

E. Research Gap

Prior work provides strong components—social registries, data integration, dimensional warehouses, entity resolution, privacy-preserving linkage, rule engines, and explainability—but an integrated welfare-centric architecture connecting all of them to cross-program scheme alignment and policy analytics is less explicit. The gap addressed here is therefore

architectural: how to structure a warehouse so that eligibility decisions, unresolved evidence, provenance, historical state, and analytical measures coexist without turning the warehouse into either a black-box predictor or a replica of fragmented operational systems.

3. Problem Formulation and Design Principles

Let B denote beneficiaries or households and S denote welfare schemes. Each beneficiary $b \in B$ is represented by a time-indexed evidence vector

$$\mathbf{x}_b(t) = \{x_{b1}(t), x_{b2}(t), \dots, x_{bm}(t)\}, \quad (1)$$

where each attribute also carries provenance metadata $p = (\text{source}, \text{timestamp}, \text{quality}, \text{transform})$. Each scheme $s \in S$ contains a versioned set of eligibility predicates $R_s = \{r_{s1}, \dots, r_{sk}\}$.

The design follows six principles. First, operational data remain owned by source agencies; the warehouse stores gov-erned analytical copies and histories. Second, identity resolution is probabilistic only where exact linkage is insufficient and ambiguous links remain reviewable. Third, eligibility rules are versioned and auditable. Fourth, missing evidence yields an *unknown* state rather than automatic failure. Fifth, every alignment result has provenance. Sixth, policy analytics use aggregate measures and what-if scenarios without rewriting historical decisions.

4. Proposed WCD-SBAS Architecture

Figure 1 shows the proposed architecture.

A. Interoperability Gateway and Staging

Incoming extracts, APIs, or event streams are mapped into a canonical exchange schema. Each attribute receives a source identifier, extraction time, effective time, quality flag, and transformation lineage. Schema validation rejects structurally invalid records, while a quarantine area isolates conflicting identifiers, impossible dates, malformed geography codes, and other data-quality anomalies. This separation prevents data cleaning from silently overwriting source evidence.

B. Privacy-Aware Entity Resolution

For two candidate records i and j , the linkage score is

$$S_{ij} = \frac{\sum_{k=1}^m w_k q_{ik} q_{jk} \text{sim}_k(i, j)}{\sum_{k=1}^m w_k q_{ik} q_{jk}}, \quad (2)$$

where w_k is field importance, q_{ik} is a quality/reliability factor, and sim_k is an exact, phonetic, numeric, token, or privacy-preserving similarity. Strong exact identifiers can short-circuit the score where legally and technically appropriate. Candidate generation uses blocking keys to reduce comparison cost. Links above τ_H are accepted, scores below τ_L are rejected, and the interval $[\tau_L, \tau_H)$ is routed to review. The linkage object stores the compared fields and evidence without exposing protected identifiers to ordinary analytical users.

TABLE I
CORE WELFARE-CENTRIC DIMENSIONAL MODEL

Object	Representative content
FactEligibility	beneficiary, scheme, evaluation date, outcome, rule version, evidence status
FactPayment	beneficiary, scheme, payment cycle, amount/status, failure reason
FactGrievance	beneficiary/household, scheme, category, opened/closed dates, resolution status
FactVerification	case, missing/conflicting evidence, queue state, reviewer outcome
Conformed dimensions	beneficiary, household, scheme, geography, agency, date/time

C. Versioned Beneficiary and Household Master

The canonical master assigns surrogate keys to persons and households and preserves history using slowly changing dimensions. Conflicting attributes are not collapsed into a single value without policy; instead, the current analytical value is accompanied by its selected-source rule, timestamps, and alternates. This allows the warehouse to answer both “what is the current value?” and “what evidence existed when the decision was made?”

D. Welfare-Centric Dimensional Model

The warehouse uses conformed beneficiary, household, scheme, geography, agency, and time dimensions so that enrollment, payment, grievance, verification, and eligibility events can be compared consistently across programs. Rather than adding a second architecture-style figure, Table I summarizes the minimum analytical grain required by the design. This keeps the manuscript focused on the proposed end-to-end architecture in Fig. 1 while still specifying the warehouse structure needed for implementation.

E. Evidence-Aware Eligibility Engine

Each predicate returns one of three values:

$$g_{sj}(b) \in \{T, F, U\}, \quad (3)$$

where U denotes unknown or insufficiently fresh evidence. For conjunctive mandatory rules, a single false predicate produces ineligibility; if no predicate is false but at least one is unknown, the case is routed to verification. Only when all mandatory predicates are true is the beneficiary automatically marked eligible. Optional scoring criteria may then rank eligible programs, but do not override statutory rules.

For ranking among schemes that are already statutorily eligible, non-binding administrative criteria such as evidence completeness, recency, or outreach priority may be used. Such ranking must never override the mandatory eligibility predicates or convert an unknown mandatory predicate into eligibility or ineligibility.

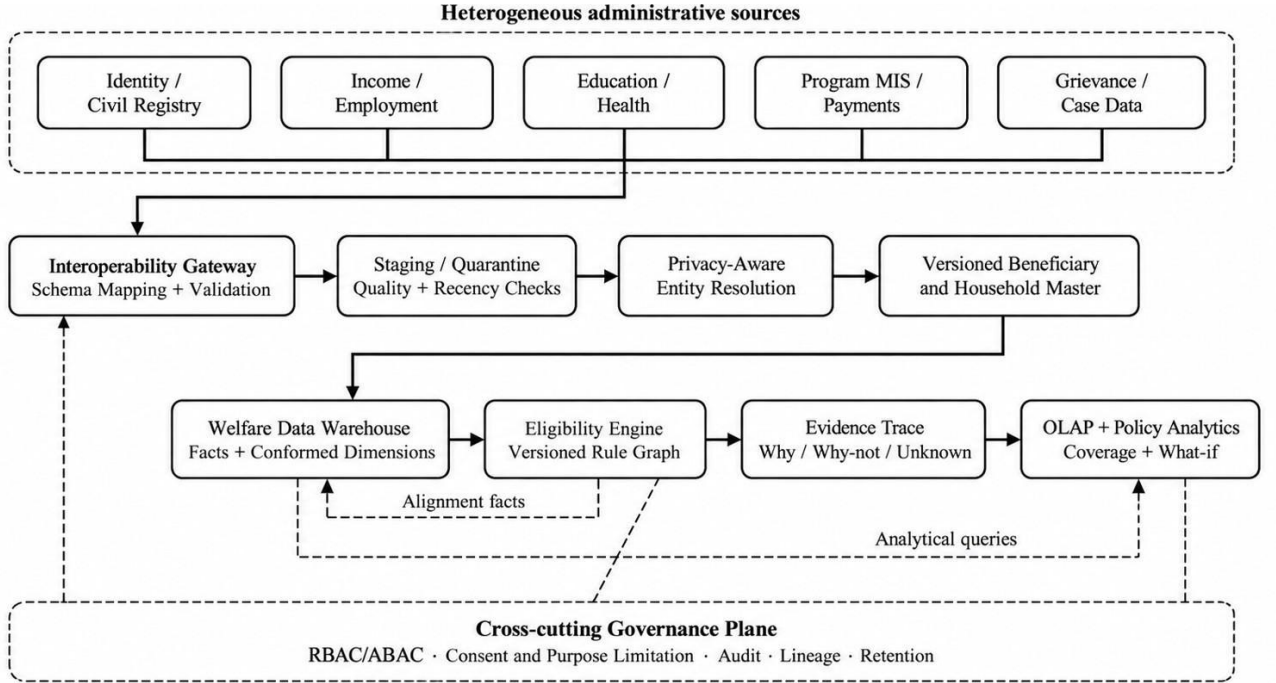


Fig. 1. WCD-SBAS architecture showing administrative data integration, the versioned beneficiary and household master, the welfare data warehouse, eligibility evaluation, evidence traces, policy analytics, and cross-cutting governance.

F. Native Explanation and Provenance

For each beneficiary–scheme evaluation, WCD-SBAS stores a structured decision trace containing the predicate identifier, observed value, threshold or category condition, true/false/unknown outcome, source system, evidence times-tamp, transformation lineage, and rule version. This directly supports “why eligible,” “why not

eligible,” and “what information is missing” explanations. Because the explanation is generated from the executed rule and the actual evidence used for the decision, it is auditable by construction rather than being a post-hoc approximation. If predictive components are later introduced for non-binding prioritization, SHAP or LIME can supplement—but not replace—the statutory rule trace [14], [15].

G. OLAP and Policy Decision Support

The analytical layer exposes measures including estimated eligible population, enrolled population, paid population, un-resolved verification cases, duplicate-suspect count, grievance aging, payment failures, and evidence freshness. A coverage gap is computed as the difference between the eligible-or-potentially-eligible count and the enrolled/benefited count for the same scheme, geography, and time period. The what-if engine changes a rule parameter in a sandbox and recomputes aggregate counts without altering historical or production decisions.

5. Eligibility Evaluation Procedure

The operational procedure follows the three-valued logic described above and does not require a separate pseudocode listing. For a selected beneficiary and versioned scheme, the engine retrieves each mandatory predicate and the best admissible evidence with provenance. Missing or stale evidence is marked UNKNOWN; any failed mandatory predicate produces NOTELIGIBLE; if no predicate fails but at least one remains unknown, the case is placed in the verification queue; only when every mandatory predicate is satisfied is the pair marked ELIGIBLE. The stored trace records the rule, evidence, source, timestamp, and outcome used at each step. This compact description avoids duplicating the same decision logic in both equations and an algorithm block.

6. Security, Privacy, and Governance

Welfare data are sensitive and often linkable across domains. The architecture therefore separates identity-linkage services from ordinary analytics, applies least-privilege role-or attribute-based access, records immutable audit events, and binds data access to declared purpose. Direct identifiers need not appear in dimensional marts; surrogate keys are sufficient for most analytics. Cross-agency linkage should use privacy-preserving techniques where necessary, while acknowledging that PPRL introduces its own trade-offs between privacy, linkage accuracy, complexity, and attack resistance [10], [11]. Aggregate releases can additionally apply disclosure controls or differential-privacy mechanisms where appropriate [17]. Governance must also define retention, correction, appeal, grievance, and data-conflict procedures rather than treating them as purely technical concerns.

7. Experimental Methodology

A. Scope of Validation

No real administrative beneficiary records or live government program data are used in this evaluation. The experiments are deterministic synthetic workflow validations intended to test architecture behavior, robustness, and reporting logic, not to estimate real inclusion, exclusion, fraud, or program-coverage rates. All numerical values in this section therefore characterize the synthetic generator and should be replaced or supplemented by authorized empirical evaluation before any deployment claim is made.

B. Experimental Design Summary

The experiment is organized as a traceable sequence rather than as a collection of disconnected headline values. First, the deterministic generator creates virtual beneficiaries, source-system records, scheme rules, field corruption, and missing evidence. Second, candidate entity pairs are evaluated by four linkage methods under identical synthetic inputs. Third, beneficiary–scheme pairs are processed by the fragmented-rule baseline and the three-valued WCD-SBAS engine. Fourth, five analytical query classes are executed under both the fragmented operational-query latency model and the dimensional-warehouse latency model. Finally, robustness sweeps vary record corruption and evidence missingness, while repeated runs and bootstrap resampling quantify run-level variation. All tables and figures in the

Results section are generated from these controlled stages; none is copied from external literature or a real welfare database.

C. Base Synthetic Workload

The base generator uses seed 20260925 and creates 120,000 virtual beneficiaries, 12 source systems, and 36 synthetic schemes. The entity-resolution benchmark contains 60,000 positive and 240,000 negative candidate pairs with controlled identifier and attribute corruption. The scheme-alignment benchmark samples 360,000 beneficiary–scheme pairs. Miss-ingness is injected into income, age, and categorical evidence to emulate fragmented source availability. Five representative OLAP workloads are exercised repeatedly, and a scalability model covers 50,000 to 1,000,000 input records. The source package includes the generators, assumptions, figures, and CSV outputs.

D. Robustness and Repeated Experiments

To avoid drawing conclusions from an artificially clean linkage setting, positive-pair field agreement is deliberately degraded and the main entity-resolution comparison is anchored at the stringent 40% corruption condition. The 20% condition is retained as a moderate-stress sensitivity point. Twenty independent repetitions are executed at each corruption level. Required eligibility evidence is independently masked at 0, 5, 10, 20, and 30%, again with 20 repetitions. For OLAP evaluation, each of the five query classes is simulated for 100 repeated runs per storage design. Median speedup confidence intervals are estimated by 3,000 bootstrap resamples. The clean

TABLE II
Primary Entity-Resolution Stress Test at 40% Field Corruption

Method	F1 (mean $\pm$ SD)
Exact-ID join	0.674 $\pm$ 0.002
Deterministic multi-field	0.810 $\pm$ 0.003
Weighted probabilistic ER	0.808 $\pm$ 0.002
WCD-SBAS hybrid ER	0.817 $\pm$ 0.002

linkage condition is treated only as a simulator sanity check and is not used as the headline performance claim.

E. Baselines and Metrics

Entity resolution is compared against exact-ID matching, deterministic multi-field rules, and a weighted probabilistic matcher. F1 is used as the principal linkage metric because it balances precision and recall; larger values are better. Following the conservative reporting protocol, the primary comparison uses the 40% field-corruption condition instead of the near-perfect clean condition. Eligibility alignment compares fragmented source-local rule execution against the evidence-aware three-valued rule engine. Because WCD-SBAS can return VERIFY, the paper reports automatic-decision coverage and verification rate rather than emphasizing an automatic-decision F1 that is exact by construction in the simulator. OLAP performance is summarized by median latency, 95th-percentile latency, and relative speedup, where speedup is the fragmented-model median latency divided by the warehouse median latency. Thus, values greater than 1 indicate faster modeled warehouse queries; for example, an 8 $\times$ value means approximately one-eighth of the modeled latency. Scalability is reported as modeled processing rate versus record volume.

8. Experimental Results

A. Primary Entity-Resolution Stress Test

The primary linkage comparison intentionally uses the stringent 40% field-corruption condition rather than the clean synthetic operating point. This choice avoids making the paper depend on near-perfect scores produced by easily

separable simulated records. As shown in Table II, WCD-SBAS obtains F1 0.817 ± 0.002 , compared with 0.810 ± 0.003 for deterministic multi-field rules, 0.808 ± 0.002 for the weighted probabilistic matcher, and 0.674 ± 0.002 for exact-ID matching. The advantage over the strongest baseline is therefore modest, which is a more credible characterization of the architecture under difficult evidence quality.

B. Robustness to Record Corruption

Table III and Fig. 2 compare the moderate 20% and stringent 40% corruption settings. At 20% corruption, WCD-SBAS reaches F1 0.930 ± 0.002 , only slightly above the weighted probabilistic matcher (0.926 ± 0.002) and deterministic rules (0.912 ± 0.002). At 40% corruption, all methods degrade substantially and the hybrid method retains F1 0.817 ± 0.002 .

TABLE III
Entity-Resolution Robustness Under Corrupted Evidence

Method	20% corruption	40% corruption
Exact-ID	0.762 ± 0.002	0.674 ± 0.002
Deterministic	0.912 ± 0.002	0.810 ± 0.003
Weighted probabilistic	0.926 ± 0.002	0.808 ± 0.002
WCD-SBAS hybrid	0.930 ± 0.002	0.817 ± 0.002

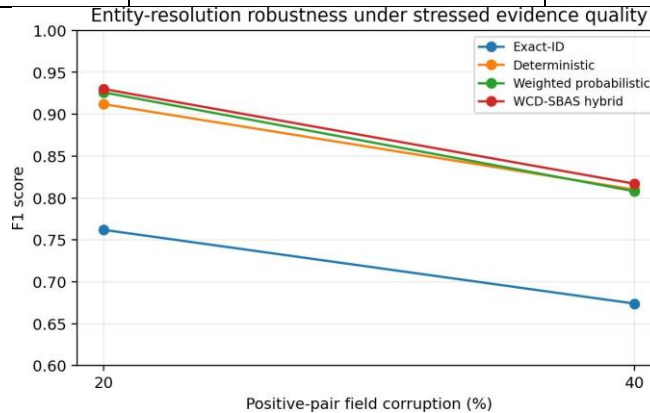


Fig. 2. Entity-resolution robustness under moderate and stringent positive-pair field corruption. The 40% condition is used as the primary reported stress test.

The important result is therefore not an unusually high clean-condition score but the relative stability of the workflow as evidence quality worsens.

C. Evidence-Aware Scheme Alignment

The scheme-alignment experiment is evaluated primarily as a decision-coverage problem. The fragmented baseline forces a binary result even when mandatory evidence is absent, whereas WCD-SBAS can return VERIFY. In the base workload, 93.3% of sampled beneficiary-scheme pairs contain sufficient evidence for an automatic decision and 6.7% are routed to verification. These values must not be read as 93.3% eligibility accuracy. Because the simulator derives rule ground truth from the same complete attributes used by the rule engine, an automatic-decision F1 would be artificially perfect and is therefore omitted from the primary table.

D. Robustness to Missing Eligibility Evidence

The second robustness sweep increases the probability that a required age, income, or categorical attribute is unavailable. As shown in Table V and Fig. 3, the fragmented rule baseline's recall drops from 1.000 with no missing

evidence to 0.373 at 30% missingness because unknown evidence is effectively converted into denial. The three-valued WCD-SBAS engine instead reduces automatic coverage from 1.000 to 0.799 and increases verification to 0.201. Automatic-decision F1 is intentionally not reported because its perfect value would be a construction property of the synthetic rule generator rather than meaningful evidence of real-world decision accuracy.

TABLE IV

Base Synthetic Scheme-Alignment Decision Coverage Method	Auto decision coverage	Verify rate
Missing-evidence handling		
Fragmented local rules	1.000	0.000
WCD-SBAS rule engine	0.933	0.067
		Forced binary decision
		Explicit verification

TABLE V

Missing-Evidence Robustness	Missing Frag.	Recall WCD auto cov.	Verify
0%	1.000	1.000	0.000
5%	0.867	0.970	0.030
10%	0.746	0.939	0.061
20%	0.537	0.873	0.127
30%	0.373	0.799	0.201

E. Analytical Query Performance

Table VI summarizes 100 repeated synthetic runs for each analytical workload. Here, *lower latency is better*. Relative speedup is computed as raw/fragmented median latency di-

vided by dimensional-warehouse median latency. The resulting values range from $7.75\times$ to $8.48\times$; therefore an $8\times$ value means the modeled warehouse query takes about one-eighth of the comparison latency for that workload. Bootstrapped 95% intervals remain above one for all five workloads, showing

that this simulated difference is stable to the run-level jitter encoded by the generator. These are not measurements of PostgreSQL, SQL Server, Oracle, or any deployed government platform; they quantify only the controlled latency model used in this evaluation.

F. Scalability Envelope

Figure 5 shows the modeled processing envelope from 50,000 to 1,000,000 input records. ETL throughput declines from approximately 40.6k to 34.7k rows/s, entity-resolution candidate comparisons from 28.1k to 22.3k comparisons/s, and eligibility evaluation from 75.5k to 67.9k pairs/s. The degradation is gradual rather than abrupt in this synthetic model, but these values are not production throughput measurements and depend directly on the capacity functions encoded by the simulator.

G. Policy What-If Analysis

For a representative synthetic rural-assistance scheme with a baseline income threshold of 200,000 units, reducing the threshold by 10% changes the potentially eligible count from **41,553 to 37,353** (-10.1%), while increasing it by 10% changes the count to 45,205 ($+8.8\%$). A 20% increase produces 48,426 potential matches ($+16.5\%$). Figure 6 visualizes this nonlinear response. The analysis is descriptive and does not recommend any threshold change; it demonstrates how the warehouse can quantify the distributional consequences of candidate parameters before a human policy decision.

9. Discussion

The principal architectural choice in WCD-SBAS is to keep statutory eligibility deterministic and explainable while

TABLE VI
Repeated Synthetic Analytical Workload Latency

Query	Raw med.	Raw P95	DW med.	DW P95	Speedup 95% CI
Eligible population by scheme/district	836	921	103	116	$8.08 \times$ [7.86, 8.29]
Payment coverage trend	696	778	82	92	$8.48 \times$ [8.30, 8.80]
Duplicate-beneficiary drill-down	1336	1541	169	190	$7.92 \times$ [7.70, 8.21]
Grievance aging by category	744	865	96	105	$7.75 \times$ [7.59, 8.03]
What-if income-threshold analysis	1800	2064	214	233	$8.41 \times$ [8.23, 8.62]

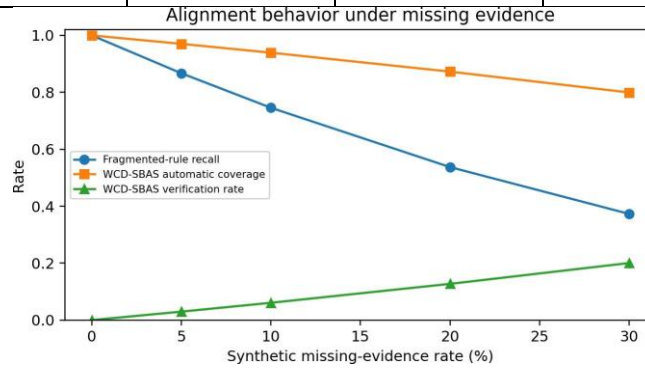


Fig. 3. Synthetic behavior under missing eligibility evidence. WCD-SBAS converts unresolved evidence into verification rather than automatic denial.

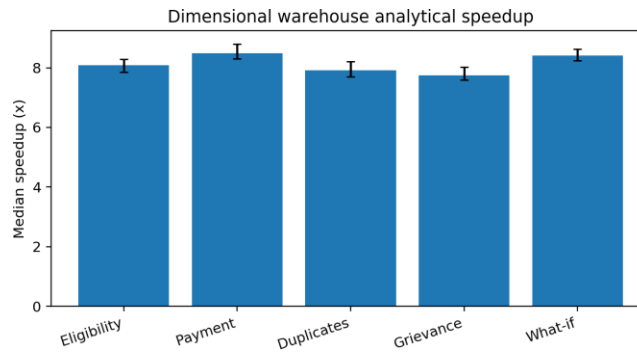


Fig. 4. Median analytical speedup of the dimensional warehouse over the fragmented operational-query model. Error bars show bootstrapped 95% intervals.

using data warehousing and analytics to improve evidence availability, cross-program visibility, and policy analysis. This separation reduces the risk that a learned model silently replaces legal or administrative criteria. Predictive methods can still be used for non-binding tasks such as anomaly prior-ization, workload forecasting, or outreach ranking, provided they are governed and evaluated independently.

The three-valued eligibility model is particularly important. Traditional database logic often collapses missing values into nulls that are later treated inconsistently by application code. In welfare delivery, an unknown income, disability certificate, household member, or residence status can be outcome-determinative. The architecture therefore makes “unknown” a first-class state and connects it to a verification queue. This

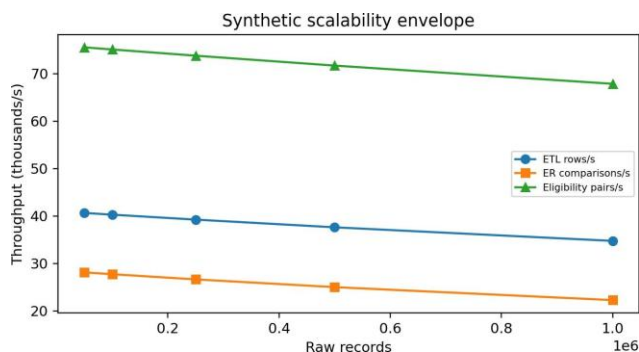


Fig. 5. Synthetic scalability envelope for ETL, entity-resolution comparisons, and beneficiary–scheme eligibility evaluation.

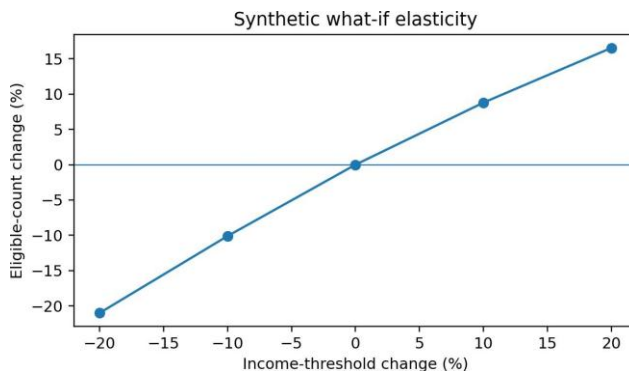


Fig. 6. Synthetic what-if elasticity for the representative income threshold. The horizontal axis is a candidate parameter change, not a policy recommendation.

creates a measurable verification backlog and avoids conflating missing data with non-eligibility.

Provenance also changes the role of the warehouse. Instead of storing only a current “best” value, WCD-SBAS keeps the selected value’s origin, effective time, processing path, and rule context. This supports appeals, audit, correction, and retrospective analysis. Provenance concepts are well established in database research [13]; the contribution here is to bind them directly to eligibility explanations.

From a policy-analytics perspective, the star schema enables drill-down and drill-across across schemes, geography, demographic bands, agencies, and time. Analysts can identify areas with high potential-eligibility but low enrollment, large

verification backlogs, persistent payment failure, or recurring grievance patterns. These are descriptive decision-support outputs, not automated political recommendations. Human authorities remain responsible for interpreting trade-offs, validating data, and deciding any policy change.

10. Threats to Validity and Limitations

The current evaluation is synthetic. The generator simplifies data corruption, household structure, source reliability, scheme complexity, and administrative workflows. Real welfare systems may contain multilingual names, address ambiguity, household splits and mergers, changing geography, policy exceptions, legal appeals, stale data, intentional misreporting, and cross-border or inter-state identities that are not captured here. Real rule sets can also include dependency chains and documentary conditions that are difficult to express as simple predicates.

The entity-resolution benchmark is based on simulated field agreements rather than real privacy-preserving encodings. The OLAP latency experiment is a controlled workload model and does not substitute for benchmarking a specific database engine, storage architecture, concurrency level, or hardware configuration. Similarly, the what-if analysis illustrates computational capability only and should not be interpreted as a policy recommendation.

Deployment would require a jurisdiction-specific governance review covering lawful basis, purpose limitation, data-sharing agreements, security, retention, correction, grievance, accessibility, and human oversight. Empirical validation should use authorized, de-identified or appropriately governed administrative data, with separate evaluation of linkage error, eligibility-rule fidelity, subgroup error, timeliness, explanation usability, and operational scalability.

11. Conclusion

This paper presented WCD-SBAS, a welfare-centric architecture that combines governed data integration, beneficiary/household master data, a dimensional warehouse, three-valued eligibility rules, evidence-linked explanations, and OLAP decision support. The central design choice is to separate deterministic statutory eligibility from non-binding analytics: missing mandatory evidence is routed to verification rather than treated as automatic ineligibility, and every decision can be traced to a rule version and source evidence. The numerical evaluation is intentionally synthetic. Its purpose is to verify that the proposed workflow behaves as designed under controlled corruption, missing evidence, repeated query workloads, and scalability stress. The primary linkage result is deliberately reported at the stringent 40% corruption condition, where WCD-SBAS obtains $F1\ 0.817 \pm 0.002$ rather than relying on the near-perfect clean-condition score. The 93.3% automatic-decision coverage and the $7.75\text{--}8.48\times$ modeled query speedups likewise characterize the simulator, not a real welfare population or production database. The robustness experiments are emphasized because they show how performance changes when evidence quality is degraded. Future work should replace or supplement the synthetic workload with authorized administrative data, manually adjudicated linkage labels, jurisdiction-specific rule validation, real database benchmarks, subgroup error analysis, case-worker explanation studies, and formal governance review before any deployment claim.

REFERENCES

1. World Bank, *Sourcebook on the Foundations of Social Protection Delivery Systems*. Washington, DC, USA: World Bank, 2020.
2. World Bank, “Digital Delivery Systems for Social Protection,” Sept. 2025.
3. World Bank, “Social Registries as Digital Platforms for Inclusive Service Delivery,” 2025.
4. World Bank, “Global Digital Public Infrastructure Program: From Foundations to Scale,” 2026.
5. S. Chaudhuri and U. Dayal, “An overview of data warehousing and OLAP technology,” *ACM SIGMOD Record*, vol. 26, no. 1, pp. 65–74, 1997, doi: 10.1145/248603.248616.
6. R. Kimball and M. Ross, *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd ed. Indianapolis, IN, USA: Wiley, 2013.
7. M. Lenzerini, “Data integration: A theoretical perspective,” in *Proc. 21st ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems*, 2002, pp. 233–246, doi: 10.1145/543613.543644.
8. P. Christen, *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Berlin, Germany: Springer, 2012, doi: 10.1007/978-3-642-31164-2.
9. P. Christen, “A survey of indexing techniques for scalable record linkage and deduplication,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 9, pp. 1537–1555, 2012, doi: 10.1109/TKDE.2011.127.
10. D. Vatsalan, P. Christen, and V. S. Verykios, “A taxonomy of privacy-preserving record linkage techniques,” *Information Systems*, vol. 38, no. 6, pp. 946–969, 2013, doi: 10.1016/j.is.2012.11.005.

11. A. Gkoulalas-Divanis, D. Vatsalan, D. Karapiperis, and M. Kantarcioglu, “Modern privacy-preserving record linkage techniques: An overview,” *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4966–4987, 2021, doi: 10.1109/TIFS.2021.3114026.
12. C. L. Forgy, “Rete: A fast algorithm for the many pattern/many object pattern match problem,” *Artificial Intelligence*, vol. 19, no. 1, pp. 17–37, 1982, doi: 10.1016/0004-3702(82)90020-0.
13. P. Buneman, S. Khanna, and W.-C. Tan, “Why and where: A characterization of data provenance,” in *Database Theory—ICDT 2001*, LNCS 1973, pp. 316–330, 2001, doi: 10.1007/3-540-44503-X20.
14. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems 30*, 2017.
15. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
16. E. Tabassi, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, National Institute of Standards and Technology, 2023, doi: 10.6028/NIST.AI.100-1.
17. C. Dwork, “Differential privacy,” in *Automata, Languages and Program-ming*, LNCS 4052, 2006, pp. 1–12.
18. R. Evani, “ReliefAI: A hybrid AI-powered decision support system for public resource eligibility across food, shelter, and clothing needs,” *IEEE Access*, vol. 13, pp. 216152–216160, 2025, doi: 10.1109/AC-CESS.2025.3646461.