CogniCrest

Anonymization Pipeline — Technical Benchmark Report

HRMS · Workforce Analytics Without Exposing Employees · synthetic data, generated for demonstration purposes

Important note: This report uses a fully synthetic dataset generated for demonstration purposes only — no real customers, patients, policyholders, or employees are involved, and no real company's data was used. This is a technical benchmark, not a client case study.

Workforce Analytics Without Exposing Employees

PROBLEM An HR platform wants pay-equity and workforce analytics across departments, but raw salary and performance data can't reach the analytics team without exposing individual employees.

APPROACH Tokenize employee identifiers, generalize salary bands and performance scores — enough signal for aggregate analysis, without exposing individuals.

GOAL Analytics proceeds. Employee privacy stays protected.

Methodology

A synthetic HRMS dataset of 400 employee records was generated, following the end-to-end process common to workforce data anonymization: ingestion, schema normalization, field classification, policy selection, privacy transformation, risk evaluation, utility evaluation, and a governance gate before delivery. Direct identifiers were tokenized. Quasi-identifiers (age, gender, region, department) were generalized using an adaptive suppression ladder until every record achieved k-anonymity of at least 5. Free-text HR case notes were screened for embedded names and addresses and redacted separately from the structured fields.

Field Classification

Fields are classified into four categories: direct identifiers, quasi-identifiers, sensitive attributes, and operational fields.

Field	Classification
employee_id	Direct identifier
full_name	Direct identifier
email	Direct identifier
phone	Direct identifier
street_address	Direct identifier

age	Quasi-identifier
gender	Quasi-identifier
zip_code	Quasi-identifier
department	Quasi-identifier
flight_risk_tier	Sensitive attribute
salary	Sensitive attribute
hr_note	Sensitive attribute (free text)
tenure_years	Operational field
performance_rating	Operational field
attrition	Target / operational field

Text Redaction — HR Case Notes

Free-text HR notes — performance discussions, compensation reviews, exit interviews — are a common privacy gap in workforce data: employee names and personal details are routinely typed directly into notes fields, bypassing whatever protection was applied to the structured columns. The pipeline screens narrative fields separately and redacts embedded direct identifiers before the note is included in any anonymized release.

BEFORE:

```
Discussed compensation and performance with David Kim (residence: 77 Cedar Court, Seattle). Flight risk assessment: Medium. Department: Engineering.
```

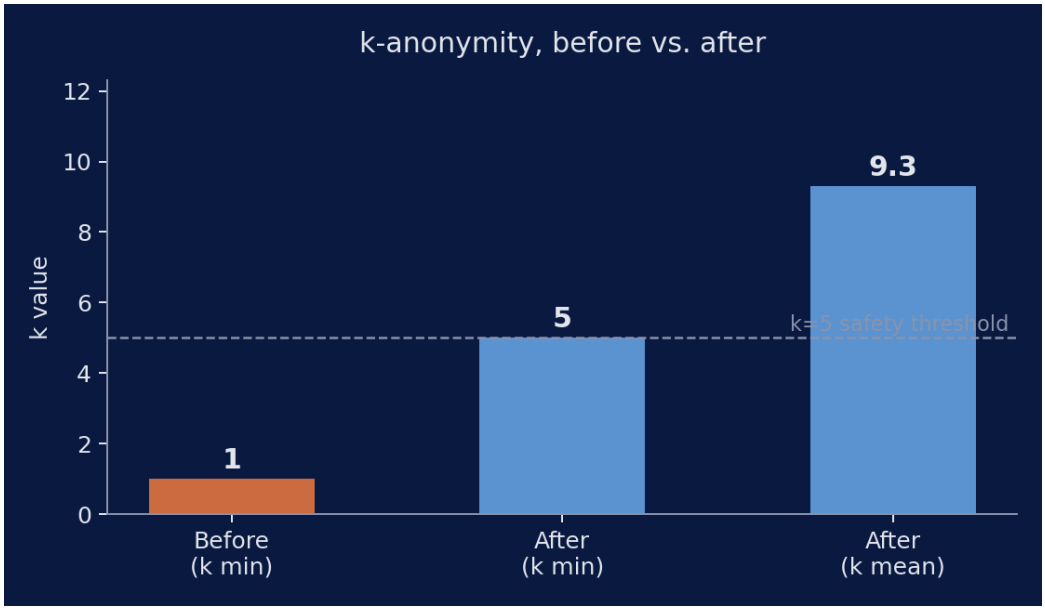
AFTER:

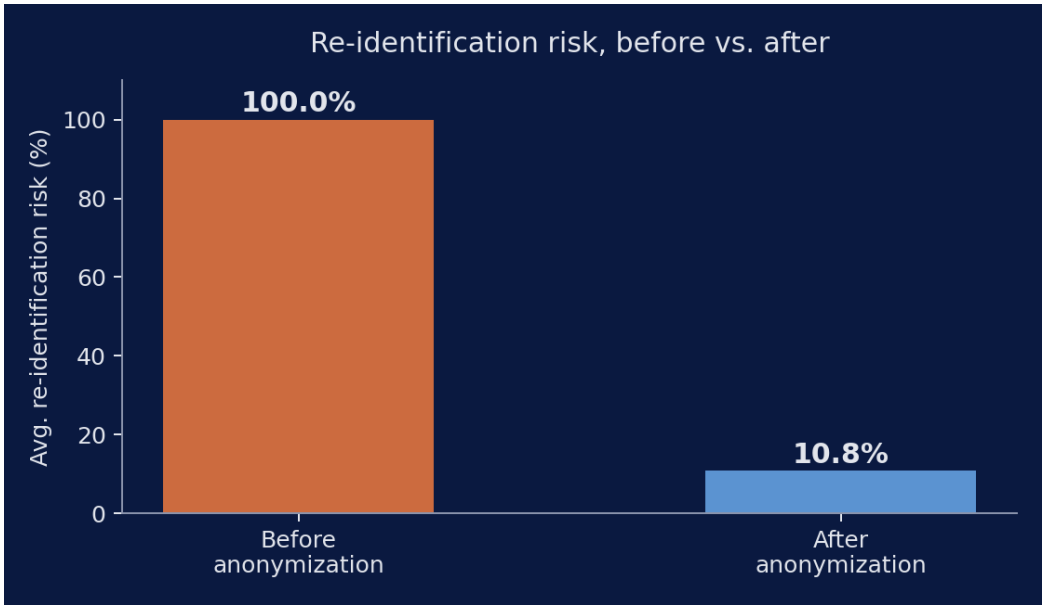
```
Discussed compensation and performance with [NAME REDACTED] (residence: [LOCATION REDACTED], Seattle). Flight risk assessment: Medium. Department: Engineering.
```

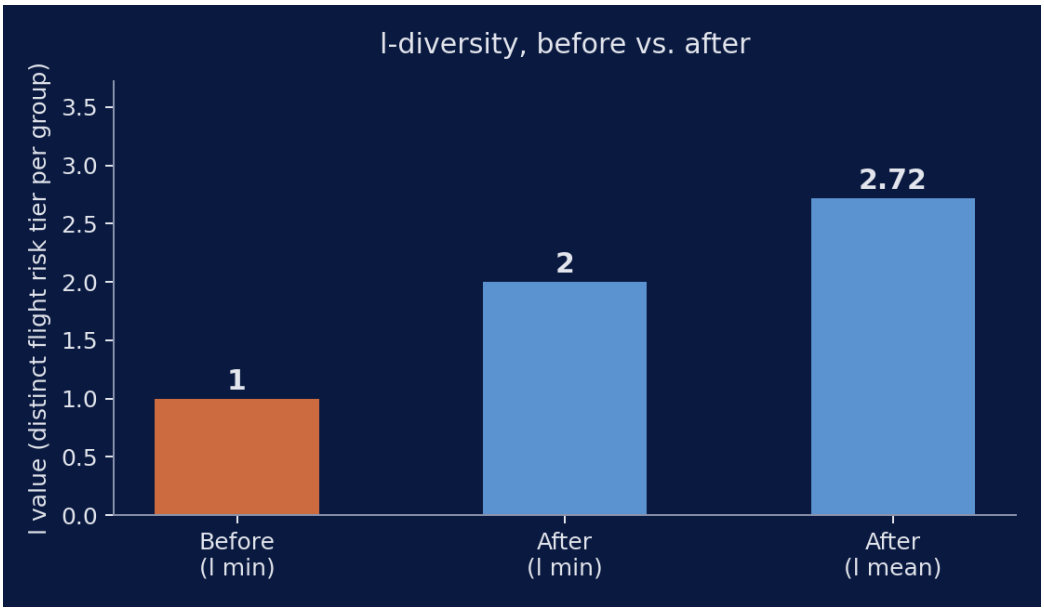
Note that city-level location ("Seattle") was deliberately retained — it's already generalized to region elsewhere in the pipeline, while the specific home address (a direct identifier on its own) is removed.

Anonymity Metrics

Two metrics are reported together deliberately. k-anonymity alone can be satisfied while still leaking sensitive information: if every record in a group shares the same sensitive value, an attacker who narrows a target to that group learns the answer with certainty even without picking out the exact record. l-diversity checks for that specific gap by measuring how many distinct flight risk tier values exist inside each k-anonymous group.







Metric	Before	After
Minimum k (k-min)	1	5
Mean k (k-mean)	1.0	9.3
Average re-identification risk	100.0%	10.8%
Minimum l-diversity (l-min)	1	2
Mean l-diversity (l-mean)	1.0	2.72
Groups with l=1 (attribute disclosure risk remains)	100.0%	0.0%

Honest finding: 0.0% of equivalence classes still have l=1 after anonymization — meaning k-anonymity was satisfied for those groups but attribute disclosure risk on flight risk tier was not fully eliminated. This is a known, real limitation of k-anonymity alone, not a bug: it's exactly why both metrics are measured and reported together rather than stopping at k-anonymity and calling the dataset safe. A production release gate would flag these specific residual groups for additional generalization or exclusion before delivery to a recipient sensitive to attribute disclosure.

Utility Metrics

Coverage Retention

100% of input records are present in the anonymized output. Where a record's quasi-identifier combination couldn't reach k-anonymity through normal generalization, the pipeline generalized its fields further rather than deleting the row — no record is ever silently dropped.

KPI Fidelity

Salaries are right-skewed (true mean $74,340 pulled up by senior-level compensation; median is lower). Two reconstruction methods were tested against this specific challenge:

Method	Reconstructed Mean	Error %
Naive band midpoint (no aggregate published)	$76,017.53	2.26%
Published bucket mean (aggregation release)	$74,340.34	0.0%

Honest finding: guessing from the band midpoint alone overestimates the true average salary by 2.26% — modest in this case, since HRMS compensation data is less extremely skewed than claims or transaction amounts, but the same underlying issue applies at any scale. Publishing the true per-bucket average alongside the generalized label — standard practice for aggregate reporting releases, such as pay-equity dashboards — reconstructs the KPI exactly, with no individual salary exposed.
