International Journal of Humanities and Information Technology (IJHIT)
e-ISSN: 2456 –1142, Volume 4, Issue 1-3, (August 2022), ijhit.info

Privacy by Architecture: Operationalizing GDPR Records of Processing Activities and Right-to-Erasure in Enterprise Analytics Platforms

Umashanker Bandari, Kalyan Kondisetty

Data, Analytics & AI Solutions Architect, Chicago, Illinois, USA

Associate Director, Wavicle Data Solutions, Chicago, Illinois, USA

ABSTRACT: The General Data Protection Regulation (GDPR) obliges controllers to keep a record of processing activities (ROPA) under Article 30 and to honour the right to erasure under Article 17. In many enterprises both duties are served by spreadsheets and ticket queues that have no executable link to the warehouses, data lakes and business-intelligence models in which personal data actually resides. This paper presents a privacy-by-architecture framework in which the ROPA becomes a machine-readable control plane. Processing activities are linked through a metadata graph to physical assets, columns, retention rules and subject-identity keys, and a deletion orchestrator derives, executes and evidences per-asset erasure plans across storage tiers. We describe the metadata model, the identity-resolution and key-management design, the operator self-check cycle, a hash-chained audit evidence trail and the penetration and erasure-verification testing approach, drawing on an implementation at a multinational manufacturing group operating on Azure and Snowflake. Particular attention is given to storage that resists deletion: append-only lake tables, time-travel and fail-safe retention, backups, derived aggregates and BI extracts, for which we combine crypto-shredding, tombstoning with deferred compaction and explicitly bounded residual-exposure windows. On representative data from the deployment, ROPA-to-asset coverage reached 87% of 1,840 in-scope assets within twelve months, median erasure turnaround fell from roughly three weeks to four days, and analyst effort per request fell by about 78%. We close by showing how the same metadata supports CCPA/CPRA, LGPD and PIPL obligations.

Keywords: GDPR, record of processing activities, right to erasure, crypto-shredding, metadata-driven governance, data lakehouse, privacy by design

DOI: 10.21590/ijhit.04.1-3.23

I. INTRODUCTION

Since the General Data Protection Regulation became applicable in May 2018 [1], every controller has carried two obligations that sound administrative but are, at bottom, systems problems. Article 30 requires a record of processing activities (ROPA) stating, for each activity, its purposes, categories of data subjects and personal data, recipients, third-country transfers and "where possible, the envisaged time limits for erasure". Article 17 grants a right to erasure without undue delay when, among other grounds, the data are no longer necessary or consent is withdrawn. Articles 5(2) and 25 require demonstrable compliance and data protection by design [2].

In practice both obligations live far from the systems they govern. The ROPA is a spreadsheet curated by the privacy office and refreshed annually; erasure requests become tickets fulfilled by engineers who search whichever systems they remember. Analytics platforms widen the gap: warehouses keep slowly changing history, lakes favour append-only and immutable storage, cloud services retain time-travel versions and backups, and BI tools cache imported copies and allow exports. Shastri et al. call such properties the "sins" of personal-data systems [3] and show that retrofitting compliance into existing databases is costly [4].

At the start of this work the case organization, a global Tier-1 industrial manufacturer operating in more than a dozen countries, held 146 processing activities in per-entity spreadsheets, none of which referenced a physical table, lake path or BI dataset. The analytics estate held roughly 1,840 assets containing or derived from personal data about employees, former employees, business-partner contacts and consumers who registered products. Erasure took about three weeks at the median and more than a month at the 90th percentile, and nobody could say with confidence that an erased subject was absent from backups or exported spreadsheets.

This paper argues that the ROPA should be a control plane for the platform rather than documentation about it: a machine-readable inventory from which asset mappings, retention schedules, erasure plans and audit evidence are derived. The contributions are:

1) A reference architecture separating governance, control, data and assurance planes around a metadata graph (Section 3.B).
2) A metadata model binding processing activities to data categories, retention rules, physical assets, columns and pseudonymous subject keys, with a coverage metric that makes ROPA completeness measurable (Section 3.C).
3) A catalogue of erasure strategies per storage tier, including crypto-shredding, tombstoning with deferred compaction and an explicit residual-exposure window (Section 3.F).
4) An assurance design: operator self-check, a hash-chained evidence trail that does not re-identify subjects, and penetration and canary-probe testing (Sections 3.G to 3.I).
5) A representative evaluation from a two-year deployment and a mapping to CCPA/CPRA, LGPD and PIPL (Sections 4 and 5).

II. BACKGROUND AND RELATED WORK

A. Regulatory obligations and standards

Article 17(3) lists exceptions to erasure, of which two dominate enterprise analytics: compliance with a legal obligation (for example, statutory retention of payroll and tax records) and the defence of legal claims. Article 12(3) requires a response within one month, extendable by two further months for complex or numerous requests, and Article 19 obliges the controller to inform each recipient of the erasure. Storage limitation (Article 5(1)(e)) also makes erasure due when retention ends. The ROPA is also the natural input for DPIA screening [5]. Pseudonymised data remain personal data (Recital 26), and the Working Party cautions that aggregation and generalisation reduce but do not automatically remove identifiability [6]. ISO/IEC 27701 adds privacy controls, including records of processing and deletion at the end of processing, to an ISO/IEC 27001 management system [7]; it assumes an accurate inventory, which spreadsheet ROPAs cannot guarantee.

B. Systems research on GDPR compliance

Shastri et al. identify seven patterns of data systems that conflict with GDPR, such as data without a time-to-live and no way to find all of a person's data [3]; their GDPRbench study shows that adding metadata indexing, time-to-live and logging to general-purpose databases carries large performance penalties [4]. Schwarzkopf et al. propose compliance by construction, with each user's data in a logical shard and application state as materialized views over the shards, so that erasure removes a shard [8]. Kraska et al. sketch SchengenDB, which enforces purpose-based access, auditing and deletion natively [9]. Such greenfield designs cannot be imposed on heterogeneous commercial platforms. Our framework instead retrofits comparable guarantees as a control layer, accepting bounded rather than instantaneous deletion where the storage engine forces it.

C. Deletion in modern storage and metadata management

Reardon, Basin and Capkun show that deletion at one layer of the storage stack rarely implies deletion at the layers below [10]. Boneh and Lipton made backups revocable by encrypting data under keys that are later destroyed [11], the ancestor of crypto-shredding, and NIST SP 800-88 recognises cryptographic erase as sanitization when keys are properly managed [12]. Log-structured storage applies deletes as tombstones at compaction time [13]; Delta Lake similarly writes new files and a log entry on DELETE while superseded files stay readable through time travel until VACUUM removes them [14]. Snowflake retains changed data for a configurable Time Travel period plus a seven-day Fail-safe that customers cannot shorten [15], and dimensional warehouses keep history in type 2 slowly changing dimensions [16]. On the metadata side, DAMA-DMBOK recommends integrating business, technical and operational metadata [17], and catalogues such as Azure Purview, generally available since September 2021, scan sources, classify sensitive data and capture lineage [18]. Catalogues know nothing of processing activities, lawful bases or erasure; we use them only as a discovery feed.

III. THE ROPA-AS-CONTROL-PLANE FRAMEWORK

A. Design principles

Six principles, agreed with the data protection officer (DPO), were acceptance criteria:

- **Single authoring point.** Activities are authored once; the Article 30 export, retention schedules and erasure plans are derived artefacts.
- **Asset-level binding.** Every in-scope asset is bound to at least one activity or marked out of scope with a reason; an unbound asset is a finding.
- **Strategy belongs to storage.** How a subject is erased depends on the asset's storage mode, not on the request.
- **Evidence without re-identification.** The audit trail proves what was done without storing clear identifiers.

- **Bounded residuals.** Where immediate physical removal is infeasible, the residual window is computed and accepted in advance.
- **A weak deleter.** The deleting identity holds only the privilege the metadata justifies.

B. Architecture

Fig. 1 shows the four planes. The **governance plane** holds what privacy and legal staff own: the ROPA, the DPIA and retention register (with exemptions and legal holds), and request intake with identity verification. The **control plane** turns this into executable metadata. A metadata graph links activities to data categories, assets and columns; a catalog scanner (Purview scans plus nightly schema-difference jobs) proposes mappings; a retention scheduler converts rules into due dates; and an identity-resolution service maps request identifiers to pseudonymous subject tokens and per-subject data-encryption keys (DEKs). The deletion orchestrator consumes all three and fans tasks out to the **data plane**, where one connector per tier implements the strategies of Section 3.F. The **assurance plane** receives hash-chained receipts, runs the self-check and hosts testing.

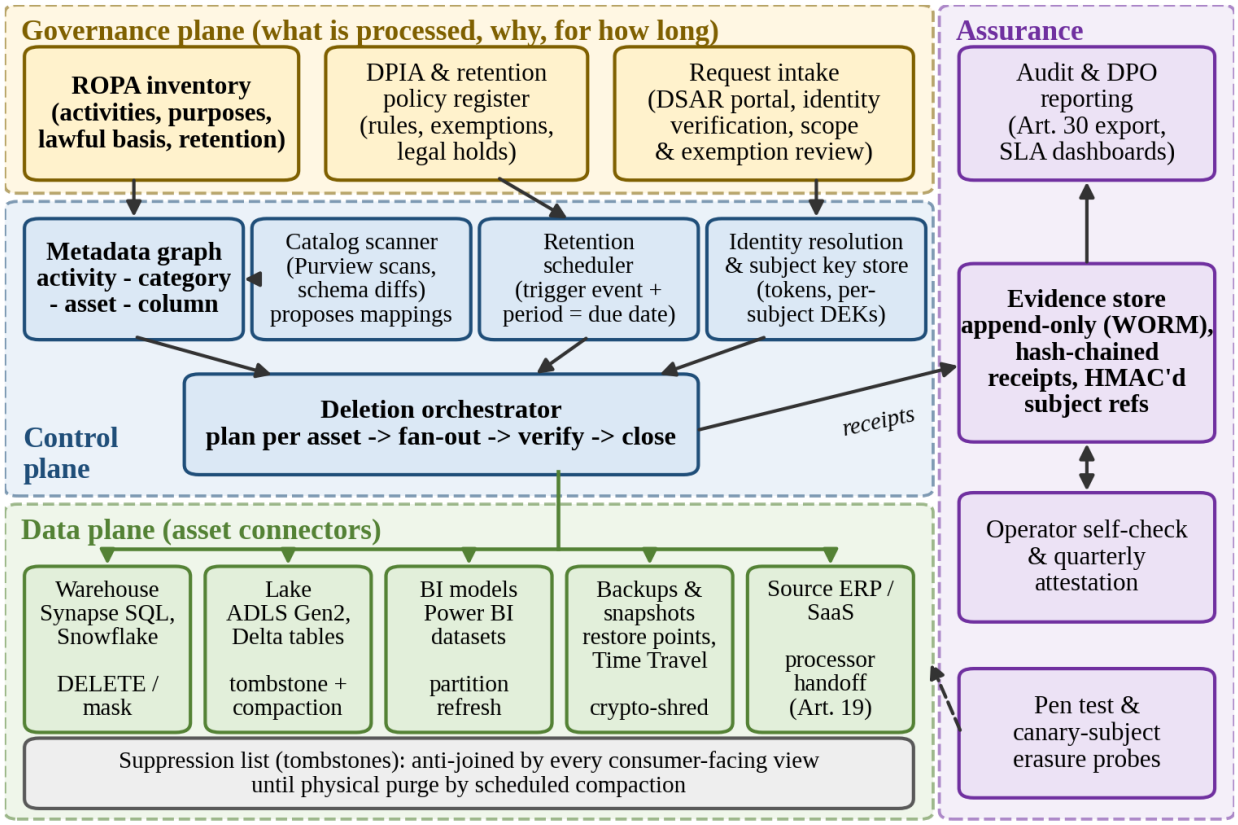


Fig.1. ROPA-as-control-plane architecture: governance metadata is compiled into a metadata graph that drives per-tier asset connectors, with receipts flowing to an append-only evidence store.

In the author's deployment the graph is a relational schema with edge tables in Azure SQL Database behind an internal REST API; a graph database was rejected because traversals are shallow (at most four hops) and the privacy office needed ordinary reporting access. Purview's Apache Atlas-compatible API supplies scanned assets and classifications. The orchestrator is a set of Azure Data Factory pipelines calling one Azure Function per connector, with a durable retry queue. Evidence lands in an immutable ADLS Gen2 container; per-subject DEKs are wrapped by a key-encryption key (KEK) in Azure Key Vault.

C. Metadata model

Table 1 summarises the model. First, the activity-to-asset link is many-to-many and carries a **role** (system of record, replica, derived dataset or extract) that determines erasure ordering and whether a lawful basis is inherited. Second, classification lives on the AssetElement (column, lake path segment or BI measure), because one wide table can mix personal and non-personal attributes. Third, scanner mappings enter a **proposed** state and become **confirmed** only after

a data steward accepts them; only confirmed mappings count or drive deletion. This stopped part numbers that matched a national-ID pattern from generating erasure tasks against engineering tables.

Table.1. Metadata model entities, key attributes and uses

Entity	Key attributes	Used for
ProcessingActivity	purpose, lawful basis, controller, subject and data categories, recipients, transfers, DPIA reference	Art. 30 export; DPIA screening
DataCategory	name, special-category flag (Art. 9), sensitivity tier, pseudonymisable	Classification; strategy choice
RetentionRule	activity x category, trigger event, period, exemption code, legal hold	Scheduler; exemption screening
DataAsset	tier, platform, qualified name, owner, storage mode, strategy, time-travel and backup retention	Planning; residual window
AssetElement	column/path/measure, category, subject-key flag, encrypted flag, classification source	Task targeting; coverage
ActivityAsset (link)	activity, asset, role (SoR, replica, derived, extract), state (proposed, confirmed)	Coverage; ordering
SubjectKey	subject token, identifier HMACs by type, DEK reference, systems seen, status	Identity resolution; crypto-shredding
ErasureRequest / Task	regime, receipt and due timestamps, exemptions; per-asset strategy, attempts, row counts, state	SLA tracking; verification
EvidenceRecord	task, executor, query hash, result digest, previous and own hash, residual-window end	Audit; attestation

Coverage, the programme's headline metric, is defined over the set A of in-scope assets (those with at least one element classified into a personal-data category); it is reported per tier and legal entity, because an unmapped asset is invisible to every downstream control:

$$C = | \{ a \in A : \exists\, p, \text{confirmed link}(p, a) \} | / | A | \quad (1)$$

D. Identity resolution and subject keys

A request arrives with a name or e-mail address; in the data the same person may appear as an HR personnel number, an ERP business-partner number or a CRM contact identifier. The identity service keeps one SubjectKey per person with a random token and identifier HMACs: each identifier type is normalised and hashed with HMAC-SHA-256 under a type-specific secret in Key Vault, so matching needs no clear identifiers. Matching is deterministic on HMAC equality; probabilistic candidates go to a privacy analyst and are never deleted automatically, since erasing another person's data is itself a breach. At first ingestion into the persisted lake zone the service creates a random 256-bit DEK for the subject, wrapped under the KEK, and direct identifiers in that zone are encrypted under it.

E. Deletion orchestration

For each verified request the orchestrator executes six steps:

1) **Due date.** Derived from the request's regime; for GDPR, t_due = t_receipt + 1 month, with an extension flag requiring DPO approval.
2) **Exemption screening.** Rules with an exemption code or legal hold remove asset-category pairs from scope; each exclusion is written to evidence with its legal reason.
3) **Suppression first.** The subject token is added to a global suppression list that every consumer-facing view anti-joins. This takes effect in minutes and stops re-propagation from lagging upstream loads while physical work proceeds.
4) **Planning.** The planner traverses the graph from the token's systems and from subject-key elements, producing one task per (asset, key element) with the asset's strategy, ordered upstream to downstream (sources and processors, lake, warehouse, BI) so that no refresh re-imports removed data.
5) **Execution and verification.** Connectors run idempotent operations and record row counts; a re-probe with the subject's key values must return nothing, otherwise the task retries and, after three attempts, escalates to the asset owner.

6) **Closure.** The subject is answered once all logical tasks are verified; deferred physical purges stay pending until their residual window (Equation 2) elapses and a final probe confirms absence.

F. Erasure strategies per storage tier

Table 2 lists the strategy per storage type. Strategies differ mainly in how quickly the last physical copy disappears, captured as a residual-exposure window for asset a:

$$W(a) = \Delta c(a) + Rtt(a) + Rb(a) \quad (2)$$

where Δc is the interval to the next compaction or rewrite, Rtt the time-travel or version retention and Rb the backup retention (for crypto-shredded data, the retention of the key store's own backups). W is computed from metadata, approved by the DPO at design time and stamped on each evidence record.

Table.2. Erasure strategy per storage type in the case deployment

Storage type	**Constraint**	**Strategy**	**Residual window, evidence**
Warehouse tables (Synapse SQL, Snowflake)	Versions kept by Time Travel or restore points	Set-based DELETE by subject key; facts re-keyed to an anonymous member	Snowflake 1 d + 7 d Fail-safe; Synapse 7 d; row counts
Type 2 dimensions	History is the purpose	Delete all subject versions; re-key dependent facts	As above; version counts
Lake Delta tables	Copy-on-write; log versions	Tombstone now; weekly batched MERGE; VACUUM at 168 h	≤ 14 d; log version ids
Lake raw zone (WORM)	Immutable by policy	Identifiers tokenised and encrypted under per-subject DEK; crypto-shred	Key-store backups, 14 d; DEK receipt
Backups, snapshots	Not record-addressable	Crypto-shred if encrypted; else expire, replaying suppression on restore	≤ 35 d; restore runbook
Aggregates, marts	Contribution inseparable	Keep if all cells ≥ k = 10 and no subject key; else recompute	Next rebuild; cell-size check
BI import models	Incremental refresh freezes history	Refresh affected partitions via XMLA endpoint	≤ 1 d after source; partitions
BI extracts, exports	Outside the platform	Export disabled on PII datasets; activity-log discovery; attestation	Attestation; findings
Source ERP, SaaS processors	Owned by others	Article 19 ticket with SLA	Processor confirmation

Warehouses. Deleting a contact from a dimension would orphan fact rows whose measures (order quantities, warranty costs) serve legitimate non-personal purposes, so such facts are re-keyed to a reserved anonymous member, removing the link to the person while preserving totals. Snowflake Time Travel on PII schemas was cut to one day, leaving the seven-day Fail-safe as a documented residual [15].

Lake tables: tombstoning with deferred compaction. A Delta DELETE rewrites every file containing a matching row [14]; per request this meant rewriting many files for a handful of rows, which was slow and costly in the pilot. Instead the suppression list acts as a tombstone, as in LSM trees [13]: curated views exclude the subject at once, a weekly job applies all pending tombstones in one MERGE per table, and VACUUM with the default 168-hour retention removes superseded files. The DPO accepted a 14-day maximum residual in exchange for predictable cost and for keeping time travel, on which engineering teams rely for recovery.

Immutable storage and backups: crypto-shredding. The persisted raw zone is under a WORM policy for audit reasons, and backups are not addressable per record. There, identifiers are replaced by the subject token and, where the original must remain recoverable, encrypted under the subject's DEK; erasure destroys the wrapped DEK, following the revocable-backup idea [11] and cryptographic erase [12]. Two caveats are recorded: the true residual is the key store's own backup retention (14 days), and ciphertext remains pseudonymised rather than anonymised data [6], so crypto-

shredding is presented as putting data beyond use. Restores of unencrypted backups replay the suppression list before any copy becomes readable.

Aggregates, BI models and extracts. Aggregates are kept when every cell covers at least k = 10 subjects and no subject key survives, following k-anonymity [19]; smaller cells, such as headcount at a small site, are recomputed. Differentially private release [20] was considered but deferred. Import-mode Power BI datasets with incremental refresh never refresh historical partitions, so a subject deleted at source would persist for years; the connector refreshes only the affected partitions through the Premium XMLA endpoint. Export was disabled on datasets with personal elements, and historical exports were traced in the activity log and attested by owners.

G. Operator self-check

Automation cannot certify that people have not made shadow copies, so each asset owner completes a quarterly self-check. The control plane generates a worksheet of automated findings in six categories: unmapped assets, orphan BI extracts, stale retention rules (records older than the rule permits), evidence gaps, activities missing a lawful basis, and over-broad delete rights (connector permissions exceeding the mapped asset set). Owners resolve or justify each finding and sign an attestation. The cycle mirrors the plan-do-check-act loop of a privacy information management system [7] and the stewardship roles recommended by DAMA [17].

H. Audit evidence trail

Each task writes an EvidenceRecord with request and task identifiers, executor identity, a hash of the executed statement, a result digest (row counts, partition or file identifiers), the residual-window end and a chained hash:

$$h_i = \text{SHA-256}(h_{(i-1)} \,\|\, \text{canon}(r_i)) \qquad (3)$$

where canon(r_i) canonically serialises record i. The chain head is anchored daily in the immutable container, so alteration or deletion is detectable. Subjects are referenced only by HMAC. Retaining even an HMAC of an erased subject is a deliberate, documented decision: without it the controller could neither demonstrate erasure (Article 5(2)) nor block re-ingestion from a lagging source. It is kept for three years and is useless without the annually rotated secret.

I. Penetration testing and erasure verification

The erasure machinery is new attack surface. The threat model covered fraudulent requests that destroy another person's data, abuse of the orchestrator's estate-wide delete identity, evidence tampering, portal enumeration, and recovery from residual copies. The portal was tested against the OWASP Web Security Testing Guide [21], connector identities were reviewed against the graph, and evidence tampering was simulated to confirm that chain verification fails. Residual recovery is tested with **canary subjects**: synthetic persons with realistic identifiers injected at source, propagated by normal pipelines, erased through the ordinary request path, and then probed for in every mapped asset, BI dataset and accessible version history.

IV. EVALUATION

A. Setup and metrics

The framework was introduced as a manual baseline in 2020, a warehouse-only pilot in the first quarter of 2021, and orchestration across all tiers from the second quarter of 2021. The in-scope estate comprised 1,840 assets: 620 warehouse tables and views, 910 lake datasets and 310 BI datasets and managed extracts. We report coverage (Equation 1) per tier; turnaround from receipt to closure of all logical tasks (median and 90th percentile per quarter); analyst effort per request by tier; and the residual rate, the fraction of canary probe locations still returning a canary after closure.

Evaluation setup and data disclosure: all figures in this section are representative values derived from a small evaluation in the case organization and from practitioner observation, normalized and anonymized to protect confidential organizational data. Turnaround covers 336 verified erasure requests received between January 2020 and December 2021, measured from ticket and orchestrator timestamps. Effort is based on time logs for a sample of 60 requests (30 manual, 30 orchestrated). The residual rate comes from two canary runs of 40 synthetic subjects each, seeded into an average of 12 assets per subject (480 probe locations per run). Coverage is taken from monthly snapshots of the metadata graph. The values are indicative of the magnitude of change and are not statistically generalizable to other organizations or platforms.

B. Coverage

Fig. 2 shows coverage during the 2021 rollout. The warehouse tier started highest, because its assets were already documented in dimensional models, and reached 95%. BI models reached 88%, helped by lineage to the warehouse in pipeline metadata. The lake ended at 82%: its assets are path-based, schema-on-read and numerous, and many were sandbox or archived landing folders whose owners had moved on. Weighted by tier, coverage reached 87%; the remaining 13% is the most important caveat of the evaluation, because every downstream control, including the canary probes, is blind to unmapped assets.

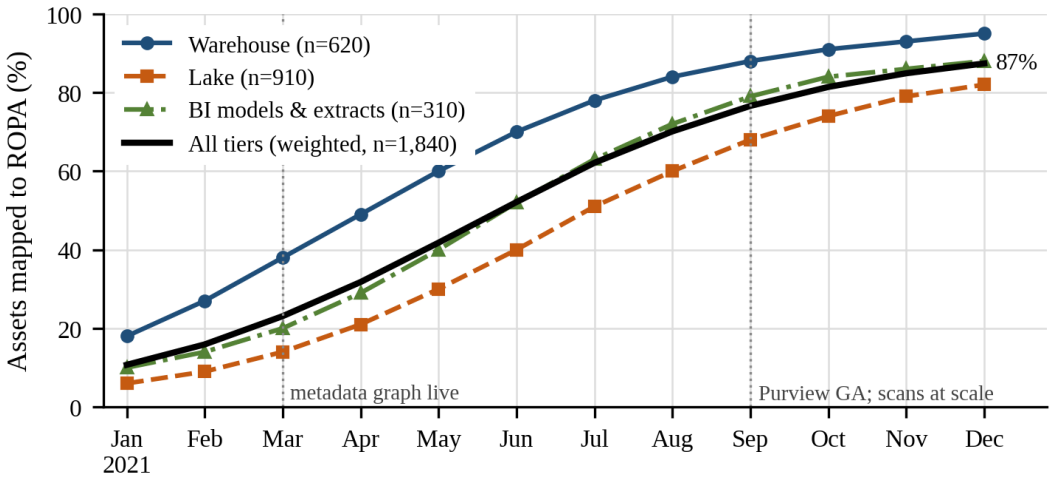


Fig.2. Percentage of in-scope assets with a confirmed ROPA mapping during the 2021 rollout, by storage tier and weighted overall (representative data).

C. Turnaround and effort

Fig. 3 reports turnaround per quarter. Under the manual process the median ranged from 19 to 24 days and the 90th percentile from 35 to 44 days, beyond the one-month limit; extensions under Article 12(3) were invoked for 17 of the 135 requests of 2020. The pilot cut the median to 14 days. With full orchestration the median fell to 4 days and the 90th percentile to 9 days by the fourth quarter of 2021, with no extensions among the 157 orchestrated requests, although quarterly volume nearly doubled over the two years. Physical closure is tracked separately: computed residual windows ranged from 8 days (Snowflake: one day of Time Travel plus seven of Fail-safe) to 35 days (stores with 35-day point-in-time backups), with a median of 14 days set by the weekly lake compaction and VACUUM cycle.

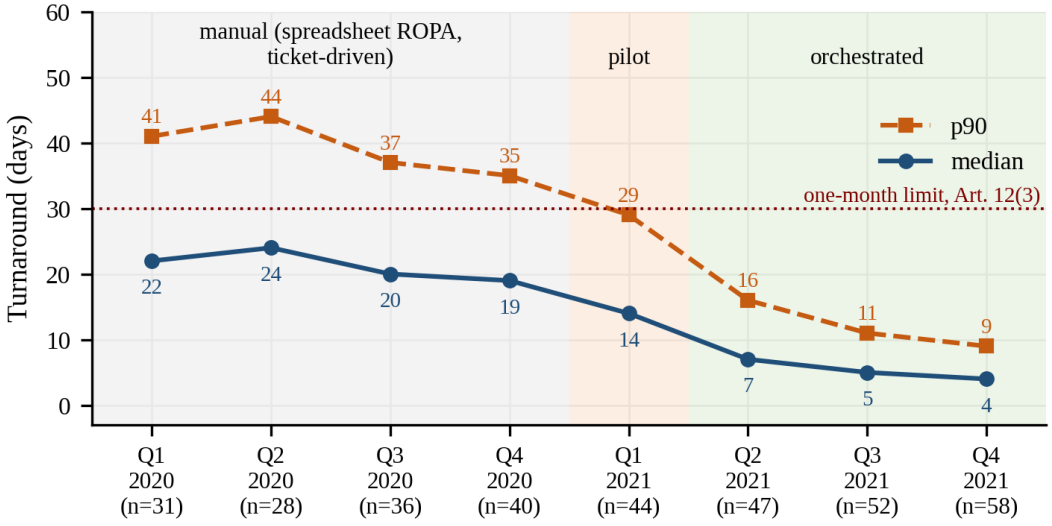


Fig.3. Erasure-request turnaround (median and 90th percentile, days) per quarter before and after orchestration; n is the number of verified requests (representative data).

Fig. 4(a) breaks down analyst effort, which fell from 14.3 to 3.2 person-hours per request, about 78%. The lake shows the largest absolute reduction (5.2 to 0.9 hours), because manual lake erasure had meant writing ad hoc Spark jobs per request. Source and processor hand-offs gained least (2.0 to 1.1 hours); they remain conversations with other system owners and are the next target for integration.

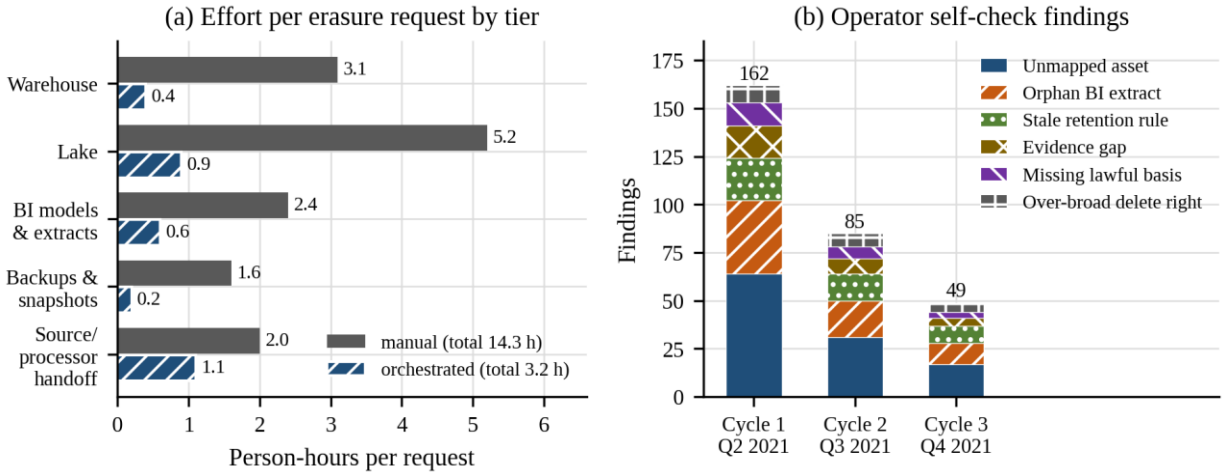


Fig.4. (a) Analyst effort per erasure request by storage tier, manual versus orchestrated (sample of 60 requests); (b) operator self-check findings per quarterly cycle by category (representative data).

D. Self-check, canary probes and penetration test

Fig. 4(b) shows self-check findings falling from 162 to 85 to 49 over three cycles. Unmapped assets (64) and orphan BI extracts (38) dominated the first cycle and both fell by more than two thirds. Over-broad delete rights fell only from 9 to 5, since each fix requires a negotiated permissions change.

The first canary run (second quarter of 2021) found canaries at 37 of 480 probe locations after closure (7.7%): 15 in BI extracts and personal workspaces, 11 in the raw lake zone before per-subject encryption was introduced there, 6 in type 2 dimension history not yet handled by the warehouse connector, and 5 in Power BI partitions outside the incremental-refresh window. The second run (fourth quarter) found 6 of 480 (1.25%): 4 superseded Delta file versions awaiting VACUUM and 2 Snowflake Time Travel versions, all inside documented residual windows; a re-probe after the windows elapsed found none. This measures completeness on mapped assets only.

The penetration test in the third quarter of 2021 produced 13 findings: none critical, 2 high, 6 medium and 5 low. The high findings were an orchestrator service principal with owner rights on two Synapse workspaces far beyond its mapped assets, and an insecure direct object reference in the portal's status endpoint allowing enumeration of requests by sequential identifier. Medium findings included e-mail-only verification for former employees and clear e-mail addresses in retry-queue messages; high findings were fixed before the second canary run.

V. DISCUSSION

A. Lessons learned

- **Residual windows, not turnaround, measure forgetting.** Most turnaround gain came from knowing where to look; how completely a platform forgets is set by Time Travel, VACUUM and backup settings usually chosen for recovery alone.
- **Coverage is a stewardship problem.** Coverage rose fastest where a named steward understood each asset's purpose.
- **Encrypt at ingestion.** Crypto-shredding cannot be retrofitted onto immutable data; zones predating per-subject encryption had to be re-ingested.
- **BI is the long tail.** Import models and extracts produced more residuals than the far larger lake.

B. Preparing for regulations beyond GDPR

Because obligations are metadata attributes rather than code paths, a new regime means new reference data and a few rules (Table 3). The CCPA grants a right to delete with a 45-day response period, extendable by 45 days, flowing down to service providers [22]; the CPRA, approved in November 2020 and operative from 2023, adds contractors, retention

disclosure and forthcoming risk-assessment regulations [23]. Brazil's LGPD provides for elimination of data processed on consent and requires records of processing operations [24]. China's PIPL, in force since November 2021, requires deletion when the purpose is achieved or retention expires and impact assessments retained for at least three years [25]. A request's due date is the minimum of its regimes' deadlines, and among overlapping jurisdiction-specific retention rules the shortest period not overridden by a statutory exemption wins; neither rule required connector changes.

Table.3. Deletion and record-keeping obligations across regimes and the metadata attributes that serve them

Regime	**Deletion right and timing**	**Record and assessment duties**	**Metadata used**
GDPR (EU)	Art. 17; one month, extendable by two; Art. 19 notice to recipients	Art. 30 ROPA; Art. 35 DPIA	lawful_basis, recipients, transfers, RetentionRule, dpia_ref
CCPA / CPRA (California)	§1798.105; 45 days, extendable by 45; flow-down to service providers (and contractors under CPRA)	Notice at collection; CPRA retention disclosure, risk assessments	regime, recipient role, disclosed retention
LGPD (Brazil)	Art. 18 elimination of consent-based data; deletion at end of processing	Art. 37 records; Art. 38 impact report on request	lawful_basis (Art. 7), jurisdiction
PIPL (China)	Art. 47 deletion when purpose achieved, retention expired or consent withdrawn	Arts. 55–56 assessment, kept ≥ 3 years	jurisdiction, transfer mechanism, assessment_ref

C. Threats to validity and limitations

This is a single-organization case with moderate request volume and no control group; the turnaround gain coincides with a maturing privacy office and cannot be attributed to the architecture alone. The effort sample is small and self-reported. Canary probes measure completeness only on mapped assets and known identifiers; free text such as ticket notes and scanned documents relies on incomplete classification heuristics. The aggregate threshold and the treatment of crypto-shredding as putting data beyond use are policy judgements agreed with one DPO, and a regulator could take a stricter view. The implementation reflects Azure, Snowflake and Power BI as of 2021: the design is portable, the residual windows are not. For practitioners, the central implication is to decide residual windows with the DPO when configuring Time Travel, VACUUM and backup retention, rather than after the first request.

VI. CONCLUSION

In enterprise analytics platforms, GDPR's record of processing activities and right to erasure are better treated as a control plane than as paperwork: a metadata graph that binds activities to assets, columns, retention rules and pseudonymous subject keys, and drives per-tier deletion strategies with verifiable evidence. The hard cases, including append-only lake tables, time-travel history, immutable zones, backups, aggregates and BI extracts, cannot all be erased instantly, but they can be erased within a computed, documented and tested window.

Several directions remain open. Machine-learning models trained on personal data are a further copy of it; unlearning techniques such as sharded training [26] suggest how retraining could be scoped by the same graph. Privacy-preserving release [20] could replace fixed cell-size thresholds for widely shared aggregates. Finally, expressing retention and erasure policy as versioned, reviewed code would bring the control plane under the same engineering discipline as the platforms it governs.

REFERENCES

[1] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," Official Journal of the European Union, L 119, pp. 1–88, May 2016.

[2] European Data Protection Board, "Guidelines 4/2019 on Article 25: Data protection by design and by default," Version 2.0, adopted Oct. 2020.

[3] S. Shastri, M. Wasserman, and V. Chidambaram, "The seven sins of personal-data processing systems under GDPR," in Proc. 11th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud), 2019.

[4] S. Shastri, V. Banakar, M. Wasserman, A. Kumar, and V. Chidambaram, "Understanding and benchmarking the impact of GDPR on database systems," Proc. VLDB Endowment, vol. 13, no. 7, pp. 1064–1077, 2020.
[5] Article 29 Data Protection Working Party, "Guidelines on data protection impact assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679," WP 248 rev.01, Oct. 2017.
[6] Article 29 Data Protection Working Party, "Opinion 05/2014 on anonymisation techniques," WP 216, Apr. 2014.
[7] ISO/IEC 27701:2019, "Security techniques — Extension to ISO/IEC 27001 and ISO/IEC 27002 for privacy information management — Requirements and guidelines," International Organization for Standardization, Geneva, 2019.
[8] M. Schwarzkopf, E. Kohler, M. F. Kaashoek, and R. Morris, "Position: GDPR compliance by construction," in Heterogeneous Data Management, Polystores, and Analytics for Healthcare (Poly/DMAH 2019), LNCS vol. 11721, Springer, 2019, pp. 39–53.
[9] T. Kraska, M. Stonebraker, M. Brodie, S. Servan-Schreiber, and D. Weitzner, "SchengenDB: A data protection database proposal," in Heterogeneous Data Management, Polystores, and Analytics for Healthcare (Poly/DMAH 2019), LNCS vol. 11721, Springer, 2019, pp. 24–38.
[10] J. Reardon, D. Basin, and S. Capkun, "SoK: Secure data deletion," in Proc. IEEE Symposium on Security and Privacy, 2013, pp. 301–315.
[11] D. Boneh and R. J. Lipton, "A revocable backup system," in Proc. 6th USENIX Security Symposium, San Jose, CA, USA, 1996.
[12] R. Kissel, A. Regenscheid, M. Scholl, and K. Stine, "Guidelines for media sanitization," NIST Special Publication 800-88, Rev. 1, Dec. 2014.
[13] P. O'Neil, E. Cheng, D. Gawlick, and E. O'Neil, "The log-structured merge-tree (LSM-tree)," Acta Informatica, vol. 33, no. 4, pp. 351–385, 1996.
[14] M. Armbrust et al., "Delta Lake: High-performance ACID table storage over cloud object stores," Proc. VLDB Endowment, vol. 13, no. 12, pp. 3411–3424, 2020.
[15] Snowflake Inc., "Understanding & using Time Travel" and "Understanding & viewing Fail-safe," Snowflake Documentation, 2021.
[16] R. Kimball and M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd ed. Indianapolis, IN, USA: Wiley, 2013.
[17] DAMA International, DAMA-DMBOK: Data Management Body of Knowledge, 2nd ed. Basking Ridge, NJ, USA: Technics Publications, 2017.
[18] Microsoft, "Azure Purview documentation: Introduction to Azure Purview," Microsoft Docs, 2021.
[19] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557–570, 2002.
[20] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014.
[21] OWASP Foundation, "Web Security Testing Guide," Version 4.2, Dec. 2020.
[22] California Consumer Privacy Act of 2018, Cal. Civ. Code §§ 1798.100–1798.199, 2018.
[23] California Privacy Rights Act of 2020 (Proposition 24), approved by California voters, Nov. 2020.
[24] Brazil, "Lei Geral de Proteção de Dados Pessoais (LGPD)," Law No. 13.709, Aug. 2018.
[25] Standing Committee of the National People's Congress, "Personal Information Protection Law of the People's Republic of China," adopted Aug. 20, 2021, effective Nov. 1, 2021.
[26] L. Bourtoule et al., "Machine unlearning," in Proc. IEEE Symposium on Security and Privacy, 2021, pp. 141–159.