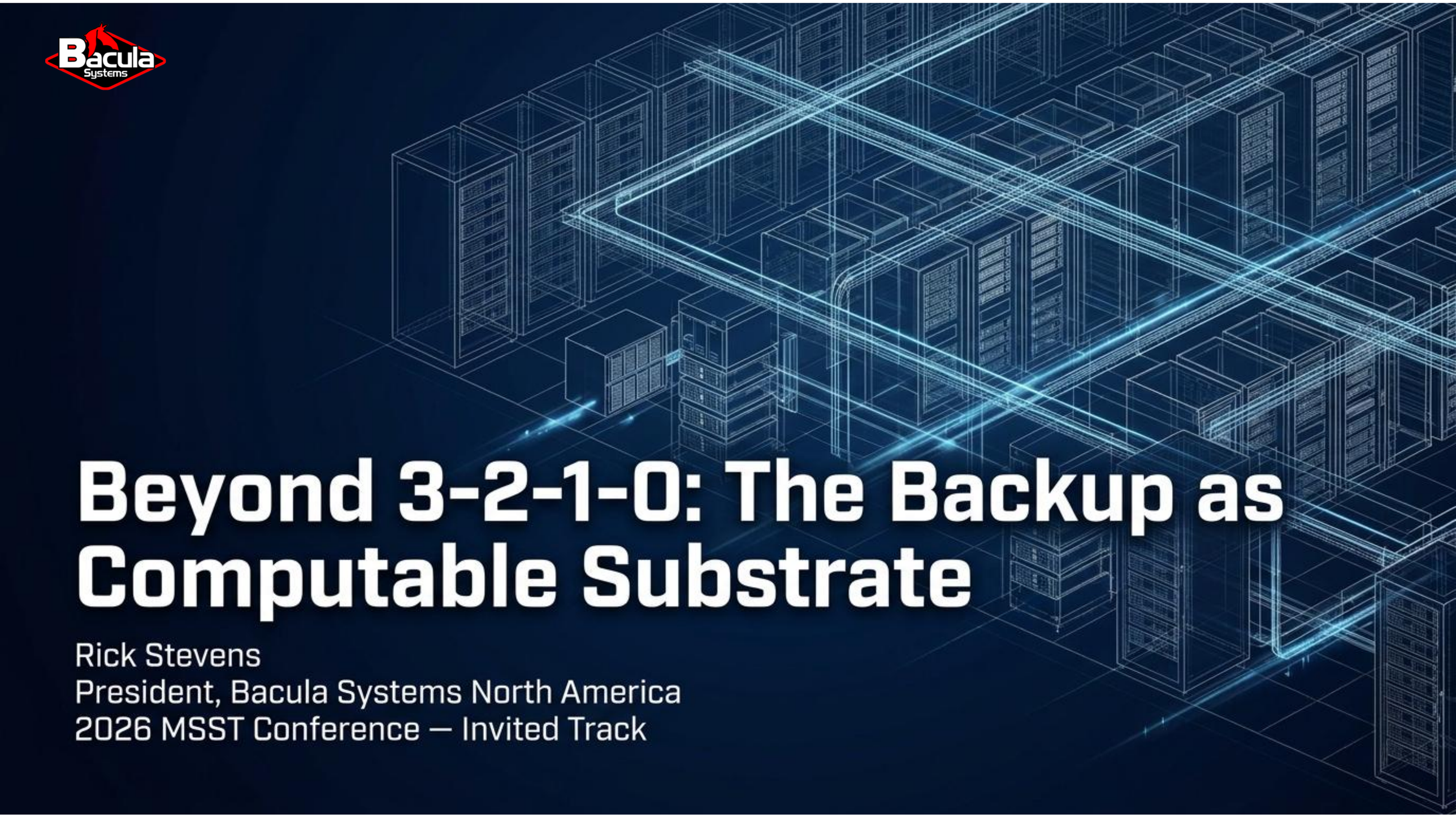


Beyond 3-2-1-0: The Backup as Computable Substrate

Rick Stevens
President, Bacula Systems North America
2026 MSST Conference — Invited Track

The Time-Indexed Mirror

The Live Tier

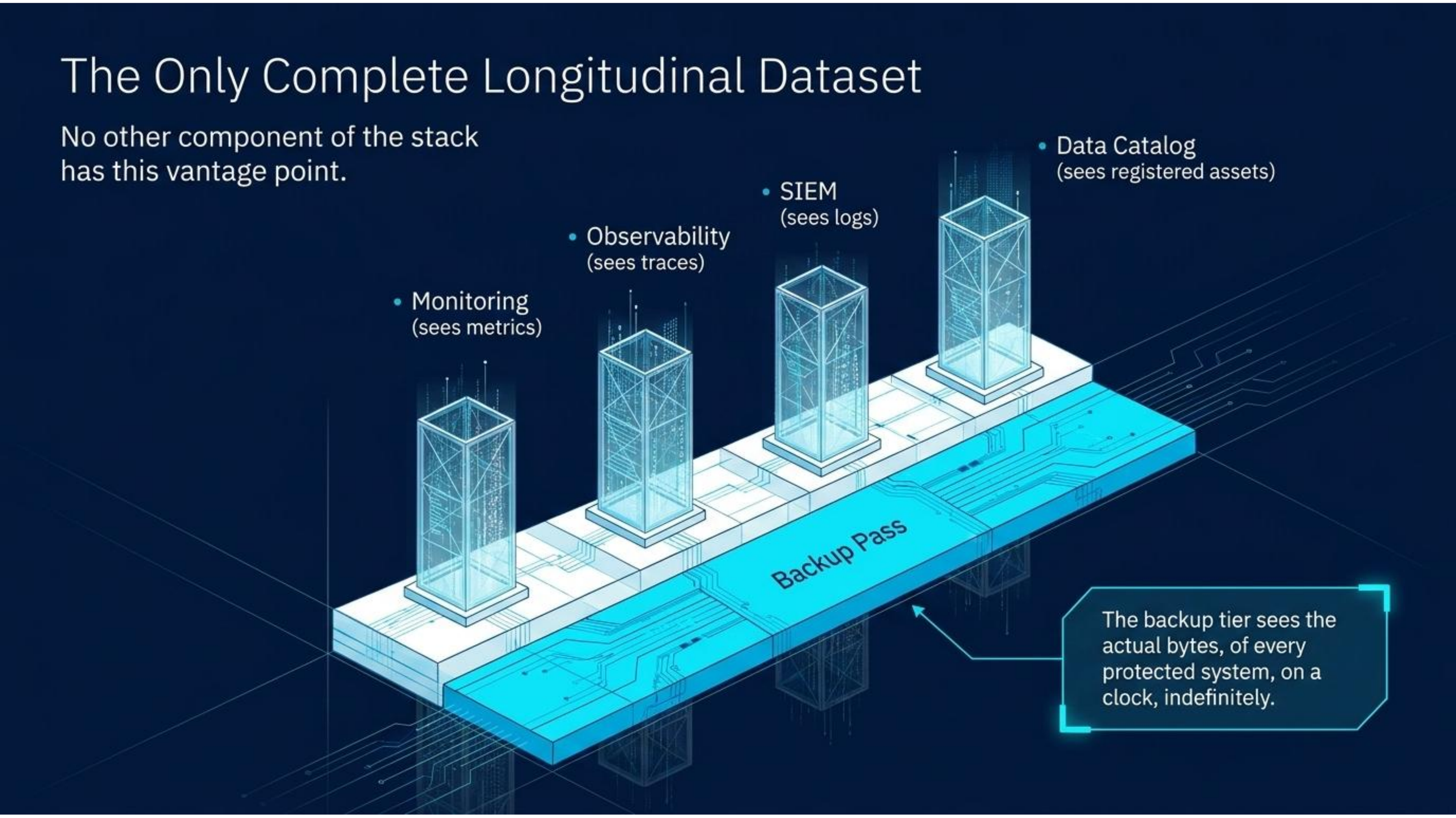
Two decades of building metadata search and query infrastructure at HPC scale.

The Secondary Tier

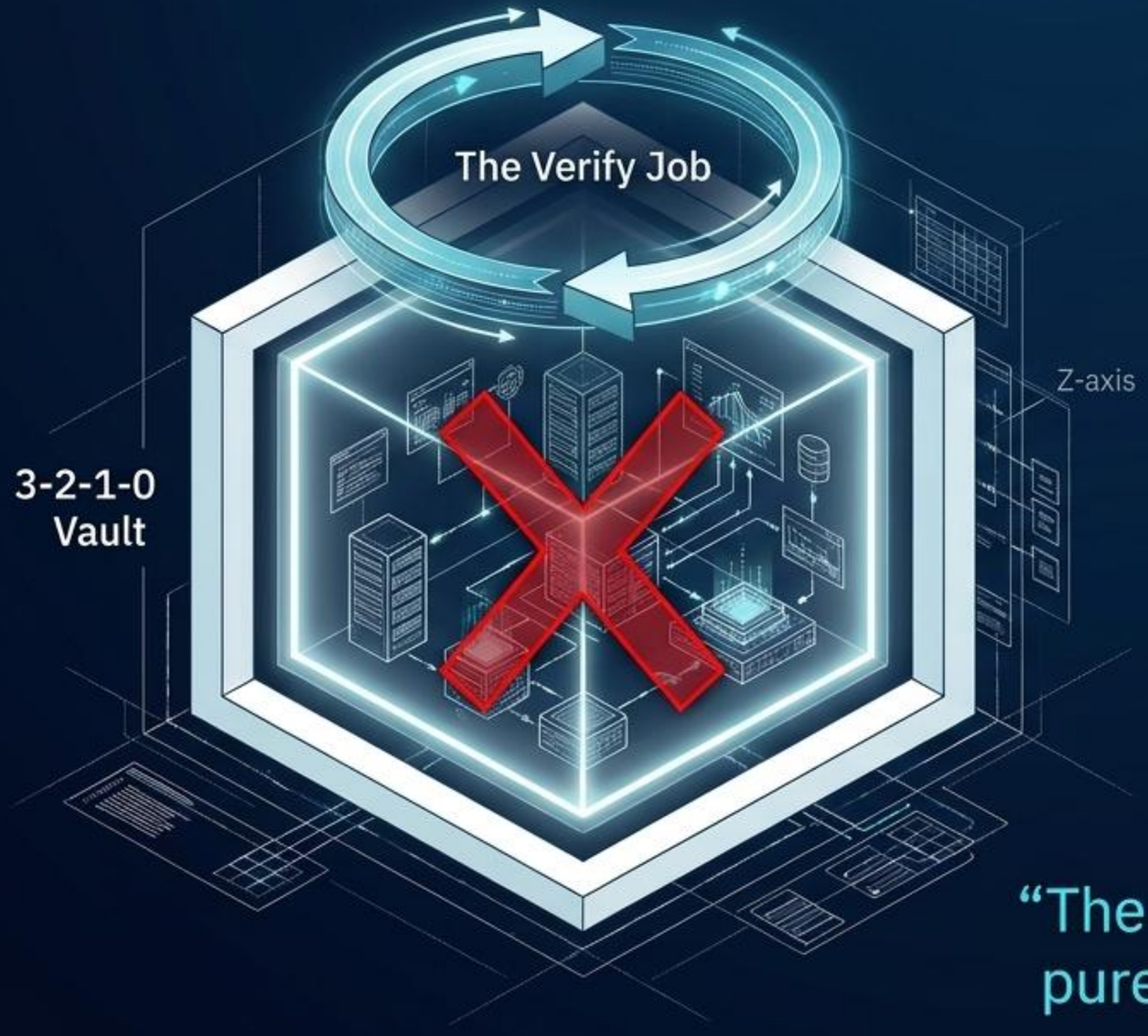
The next frontier is applying that same rigor to the time-indexed mirror of our workloads.

The Only Complete Longitudinal Dataset

No other component of the stack has this vantage point.



Establishing the Guardrails



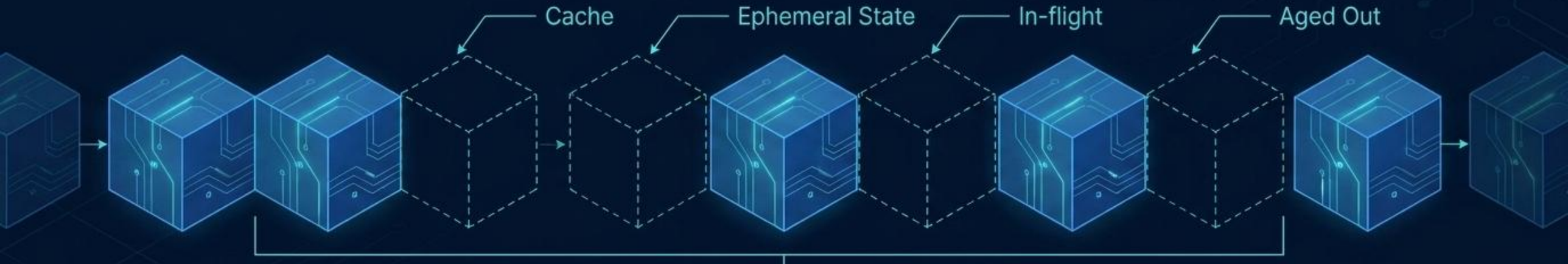
What I am not arguing:

- ❌ Hosting analytic workloads on the active recovery chain
- ❌ Turning the production repository into a queryable data lake
- ❌ Abandoning 3-2-1-0 survivability disciplines

"The zero was the field's first step beyond pure survivability. There are more steps."

A Shift in Framing

The secondary tier is a uniquely valuable longitudinal evidence layer.

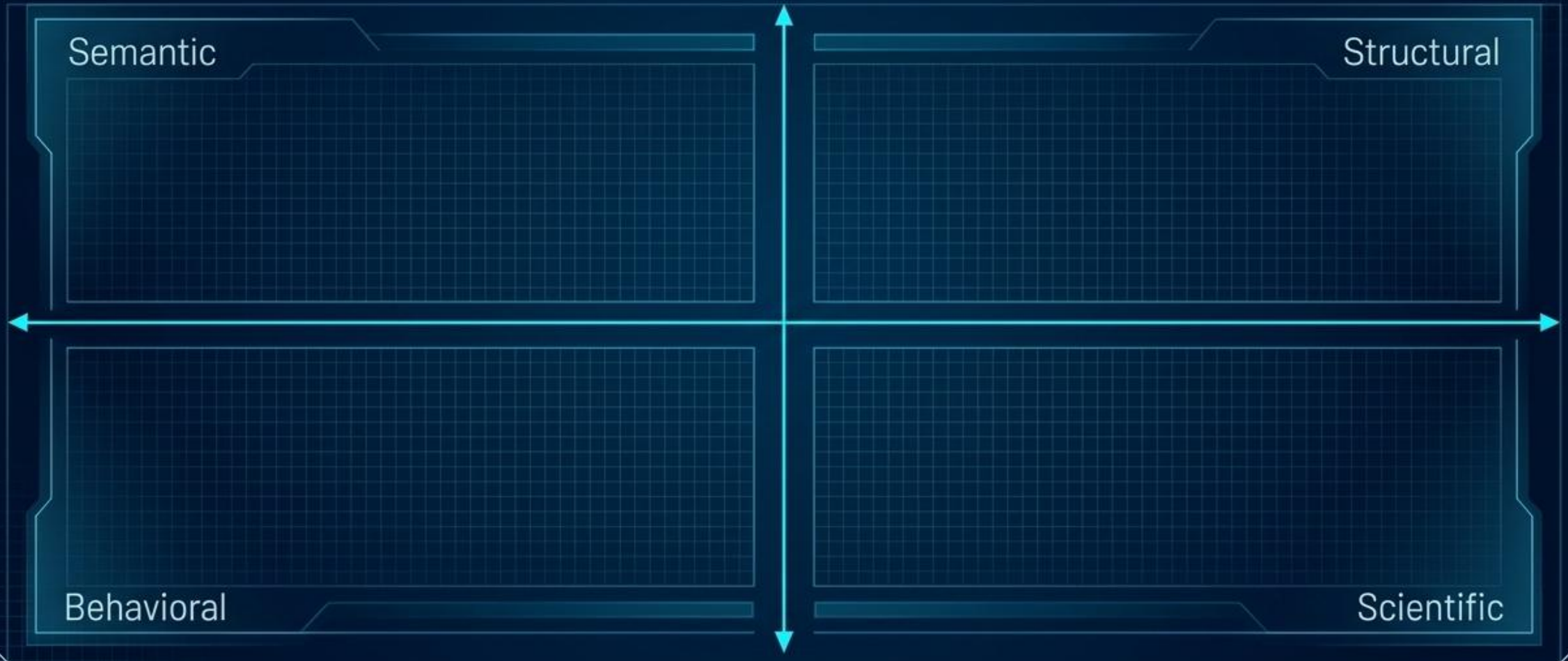


Longitudinal evidence layer — not canonical record.

Bounded by retention policies, structured by schedule capture, and complete enough across the enterprise to unlock capabilities no live system can match.

Four Dimensions of the Substrate

Treating the secondary tier as an evidence layer opens four families of problems for the storage research community.

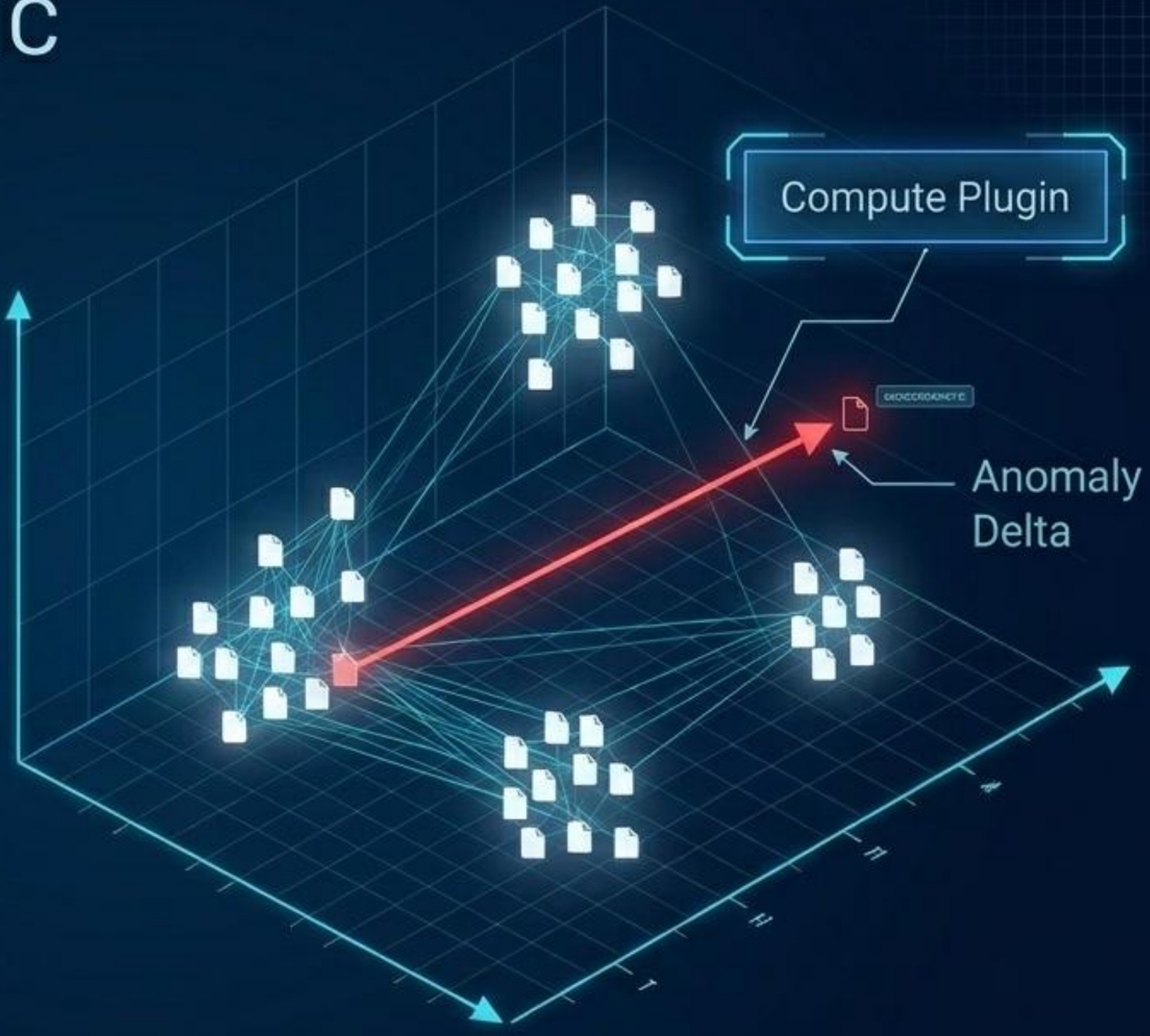


Axis One: Semantic

What do the bytes mean?

Instead of hash-and-store alone, we unlock:

1. Retrieval over history
2. Similarity indexing
3. Anomaly detection via embedding-space delta



The Architectural Honesty of Post-Backup Embedding

The decision of what to embed cannot responsibly be made in the hottest part of the ingest pipeline.
Selection must happen before embedding.

	Inline/Ingest Embedding	Post-Backup Plugin
Backup Window Impact	Severe bottleneck	Zero impact
Model Version Drift	Locked to past model	Re-computable with modern models
Legal/Eligibility Control	Forces blind processing	Selective by explicit policy
Processing Load	Hottest part of pipeline	Decoupled schedule

Substrates for Retrieval-Augmented Generation

The backup tier is not a substitute for current-tier RAG. It is the only honest substrate for use cases where history is the point.

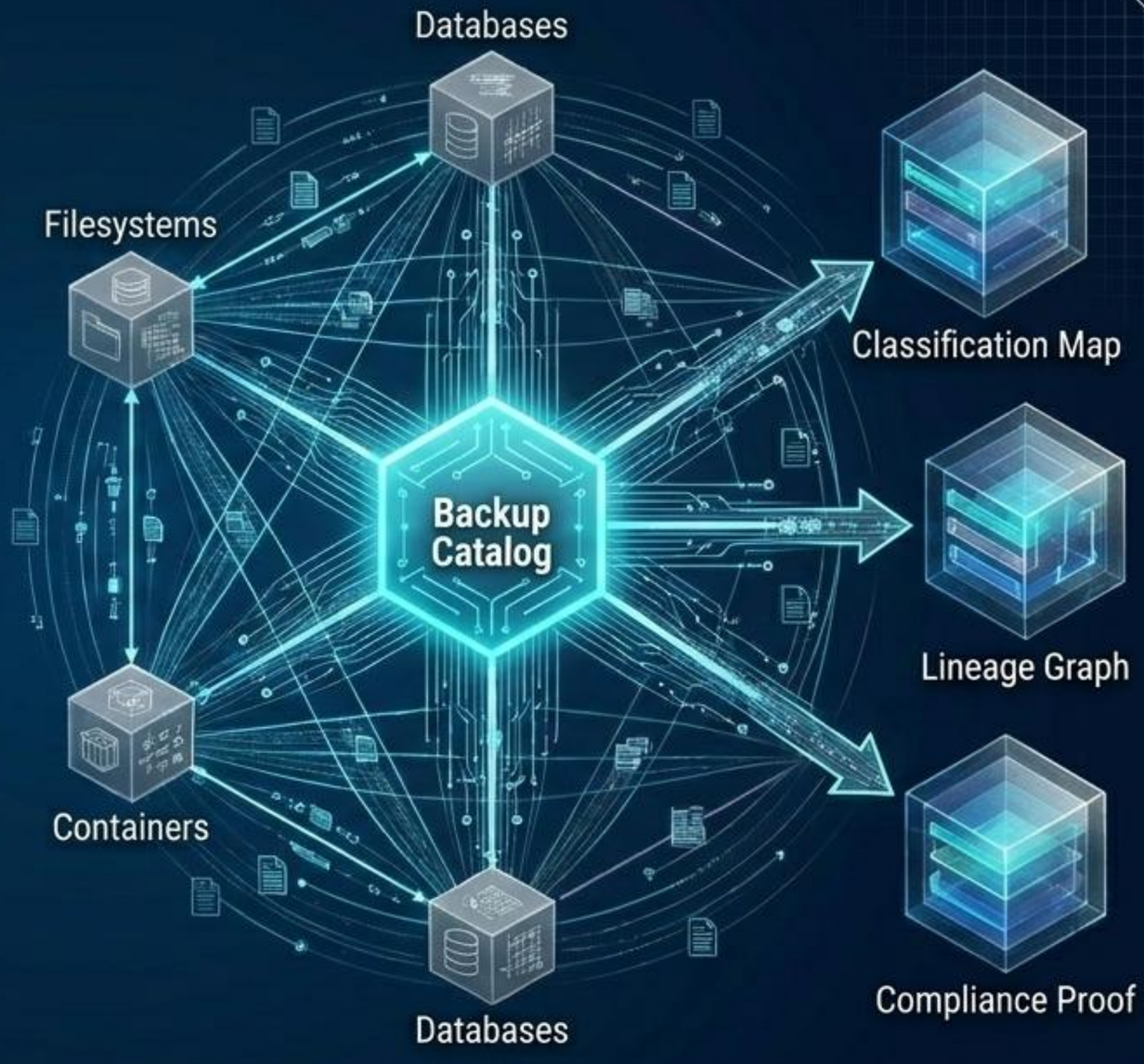
	Live-Tier Suitability	Backup-Tier Suitability
Real-time fraud detection	Yes ✓	No ✗
Customer Service	Yes ✓	No ✗
e-Discovery	No ✗	Yes ✓
Forensic Reconstruction	No ✗	Yes ✓
AI Provenance	No ✗	Yes ✓

Axis Two: Structural

What are the bytes, where did they come from, and who can see them?

- Continuous Classification (PII, PHI, CUI) derived from the index, not discovery projects.
- Edge Reconstruction for dataset lineage.
- Compliance Attestation (Immutable proofs over the audit window).

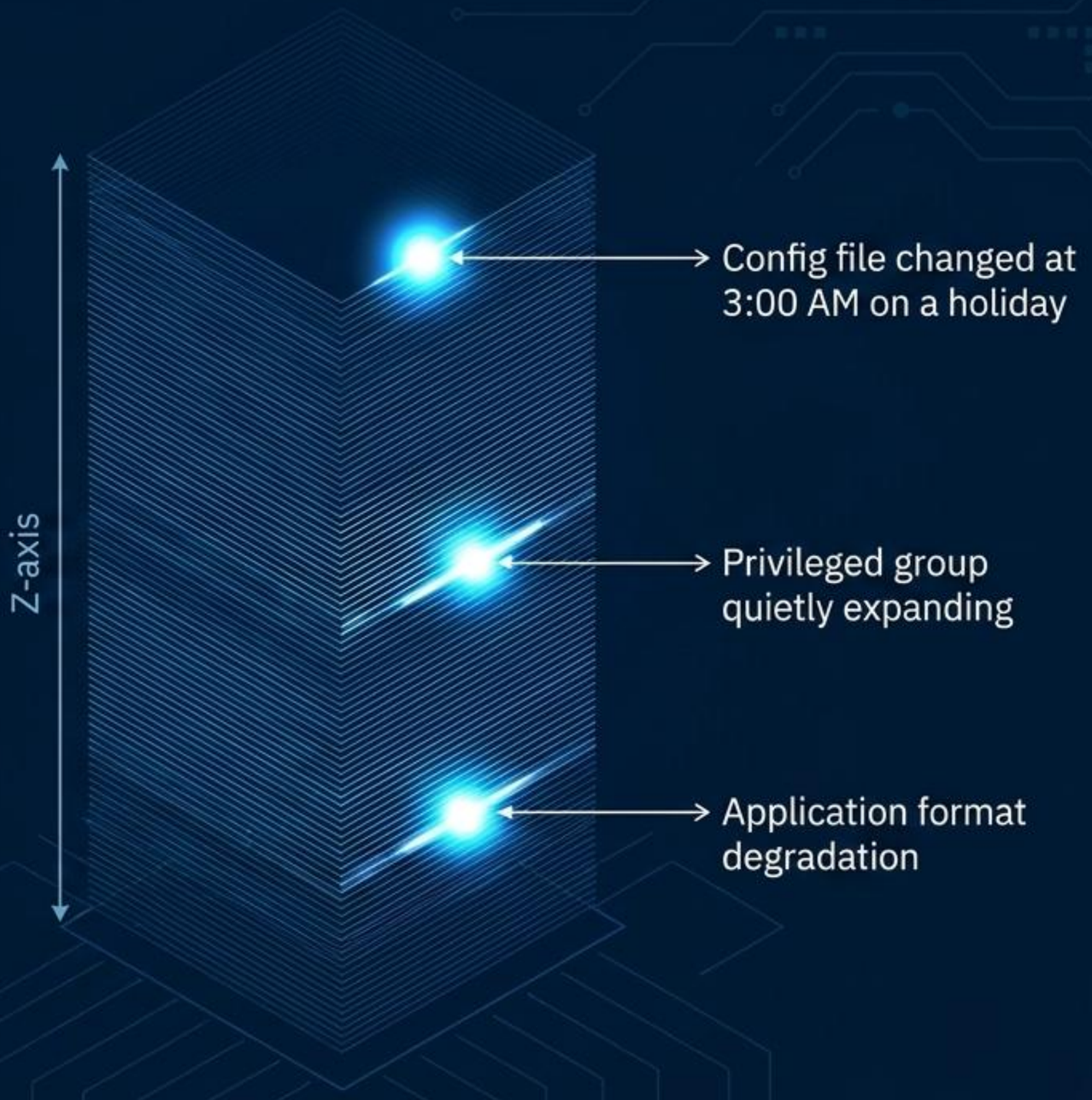
Building on prior art: The Grand Unified File Index (GUFI) demonstrated exhaustive metadata indexing is cheap at the right point in the pipeline.



Axis Three: Behavioral

The snapshot history is a four-dimensional time series: every object, every property, at every snapshot, across the retention horizon.

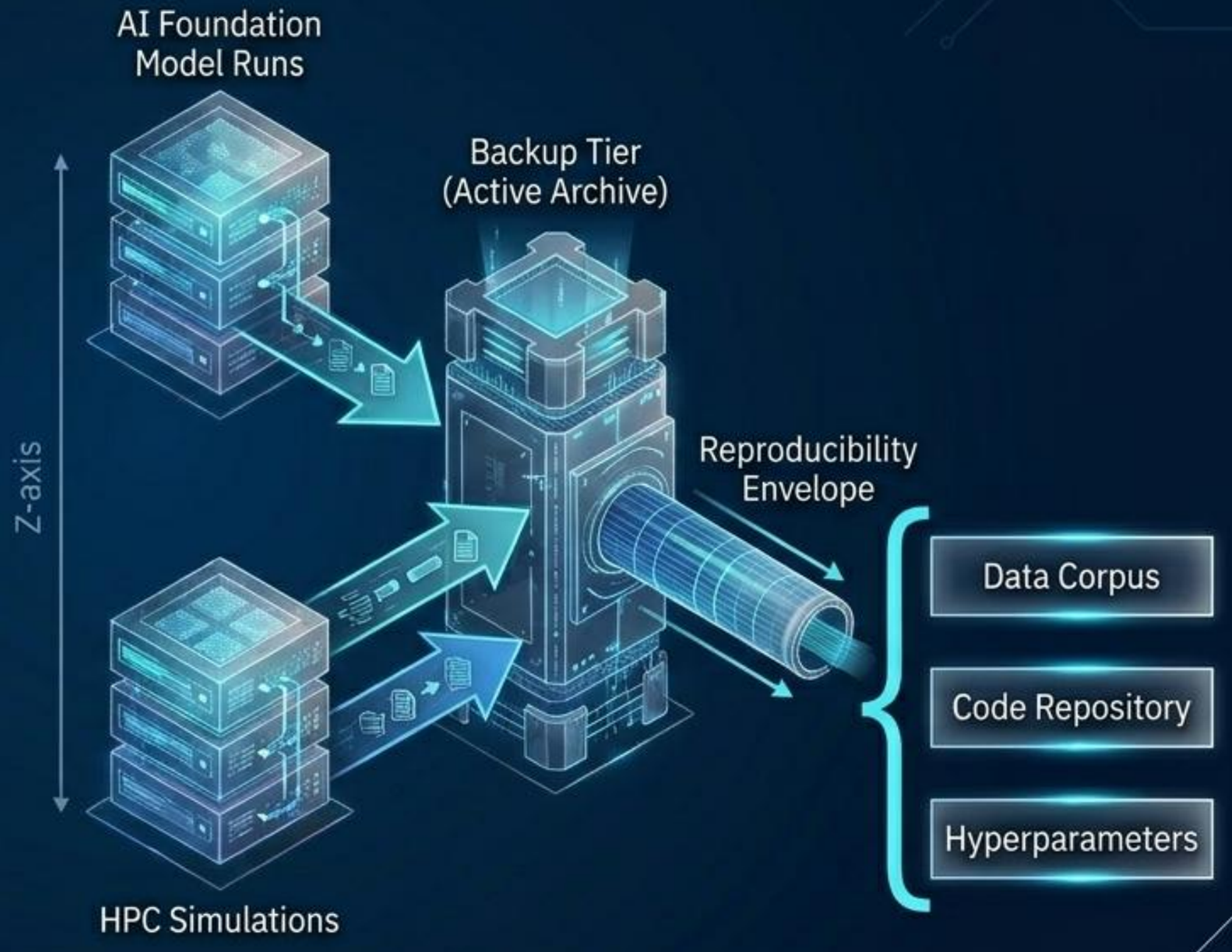
- Configuration drift analysis beneath the IaC layer
- Retrospective scanning for supply-chain compromise footprints
- Integrity and format erosion over decade-scale retention



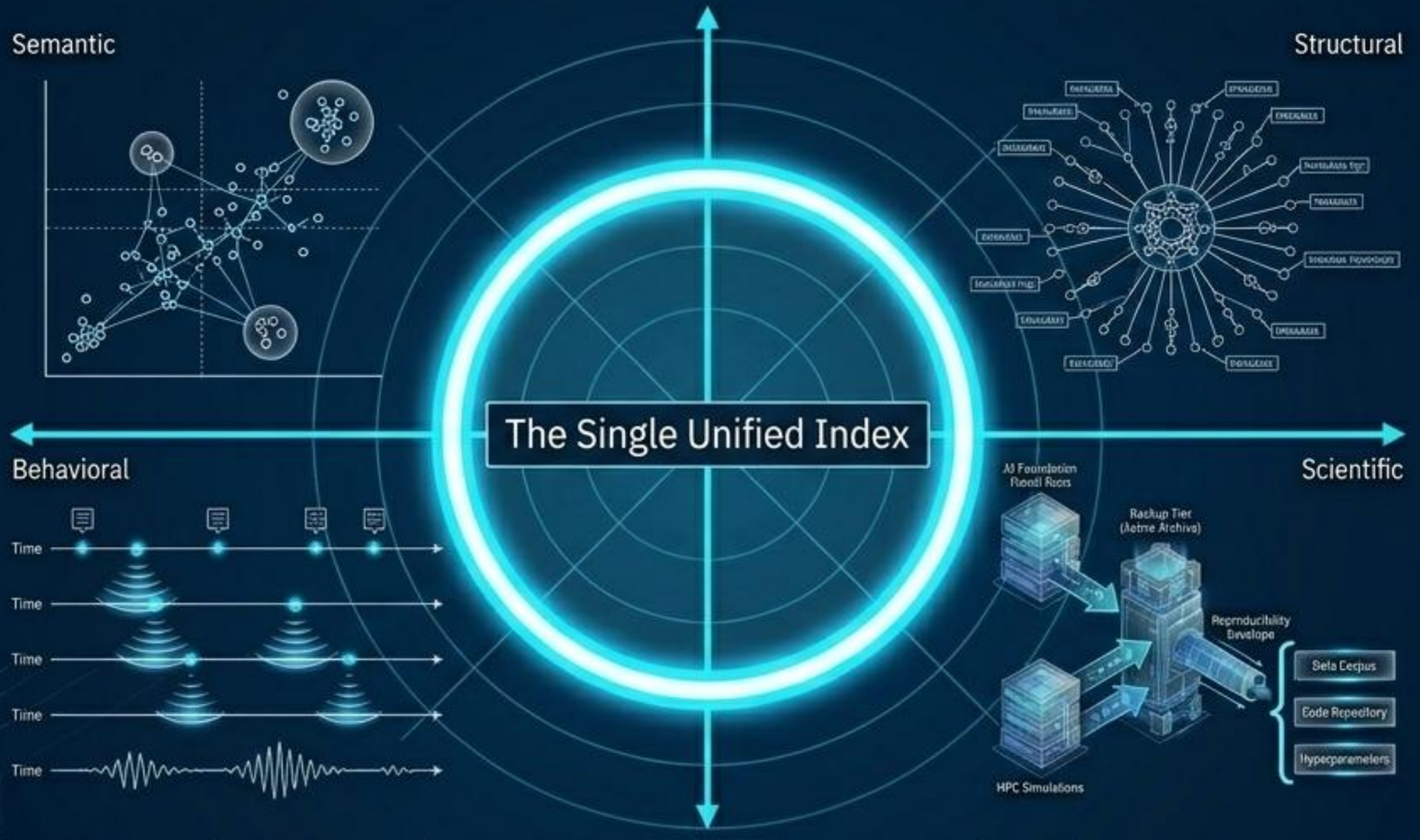
Axis Four: Scientific

For AI and HPC, the secondary tier is the de facto active archive.

- AI training is a reproducibility problem with a hardware problem layered on top.
- The backup catalog uniquely captures data, code, and configuration at a single point in time.
- Lab federated metadata and enterprise ‘data meshes’ are structurally the exact same problem.



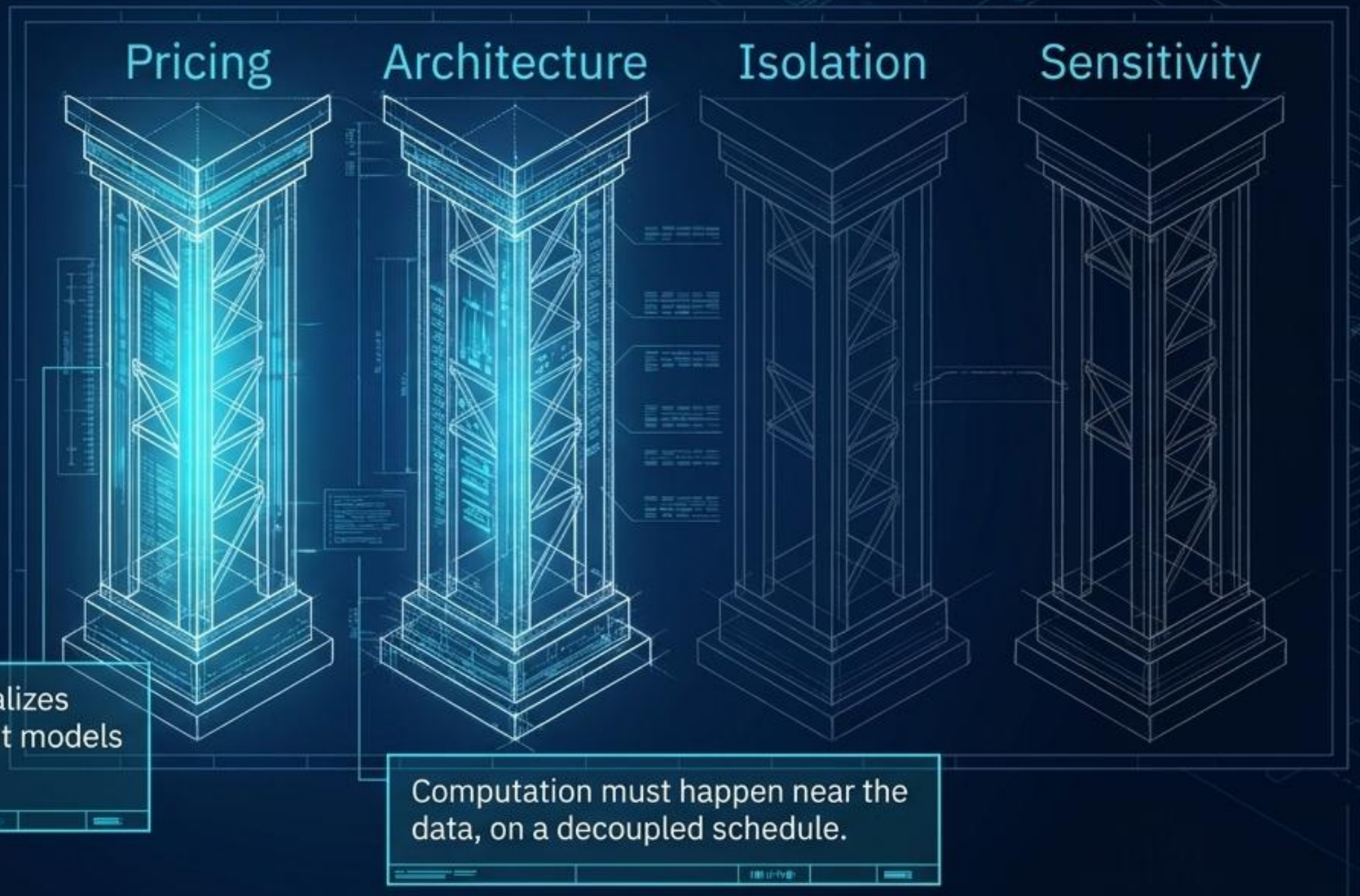
The Substrate Synthesis



AI Reproducibility (Scientific), Compliance Attestation (Structural), and Anomaly Detection (Behavioral) are not separate products. They are entirely different queries run against the exact same index, derived from the exact same semantic substrate. Solve the architecture once.

Systems Requirements

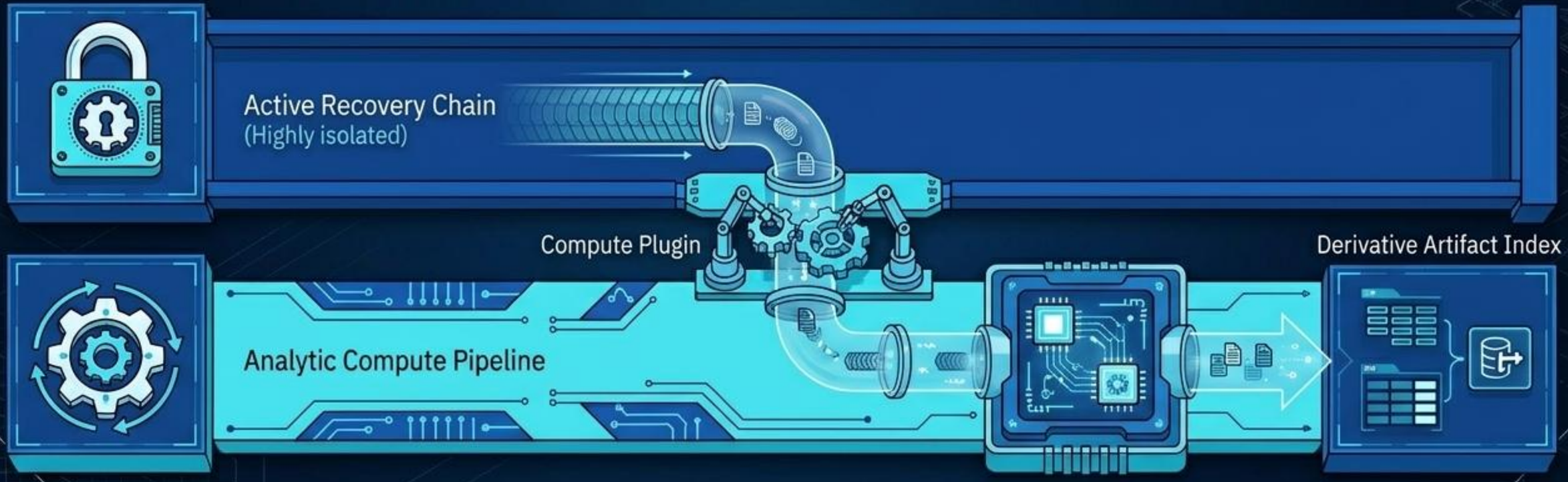
None of this is free. Advancing this framing requires solving four systems-level constraints at exabyte scale:



The Compute Plugin Pattern

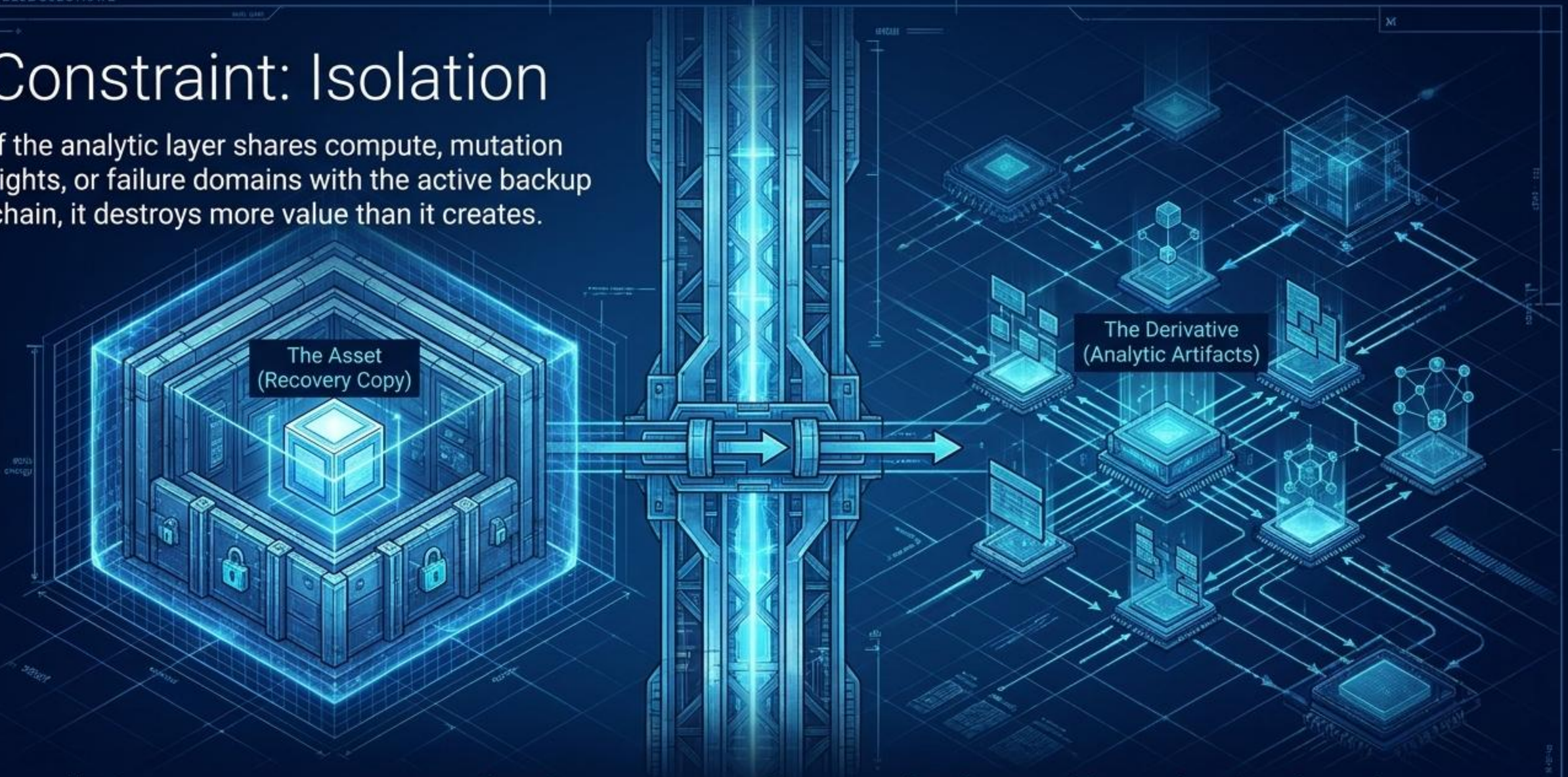
Based on 15 years of module architecture refinement:

- Runs against immutable copies.
- Decoupled from the active backup window.
- Applied selectively, dataset by dataset.



Constraint: Isolation

If the analytic layer shares compute, mutation rights, or failure domains with the active backup chain, it destroys more value than it creates.



“The recovery copy is the asset, the analytic artifact is derivative, and the two never run in the same room.”

The Burden of History

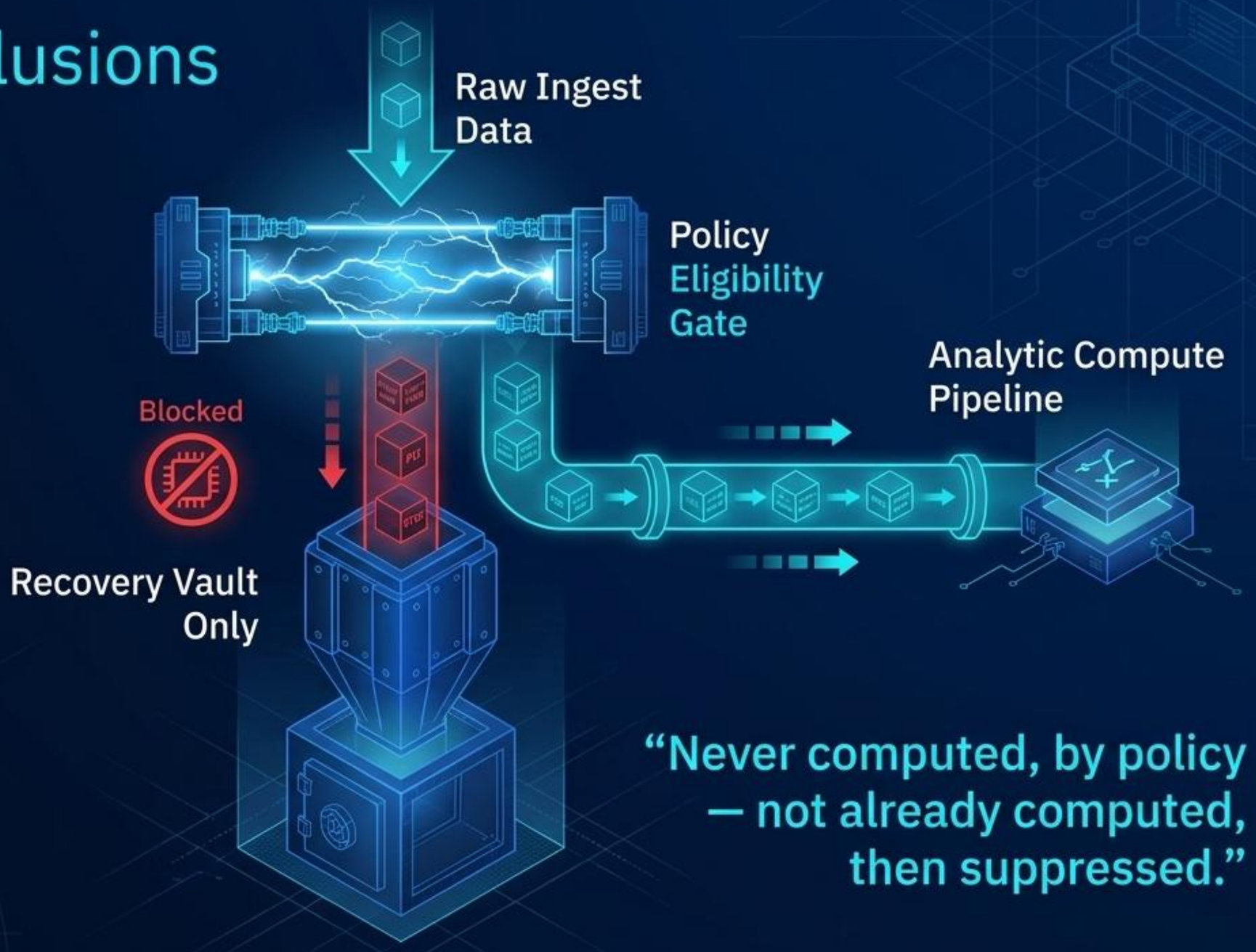
“Historical data is more sensitive than live data, not less.”

The backup tier preserves:

- Revoked entitlements
- Right-To-Be-Forgotten (RTBF) deleted records
- Declassified material in its prior classified form

Selective Exclusions

Retention eligibility and analytic eligibility are fundamentally different policies.



"Never computed, by policy — not already computed, then suppressed."

Architectural Governance

Sensitivity requires explicit enforcement beyond logical isolation.

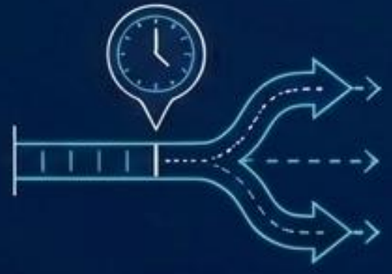
Encryption Boundaries

Derivative artifacts inherit source sensitivity and need distinct, independent keys.

Privilege Separation

Restorer cannot query the index; Analyst cannot mutate the recovery chain.

Key Management



Cryptographic posture must outlast the multi-decade retention window. Post-quantum ready.

A Research Agenda for MSST

1. Architecture & Cost Envelope: Contracts and isolation guarantees for Compute Plugins.

2. Query Primitives: Time-indexed metadata query languages acceptable to auditors.

3. False-Positive Rates: Embedding-space anomaly detection against real workloads.

6. Federation Models: Cross-institutional metadata indices without breaking trust.

5. Proof of Exclusion: Defining correct selective exclusion for regulators.

4. AI Reproducibility: Constructing envelopes purely from backup catalogs.

Building Above the Floor

3-2-1-0 is the floor, not the ceiling.
The longitudinal evidence layer already exists in your datacenter.

Thank You

Rick.stevens@baculasystems.com

"Backups remain recovery-first, regardless of any analytic processing we layer above them."