

Data Lakes Store Data.
Agentic AI Needs Data It Can Trust.



By **Kumar Vishwanathan**
Chief Product Officer

& **Siddhartha Dattagupta**
Chief Technology Officer

EXECUTIVE SUMMARY

In 2026, one in five organizations reported an AI-related security incident. Among those, 92% had no proper AI access controls in place at the time. That single statistic, from IBM's 2026 Cost of a Data Breach Report [12], is close to a direct restatement of this paper's argument: the industry's rush to deploy agentic AI has dramatically outpaced its investment in governing the data those agents actually touch.

Every enterprise and government moving into agentic AI eventually discovers the same uncomfortable fact — the data lake they built for analytics was never designed to answer the questions autonomous agents now force to the surface: who is allowed to touch this dataset, under what proven conditions, with what auditable trail, and how quickly can that access be revoked if something goes wrong. A data lake answers where data lives. It was never built to answer who may touch it, under what conditions, or how that access can be proven and revoked.

This paper makes three connected arguments:

First, that data sovereignty — the demand that data and the AI operating on it stay under a nation's or an enterprise's own control — has moved from a niche compliance concern to a default requirement over the last two years.

Second, that the data lake architecture underneath most AI initiatives was never built to satisfy the requirement of staying under the control of a nation/enterprise, and the resulting gap is a well-documented, expensive, and recurring failure pattern, not a hypothetical risk.

Third, that closing the gap between current data lake architecture and what is necessary for data sovereignty requires a distinct governance layer addresses the following challenges:

1. **Classifies** data at the point of ingestion with filters
2. **Enforces** access through a single accountable gateway
3. **Encrypts** each dataset independently rather than as one shared pool
4. **Records** an immutable record of every access

Furthermore, this layer has to be built before the AI and analytics layer on top of it, not retrofitted afterwards. AXON Datum is built around exactly that kind of structurally inescapable governance with sovereign boundaries of operations.

This paper lays out the market forces driving the sovereignty wave, the specific and well-documented ways ungoverned data platforms fail, the architecture that closes the gap, and the sequencing discipline — governance before infrastructure — that determines whether an AI program succeeds with right expected outcome or is doomed to become another abandoned initiative.

Furthermore, the enforcement of these principles at the ingestion point creates the significant cost benefit balance in operations avoiding expensive mitigations, spiraling impacts, compute overload, and expensive clean-up process at a later stage of the overall data processing pipeline. This also builds a compelling trust delegation paving the way towards true operational autonomy.

WHAT AXON DATUM IS

AXON Datum is AXON Network's structurally inescapable governance layer for AI — a single accountable gateway that every request for data must pass through, not one path among several.

AXON Datum classifies every dataset at the point it enters the platform, enforces access through a single accountable gateway, encrypts each dataset independently rather than as one shared pool, and keeps an immutable, cryptographically chained record of every access.

AXON Datum is designed to sit above and alongside an organization's existing data storage and data pipeline — not to replace it — and to be adopted before an AI or analytics layer is built on top, not retrofitted after.

1. THE SOVEREIGNTY WAVE

Two years ago, data sovereignty was a compliance checkbox for a subset of regulated industries. It is no longer that. Deloitte's 2026 Enterprise AI Report found that the share of enterprise executives naming AI sovereignty as a top priority rose from 41% in 2024 to 93% in early 2026 [1] — a shift in two years that most technology priorities take a decade to make. The same report found that 77% of companies now factor an AI vendor's country of origin into vendor selection, and 58% are building their AI stacks primarily with local vendors [1].

Capital investment is following the same curve. Gartner forecasts worldwide sovereign cloud infrastructure-as-a-service spending will reach \$80 billion in 2026, up 35.6% year-on-year [2], alongside an estimated \$100 billion in global sovereign AI compute investment in the same year [2]. Longer-range forecasts point in the same direction: the sovereign AI cloud market specifically is projected to grow from \$12.35 billion in 2025 to \$138.45 billion by 2035, a 27.34% compound annual growth rate [3]. (These figures use different scope definitions — Gartner's is IaaS spending specifically, not the full sovereign AI stack — so they should be read as directional, not additive.)

Regulation has caught up faster than most vendors planned for. The EU AI Act's Article 10 — which requires formal data governance documentation for any high-risk AI system — began enforcement on August 2, 2026, with fines of up to €35 million or 6% of global turnover, whichever is larger [4]. That is not a future deadline to plan around; it is now the operating reality for any organization running high-risk AI in or into the EU.

None of this is confined to a handful of governments experimenting at the margins. Every major economy now has a live sovereign AI initiative under way [6]: France's Mistral, the UAE's G42 and its Falcon models, Saudi Arabia's HUMAIN, Singapore's SEA-LION, India's BharatGen and its AIKosha data-governance initiative, and Japan's LLM-jp, alongside infrastructure commitments like the EU's Common European Language Data Space, Canada's \$925.6 million pledge over five years for sovereign public AI infrastructure, and the EU's €20 billion AI Gigafactory program. The pattern across nearly all of them is the same: a national fund, a compute partnership, and — in every serious version — an explicit data governance and sovereignty clause [5].

The takeaway for any organization or institution building AI capability today, sovereign or enterprise, is that governed, jurisdiction-controlled data is no longer a differentiator. It is table stakes, and the gap between organizations/institutions that have it and those still catching up is widening, not narrowing.

2. WHY DATA LAKES WEREN'T BUILT FOR THIS

The data lake, as a concept, was never designed to make governance an inherent property of the architecture. Object- and file-level access controls have existed in most data lake stacks for years, but almost always as an optional, separately configured layer that a team has to choose to buy, integrate, and maintain — not something the platform enforces by default. Classification of what a dataset actually contains, including personally identifiable or otherwise sensitive information, is rarer still as a native capability; it depends on someone tagging it correctly, and nothing in the architecture itself stops that tag from being wrong, missing, or silently dropped later. The lake was designed around a different goal entirely: cheap, flexible, schema-free storage that could absorb any kind of data without the upfront modeling cost of a data warehouse. That trade-off made sense when the primary consumer of a data lake was a human analyst who could apply judgment, notice something looked wrong, and ask a question before acting on it. It makes much less sense when the primary consumer is an autonomous agent that acts on what it finds.

The scale of the resulting problem is well-documented and larger than most organizations admit publicly. An estimated 55% of enterprise data is now “dark” — stored but never used for analysis or decisions — and organizations collectively spend an estimated \$350 billion a year storing and managing data that delivers no value in return [8]. A 2024 Informatica survey found that 68% of enterprises had at least one abandoned data lake initiative, with metadata chaos — the inability to reliably say what a given dataset is, who owns it, or whether it can be trusted — cited as the recurring cause [7].

A useful way to name this failure pattern comes from a 2026 independent analysis of data lake deployments [9] (cited here as a practitioner synthesis, not a peer-reviewed or statistically validated study): “Governance Debt” — the idea that every deferred governance decision compounds like unpaid technical debt — and “Governance Gravity” — the tendency of organizations to drift back toward ungoverned patterns the moment governance becomes inconvenient. Both terms describe something every practitioner who has run a data platform for more than a year will recognize, even without a formal study behind them.

The core problem this paper is concerned with is what happens when Governance Debt meets an autonomous agent instead of a human analyst. A person browsing an ungoverned, poorly labeled dataset might pause and ask a colleague whether it’s safe to use. An agent executing a task does not pause by default — it acts on whatever it can reach, at whatever speed it operates, unless something in the architecture itself stops it. Fifteen years of data lake practice tolerated ungoverned data because a human being was always the last line of defense. Agentic AI removes that line of defense, which is precisely why the failure pattern that has been manageable — if expensive — for over a decade is becoming acute now.

3. THE GOVERNANCE LAYER AGENTIC AI ACTUALLY NEEDS

The fix is not a better data lake. Commodity object storage — the technology underneath most data lakes — is not going away, and it does not need to. What is missing is not access control in the abstract — as the previous section noted, pieces of it already exist in most stacks. What is missing is for that control, classification, and audit to be inherent to the architecture rather than inconsistent, retroactive, and “bypassable” add-ons that only work if someone configures and maintains them correctly every time. A data lake answers where data lives; on its own, it does not reliably answer which data gets into this, who can touch it, under what proven conditions, or

with what audit trail. The fix is a distinct governance layer — AXON Datum — that sits above and alongside storage, making that “who, under what conditions, with what proof” question structurally inescapable rather than optional.

3.1 Classify At Ingestion, Not After The Fact

Every dataset is assigned a sensitivity classification the moment it enters Datum — not retroactively, and not as a separate cataloguing project layered on afterward. That label travels with the data for its entire lifecycle and cannot be removed or downgraded without a formal, logged authorization step. This is Datum’s direct answer to the metadata chaos that drives the majority of abandoned data lake initiatives: classification is not optional metadata someone might get around to adding — it is a precondition of the data entering the platform at all.

3.2 Govern Access By Policy, Enforced Automatically

Access rules in Datum are defined by the people who own the data, not by whoever happens to operate the platform. Routine access, once policy is set, requires no manual approval chain — but access across an organizational boundary a policy doesn’t already cover triggers a structured authorization workflow rather than a default allow or a default block. This is what makes the model usable in practice: most access is fast and automatic, while the genuinely sensitive cases get real human judgment instead of either blanket denial or blanket trust.

3.3 Enforce At A Single, Accountable Gateway

Every request for data — including requests from AXON’s own AI components — passes through Datum’s single enforcement point, which checks identity, classification, and policy before any data moves. There is no secondary path that bypasses this check. This matters more than it might sound like it does: a governance model with multiple ways to reach the same data is a governance model with multiple places a mistake or a compromise can slip through undetected.

3.4 Audit Immutably, Not Just Log

Every access, key operation, and authorization decision inside Datum is written to a cryptographically chained record that cannot be edited or deleted after the fact — not a conventional access log that a sufficiently privileged actor could alter, but a forensically provable trail. When something goes wrong, the question “what actually happened” has a definite, tamper-evident answer rather than a log file whose integrity is itself in question.

4. THE SECURITY CASE: WHY PER-DATASET ENCRYPTION BEATS ACCESS RULES ALONE

A reasonable objection to the architecture above is that most of it could be implemented with access-control rules on top of a conventional data lake, without a separate governance layer like Datum at all. In practice, that objection undersells how often access-control logic fails in production, and why a purely rule-based defense leaves a well-documented category of failure completely unaddressed.

Broken Object-Level Authorization — a failure mode where an application correctly verifies that a caller is authenticated but never verifies that the caller is authorized for the specific record being requested — has topped the OWASP API Security Top 10 since it was first published [10], and by recent industry tracking now accounts for roughly 40% of all API attacks [11]. It is not a rare or exotic bug class. It is the single most common way access control actually fails, precisely because it is easy to introduce and hard to catch with conventional testing: the authentication check passes, the request looks legitimate, and the bug only shows up when someone requests a record they should never have been able to see.

This is exactly the scenario where policy-only enforcement and Datum's cryptographic enforcement diverge. If an access-control rule is the only thing standing between a request and the data, a single authorization bug — the kind that accounts for 40% of API attacks industry-wide — results in an actual, immediate data exposure, because passing that one policy check is all that separates the request from the plaintext. Datum adds a second control that sits in a different part of the system: each dataset is encrypted with its own independent key, issued only at the moment a request clears classification and policy, and never reused for any other request or dataset. An authorization bug that wrongly lets a request through still does not hand over a readable record — it hands over ciphertext, because the request never legitimately went through the separate, cryptographic key-issuance step tied to that specific dataset. The two controls can fail independently because they are different kinds of control, not two names for the same check: one is a policy decision, the other is a cryptographic operation that decision merely gates. The vulnerability still exists; its consequence does not.

The cost of getting this wrong is not abstract. IBM's 2026 Cost of a Data Breach Report puts the global average cost of a data breach at \$4.99 million, with malicious insider incidents — attacks that exploit legitimate, already-authorized access — running \$4.92 million, the highest average cost of any initial attack vector in the report, precisely because legitimate-looking access is the hardest kind to detect quickly [12]. Among the AI-specific incidents in the same report, model inversion attacks averaged \$6.07 million and prompt injection attacks averaged \$5.89 million per breach [12]. And the finding this paper opened with bears repeating in this context: 92% of organizations that had an AI-related breach had no proper AI access controls in place when it happened [12].

The honest framing is not that encryption replaces access policy — the policy engine still has to authorize the key-issuance step in the first place. It is that relying on access rules alone leaves an organization exposed to its single most common documented vulnerability class with no second line of defense. Per-dataset encryption is that second line: two independent controls failing at the same time is a far lower-probability event than one control failing alone.

5. GOVERNANCE BEFORE INFRASTRUCTURE

The most consequential decision in any sovereign or enterprise AI program is not which model to use or which cloud to build on. It is the order of operations: whether governance is designed and signed off before the data platform is built or bolted-on after the platform — and often the AI layer on top of it — already exists.

The second path is far more common, and it is exactly the path that produces Governance Debt: each deferred classification decision, each dataset onboarded without a clear access policy, each shortcut taken to hit a launch date, compounds into a platform that is difficult and expensive to govern retroactively, if it can be governed retroactively at all.

Once an AI system is already consuming ungoverned data in production, adding governance means auditing everything the system has already touched – a materially harder problem than governing data before anything was built on top of it.

The alternative – and the sequencing this paper recommends without qualification – is to treat governance as the first deliverable, not a parallel workstream. Concretely: agree on a classification scheme and the authorities who own each classification tier before any data is ingested; stand up the access-policy and audit infrastructure before the first application goes live against it; and validate that the model works on a deliberately small, well-understood slice of data before scaling it across an entire organization or every government ministry at once. Every one of the sovereign AI initiatives named in Section 1 that has published real implementation detail follows some version of this order. None of the credible ones start with model selection or compute procurement and figure out governance afterward.

This is a discipline, not a technology purchase, and it is the reason a structurally inescapable governance layer like Datum must be a foundational decision rather than an add-on module considered once the rest of the stack is already in production.

6. CONCLUSION

The sovereignty wave described in Section 1 is not a passing trend – it is backed by regulation with real enforcement now underway, by capital investment moving at a pace few technology shifts have matched, and by a near-unanimous shift in executive priority in under two years. The organizations that will capture the upside of agentic AI are not the ones with the newest model or the largest compute budget. They are the ones whose data was governed before an agent was ever pointed at it.

The 92% statistic this paper opened with is the clearest available evidence that most organizations moving into agentic AI have not yet made that sequencing decision correctly. The gap between a data lake and a governed data platform is not a matter of scale or budget – it is a matter of what the platform was designed to answer in the first place. A data lake was designed to answer “Where”. Agentic AI needs a platform that can also answer “Who”, “under what conditions”, and “with what proof”.

AXON Datum is built to be that platform: classification with filters at ingestion, policy-governed access, single-gateway enforcement, per-dataset cryptographic isolation, and an immutable audit trail – designed to be the governance foundation an organization builds before its AI layer, not the retrofit it reaches for after something goes wrong.

None of this operates in isolation. In AXON's own architecture, Datum is the governance layer the rest of the autonomous-networking stack depends on, not a parallel compliance exercise running alongside it. Every reading that feeds AXON Maestro's real-time Digital Twin arrives already classified and provenance-checked by Datum, and every action AXON's Neura reasoning engine takes is gated by the same policy and audit trail Datum enforces at ingestion. Datum and Maestro are complementary – Datum governs the data, Maestro operates the network – but an autonomous system that acts on the world runs on a trust delegation model from DataOps to AIOps and is only ever as trustworthy as the governance underneath the data it acts on. That is the connection this paper has been building toward throughout: sovereignty and governance are not a workstream separate from agentic autonomy. They are the foundation it has to stand on.

7. REFERENCES

1. **Deloitte** – “2026 Enterprise AI Report” (AI sovereignty as executive priority, vendor country-of-origin selection criteria). Deloitte, 2026.
2. **Gartner, Inc.** – “Gartner Says Worldwide Sovereign Cloud IaaS Spending Will Total \$80 Billion in 2026” (35.6% year-on-year growth). Gartner press release, February 2026. Global sovereign AI compute investment figure (\$100B, 2026) drawn from related 2026 industry investment tracking.
3. **Precedence Research** – “Sovereign AI Infrastructure Market” (\$12.35B 2025 to \$138.45B by 2035, 27.34% CAGR). 2026.
4. **European Union** – AI Act, Article 10 (data governance requirements for high-risk AI systems), enforcement effective August 2, 2026.
5. **McKinsey** – “Sovereign AI ecosystems for strategic resilience and economic impact.” McKinsey & Company, 2026.
6. **National Sovereign AI Initiative Reporting:** France (Mistral), UAE (G42/Falcon), Saudi Arabia (HUMAIN), Singapore (SEA-LION), India (BharatGen, AIKosha), Japan (LLM-jp), EU Common European Language Data Space, Canada’s sovereign public AI infrastructure pledge (\$925.6M/5yr), EU AI Gigafactory program (€20B) – aggregated from 2026 industry and government reporting.
7. **Informatica** – 2024 enterprise data lake survey (abandoned data lake initiatives, metadata chaos).
8. **Dark data volume and cost estimates** (55% of enterprise data unused; \$350B annual cost of unused data storage/management) – 2025–2026 industry data aggregation.
9. **Gahi, Y.** – “What Went Wrong with Data Lakes? A 15-Year Reality Check from the Field.” arXiv:2606.08266, 2026. (Independent practitioner analysis; not peer-reviewed – cited for framing terminology, not as validated statistical research.)
10. **OWASP API Security Top 10** – Broken Object-Level Authorization (API1). OWASP Foundation.
11. **Wallarm** – API ThreatStats Report, Q2 2025 (BOLA share of API attacks, exploited known vulnerabilities).
12. **IBM** – “Cost of a Data Breach Report 2026” (global average breach cost, malicious insider cost, AI-related incident access-control findings, model inversion/prompt injection costs). IBM Security, 2026.

By **Kumar Vishwanathan**
Chief Product Officer

& **Siddhartha Dattagupta**
Chief Technology Officer