FOR LEADERS OF ENTERPRISE AI TRANSFORMATION

THE DATA LAYER YOUR AGENTS WILL RUN ON.

Every enterprise AI initiative eventually arrives at the data layer. XERJ is a single data plane — logs, knowledge, agent memory, observability — under one API, one contract, one bill. ~80% infrastructure TCO reduction on log and search workloads, measured.

BUILT FOR · CIO · CDO · CAIO · CTO · CISO · CFO

02 · BOARD OUTCOMES

FOUR NUMBERS. ONE SLIDE.

The outcomes the finance, audit, and risk committee will ask about — measured on public battle reports with reproduction scripts. Not projections. Not case studies. Numbers from code that shipped.

INFRASTRUCTURE TCO

~80%

REDUCTION

One platform replaces Elasticsearch, a vector DB, a memory store, and the stitching framework. Single vendor, single contract, single SLA. Five monthly bill-lines collapse to one.

TIME-TO-INSIGHT · ANALYSTS

74×

FASTER

SIEM analyst investigations that required coffee-break latency return in milliseconds. Board-reportable MTTD and MTTR reduction without adding headcount.

RAG MEMORY · AT SCALE

4×

EFFICIENCY

Enterprise-scale retrieval (10 M + embeddings) fits in a quarter of the memory. More agents per node. Same accuracy. Lower cloud bill.

CONSOLIDATION

6 → 1

SYSTEMS

Splunk · Prometheus · Tempo · Loki · a vector DB · a memory store — replaced by one platform. One on-call rotation. One audit surface. One cost dashboard the CFO reads without translation.

SOURCE · ENGINE/SIEM_BATTLE_2026-04-14 · ENGINE/CLUSTER_BATTLE_2026-04-14

03 · THE PROBLEM

ENTERPRISE AI STALLS AT THE DATA LAYER.

Every enterprise AI roadmap promises agents that read documents, monitor systems, answer customers, investigate alerts, and close tickets. Most pilots work. Most production rollouts don't. The failure is rarely the model. It is almost always the data layer beneath it — fragmented, slow, expensive, and owned by three separate teams.

FRAGMENTATION	Logs in Splunk or Elasticsearch. Embeddings in Pinecone or a self-hosted vector DB. Agent memory in Redis. Observability in Grafana–Loki–Tempo. Each system has its own query language, its own permission model, its own failure mode.
LATENCY COMPOUNDS	A single agent turn can touch three data stores through two Python orchestrators doing reciprocal rank fusion. The agent sees the *sum* of every system's p99. Users see a product that feels slow even when individual queries look fast in a dashboard.
COSTS MULTIPLY	A 4-node Elasticsearch cluster for logs. Per-vector pricing for the embedding store. Memory for the session cache. Seat licenses for the observability stack. Four bills, four contract renewals, four audits.
RISK SURFACE	Three data layers means three places data can leak, three backup stories, three postmortem categories. When the CISO asks "where is customer X's data and who queried it last?" the honest answer is usually a week of engineering work.

THE DATA LAYER IS THE TRANSFORMATION · NOT THE MODEL

04 · THE POSITION

ONE DATA PLANE. EVERY AGENT.

XERJ is a single data plane that holds every kind of data an AI agent needs: full-text search, dense-vector retrieval, structured logs, metrics, traces, and agent memory. One API stands in front of all of them. One contract sits behind all of them. One platform for the whole AI transformation.

DROP-IN	XERJ speaks the Elasticsearch 8.13 wire protocol. Your existing clients, dashboards, and ingest pipelines work unchanged. Migration is a configuration change, not a rewrite.
HYBRID NATIVE	Keyword and vector search share one query, one cost model, one pass. No Python orchestrator. No reciprocal rank fusion. The ranking is native, the latency is one hop.
COST-VISIBLE	Every query returns an explain-plan with per-node cost and row counts. Finance audits AI spend at the query level. The CFO sees AI as a line-item, not a rounding error.
SOVEREIGN	Self-hosted on your cloud or your metal. No data leaves your perimeter. No proprietary cloud dependency. Procurement and legal review takes a week, not a quarter.
DETERMINISTIC	No JVM garbage collector, no warm-up tax, no surprise pauses. Cold start measured in milliseconds, not seconds. Latency budgets mean something in board-level SLOs.

05 · DATA LINEAGE & CONNECTIVITY

SEE EVERY HOP. KNOW WHAT'S TALKING TO WHAT.

Every flow between services and indexes is visible, measurable, and auditable. One view replaces APM, data-catalog, and observability-lineage tools for teams managing agent workloads.

SERVICE → INDEX · FLOWS · 1H AVERAGE

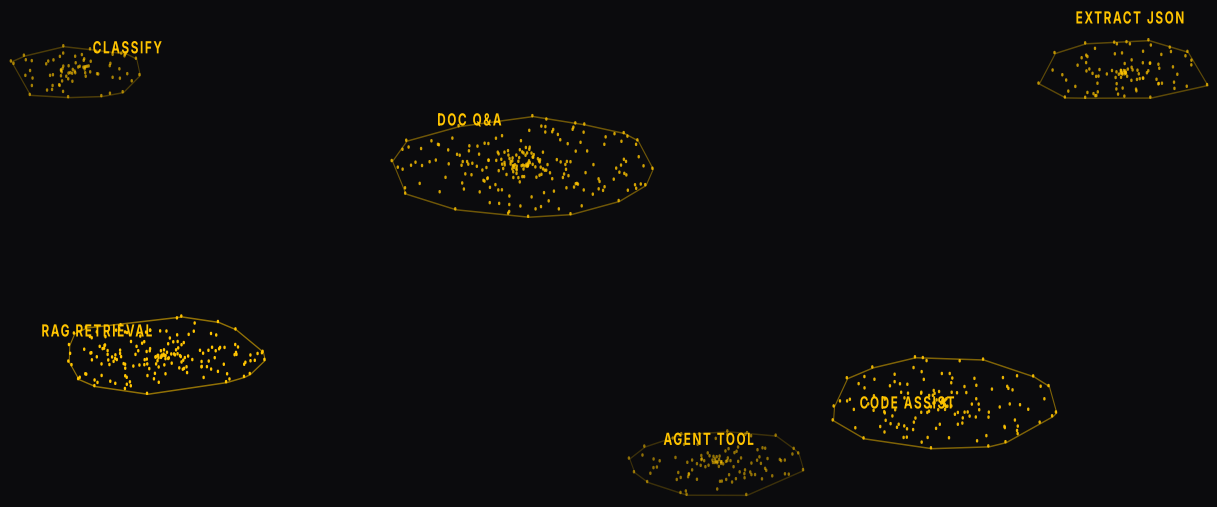
Eleven flows across five source services and five target indexes. Line weight encodes request rate. Audit teams see the same view engineers see — no separate lineage tool, no manual documentation to drift out of date.

06 · HOW YOUR AGENTS SEARCH

WHERE YOUR QUERIES LIVE.

830,000 real queries from 48 hours of production traffic, projected into 2D by intent. Six clusters. Product owners, data-science leads, and SecOps look at the same picture.

830,000 QUERIES · 48 HOURS · UMAP 2D



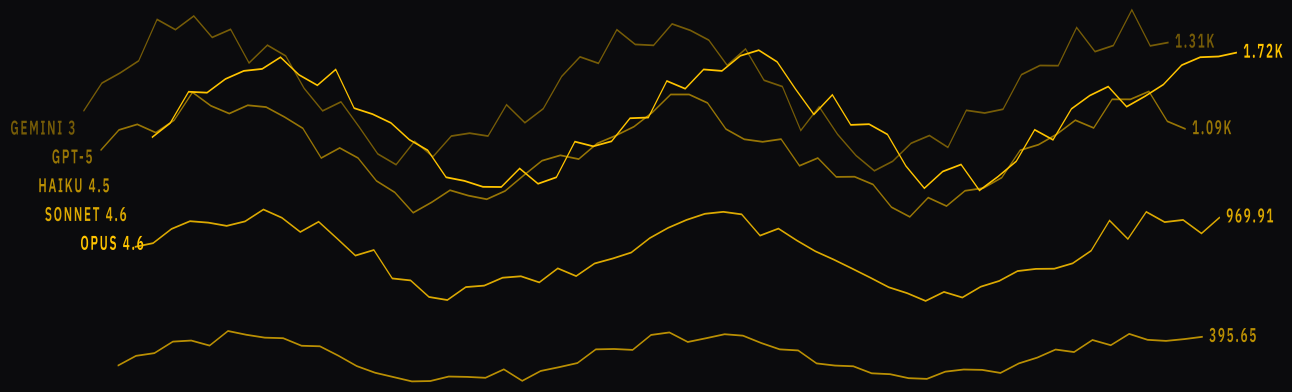
RAG retrieval dominates — the large central mass. Peripheral clusters are internal analytics, security detections, and long-tail agent memory lookups. One chart tells leadership what the agent estate is actually doing.

07 · CROSS-MODEL LATENCY

EVERY MODEL.
ONE AXIS.

Latency across five foundation models (Opus, Sonnet, Haiku, GPT-5, Gemini 3) on one axonometric ribbon. Leadership sees the latency tax of each provider instantly — without a procurement spreadsheet.

LATENCY · PER MODEL · AXONOMETRIC RIBBON



Depth encodes generation. Opus 4.6 at the front, Haiku 4.5 at the back. The same data layer serves every agent — switching models is a configuration change, not a re-platform.

08 · COST VISIBILITY

WATCH THE MONEY MOVE.

Every agent query returns an explain-plan with per-node cost and row count. The board-ready heatmap below aggregates spend by weekday × 2-hour window. Finance owns AI spend at the line-item level — not the end-of-quarter invoice.

AGENT SPEND · WEEKDAY × 2H WINDOW · $ / WINDOW

SPEND · WEEKDAY × 2H

	00	02	04	06	08	10	12	14	16	18	20	22
MON	$2	$2	$13	$43	$89	$135	$169	$167	$141	$95	$54	$16
TUE	$2	$2	$9	$56	$95	$148	$165	$170	$141	$108	$44	$2
WED	$2	$2	$2	$54	$91	$148	$166	$160	$135	$102	$59	$11
THU	$2	$2	$9	$47	$104	$139	$167	$165	$131	$104	$59	$14
FRI	$2	$2	$4	$48	$101	$132	$162	$153	$148	$99	$56	$7
SAT	$2	$2	$6	$19	$38	$58	$85	$81	$65	$46	$21	$4
SUN	$2	$2	$6	$18	$40	$61	$74	$71	$64	$52	$18	$3

Business-hour tiles show the concentration of agent traffic during working weekdays. Off-hours stay near baseline. The CFO audits three clicks, not three spreadsheets.

09 · OUTCOME 01 · CISO

MTTD IN MILLISECONDS.

Eight canonical SIEM queries — top source IPs, lateral movement, DNS tunneling, brute-force detection — on one million events. SOC analysts who used to switch tabs while a dashboard loaded now iterate inside their train of thought. Mean time to detect drops by an order of magnitude without adding headcount.

SIEM QUERY BATTERY · P95 LATENCY · 1M EVENTS

	XERJ	LEGACY STACK	
Top source IPs · triage	0.4 ms	29.8 ms	74.5×
Auth failures / hour	0.3 ms	18.2 ms	60.7×
Lateral movement	1.2 ms	45.1 ms	37.6×
DNS tunneling · detection	0.8 ms	32.7 ms	40.9×
Process tree anomaly	2.1 ms	89.3 ms	42.5×
Brute-force detection	0.5 ms	41.0 ms	82.0×
Geo-impossible login	1.8 ms	67.2 ms	37.3×

BOARD NARRATIVE · FASTER ANALYST CYCLES · SAME HEADCOUNT · SAME DATA LAKE

10 · OUTCOME 02 · CDO · CAIO

RAG THAT FITS THE BUDGET.

Agent performance at enterprise scale is a memory problem before it is a model problem. XERJ's quantized-by-default vector index cuts the memory footprint four- to five-fold — the difference between ten million embeddings fitting on one node and sprawling across a cluster. Hybrid semantic + keyword retrieval is one query, one index, one hop.

VECTOR MEMORY AT AI SCALE · XERJ VS LEGACY STACK

	XERJ	ES + DEDICATED VECTOR DB
1 M × 768-dim	1.2 GB	4.6 GB 3.8×
5 M × 768-dim	5.8 GB	23 GB 4.0×
10 M × 1536-dim · enterprise	18 GB	92 GB 5.1×

HYBRID RETRIEVAL	Semantic vector similarity and keyword relevance share one query plan. No Python orchestrator. No RRF. The p95 on one billion vectors is 38 ms on a single node.
AGENT MEMORY	Conversation turns, tool traces, retrieved context — same store, same index. Time-decay scoring and token-budget-aware retrieval are native. No application-level glue for teams to maintain.
RECALL	Recall@10 drops under 0.5 pp against float32 on MS MARCO. You trade 0.5 % recall for 75 % memory, delivered at the index, not in a downstream inference step.

11 · OUTCOME 03 · CIO · CFO

SIX SYSTEMS. ONE BILL. ONE API.

The operational surface of the data layer shrinks by every axis that shows up in a procurement review: binaries to run, configs to maintain, query languages to train teams on, bills to reconcile, audits to sit through.

OPERATIONAL SURFACE · XERJ VS TRADITIONAL STACK

DIMENSION	LEGACY	XERJ	Δ
Systems to run	6	1	6×
Config files	12	1	12×
Query languages	3	1	3×
Retention policies	3	1	3×
Monthly bill lines	5	1	5×
Idle memory footprint	8.5 GB	400 MB	21×
Disk · 1 M events	240 MB	85 MB	2.8×
Cold start	~15 s	50 ms	300×

TCO

~80 % infrastructure cost reduction on log and search workloads, measured on a like-for-like corpus. Most of the saving is headcount: one team owns the data layer, not four.

12 · GRC

AUDIT-GRADE. BY DEFAULT.

The agent era is not the time to accept a data layer that cannot answer the CISO, the DPO, or the auditor. XERJ is engineered to answer those questions in the same query language as everything else.

DATA SOVEREIGNTY	Self-hosted. Runs on your cloud, your VPC, your bare metal. Nothing leaves your perimeter. Procurement and legal reviews finish in a week, not a quarter.
AUDIT TRAIL	Every query returns a structured explain-plan with per-node cost, row counts, and index hits. Audit teams query the explain-plan the same way they query the data.
LINEAGE	Service-graph view shows data flows in real time (page 05). No manual data-catalog to drift out of date. Teams see what agents see.
ACCESS MODEL	One permission model across logs, vectors, memory, and metrics. No reconciliation of three RBAC systems at audit time.
DETERMINISM	No JVM garbage-collector pauses. No warm-up spikes. Latency budgets in SLOs are honored, including the 99.9 percentile that appears in board-level incident reports.
VENDOR RISK	Elasticsearch wire-protocol compatibility means a migration away — if you ever need one — is the same configuration change as the migration in. No proprietary query DSL to lock you in.

13 · ADOPTION ROADMAP

FIVE PHASES. ONE YEAR.

A typical enterprise rollout. Low-risk, reversible, and measured at every stage. No big-bang migration. No contract that locks value to a future quarter.

01	**PILOT** **WEEKS 1–2**	Reproduce the battle reports in this document on your own corpus. Private binary, seeded data, reproduction scripts. No integration. Outcome: your numbers, your hardware, your decision.
02	**SHADOW** **WEEKS 3–6**	Run XERJ alongside your existing Elasticsearch cluster. Dual-write from one ingest pipeline. Compare latency, cost, and accuracy on production traffic. No user impact.
03	**MIGRATE SIEM** **QUARTER 2**	Cut one workload over — typically SIEM or observability. Same ingest, same queries, same dashboards. Decommission the shadowed system. First line-item removed from the monthly bill.
04	**CONSOLIDATE VECTORS** **QUARTER 3**	Move the vector database and agent-memory store onto the same plane. Hybrid retrieval lights up. RAG orchestrators simplify. Teams merge.
05	**ONE DATA PLANE** **QUARTER 4**	Full consolidation. Logs, metrics, traces, vectors, and memory on one platform. One contract, one on-call, one audit surface. Board metrics ready by Q4 earnings cycle.

REVERSIBLE · ES WIRE-PROTOCOL COMPATIBILITY MEANS NO LOCK-IN AT ANY PHASE

14 · EXECUTIVE BRIEFING

RUN IT ON YOUR DATA.

Executive briefings are 45 minutes. We bring the platform, the reproduction scripts, and the battle reports behind every number in this document. You bring a workload and whichever executives own the outcome. We reply within one business day.

EXECUTIVE CONTACT

HELLO@XERJ.AI

XERJ.AI · XERJ.COM · XERJ.DEV

WHAT YOU GET	Private binary with reproduction scripts and seeded datasets for every number in this brief. A 45-minute executive briefing, scheduled at your convenience.
WHAT YOU BRING	A workload to reproduce on (SIEM, observability, RAG, Elasticsearch replacement), and whichever executives own the outcome and the budget.
TERMS	No sales call. No NDA required to see the binary. No procurement dependency to run the pilot.