

INDUSTRY BRIEF · FOR CIO · CDO · CISO · HEAD OF AI · RISK

AI IN FINANCIAL SERVICES. ON YOUR DATA.

Banks, insurers, and asset managers run AI under the tightest data controls in the world. XERJ is a self-hosted data plane that runs inside your perimeter — full-text, vector, and agent memory on one platform — so the AI transformation never requires sending data to a third-party retrieval vendor.

SELF-HOSTED · DATA-SOVEREIGN · AUDIT-GRADE

02 · WHY THIS MATTERS NOW

AI TRANSFORMATION MEETS THE DATA RULE BOOK.

Every board agenda in banking and insurance names AI productivity as a top-three priority. Every regulator has responded with guidance — SR 11-7, EBA on model risk, PRA SS1/23, NAIC Bulletin. The AI platform conversation is no longer about what model to use; it is about whether the retrieval layer and the audit trail beneath it meet the rule book.

DATA SOVEREIGNTY	GDPR, DORA, cross-border data rules, national privacy acts. Your retrieval layer cannot call a vendor API in a different jurisdiction. Self-hosted is not a preference — it is a constraint.
MODEL-RISK GOVERNANCE	SR 11-7 and equivalents require documented inputs, outputs, and validation evidence for every AI decision. The retrieval layer is part of the model chain and must produce its own audit artifact per query.
FRAUD & AML LATENCY	Transaction monitoring and fraud decisions increasingly rely on vector similarity to historical patterns. The retrieval has to fit inside the payment-authorization budget — milliseconds, not seconds.
LEGACY STACK DRAG	Three separate retrieval stacks for security, compliance, and analyst productivity. Each has a dedicated team, contract, and audit. AI programs inherit the sum of their complexity.

03 · UNIFIED AI DATA STACK · FINSERV-READY

ONE PLATFORM. INSIDE YOUR PERIMETER.

XERJ is a single Rust binary that hosts logs, vectors, agent memory, and hybrid retrieval — under one RBAC model, one audit trail, one BYOK story. Deploys inside your VPC, your data centre, or an air-gapped enclave. Nothing leaves the perimeter unless an embedder is deliberately pointed outside.

SOVEREIGNTY 100 % SELF-HOSTED BY DEFAULT Your cloud, your VPC, your data centre. Air-gap supported. No XERJ-side telemetry egress in AIRGAP=1 mode.	AUDIT PER-QUERY EXPLAIN-PLAN FIRST-CLASS Every retrieval returns a structured audit artifact — query, principal, cost, rows. Model-risk evidence by default.
FRAUD LATENCY	0.4 MS

TERMS AGG P95

CONSOLIDATION4 → 1RETRIEVAL STACKS

UNIFIEDCompliance logs, fraud vectors, analyst RAG, SIEM — one platform, one audit surface.

Retrieval inside the payment-authorization budget. Hybrid similarity + keyword filter in one query.

04 · FINSERV BLUEPRINT

ONE DATA PLANE. MANY WORKLOADS.

Three XERJ clusters — one per security domain — serving distinct finserv workloads. Same binary, same audit model, different data-classification boundaries.

DOMAIN A · RETAIL BANKING VPC · production cardholder environment

Fraud events · transaction logs · card-auth streams

|



XERJ cluster · 3-node HA · PCI-DSS-scoped

Workloads: fraud similarity · AML case retrieval · tx-log SIEM

Retention: hot 30 d · warm 90 d · cold 7 y (regulatory)

DOMAIN B · CORPORATE & INVESTMENT BANKING VPC · restricted

Research notes · term sheets · trade tickets · client correspondence

|



XERJ cluster · 3-node HA · analyst RAG

Workloads: research RAG · trade-ticket search · compliance surveillance

Retention: hot 90 d · warm 1 y · cold 7 y (regulatory)

DOMAIN C · RISK & MODEL GOVERNANCE VPC · model-ops

Model outputs · feature vectors · validation evidence · audit trails

|



XERJ cluster · 3-node HA · model-risk archive

Workloads: SR 11-7 evidence store · validation retrieval · explain-plan audit

Retention: model lifecycle + 7 y

SHARED BYOK via enterprise KMS · SAML federated identity (roadmap Q2)

Regional deployment · no cross-border data flow · tenant isolation

05 · WHERE XERJ APPLIES

FIVE PATTERNS. ONE PLATFORM.

The five AI-era workloads that show up in every finserv roadmap. Each maps cleanly to one XERJ deployment pattern.

WORKLOAD	PATTERN	STORE
FRAUD DETECTION	Vector similarity against historical fraud patterns + keyword filter on merchant / geography. Result inside payment-auth budget.	HYBRID
AML / SURVEILLANCE	Retrieval-augmented case review. Analyst queries unified across transaction logs, KYC notes, prior-case narratives.	HYBRID
ANALYST RAG	Research RAG over internal + external corpora. Citation fidelity matters · recall @ 10 tracked · re-embed on drift.	VECTOR
COMPLIANCE SEARCH	eDiscovery-style search across regulated correspondence · legal hold markers · tenant-isolated indices per jurisdiction.	FTS
MODEL-RISK AUDIT	Store of model inputs, outputs, and validation evidence · queryable by auditors · SR 11-7-aligned · explain-plan audit.	COLUMNAR
CUSTOMER SUPPORT	Agent-assisted servicing over transcripts + KB + policy. Agent memory scoped per session. Time-decay retrieval for recency.	AGENT MEM
SIEM	Unified SOC data plane across cloud + on-prem + identity. 74× faster p95 on SIEM aggregation batteries.	LOGS

06 · REGULATORY POSTURE

AUDIT-READY. BY DEFAULT.

Security controls an enterprise architect signs off on day one; certifications on a published roadmap. Self-hosted means most regulatory scope is inherited from customer controls — XERJ's certification track reinforces rather than replaces them.

STANDARDS & ATTESTATIONS

STANDARD	RELEVANCE	STATUS
SOC 2 TYPE I	Platform controls · access mgmt · incident response · change mgmt	ROAD Q3 2026
SOC 2 TYPE II	12-month operating effectiveness · follows Type I	ROAD Q2 2027
ISO 27001	ISMS over platform engineering + SaaS ops	ROAD Q4 2026
PCI-DSS v4.0	Segments + WAL encrypted at rest · BYOK · RBAC · audit trail	INHERITED (SELF-HOSTED)
GDPR / DORA	Self-hosted in customer region · no cross-border data transfer	GA
SR 11-7 MODEL RISK	Per-query explain-plan + audit trail = model evidence artifact	GA
FIPS 140-2	AES-256-GCM via KMS · TLS 1.3 restricted ciphers · cert roadmap	ROAD 2027
GDPR DATA-SUBJECT RIGHTS	Tenant-isolated indices · per-doc delete · tombstones · replayable audit	GA
KEY MGMT	BYOK via AWS KMS / GCP KMS / HashiCorp Vault (roadmap Q2). Rotation without downtime. No XERJ-side plaintext key access.	
NETWORK	TLS 1.3 all listeners. Modern AEAD ciphers. Certificate rotation without restart. Air-gap mode disables all telemetry egress.	

07 · FINSERV STACK · INTEGRATED

BANK STACK. UNCHANGED.

The tooling a bank or insurer already runs — ES-backed SIEM, Kafka streams, KMS-managed keys, federated identity — works out of the box.

CATEGORY	TOOLS	STATUS
MARKET DATA	Bloomberg / Refinitiv feeds → normalize → ingest via Vector · Kafka source (roadmap Q3)	GA
CARD NETWORKS	Visa / Mastercard event streams · ISO 8583 message ingestion via custom adapter	GA
IDENTITY	OIDC · SAML · enterprise AD · Okta · Azure AD · Ping · SailPoint · RBAC mapped to LDAP groups	ROAD Q2
KMS	AWS KMS · GCP KMS · Azure Key Vault · HashiCorp Vault · Thales Luna HSM · BYOK via industry-standard KMIP	ROAD Q2
COMPLIANCE	eDiscovery tools via ES API · legal-hold markers · tenant isolation · immutable audit trail	GA
SIEM	Wazuh · Elastic Security · Sigma rules · splunk-compatible exports via scheduled job	GA
RAG FRAMEWORKS	LangChain / LlamaIndex / Haystack / DSPy · ES retriever · pluggable embedder (in-house / vendor)	GA
CORE BANKING	Temenos · Finastra · FIS · Fiserv · via change-data-capture into XERJ log index	GA
ANALYTICS	Tableau · Power BI · via SQL frontend (roadmap Q4); today via ES query API	ROAD Q4

08 · MEASURED VS LEGACY

NUMBERS THE BOARD READS. OUT OF THE BOX.

Finserv retrieval workloads are where the operational delta is most measurable. Same corpus, same hardware class, same query shape.

METRIC	XERJ	LEGACY STACK
FRAUD · VECTOR LOOKUP P95	1.2 ms	45 ms (ES + Pinecone)
AML CASE RETRIEVAL P95	0.4 ms	29.8 ms (ES 8.13)
ANALYST RAG · RECALL @ 10	< 0.5 pp loss vs float32	baseline (float32)
VECTOR MEMORY · 10 M × 1.5 K	18 GB	92 GB — 5.1×
BINARY DEPLOY FOOTPRINT	13 MB	~620 MB JAR + JRE
COLD START	50 ms	~15 s JVM warm-up
PER-QUERY AUDIT ARTIFACT	yes · explain-plan	application-level plumbing
CROSS-BORDER DATA EGRESS	0 (self-hosted)	depends on vendor region
TCO IMPACT	~80 % infrastructure cost reduction on log + search + vector workloads. Most savings are headcount and vendor consolidation: one platform team, one contract renewal cycle, one SOC 2 audit scope.	
GRC IMPACT	Per-query explain-plan becomes a first-class audit artifact. Model-risk evidence production drops from a ticket-per-query exercise to a queryable index.	

09 · FIVE PHASES · ONE YEAR

LOW-RISK ROLLOUT. REGULATED.

A typical finserv rollout starts with a non-regulated pilot, then migrates into the regulated perimeter once the security posture is signed off by the CISO's office.

01	NON-PROD PILOT WEEKS 1-2	Dev VPC · sanitized corpus · reproduce SIEM_BATTLE + RAG_BENCH · CISO-office review of architecture + secure-by-default controls.
02	UAT · REGULATED- ADJACENT WEEKS 3-10	Staging VPC · real data in a regulated-adjacent classification · shadow-deploy one fraud-detection query · compare p95 + recall with production stack.
03	DOMAIN A PRODUCTION QUARTER 2	Deploy into retail banking VPC · PCI-DSS scope · fraud + AML workloads primary · decommission shadowed legacy systems.
04	DOMAIN B & C QUARTER 3	Corporate & investment banking analyst RAG · model-risk archive · SR 11-7 audit-evidence store. Federated identity + BYOK complete.
05	CONSOLIDATION QUARTER 4	Legacy retrieval contracts retired · one platform team owns the AI data plane · board-level TCO + model-risk dashboards live.

REVERSIBLE AT EVERY PHASE · ES WIRE MEANS NO LOCK-IN

10 · NEXT STEPS

PROVE IT. IN YOUR VPC.

A two-week CISO-signoff pilot · sanitized data · private binary · architecture + controls review. We can bring a mutual NDA day one; binary delivery does not require one.

FINSERV CONTACT

HELLO@XERJ.AI

XERJ.AI · XERJ.COM · XERJ.DEV

COMPANION
BRIEFS

xerj-exec-brief.pdf (board) · xerj-tech-brief.pdf (architect) · xerj-usecase-security.pdf (CISO) · xerj-usecase-rag.pdf (CDO)

CISO REVIEW
PACKAGE

Secure-by-default controls matrix · BYOK + KMS integration detail · air-gap deployment checklist · SOC 2 roadmap with dated milestones.
