INDUSTRY BRIEF · FOR CIO · CTO · HEAD OF AI · HEAD OF DATA · CDO

PRODUCT SEARCH. RAG. ONE STORE.

Retail AI lives at the seam between product catalog, user behavior, and conversational commerce. XERJ runs the catalog search, the embedding index, the session memory, and the observability data plane in one self-hosted process — compatible with the Elasticsearch your e-commerce stack already expects.

CATALOG + RAG + SESSIONS · ONE PLATFORM · DROP-IN

02 · WHY THIS MATTERS NOW

SEARCH USED TO BE BM25. NOW IT IS AI.

Conversational commerce, agent-assisted servicing, semantic catalog search — every retail AI initiative demands the same retrieval primitive: a keyword filter plus a semantic similarity plus a personalization signal, executed inside the page-render budget. Most retailers' existing Elasticsearch cluster cannot do that without a Python orchestrator.

CONVERSATIONAL SHOPPING	Conversational commerce replaces facets with natural-language queries. The retrieval layer must combine semantic intent, inventory, geography, and personalization in one hop.
CATALOG SIZE	Millions of SKUs · hundreds of thousands of new product descriptions per quarter. Embedding + re-embedding at that scale crushes general-purpose vector DBs.
AGENT MEMORY	Customer-service copilots, ambient shopping assistants, on-site agents — all need session memory that persists across channels. Redis-as-memory is a stopgap; session-scoped retrieval is native.
OBSERVABILITY + FRAUD	Retail runs high-volume log fleets and fraud-detection workloads on the same infrastructure. Consolidation of observability + commerce retrieval is a natural win.

03 · UNIFIED AI DATA STACK · RETAIL

CATALOG. AGENTS. ONE ENGINE.

XERJ hosts the product catalog's BM25 index, the semantic-search HNSW index, the session-scoped agent memory, and the observability log plane on a single binary. Drop-in for existing ES clients — your e-commerce stack continues to work while the retrieval layer underneath upgrades.

RETRIEVAL LATENCY

≤ 40 MS

HYBRID P95 PAGE-RENDER

Inside the page-render budget. Conversational shopping keeps up with the customer.

CATALOG SCALE

10 M + SKU

EMBEDDINGS PER NODE

SQ8 quantization · 4× memory over float32 · re-embedding at catalog refresh cadence.

SESSION MEMORY

NATIVE

TIME-DECAY TOKEN-AWARE

Agent memory is a first-class store · per-session indices · no Redis TTL glue code.

CONSOLIDATION

5 → 1

RETRIEVAL STACKS UNIFIED

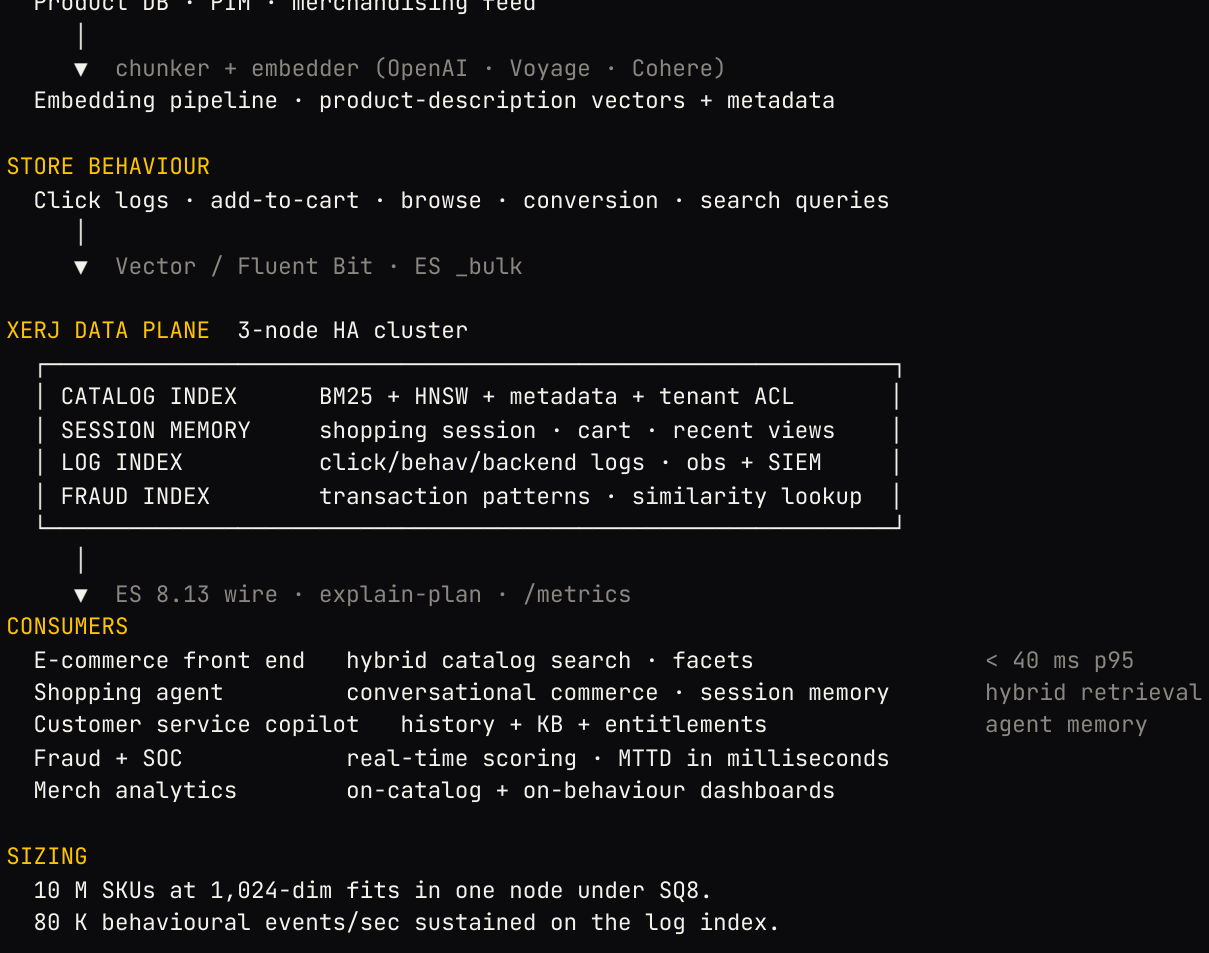
Catalog search + RAG + recommendations + sessions + observability on one platform.

DROP-IN ES 8.13 · LANGCHAIN + LLAMAINDEX · COMMERCE SDKS UNCHANGED

04 · RETAIL BLUEPRINT

CATALOG. COPILOT. OBSERVABILITY.

A three-node XERJ cluster for an e-commerce platform. Serves catalog search, shopping agents, and backend observability from the same data plane.



SIZING

10 M SKUs at 1,024-dim fits in one node under SQ8.
80 K behavioural events/sec sustained on the log index.

ONE PLATFORM · ONE QUERY API · FIVE RETAIL WORKLOADS

05 · WHERE XERJ APPLIES

SHOPPING. OPERATIONS. GROWTH.

The AI workloads that matter in retail and consumer roadmaps. Each maps to one XERJ deployment pattern.

WORKLOAD	PATTERN	STORE
CATALOG SEARCH	Hybrid BM25 + semantic search on product corpus. Facet filter · price · inventory · geography · personalization signal.	HYBRID
CONVERSATIONAL COMMERCE	Natural-language shopping agent. Session memory persists across channels. Time-decay on recent views · token-budget retrieval.	HYBRID
CUSTOMER-SERVICE COPILOT	Agent retrieves order history, policy, KB, entitlements. One unified retrieval across customer / order / product / KB.	HYBRID
RECOMMENDATIONS	Vector similarity to recent views · keyword filter for availability · personalization via metadata. Real-time, per-request.	VECTOR
FRAUD DETECTION	Vector similarity to historical fraud patterns + keyword on card / merchant / geography. Retrieval inside auth budget.	HYBRID
MERCHANDISING ANALYTICS	Behavioural events aggregated with millisecond latency. Merch teams query live. Agents generate narratives.	COLUMNAR
OBSERVABILITY + SIEM	E-commerce backend log fleet · MTTD in milliseconds · unified with application and identity logs.	LOGS

SEVEN RETAIL WORKLOADS · ONE PLATFORM · ONE CONTRACT

06 · PRIVACY POSTURE

CUSTOMER DATA. TENANT-ISOLATED.

Retail data is PII, payment data is PCI, and every jurisdiction has its own consumer-privacy rules. XERJ's controls are designed so that the same platform works under CCPA, GDPR, LGPD, PIPEDA, or national equivalents.

STANDARD	RELEVANCE	STATUS
PCI-DSS v4.0	Segments + WAL encrypted at rest · BYOK · RBAC · audit · tenant isolation for card-adjacent data	INHERITED (SELF-HOSTED)
SOC 2 TYPE I	Platform controls for SaaS offering	ROAD Q3 2026
ISO 27001	ISMS for platform engineering	ROAD Q4 2026
GDPR	Self-hosted in-region · right-to-erasure via tombstone · consent-tagged retention	GA
CCPA · CPRA	Right-to-know / right-to-delete enforced at index level · tenant-scoped indices per market	GA
LGPD · PIPEDA	Same mechanics as GDPR · self-hosted in appropriate region	GA
AGE-GATING · CHILDREN'S DATA	Per-index RBAC + tenant isolation · age-gated-content tagging · COPPA alignment	GA
AI GOVERNANCE	Per-query explain-plan = retrieval-layer audit artifact · supports EU AI Act traceability	GA

PERSONALIZATION	Customer-level personalization vectors live in per-tenant indices. Consent-scoped retrieval: if consent is revoked, the index removes the doc and audit trail records the deletion.
CARDHOLDER	Card-adjacent data (fraud patterns, tokens) lives in a PCI-DSS-scoped index with dedicated RBAC. BYOK keys scoped to the PCI perimeter.

07 · RETAIL STACK · INTEGRATED

COMMERCE. ANALYTICS. UNCHANGED.

The commerce, analytics, and customer-data tools already in a retailer's stack keep working. ES-wire-compatible; standard e-commerce integrations bolt on.

CATEGORY	TOOLS	STATUS
COMMERCE PLATFORMS	Shopify · Salesforce Commerce Cloud · Adobe Commerce · commercetools · BigCommerce — ES-compatible search backends	GA
PIM / CATALOG	Akeneo · Salsify · inriver · contentful-style product feeds · CDC ingest via Vector / Kafka source	GA
BEHAVIOURAL	Segment · mParticle · Rudderstack · custom event pipes · via ES _bulk or Kafka source (roadmap Q3)	GA
CUSTOMER DATA	Salesforce CDP · Treasure Data · Tealium · ADX · tenant-scoped customer profile index	GA
RECOMMENDATION	Custom ranker over XERJ retrieval · LTR hooks · LLM re-rank for zero-shot use cases	GA
AGENT FRAMEWORKS	LangChain · LlamaIndex · Haystack · via ES retriever · pluggable embedder (OpenAI / Voyage / Cohere)	GA
OBSERVABILITY	Datadog ES source · New Relic ES source · Grafana · Kibana — pointed at XERJ	GA
FRAUD	Signifyd · Riskified · Kount · via shared XERJ fraud-pattern index + ES query API	GA
IDENTITY	OIDC · SAML · auth0 · Okta · Auth.js · tenant-per-brand isolation	ROAD Q2

08 · MEASURED VS LEGACY

PAGE RENDER. AGENT TURN. MEASURED.

The retail benchmarks that decide whether AI commerce ships or stays in pilot. XERJ against a legacy ES + Pinecone + Redis reference stack doing the same job.

METRIC	XERJ	LEGACY STACK
CATALOG HYBRID SEARCH P95	< 40 ms	100–250 ms (orchestrator)
SHOPPING-AGENT TURN (10 HOPS)	~400 ms	~2 s (incl. LLM)
VECTOR MEMORY · 10 M × 1.5 K	18 GB	92 GB — 5.1×
RECALL @ 10 VS FLOAT32	< 0.5 pp loss	baseline
BEHAVIOURAL LOG INGEST	80 K/sec sustained	varies · typically 20–40 K
COLD START	50 ms	~15 s
CONSENT-DRIVEN DELETE	segment-level tombstone	varies · tooling gaps
RETRIEVAL VENDORS	1	3–4 (ES + vector + sessions + fraud)

CONVERSION IMPACT	Faster page renders + semantically better catalog retrieval move conversion. Shopping-agent turn times that fit inside attention budgets keep customers on-task.
OPERATIONAL IMPACT	Vector-DB contract + session-cache contract + log-store contract collapse. ~80 % infrastructure cost reduction on consolidated retrieval; one on-call rotation for the data plane.

09 · FIVE PHASES · ONE YEAR

FROM CATALOG TO THE WHOLE STACK.

A typical retail rollout. Catalog search is the lowest-risk, highest-visibility win — ideal first workload. Conversational commerce and agent memory follow once retrieval latency is validated at production volume.

01	CATALOG PILOT WEEKS 1–2	Single node · 1 M SKU embed · reproduce RAG_BENCH · compare hybrid p95 against the production ES cluster on the same corpus · no customer impact.
02	CATALOG SHADOW WEEKS 3–6	3-node HA · dual-write catalog embeddings · A/B on 1–5 % of search traffic · measure CTR + conversion deltas alongside p95 + recall.
03	CATALOG PRIMARY QUARTER 2	Primary search serving XERJ. Merch teams use the live analytics API. Decommission dedicated vector-DB contract for catalog workload.
04	AGENTS + SESSIONS QUARTER 3	Shopping copilot + customer-service copilot launched over XERJ. Session memory native · Redis retired. Fraud-pattern index migrated.
05	OBSERVABILITY QUARTER 4	Backend log fleet + SIEM consolidate onto the same platform. One retrieval platform, one audit trail, one on-call rotation for the data layer.

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

10 · NEXT STEPS

RUN HYBRID. ON YOUR CATALOG.

A two-week pilot on a real catalog corpus. We ship the binary, the embedding pipeline, and a sample shopping-agent template. You bring a catalog, your current p95, and the CTR baseline.

RETAIL CONTACT

HELLO@XERJ.AI

XERJ.AI · XERJ.COM · XERJ.DEV

WHAT YOU GET	Private binary · catalog-embedding pipeline · LangChain retriever example · shopping-agent template · merch-analytics dashboard JSON.
WHAT YOU BRING	One catalog corpus · current search p95 · a CTR baseline for one search query · optionally one fraud-pattern corpus.
COMPANION BRIEFS	`xerj-usecase-rag.pdf` (RAG) · `xerj-usecase-observability.pdf` (obs) · `xerj-tech-brief.pdf` (architect)