StreamKernel · Air-Gap Series

No cloud.
No calls home.
No compromise.

Most AI inference platforms require internet connectivity, cloud APIs, or managed model registries. StreamKernel was built without that assumption.

THE CHALLENGE

Regulated environments reject phone-home architectures

Defense, intelligence, healthcare, and critical infrastructure ops can't afford a model that dials out to validate a license or fetch weights at runtime.

Network isolation

SCIFs, classified networks, and OT environments have no egress. Period.

Data sovereignty

ITAR, FedRAMP, and HIPAA restrict where data can travel—or forbid it entirely.

Latency intolerance

Real-time sensor fusion and threat detection can't afford a round-trip to a remote API.

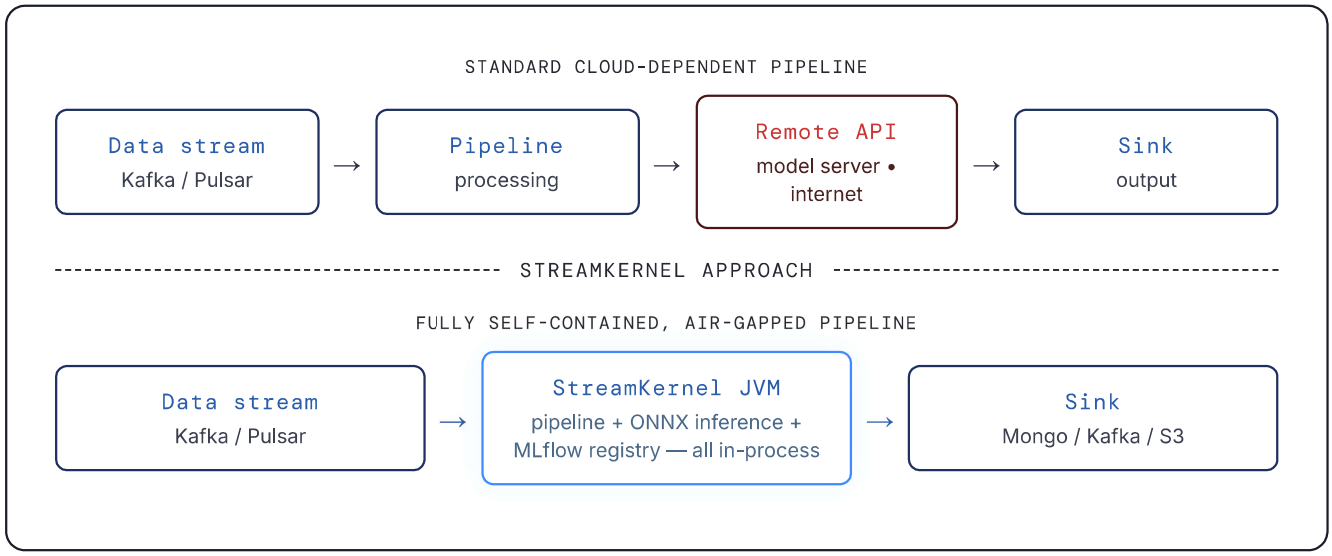
Disconnected ops

Ships, forward bases, and edge sites operate with intermittent or zero connectivity.

HOW IT WORKS

Everything runs in-process. No agents. No sidecars.

StreamKernel embeds ONNX/DJL AI inference directly inside the JVM pipeline runtime — no external model server, no gRPC call, no network hop.



BENCHMARK EVIDENCE

Real numbers from a sealed stack

Validated on a single-node JVM with zero cloud dependencies. These aren't lab conditions — this is the production-candidate configuration.

336.8 RECORDS/SEC AI EMBEDDING THROUGHPUT	**28x** THROUGHPUT IMPROVEMENT FROM BASELINE RUN	**0** RECORD LOSS AT 200K+ DOCS / RUN
163K DOCS/SEC AVG RAW MONGODB WRITE	**<30s** MODEL SWAP CYCLE PROMOTE → ROLLBACK	**1 JAR** ENTIRE RUNTIME NO EXTERNAL DEPS

AIR-GAP READINESS

Every requirement. Checked by design.

- ☑ **No outbound network calls at runtime**
  ONNX weights and MLflow model files are loaded from local disk or internal registry. Zero internet dependency post-deploy.
- ☑ **Single-JAR deployment artifact**
  One file. Copy it. Run it. No container registry pull, no cloud bootstrap, no license server ping.
- ☑ **Local MLflow model governance**
  Promote, demote, and roll back models against an air-gapped MLflow instance. Full lineage. Full audit trail. No Databricks managed cloud.
- ☑ **mTLS + OPA security profile**
  Validated at 366K ops/sec with TLSv1.3 and OPA policy enforcement active simultaneously. Hardened for classified transport.

☑ **Per-event model lineage stamping**

Every enriched record carries metadata identifying the exact model version that produced it — auditability without a cloud telemetry pipeline.

COMPETITIVE LANDSCAPE

The rest weren't built for dark environments

REQUIREMENT	TYPICAL ML PLATFORMS	STREAMKERNEL
Runtime connectivity	Required — license server, model hub, or telemetry	None — fully offline after deployment
Model governance	Cloud-managed registry (SaaS dependency)	Local MLflow — self-hosted, air-gapped
Inference location	Remote API call or sidecar process	In-process, within the JVM — no hop
Deployment artifact	Container images + orchestrator + registry	Single JAR, runs on bare JVM
Security posture	Varies — TLS optional, policy not enforced	mTLS + OPA enforced at 366K ops/sec
Audit trail	Cloud-side logs only	Per-event lineage in every output record

SK

AI inference at stream speed. Inside the perimeter.

StreamKernel is the only JVM pipeline runtime with embedded ONNX/DJL inference and local MLflow governance — built for environments where the cloud is not an option.

github.com/IntuitiveDesigns/StreamKernel-io

lopezstevie@gmail.com