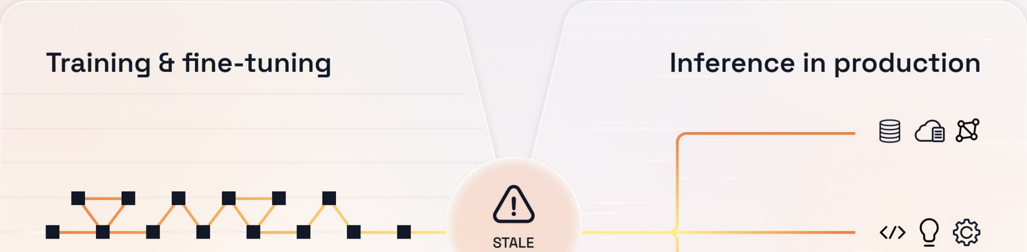
Redpanda

AI, real-time data, and why streaming matters

How batch AI architectures break in production.

AI systems break when training and inference rely on different data


Training & fine-tuning
Inference in production
STALE

Legacy batch = AI technical debt

Batch pipelines don't just slow data—they slow your ability to ship new AI products.

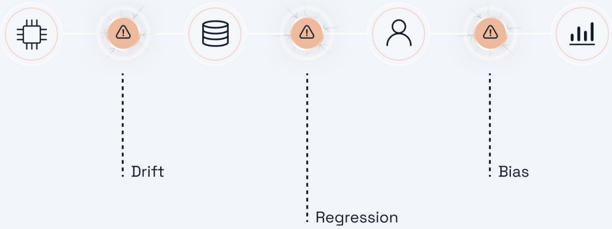
- ✓ Training relies on replayable event history
- ✓ Inference depends on fresh, real-time data
- ✓ When these fall out of sync, predictions lose accuracy

!

Most systems separate training and inference into different pipelines. That creates duplication and inconsistent data.

A single streaming log can support both. It allows replay for training while delivering live data for inference.

Why *scaling AI* is harder than training AI


Drift
Regression
Bias

- Model quality degrades over time
- Bigger models don't guarantee better outcomes
- Static retraining cycles can't keep up
- Hallucinations persist in production
- Feedback loops reinforce bias

! **Without continuous feedback, AI systems drift out of alignment**

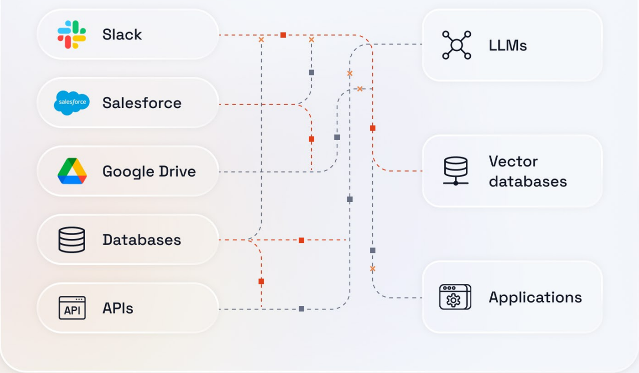
AI performance depends on access to the right context

⚠ Siloed data sources

⚠ Delayed ingestion

⚠ Incomplete context

Important context is spread across many systems



Slack

Salesforce

Google Drive

Databases

APIs

LLMs

Vector databases

Applications

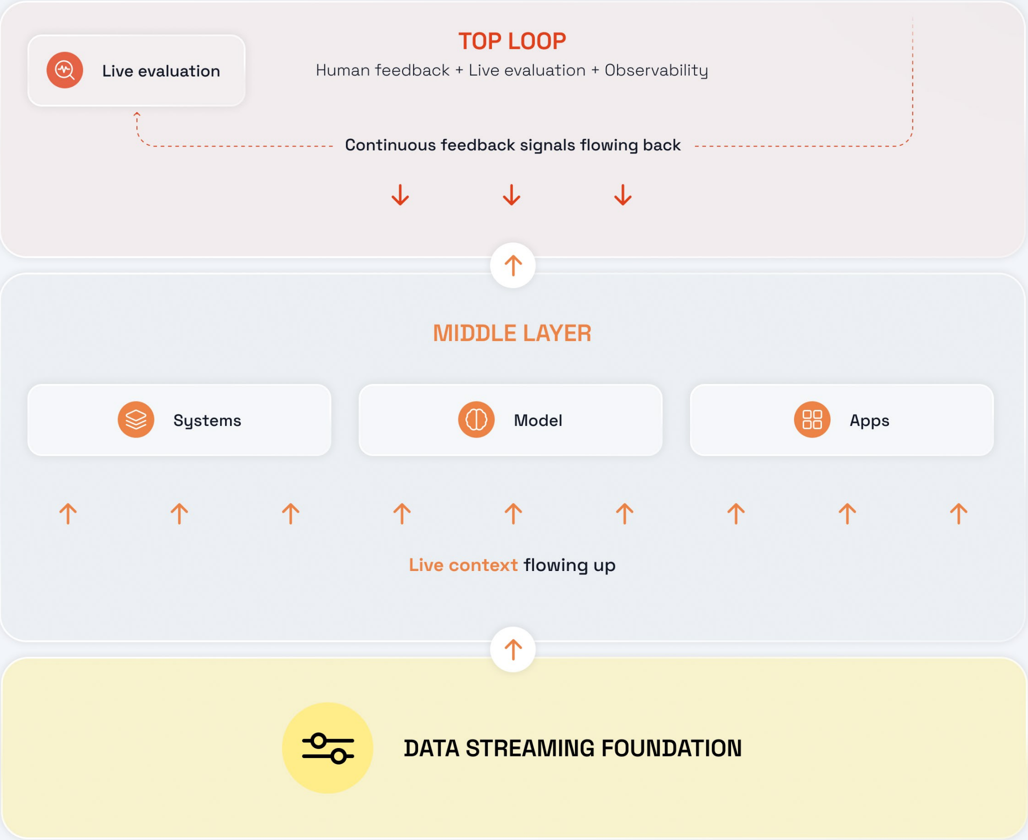
Data often arrives late or not at all

Models operate without a complete picture

If systems cannot access fresh, connected data, model outputs become unreliable.

Streaming connectors can pull data from systems like Slack, Salesforce, and Google Drive into real-time pipelines.

Production AI depends on real-time data, feedback, and visibility



TOP LOOP

Human feedback + Live evaluation + Observability

Live evaluation

Continuous feedback signals flowing back

MIDDLE LAYER

Systems

Model

Apps

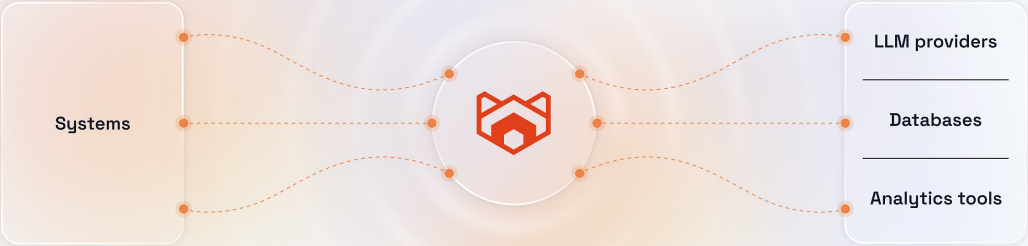
Live context flowing up

DATA STREAMING FOUNDATION

Without continuous feedback and visibility, models drift before teams can detect the issue.

The agentic data plane *connects* systems without locking you in

CONNECT STREAM EVALUATE GOVERN

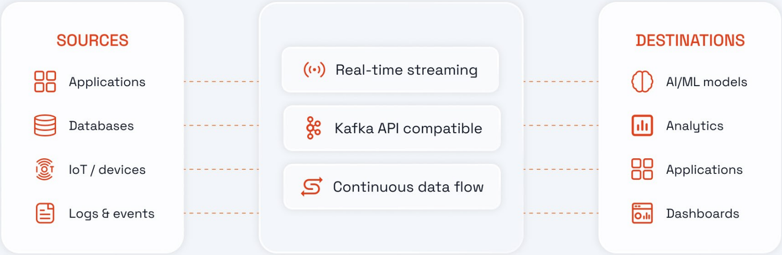

Systems
LLM providers
Databases
Analytics tools

A shared data layer connects systems, streams events continuously, and supports real-time feedback across the AI stack.

A streaming foundation gives teams flexibility. You can evolve models, tools, and vendors without rebuilding pipelines.

A streaming backbone built for real-time AI workloads

High throughput and low latency help teams get more value from modern AI infrastructure.


SOURCES
Applications
Databases
IoT / devices
Logs & events
Real-time streaming
Kafka API compatible
Continuous data flow
DESTINATIONS
AI/ML models
Analytics
Applications
Dashboards

Keep model inputs fresh with real-time data delivery

Reduce lag between events and predictions

Avoid stale context that leads to poor outputs

Support both replay for training and live streams for inference in one system

Prompt: How much delay can your model tolerate before the output stops being useful?

See how streaming supports real-time AI systems

EXPLORE REAL-TIME DATA STREAMING

LEARN HOW REDPANDA SIMPLIFIES DATA STREAMING PIPELINES

Redpanda

© 2026 Redpanda. All rights reserved.