Redpanda

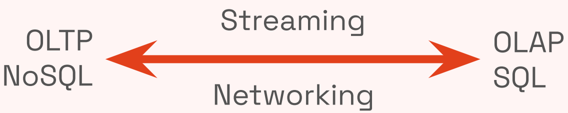
# Emerging Architectures for Real-Time Observability at Scale

**Peter Corless** *Principal Product Marketing Manager at Redpanda Data*
@PeterCorless | linkedin.com/in/petercorless

# Whoami?

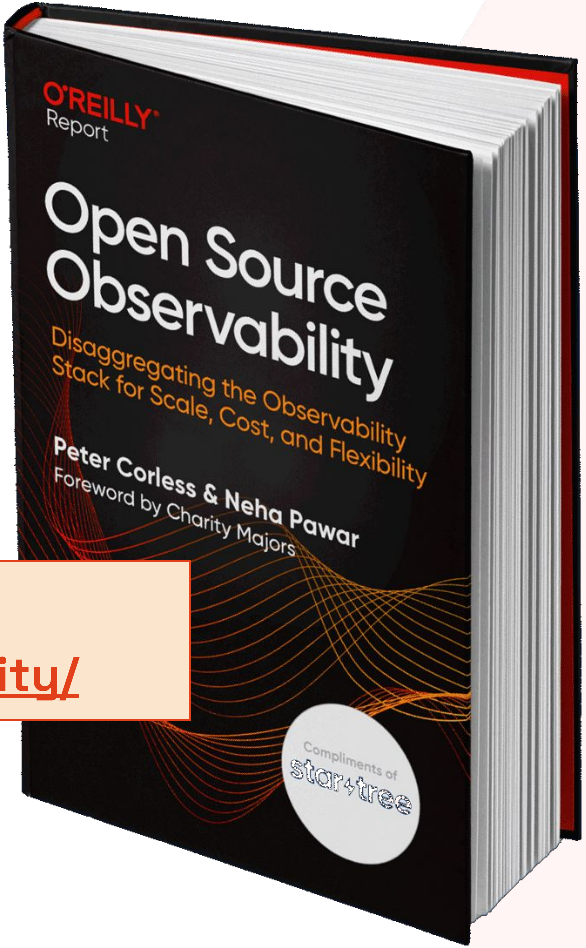
**Peter Corless**, Principal Product Marketing Manager, Redpanda Data

- Currently **Redpanda Data** — Data Streaming (Kafka-compatible)
  - Note: Not OSS, but source-available
  - Integrates with a lot of OSS tools/systems
- Formerly
  - **StarTree** (Apache Pinot) — Real-Time SQL Analytics (OLAP)
  - **ScyllaDB** — Real-Time Wide Column NoSQL (OLTP)
  - **Aerospike** — Real-Time Key-Value NoSQL (OLTP)
  - **Cisco** — Routing, Switching (back in the age of dinosaurs)
- Let's get LinkedIn: linkedin.com/in/petercorless/



Redpanda

# Free O’Reilly Book


O'REILLY
Report
Open Source Observability
Disaggregating the Observability Stack for Scale, Cost, and Flexibility
Peter Corless & Neha Pawar
Foreword by Charity Majors
Compliments of
star✦tree

“Open Source Observability”
startree.ai/solutions/observability/

# Chapter 1:
# LOCK-IN

We'll own your agents

We'll own the data collection

We'll own your data

We'll own the analysis

We'll own the dashboards

We'll own the alerting

We'll own **you**

— typical Olly vendor

# Costs? Sure! It’s simple...

$$PED = \frac{(Q_1 - Q_0) / (Q_1 + Q_0) / 2}{(P_1 - P_0) / (P_1 + P_0) / 2}$$

— typical Olly vendor

# Olly as % of IT Spending

**Depends on who you ask…**

- 15%-25% of *Infra bill* (Honeycomb, 2025)
- 10%-25% of *API ops* expenses (Gravitee, 2025)
- 7%-10% of *Cloud budget* (AWS, 2024)

*Problem: All of these are just a* ***% of a %*** *of your overall IT spending; not normalized against each other; YMMV*

Only consensus:

## ***“Observability costs are too damn high!”***

— *Shahar Azulay, 2023*

# Chapter 2:
# HIT ESCAPE

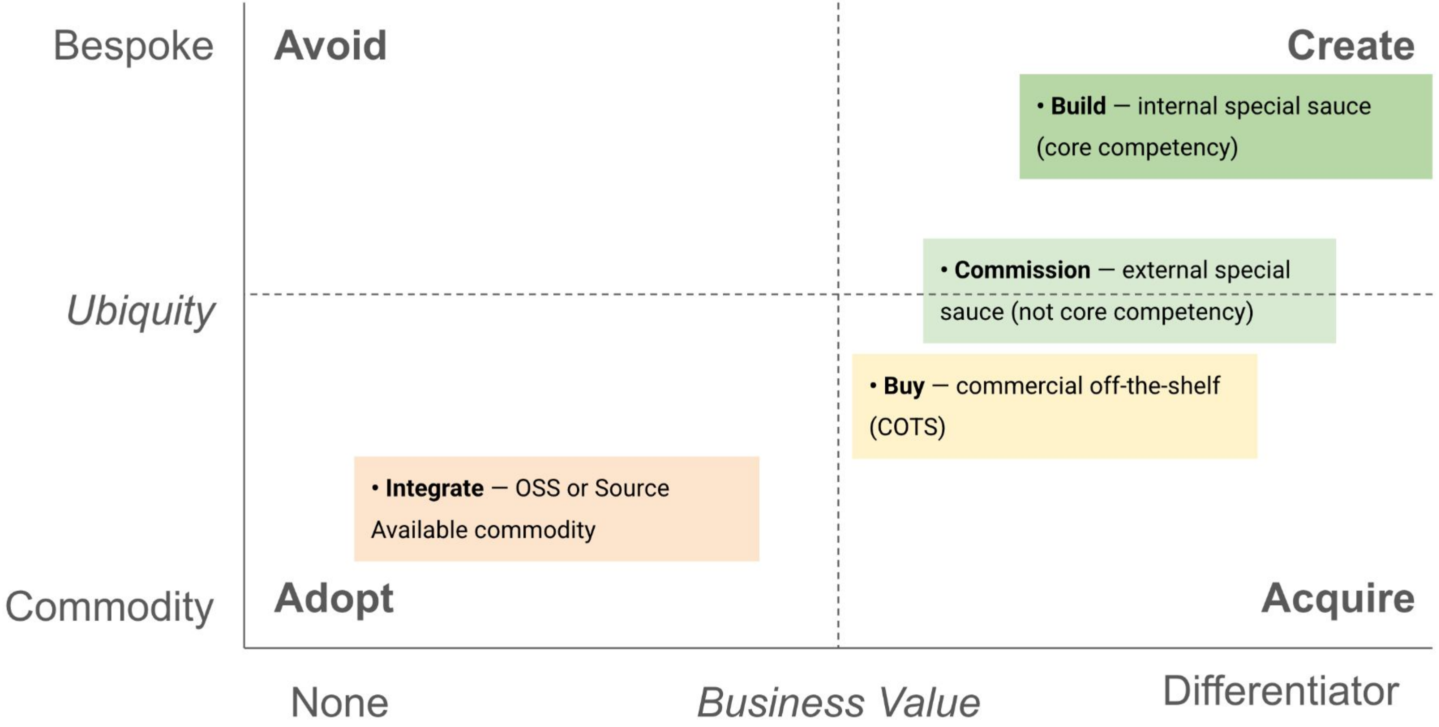
# Chapter 2: (Wait. *Can* you hit escape?)

# Typical Olly Vendor Architecture Components

- Telemetry Instrument-ation
- Collection
- Transport
- Storage
- Querying/ Analytics
- Dashboards/ Response

Redpanda

# The Choice: Build or Buy + Commission or Integrate

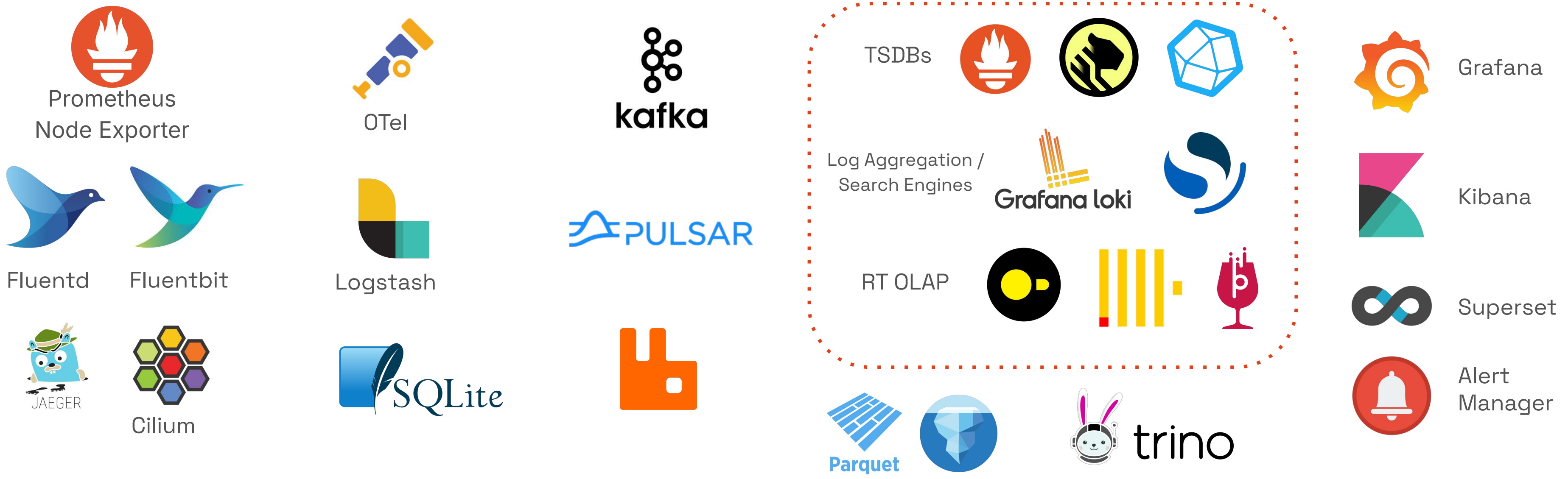


- **Build** — internal special sauce (core competency)
- **Commission** — external special sauce (not core competency)
- **Buy** — commercial off-the-shelf (COTS)
- **Integrate** — OSS or Source Available commodity

Redpanda

Original quadrant chart credit Rick Clark
Shoutout to Charity Majors, Honeycomb CTO, for inspiring conversation

# Typical OSS Observability Architecture Components



**Telemetry Instrument-ation**

Prometheus Node Exporter

Fluentd

Fluentbit

JAEGER

Cilium

**Collection**

OTel

Logstash

SQLite

**Transport**

kafka

PULSAR

**Storage** / **Querying/ Analytics**

TSDBs

Log Aggregation / Search Engines

Grafana loki

RT OLAP

Parquet

trino

**Dashboards/ Response**

Grafana

Kibana

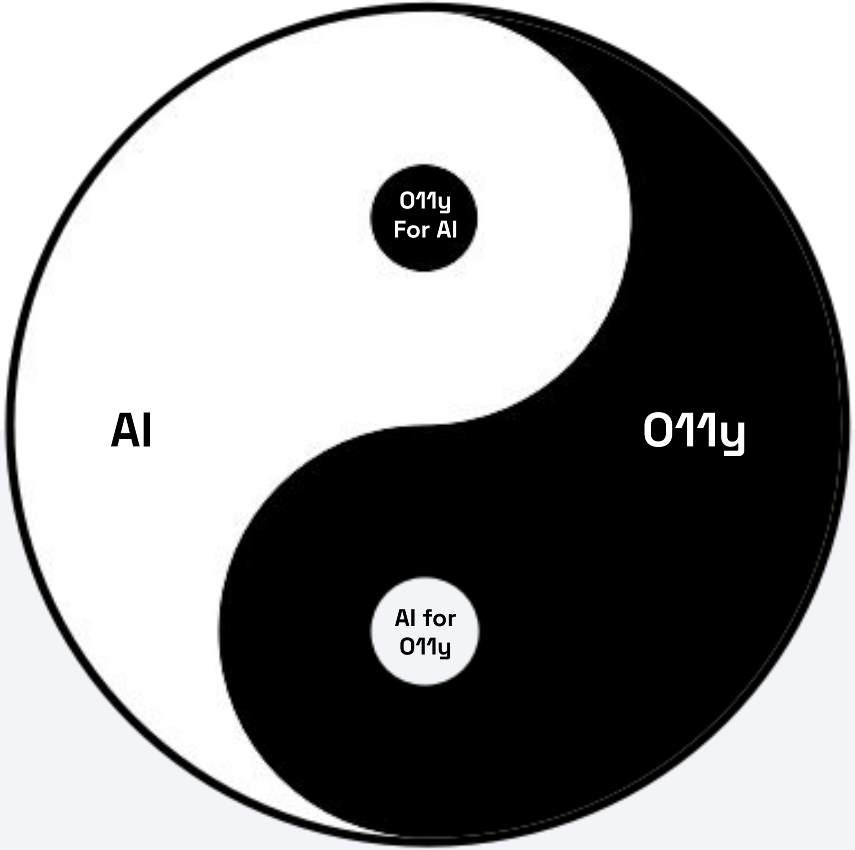
Superset

Alert Manager

Redpanda

# Chapter 3: AI is Permeating Everything, Including Olly

# Olly, AI, and Your Moment of Zen



# AI Observability & Evaluation

How to use O11y to make a *better* AI

**AI Observability**

- Model Performance Tracking
  - Accuracy
  - Precision
  - Recall
- System Resource Utilization
- Tracing
- Clustering, Visualization

**AI Evaluation**

- Data Quality Monitoring
- Fairness, Bias Detection, Ethical Compliance
- Explainability (Data Drift), Transparency, Interpretability
- Multilingual Quality (BLEU)
- Perplexity [guessability of next word in sequence], Sensibleness and Specificity Average (SSA)

# A Brief Survey of OSS AI Observability Projects

Langfuse

langfuse.com

PHOENIX

phoenix.arize.com

helicone

helicone.ai

opik by comet

comet.com/site/products/opik

laminar

laminar.sh

LangChain

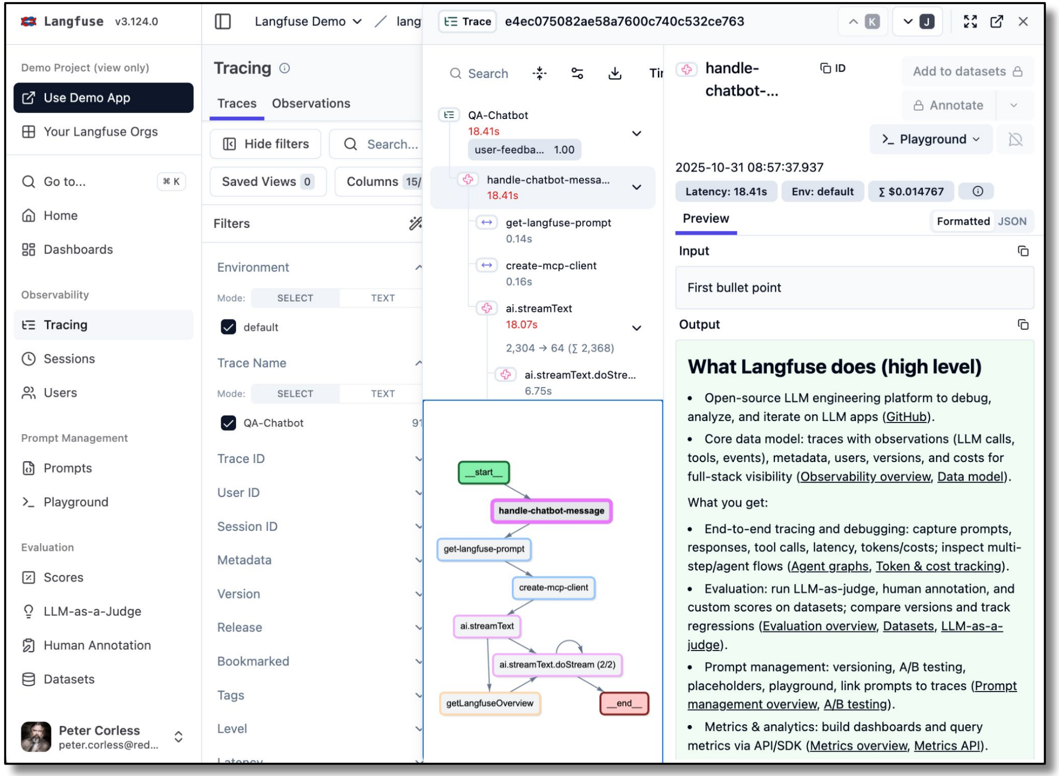
LangSmith Observability

langchain.com/langsmith/observability

Not OSS
SDK only!

# Langfuse: Tracing, Sessions, and User Analysis

LLM tracing in Langfuse lets you understand how you got the answer you got. You can see the step-by-step latency of the reply to your prompts, as well as who's using tons of tokens



First bullet point

### What Langfuse does (high level)

- Open-source LLM engineering platform to debug, analyze, and iterate on LLM apps (GitHub).
- Core data model: traces with observations (LLM calls, tools, events), metadata, users, versions, and costs for full-stack visibility (Observability overview, Data model).

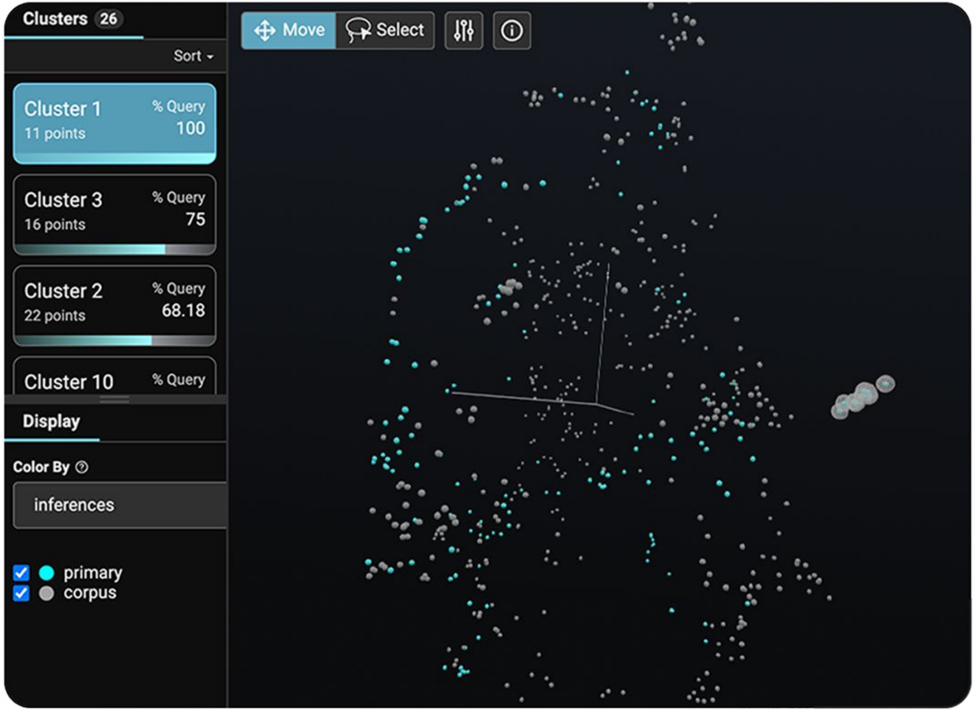
What you get:

- End-to-end tracing and debugging: capture prompts, responses, tool calls, latency, tokens/costs; inspect multi-step/agent flows (Agent graphs, Token & cost tracking).
- Evaluation: run LLM-as-judge, human annotation, and custom scores on datasets; compare versions and track regressions (Evaluation overview, Datasets, LLM-as-a-judge).
- Prompt management: versioning, A/B testing, placeholders, playground, link prompts to traces (Prompt management overview, A/B testing).
- Metrics & analytics: build dashboards and query metrics via API/SDK (Metrics overview, Metrics API).

Redpanda

# Arize Phoenix: Dataset Clustering Visualization

Uncover semantically similar questions, document chunks, and responses using embeddings to isolate poor performance.



Clusters 26
Sort
Cluster 1 11 points % Query 100
Cluster 3 16 points % Query 75
Cluster 2 22 points % Query 68.18
Cluster 10 % Query
Display
Color By
inferences
☑ ● primary
☑ ● corpus
Move
Select

Redpanda

# Chapter 4:
# Olly is Everywhere

# Data Streaming is Everywhere

Redpanda currently has customers in each of these industries

- Financial Services
- Manufacturing
- Cybersecurity
- Adtech
- Gaming & Gambling
- Energy
- Telecom
- Business Services
- Sports
- Operations
- Software Development
- Human Resources
- Insurance
- Retail
- Supply Chain
- Transportation
- Hospitality
- IT
- Blockchain
- Analytics
- Media & Entertainment
- Art & Design
- Nonprofits
- Healthcare
- Martech

# Redpanda's Core Verticals

Where data-in-motion is vital for success

- Financial Services
- Manufacturing
- Cybersecurity
- Adtech
- Gaming

# Observability is Everywhere!

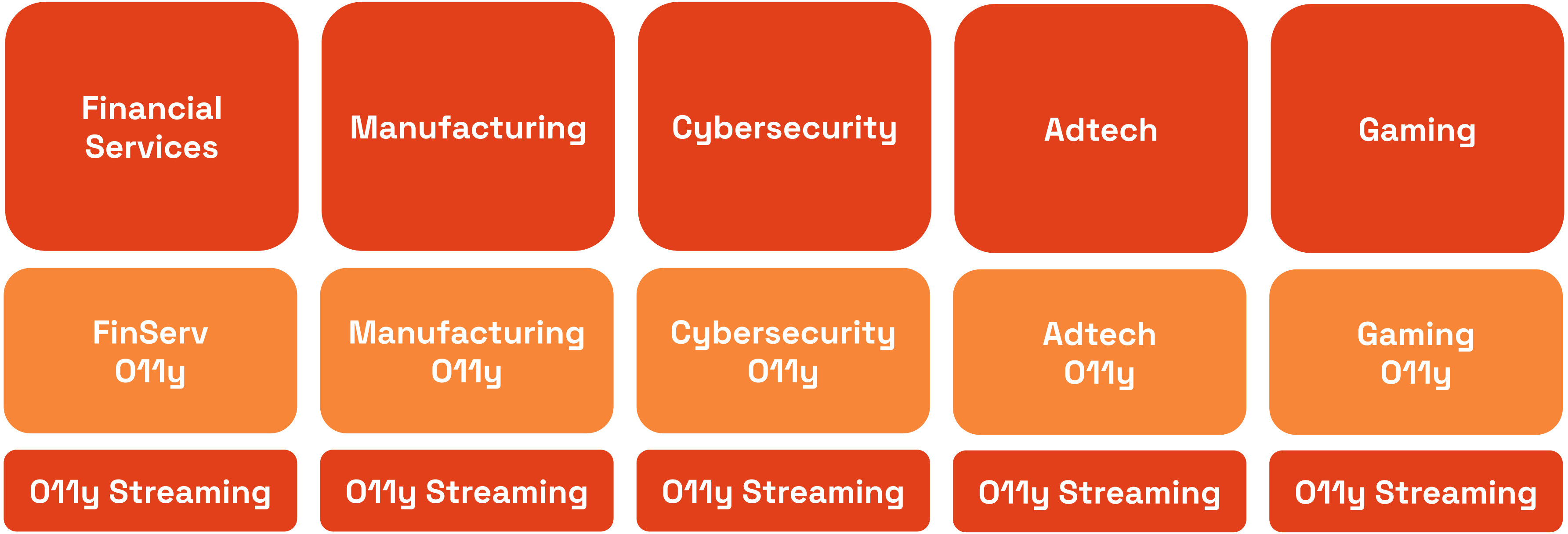
O11y is a shadow deployment under every IT use case

| | | | | |
|---|---|---|---|---|
| Financial Services | Manufacturing | Cybersecurity | Adtech | Gaming |
| FinServ O11y | Manufacturing O11y | Cybersecurity O11y | Adtech O11y | Gaming O11y |

Redpanda

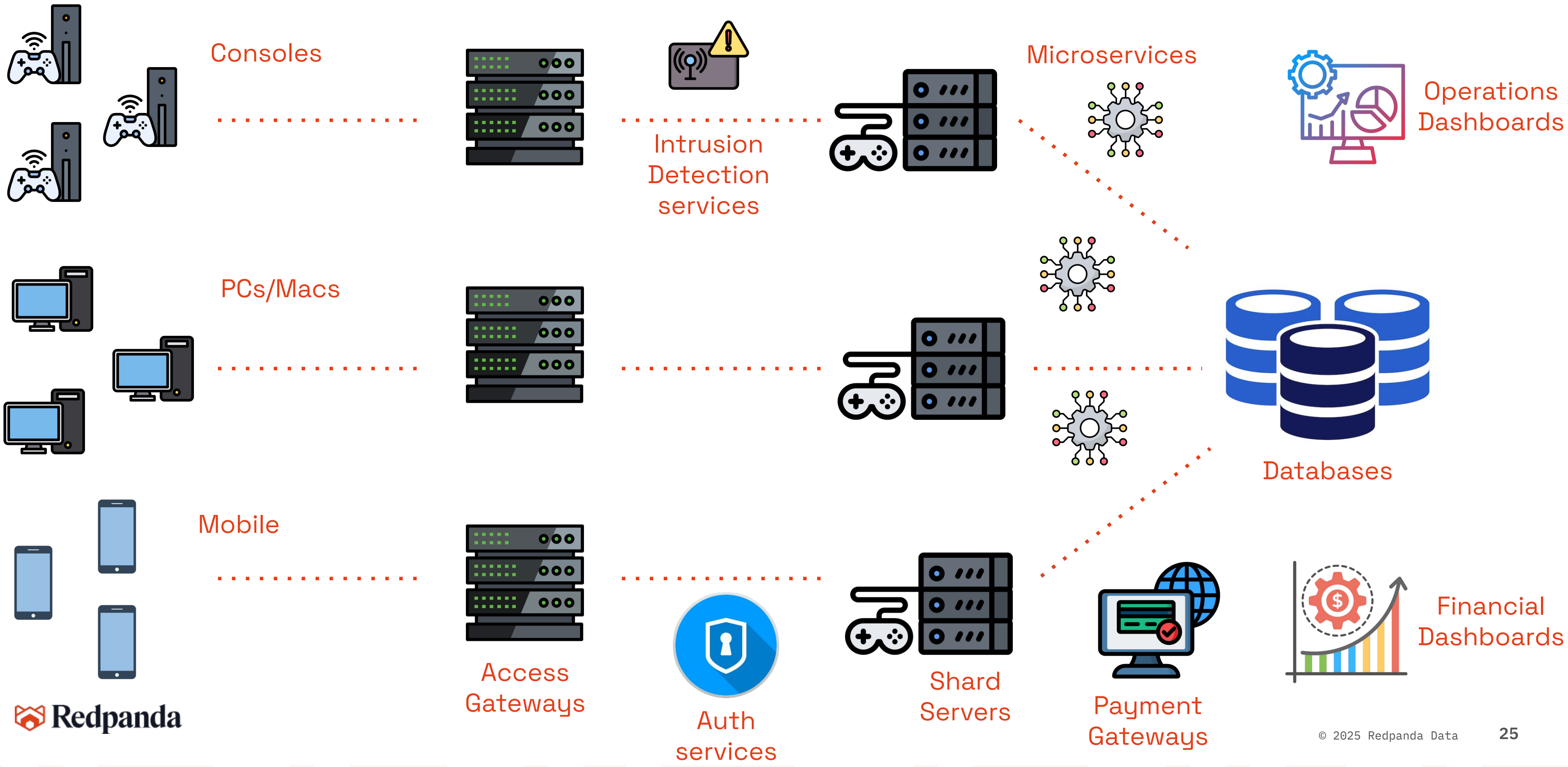
# Recursion! Data Streaming supports Olly

Streaming is a necessity for pretty much every O11y use case at scale



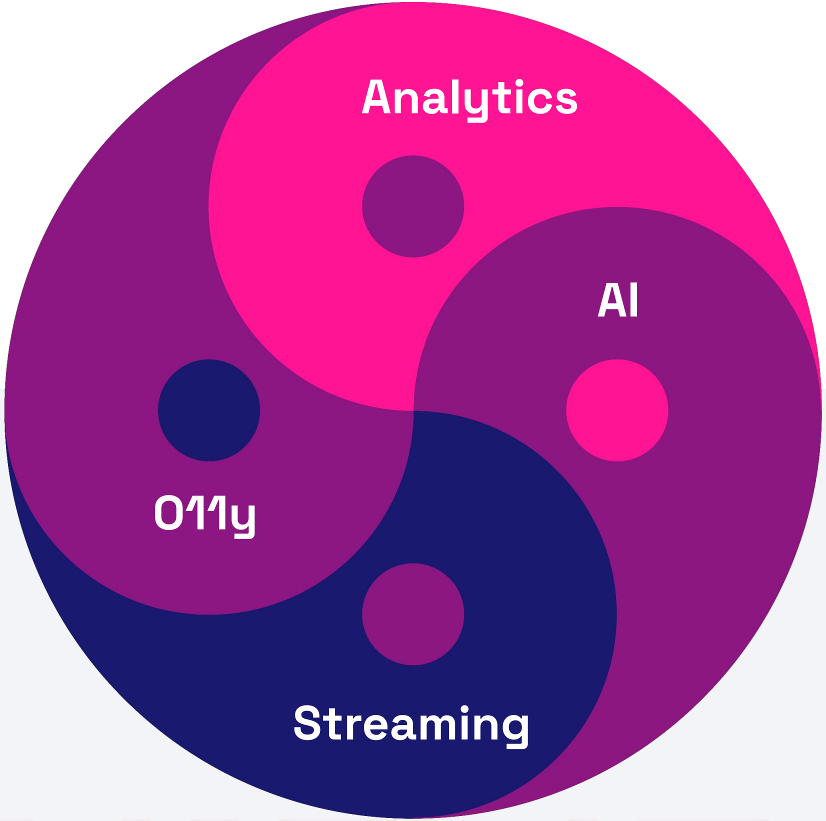
| | | | | |
|---|---|---|---|---|
| Financial Services | Manufacturing | Cybersecurity | Adtech | Gaming |
| FinServ O11y | Manufacturing O11y | Cybersecurity O11y | Adtech O11y | Gaming O11y |
| O11y Streaming | O11y Streaming | O11y Streaming | O11y Streaming | O11y Streaming |

# Example: Typical Game Company Observability



Redpanda

# AI, Olly, Streaming — and Analytics on top!



Redpanda

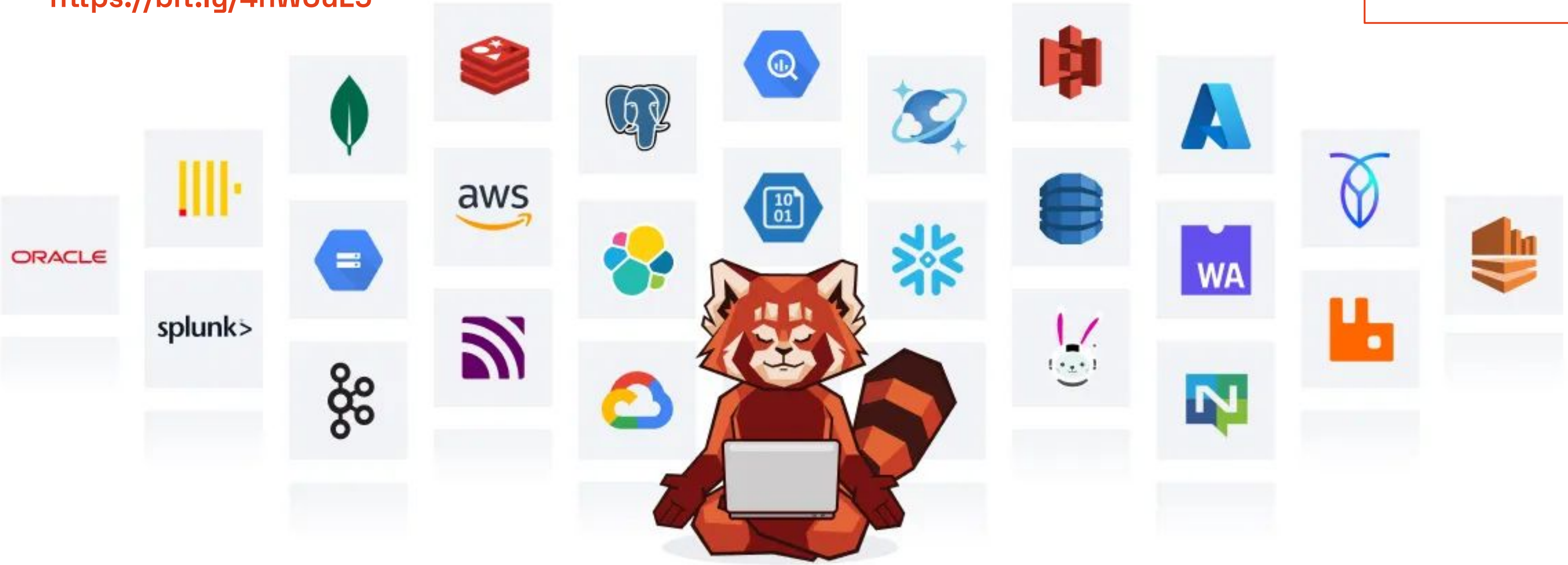
# Redpanda Connect

Hundreds of connectors to upstream and downstream systems vital for O11y

Almost all are free, OSS under Apache 2.0

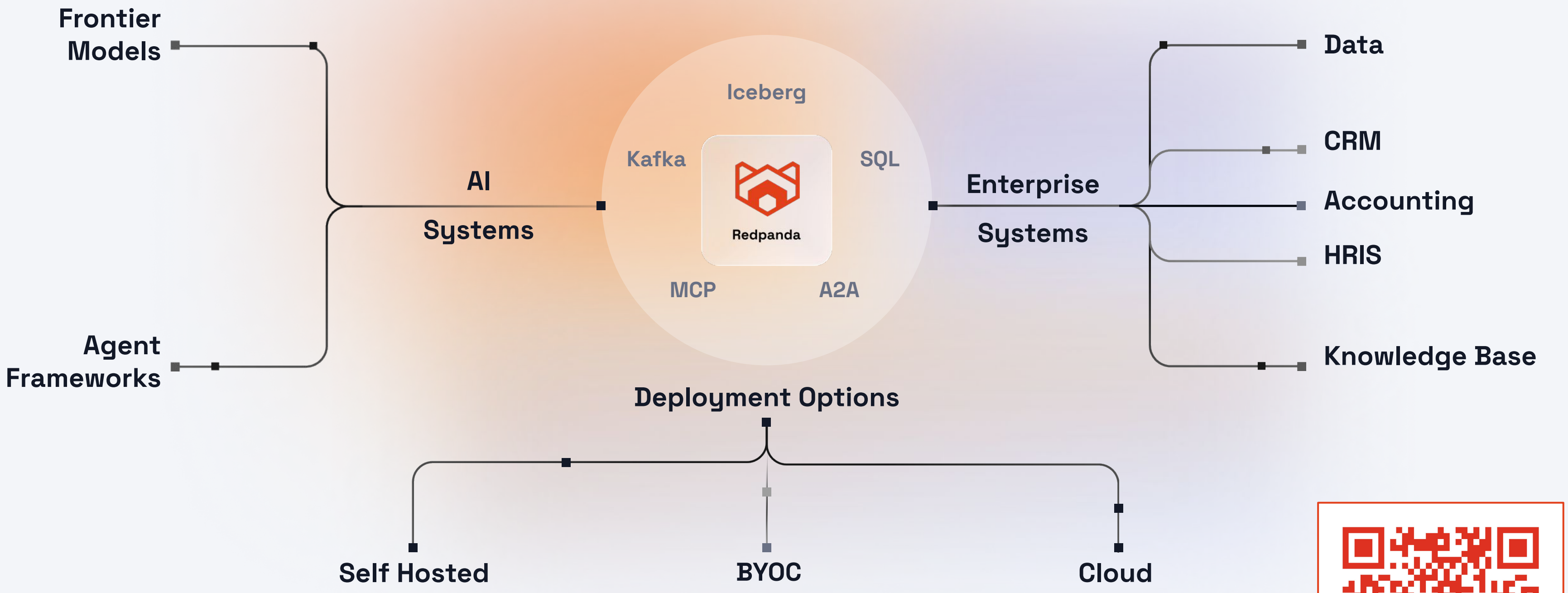
Including 16 AI connectors

https://bit.ly/4hWOuE9



Redpanda

# The Redpanda Agentic Data Plane (ADP)



Frontier Models

Agent Frameworks

AI Systems

Iceberg

Kafka

SQL

Redpanda

MCP

A2A

Enterprise Systems

Data

CRM

Accounting

HRIS

Knowledge Base

Deployment Options

Self Hosted

BYOC

Cloud

**A whole new concept; will integrate with OSS analytics, observability & AI systems**

*A unified, governed access layer that connects all your data systems and mediates every agentic interaction.*

Redpanda

# Thanks for joining!

Let's keep in touch

🐦 @redpandadata

GitHub redpanda-data

in redpanda-data

✉ hello@redpanda.com

Redpanda