Red Hat Summit

Connect

# AI · (Redis + Openshift) = 🥇

Turbocharging AI Pipelines with Redis and OpenShift AI:
Real-Time Data for Smarter Models

Redis | Red Hat

# Who we are!

Stefano Mancino

Senior Solution Architect

Luigi Fugaro

Senior Field Engineer

**Redis lets organizations deliver real-time experiences in a highly reliable and scalable manner. Redis is the world's fastest in-memory database.**

**Redis is consistently ranked as a leader in top analyst reports on NoSQL, in-memory databases, operational databases, and database-as-a-service, and trusted by over 10,000 enterprise customers.**

**redis.io**

**Founded:** 2011

**Headquarters:** Mountain View, CA

**Offices:** London Tel-Aviv, Bengaluru, Austin TX

**Employees:** 900+

**Active Regions:** NA, EMEA, APJ

**Red Hat ISV Ready Partner since 2019**

## Redis + Red Hat Joint Value Proposition

Redis is fully integrated with Red Hat OpenShift to deliver blazing fast performance with failsafe high availability, scalability, and built-in persistence reinforced with security controls, backups, and auto-recovery. Redis can be used as an in-memory database to speed up microservices applications, power real-time search and query, or enable new AI applications by using Redis for vector search or semantic caching.

Together, developers are empowered with a modern cloud-native platform to efficiently build, deploy and manage highly scalable and reliable applications with tremendous agility and lower cost.

## Joint Solutions

Redis with Red Hat OpenShift Container Platform.

## Product Certifications

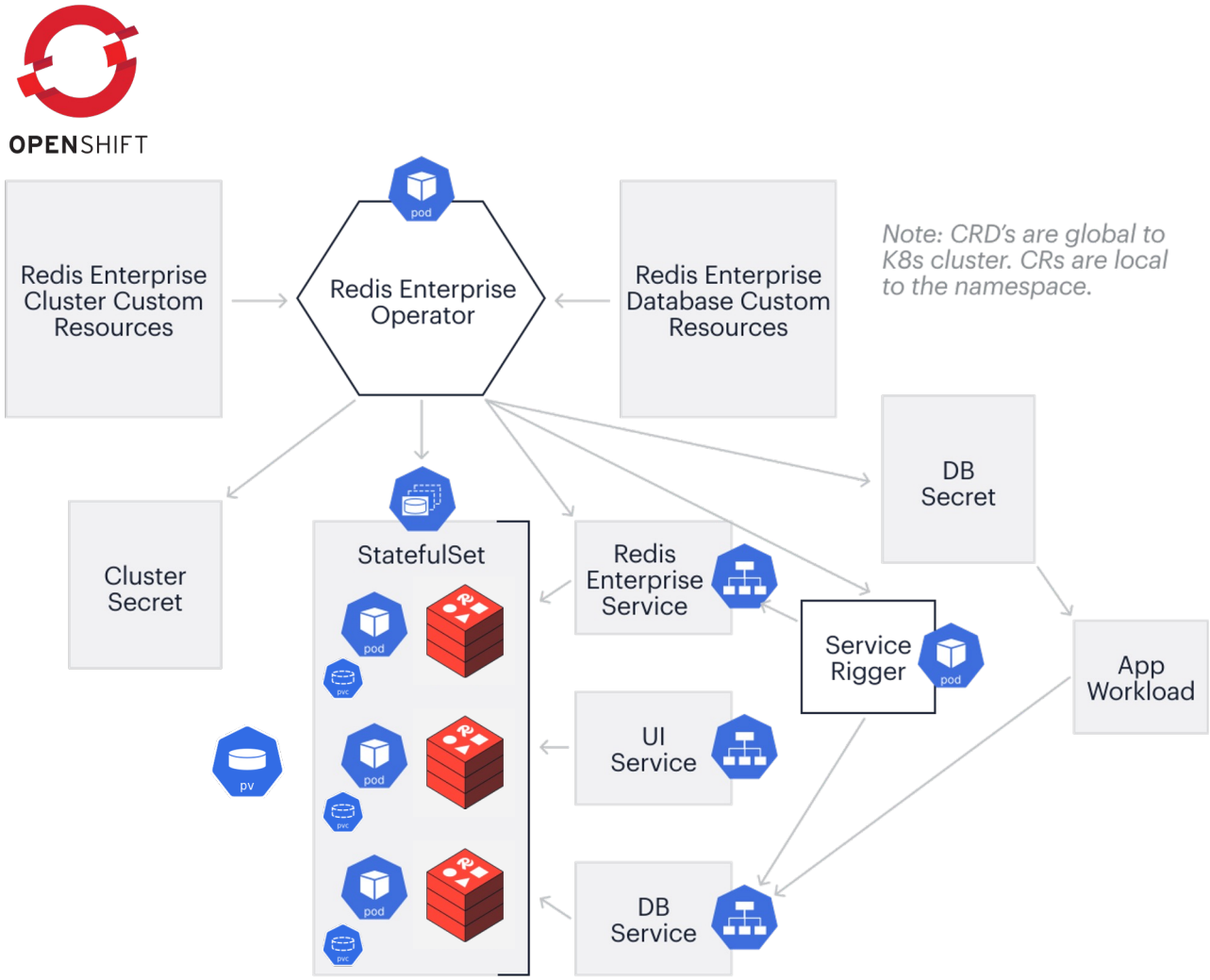
OPENSHIFT

## Highlights

- Automation - Redis Enterprise Operator for day-2 operations
- DR & BC - Active-Active across data centers for geo-redundancy

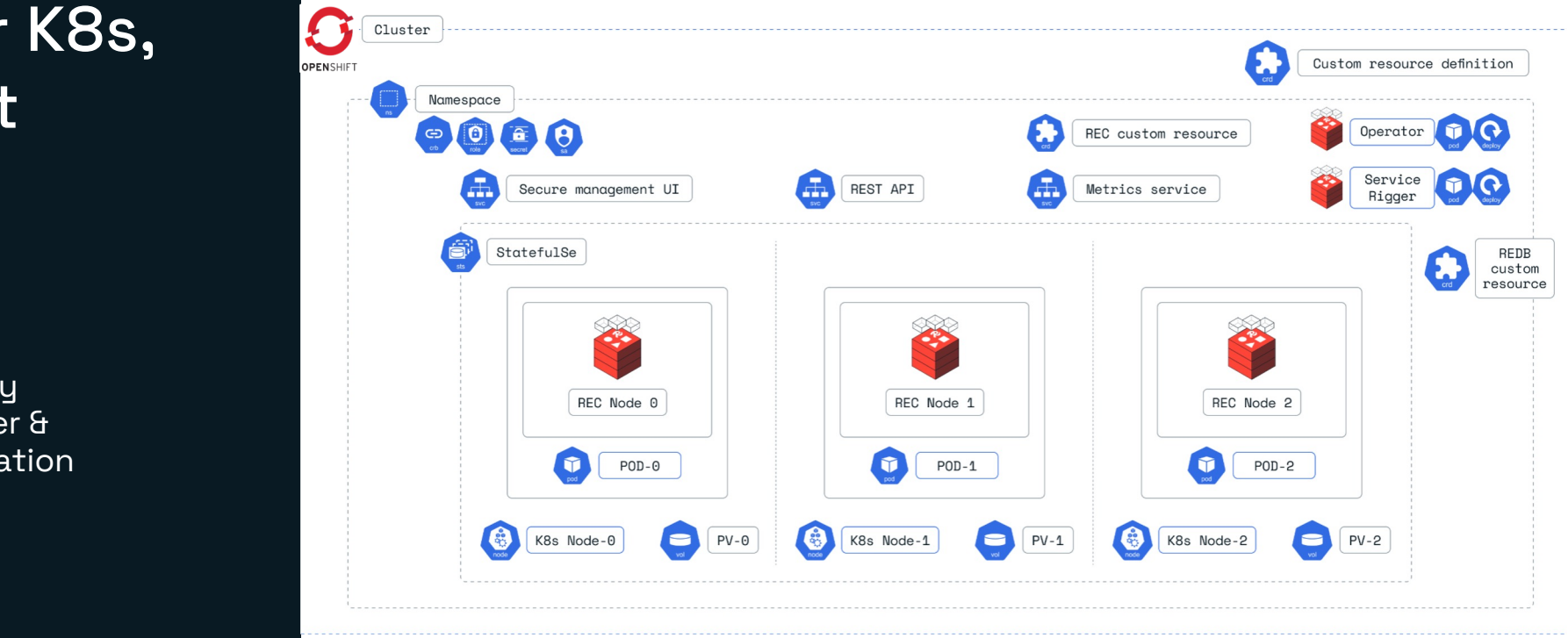
# Container native for K8s, OpenShift

Multiple tenancy model for cluster & namespace isolation



# Container native for K8s, OpenShift

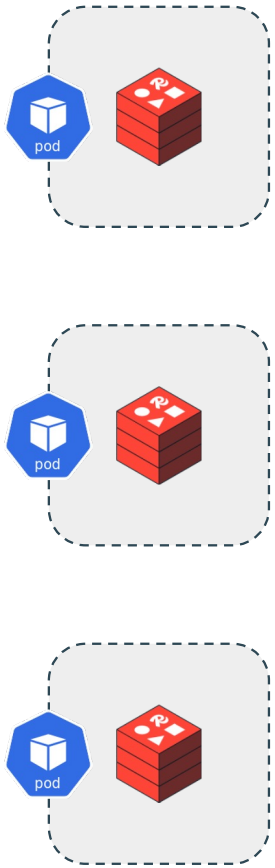
Multiple tenancy model for cluster & namespace isolation



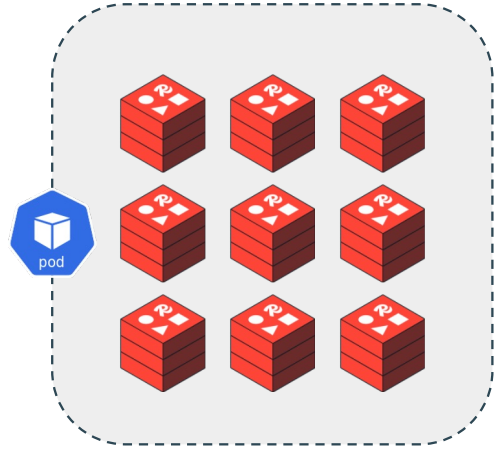
# Redis Enterprise on OpenShift A new approach

Multiple Redis Enterprise database instances on a single pod for better usage of hardware resources, keeping the same level of isolation.

## Traditional




## Redis Enterprise



# Redis Operator

*Automating Redis on OpenShift*

**Certified Operator**

Consistent packaging, deployment and life cycle management across Openshift footprints

Redis provides product support.

When Red Hat publishes a security advisory, Red Hat scans partner container images for important vulnerabilities

Runs on OpenShift

Certified operators

Fully containerized

Vendor supported

Vulnerability scans

Self-service access to application workloads, managed service-like experience.

Extends and orchestrates Kubernetes. Streamline and automate installation, updates, back-ups, and maintenance of container-based services.

## Operator capability level

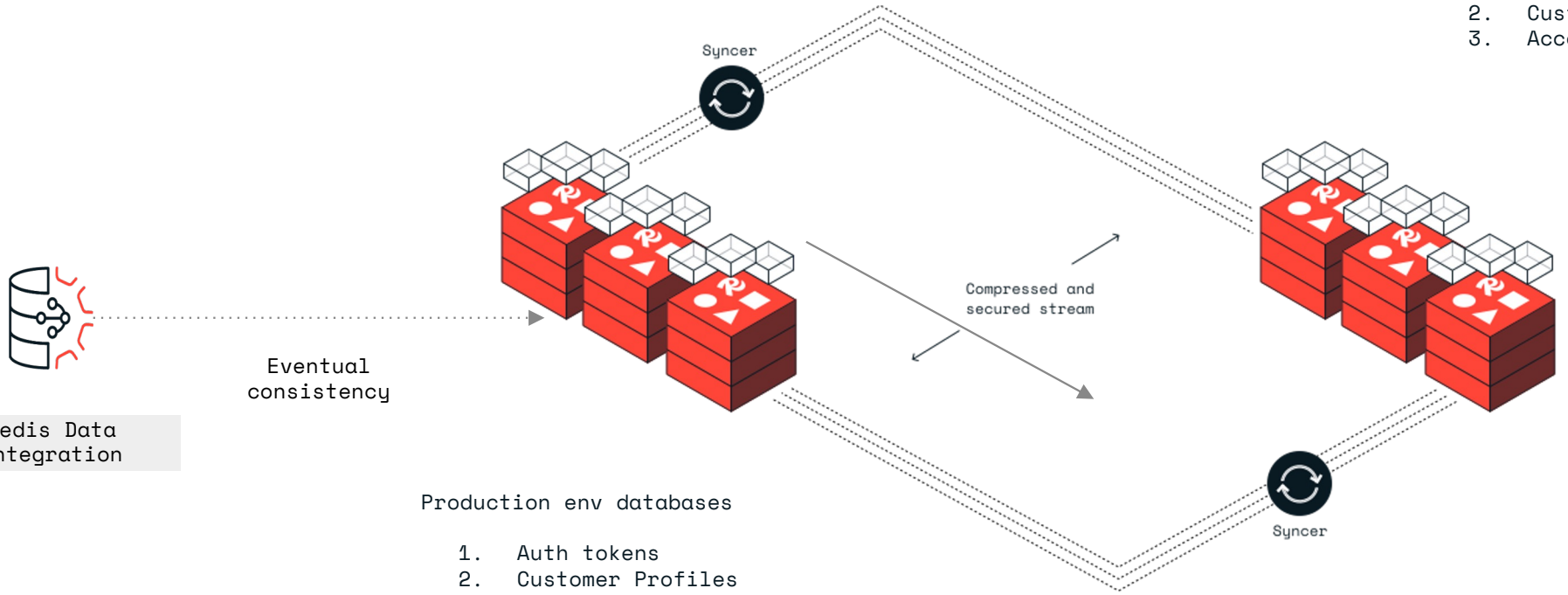
| Level I | Level II | Level III | Level IV | Level V |
|---|---|---|---|---|
| **Basic Install** Automated application provisioning and configuration management | **Seamless Upgrades** Patch and minor version upgrades supported | **Full Lifecycle** Application lifecycle, storage lifecycle (backup, failure, recovery) | **Deep Insights** Metrics, alerts, log processing and workload analysis | **Auto Pilot** Horizontal/vertical scaling, auto-config tuning, abnormal detection, schedule tuning |

# Production & Disaster recovery environments

## Active-Active GEO replication ensures 99.999% uptime

DR env databases

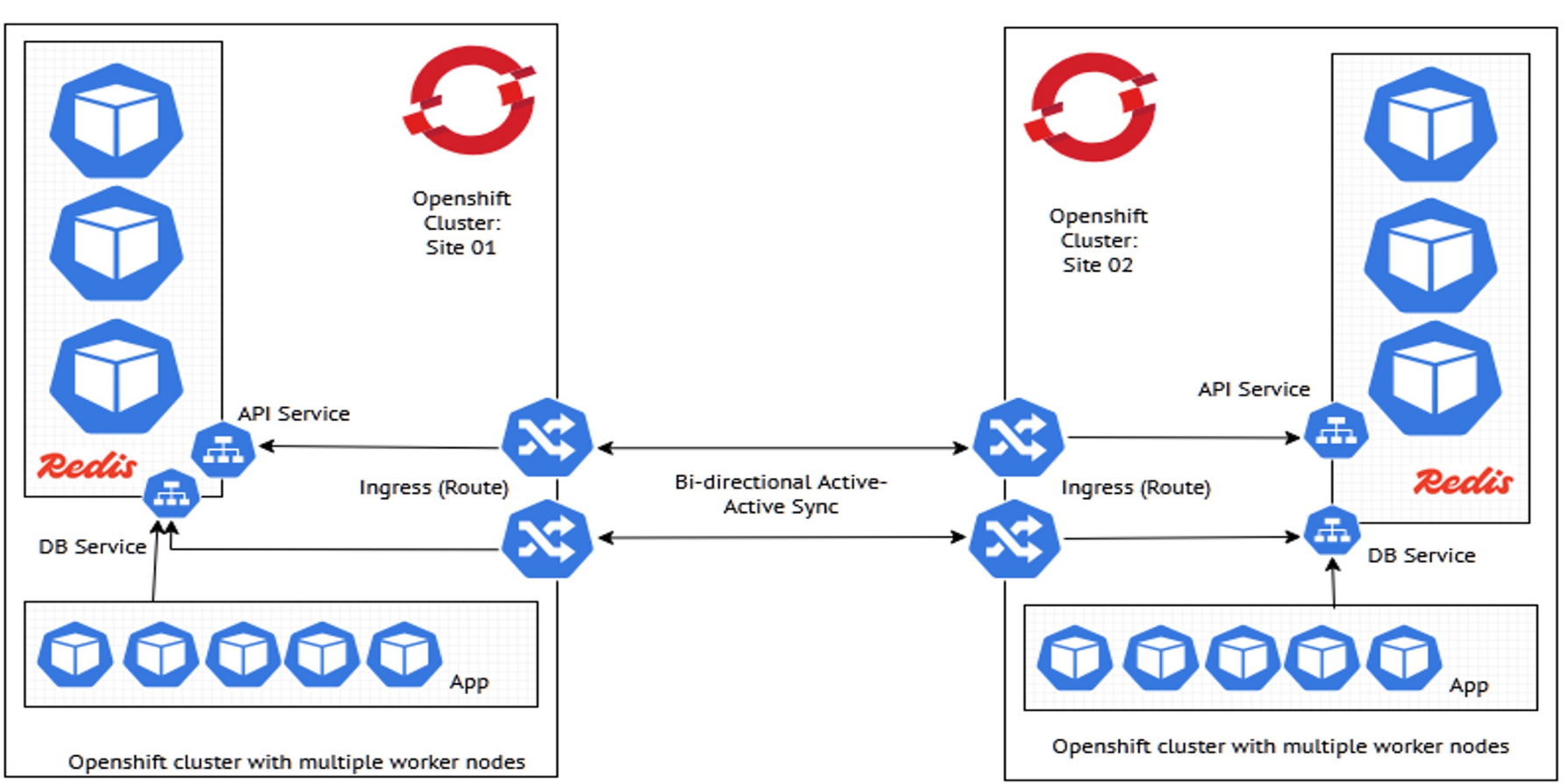
1. Auth tokens
2. Customer Profiles
3. Account Dash




Production env databases

1. Auth tokens
2. Customer Profiles
3. Account Dash

# Redis Active-Active geo distributed Architecture on Openshift




*An active-active Database architecture running on 2 independent openshift clusters while bi-directionally syncing data (while applications connect locally to individual instances within the cluster)*

# Red Hat OpenShift AI

## Integrated AI platform

Create and deliver gen AI and predictive models at scale across hybrid cloud environments.

Available as

- Fully managed cloud service
- Traditional software product on-site or in the cloud!

### Model development

Bring your own models or customize Granite models to your use case with your data. Supports integration of multiple AI/ML libraries, frameworks, and runtimes.

### Model serving and monitoring

Deploy models across any OpenShift footprint and centrally monitor their performance.

### Lifecycle management

Expand DevOps practices to MLOps to manage the entire AI/ML lifecycle.

### Resource optimization and management

Scale to meet workload demands of gen AI and predictive models. Share resources, projects, and models across environments.



Red Hat

# Detailed look integrating our partner ecosystem

Gather and prepare data | Develop models | Deploy models in an application | Model monitoring and management

**ISV software and services**

Customer managed applications

elastic, Starburst, Pachyderm, mongoDB, crunchy data, Intel AI Analytics Toolkit, CLOUDERA, run:ai, H2O.ai, Redis, SAS, OpenVINO, dynatrace, avanseus, EDB POSTGRES, VEEAM, PURESTORAGE, CognitiveScale THE TRUSTED AI COMPANY, cnvrg.io

**3rd party models and repositories**

stability.ai, NVIDIA NIM, HUGGING FACE

**Red Hat software and cloud services**

Red Hat Application Foundations

Red Hat Developer Hub

Red Hat OpenShift AI

Granite models

Red Hat Enterprise Linux AI

InstructLab

**Red Hat cloud platform**

Application platform: Red Hat Enterprise Linux; Red Hat OpenShift

Accelerators: NVIDIA, intel, AMD

Infrastructure: Physical, Virtual, Edge, aws, Google Cloud, Azure, IBM Cloud

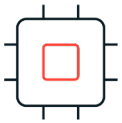


Red Hat

# Supercharge your AI with OpenShift AI and Redis

## RAG

Search for relevant text sources from knowledge bases and provide them as context for the LLM.



## Semantic caching

Search for semantically similar prompts LLM (entries).

## LLM Memory Session

Improve the quality of prompts and the personalization of LLM calls.

## Routing

Fast decision analysis using vector search to route queries based on semantic similarity

## Rate Limiting

Enforce usage limits

## Feature store

Store ML features for fast data retrieval

## Redis Data Integration

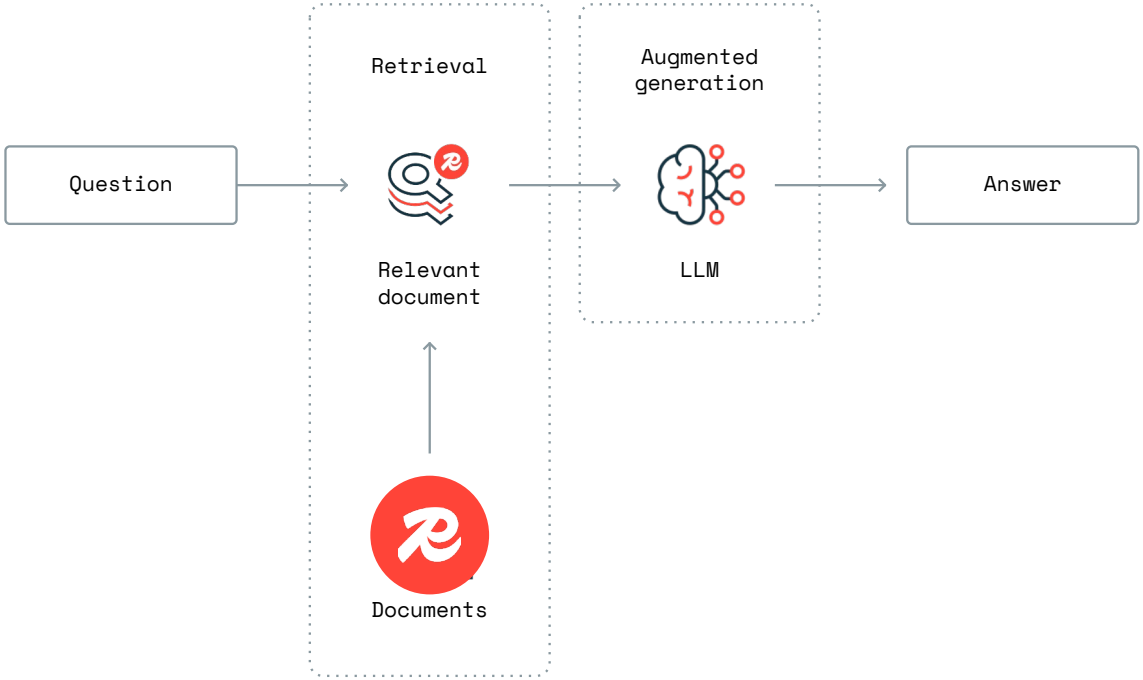
CDC data from RDBMs to Redis

# RAG

## Retrieval-augmented generation

A pattern where any and all related content is retrieved from a trusted data source, augmented with a user request, and sent to an LLM to generate a response.

- → Reduce hallucinations by inserting relevant info into the LLM context
- → Stay fresh by adding up-to-date details and proprietary info into LLM responses

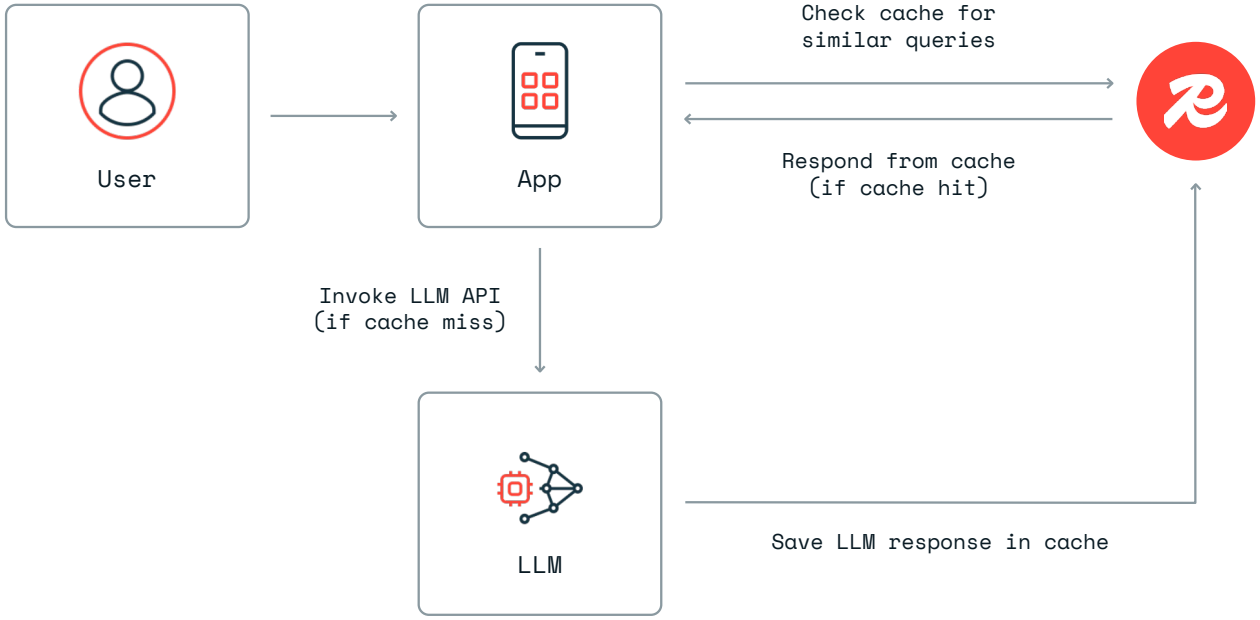


# Semantic caching

Caching with AI

Semantic caching reduces the external calls to LLMs, saving money and decreasing app response times.

- → **15x faster** than complete calling LLM API
- → **Up to 90% less cost** from calls to LLMs



# Redis Agent Memory Server

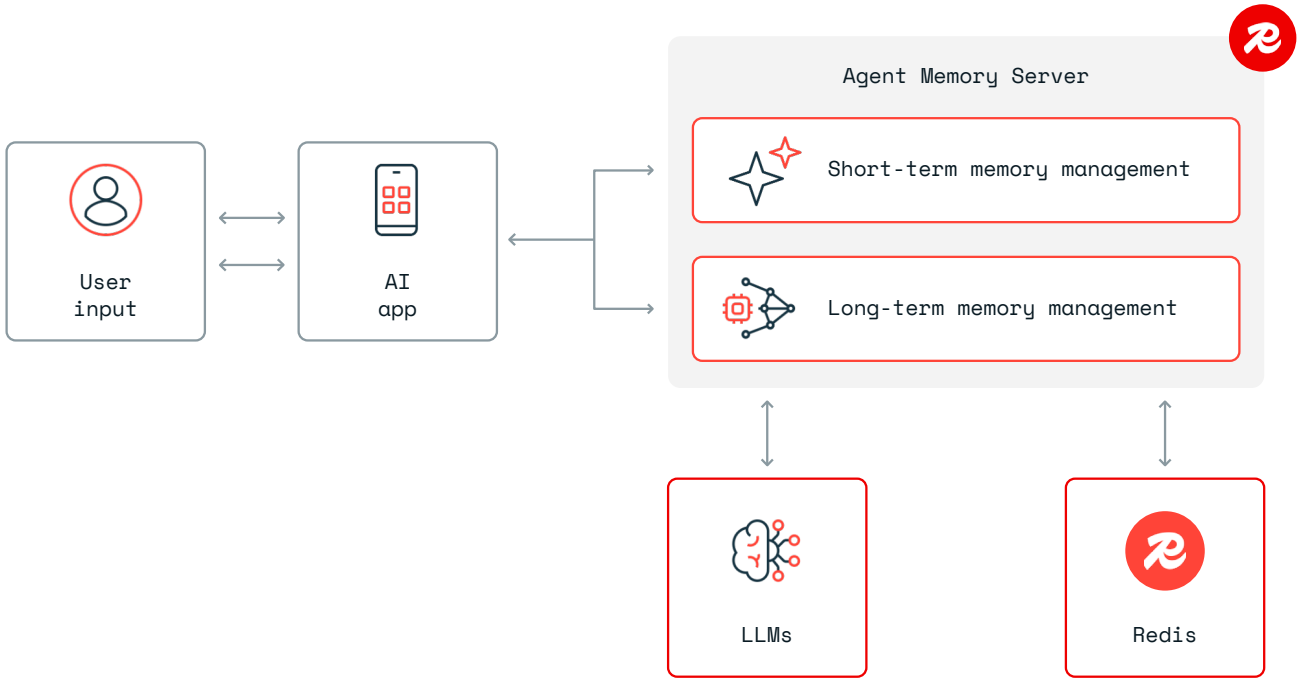
Our Agent Memory Server makes LLM responses more relevant & useful by managing short-term and long-term memory.

Short-term memory

- → **Automatic summarization**
- → **Configurable window sizes** for recent messages

Long-term memory

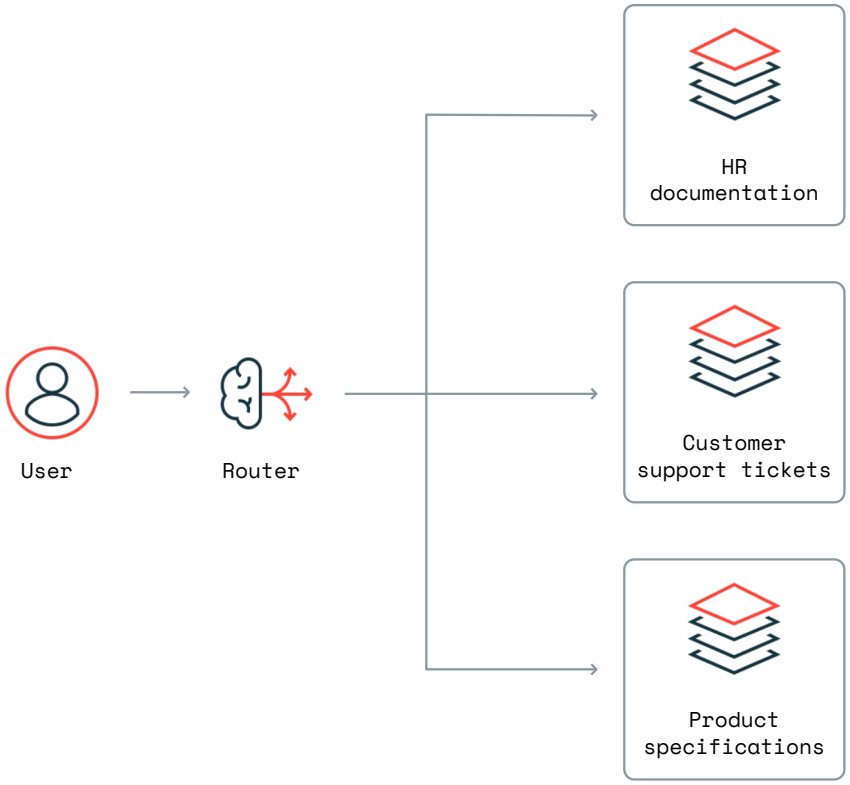
- → **Search** for relevant memories
- → **Extract** topic & named entity recognition
- → **Namespace support** for proper isolation



# Semantic routing

Direct AI queries based on meaning, not just keywords. Using vector search, apps understands intent and routes queries to the best data source, tool, model, or processing route.

- → **Reduce tokens** by choosing the optimal LLM
- → **Protect your company** by adding guardrails for bad behavior

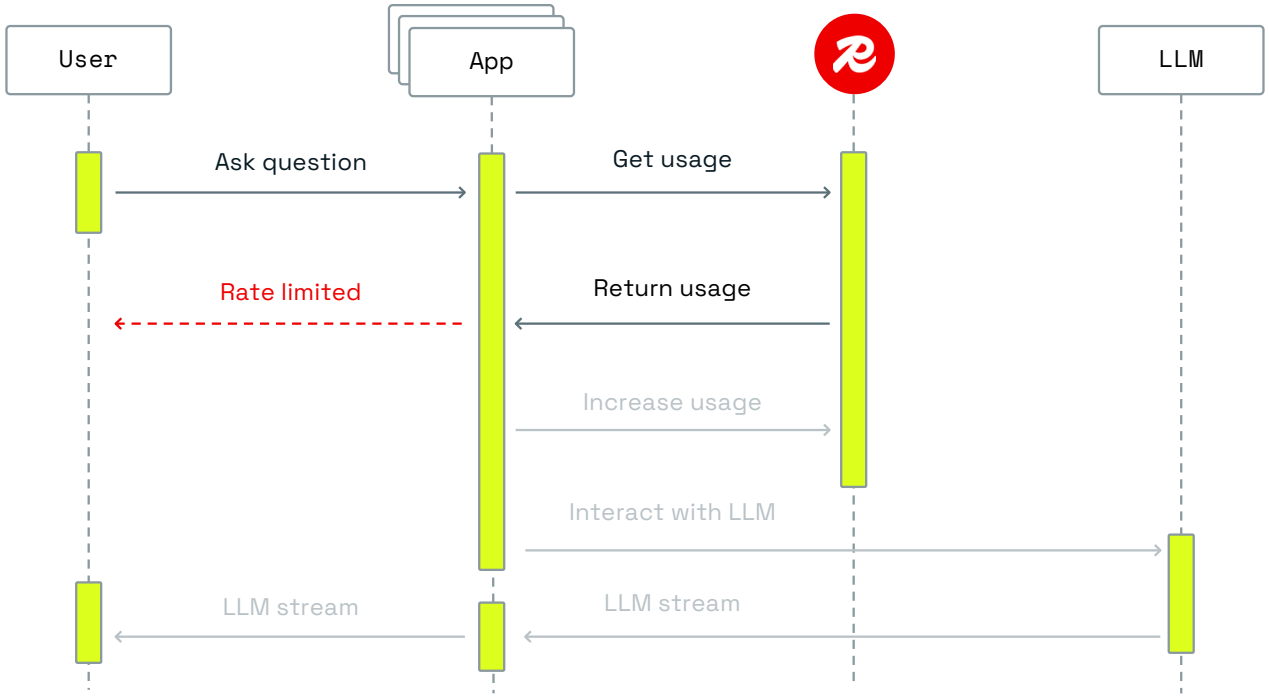


# Rate limiting

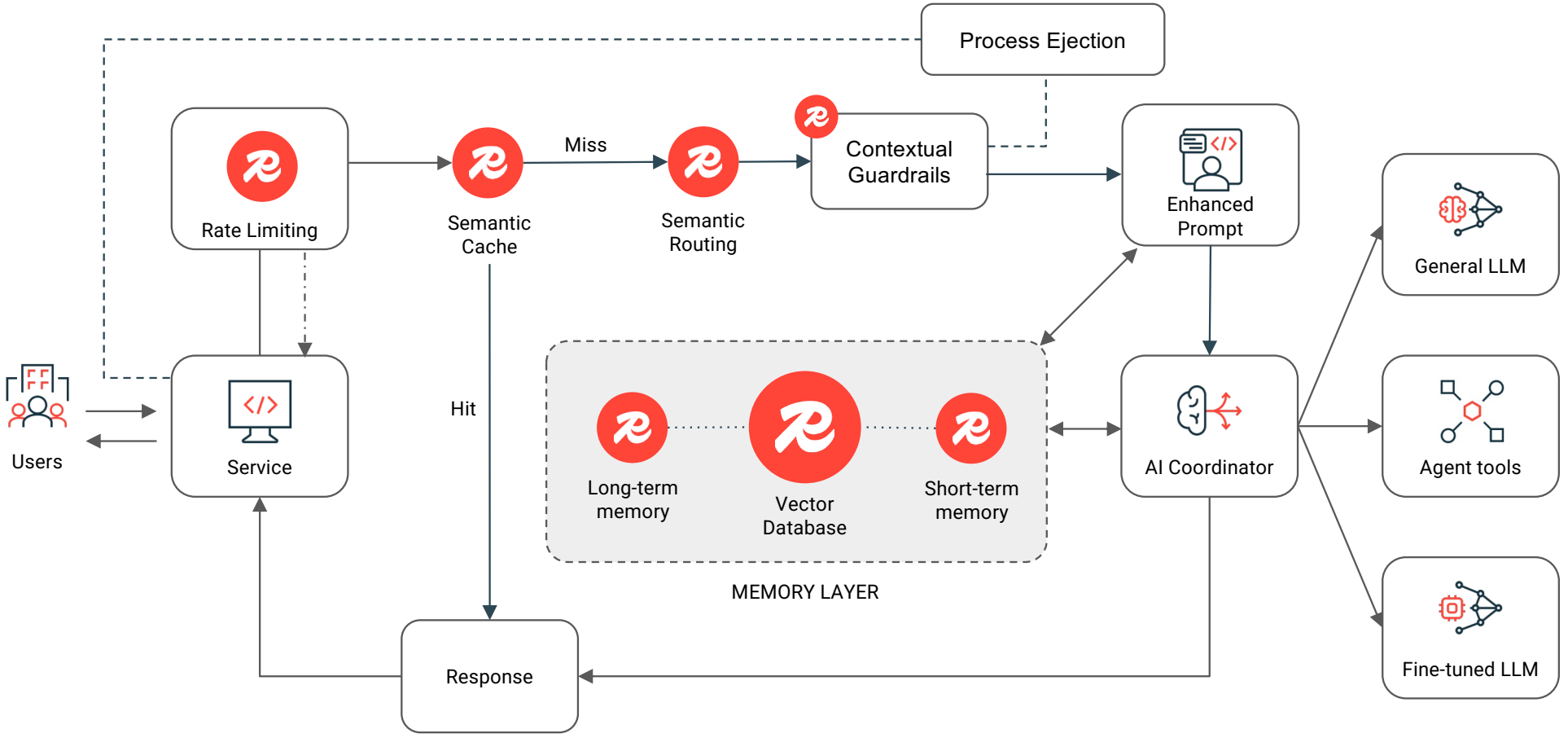
A technique to control the rate at which requests are sent or processed in order to maintain system stability and reduce LLM costs.

→ **Balance loads** across LLMs

→ **Prevent abuse** from bad actors or rogue apps



# Redis Architecture for AI



Red Hat Summit

Connect

# Grazie

 linkedin.com/company/red-hat

 facebook.com/redhatinc

 youtube.com/user/RedHatVideos

 twitter.com/RedHat