Red Hat Summit

Connect

From Zero to Hero Redis RAG Architecture

Luigi Fugaro, Solution Architect
Redis

Rome, November 7th, 2024

Redis | Red Hat

Red Hat Summit

Connect

From Zero to Hero Redis RAG Architecture

Luigi Fugaro, Solution Architect
Redis

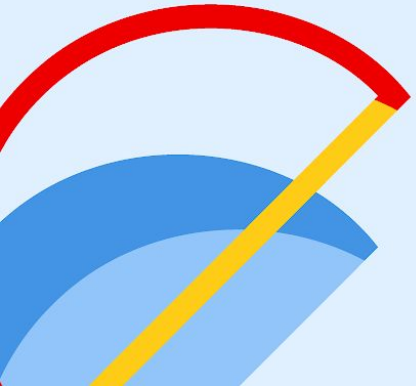
Milan, November 19th, 2024

Redis | Red Hat

Agenda

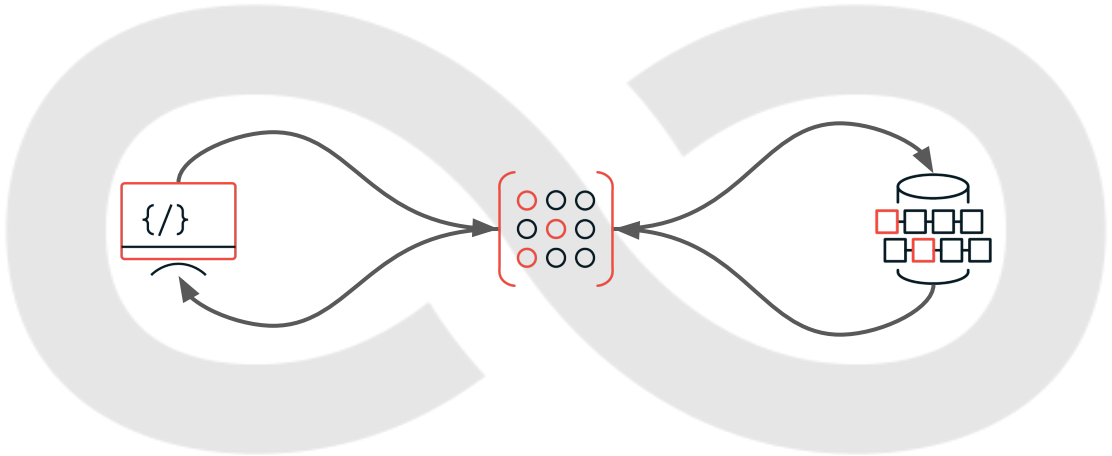
- RAG, key concepts
- Infrastructure
- OpenShift AI
- Redis Enterprise for AI
- Advantages
- Q/A

Retrieval Augmented Generation



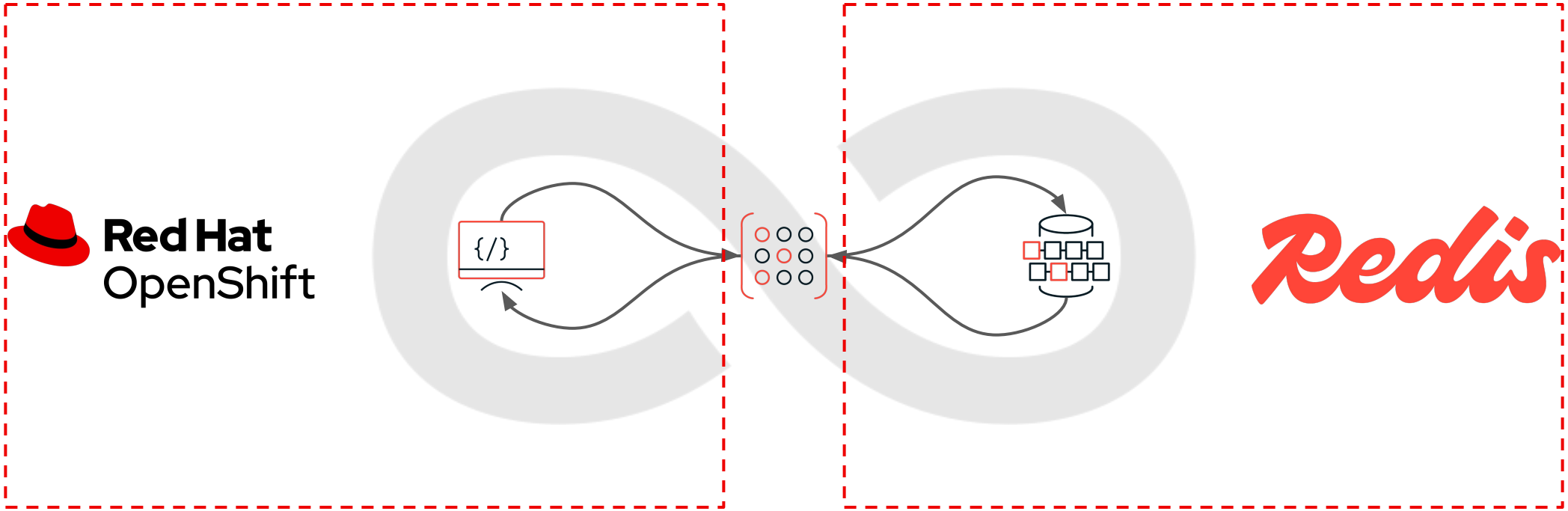
Redis | Red Hat

Retrieval Augmented Generation



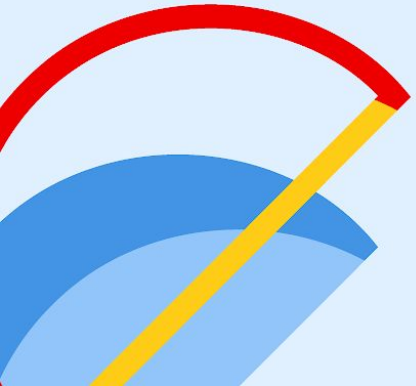
Macro components: Application, Model, Vector database

Retrieval Augmented Generation



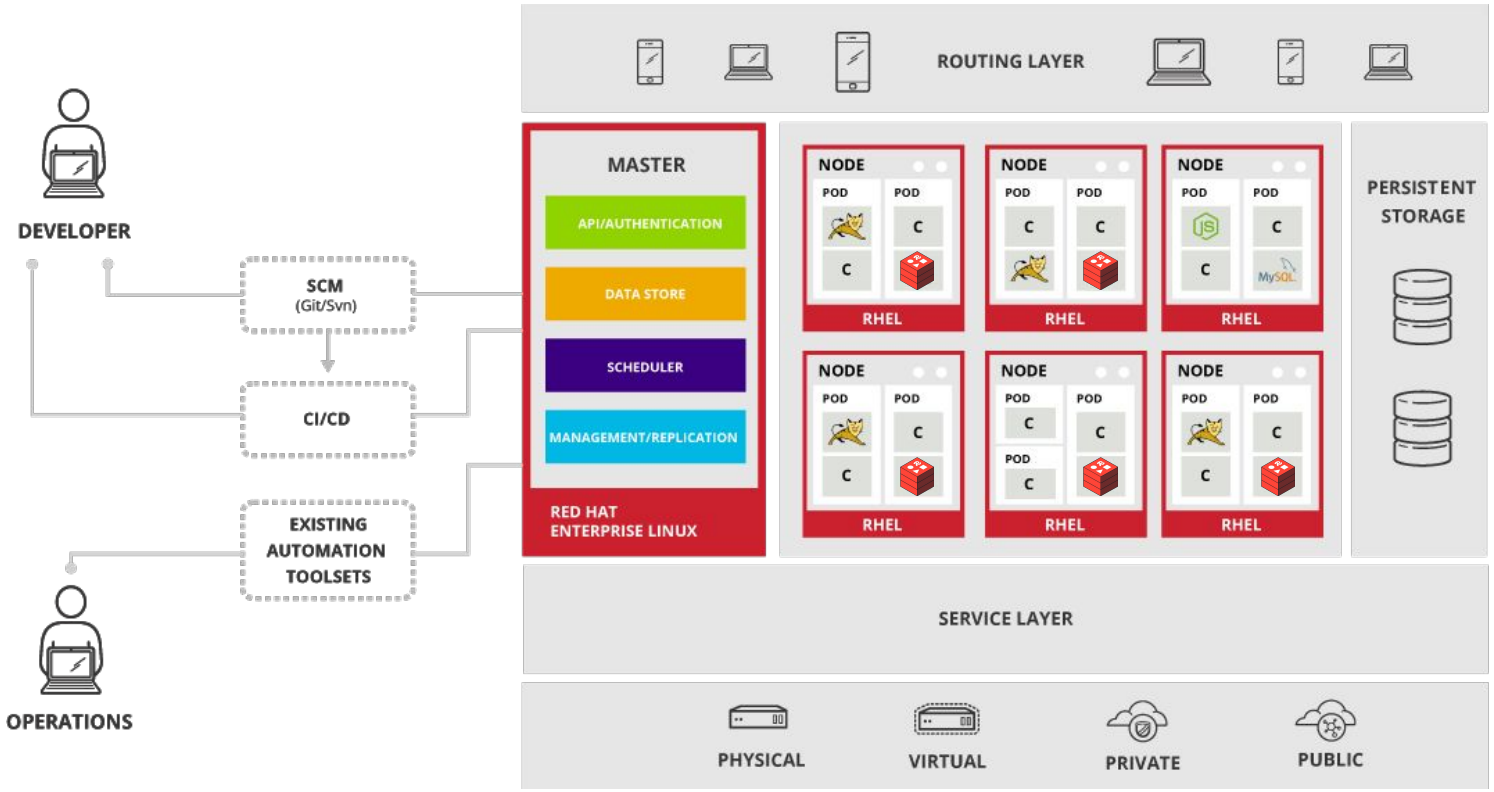
Macro components: Application, Model, Vector database

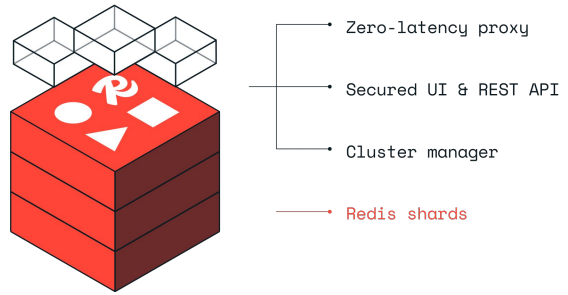
Infrastructure



Redis | Red Hat

Infrastructure

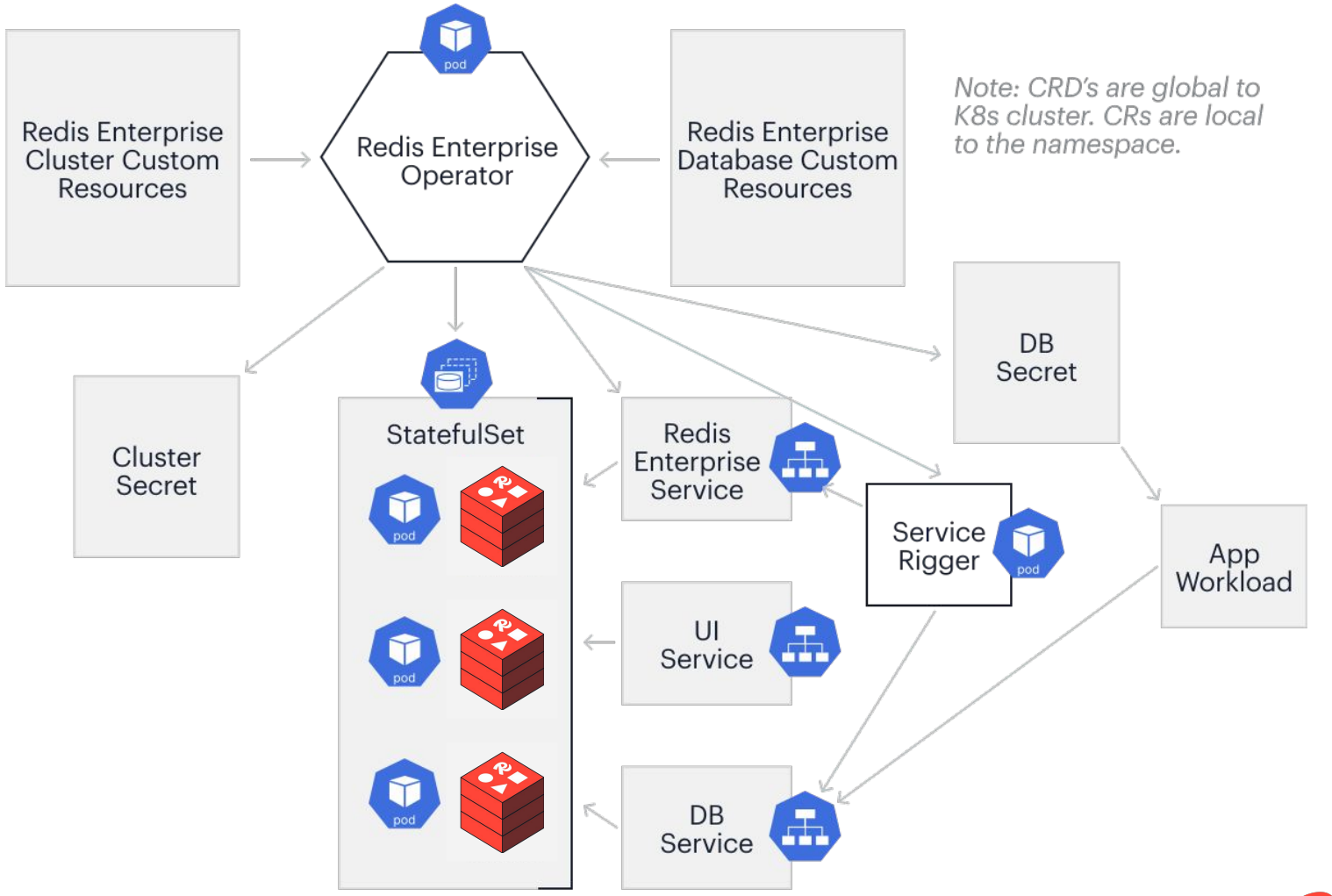




Macro components: Application, Model, Vector database

All LLMs are models, but not all models are LLMs.

Infrastructure



Infrastructure

Certified enterprise ready

About certification

Redis Enterprise Software

By Redis Enterprise

A real-time data platform with high performance caching. The best version of the most loved database in the world. Combine your caching layer with a real-time database to provide instantaneous access to API responses, session data, and DBMS data.

Software version	Runs on	Delivery method
7.4	**OpenShift 4.6 - 4.18**	**Operator**

Purchase

Free trial

Contact sales

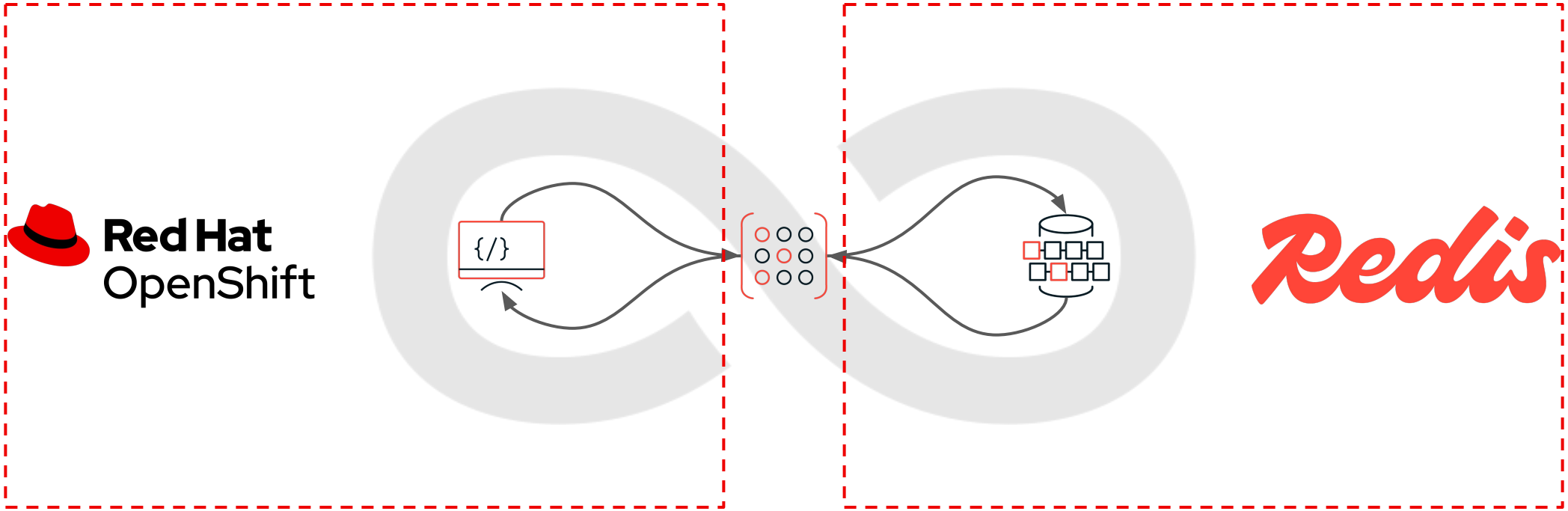
Highlights

- Runs on OpenShift
- Certified operators
- Fully containerized
- Vendor supported
- Vulnerability scans

Overview | Documentation | Pricing | Help

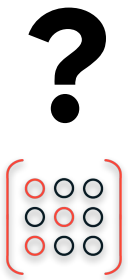
A robust, in-memory database platform built by the same people who develop open source Redis. It maintains the simplicity and high performance of Redis and adds many enterprise-grade capabilities for companies running their business in the cloud, on-prem and hybrid models. Redis Enterprise provides customers real-time performance with linear scaling to hundreds of millions of operations per second while providing local latency in a global Active-Active deployment with 99.999% uptime.

Retrieval Augmented Generation



Macro components: Application, Model, Vector database

Retrieval Augmented Generation



Macro components: Application, Model, Vector database

All LLMs are models, but not all models are LLMs.

Retrieval Augmented Generation

Architecture

Output

Parameter

Performance metrics

Training Data

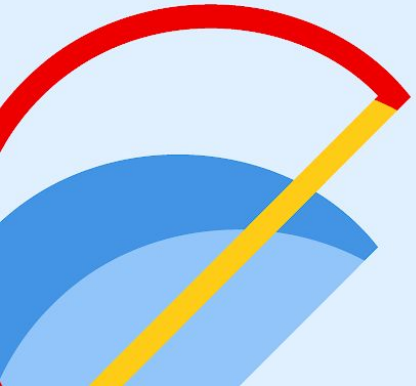
Inference Speed

Version

Macro components: Application, Model, Vector database

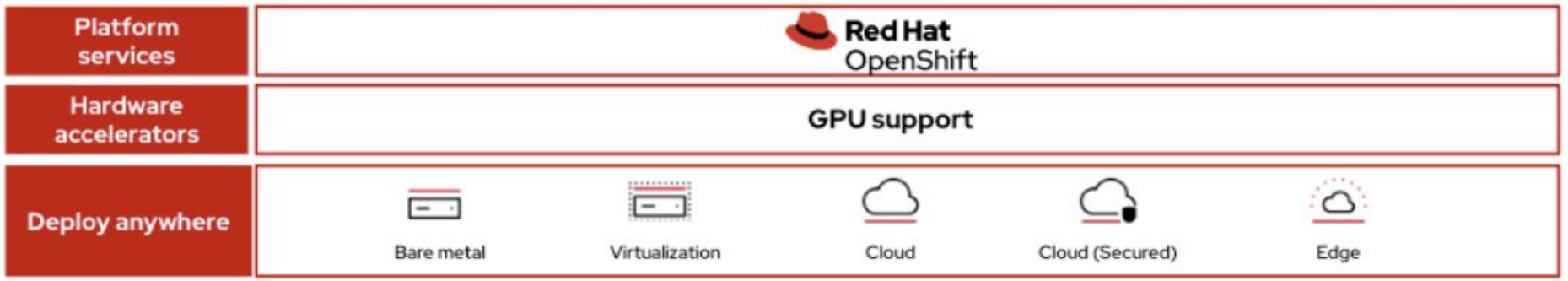
All LLMs are models, but not all models are LLMs.

OpenShift AI



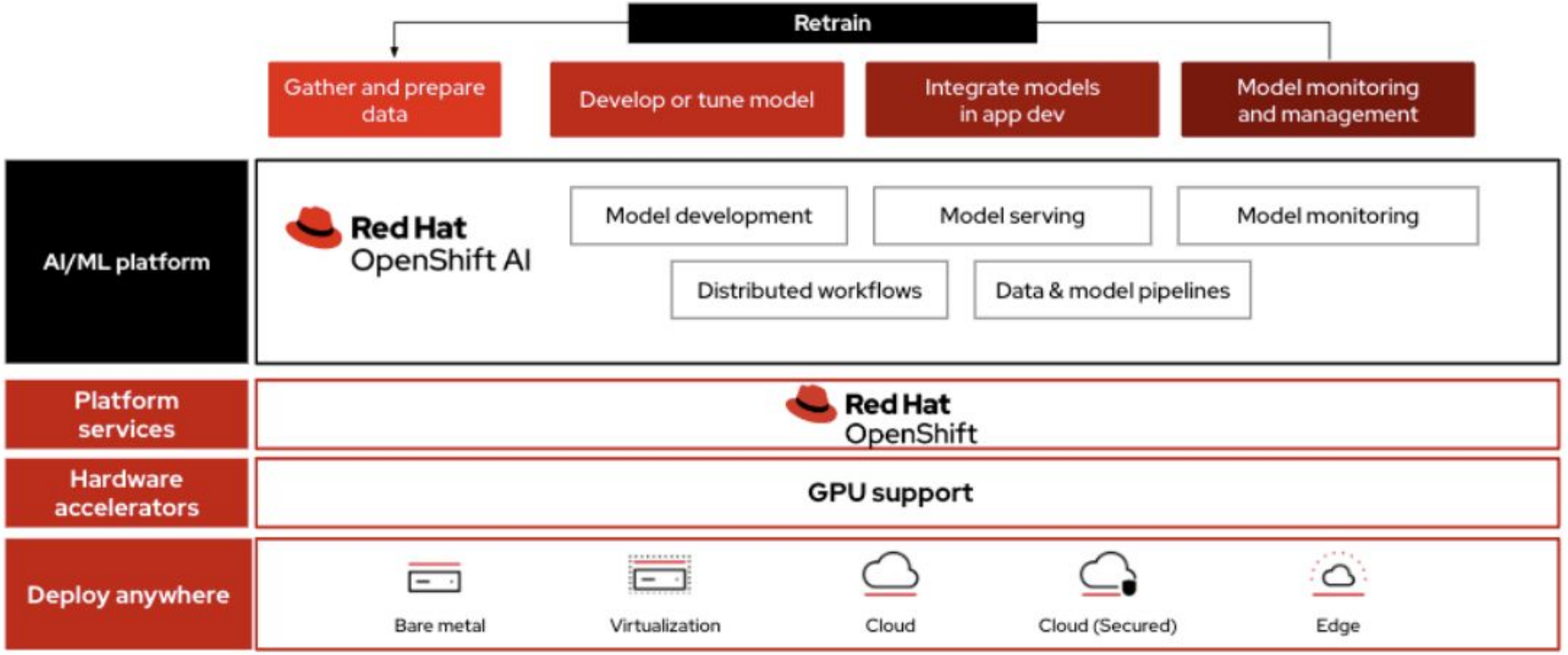
Redis | Red Hat

OpenShift



OpenShift - Enterprise Kubernetes Distribution.

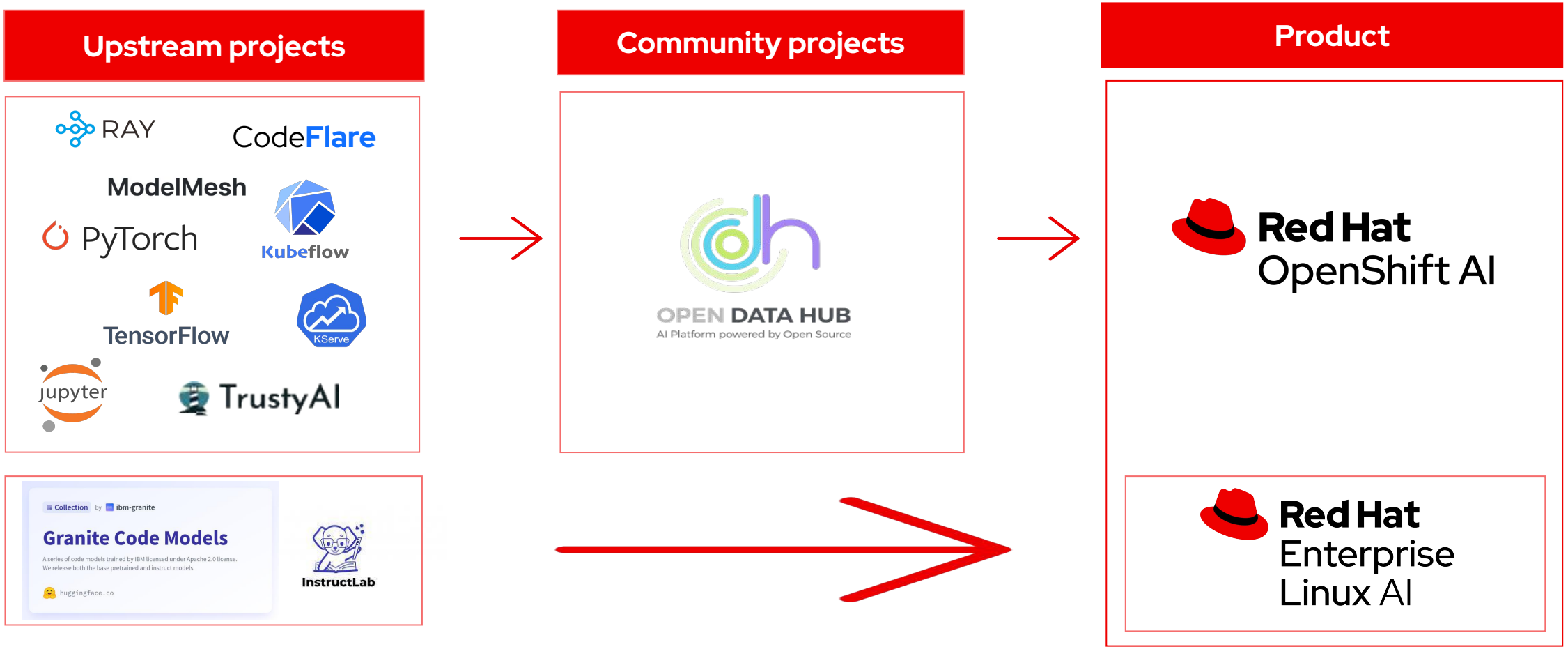
OpenShift AI



OpenShift - Enterprise Kubernetes Distribution.

OpenShift AI - MLOps platform managing models life cycle.

OpenShift AI



OpenShift - Enterprise Kubernetes Distribution.

OpenShift AI - MLOps platform managing models life cycle.

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.

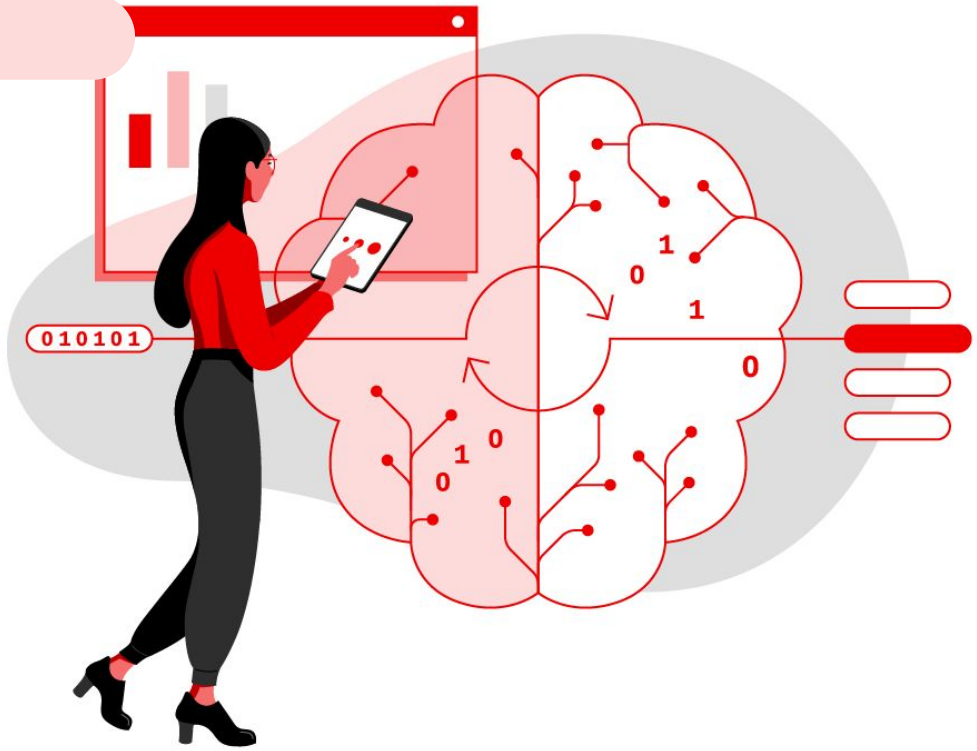
Lifecycle management

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.

Make model serving more flexible and transparent



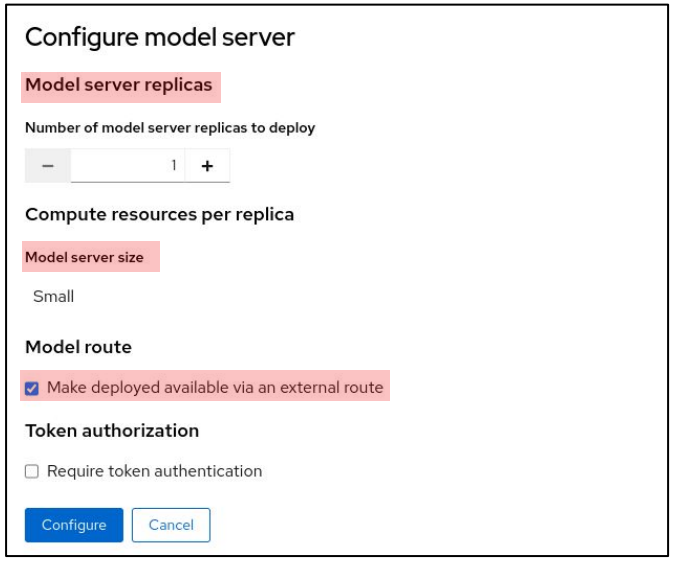
- **Use model-serving user interface (UI)**
  integrated within product dashboard and projects workspace
- **Serve models**
  from providers like Hugging Face and Stability AI or Granite models
- **Support a variety of model frameworks**
  including TensorFlow, PyTorch, and ONNX
- **Choose inference servers**
  either out-of-the-box options optimized for model types
  or your own custom inference server
- **Scale cluster resources**
  up or down as your workload requires

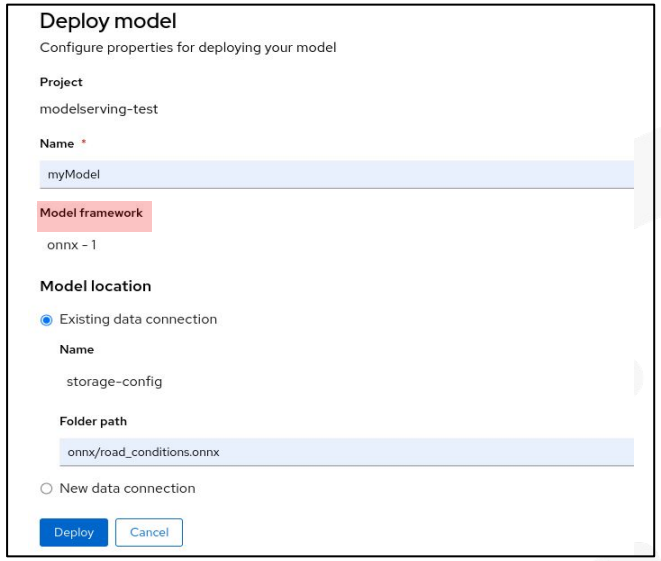
Red Hat

Serve, scale, and monitor your models

Select the required resources and scale model serving as needed

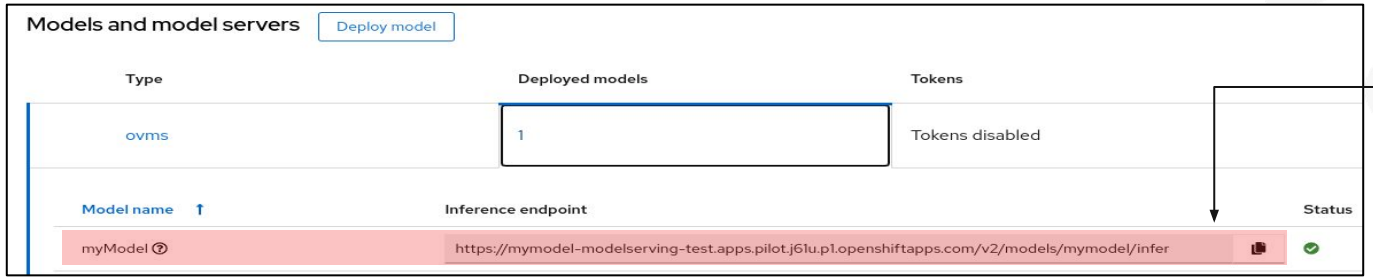
Make your model public and secure





Select your model framework

View your deployed model fleet endpoints



Red Hat

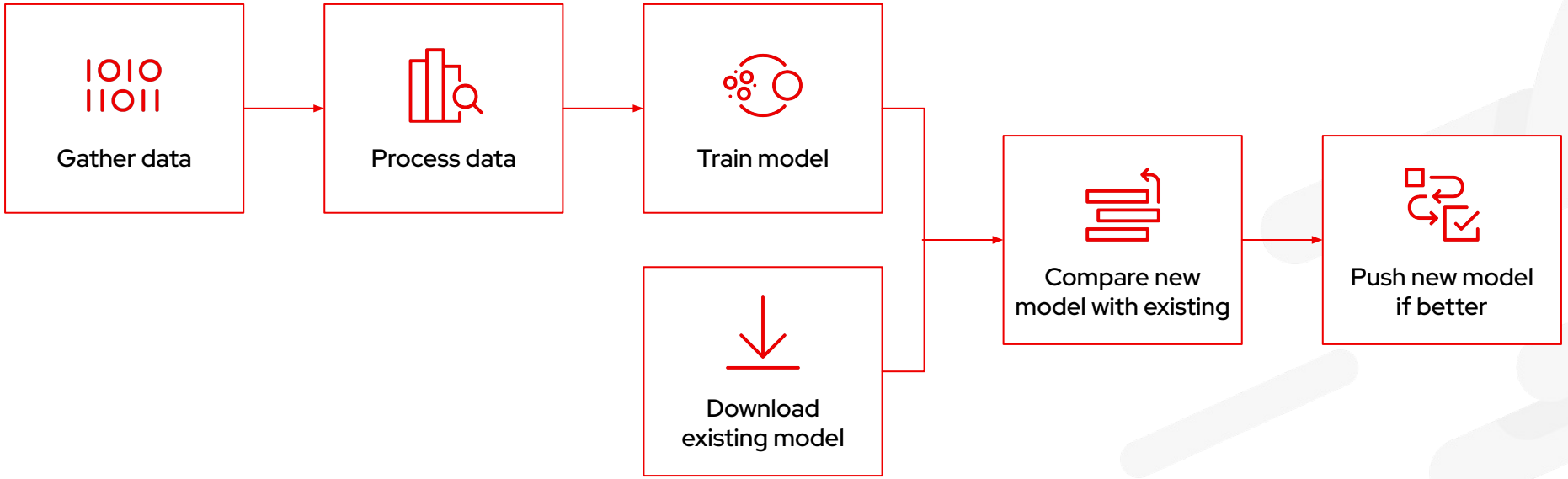
Model performance metrics

Access a range of model performance metrics to build your own visualizations or integrate data with other observability services

- Out-of-the-box visualizations for performance and operations metrics

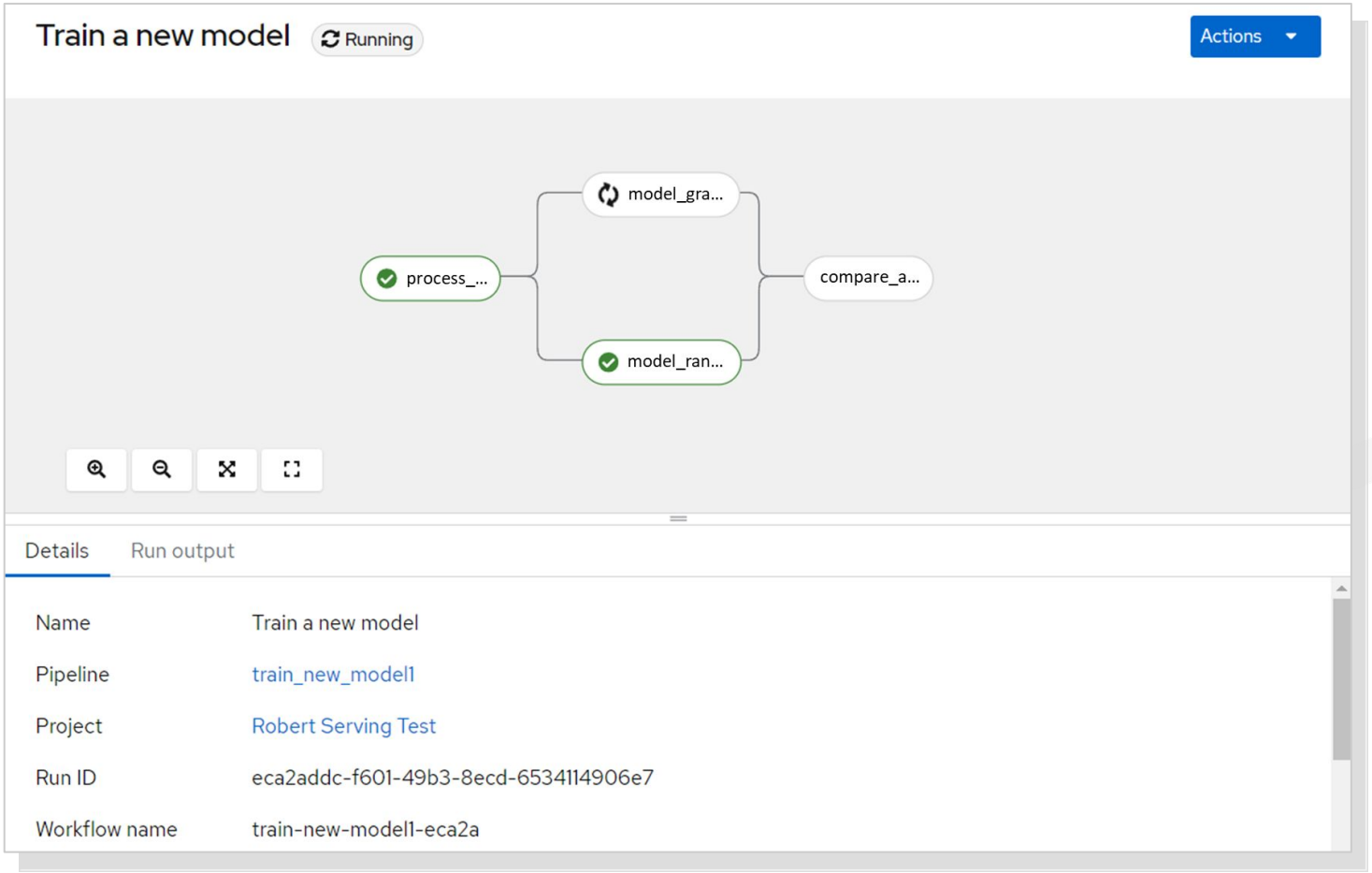
Red Hat

Data science pipelines component



- Continuously deliver and test models in production
- Schedule, track, and manage pipeline runs
- Easily build pipelines using graphical front end
- Orchestrate data science tasks into pipelines
- Chain together processes like data prep, build models, and serve models
- Data science pipelines are based on upstream Kubeflow pipelines

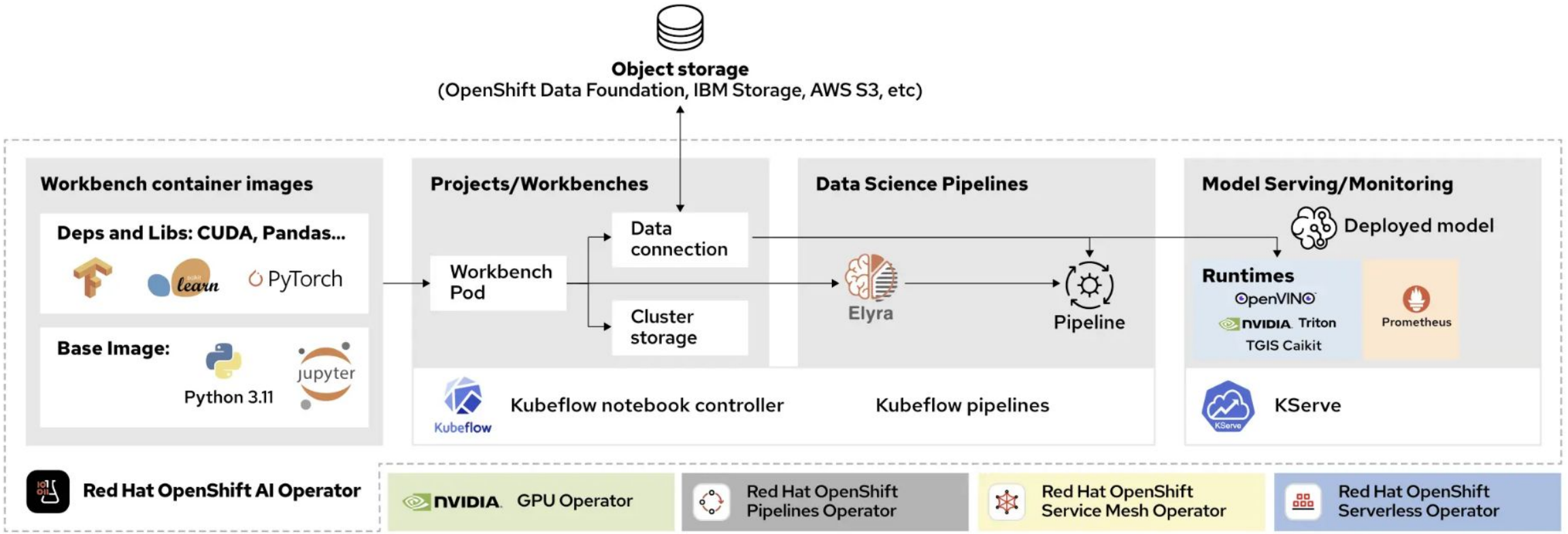
Red Hat OpenShift data science pipelines user interface



The OpenShift AI user interface enables you to track and manage pipelines and pipeline runs.

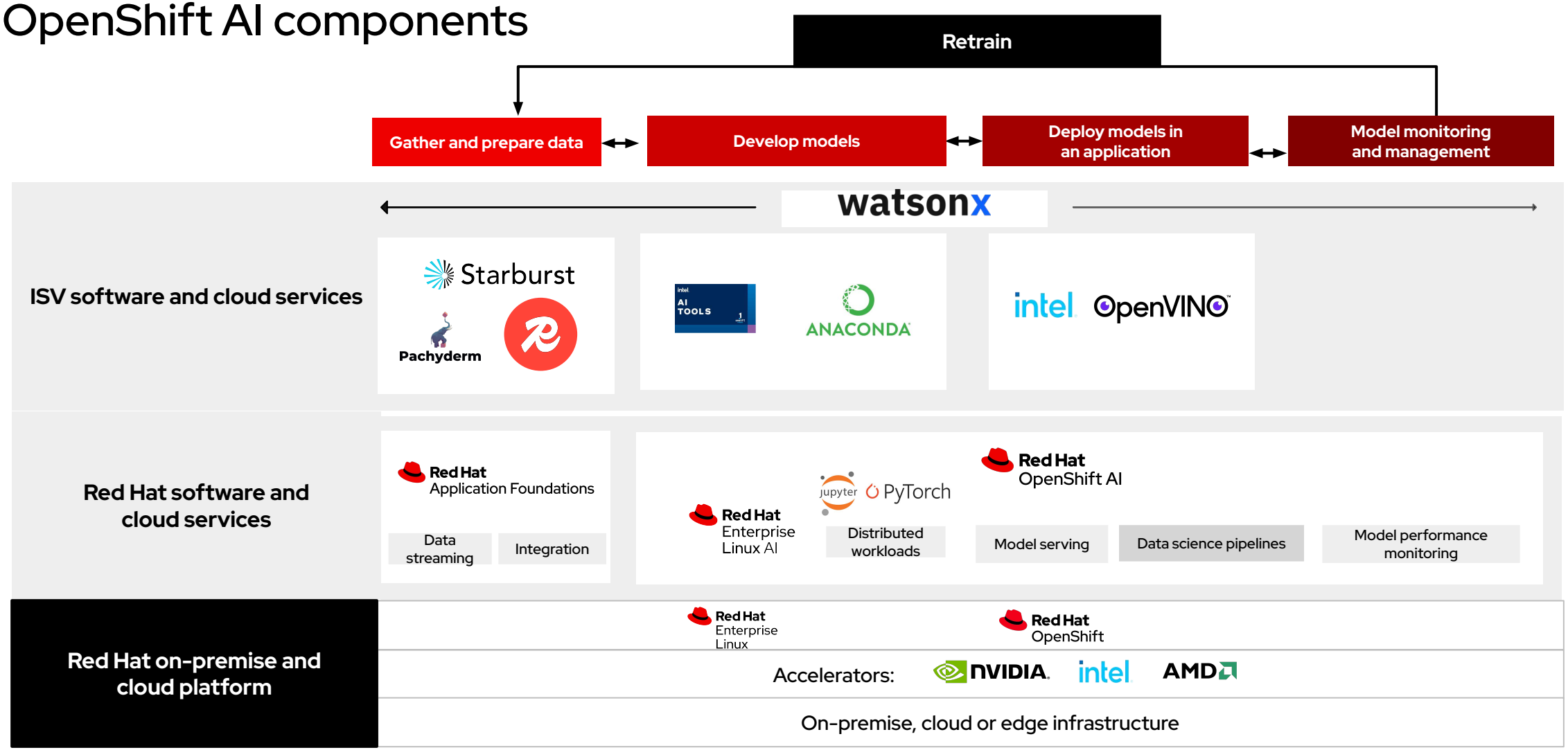
Red Hat

Red Hat OpenShift AI



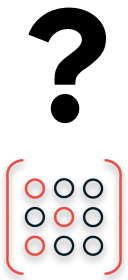
Red Hat

OpenShift AI components



Redis | Red Hat

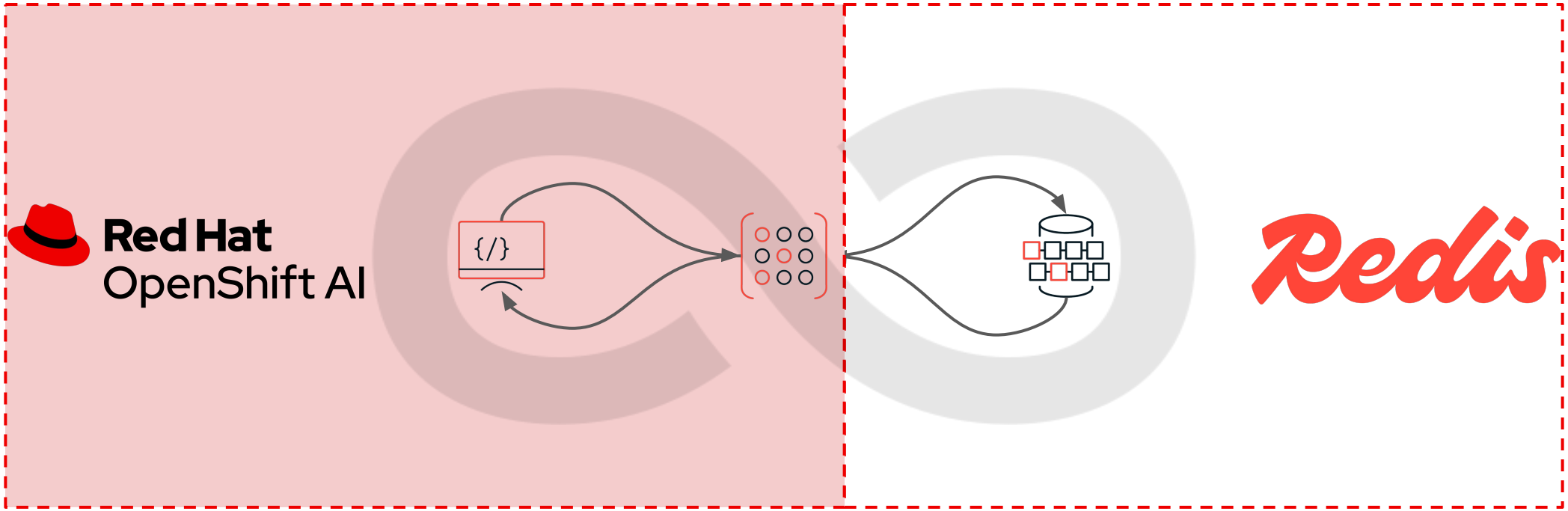
Retrieval Augmented Generation



Macro components: Application, Model, Vector database

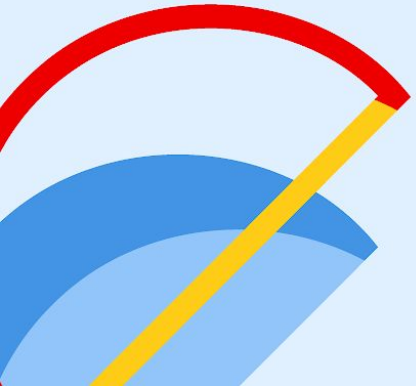
All LLMs are models, but not all models are LLMs.

Retrieval Augmented Generation



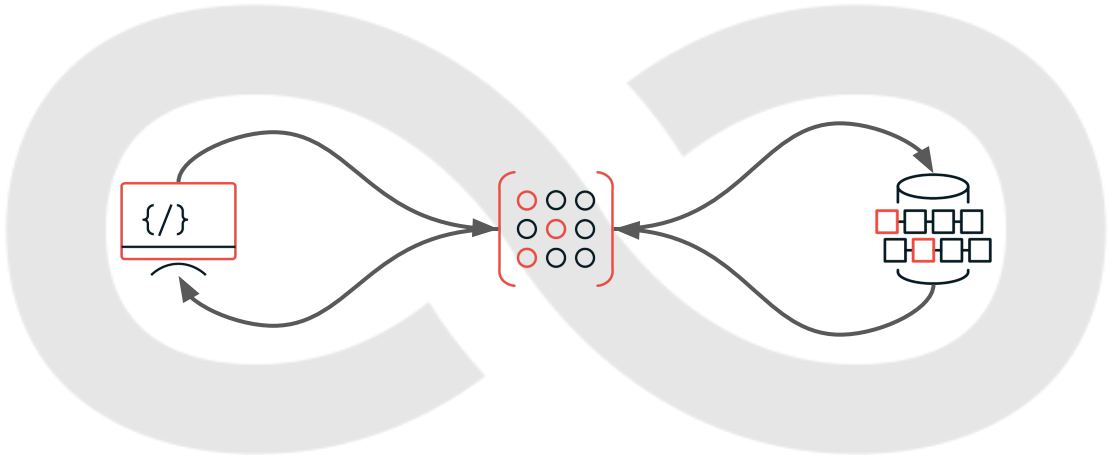
Macro components: Application, Model, Vector database

Redis Enterprise for AI



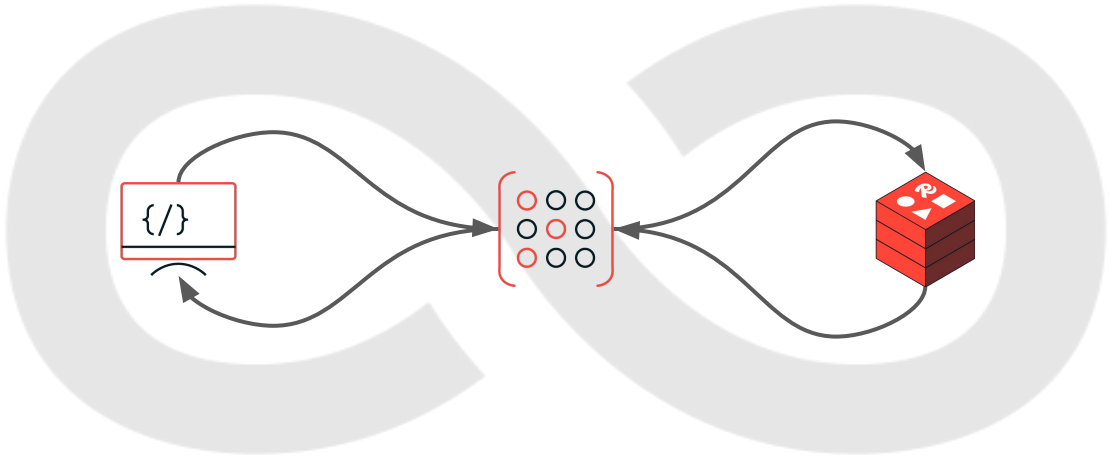
Redis | Red Hat

Retrieval Augmented Generation



Macro components: Application, Model, Vector database

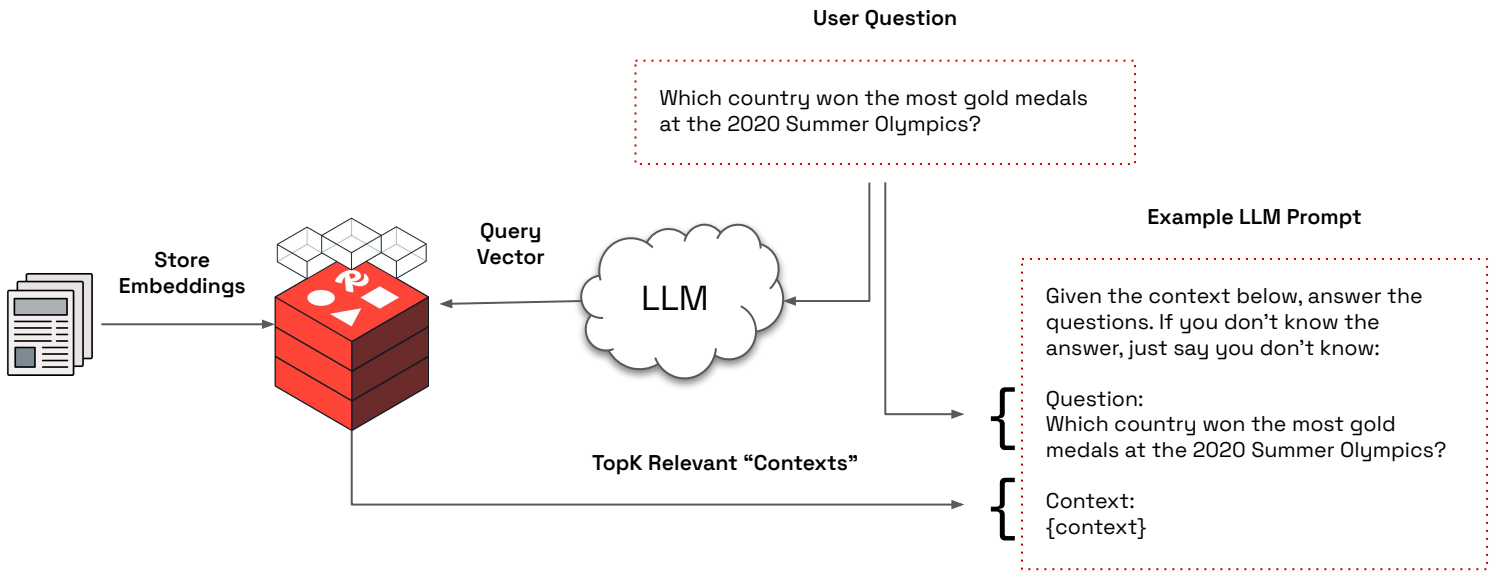
Retrieval Augmented Generation



Macro components: Application, Model, Vector database

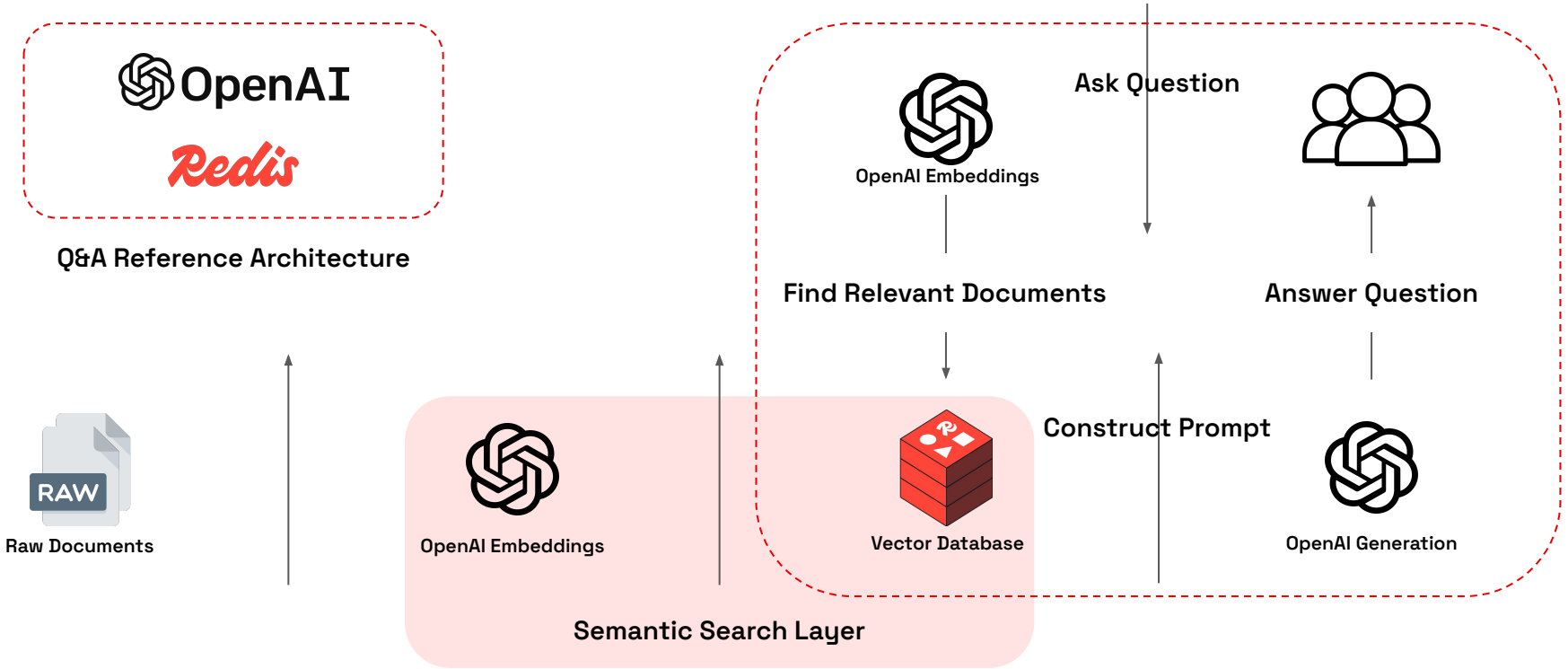
Retrieval Augmented Generation

Context Retrieval



Macro components: Application, Model, Vector database

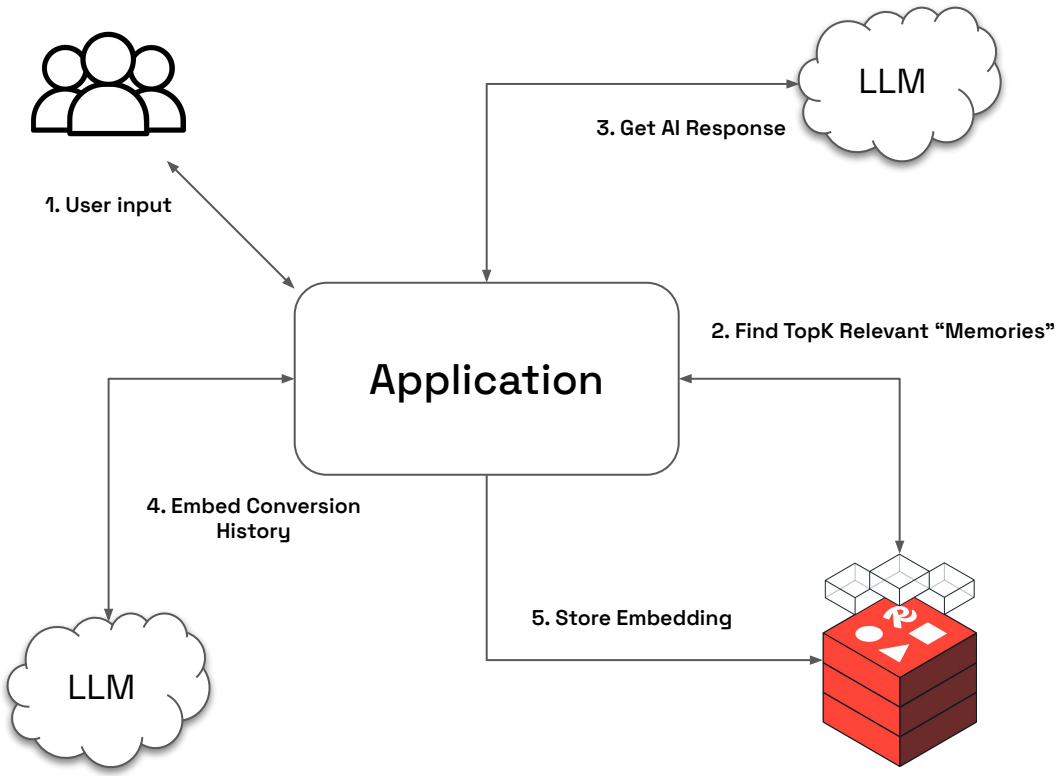
Retrieval Augmented Generation
Question & Answering



Macro components: Application, Model, Vector database

Retrieval Augmented Generation

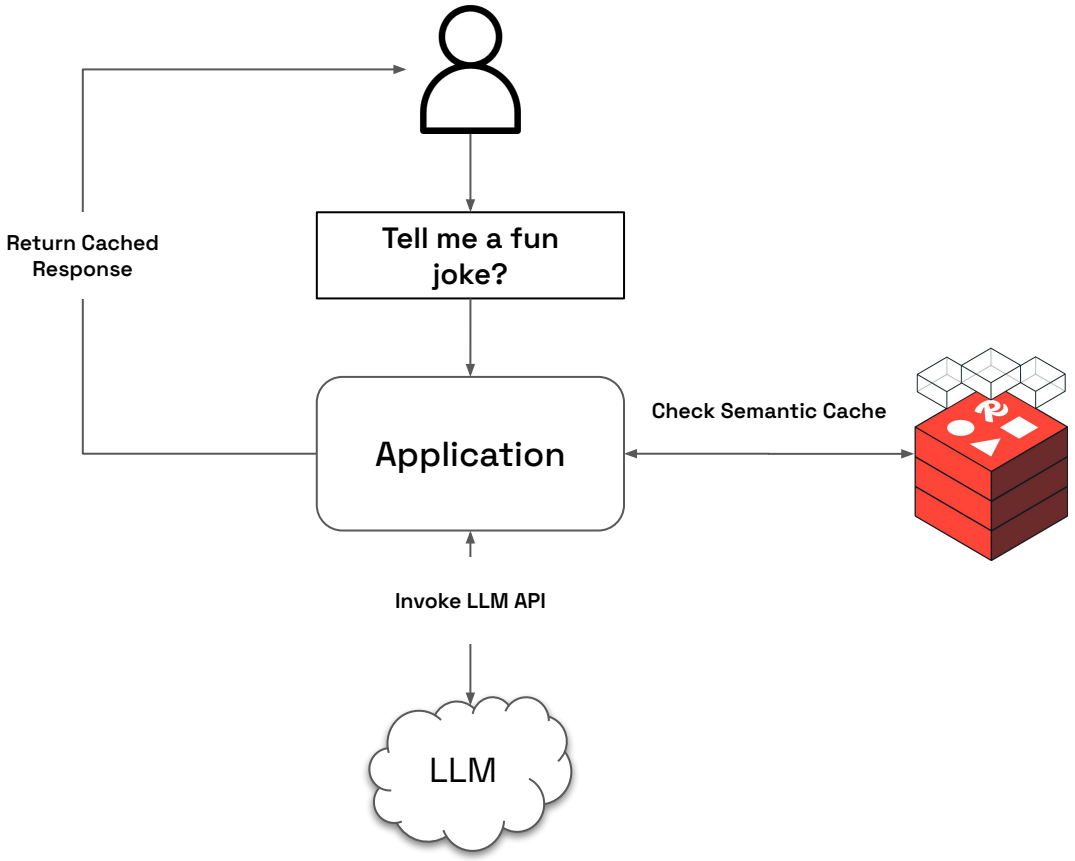
LLM Conversion Memory



Macro components: Application, Model, Vector database

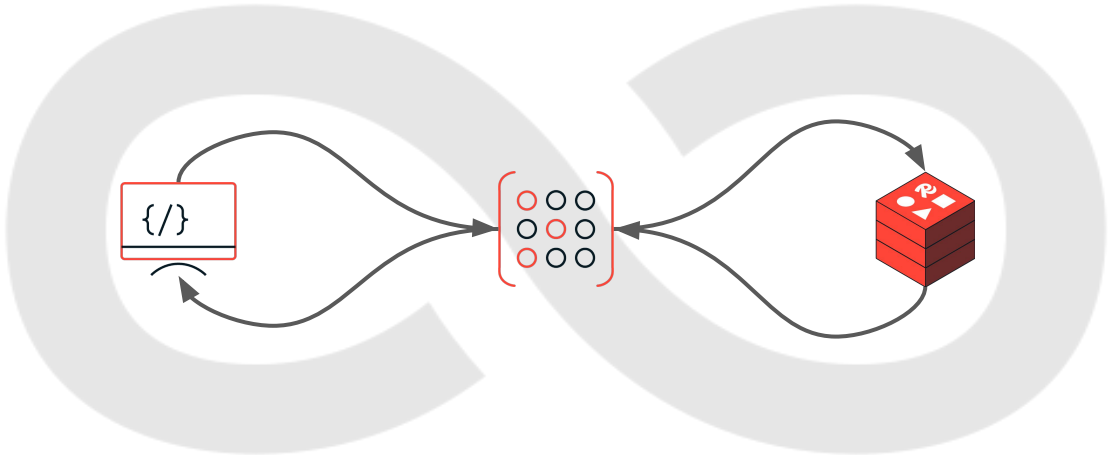
Retrieval Augmented Generation

Semantic Caching



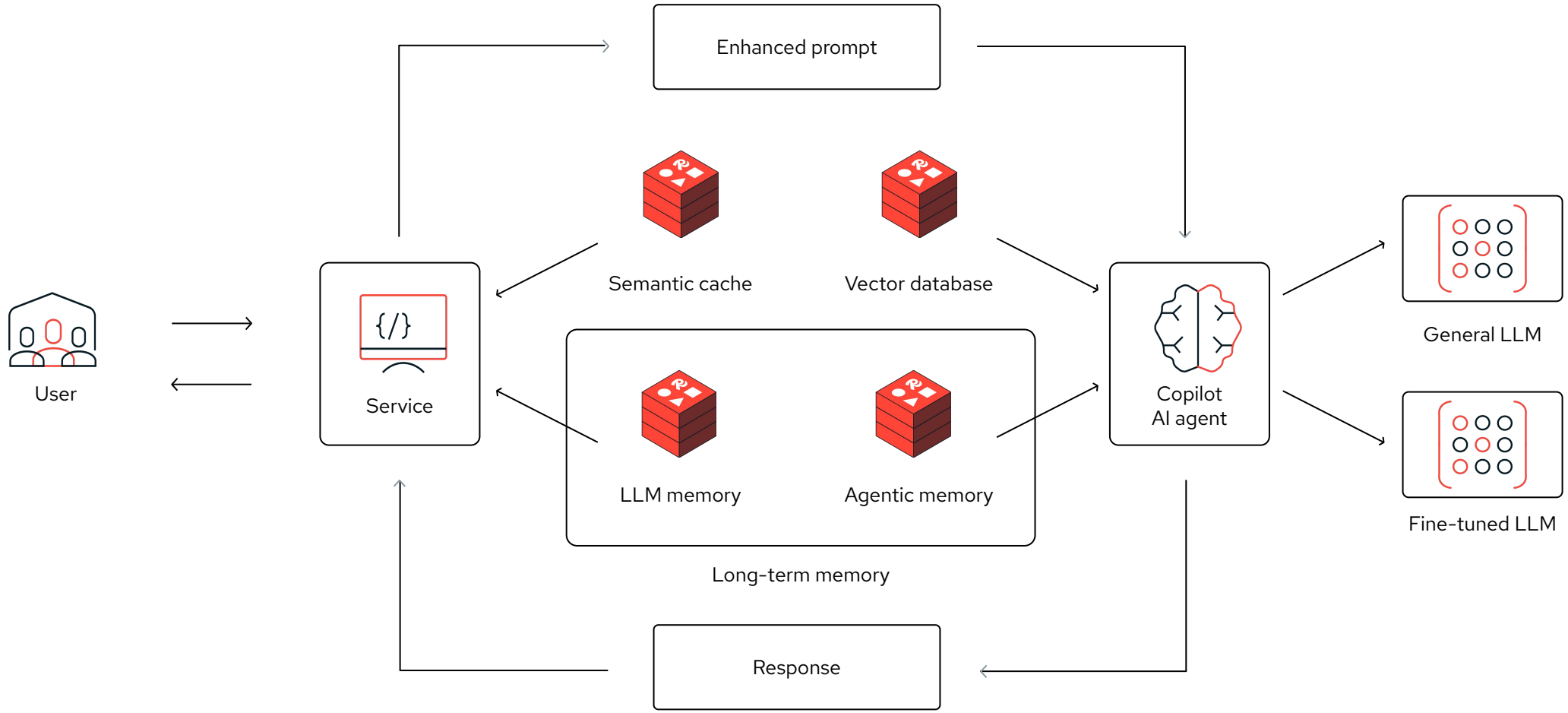
Macro components: Application, Model, Vector database

Retrieval Augmented Generation



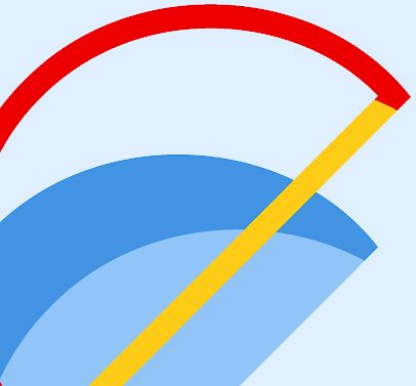
Macro components: Application, Model, Vector database

Retrieval Augmented Generation



Macro components: Application, Model, Vector database

Advantages



Retrieval Augmented Generation

Accurate
Less hallucinations

Transparent
Allows quotation/citation

Up to date
Knowledge without model training

Secure
Respects user level CRUD privileges

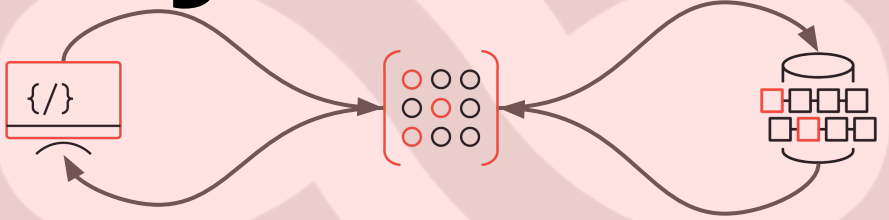
Cost effective
Smaller and more efficient models

Standard models
Standardized models keeping differentiation

Macro components: Application, Model, Vector database

Retrieval Augmented Generation

Red Hat OpenShift



Redis

Fully Featured

Private AI

Macro components: Application, Model, Vector database

Red Hat Summit

Connect

Q&A

Redis | Red Hat

Red Hat Summit

Connect

Thank you

Redis | Red Hat