

Red Hat
Summit

Connect

From Zero to Hero Redis RAG Architecture

Luigi Fugaro, Solution Architect
Redis

Milan, November 19th, 2024

Redis

 Red Hat

Red Hat
Summit

Connect

From Zero to Hero Redis RAG Architecture

Luigi Fugaro, Solution Architect
Redis

Milan, November 19th, 2024

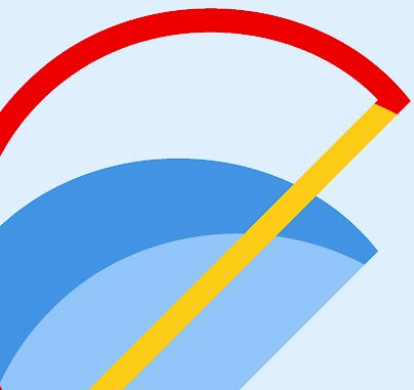
Redis

 Red Hat

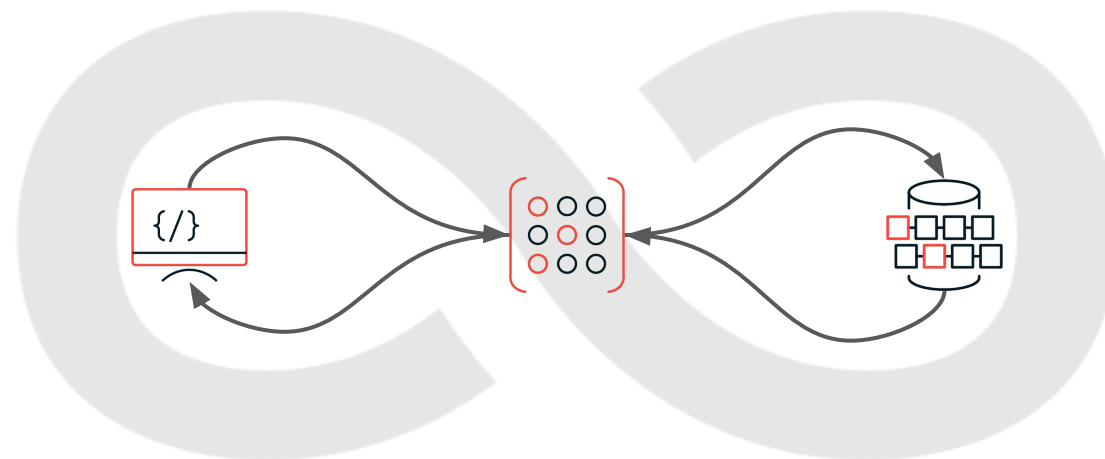
Agenda

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

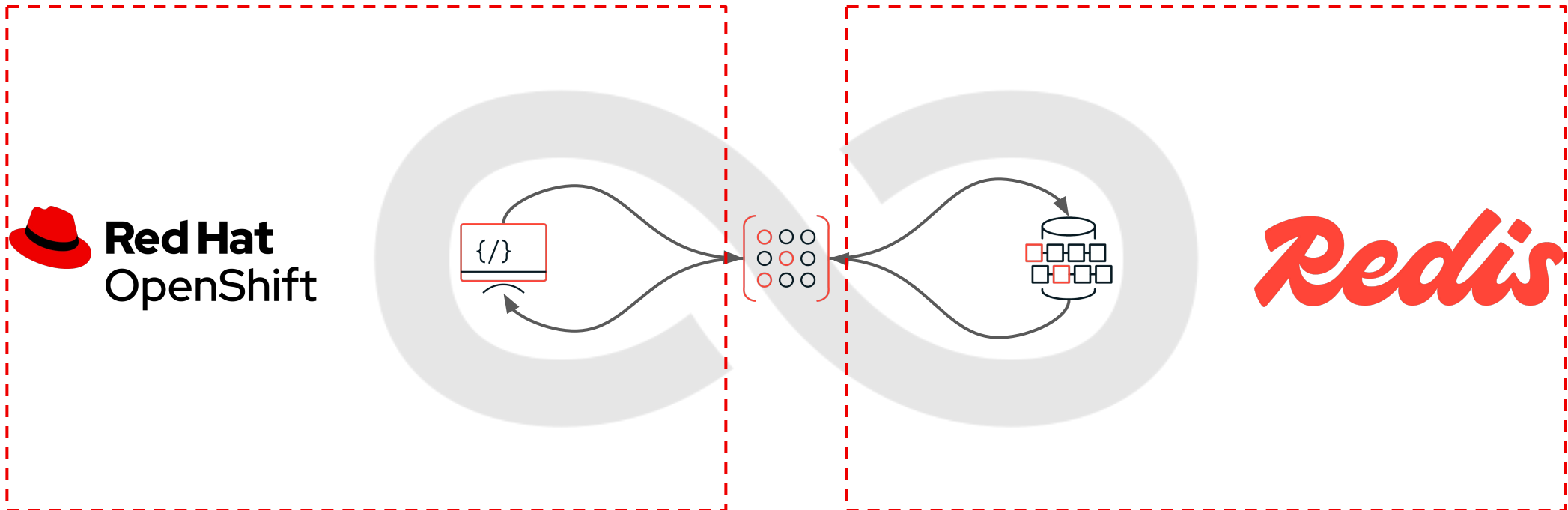
Retrieval Augmented Generation



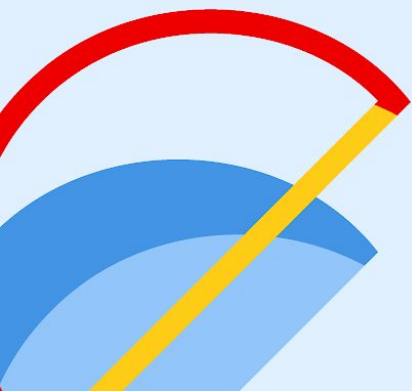
Retrieval Augmented Generation



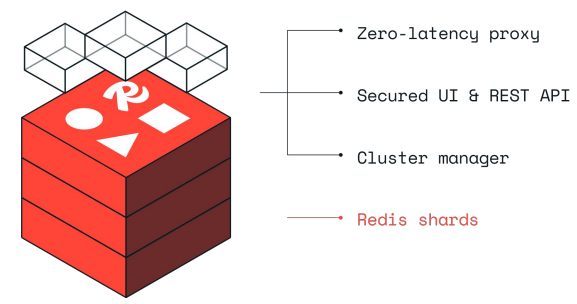
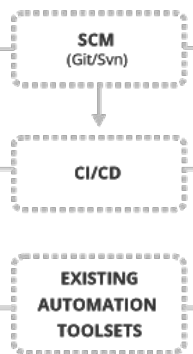
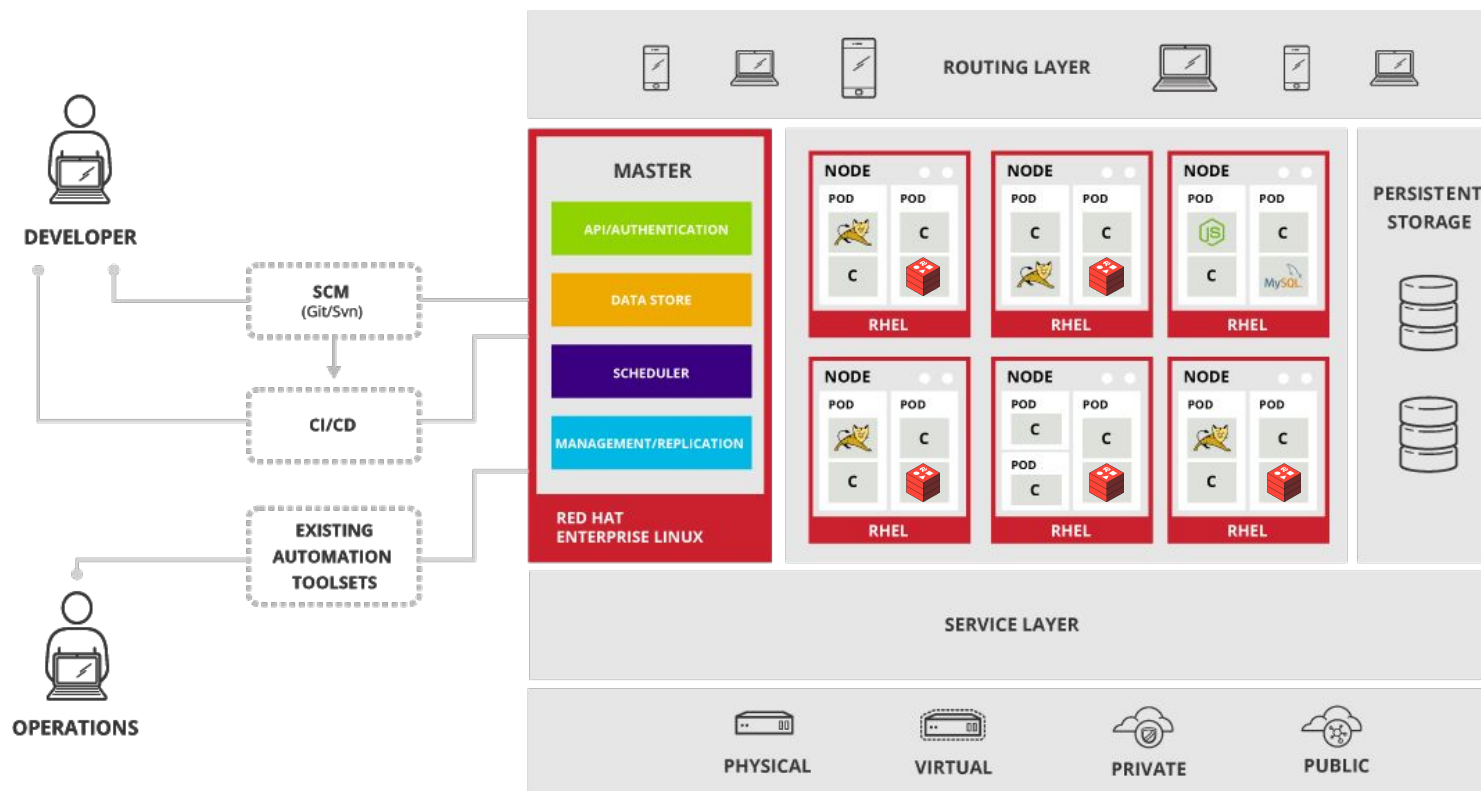
Retrieval Augmented Generation



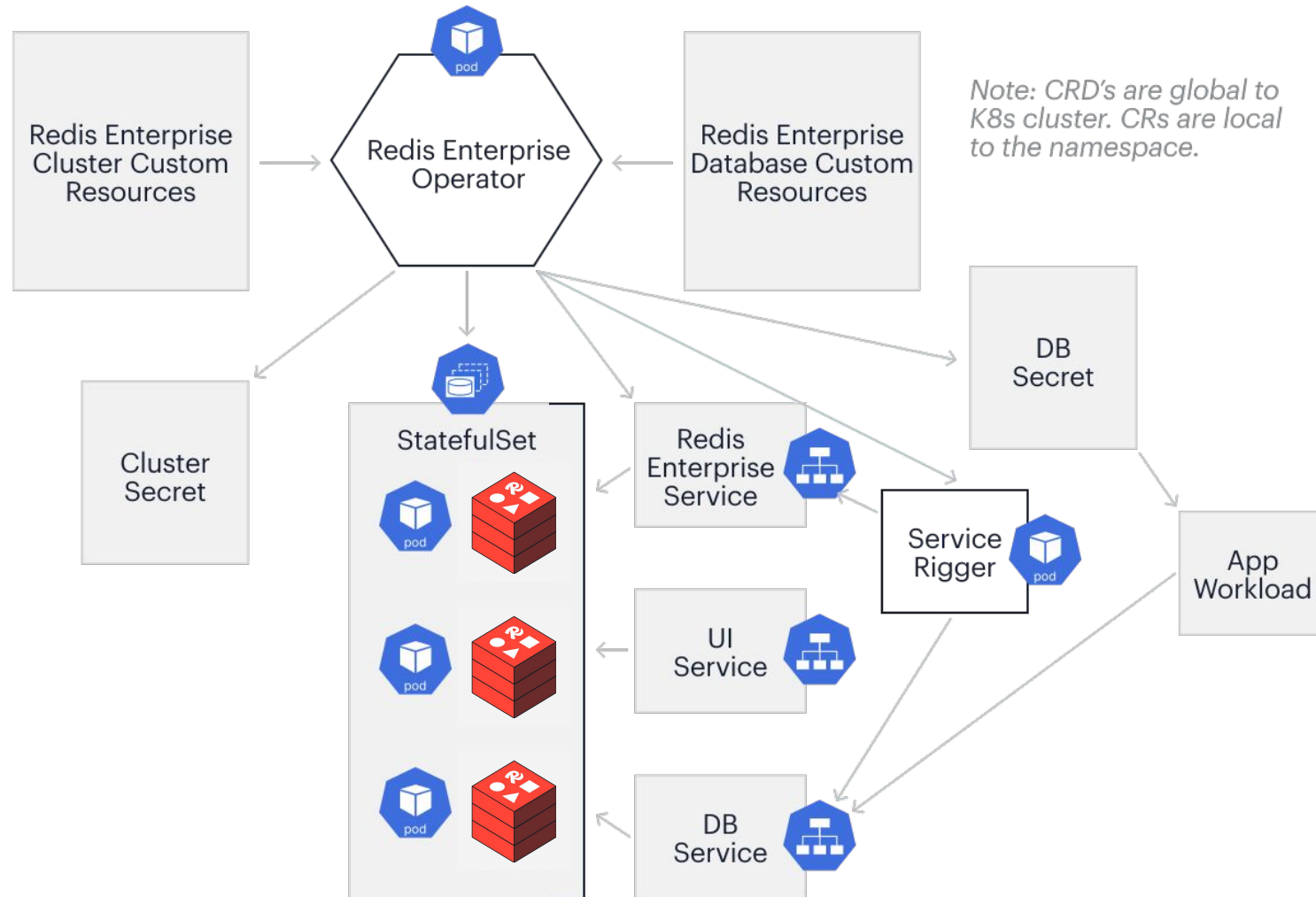
Infrastructure



Infrastructure



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

7.4

Runs on

OpenShift 4.6 - 4.18

Delivery method

Operator

[Purchase](#)

[Free trial](#)

[Contact sales](#)

Highlights

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

[Overview](#)

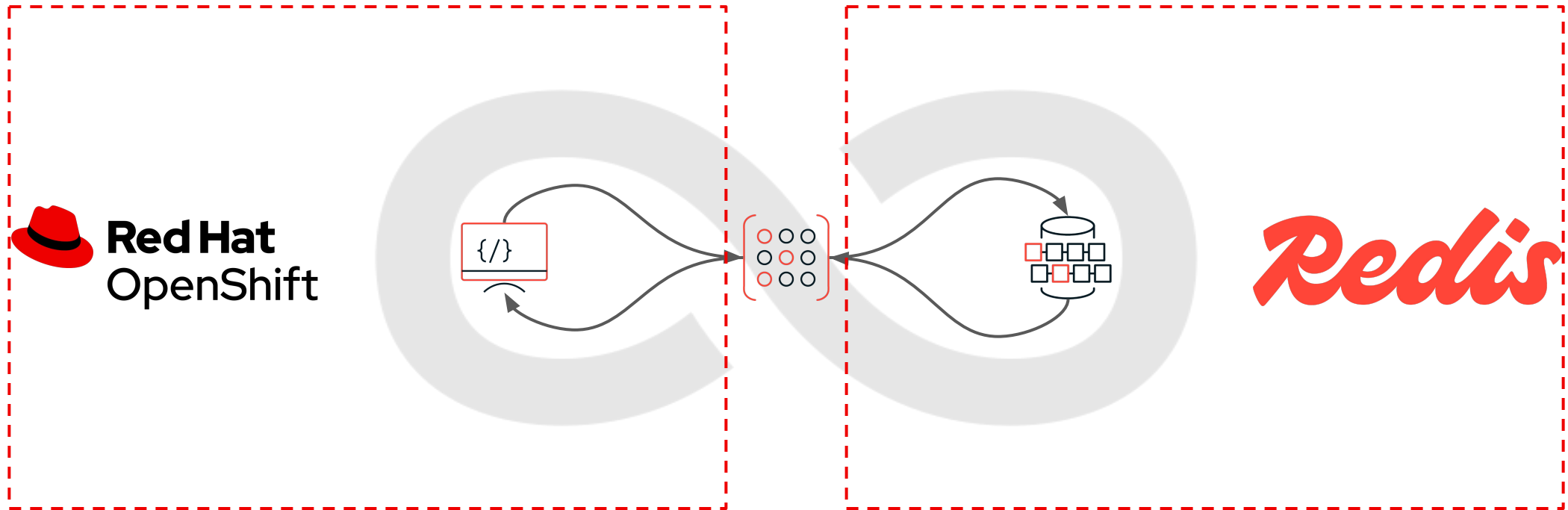
[Documentation](#)

[Pricing](#)

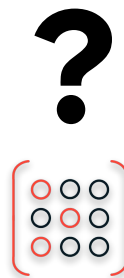
[Help](#)

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.

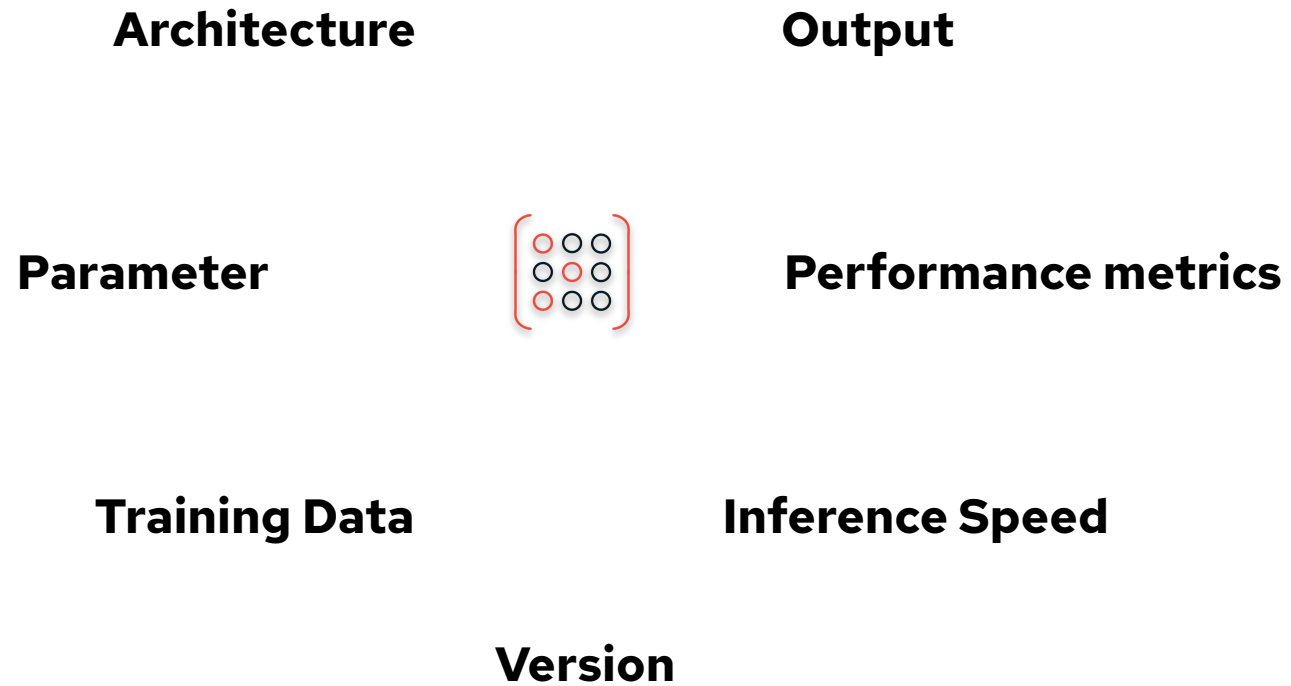
Retrieval Augmented Generation



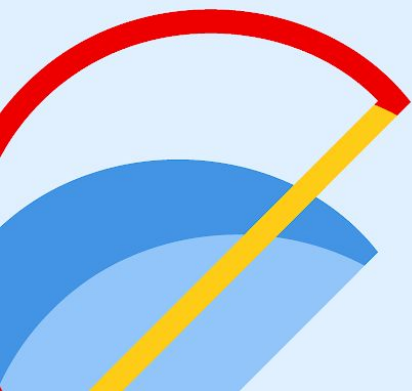
Retrieval Augmented Generation



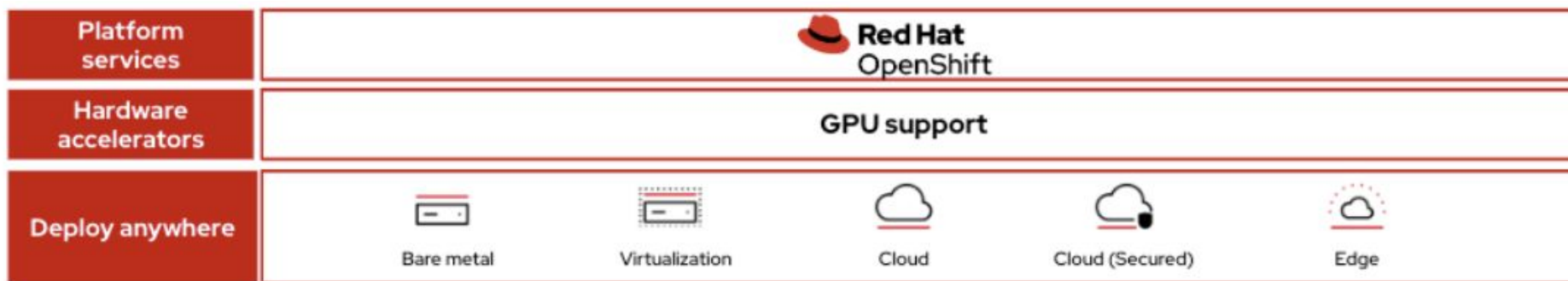
Retrieval Augmented Generation



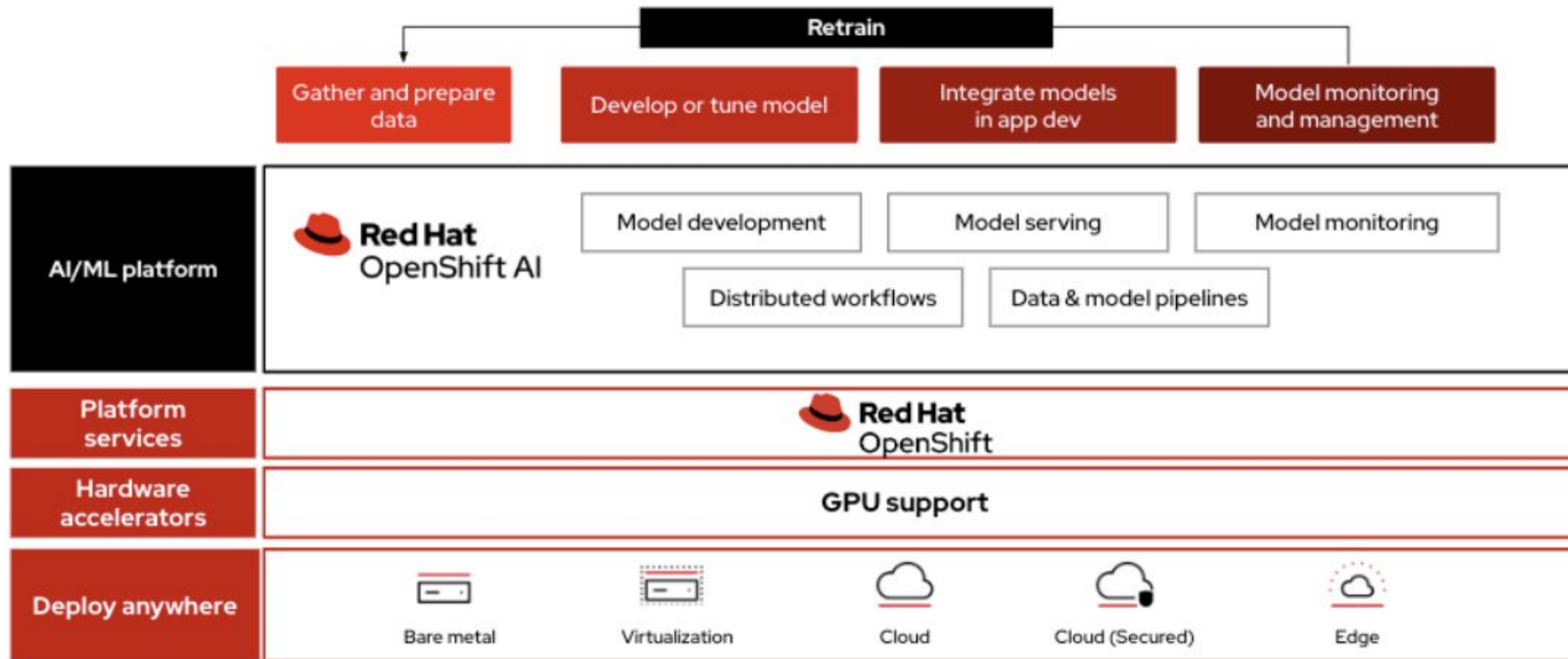
OpenShift AI



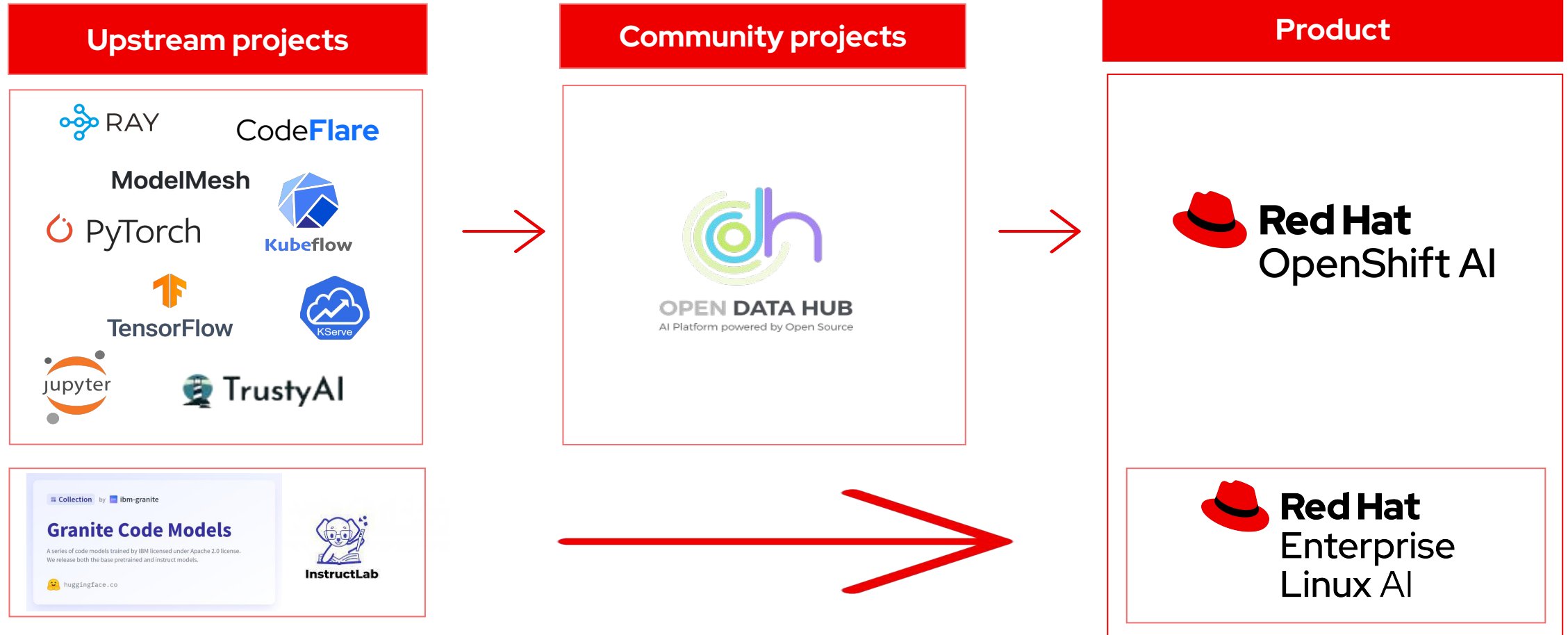
OpenShift



OpenShift AI



OpenShift AI





Integrated AI platform

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



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.



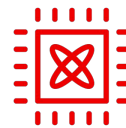
Model serving and monitoring

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.



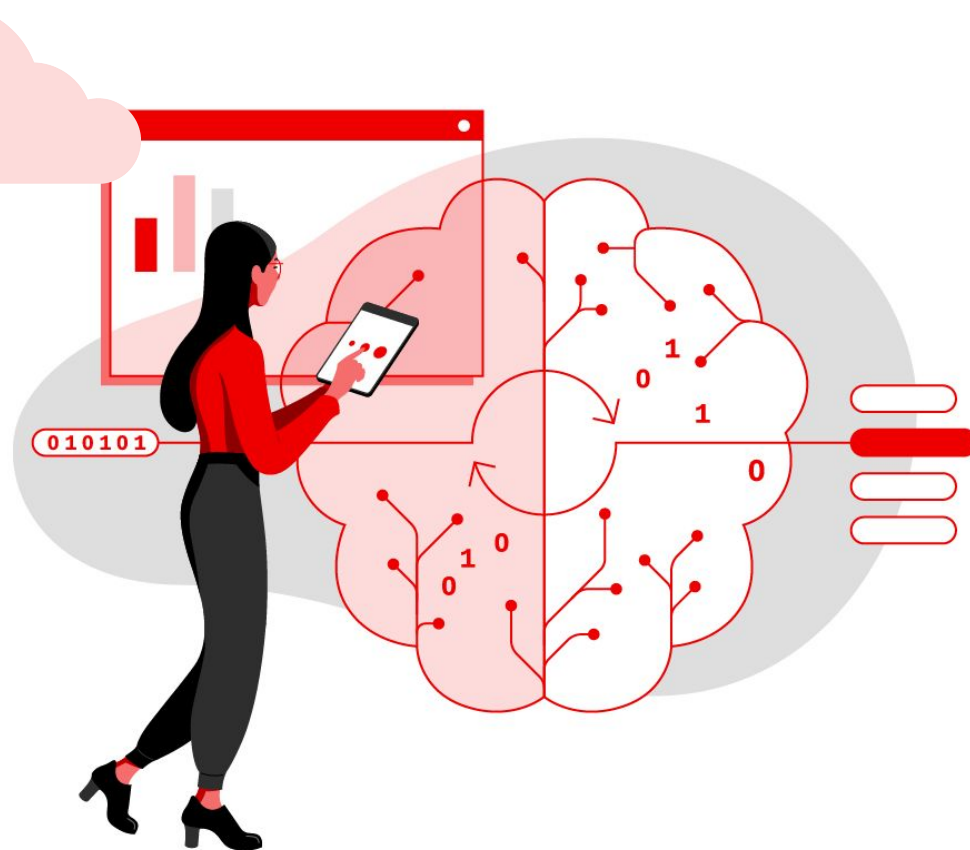
Resource optimization and management

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

Available as

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

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

Serve, scale, and monitor your models

Select the required resources and scale model serving as needed

Make your model public and secure

Configure model server

Model server replicas

Number of model server replicas to deploy

- 1 +

Compute resources per replica

Model server size

Small

Model route

Make deployed available via an external route

Token authorization

Require token authentication

Deploy model

Configure properties for deploying your model

Project

modelserving-test

Name *

myModel

Model framework

onnx - 1

Model location

Existing data connection

Name

storage-config

Folder path

onnx/road_conditions.onnx

New data connection

Select your model framework

Models and model servers

Type	Deployed models	Tokens
ovms	1	Tokens disabled

Model name	Inference endpoint	Status
myModel	https://mymodel-modelserving-test.apps.pilot.j61u.p1.openshiftapps.com/v2/models/mymodel/infer	

View your deployed model fleet endpoints

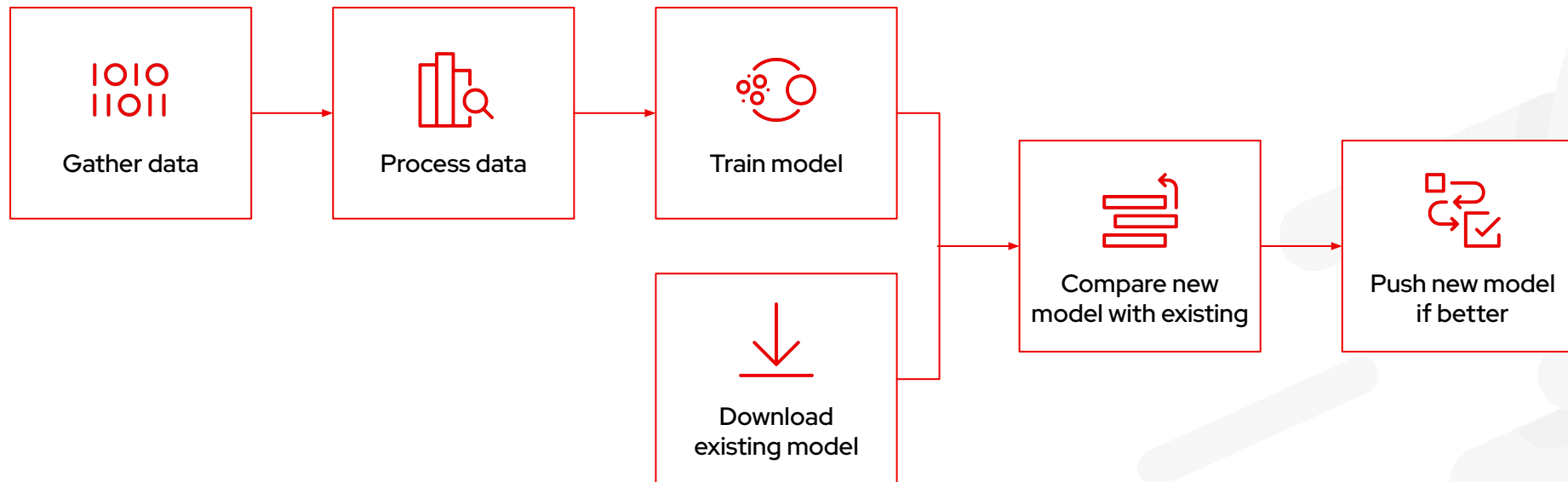
Model performance metrics

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



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

Train a new model Running Actions

```
graph LR; process_...[process_...] --> model_gra...[model_gra...]; process_... --> model_ran...[model_ran...]; model_gra... --> compare_a...[compare_a...]; model_ran... --> compare_a...;
```

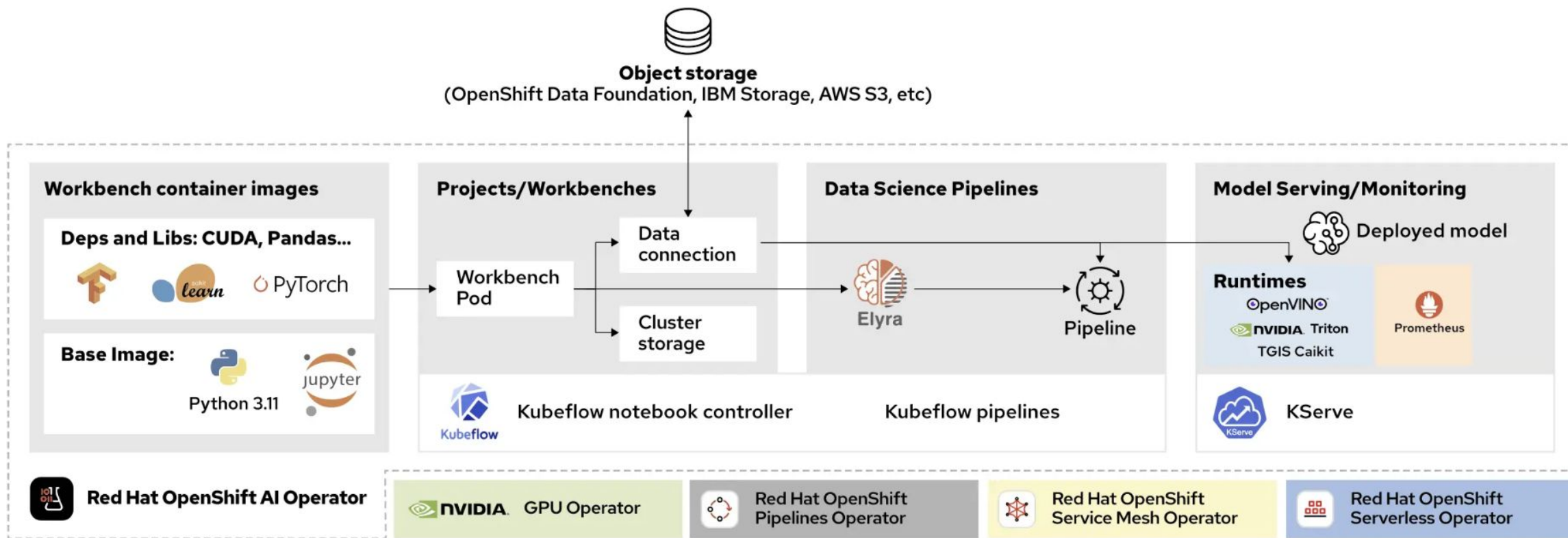
Details | Run output

Name	Train a new model
Pipeline	train_new_model1
Project	Robert Serving Test
Run ID	eca2addc-f601-49b3-8ecd-6534114906e7
Workflow name	train-new-model1-eca2a

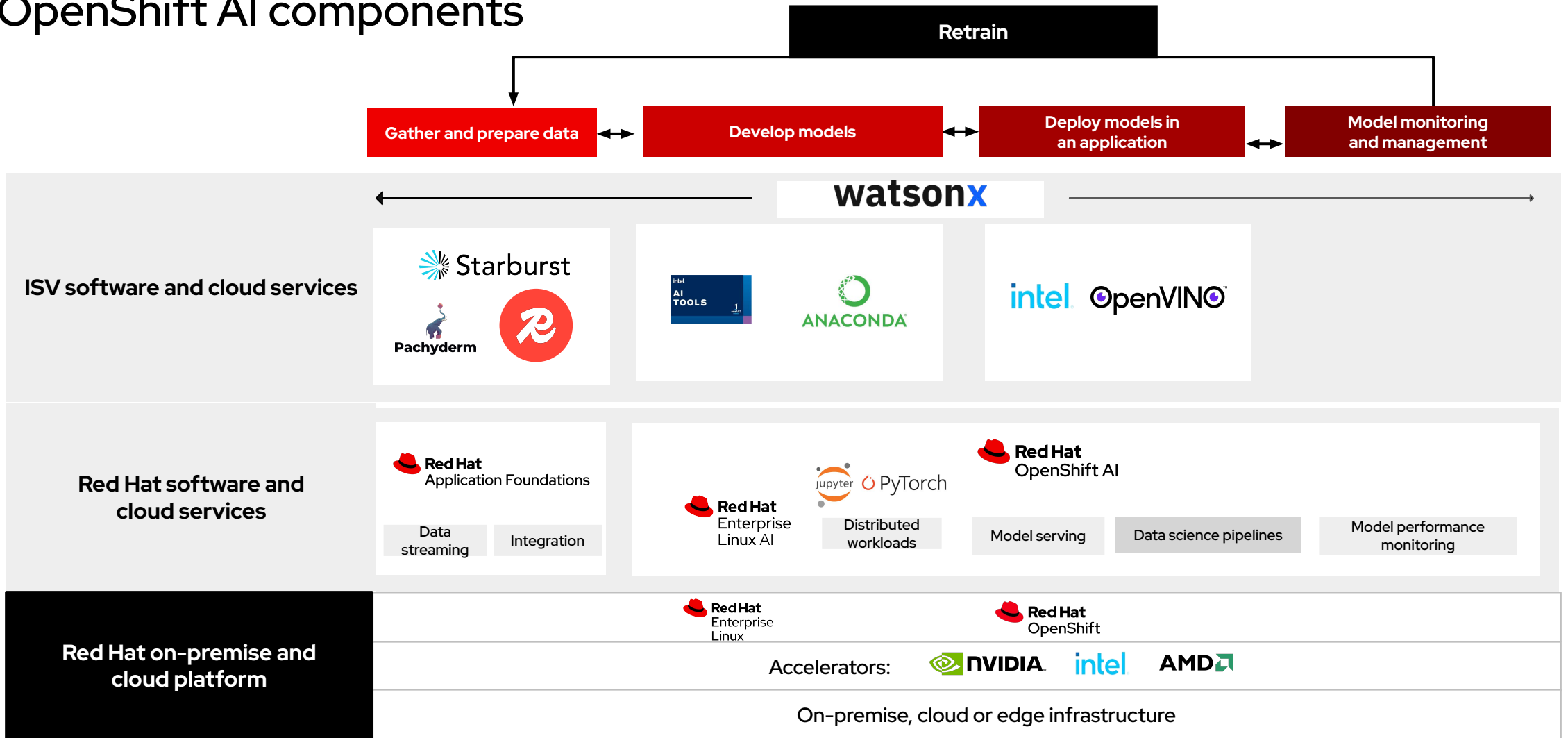


The OpenShift AI user interface enables you to track and manage pipelines and pipeline runs.

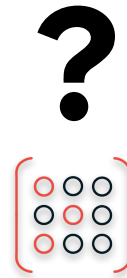
Red Hat OpenShift AI



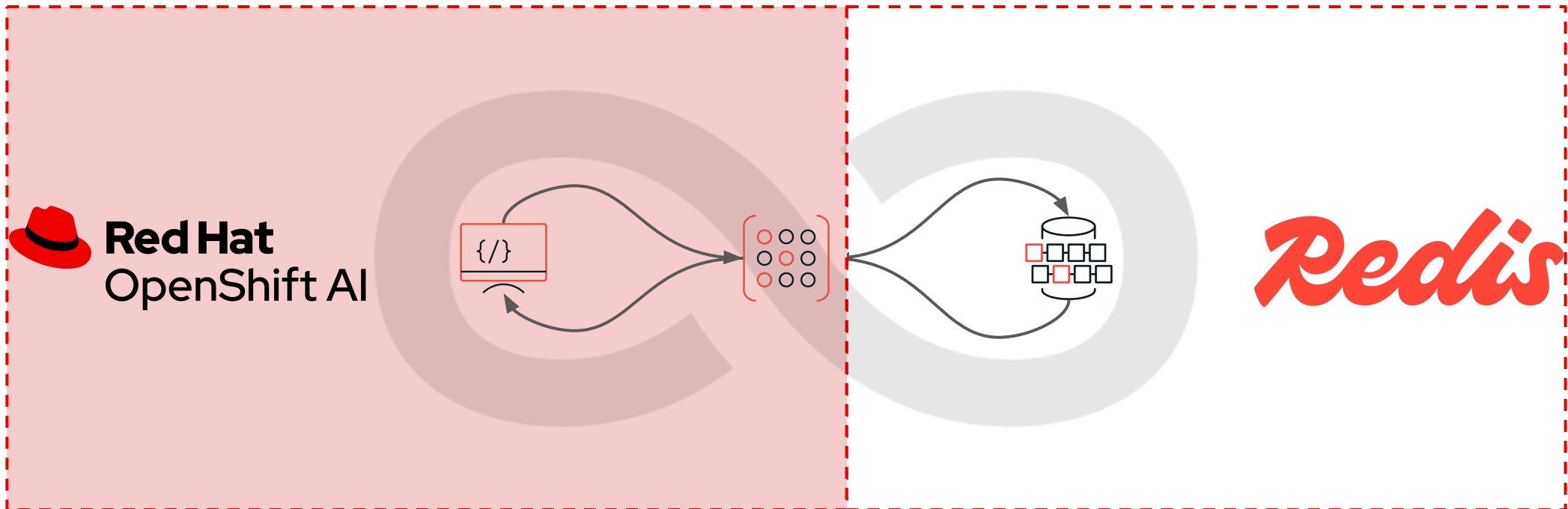
OpenShift AI components



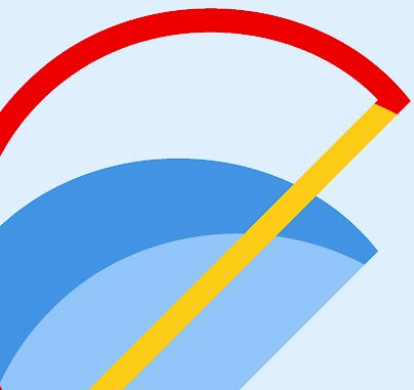
Retrieval Augmented Generation



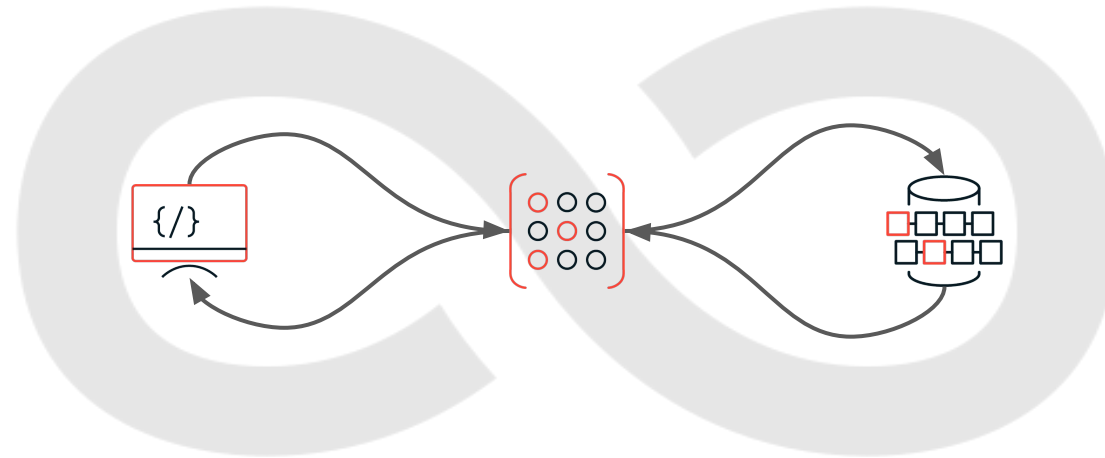
Retrieval Augmented Generation



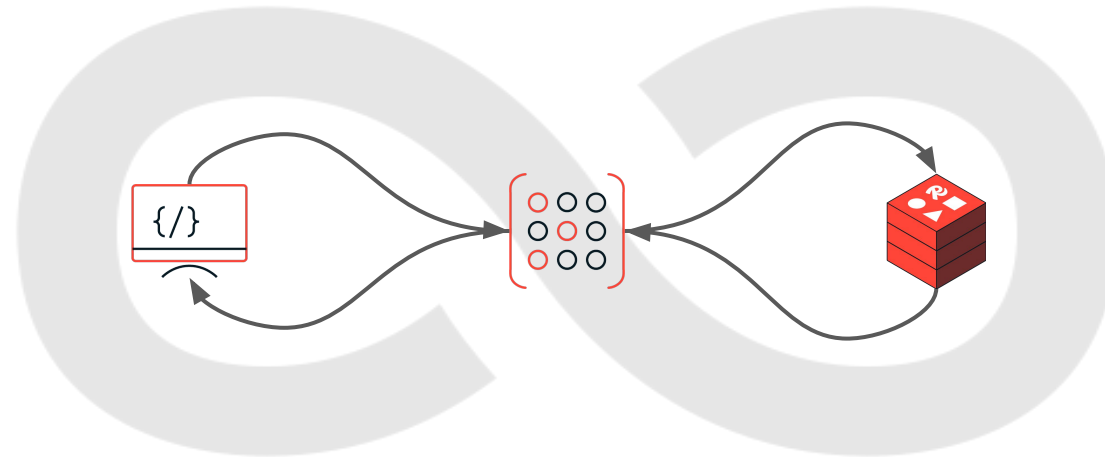
Redis Enterprise for AI



Retrieval Augmented Generation

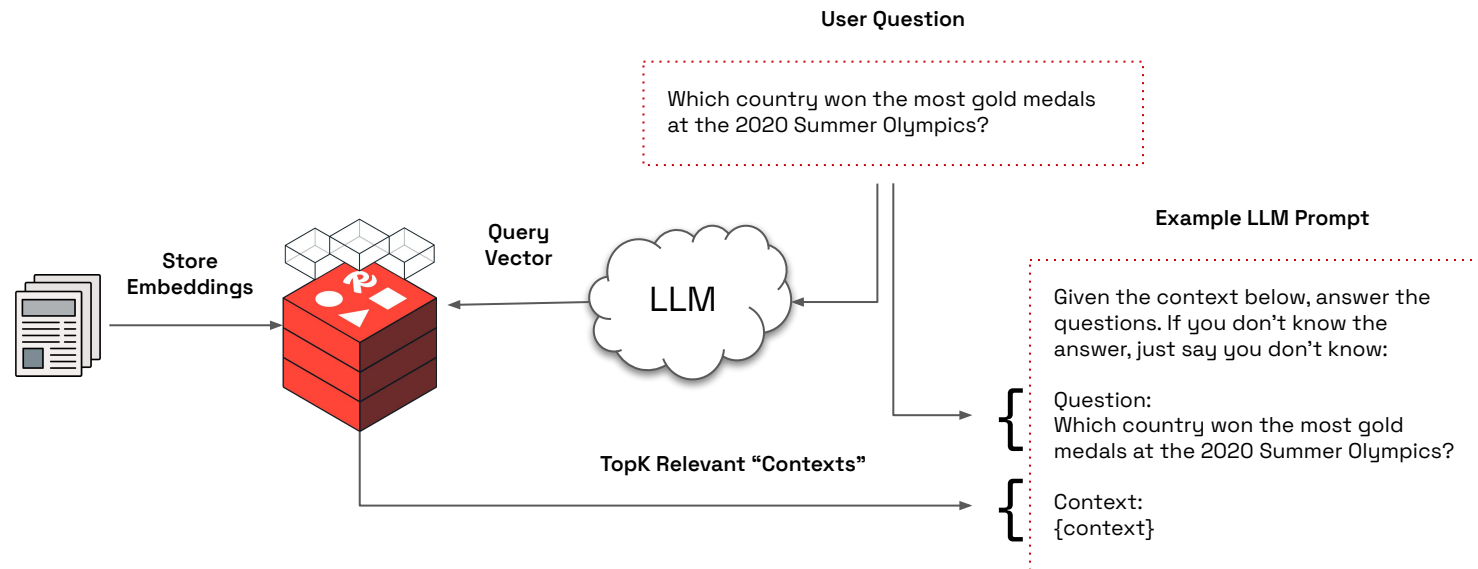


Retrieval Augmented Generation

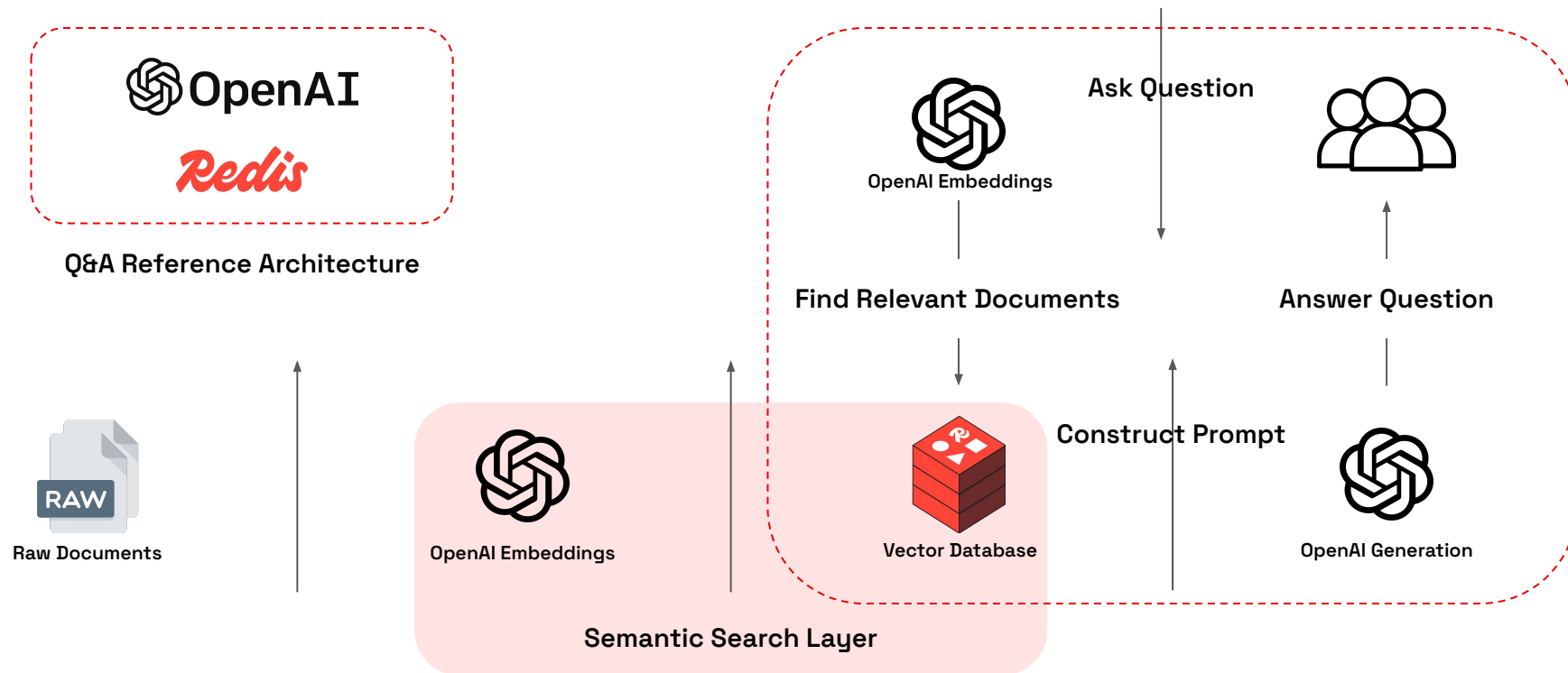


Retrieval Augmented Generation

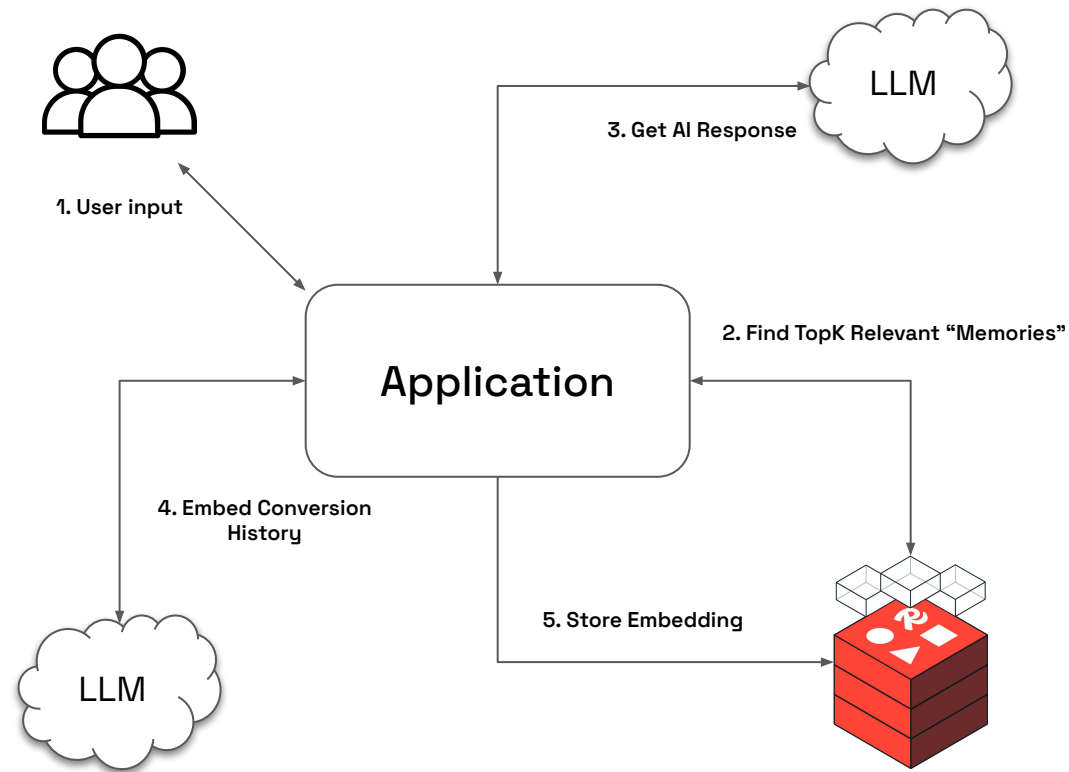
Context Retrieval



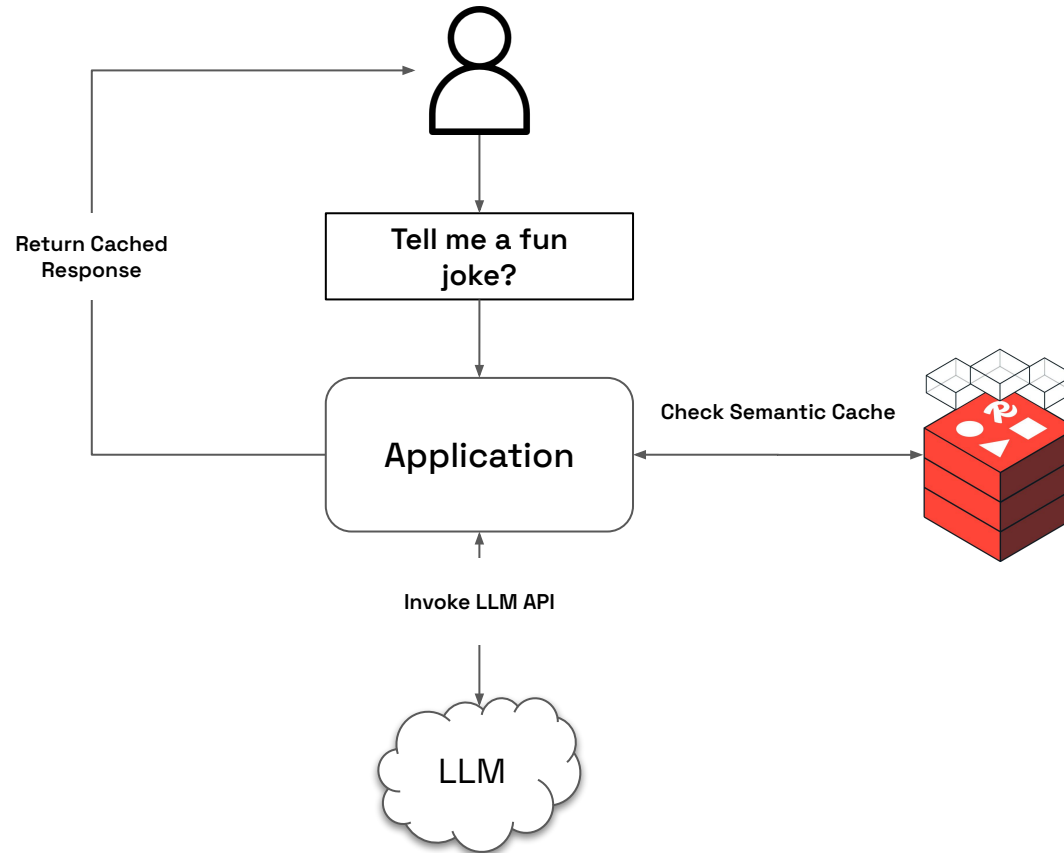
Retrieval Augmented Generation Question & Answering



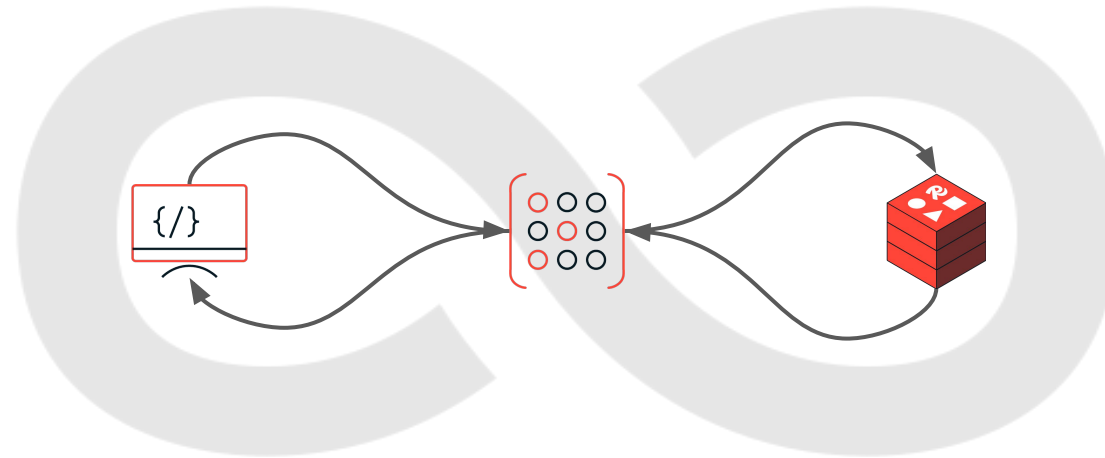
Retrieval Augmented Generation LLM Conversation Memory



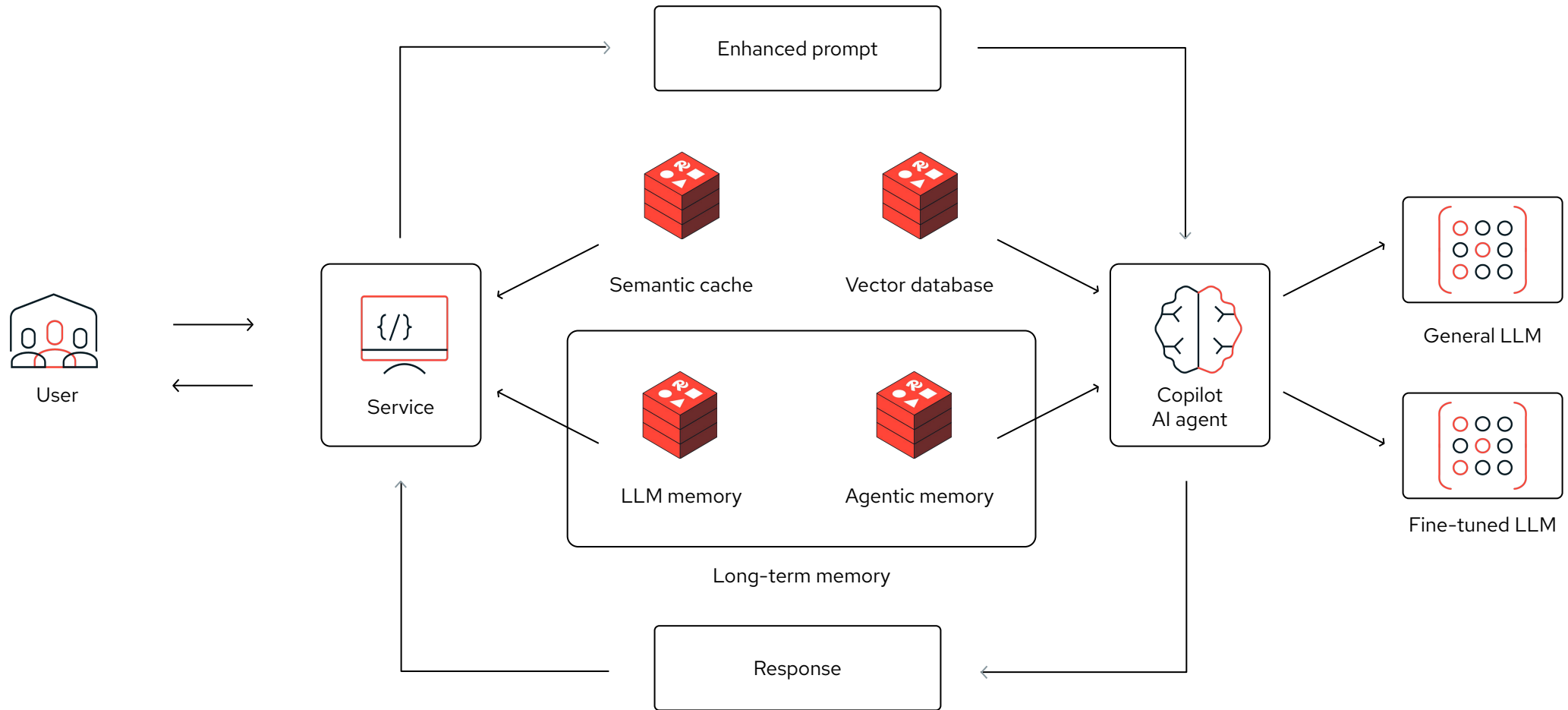
Retrieval Augmented Generation Semantic Caching



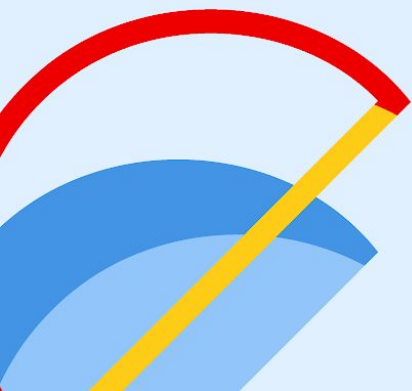
Retrieval Augmented Generation



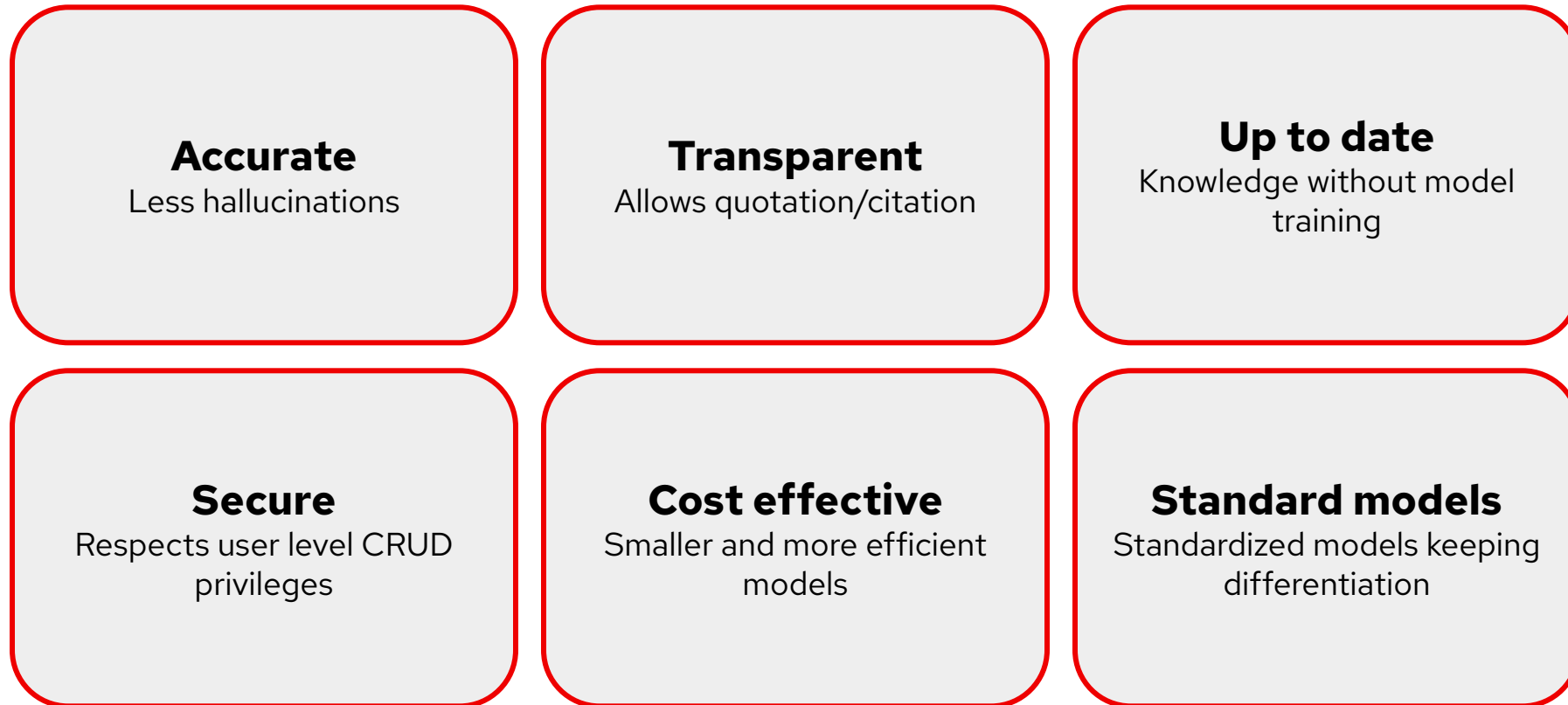
Retrieval Augmented Generation



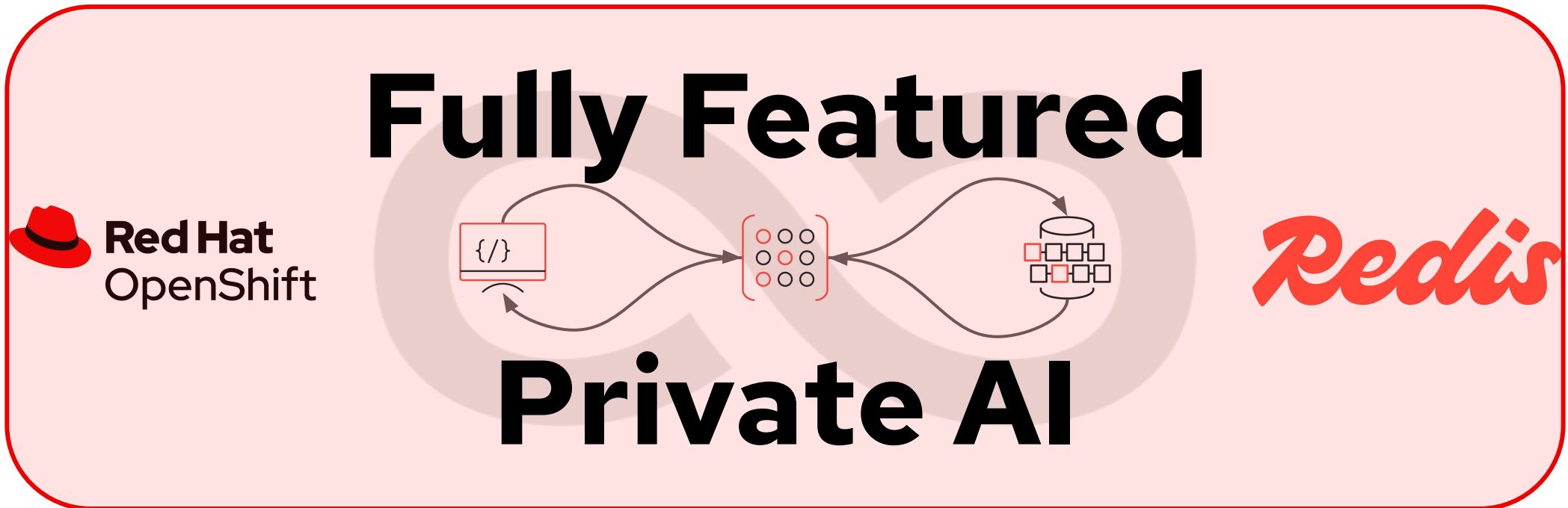
Advantages



Retrieval Augmented Generation



Retrieval Augmented Generation



Red Hat
Summit

Connect

Q&A

Redis

 Red Hat

Red Hat
Summit

Connect

Thank you

Redis

 Red Hat