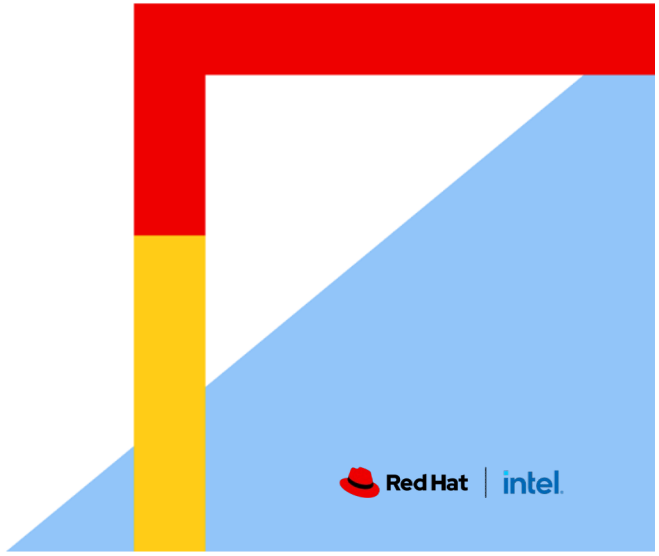
OPEA-based Retrieval Augmented Generation (RAG) on Intel® Gaudi with OpenShift AI



Red Hat + intel

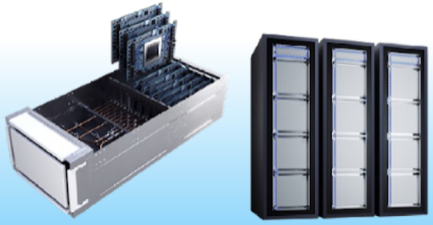
Over **25** Years of Collaboration

Bringing AI Everywhere

Intel's AI Strategy

AI PC Node
AI Developer Productivity & Light Inference

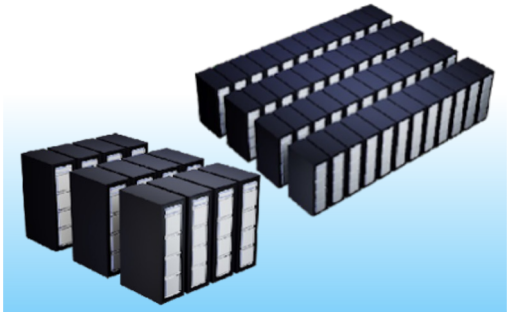
AI PC
Broadest AI SW Ecosystem



Node
Fine-tuning, Inference

Cluster
Light Training, Tuning, Peak Inference

ENTERPRISE AI & EDGE AI
Open Standard, "Ready to Use"

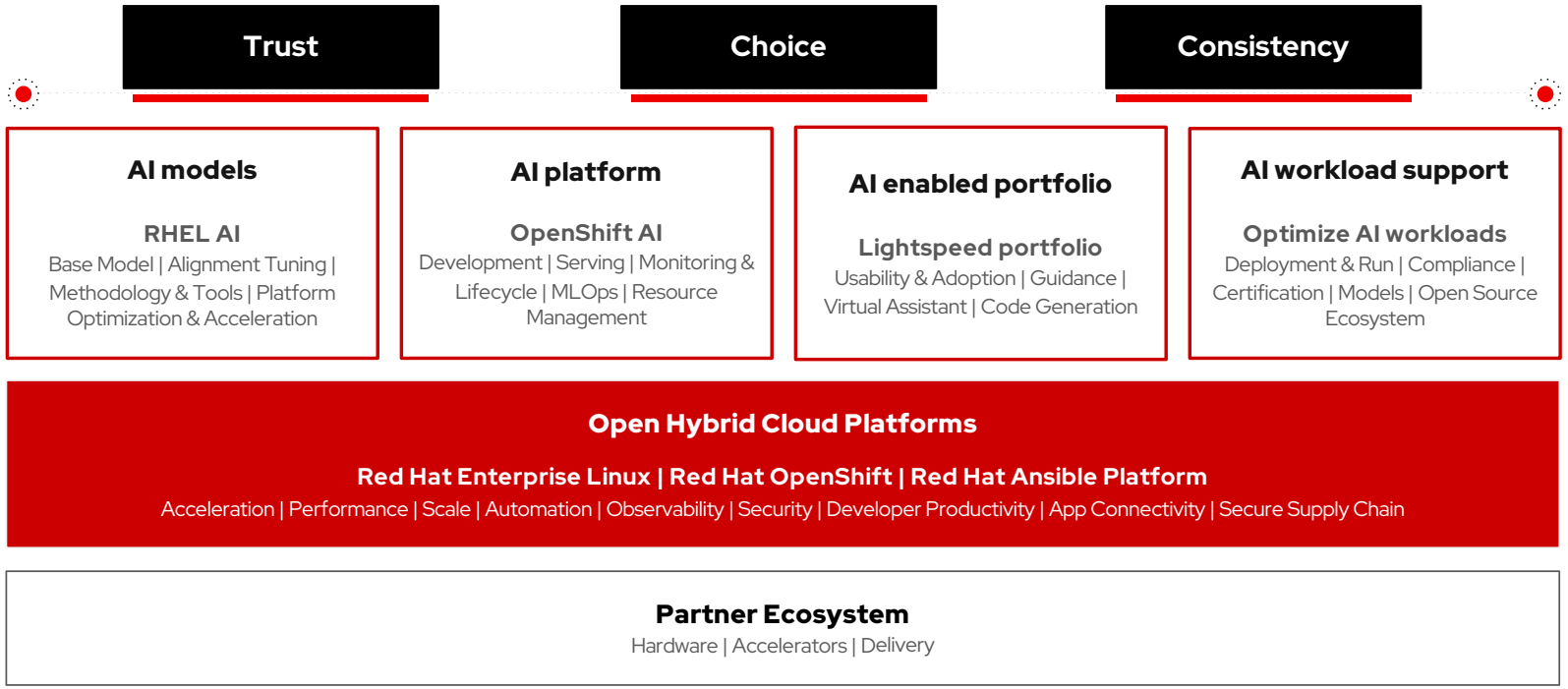


Super Cluster
Training, Tuning, Peak Inference

Mega Cluster
Large Scale Training & Inference

DATA CENTER AI
AI Open, Scalable Systems & Reference Arch

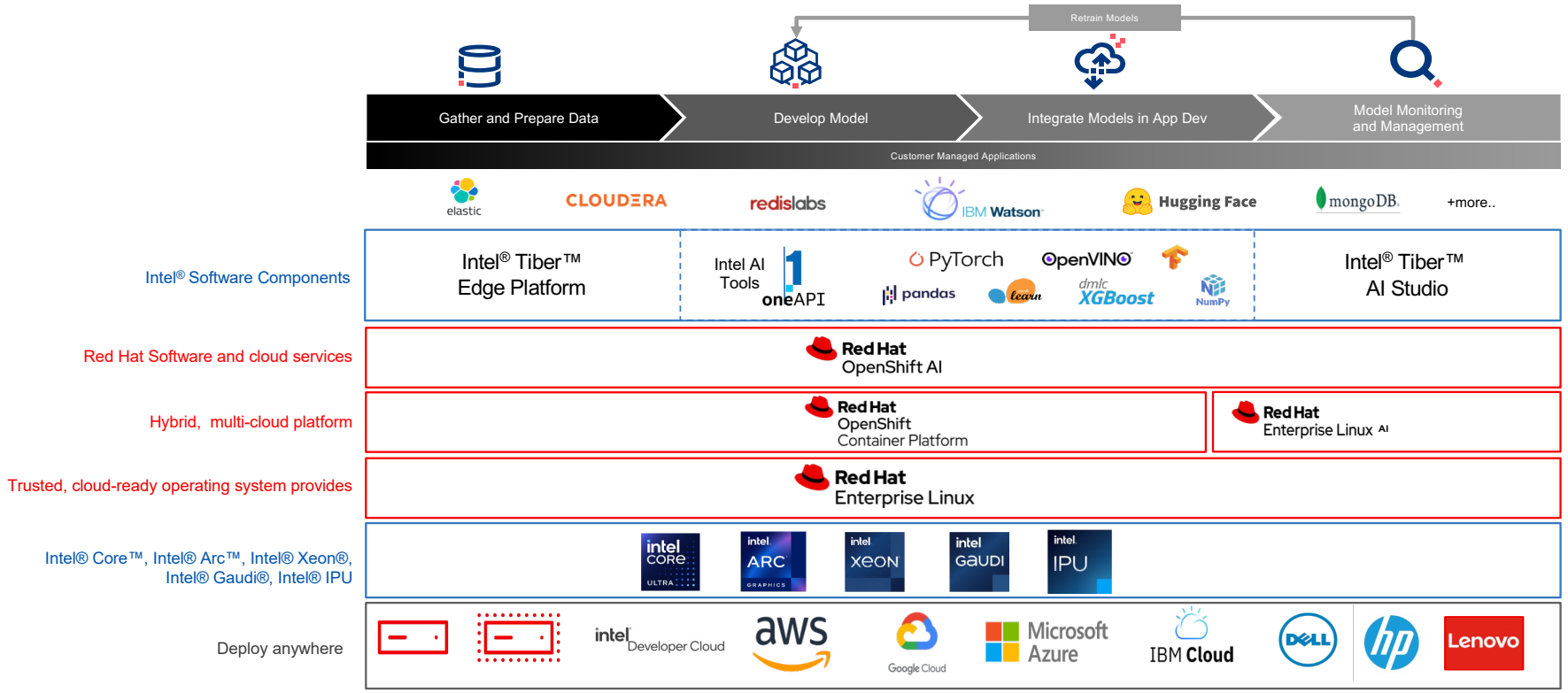
Red Hat's AI Strategy



Red Hat | intel.

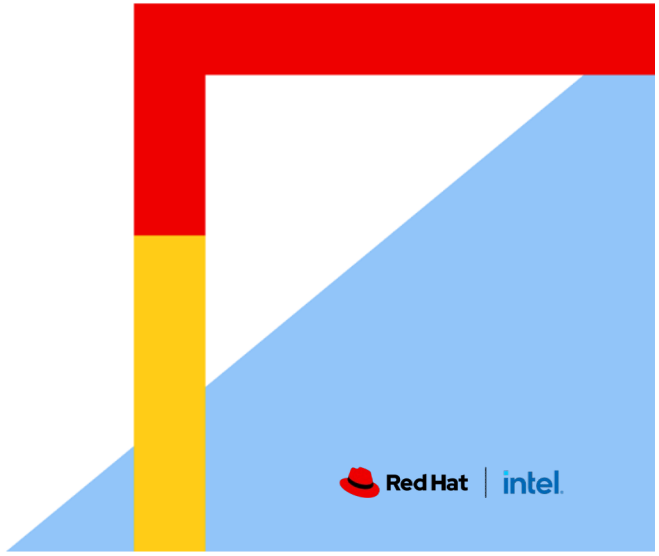
Joint Ecosystem for AI Transformation

Intel Enterprise AI with Red Hat® OpenShift® AI



Red Hat | intel.

OPEA – Open Platform for Enterprise AI



Open Platform for Enterprise AI

OPEA - Open Platform for Enterprise AI

By The Linux Foundation

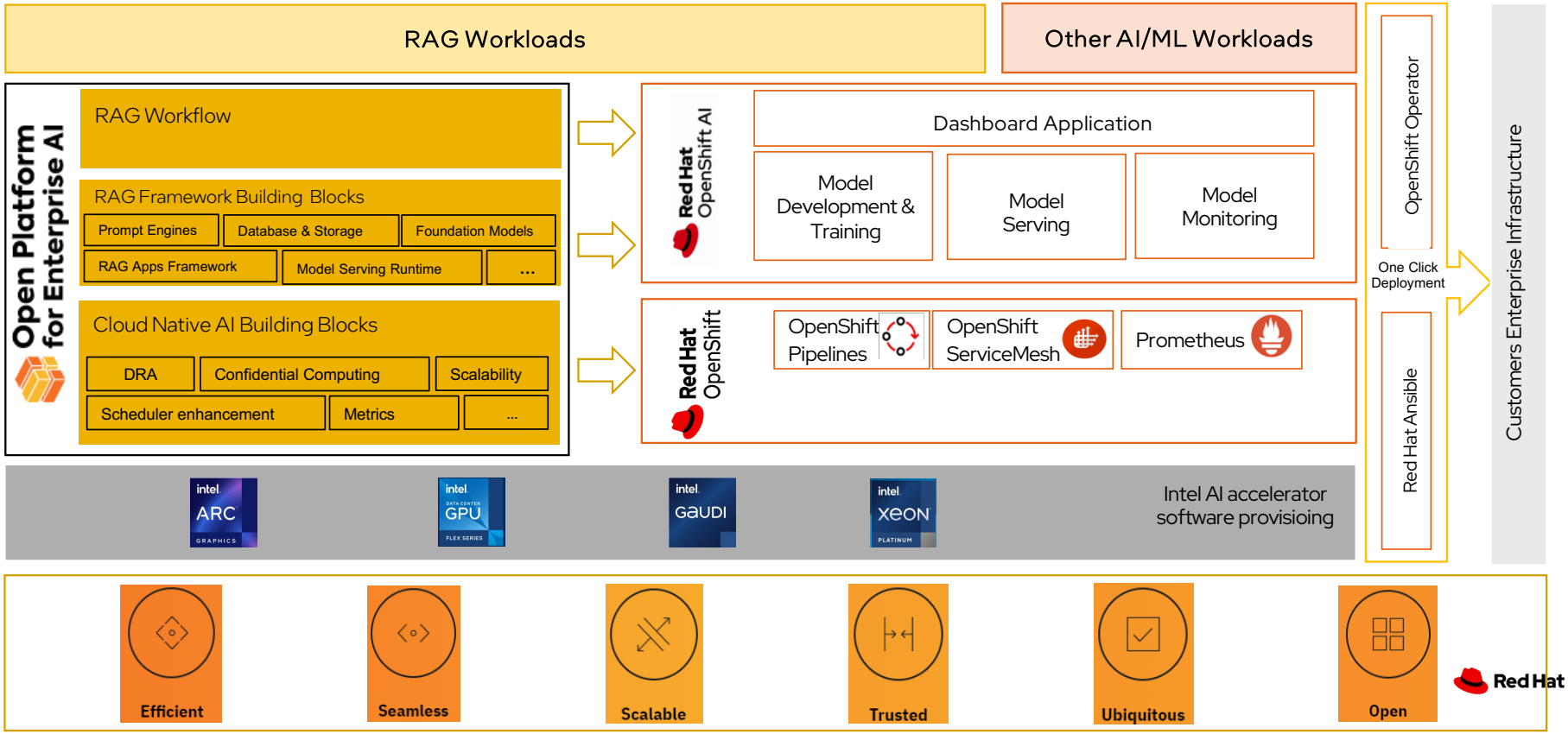
- Ecosystem orchestration framework for GenAI
- The OPEA platform includes:
  - Detailed frameworks of composable building blocks
  - Architectural blueprints of retrieval-augmented generative AI stacks
  - End-to-end workflows
  - Assessment tools
- GitHub: https://github.com/opea-project
- Contributors:

Red Hat | intel.

Open Platform for Enterprise AI

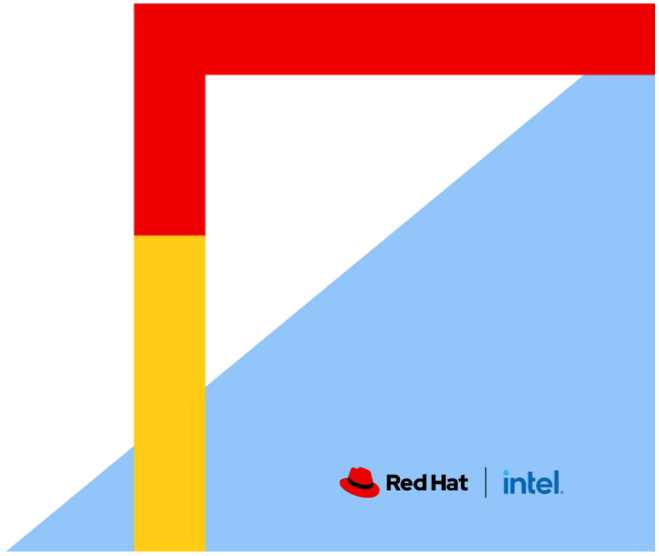
OPEA with OpenShift AI

OpenShift AI makes OPEA more enterprise ready



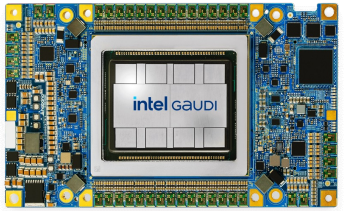
Red Hat | intel.

Intel Gaudi
AI Accelerators



Introducing the Intel® Gaudi® 3 Accelerator

Breaking benchmarks, not budgets



Competitive Gen AI Performance over H100

- Projected **50% faster time to train**[1]
- Projected **50% faster inferencing**[2]
- Projected **40% better power efficiency**[3]

Freedom to Scale without Lock-in

- Open standard ethernet networking vs proprietary InfiniBand
- 24x200 GbE ports of industry-standard RoCE on every Gaudi® [3]
- 33% more I/O peak throughput vs H100 for massive scale-up within the server[4]

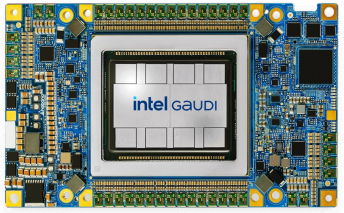
Open Development on GenAI platforms

- Integrated open-source PyTorch framework with optimized model library on Hugging Face
- Migrate models on open software from H100 with as few as 3 lines of code

[1] NV H100 comparison based on : https://developer.nvidia.com/deep-learning-performance-training-inference/training, Mar 28th 2024 -> "Large Language Model" tab.
[2] Source: NV H100 comparison based on https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 , Mar 28th, 2024. Reported numbers are per GPU.
[3] Source: NV comparison based on https://nvidia.github.io/TensorRT-LLM/performance.html#h100-gpus-fp8 , Mar 28th, 2024. Reported numbers are per GPU.
[1-3] Vs Intel® Gaudi® 3 projections for LLAMA2-7B, LLAMA2-70B & Falcon 180B Power efficiency for both Nvidia and Gaudi3 based on internal estimates. Results may vary.
[4] 900 GB/s NVLink connectivity on H100 vs. 1200 GB/s on Gaudi 3

Red Hat | intel

Intel Gaudi AI Accelerators



Broad Application Support with Focus on Multi-Modal, LLM and RAG

AI Applications

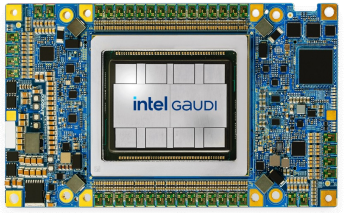
AI Functions

3D Generation	Text Generation	Classification
Video Generation	Sentiment	Translation
Image Generation	Summarization	Q&A

Core Capabilities

- Multi-modal Models
- LLM
- RAG

Intel® Gaudi® 3 AI Accelerator



Launch Partners

DELL Technologies

Lenovo

Hewlett Packard Enterprise

SUPERMICRO

ASUS

Inventec

QCT Quanta CLOUD TECHNOLOGY

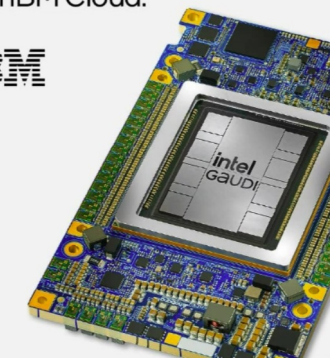
GIGABYTE TECHNOLOGY

ingrasys

wistron

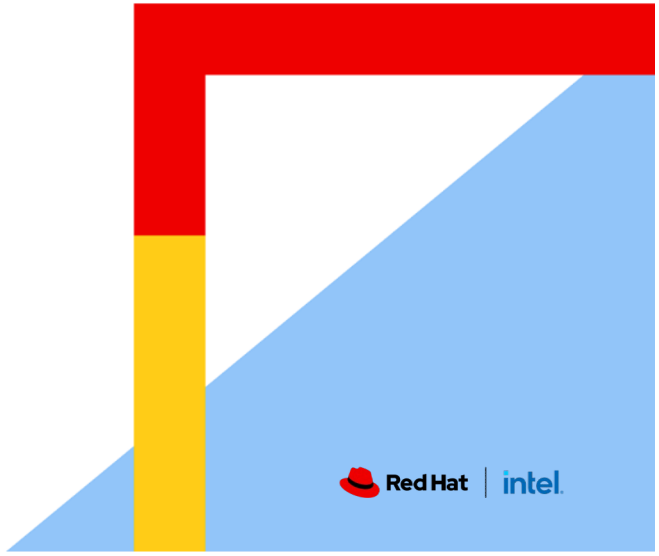
IBM and Intel announce a global collaboration to integrate Intel® Gaudi® 3 accelerators with watsonx on IBM Cloud.

intel | IBM



Red Hat | intel.

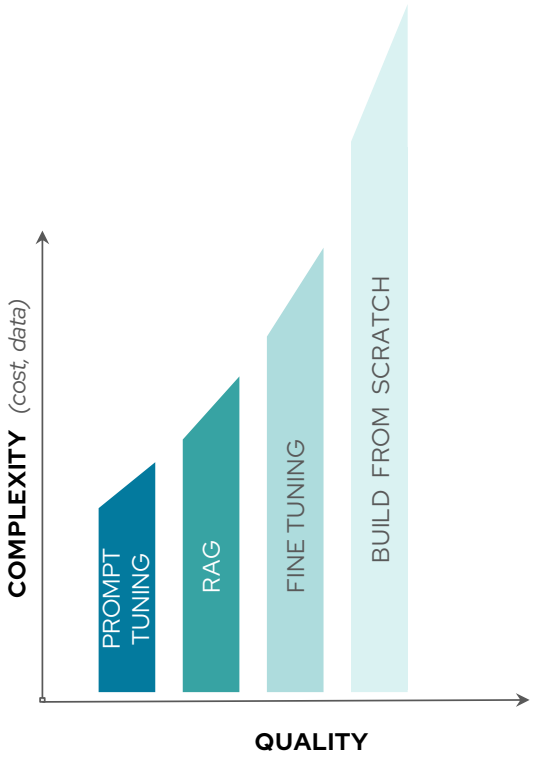
Retrieval Augmented Generation (RAG) Explained



The balancing act of using foundation models

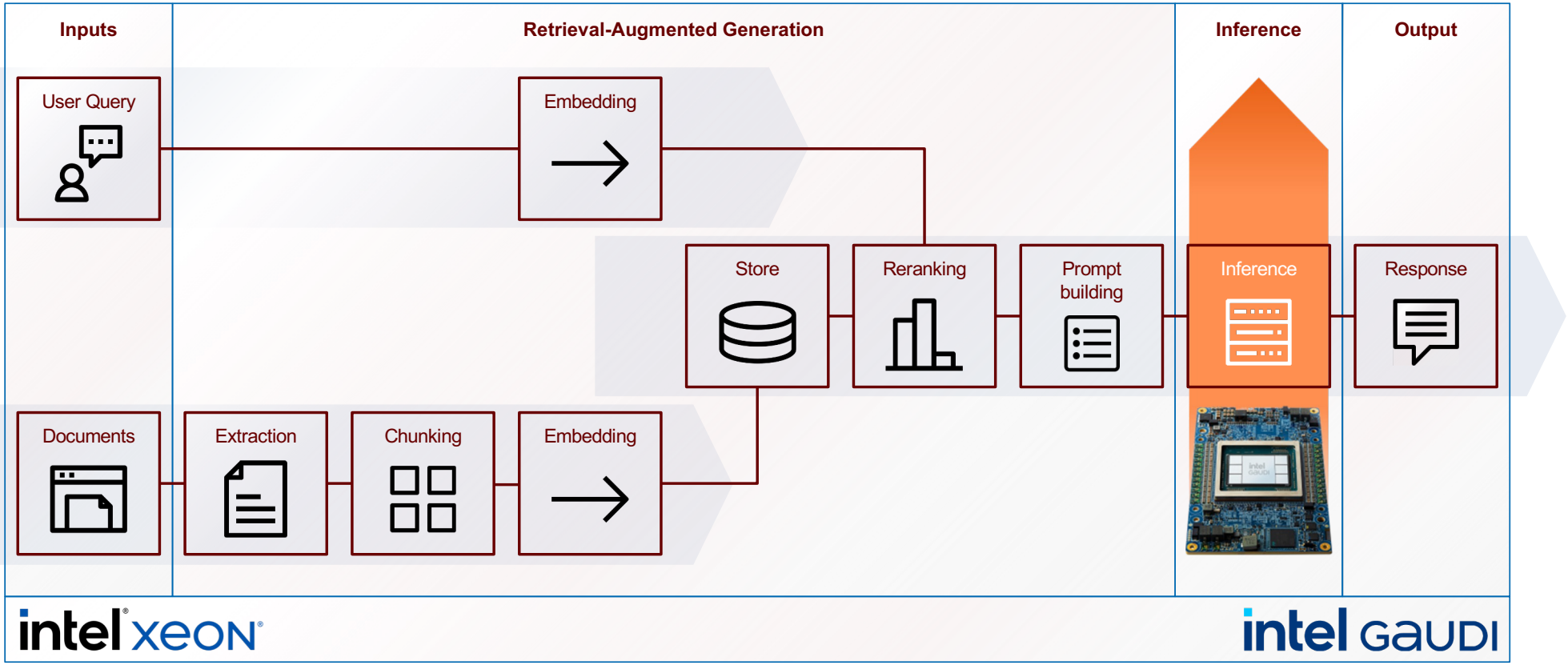
Foundation models will still need more work to be useful

- **Prompt tuning** allows to adapt models offering 'good enough' accuracy but doing it with less resources
- **Retrieval augmented generation (RAG)** allows training models with targeted information without modifying the underlying model itself
- **Fine tuning foundation models** requires a high amount of resources (data, hardware, people)
- **Training a Foundation Model from scratch** requires un-realistic amount of computing



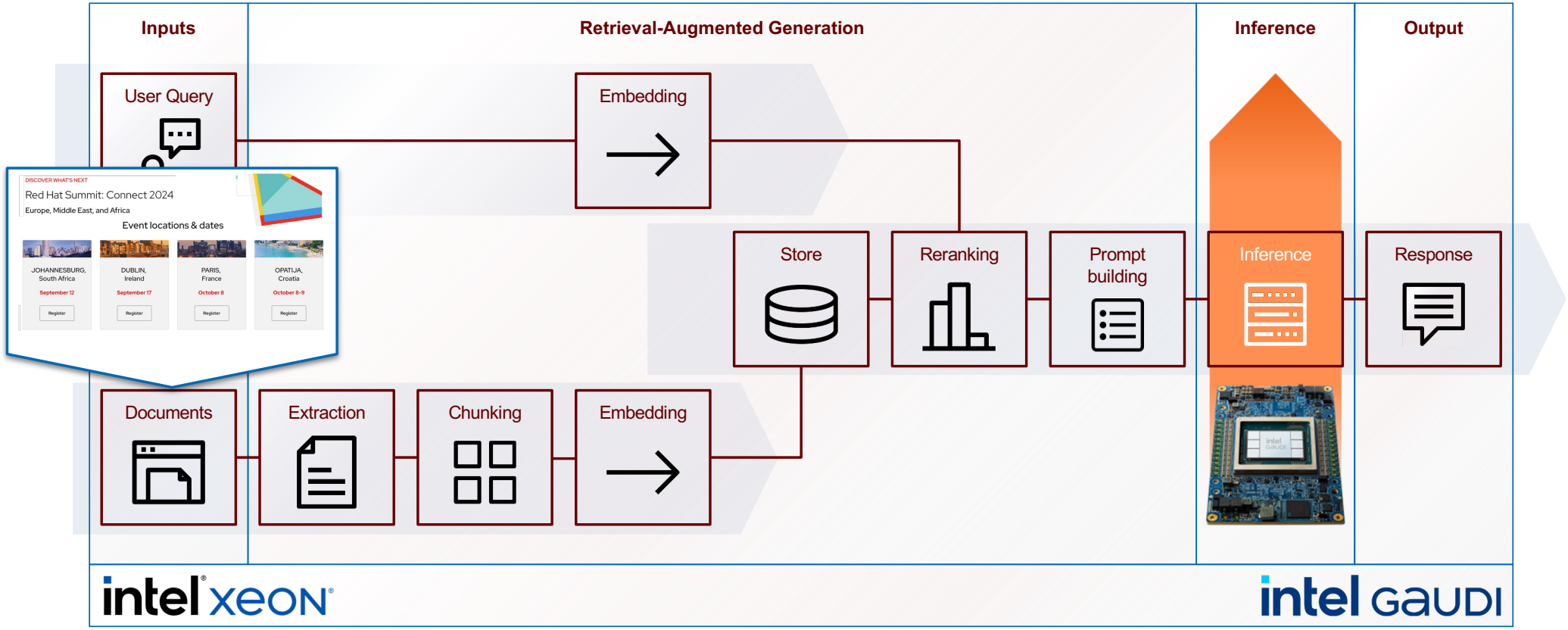
Red Hat | intel.

Retrieval Augmented Generation (RAG)



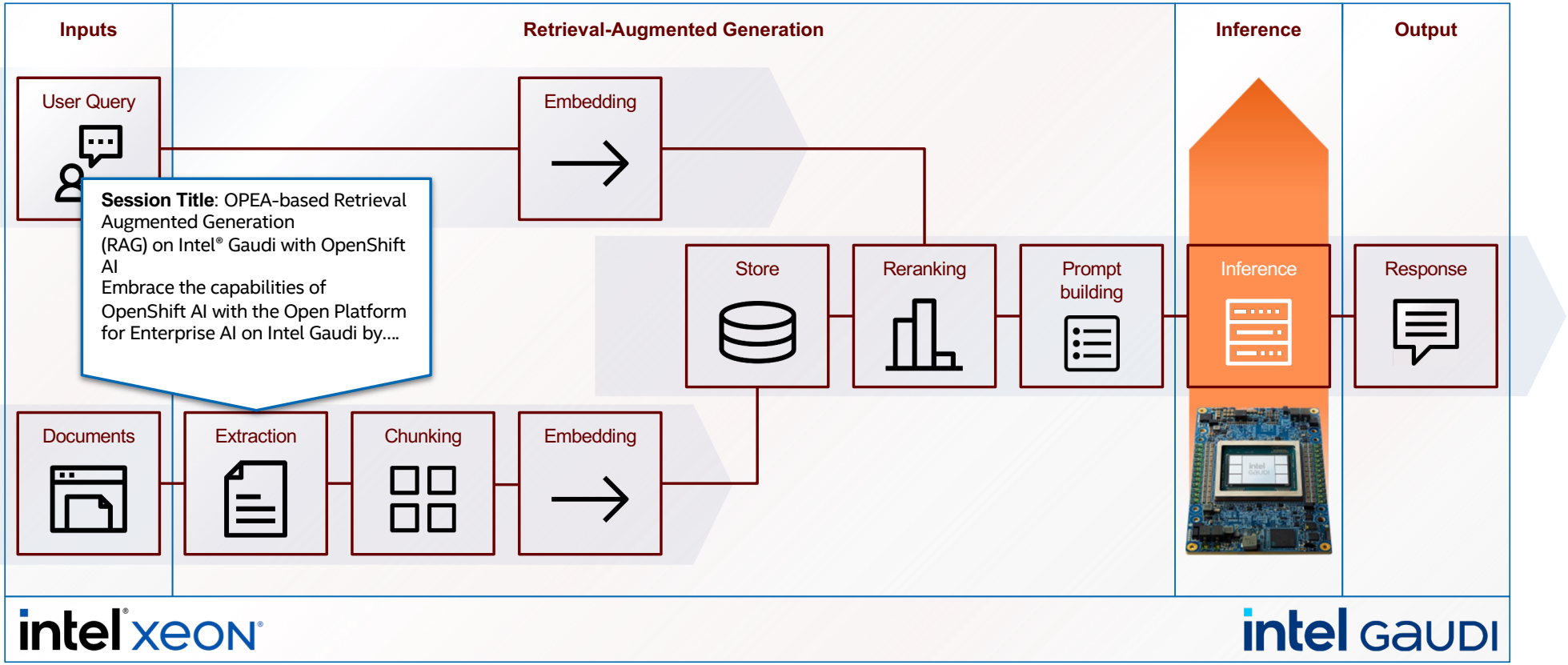
Red Hat | intel

Retrieval Augmented Generation (RAG)



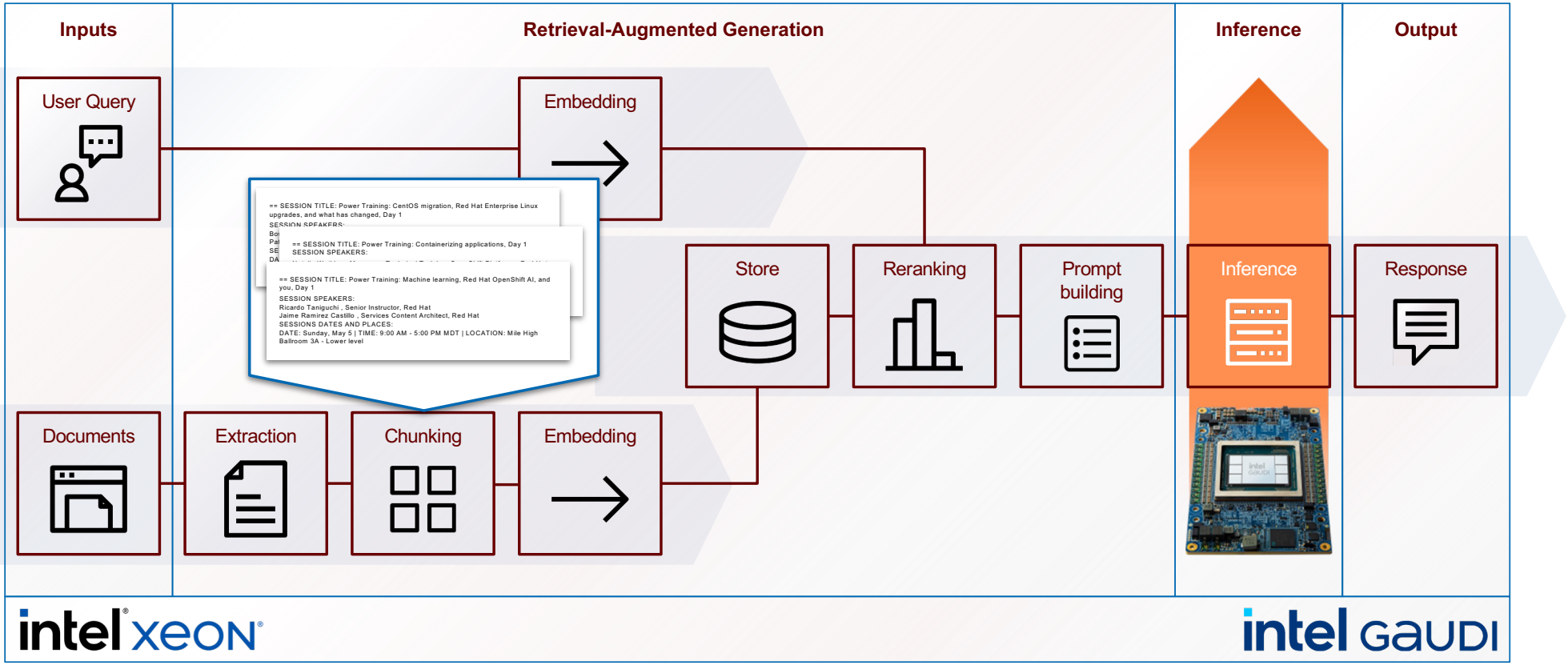
Red Hat | intel

Retrieval Augmented Generation (RAG)



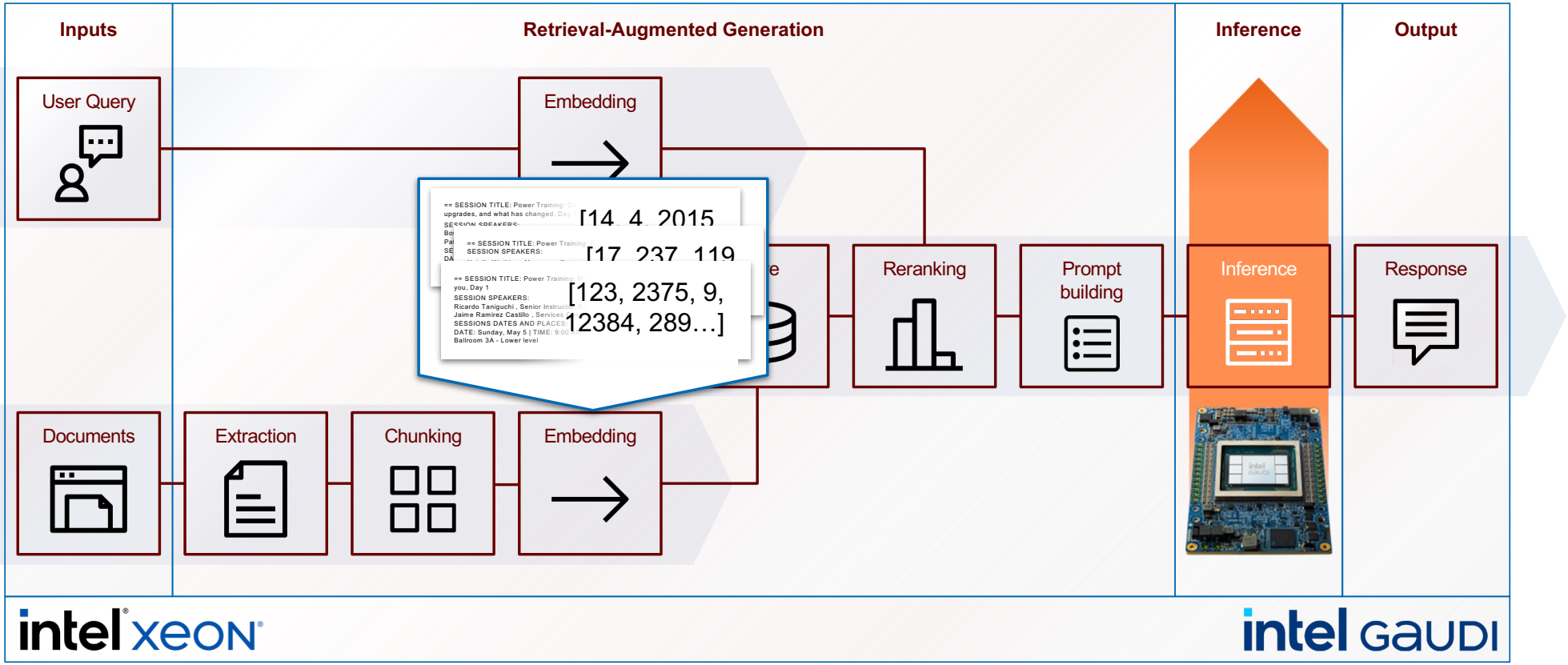
Red Hat | intel

Retrieval Augmented Generation (RAG)



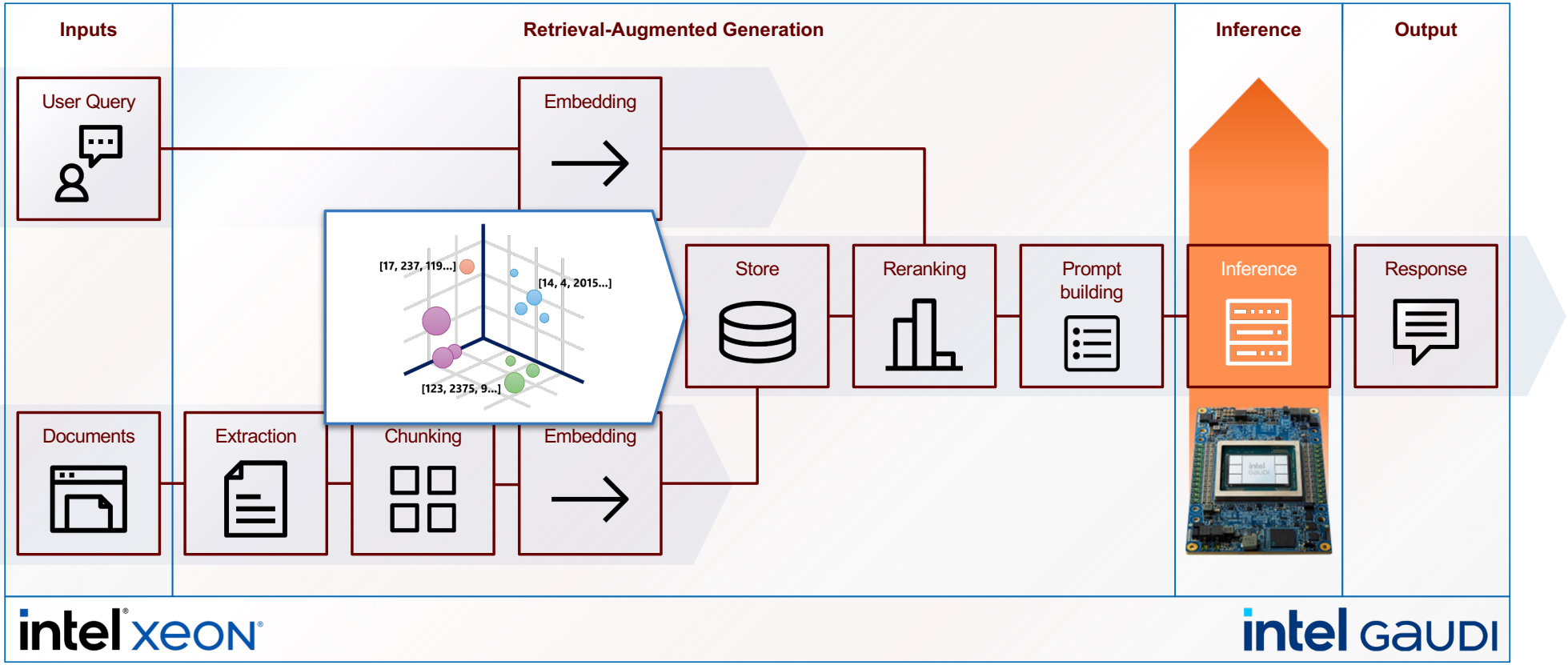
Red Hat | intel

Retrieval Augmented Generation (RAG)



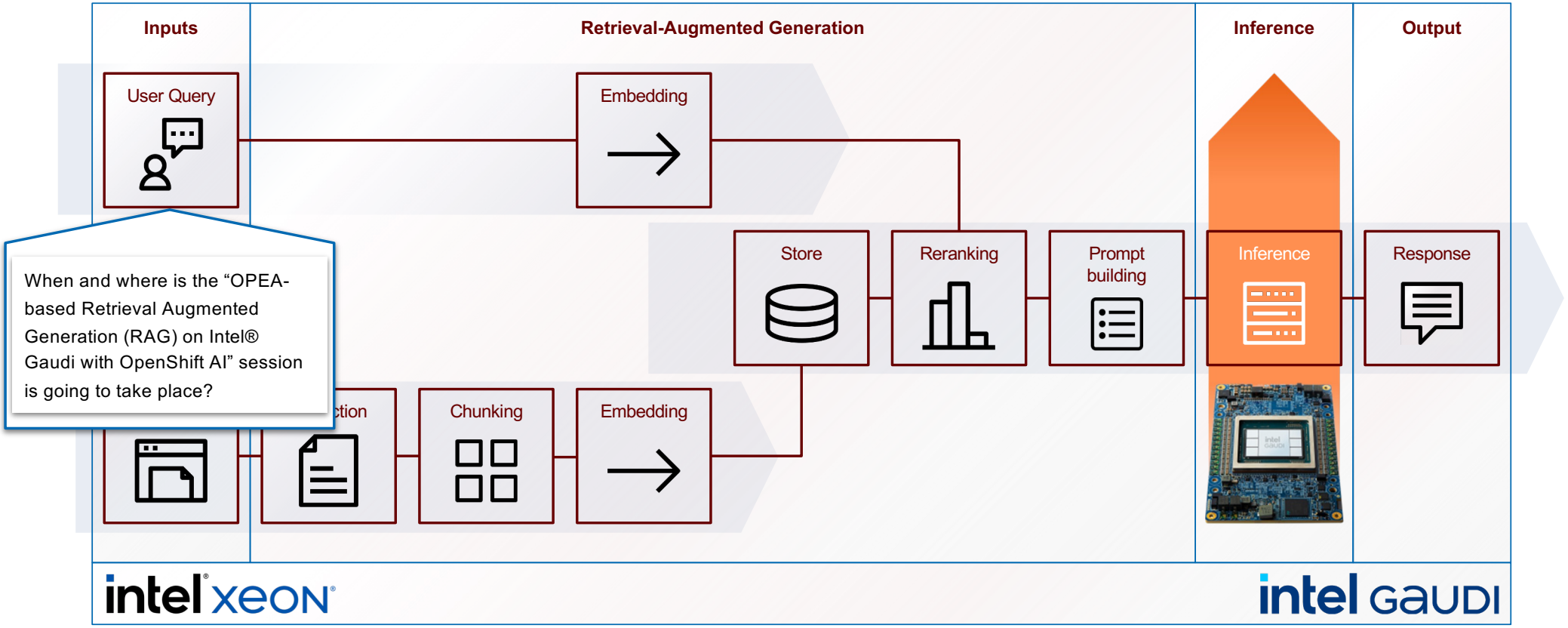
Red Hat | intel

Retrieval Augmented Generation (RAG)



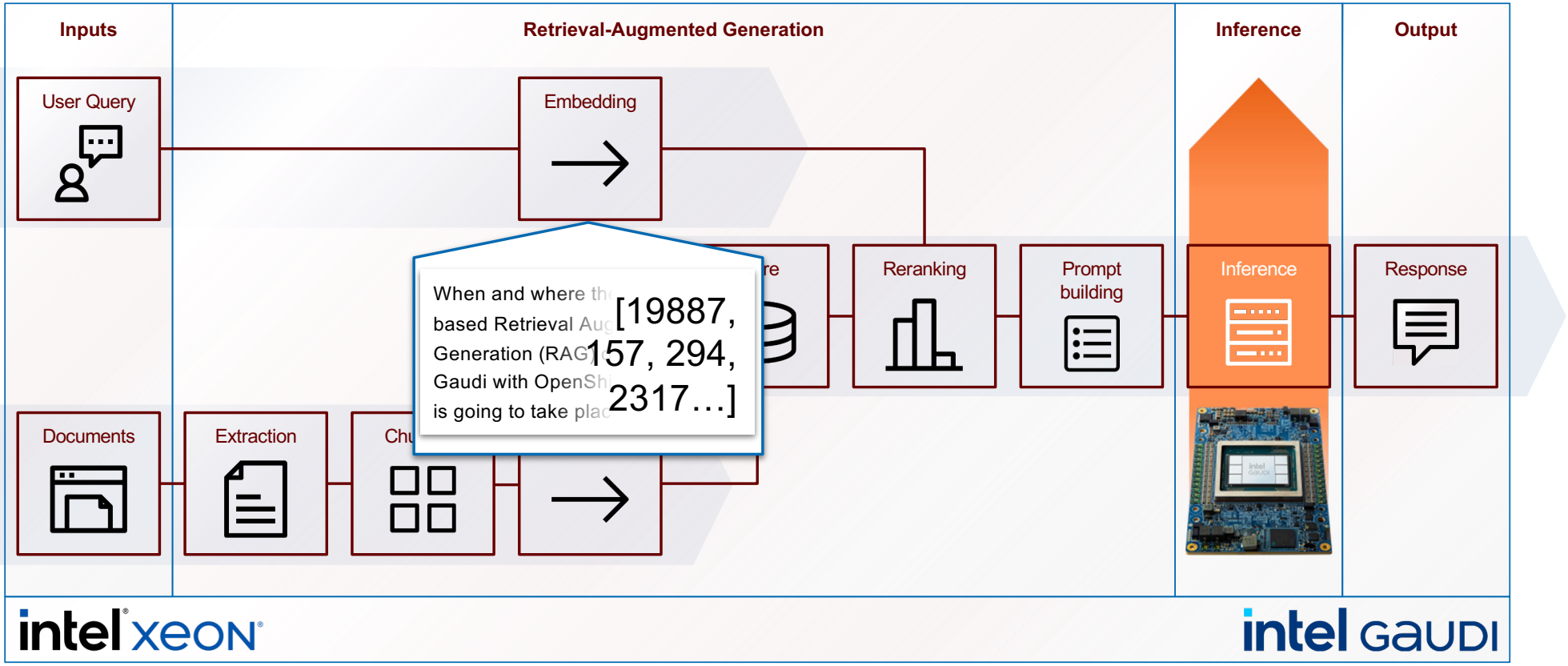
Red Hat | intel.

Retrieval Augmented Generation (RAG)



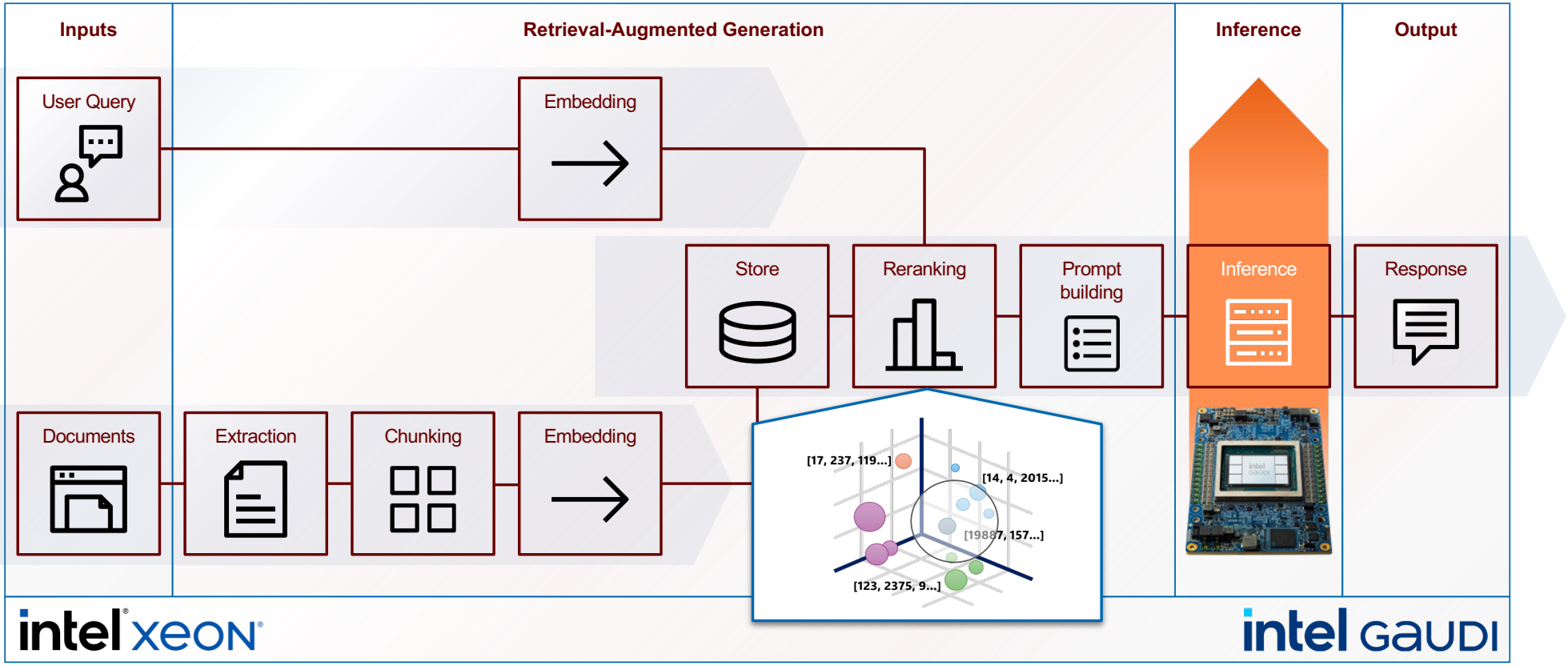
Red Hat | intel

Retrieval Augmented Generation (RAG)



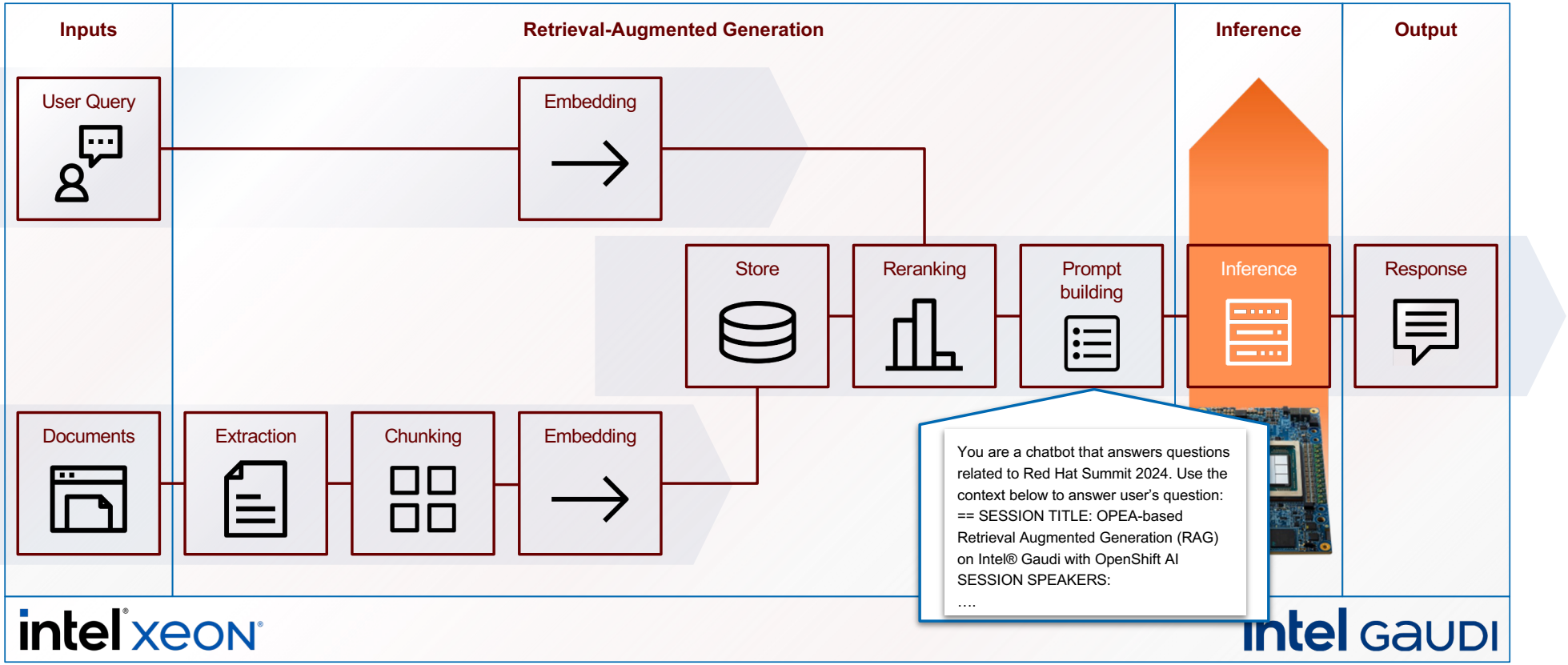
Red Hat | intel.

Retrieval Augmented Generation (RAG)



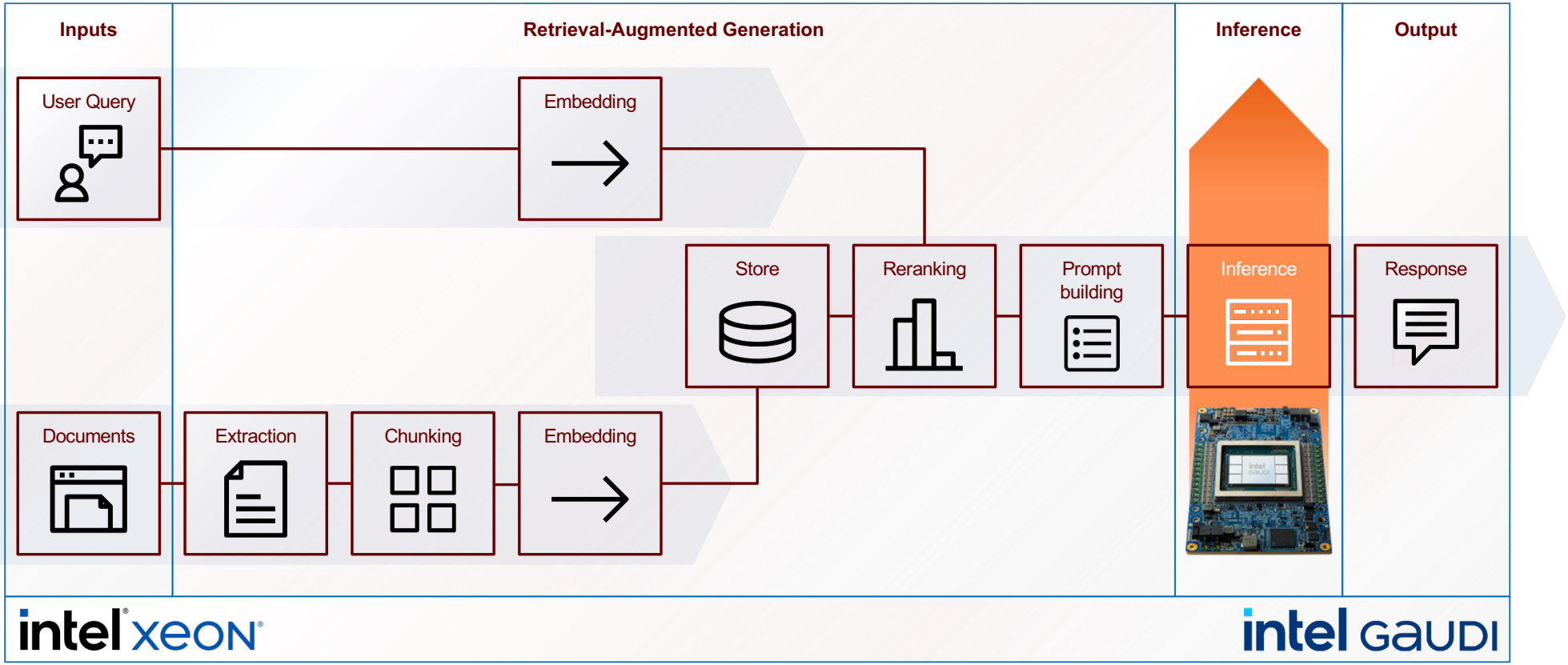
Red Hat | intel.

Retrieval Augmented Generation (RAG)



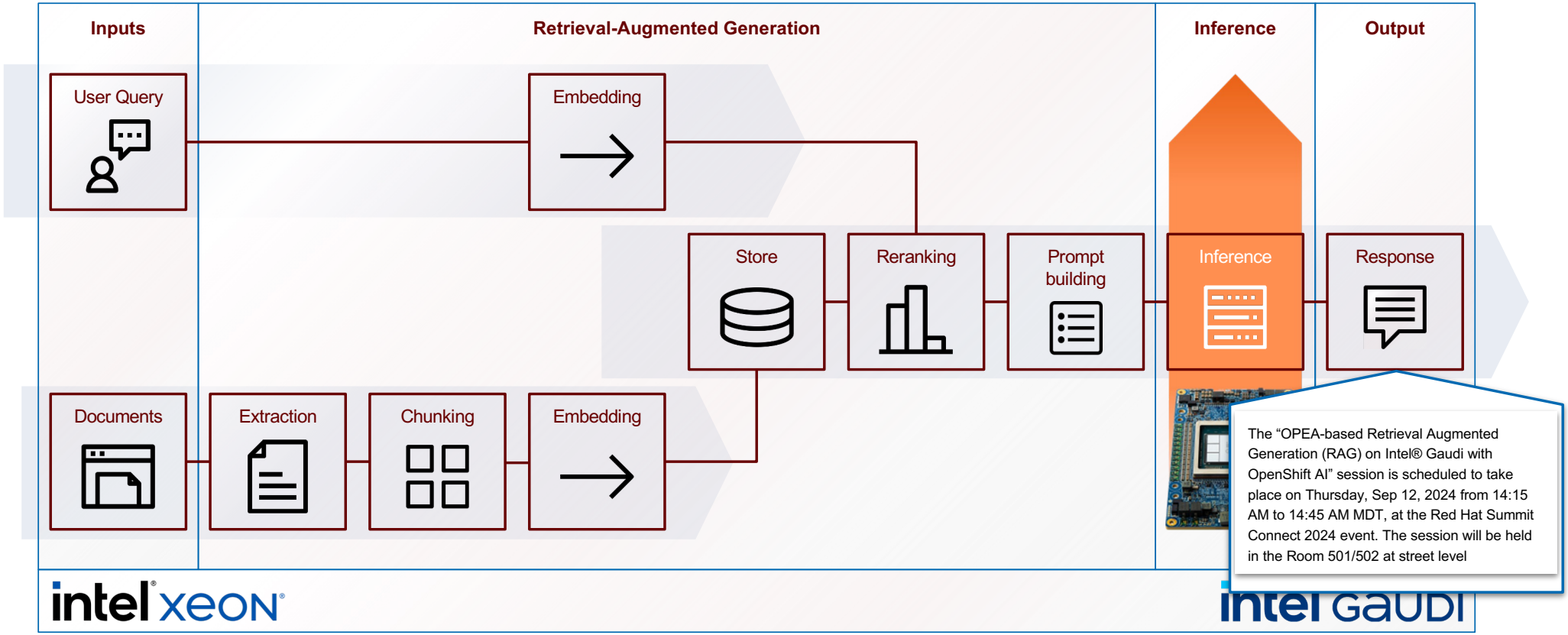
Red Hat | intel.

Retrieval Augmented Generation (RAG)



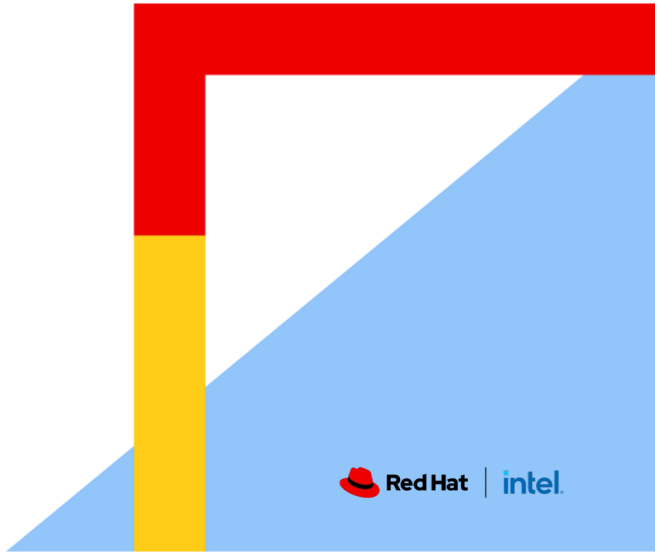
Red Hat | intel

Retrieval Augmented Generation (RAG)

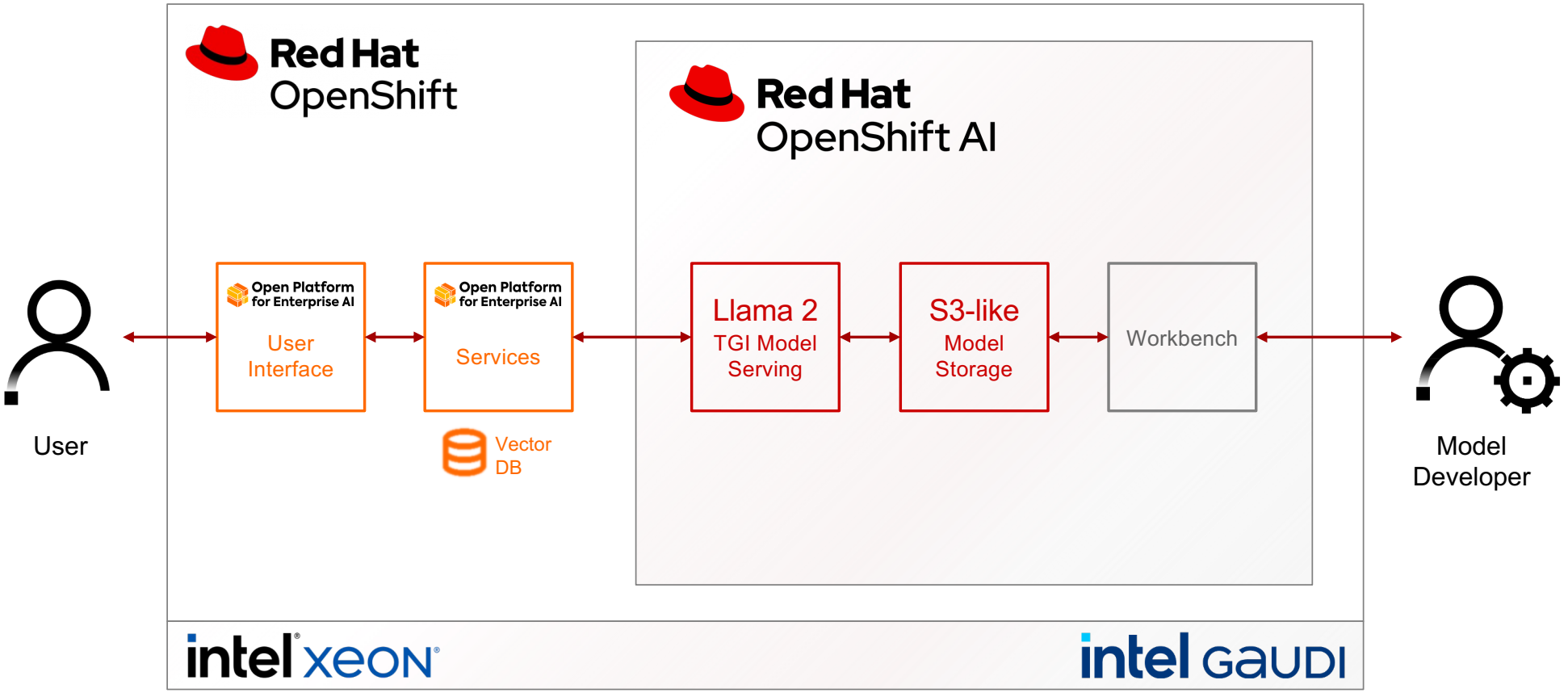


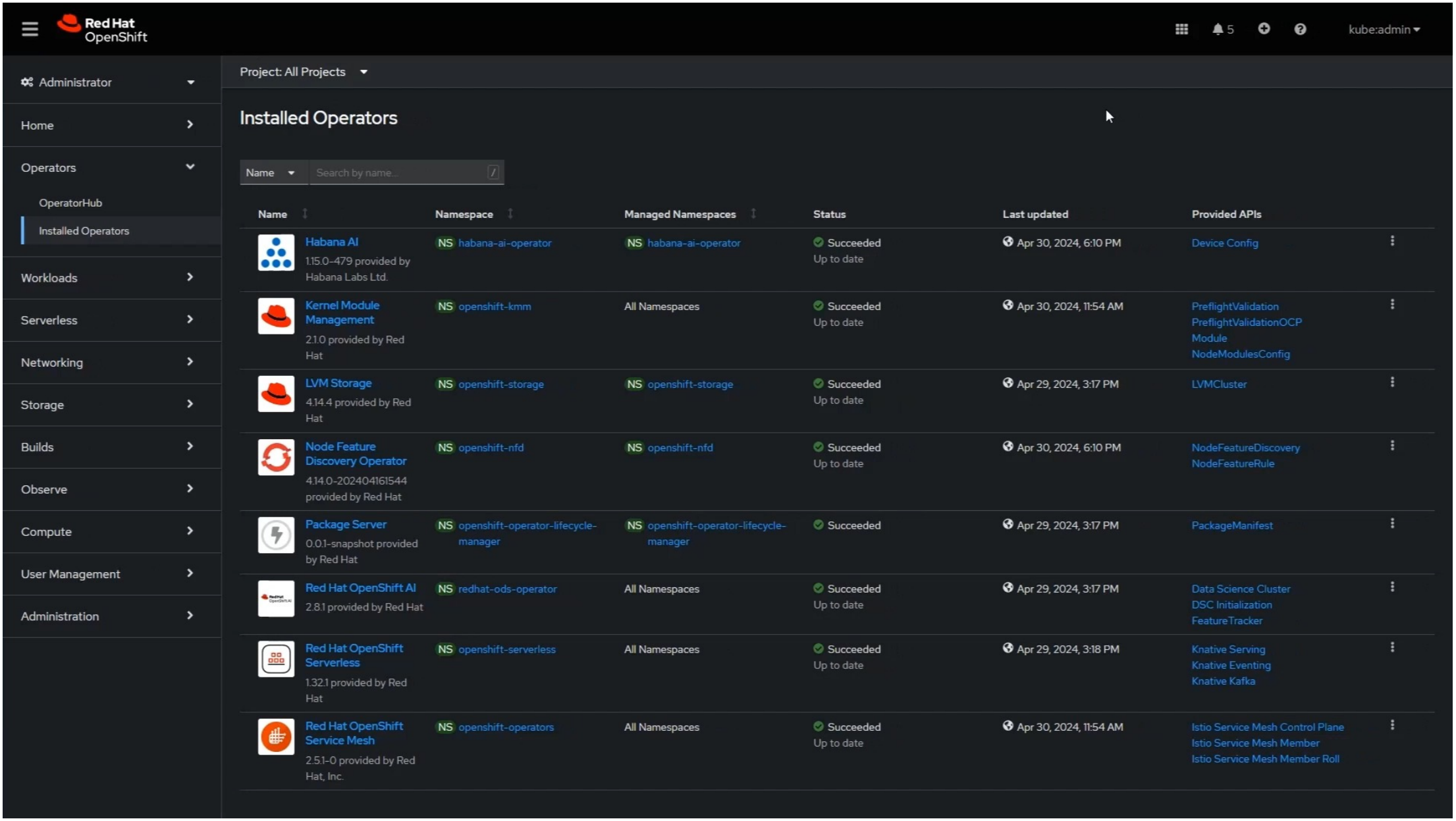
Red Hat | intel.

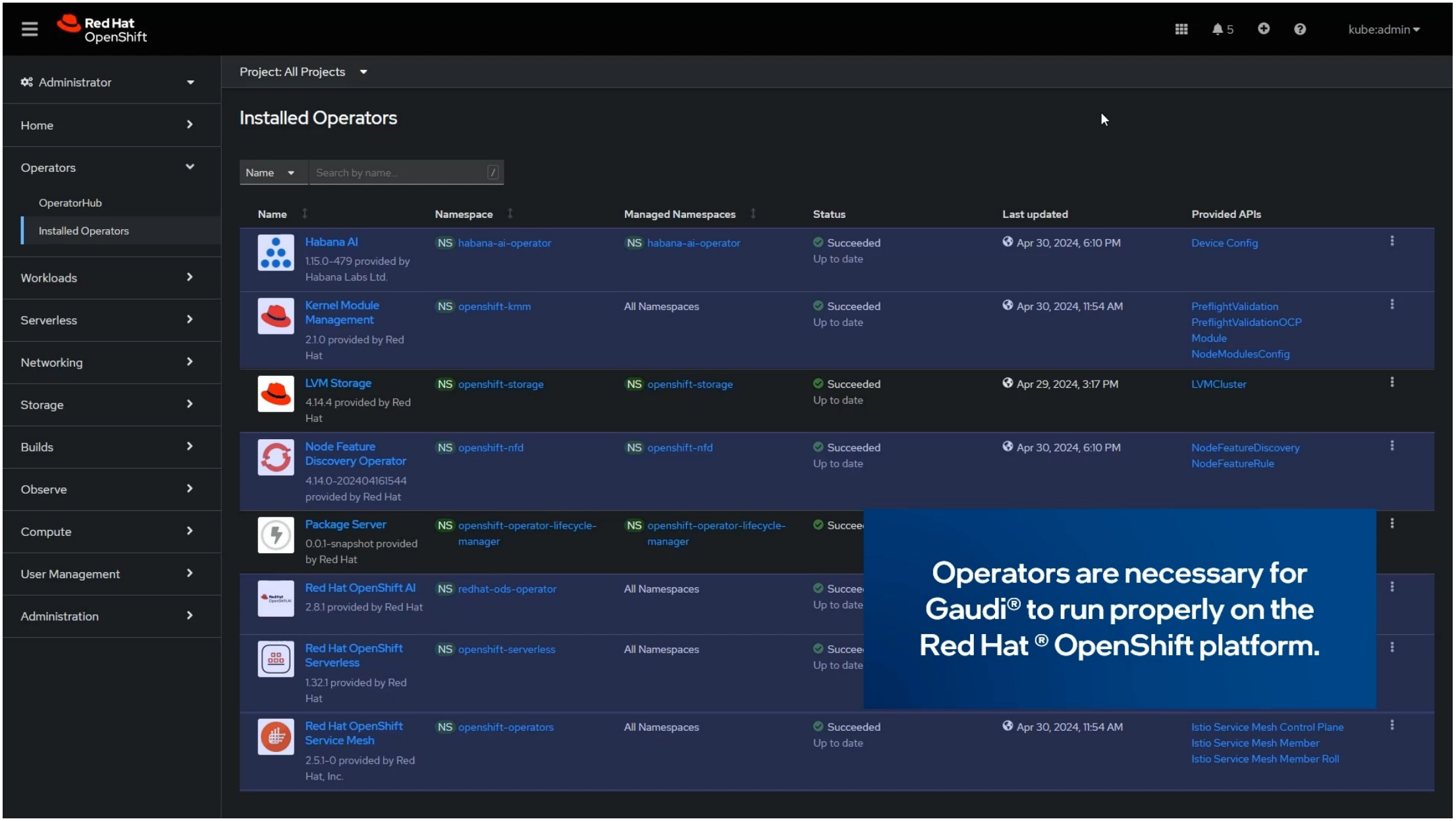
Retrieval Augmented Generation (RAG) ChatBot Demo

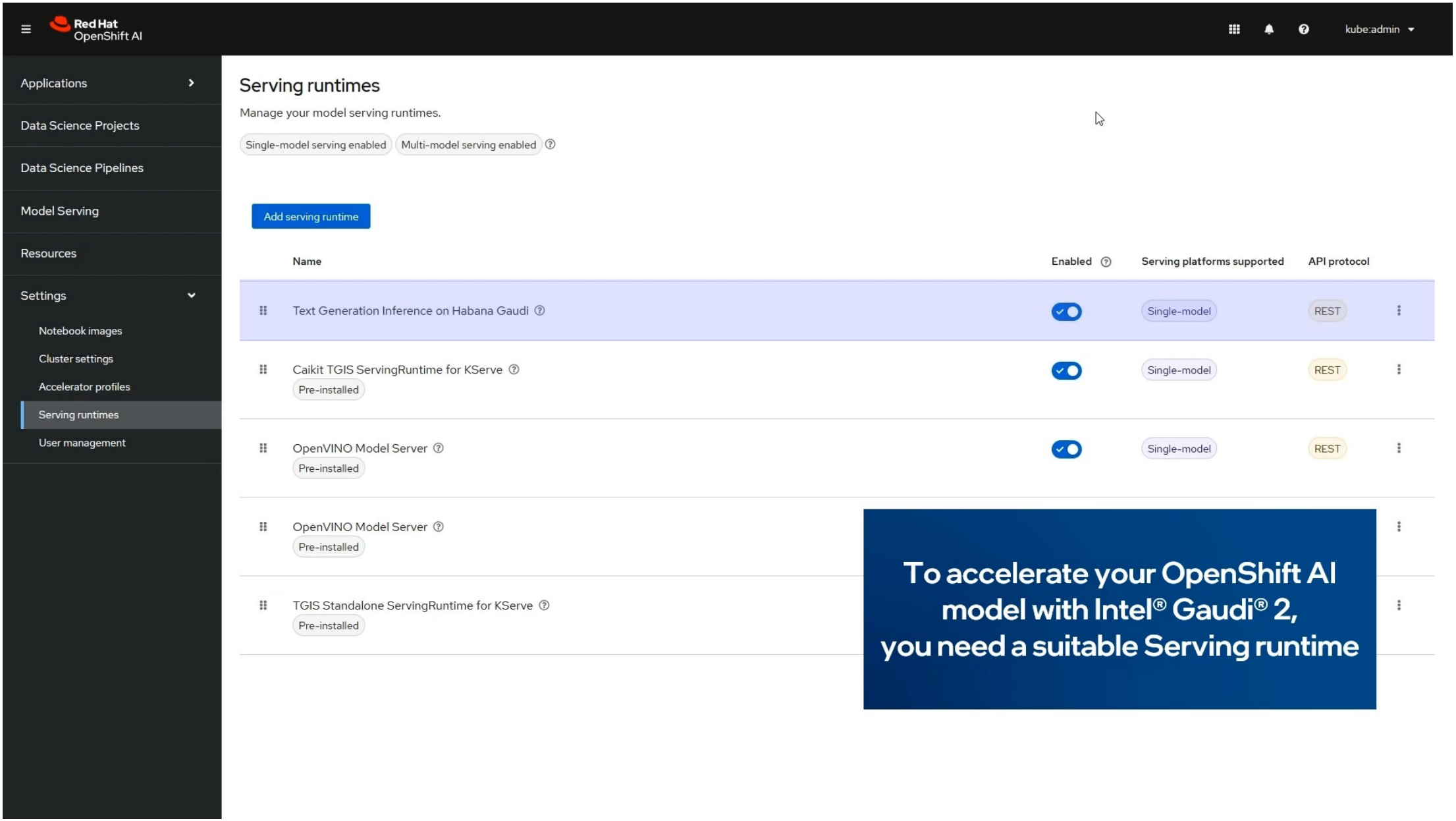


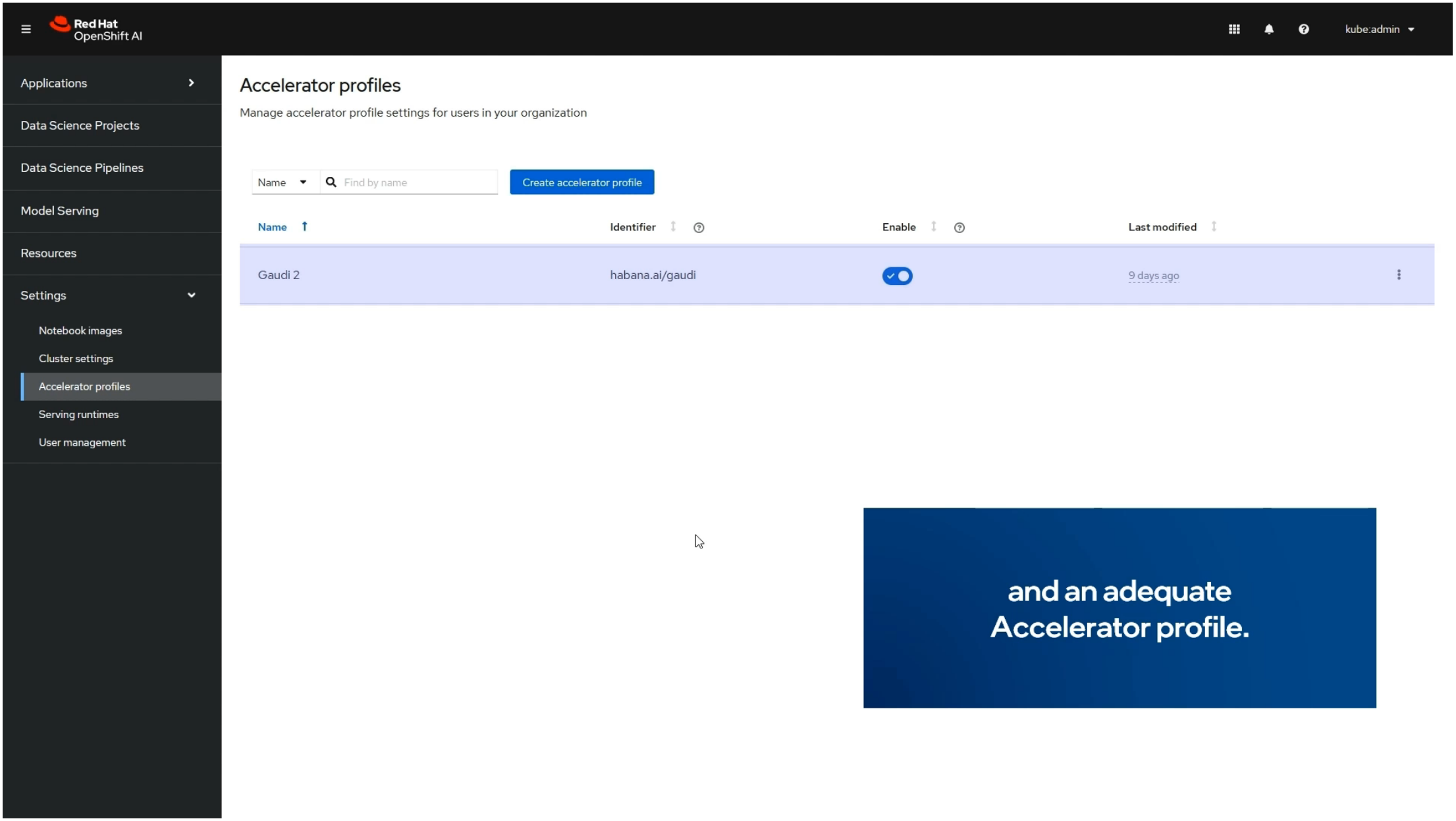
OPEA Chat QnA on OpenShift – demo design

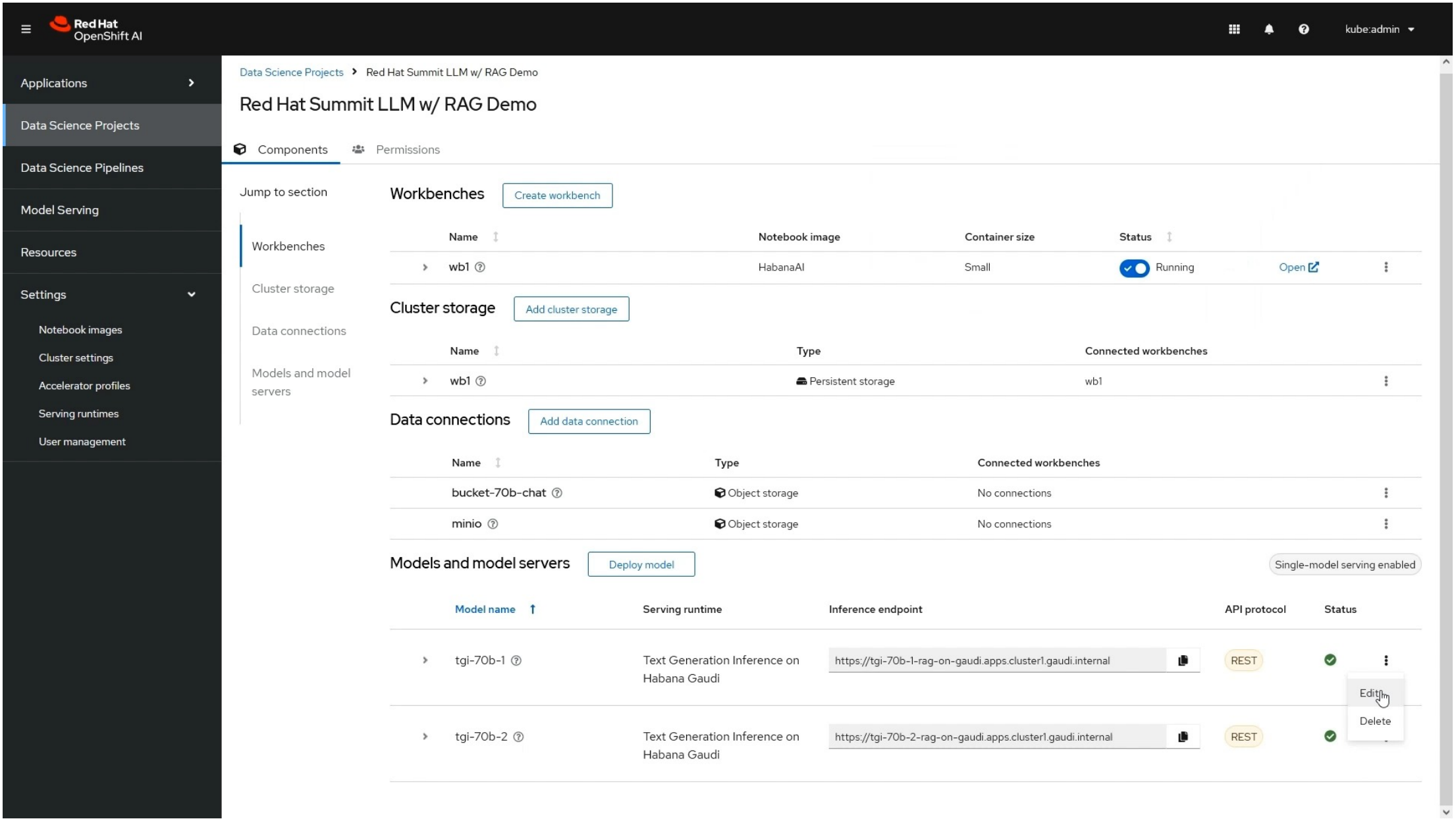


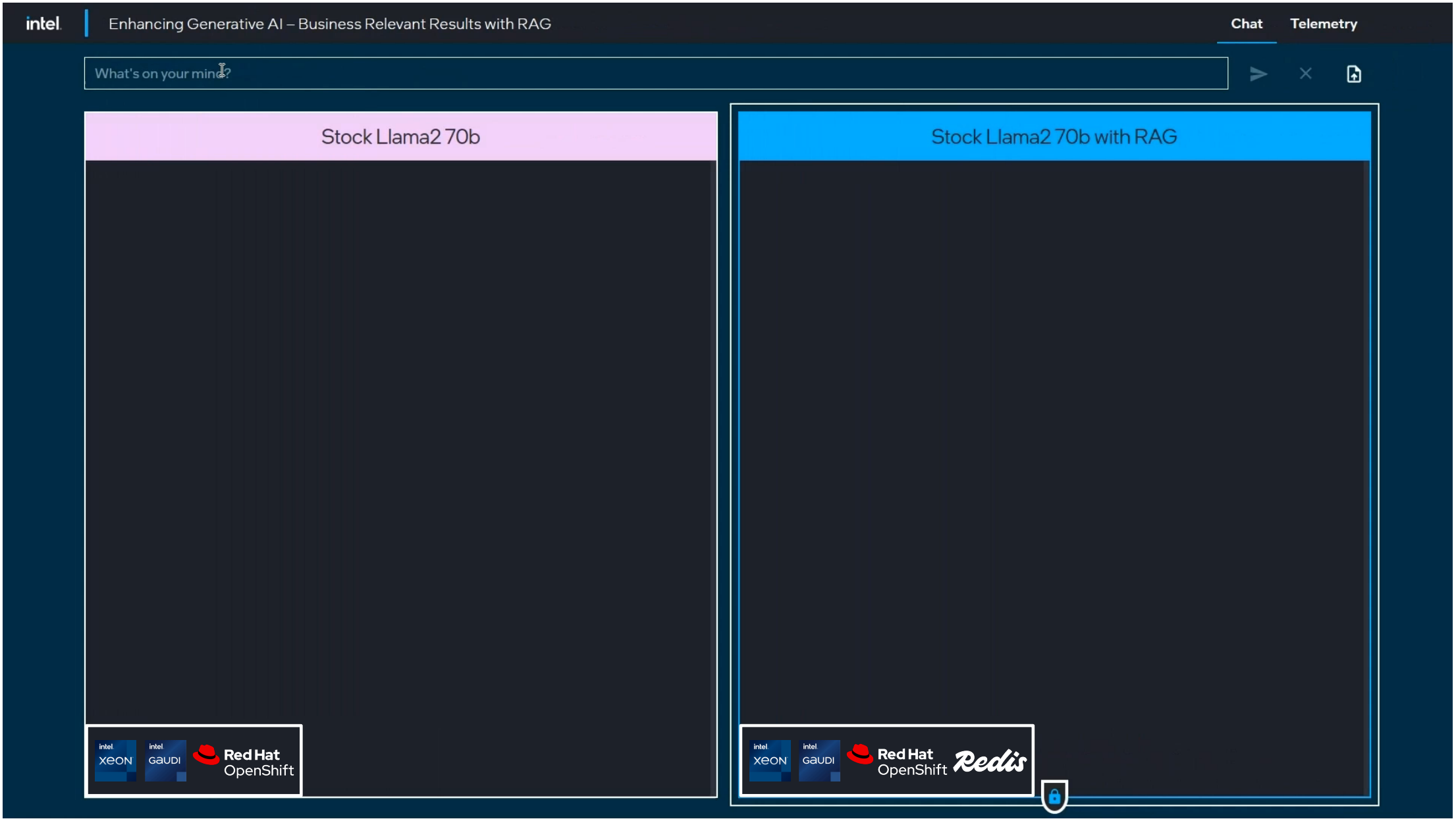

Red Hat OpenShift
5
kube:admin
Administrator
Home
Operators
OperatorHub
Installed Operators
Workloads
Serverless
Networking
Storage
Builds
Observe
Compute
User Management
Administration
Project: All Projects
Installed Operators
Name
Search by name...
Name
Namespace
Managed Namespaces
Status
Last updated
Provided APIs
Habana AI
1.15.0-479 provided by Habana Labs Ltd.
NS habana-ai-operator
NS habana-ai-operator
Succeeded
Up to date
Apr 30, 2024, 6:10 PM
Device Config
Kernel Module Management
2.1.0 provided by Red Hat
NS openshift-kmm
All Namespaces
Succeeded
Up to date
Apr 30, 2024, 11:54 AM
PreflightValidation
PreflightValidationOCP
Module
NodeModulesConfig
LVM Storage
4.14.4 provided by Red Hat
NS openshift-storage
NS openshift-storage
Succeeded
Up to date
Apr 29, 2024, 3:17 PM
LVMCluster
Node Feature Discovery Operator
4.14.0-202404161544 provided by Red Hat
NS openshift-nfd
NS openshift-nfd
Succeeded
Up to date
Apr 30, 2024, 6:10 PM
NodeFeatureDiscovery
NodeFeatureRule
Package Server
0.0.1-snapshot provided by Red Hat
NS openshift-operator-lifecycle-manager
NS openshift-operator-lifecycle-manager
Succeeded
Apr 29, 2024, 3:17 PM
PackageManifest
Red Hat OpenShift AI
2.8.1 provided by Red Hat
NS redhat-ods-operator
All Namespaces
Succeeded
Up to date
Apr 29, 2024, 3:17 PM
Data Science Cluster
DSC Initialization
FeatureTracker
Red Hat OpenShift Serverless
1.32.1 provided by Red Hat
NS openshift-serverless
All Namespaces
Succeeded
Up to date
Apr 29, 2024, 3:18 PM
Knative Serving
Knative Eventing
Knative Kafka
Red Hat OpenShift Service Mesh
2.5.1-0 provided by Red Hat, Inc.
NS openshift-operators
All Namespaces
Succeeded
Up to date
Apr 30, 2024, 11:54 AM
Istio Service Mesh Control Plane
Istio Service Mesh Member
Istio Service Mesh Member Roll

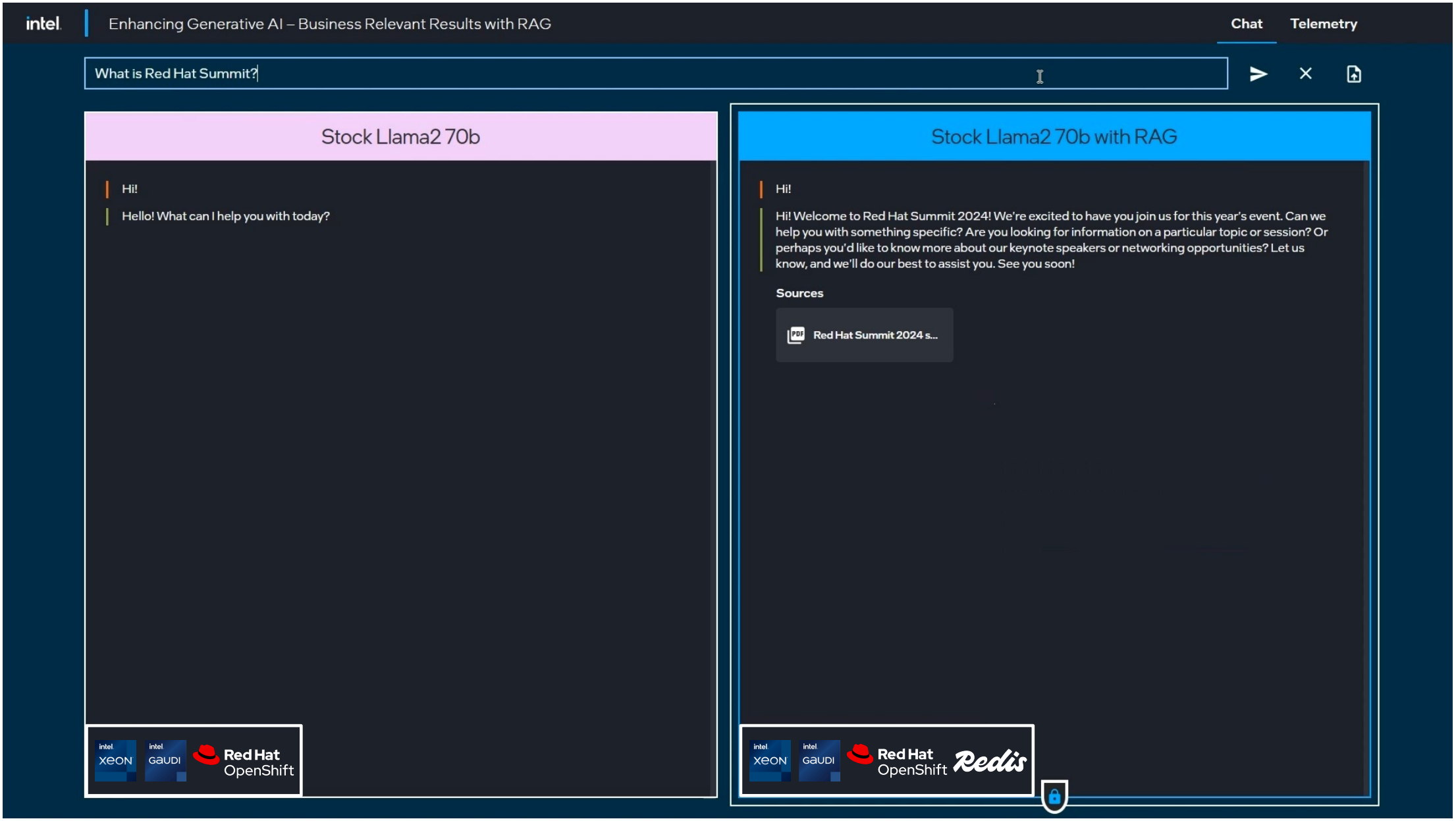

Red Hat OpenShift
5
kube:admin
Administrator
Home
Operators
OperatorHub
Installed Operators
Workloads
Serverless
Networking
Storage
Builds
Observe
Compute
User Management
Administration
Project: All Projects
Installed Operators
Name
Search by name...
Name
Namespace
Managed Namespaces
Status
Last updated
Provided APIs
Habana AI
1.15.0-479 provided by Habana Labs Ltd.
NS habana-ai-operator
NS habana-ai-operator
Succeeded
Up to date
Apr 30, 2024, 6:10 PM
Device Config
Kernel Module Management
2.1.0 provided by Red Hat
NS openshift-kmm
All Namespaces
Succeeded
Up to date
Apr 30, 2024, 11:54 AM
PreflightValidation
PreflightValidationOCP
Module
NodeModulesConfig
LVM Storage
4.14.4 provided by Red Hat
NS openshift-storage
NS openshift-storage
Succeeded
Up to date
Apr 29, 2024, 3:17 PM
LVMCluster
Node Feature Discovery Operator
4.14.0-202404161544 provided by Red Hat
NS openshift-nfd
NS openshift-nfd
Succeeded
Up to date
Apr 30, 2024, 6:10 PM
NodeFeatureDiscovery
NodeFeatureRule
Package Server
0.0.1-snapshot provided by Red Hat
NS openshift-operator-lifecycle-manager
NS openshift-operator-lifecycle-manager
Succee
Red Hat OpenShift AI
2.8.1 provided by Red Hat
NS redhat-ods-operator
All Namespaces
Succee
Up to date
Red Hat OpenShift Serverless
1.32.1 provided by Red Hat
NS openshift-serverless
All Namespaces
Succee
Up to date
Red Hat OpenShift Service Mesh
2.5.1-0 provided by Red Hat, Inc.
NS openshift-operators
All Namespaces
Succeeded
Up to date
Apr 30, 2024, 11:54 AM
Istio Service Mesh Control Plane
Istio Service Mesh Member
Istio Service Mesh Member Roll
Operators are necessary for Gaudi® to run properly on the Red Hat ® OpenShift platform.


Red Hat OpenShift AI
kube:admin
Applications
Data Science Projects
Data Science Pipelines
Model Serving
Resources
Settings
Notebook images
Cluster settings
Accelerator profiles
Serving runtimes
User management
Serving runtimes
Manage your model serving runtimes.
Single-model serving enabled
Multi-model serving enabled
Add serving runtime
Name
Enabled
Serving platforms supported
API protocol
Text Generation Inference on Habana Gaudi
Single-model
REST
Caikit TGIS ServingRuntime for KServe
Pre-installed
Single-model
REST
OpenVINO Model Server
Pre-installed
Single-model
REST
OpenVINO Model Server
Pre-installed
TGIS Standalone ServingRuntime for KServe
Pre-installed
To accelerate your OpenShift AI model with Intel® Gaudi® 2, you need a suitable Serving runtime

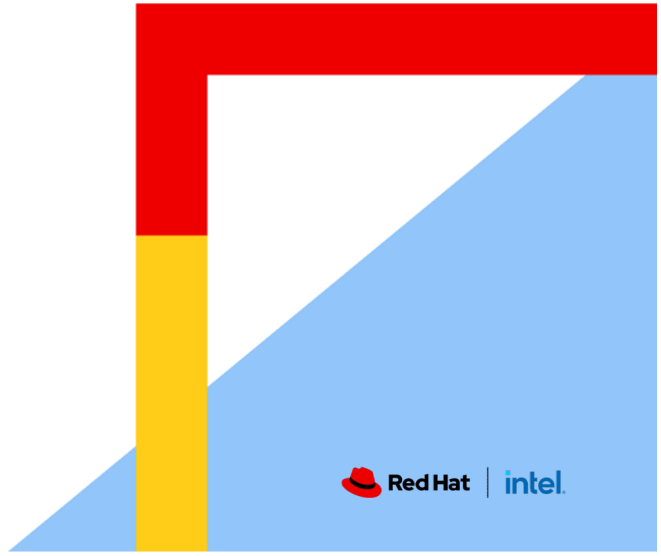

Red Hat OpenShift AI
kube:admin
Applications
Data Science Projects
Data Science Pipelines
Model Serving
Resources
Settings
Notebook images
Cluster settings
Accelerator profiles
Serving runtimes
User management
Accelerator profiles
Manage accelerator profile settings for users in your organization
Name
Find by name
Create accelerator profile
Name
Identifier
Enable
Last modified
Gaudi 2
habana.ai/gaudi
9 days ago
and an adequate
Accelerator profile.


Red Hat OpenShift AI
kube:admin
Applications
Data Science Projects
Data Science Pipelines
Model Serving
Resources
Settings
Notebook images
Cluster settings
Accelerator profiles
Serving runtimes
User management
Data Science Projects > Red Hat Summit LLM w/ RAG Demo
Red Hat Summit LLM w/ RAG Demo
Components
Permissions
Jump to section
Workbenches
Cluster storage
Data connections
Models and model servers
Workbenches
Create workbench
Name
Notebook image
Container size
Status
wb1
HabanaAI
Small
Running
Open
Cluster storage
Add cluster storage
Name
Type
Connected workbenches
wb1
Persistent storage
wb1
Data connections
Add data connection
Name
Type
Connected workbenches
bucket-70b-chat
Object storage
No connections
minio
Object storage
No connections
Models and model servers
Deploy model
Single-model serving enabled
Model name
Serving runtime
Inference endpoint
API protocol
Status
tgi-70b-1
Text Generation Inference on Habana Gaudi
https://tgi-70b-1-rag-on-gaudi.apps.cluster1.gaudi.internal
REST
Edit
Delete
tgi-70b-2
Text Generation Inference on Habana Gaudi
https://tgi-70b-2-rag-on-gaudi.apps.cluster1.gaudi.internal
REST


intel
Enhancing Generative AI – Business Relevant Results with RAG
Chat
Telemetry
What's on your mind?
Stock Llama2 70b
Stock Llama2 70b with RAG
intel Xeon
intel Gaudi
Red Hat OpenShift
intel Xeon
intel Gaudi
Red Hat OpenShift
Redis


intel
Enhancing Generative AI – Business Relevant Results with RAG
Chat
Telemetry
What is Red Hat Summit?
Stock Llama2 70b
Hi!
Hello! What can I help you with today?
intel Xeon
intel Gaudi
Red Hat OpenShift
Stock Llama2 70b with RAG
Hi!
Hi! Welcome to Red Hat Summit 2024! We're excited to have you join us for this year's event. Can we help you with something specific? Are you looking for information on a particular topic or session? Or perhaps you'd like to know more about our keynote speakers or networking opportunities? Let us know, and we'll do our best to assist you. See you soon!
Sources
Red Hat Summit 2024 s...
intel Xeon
intel Gaudi
Red Hat OpenShift
Redis

Summary



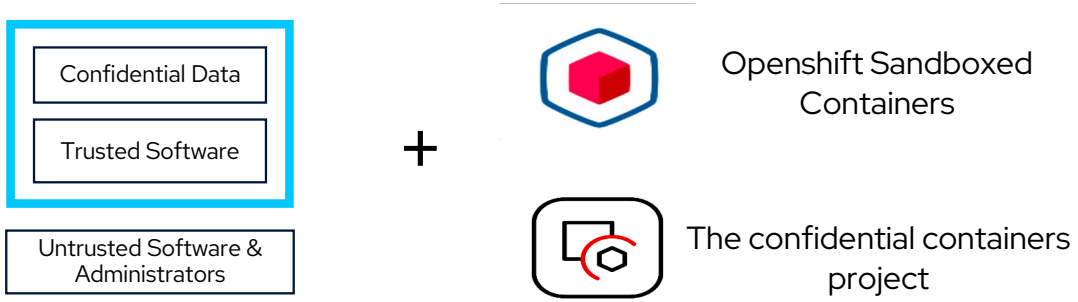
Key Takeaways

- RAG enhances AI development by integrating large language models with real-time data retrieval for more accurate and contextually relevant outputs, crucial for complex tasks.
- OPEA offers a flexible, multivendor ecosystem that simplifies AI deployment and boosts performance.
- OpenShift AI provides a scalable, cloud-native platform that integrates seamlessly into DevOps workflows.
- Intel Gaudi AI Accelerators complement these by delivering high performance and efficiency, reducing costs, making them ideal for enterprise-scale AI solutions.
- Together, these technologies create a powerful, interoperable, and efficient AI development environment.

Confidential AI Helps Protect Data & Models In-Use

Utilizing Confidential Computing for Containers with Intel TDX

Hardware-Based Protection of Data In-Use
With Intel Trusted Domain Extensions (TDX)



Confidential Computing is about protecting data in-use
You do not have to trust the system admins of the providers any longer

Red Hat | intel.

Confidential AI Helps Protect Data & Models In-Use

Utilizing Confidential Computing for Containers with Intel TDX

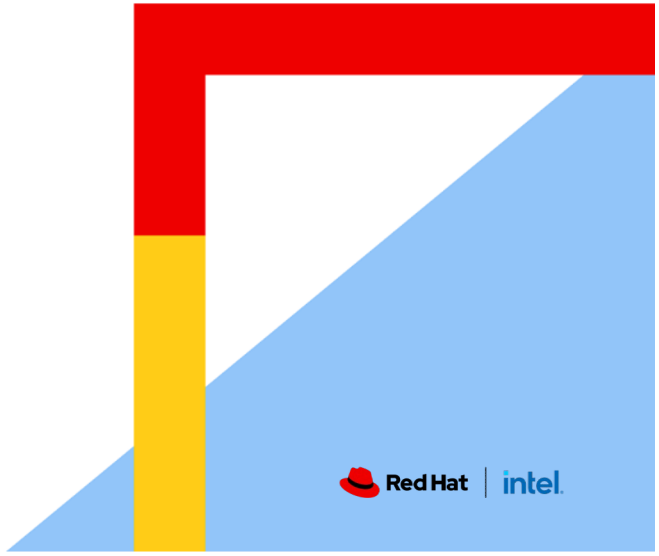
Come visit the Intel and Red Hat booth on the showfloor to learn more about Confidential Computing

Confidential Computing is about protecting data in-use

You do not have to trust the system admins of the providers any longer

Red Hat | intel.

Q&A



Red Hat Summit

Connect

Thank you

 linkedin.com/company/red-hat

 facebook.com/redhatinc

 youtube.com/user/RedHatVideos

 twitter.com/RedHat

Red Hat | intel