Red Hat

Wojciech Furmankiewicz

Head of Technology Sales, CEE

Red Hat

Piotr Grabuszyński

Cloud Software Architect

Intel

Red Hat + intel.

Over 25 Years of Collaboration

Bringing AI Everywhere

Intel's AI Strategy

AI PC Node

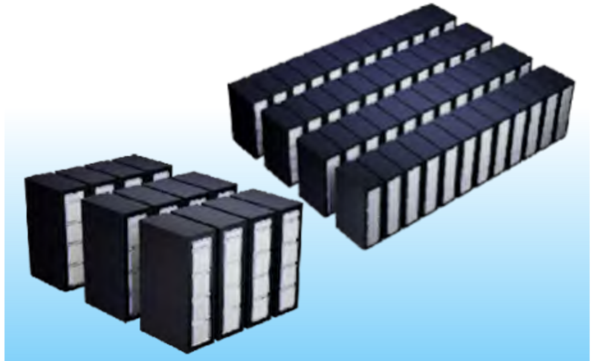
AI Developer Productivity & Light Inference

Node

Fine-tuning, Inference

Cluster

Light Training, Tuning, Peak Inference



Super Cluster

Training, Tuning, Peak Inference

Mega Cluster

Large Scale Training & Inference

AI PC

Broadest AI SW Ecosystem

ENTERPRISE AI & EDGE AI

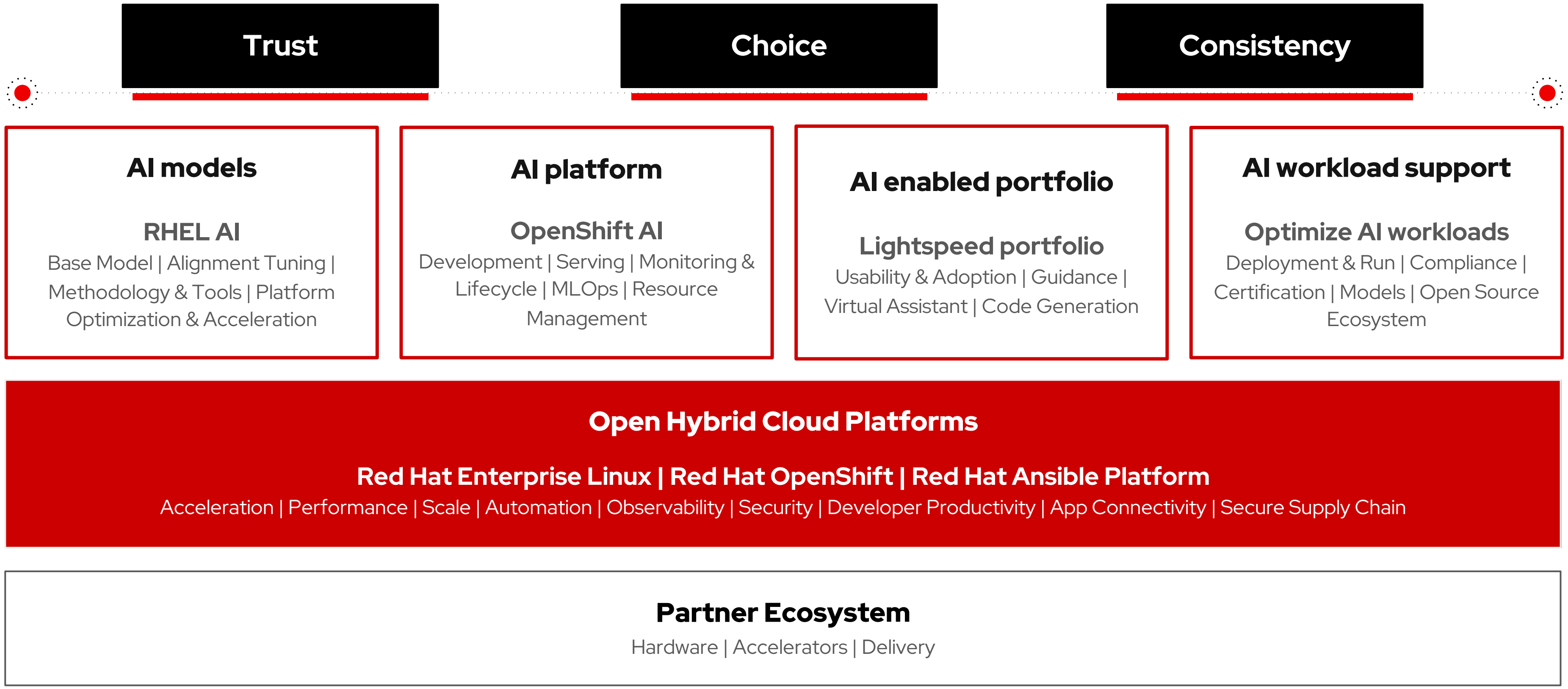
Open Standard, "Ready to Use"

DATA CENTER AI

AI Open, Scalable Systems & Reference Arch

intel CORE ULTRA | intel ARC GRAPHICS | intel CORE ULTRA | intel XEON | intel GAUDI | intel XEON | intel GAUDI | intel IPU

Red Hat's AI Strategy



Trust

Choice

Consistency

AI models

RHEL AI

Base Model | Alignment Tuning | Methodology & Tools | Platform Optimization & Acceleration

AI platform

OpenShift AI

Development | Serving | Monitoring & Lifecycle | MLOps | Resource Management

AI enabled portfolio

Lightspeed portfolio

Usability & Adoption | Guidance | Virtual Assistant | Code Generation

AI workload support

Optimize AI workloads

Deployment & Run | Compliance | Certification | Models | Open Source Ecosystem

Open Hybrid Cloud Platforms

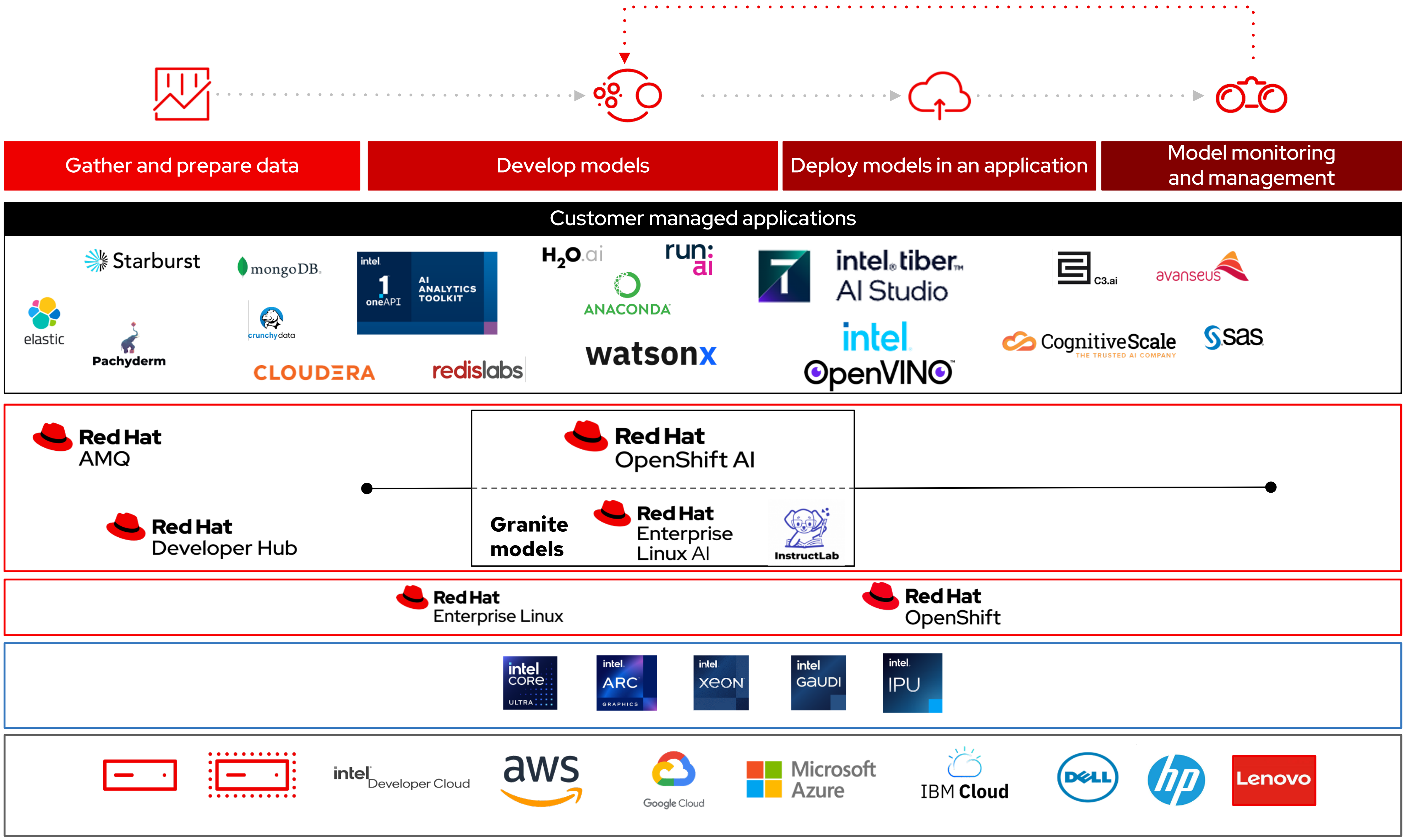
Red Hat Enterprise Linux | Red Hat OpenShift | Red Hat Ansible Platform

Acceleration | Performance | Scale | Automation | Observability | Security | Developer Productivity | App Connectivity | Secure Supply Chain

Partner Ecosystem

Hardware | Accelerators | Delivery

Intel Enterprise AI with Red Hat® OpenShift® AI



Gather and prepare data

Develop models

Deploy models in an application

Model monitoring and management

Customer managed applications

ISV software and services including INTEL

Starburst, mongoDB., intel oneAPI AI Analytics Toolkit, H2O.ai, run:ai, ANACONDA, intel. tiber. AI Studio, C3.ai, avanseus, elastic, Pachyderm, crunchy data, CLOUDERA, redislabs, watsonx, intel OpenVINO, CognitiveScale THE TRUSTED AI COMPANY, sas

Red Hat Software and cloud services Hybrid, multi-cloud platform

Red Hat AMQ

Red Hat Developer Hub

Red Hat OpenShift AI

Granite models

Red Hat Enterprise Linux AI

InstructLab

Trusted, cloud-ready platform

Red Hat Enterprise Linux

Red Hat OpenShift

Intel® Core™, Intel® Arc™, Intel® Xeon®, Intel® Gaudi®, Intel® IPU

intel CORE ULTRA, intel ARC GRAPHICS, intel XEON, intel GAUDI, intel IPU

Deploy anywhere

intel Developer Cloud, aws, Google Cloud, Microsoft Azure, IBM Cloud, DELL, hp, Lenovo

Red Hat | intel.

OPEA – Open Platform for Enterprise AI

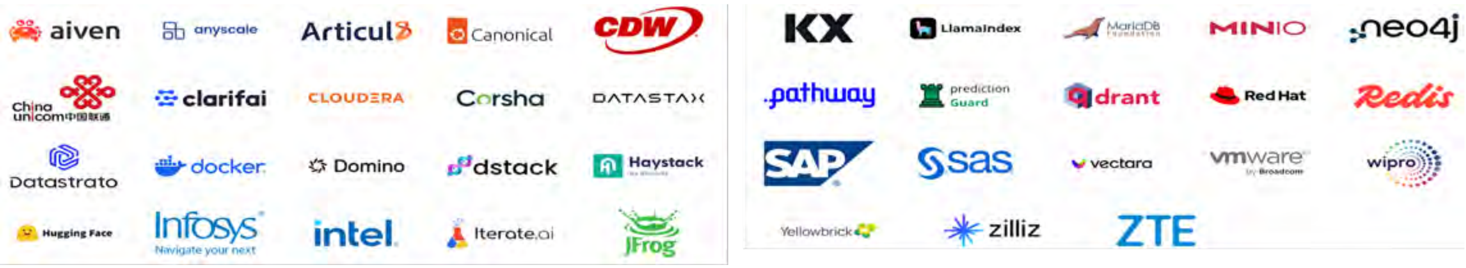
Red Hat | intel.

Open Platform for Enterprise AI

OPEA - Open Platform for Enterprise AI

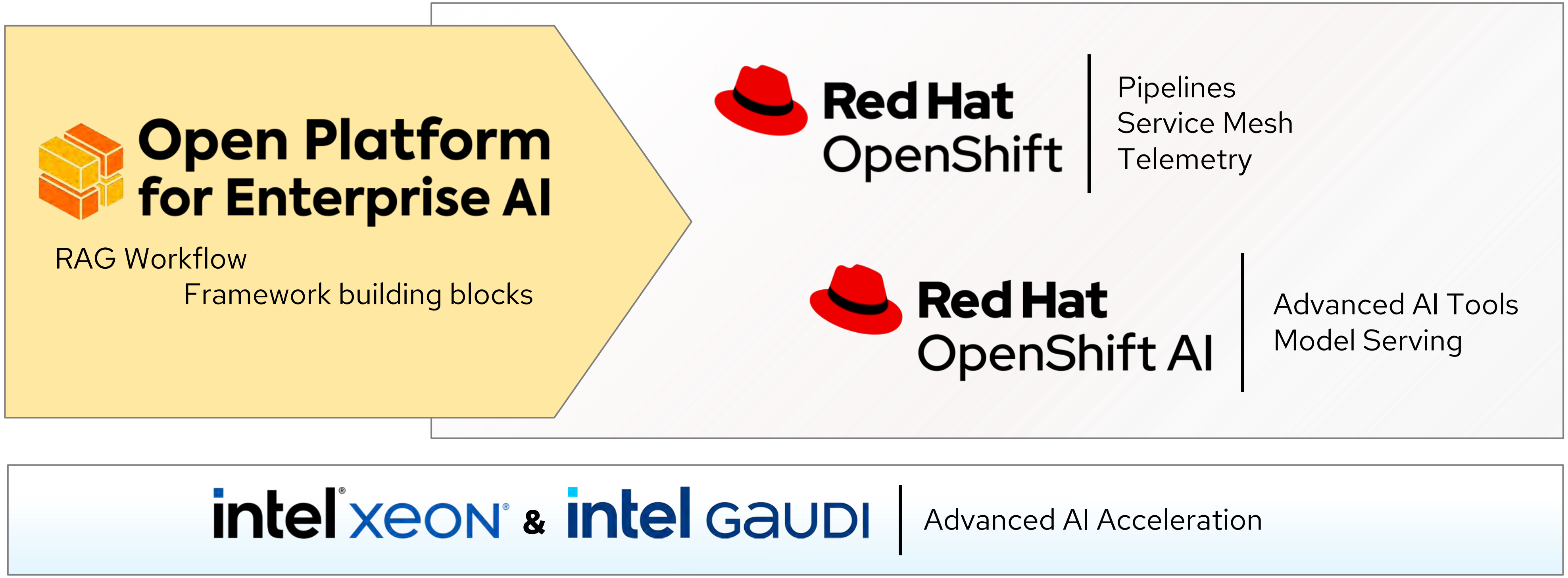
By The Linux Foundation

- Ecosystem orchestration framework for GenAI
- OPEA.dev
- GitHub: https://github.com/opea-project
- Contributors:



OPEA with OpenShift AI

OpenShift AI makes OPEA more enterprise ready



Open Platform for Enterprise AI

RAG Workflow

Framework building blocks

Red Hat OpenShift

Pipelines
Service Mesh
Telemetry

Red Hat OpenShift AI

Advanced AI Tools
Model Serving

intel xeon & intel GAUDI

Advanced AI Acceleration

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]

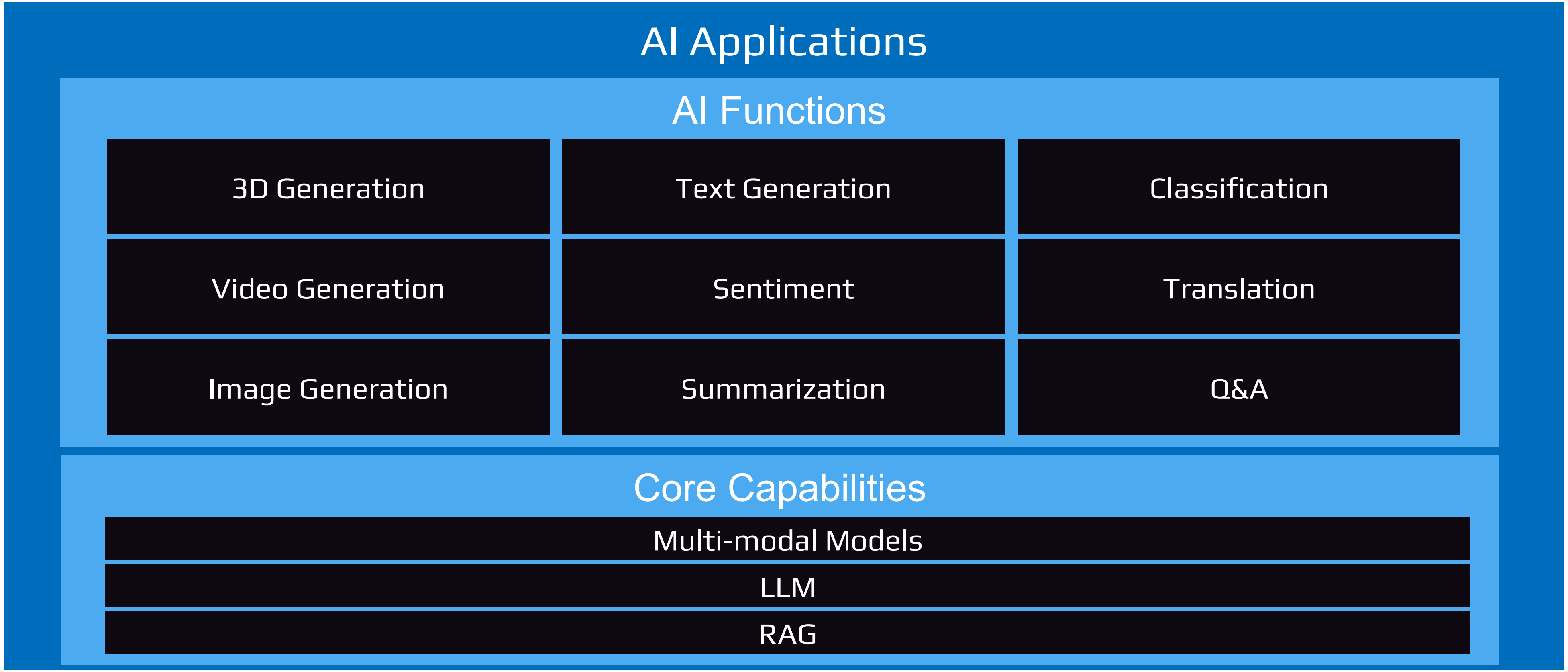
Open Development on GenAI platforms

- 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

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.

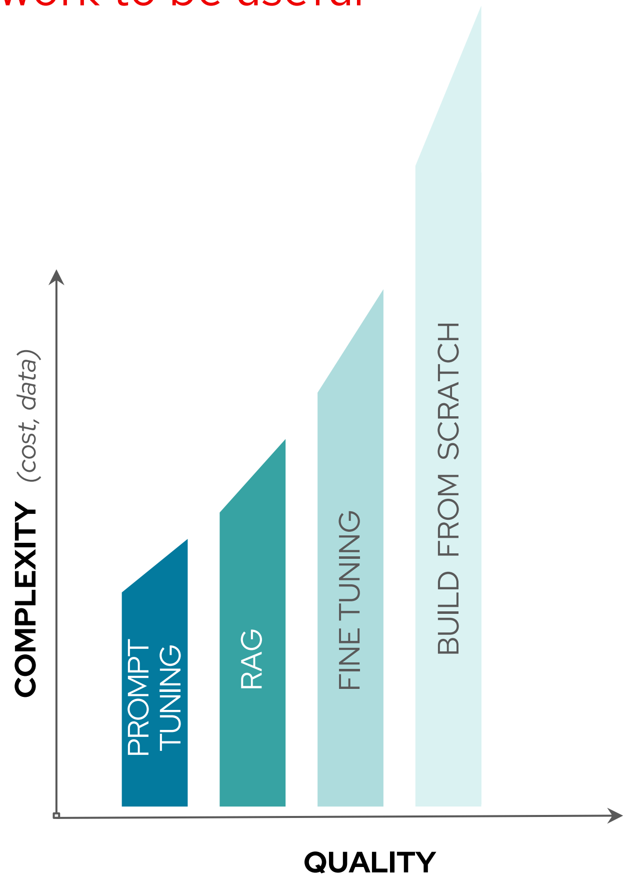
intel | IBM

Retrieval Augmented Generation (RAG) Explained

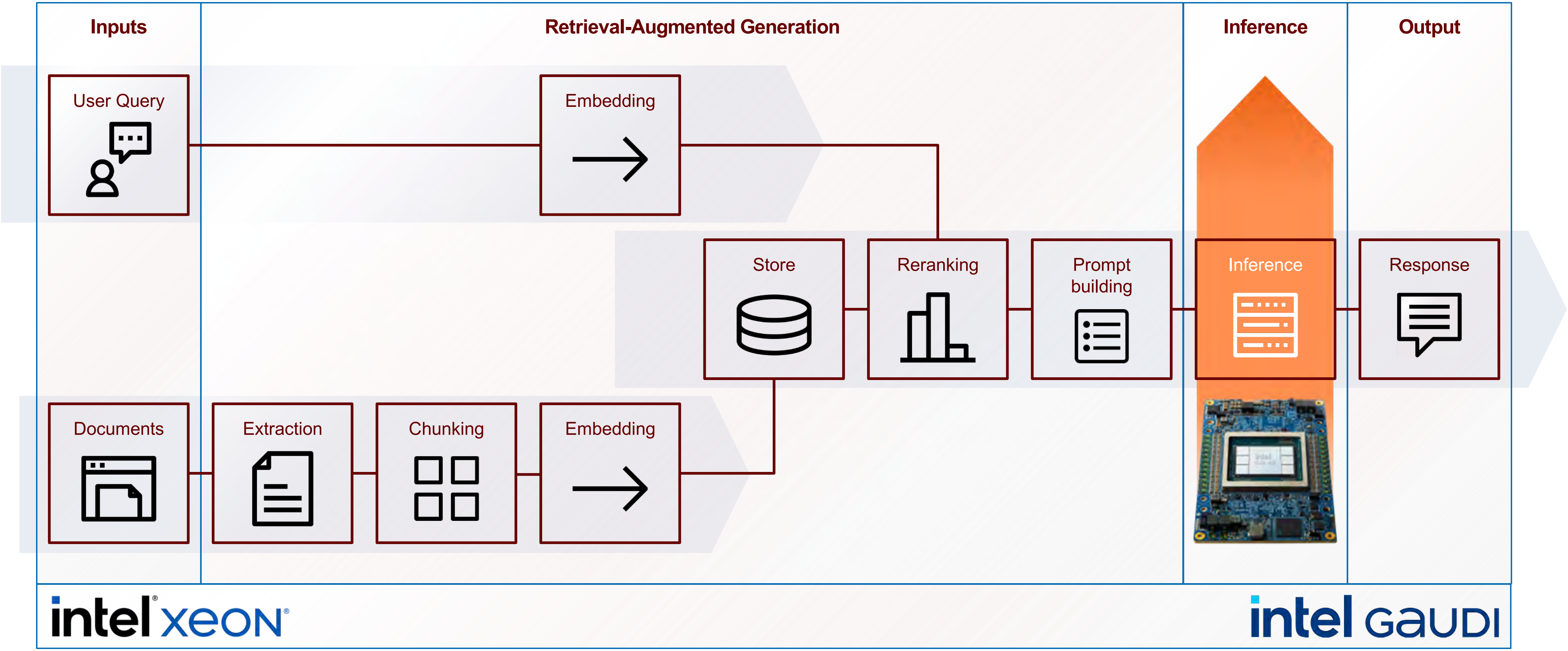
The balancing act of using foundation models

Foundation models will still need more work to be useful

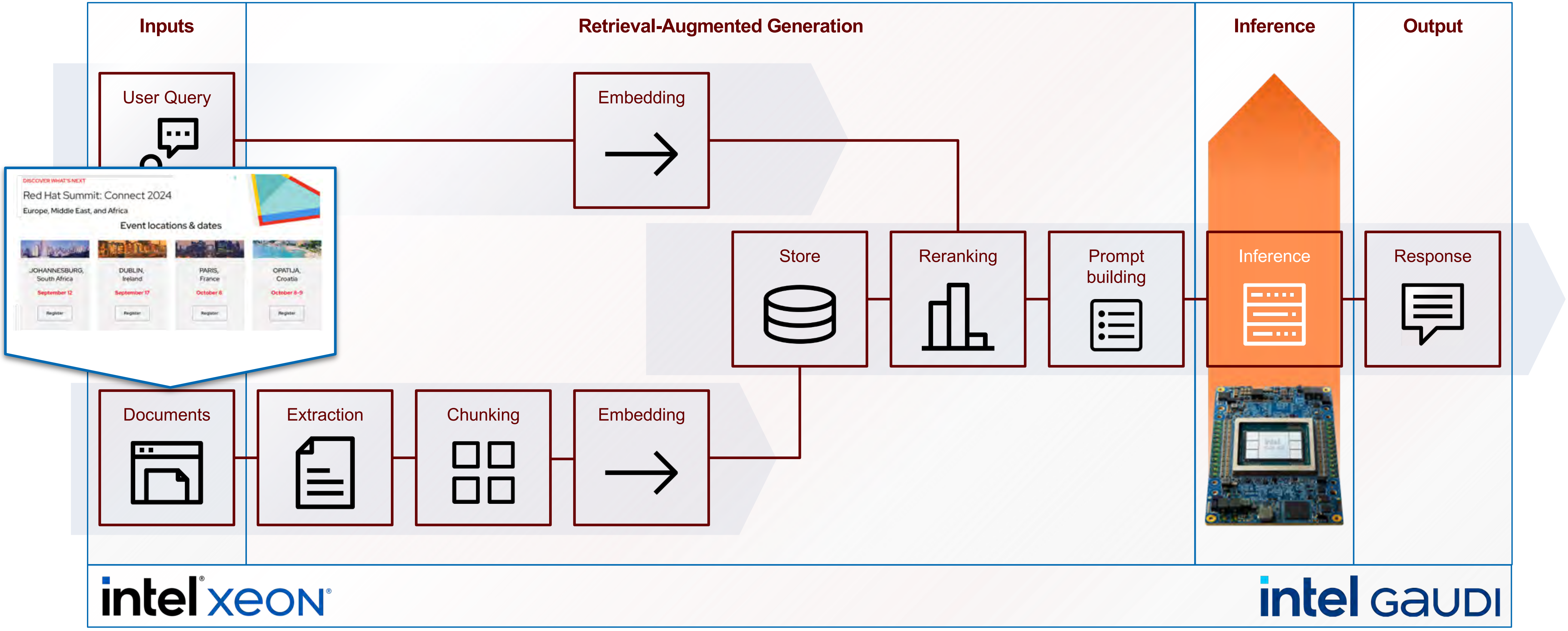
- Prompt tuning
- Retrieval-Augmented Generation (RAG)
- Fine tuning foundation models
- Training a Foundation Model from scratch



Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG)



DISCOVER WHAT'S NEXT

Red Hat Summit: Connect 2024

Europe, Middle East, and Africa

Event locations & dates

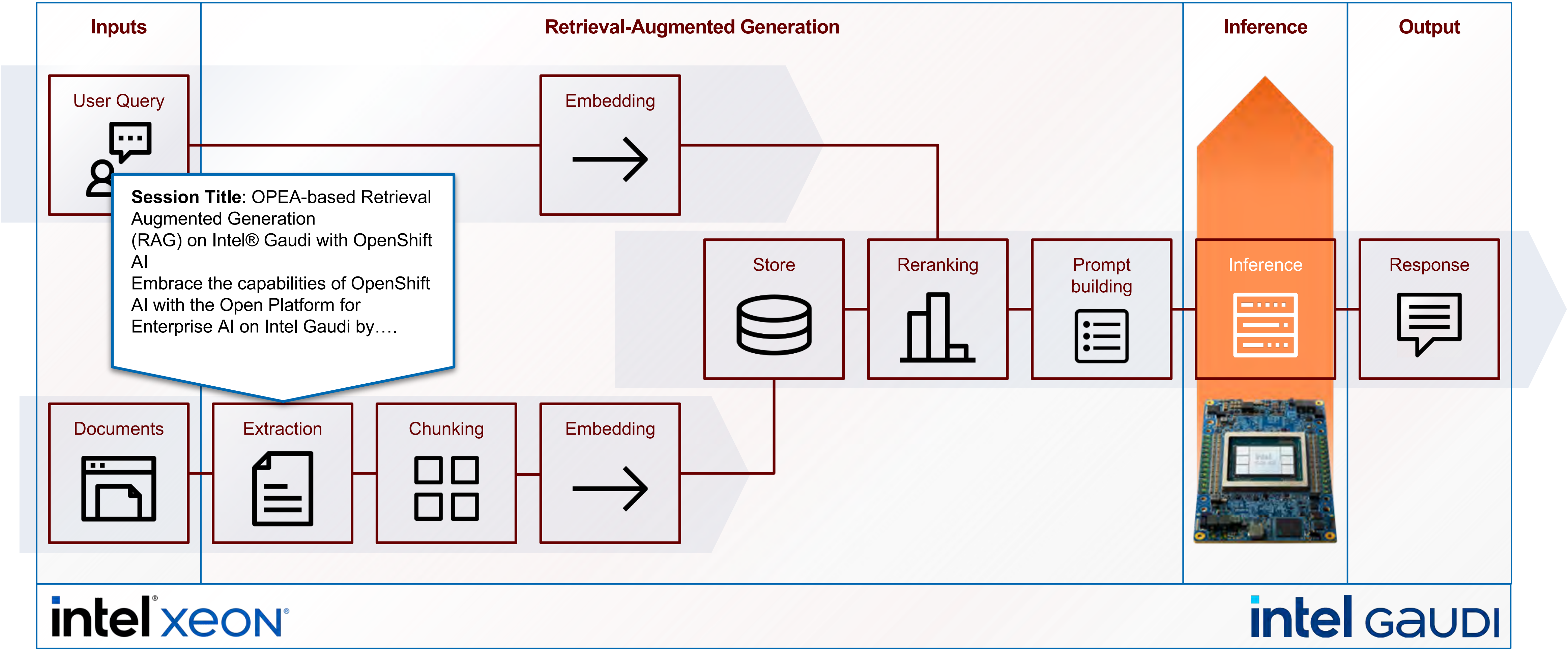
JOHANNESBURG, South Africa — September 12

DUBLIN, Ireland — September 17

PARIS, France — October 8

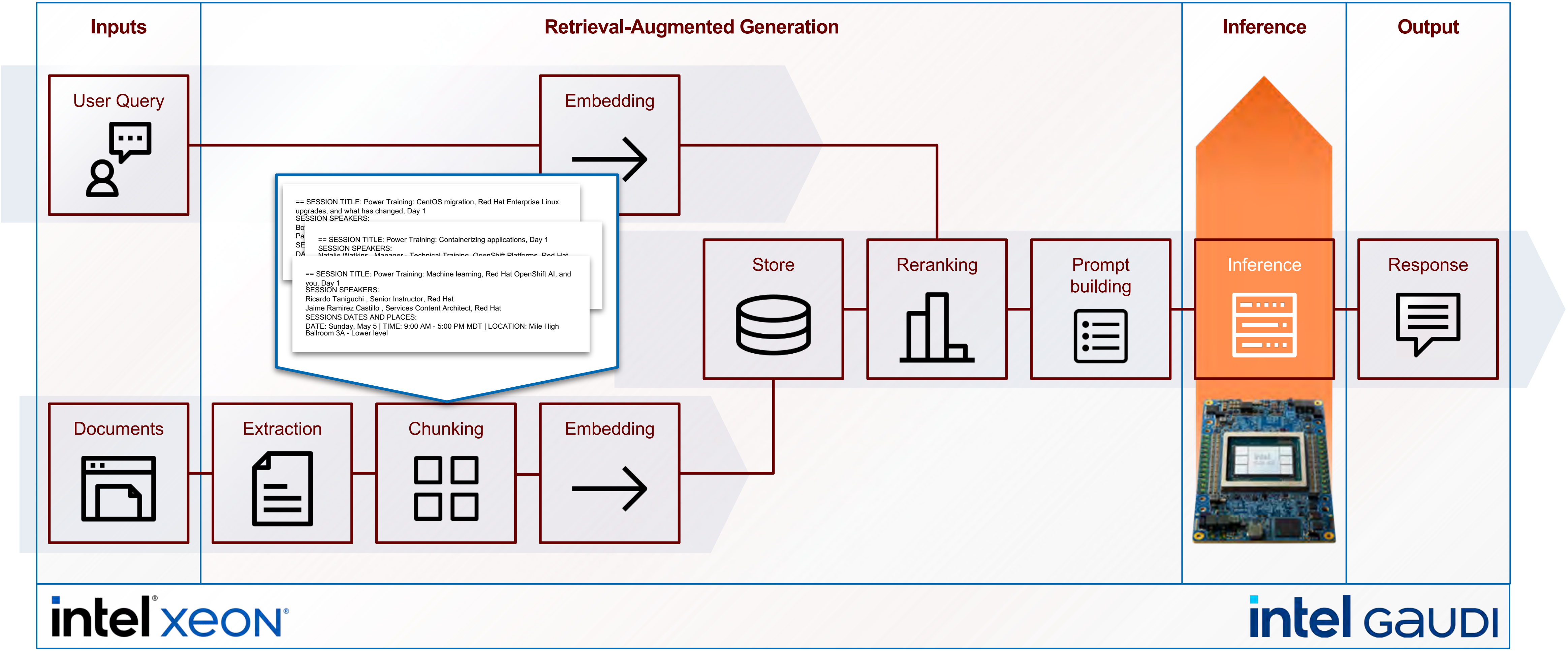
OPATIJA, Croatia — October 8-9

Retrieval Augmented Generation (RAG)



Session Title: OPEA-based Retrieval Augmented Generation (RAG) on Intel® Gaudi with OpenShift AI
Embrace the capabilities of OpenShift AI with the Open Platform for Enterprise AI on Intel Gaudi by….

Retrieval Augmented Generation (RAG)

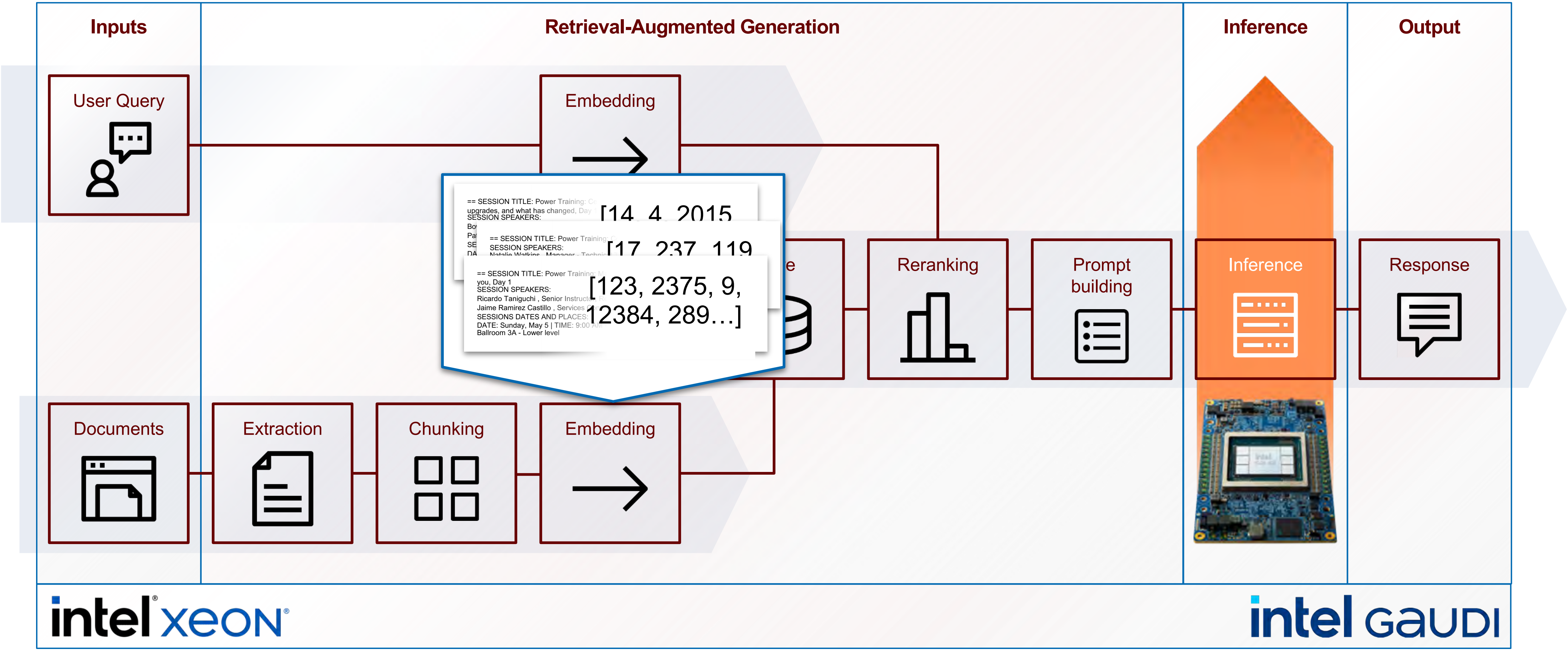


== SESSION TITLE: Power Training: CentOS migration, Red Hat Enterprise Linux upgrades, and what has changed, Day 1
SESSION SPEAKERS:
Bo
Pa
SE
DA

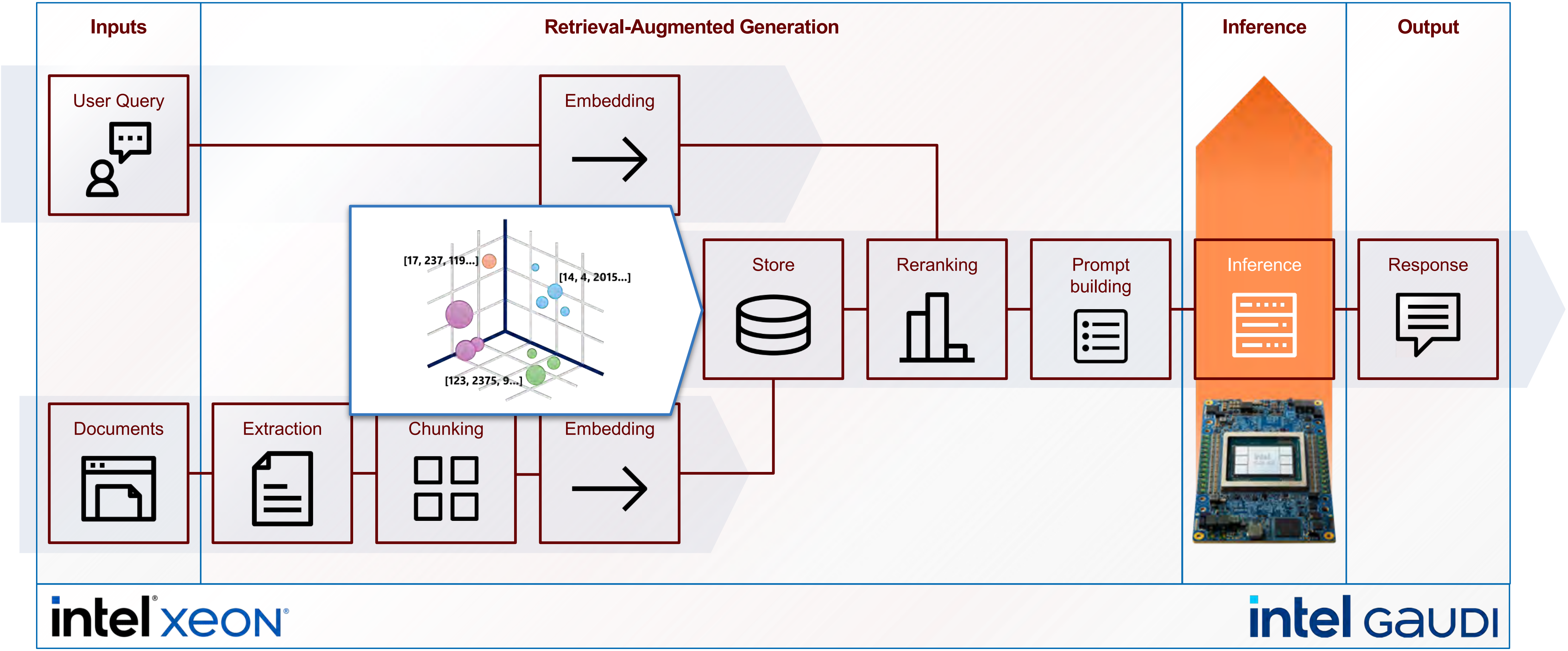
== SESSION TITLE: Power Training: Containerizing applications, Day 1
SESSION SPEAKERS:
Natalie Watkins , Manager - Technical Training, OpenShift Platforms, Red Hat

== SESSION TITLE: Power Training: Machine learning, Red Hat OpenShift AI, and you, Day 1
SESSION SPEAKERS:
Ricardo Taniguchi , Senior Instructor, Red Hat
Jaime Ramirez Castillo , Services Content Architect, Red Hat
SESSIONS DATES AND PLACES:
DATE: Sunday, May 5 | TIME: 9:00 AM - 5:00 PM MDT | LOCATION: Mile High Ballroom 3A - Lower level

Retrieval Augmented Generation (RAG)


Inputs
Retrieval-Augmented Generation
Inference
Output
User Query
Embedding
== SESSION TITLE: Power Training: Ce
upgrades, and what has changed, Day 1
SESSION SPEAKERS:
Bo
Pat
SE
DA
[14, 4, 2015
== SESSION TITLE: Power Training: C
SESSION SPEAKERS:
Natalie Watkins , Manager - Technic
[17, 237, 119
== SESSION TITLE: Power Training: Ma
you, Day 1
SESSION SPEAKERS:
Ricardo Taniguchi , Senior Instructor, Re
Jaime Ramirez Castillo , Services Cont
SESSIONS DATES AND PLACES:
DATE: Sunday, May 5 | TIME: 9:00 AM -
Ballroom 3A - Lower level
[123, 2375, 9, 12384, 289…]
e
Reranking
Prompt building
Inference
Response
Documents
Extraction
Chunking
Embedding
intel xeon
intel GAUDI

Retrieval Augmented Generation (RAG)



Inputs

Retrieval-Augmented Generation

Inference

Output

User Query

Embedding

Store

Reranking

Prompt building

Inference

Response

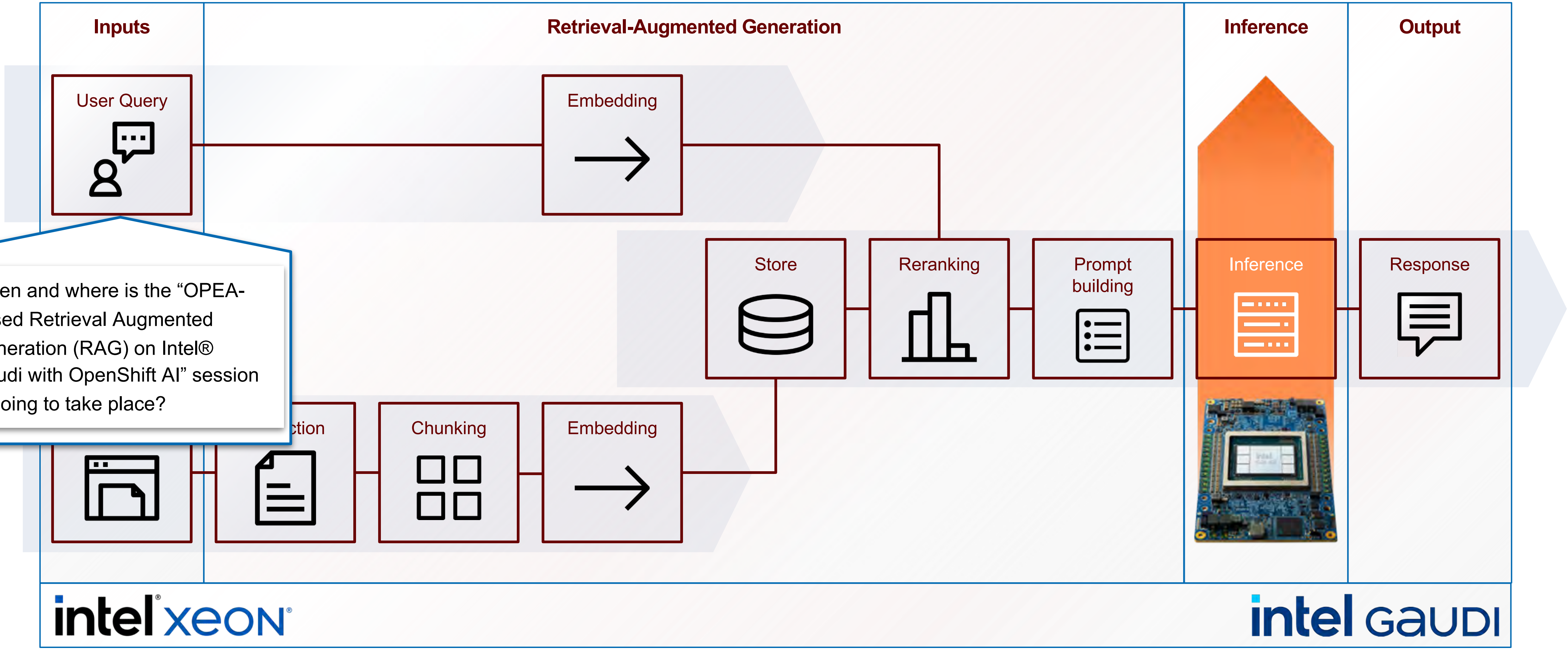
Documents

Extraction

Chunking

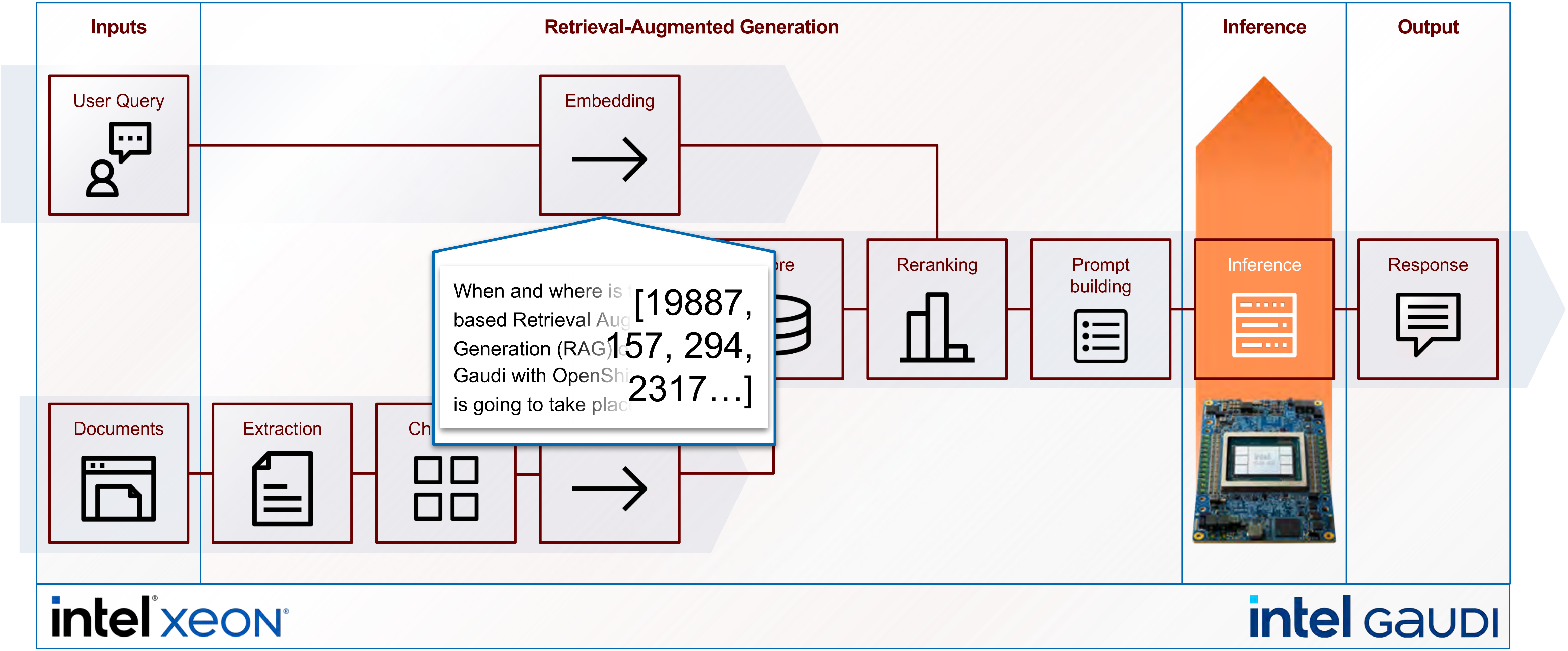
Embedding

Retrieval Augmented Generation (RAG)



When and where is the "OPEA-based Retrieval Augmented Generation (RAG) on Intel® Gaudi with OpenShift AI" session is going to take place?

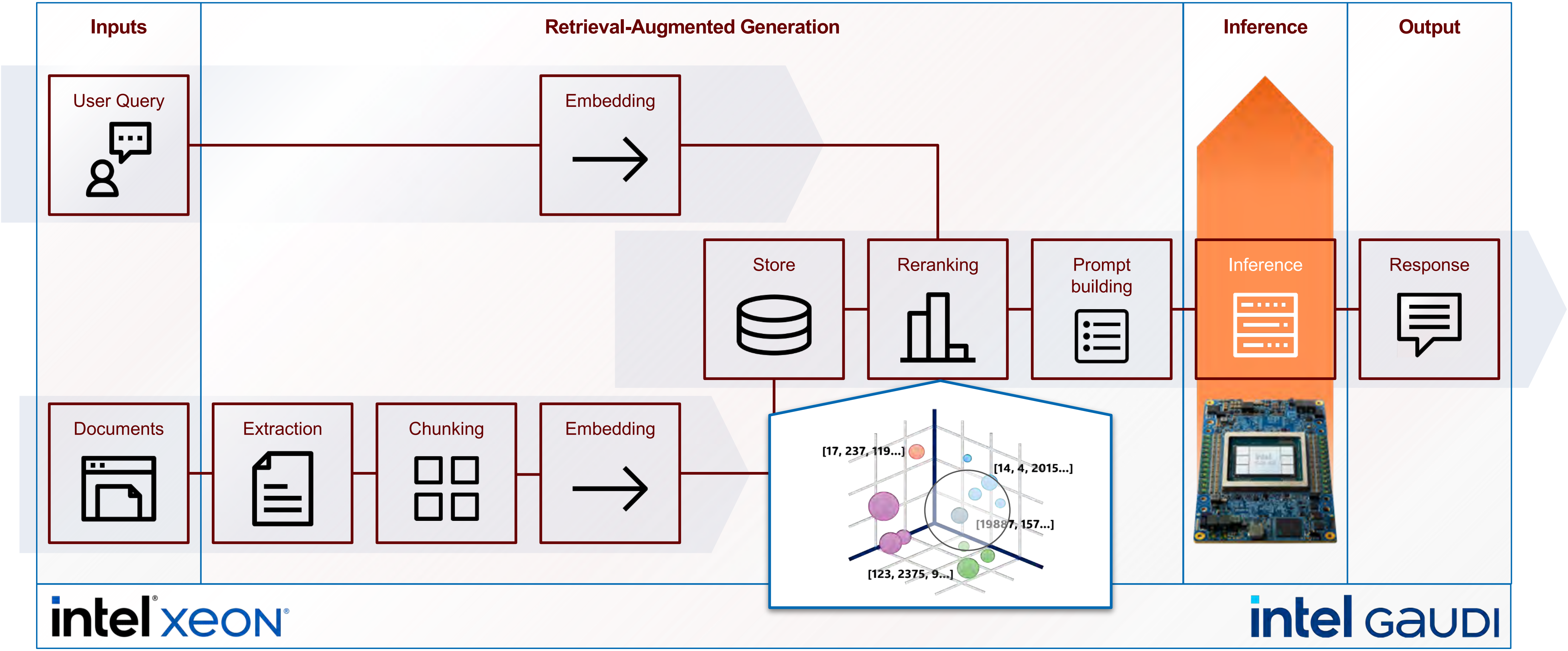
Retrieval Augmented Generation (RAG)



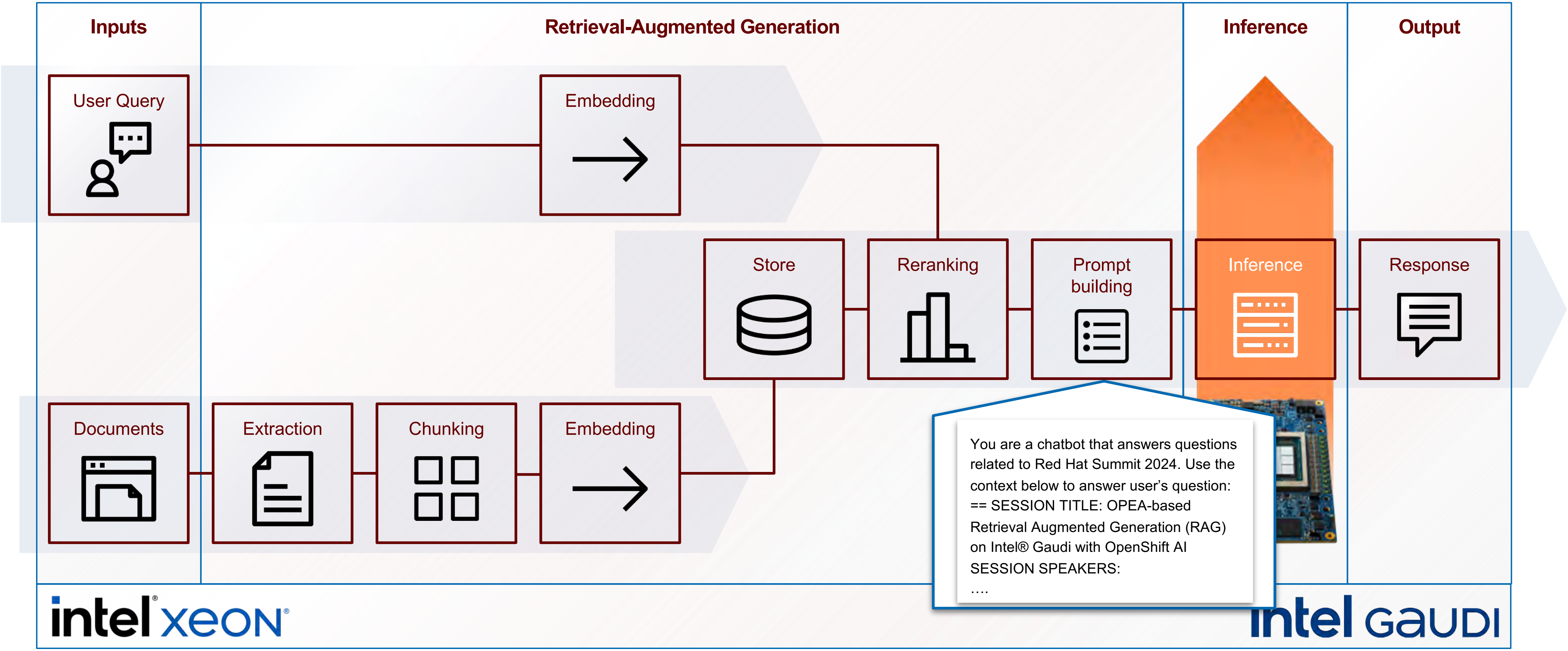
When and where is t
based Retrieval Aug
Generation (RAG) c
Gaudi with OpenShi
is going to take plac

[19887, 157, 294, 2317…]

Retrieval Augmented Generation (RAG)

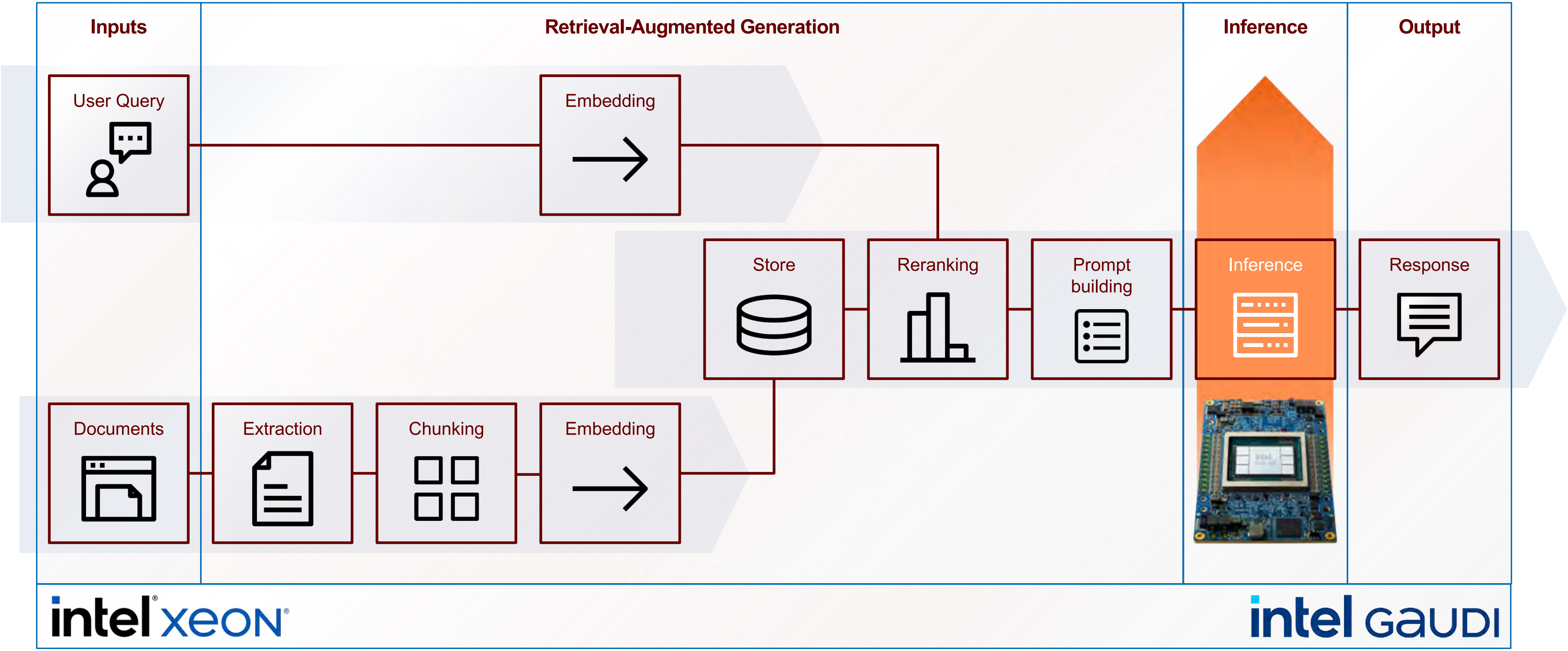


Retrieval Augmented Generation (RAG)

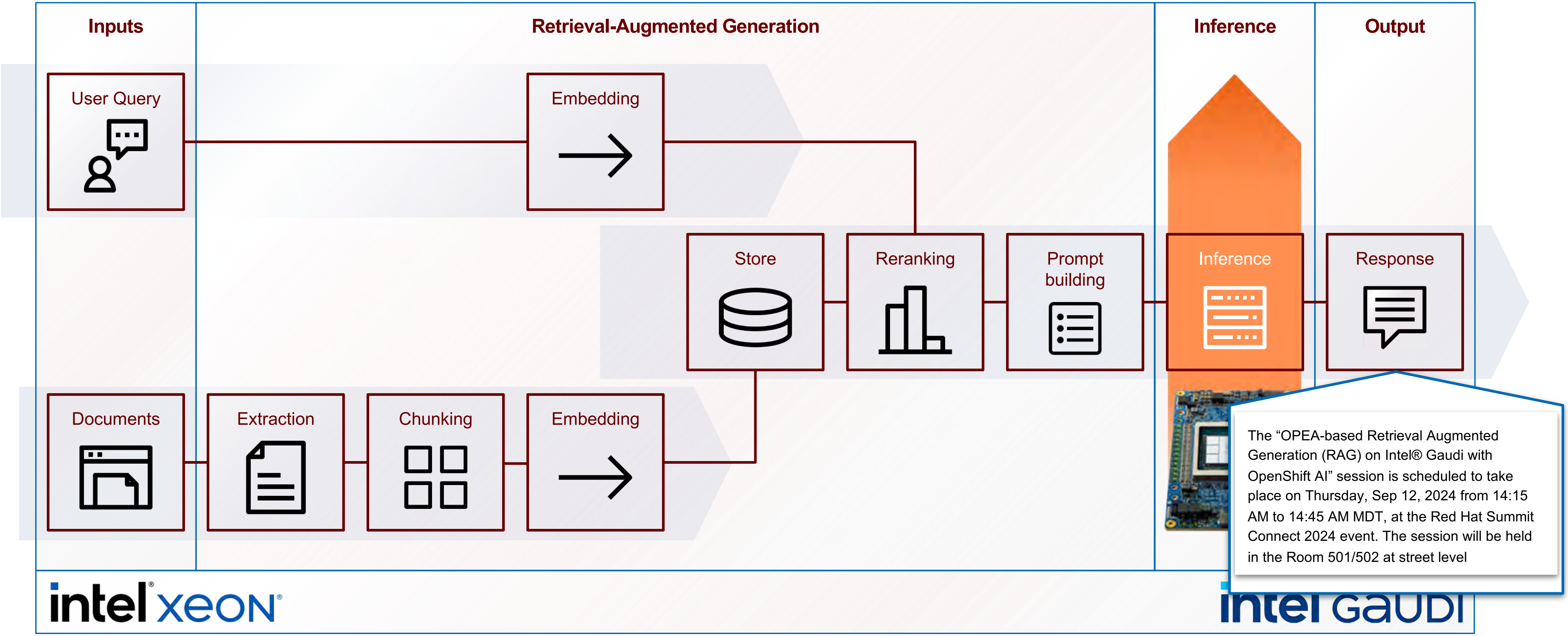


You are a chatbot that answers questions related to Red Hat Summit 2024. Use the context below to answer user's question:
== SESSION TITLE: OPEA-based Retrieval Augmented Generation (RAG) on Intel® Gaudi with OpenShift AI
SESSION SPEAKERS:
….

Retrieval Augmented Generation (RAG)



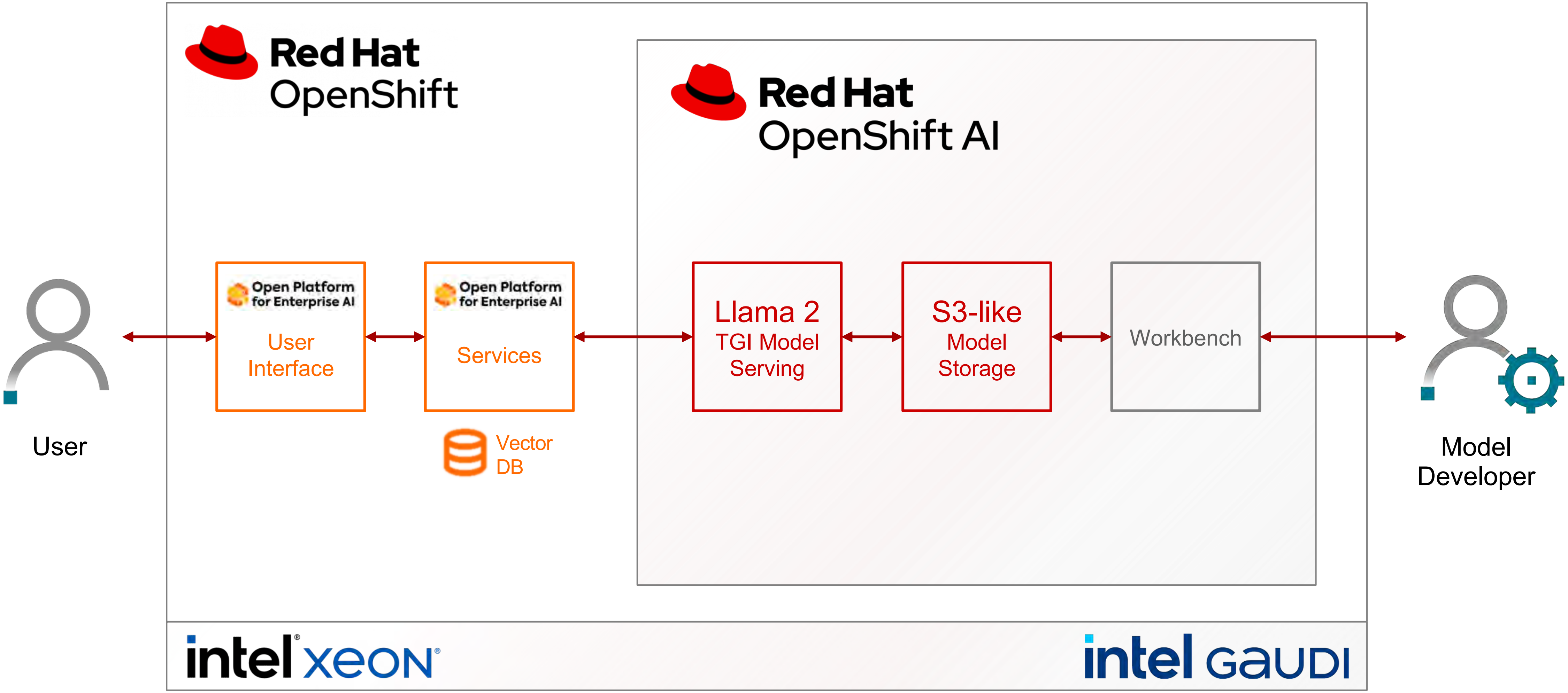
Retrieval Augmented Generation (RAG)

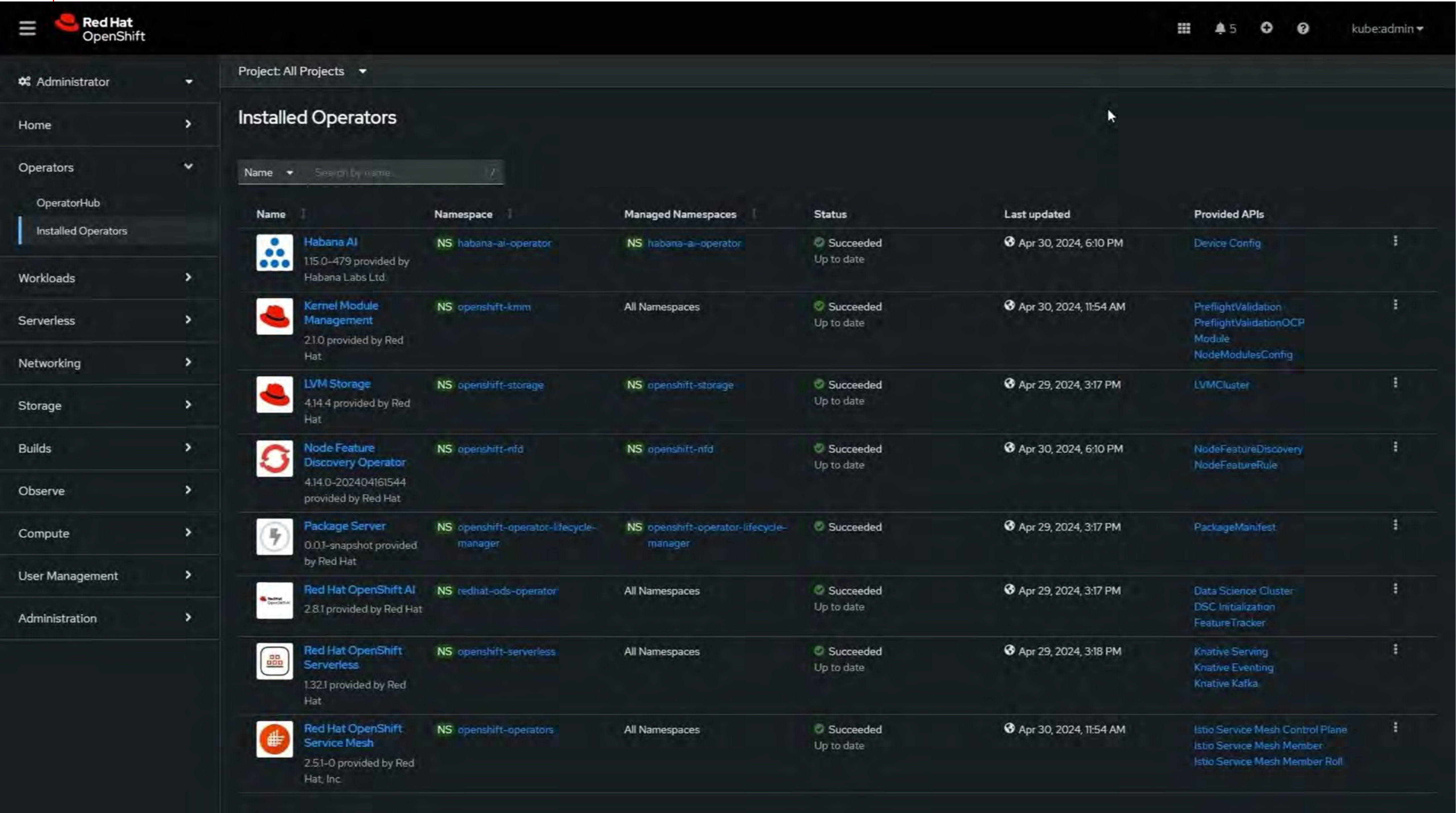


The "OPEA-based Retrieval Augmented Generation (RAG) on Intel® Gaudi with OpenShift AI" session is scheduled to take place on Thursday, Sep 12, 2024 from 14:15 AM to 14:45 AM MDT, at the Red Hat Summit Connect 2024 event. The session will be held in the Room 501/502 at street level

Retrieval Augmented Generation (RAG) Chatbot Demo

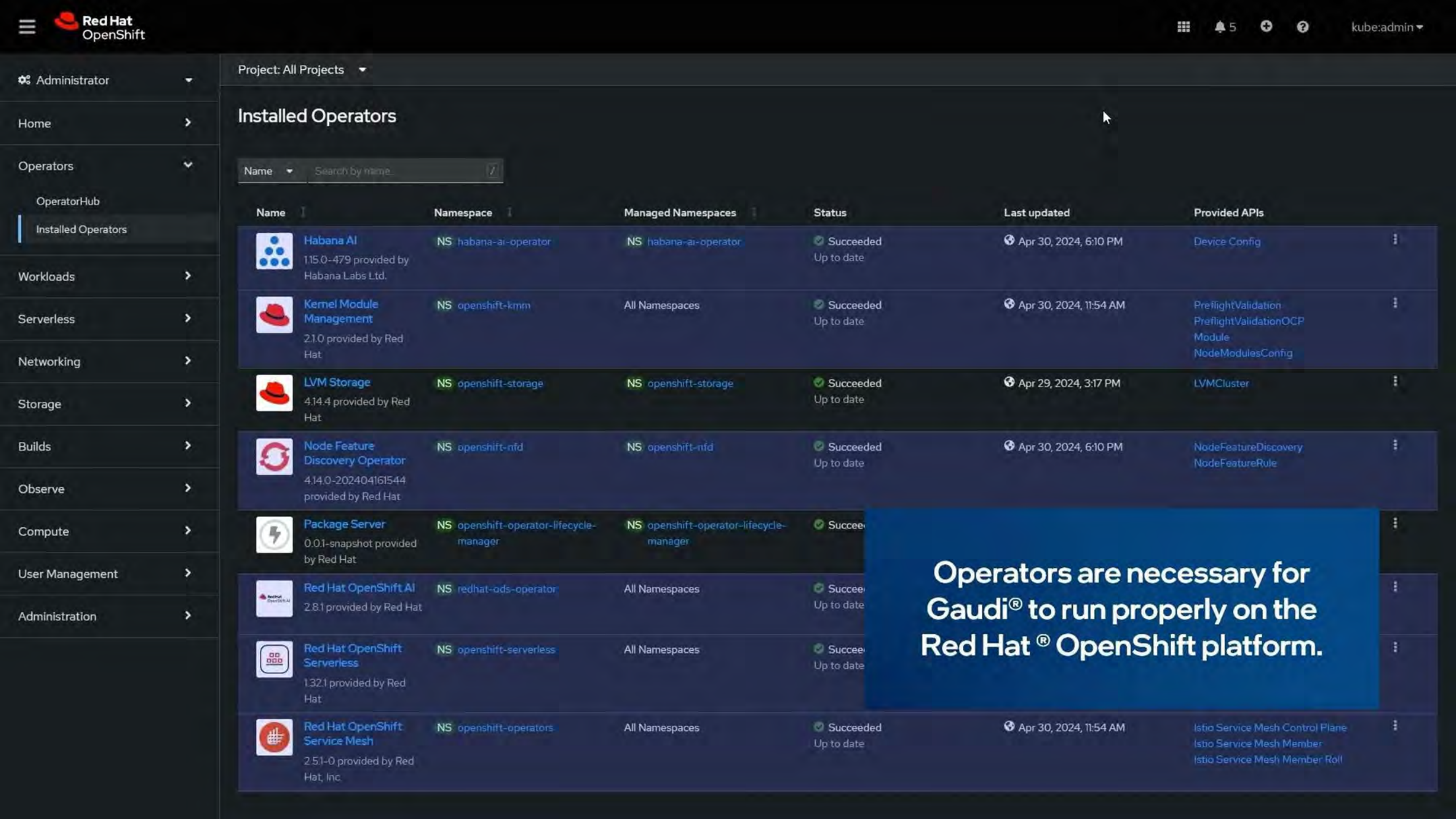
OPEA Chat QnA on OpenShift – demo design





Installed Operators

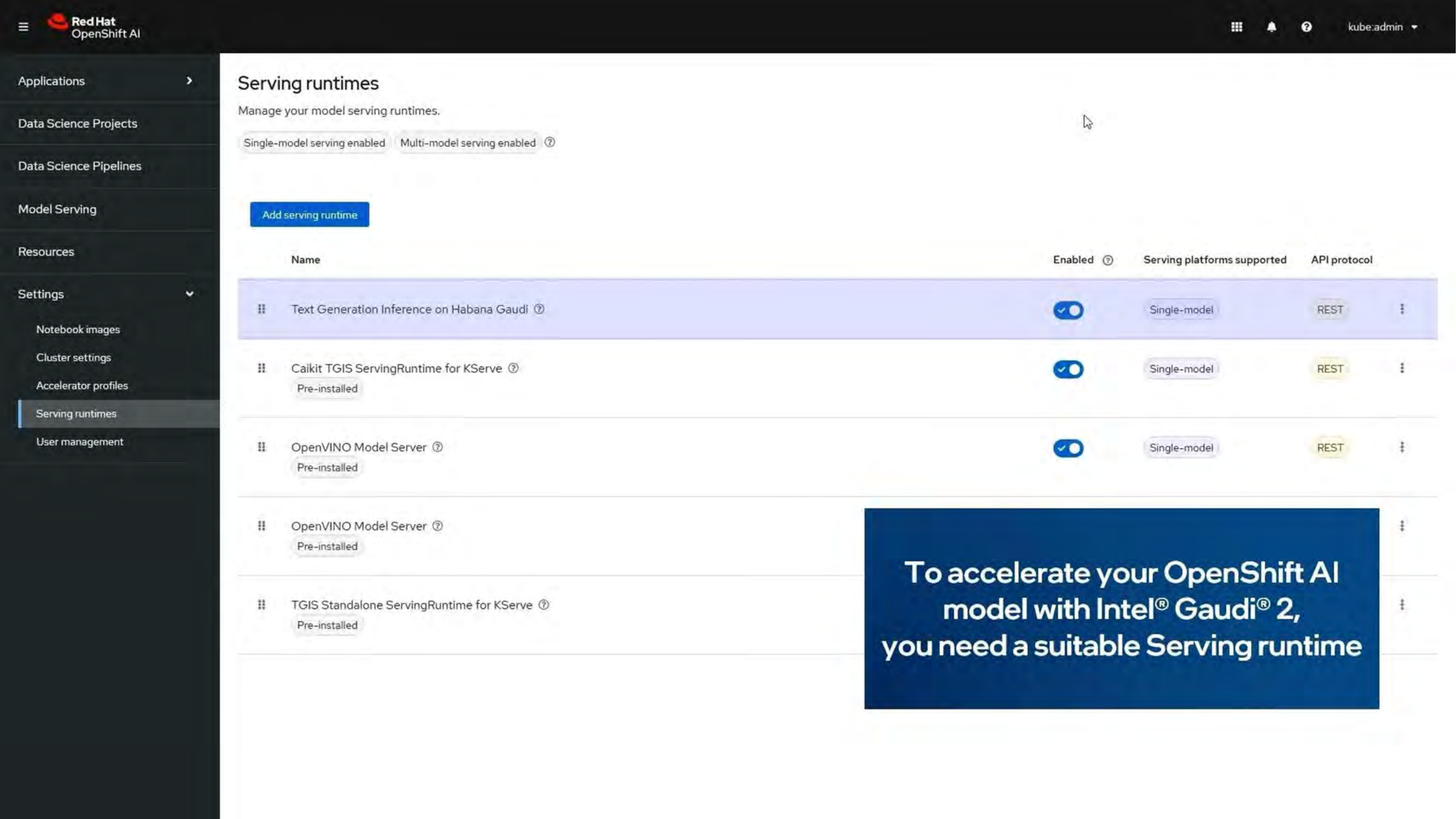
Name	Namespace	Managed Namespaces	Status	Last updated	Provided APIs
Habana AI 1.15.0-479 provided by Habana Labs Ltd.	NS habana-ai-operator	NS habana-ai-operator	Succeeded Up to date	Apr 30, 2024, 6:10 PM	Device Config
Kernel Module Management 2.1.0 provided by Red Hat	NS openshift-kmm	All Namespaces	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



Installed Operators

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.



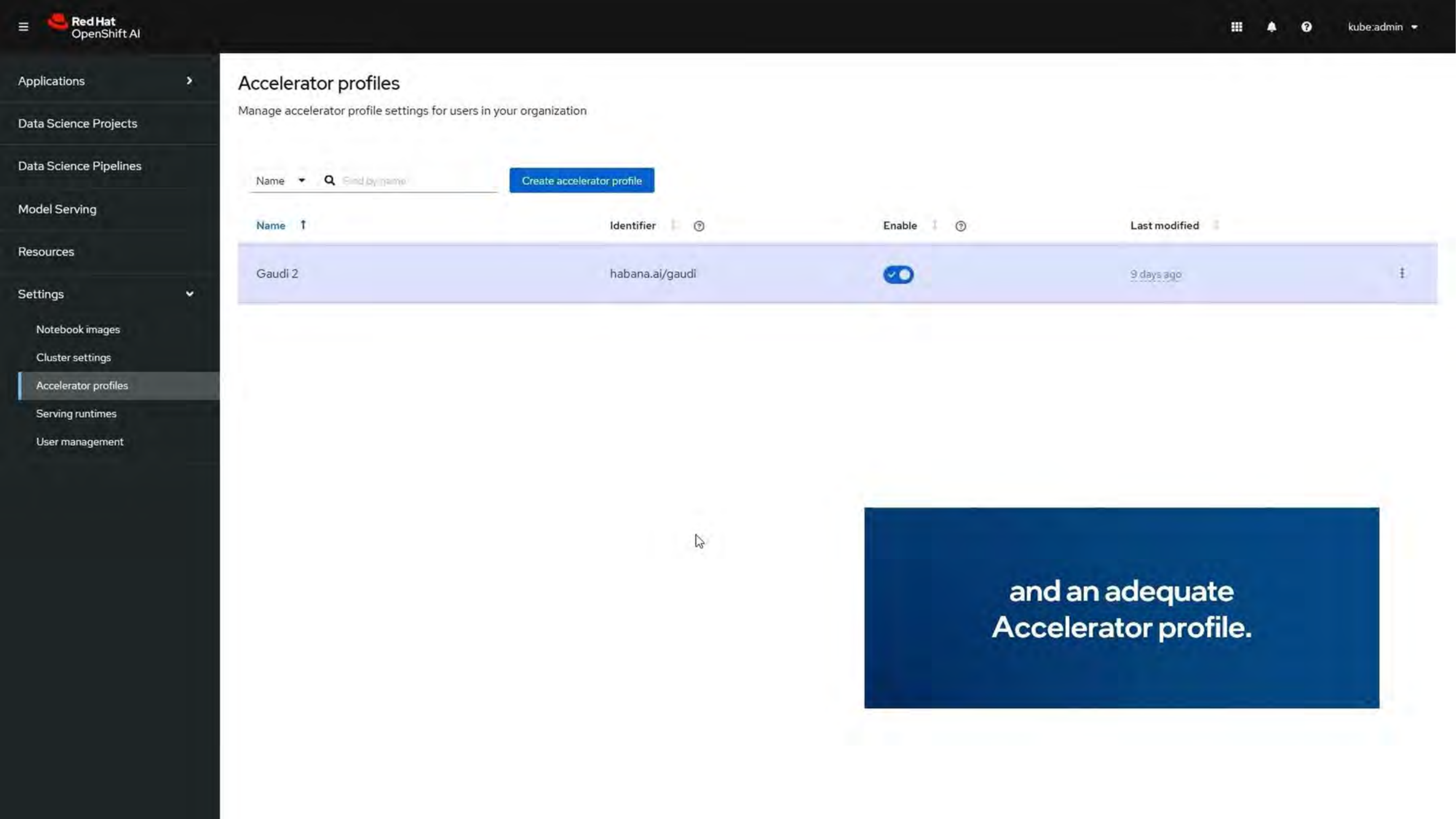
Serving runtimes

Manage your model serving runtimes.

Single-model serving enabled Multi-model serving enabled

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

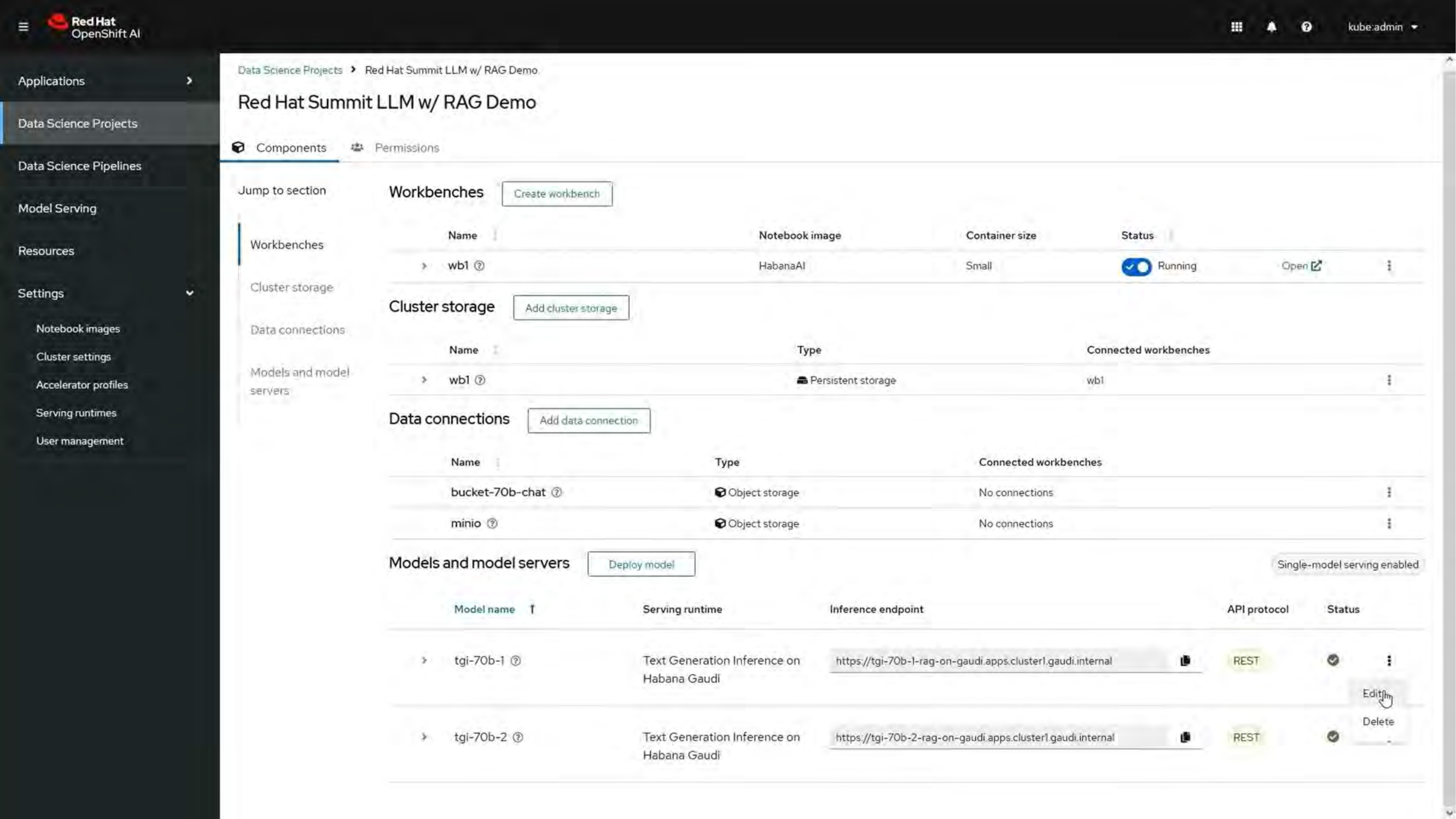


Accelerator profiles

Manage accelerator profile settings for users in your organization

Name	Identifier	Enable	Last modified
Gaudi 2	habana.ai/gaudi	✓	9 days ago

and an adequate Accelerator profile.



Red Hat Summit LLM w/ RAG Demo

Workbenches

Name	Notebook image	Container size	Status
wb1	HabanaAI	Small	Running

Cluster storage

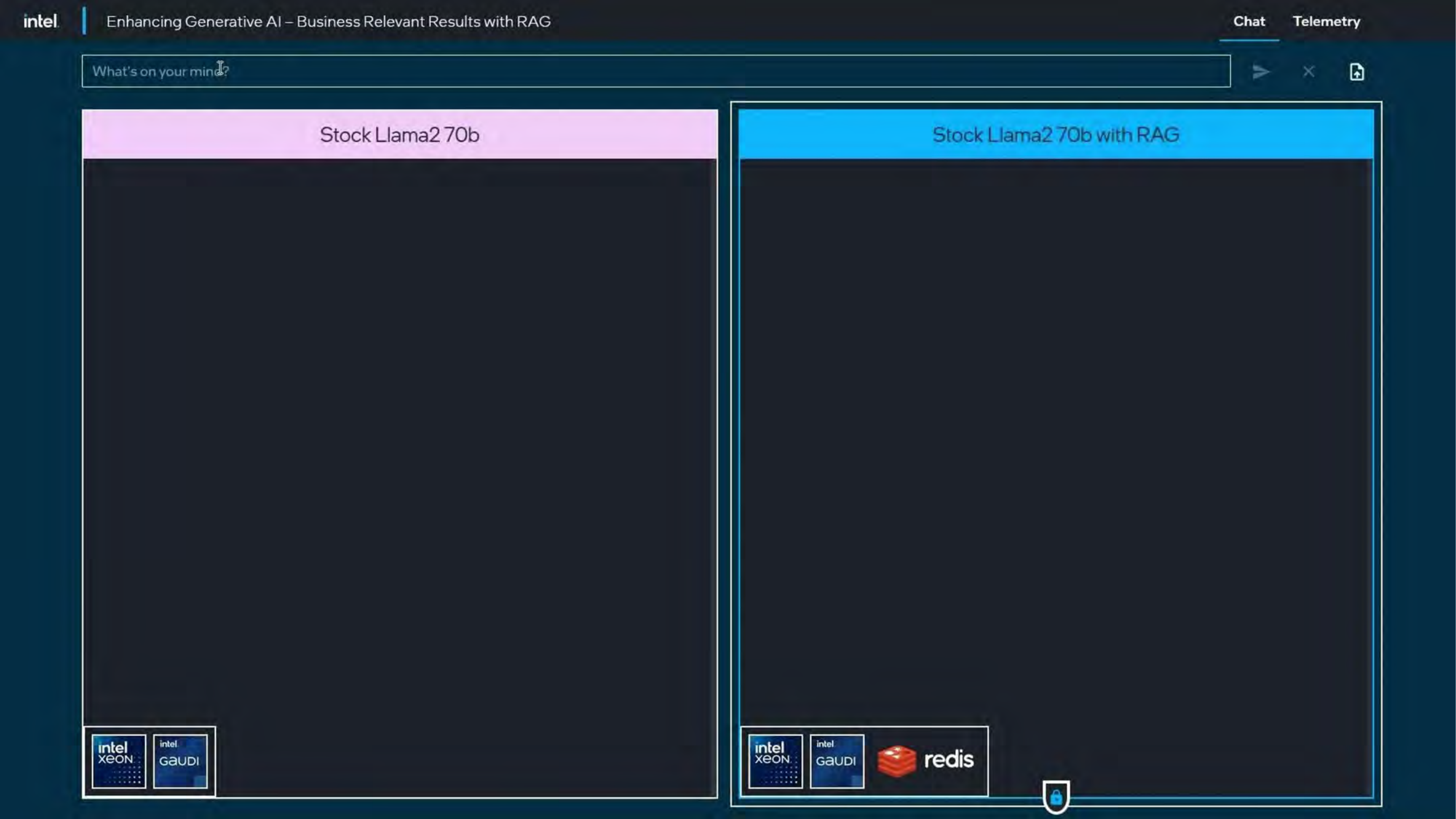
Name	Type	Connected workbenches
wb1	Persistent storage	wb1

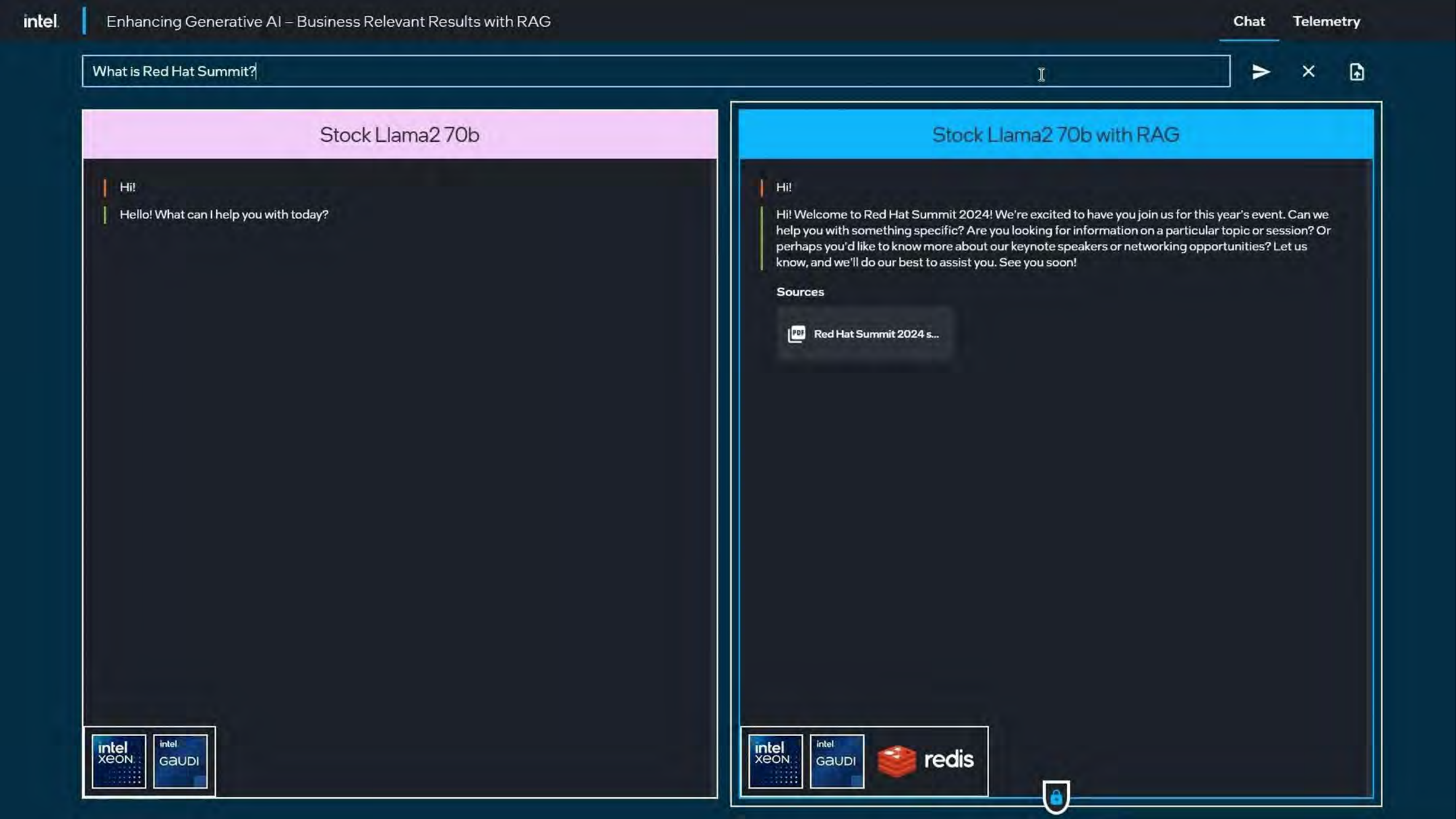
Data connections

Name	Type	Connected workbenches
bucket-70b-chat	Object storage	No connections
minio	Object storage	No connections

Models and model servers

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	✓
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
intel XEON
intel GAUDI
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?

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...

Summary

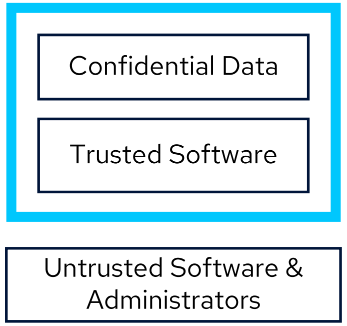
Key Takeaways

- RAG enhances AI development
- OPEA simplifies AI deployment
- OpenShift AI integrates into DevOps workflow
- Intel Gaudi 3 accelerates AI training and inference

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 Data

Trusted Software

Untrusted Software & Administrators

+

Openshift sandboxed containers

The Confidential Containers project

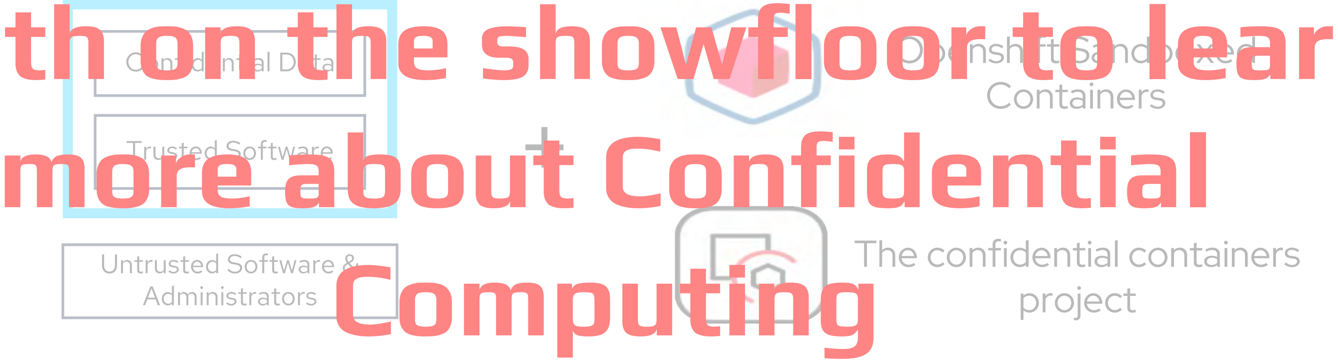
Confidential Computing is about protecting data in-use.
You do not have to trust the system admins of the providers any longer.

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

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



Confidential Data
Trusted Software
Untrusted Software & Administrators
+
OpenShift Sandboxed Containers
The confidential containers project

Learn more!

Learn more!

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

Q&A

Red Hat Summit

Connect

Thank you

in linkedin.com/company/red-hat

f facebook.com/redhatinc

youtube.com/user/RedHatVideos

twitter.com/RedHat