

Summit Recap

Q2 Michigan RHUG

Richard Hofmeister
App Platform SSA



Red Hat AI

- **Red Hat AI Inference Server**
  Available standalone or as part of Red Hat Enterprise Linux AI (RHEL AI) and Red Hat OpenShift AI
- **General Availability of Red Hat OpenShift Lightspeed**
  Offers flexible integration with model providers like OpenAI, Azure OpenAI, and watsonx, plus private AI options
- **Third-Party Validated Models for Red Hat AI, NVIDIA partnership**
  Come with deployment guidance and reproducible outcomes
- **llm-d Open Source Community**
  Builds on vLLM and Inference Gateway projects
- **Red Hat supported LLama Stack and validated LLama for Red Hat AI**
  Day 0 support for Llama models on vLLM and Red Hat AI

Red Hat

The World Changed in November 2022

ChatGPT woke the world up to the power of generative AI

CTECH 24/7 Buzz Startups VC @work Innovation Opini

by CALCALIST Homepage > Events > Tokyo 2023

Israel-Japan Conference

"The AI revolution, and especially generative AI, will completely change the world as we know it"

Takenaka Heizo, Professor Emeritus at Keio University, was speaking with Eyal Agmoni, Founder of Chartered Group, during the Israel-Japan Conference in Tokyo

Meir Orbach 11:02, 30.05.23

TAGS: Generative AI AI Israel-Japan Conference

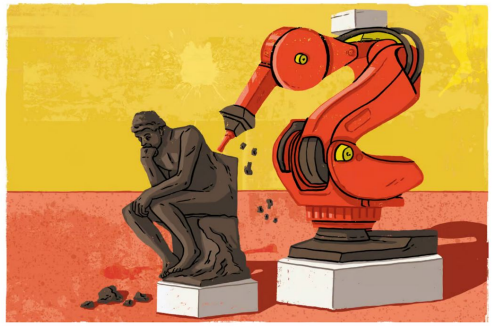


Eyal Agmoni and Takenaka Heizo (Shmulik Dudfur)

"The AI revolution, and especially generative AI, will completely change the world as we know it," said Takenaka Heizo, Professor Emeritus at Keio University, during a talk with Eyal Agmoni, Founder of Chartered Group during the Israel-Japan

TechFinitive INNOVATE COLLABORATE DATA PROTECT HARDWARE RESOURCES JOB BOARD

Home > Features > Innovate > ChatGPT is the real deal – and it's going to change the world



ChatGPT is the real deal – and it's going to change the world

Written by James O'Malley

It's strange how fast technol

Think about the last time you that somehow, a flying meta been unimaginable not so l yourself into a cramped seat

Of course, like everyone, you dropping advances in gener

Featured

CUTTING EDGE
Midjourney 5.2 delivers infinite zoom alongside powerful tools to add extra style and variety
26 Jun 2023

CUTTING EDGE
5G Standalone: The UK gets the real thing at last
23 Jun 2023

CUTTING EDGE
OpenAI's ChatGPT Plus launches for $20 per month
1 Feb 2023

CUTTING EDGE
China's largest search engine Baidu is developing ChatGPT-style AI
30 Jan 2023

BUSINESS INSIDER

HOW AI IS CHANGING EVERYTHING

Events Video Special Issues Jobs **VentureBeat** Subscribe GamesBeat

Artificial Intelligence Security Data Infrastructure Automation Enterprise Analytics More

Guest

The language revolution: How LLMs could transform the world

Shyamal Anadkat, OpenAI @shyamalanadkat

February 11, 2023 8:20 AM



Image Credit: Getty Images

Join enterprise leaders in San Francisco

Learn to leverage the power of Generative AI

Learn More

Join top executives in San Francisco on July 11-12 and learn how business leaders are getting ahead of the generative AI revolution. Learn More

We are living at a historic moment. A new revolution, comparable to the Industrial Revolution, is underway. Entire industries are going to be disrupted. The nature of creativity and knowledge work is going to change. Language is going to become the most important sense for humans. Language — specifically in the form of large language models (LLMs) — is going to reshape how we think about the world around us.

Privacy

Harvard Business Review Subscribe Sign In

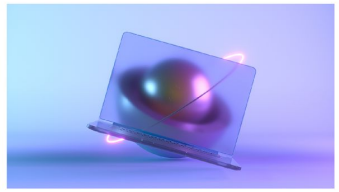
Latest Magazine Ascend Topics Podcasts Video Store The Big Idea Data & Visuals Case Selections

Technology And Analytics

Generative AI Will Change Your Business. Here's How to Adapt.

by David C. Edelman and Mark Abraham

April 12, 2023



Summary. Generative AI can "generate" text, speech, images, music, video, and especially, code. When that capability is joined with a feed of someone's own information, used to tailor the when, what, and how of an interaction, then the ease by which someone can get things... more

It's coming. Generative AI will change the nature of how we interact with all software, and given how many brands have significant software components in how they interact with customers, generative AI will drive and distinguish how more brands compete.

Harvard Business Review

Welcome to HBR!

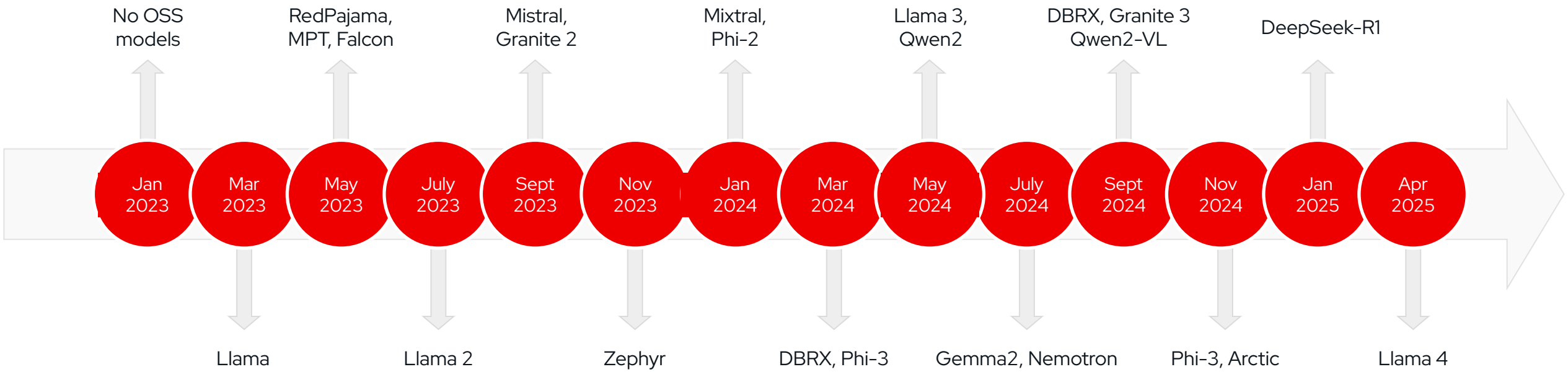
Register for more solutions, ideas, and inspiration

Register Now

Red Hat

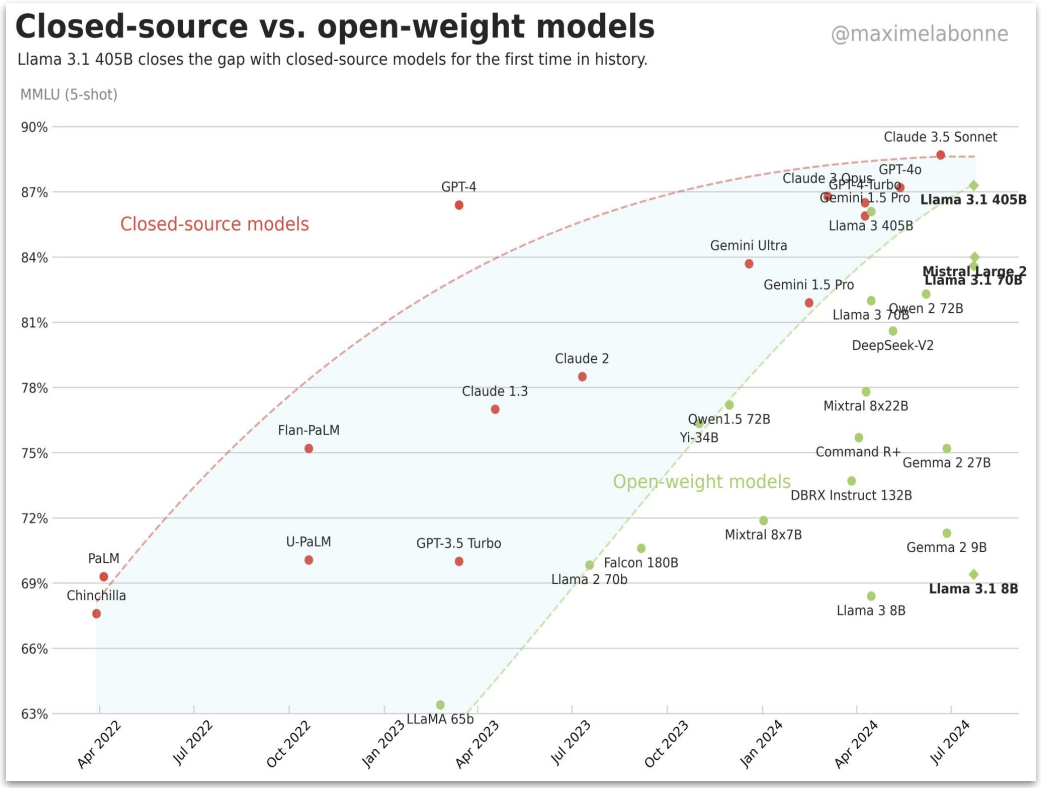
The Power of Open

There has been an explosion of capability from open-source over the last 2 years.



The Power of Open

Open models are deployment targets **today**. And the trend is not slowing down.



Headlines

Meta

Llama

- 650M downloads in 2024
- 85,000 Llama derivative models
- 1B, 3B, 8B, 70B, 405B variants
- Multilingual, Multimodal, Mobile

---

deepseek

R1

- First reasoning model on par in quality with OpenAI O1
- 1-70B parameter distilled versions
- Global market pandemonium?

Models are commoditizing → many options for diverse enterprise needs.

Advantages of Open Source Models

Open-source models play an important role in the Enterprise AI landscape



Customization

Improve accuracy and costs with task specific tuning



Security

Complete data privacy (no 3rd party APIs)



Control

Model lifecycle (no changes to the model in place) and Resources (no rate limits / API downtime)



Cost

Self managed infrastructure. 1B-405B size - match task difficulty to model



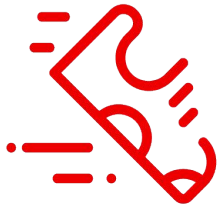
Red Hat

Inference is becoming the gravity in AI because it is where the real world value happens



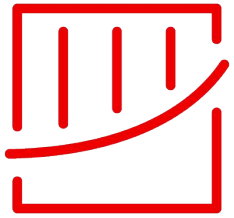
Always-on intelligence

Inference powers every user interaction 24/7, making inference the constant cost and performance driver



Latency defines experience

Low-latency and fast inference ensures optimal user experiences, essential for real-time applications and user retention



Exponential market momentum

The AI inference market is forecast to grow from USD 106.15 billion in 2025 to USD 254.98 billion by 2030, underscoring its role

Source:
What is AI Inference?, Jan 2025, Red Hat
AI Inference Market Size, Share & Trends, Feb 2025, Markets and Markets
Amazon Found Every 100ms of Latency Cost them 1% in Sales, July 2023, Gigaspaces

Fast

Cost-effective

Optimized



Red Hat AI
Inference Server

Any gen AI model

Any AI Accelerator

Any environment

Red Hat AI Inference Server

Gain consistent, fast and cost-effective inference at scale



Inference runtime for the hybrid cloud

Run your models of choice across any accelerator and any environment



Compress Models

Reduce compute and costs while preserving accuracy



Red Hat AI Hugging Face repository

Access a collection of third-party validated and optimized models ready for inference.



Certified for all Red Hat products

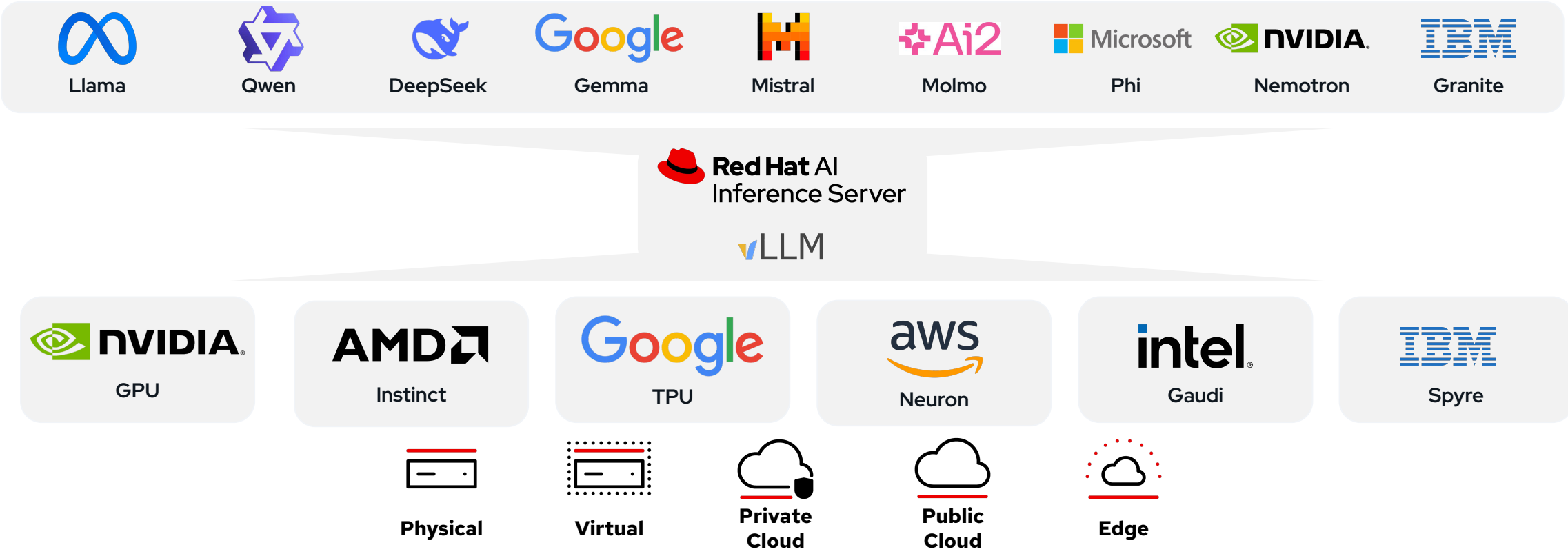
Deployable across non-Red Hat Linux and Kubernetes platforms

Red Hat AI Inference Server is included in OpenShift AI and RHEL AI

Red Hat

Red Hat AI Inference Server

vLLM connects model creators to accelerated hardware providers



Single platform to run any model, on any accelerator, on any cloud

Red Hat

Red Hat AI Inference Server

vLLM is emerging as the Linux of GenAI Inference

HIGH PERFORMANCE

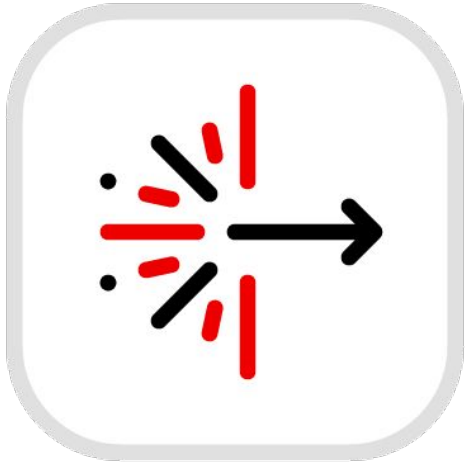
- Advanced algorithms for high QPS serving
- Single server/GPU to distributed/multi GPU
- Already comparable to Nvidia (TRT-LLM)

EASY TO USE CAPABILITIES DRIVING DEVELOPER AND IT PRODUCTIVITY

- Native Hugging Face integration
- Simple APIs for online and offline inference
- OpenAI-compatible API protocol

Scalable inference across the hybrid cloud

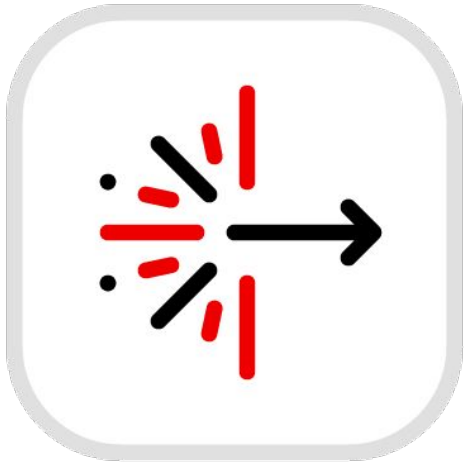
Red Hat



OpenShift Lightspeed exists to help improve productivity and accessibility for OpenShift users of all skill levels.

Red Hat

OpenShift Lightspeed



Help where you need it

Integrated directly into the Red Hat OpenShift web console

Troubleshoot, investigate and learn more

Provides assistance with troubleshooting and investigating cluster resources

Generative AI

Powerful, pluggable LLMs combined with the latest OpenShift documentation

Red Hat

Benefits for the entire team

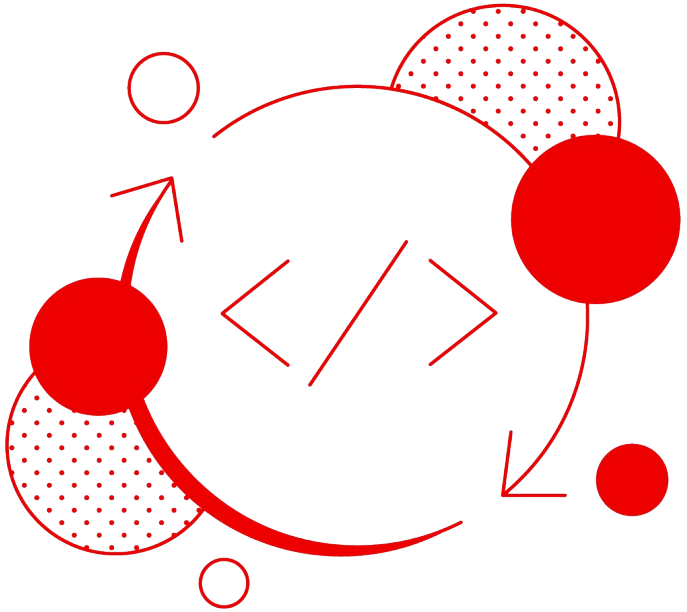
Regardless of skill level

Novices

+ Can reduce the OpenShift learning curve
+ Can provide step-by-step guidance
+ Can help with troubleshooting errors
+ Can increase skill development
+ Can help build confidence

Experts

+ Can increase productivity for day-to-day operations
+ Can help with troubleshooting errors



Red Hat

RHEL 10

- **Red Hat Image Mode**
  Immutable operating system delivered as a bootable container image
- **Post-Quantum Cryptography**
  Quantum-resistant algorithms to protect against future attacks
- **RHEL Lightspeed**
  Integrates gen AI directly into the RHEL command line
- **Red Hat Extension Repository**
  Access to new technologies, validated by Red Hat
- **WSL and Podman Desktop for RHEL**
  Support developers working with RHEL anywhere

Red Hat

Red Hat Enterprise Linux 10 will help you...



Address the Linux skills gap
with decades of Red Hat's Linux knowledge and expertise

Contain drift and accelerate delivery
with container tools and technologies

Make better decisions at build time
when it's typically easier and cheaper to make changes

Resist security attacks from hackers
when quantum computers become prevalent

Leverage Red Hat Enterprise Linux as a trusted AI foundation
with an extensive ecosystem of trusted partners and tools

Addressing the Linux skills gap

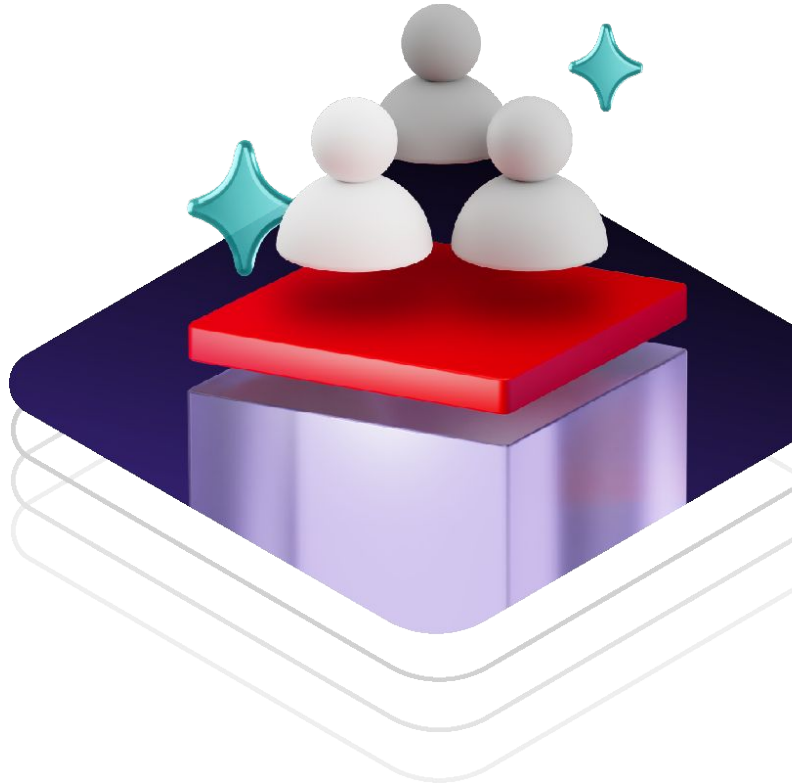
with decades of Red Hat's Linux knowledge and expertise

Red Hat Enterprise Linux Lightspeed

combines decades of Red Hat Enterprise Linux expertise with AI technologies to proactively inform and simplify how both newer and experienced IT professionals build, deploy, and manage Red Hat Enterprise Linux.

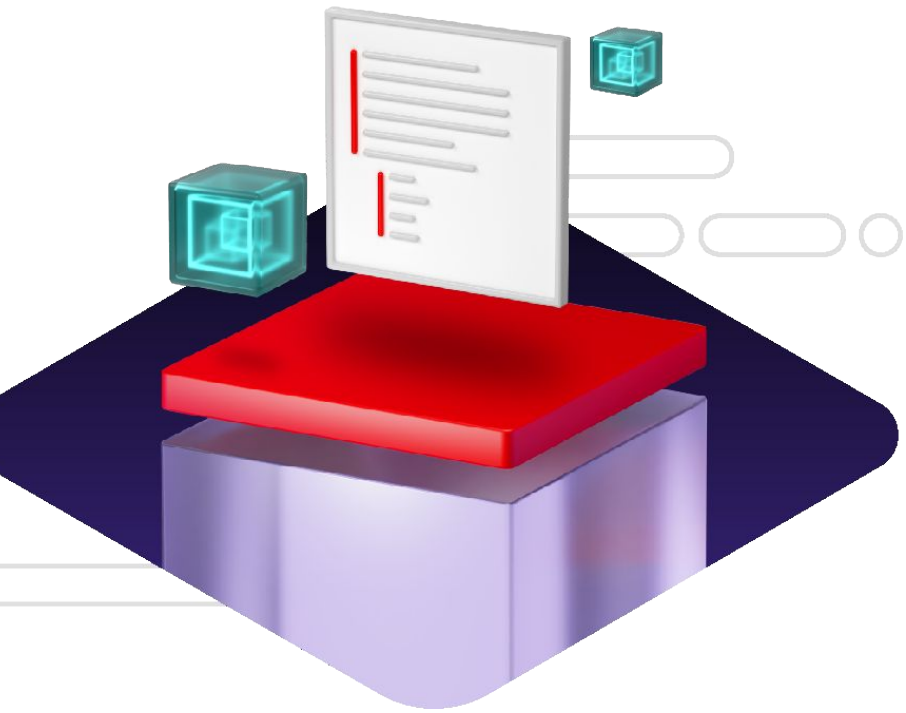
- Use plain language to simplify the way you interact with Red Hat Enterprise Linux
- Make better decisions with recommendations and actionable guidance

Simplify tasks. Amplify results.



Contain drift and accelerate delivery

using container tools and technologies



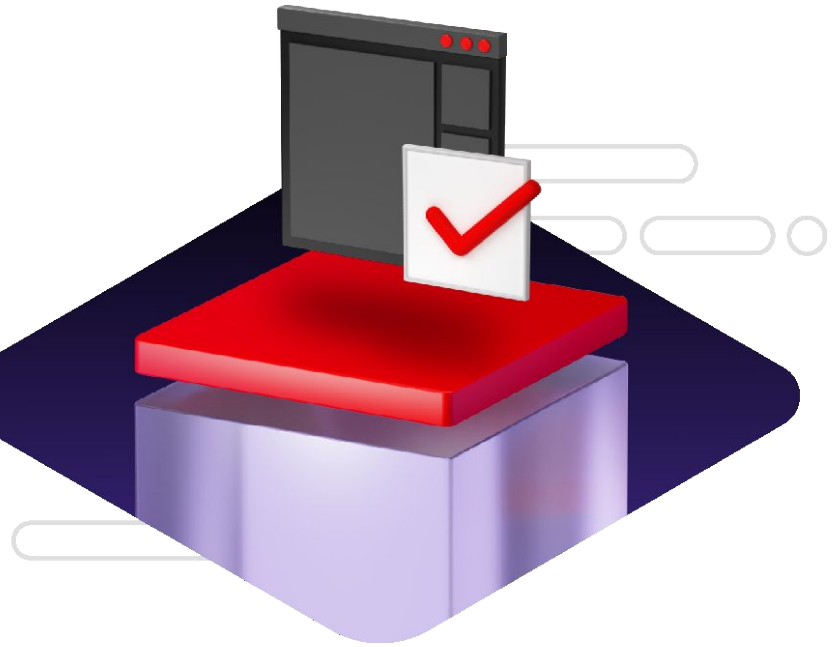
With image mode for Red Hat Enterprise Linux, you can:

- **Speed time to market**
  using DevOps and CI/CD practices, which now include the OS
- **Streamline operations**
  by automating updates and rollbacks—just like your smartphone
- **Enhance security**
  by reducing your attack surface with immutable system images
- **Simplify appliance creation**
  by combining the OS with apps and drivers for faster development and delivery

Because systems should be as easy to update as smartphones

Make better decisions at build time

when it's typically easier and cheaper to make changes



- "Shift left" with new capabilities in Red Hat Insights
  - Package recommendations powered by Red Hat Enterprise Linux Lightspeed
  - Red Hat Insights planning (Lifecycle and roadmap details)
- Use pre-hardened images with image mode to drastically reduce time required for configuring, domain-join etc.
  - Applying guardrails at build time helps users remain secure and compliant
- Securely build and automatically generate SBOM artifacts to easily prove a secure supply chain process

Smart from the start. Fast to the finish.

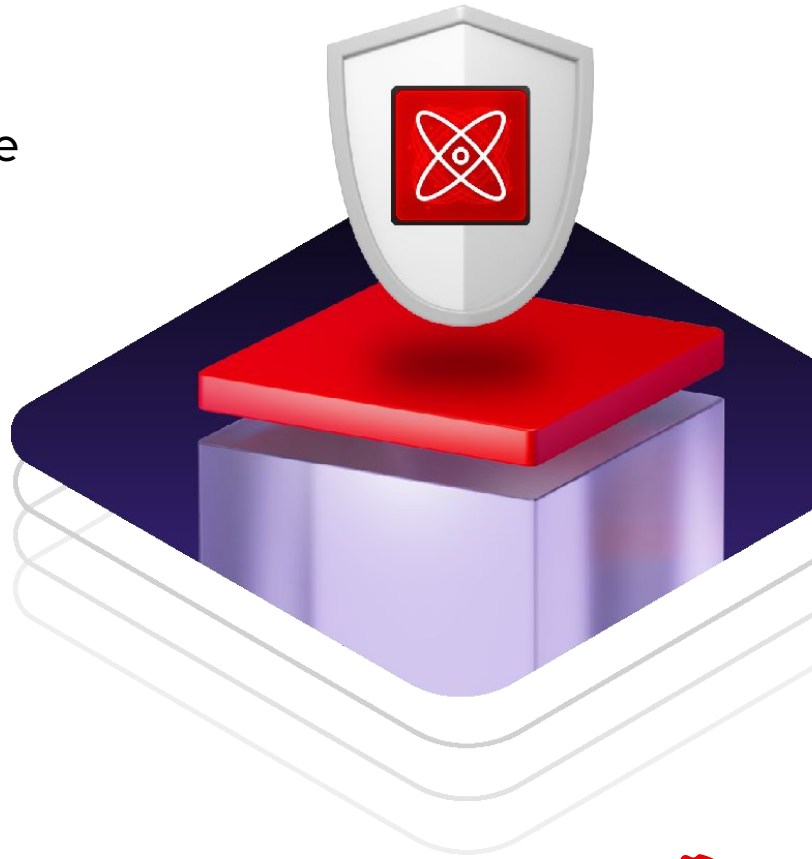
Resist security attacks from hackers

when quantum computers become prevalent

- Red Hat Enterprise Linux 10 is the first enterprise Linux distribution to be post-quantum capable with new quantum-resistant algorithms (and more to come) so you can prepare now for future compliance mandates
  - OpenSSL
  - ML-KEM (FIPS 203)
  - ML-DSA (FIPS 204)
- A new FIPS module

  simplifies development of FIPS-compliant app deployments on containers, VMs or bare metal
- FIPS cryptographic standards can be validated separately

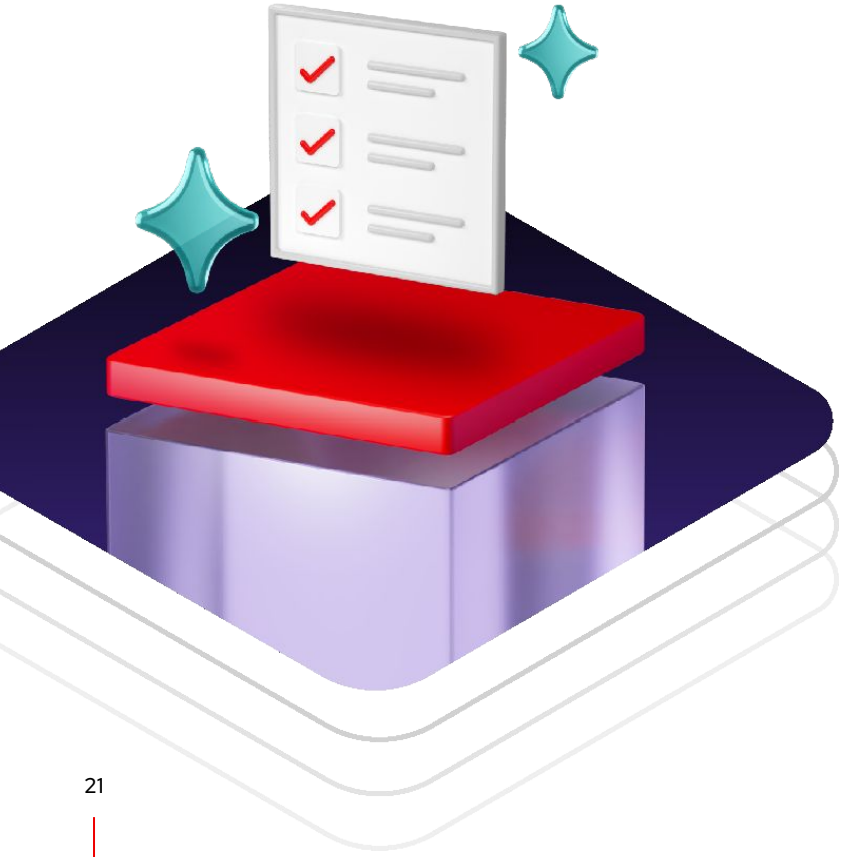
  meaning CVE fixes related to Open SSL or crypto libraries can be done *without* requiring a new FIPS validation certificate (average time to obtain is >300 days!)



The best defense is a strong partner.

Leverage Red Hat Enterprise Linux as a trusted AI foundation

with an extensive ecosystem of trusted partners and tools



Get new AI advances faster, with safety and trust

- New Partner Validation Program for the latest AI hardware/software
- New Extensions Repository for community content that is trusted by Red Hat

Experiment with and develop AI capabilities easily and efficiently

- Simplify management of containerized applications with Podman Desktop
- Use foundational building blocks like the Postgres Vector Database Extension to improve the accuracy of Generative AI and planned support for Model Context Protocol (MCP) to add agentic capabilities
- Streamline deployment of pre-configured base images with image mode

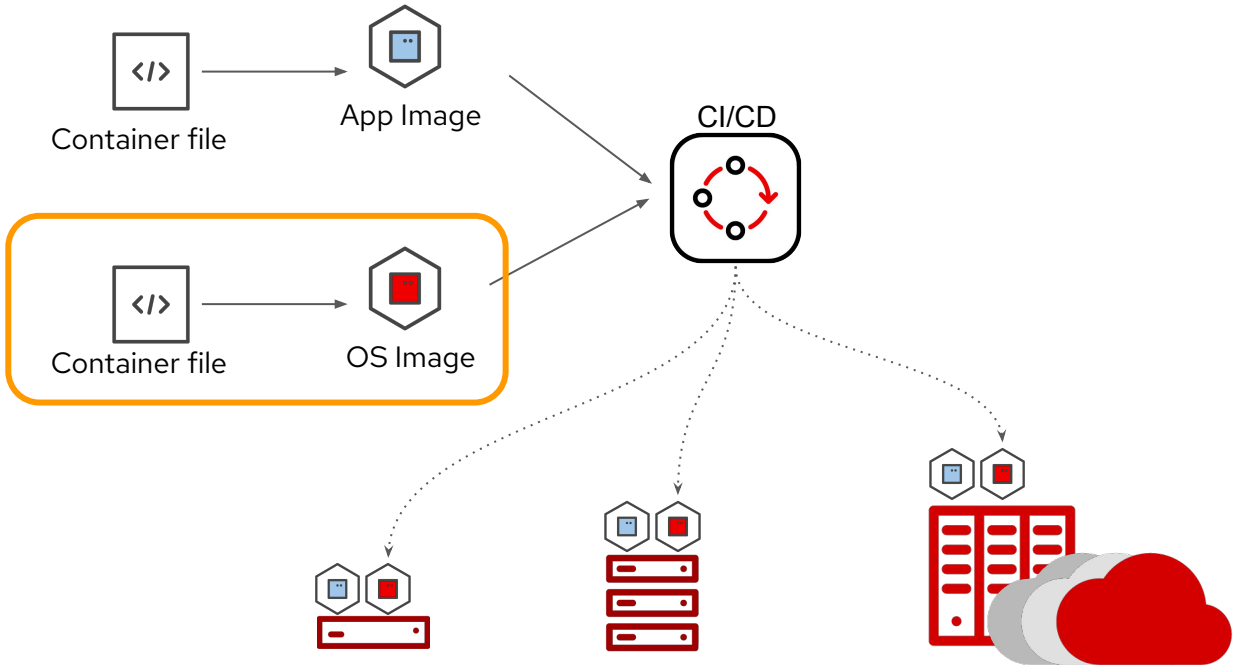
Optimize data storage and security

- Protect data in use with confidential computing support

Wait less to put AI to work.

RHEL Image Mode

RHEL enables teams, infrastructure, and ecosystems to converge on a single container-native workflow to manage everything from their applications to the underlying operating system. Build, test and distribute the operating system as if it was any other container.



- Gain **simplicity** and **portability** across traditional and hybrid cloud environments
- **Accelerate** OS deployments and last-mile challenges with existing CI/CD & GitOps workflows
- **Security benefits** include cryptographic validation, transparent provenance, and tool consolidation

Red Hat

Red Hat Satellite 6.17



Support for Red Hat Enterprise Linux 10

Inventory and manage systems running Red Hat Enterprise Linux 10

Image mode support

New support for provisioning, client management, and registry distribution for systems created with image mode for Red Hat Enterprise Linux

Support for Flatpak content

Simplify importing, managing, and deploying updates via containers with support for Flatpak content, including both custom packages and Red Hat-provided applications

Secure boot support

Enhance security during provisioning workflows on bare-metal, VMware vSphere, and Libvirt platforms

IPv6 support

Deploy Satellite in an IPv6 environment to help address the limitations of IPv4, including improved network efficiency and security

Red Hat Insights advisor

Get proactive risk detection and configuration recommendations without connecting directly to the internet via an on-premise extension (now available in tech preview)

OpenShift Virtualization

- **Red Hat and AMD strengthened strategic collaboration**
  Integrates Red Hat platforms with AMD Instinct GPUs to work with OpenShift AI
- **Red Hat powers modern virtualization on Microsoft Azure**
  OpenShift 4.16 support, enterprise-grade cluster-wide proxy, Ddsv5 instance performance optimization
- **Red Hat OpenShift Virtualization on Google Cloud**
  Now GA

Red Hat

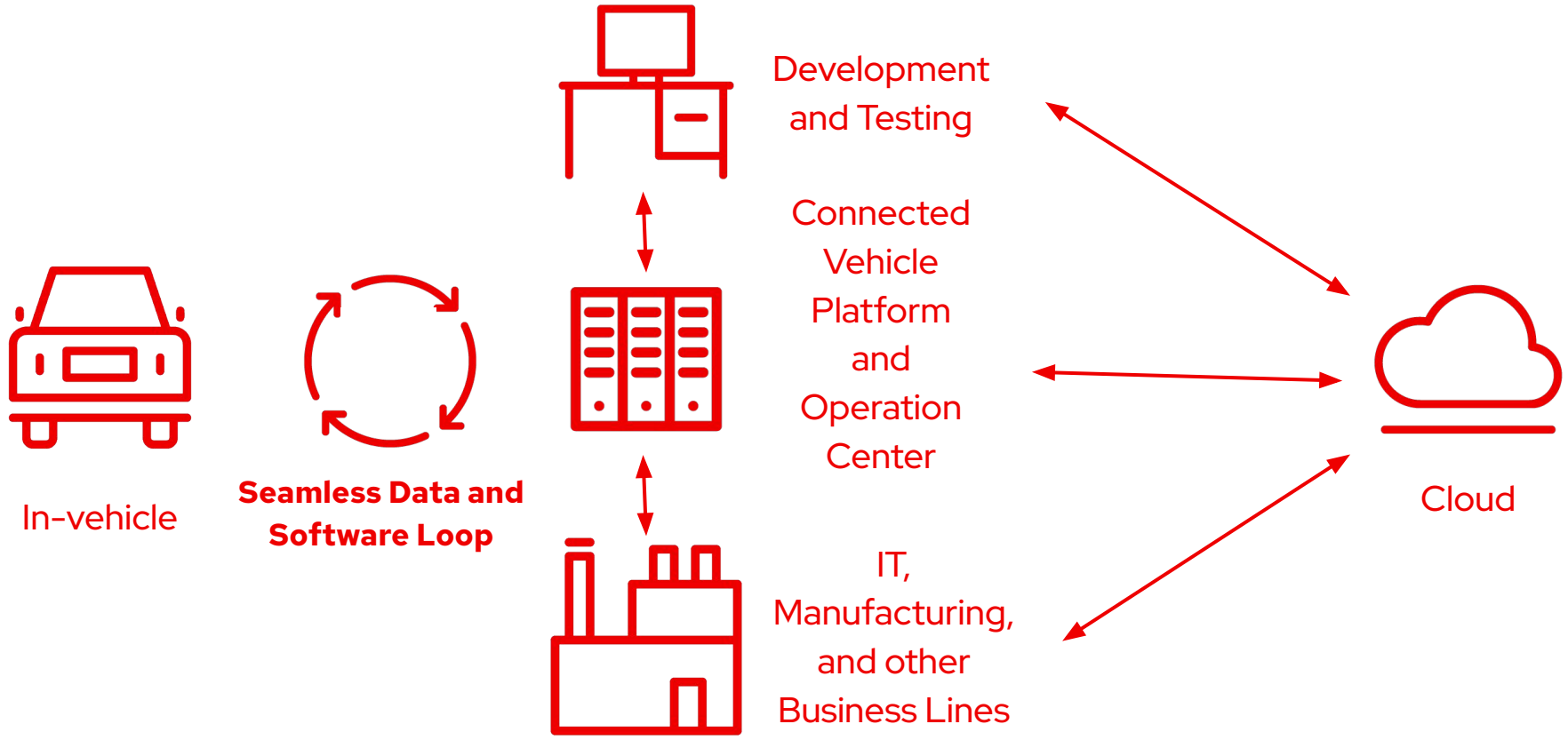
Edge and In-Vehicle OS

- **Introduced Red Hat Edge Manager**
  Comprehensive fleet management solution for edge devices at scale
- **Functional Safety Certification of Red Hat In-Vehicle Operating System**
  Certified as a Safety Element out-of-Context (SEooC) against ISO 26262 Edition 2, 2018 - Level ASIL-B

Red Hat

Red Hat confidential - for internal use only

Red Hat Focus Areas on Software Defined Vehicle



Consistent Technologies, Platforms and Operations for Linux, Containers, DevOps & Microservices

Red Hat confidential - for internal use only

Red Hat Edge portfolio

Our vision is to address the diverse needs of edge

Industry-specific solutions, co-developed with partners

Red Hat In-vehicle Operating System (RHIVOS)

Connected vehicles

Software-defined vehicles with connected services

Red Hat Device Edge

Red Hat Ansible Automation Platform

MicroShift (lightweight K8s)

Red Hat Enterprise Linux (with edge features)

Edge devices

Field-deployed endpoints and gateways with low footprint

Red Hat OpenShift*

or

Red Hat Enterprise Linux (with edge features)

Edge servers

Standard servers deployed in controlled environments

AI & ML

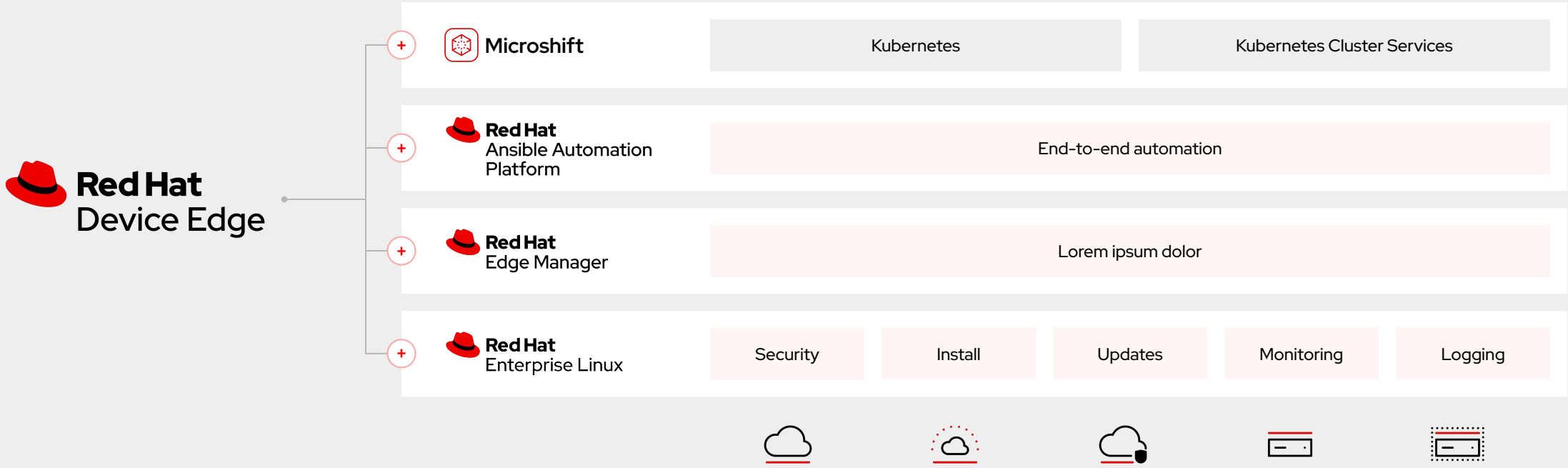
Management & Automation

Security

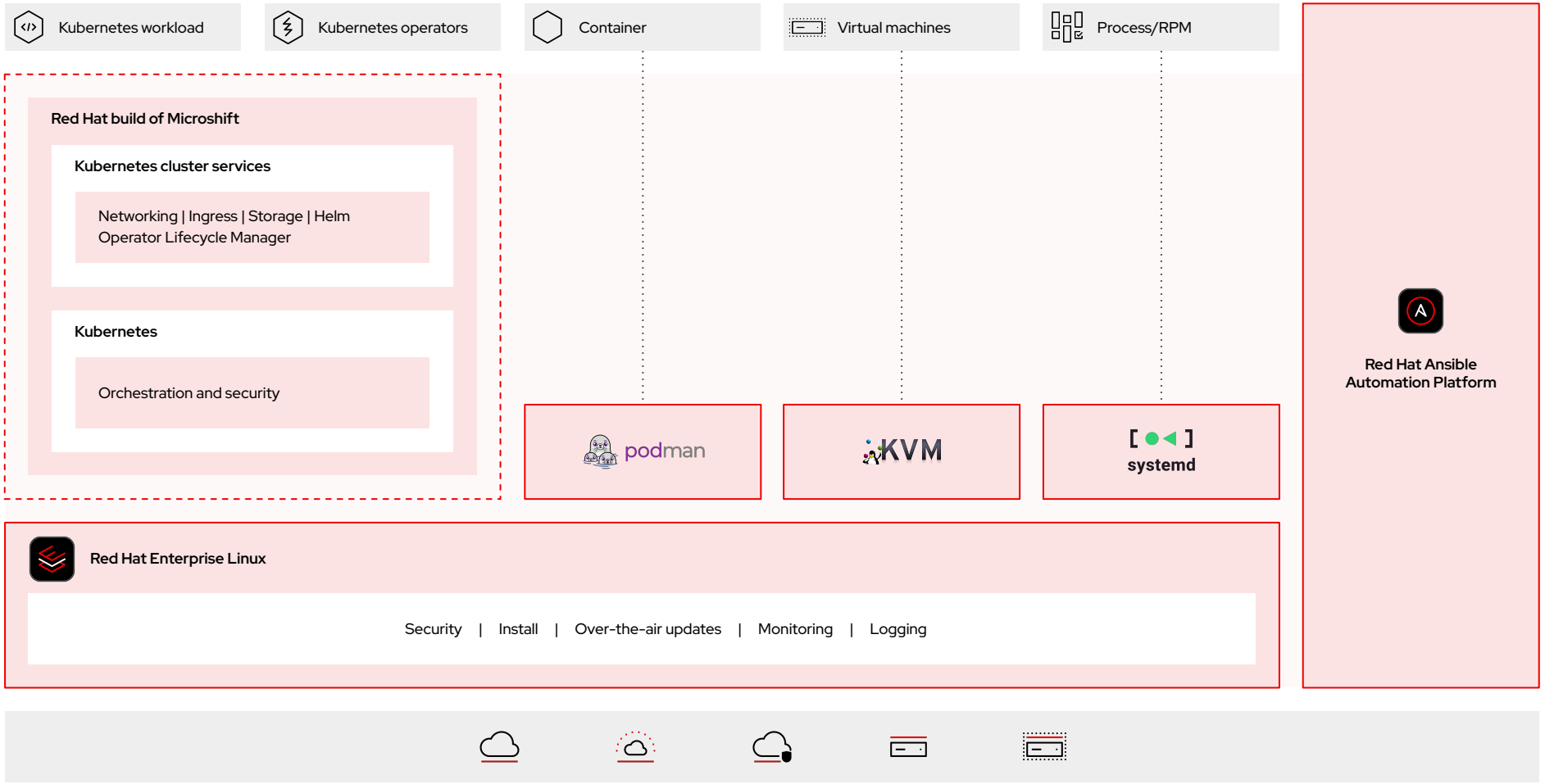
Application and data services

*** OpenShift includes RHEL CoreOS, so no separate purchase needed for the Operating System*

What is Red Hat Device Edge?

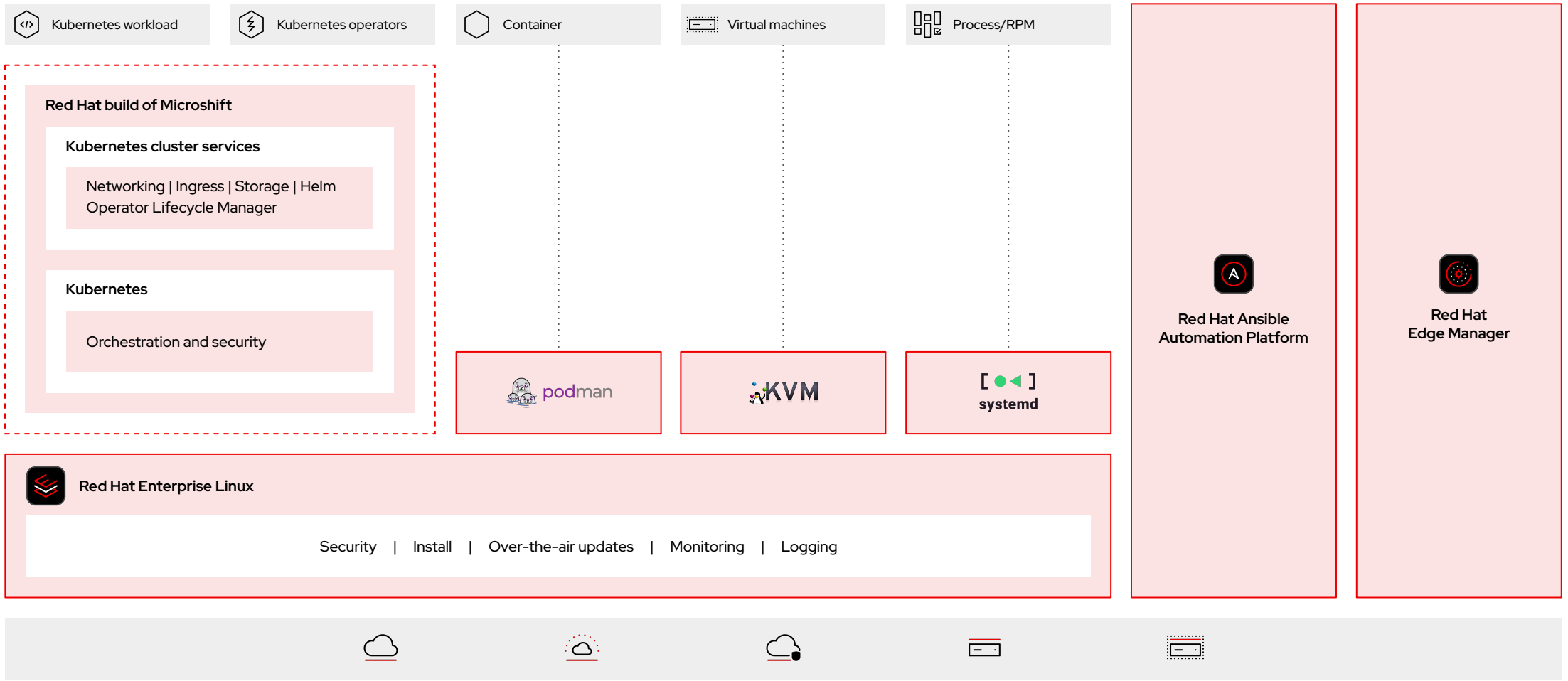


Red Hat
Device Edge



Option 2

Red Hat Device Edge



Red Hat Edge Manager

Red Hat confidential - for internal use only

Simple

Intuitive edge operations

- Bridge the IT skills gap at the edge with a user-friendly environment designed for ease of use and simplified management.

Flexible management options

- Choose the deployment model that fits your strategy, with on-premises management offering comparable or superior value to cloud alternatives.

Scalable

Policy-driven deployment

- Implement a desired-state configuration model for both applications and infrastructure, enabling consistent and scalable deployments, maximizing operational efficiency.

Resilient agent architecture

- Use a robust agent-based architecture for scalable device management, maintaining connectivity and control even in challenging network conditions, without complex network configurations.

Built for edge

Hardened device communications

- Employ mutual Transport Layer Security (mTLS) for robust and authenticated agent service communication. Establish a consistent security posture even for re-connected devices through rigorous identity verification.

Proactive device insights

- Monitor critical device compute, memory and disk resources. Capture essential metrics and logs for effective issue remediation and operational awareness.

Lifecycle management

- Provides secure onboarding, upgrades and decommissioning of applications and OS.

Announcing

Red Hat In Vehicle Operating System

The first safety-certified Linux platform for automotive software development



Launch Timeline

✓ **Coming Q3 2025**



Functional Safety Certification

✓ **Certified as a Safety Element out of Context (SEooC)**

- First **FuSa-certified Linux distribution** for automotive software development
- Comes with **Safety Guidance** defining safety scope and Assumptions of Use



Built on Red Hat Enterprise Linux (RHEL)

✓ **Tailored to automotive industry needs**

- Leverage RHEL's proven foundation and security practices
- Adapted for **long lifecycle support, real-time performance and automotive-grade resilience**

Tailored to meet the specific demands of the automotive industry

Safety Certification - ISO 26262 ASIL-B as SEooC

Mixed-Criticality Support: support hosting multiple concurrent workloads at different integrity levels

Fast Boot - Optimized startup times for automotive applications

Real-Time Capabilities - Reduced variation & preemptive latency with default PREEMPT-RT

Package Selection - All RHEL packages and automotive-specific packages included

Hardware Architecture - Wide variety of hardware platforms based on aarch64 and x86 architecture

Diverse Use Cases - From digital cockpits & ADAS to gateways and QM up to ASIL-B applications

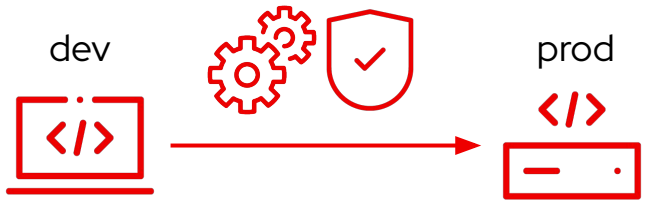
Modern SW Development - Embrace container-based development and deployment paradigm and DevOps style build/deploy/test cycles

Advanced Developer Suite

- **Integrated software supply chain security best practices**
  Holistic solution to help developers move faster while adhering to stringent security and compliance demands
- **Driving AI innovation with greater speed and enhanced security**
  Integrations with Red Hat AI Inference Server, RHEL AI and Red Hat OpenShift AI that deliver security-focused AI-enabled applications
- **Red Hat Developer Hub, Red Hat Trusted Profile Analyzer, Red Hat Trusted Artifact Signer**
  Red Hat Advanced Developer Suite will be available for purchase on July 1, 2025.

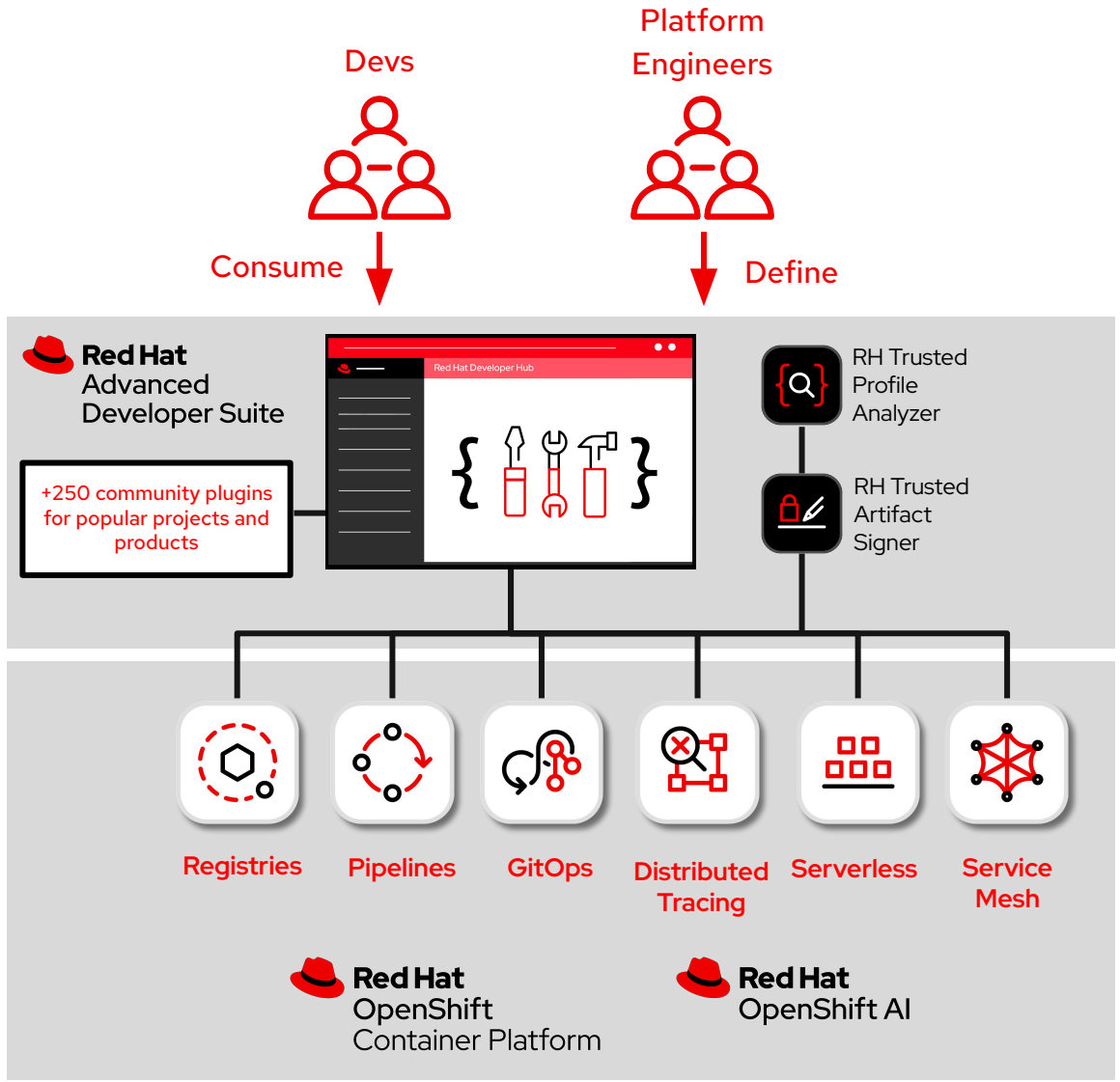
Red Hat

Red Hat Advanced Developer Suite



RHADS combines platform engineering tools with enhanced security capabilities to help accelerate and simplify application development with new enhancements to speed the adoption of AI

- Modular design allows it to be integrated with existing platforms and systems including Red Hat's platforms
- Integrates with existing security capabilities found in Red Hat Advanced Cluster Security
- Complements existing developer tools from Red Hat including OpenShift DevSpaces
- Built on leading open source projects like Backstage and Sigstore



Red Hat

Red Hat Advanced Developer Suite

IDP: Internal Developer Portal

Red Hat Developer Hub

AI/Kube Developer Desktop Experience : **Podman Desktop**

Migration toolkits

Development Tools

IDE Plugins

Red Hat OpenShift Dev Spaces

Trusted Software Supply Chain

Red Hat Trusted Artifact Signer

Red Hat Trusted Profile Analyzer

AI Enablement

Developer Lightspeed

Podman Desktop AI Lab

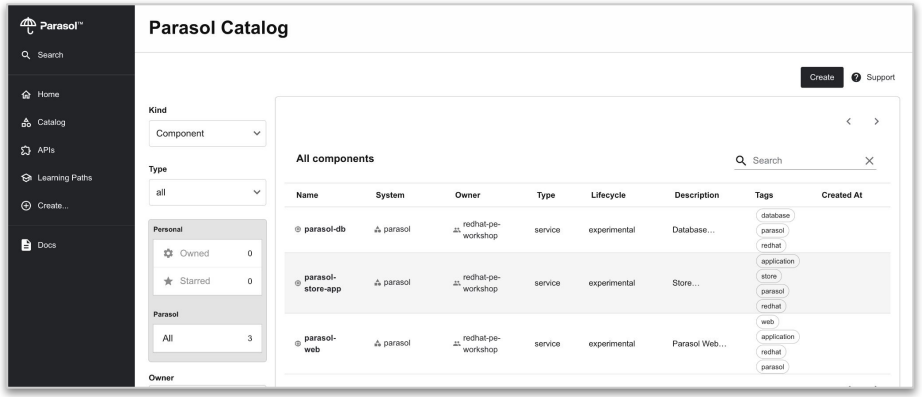
Red Hat Developer Hub AI Integration

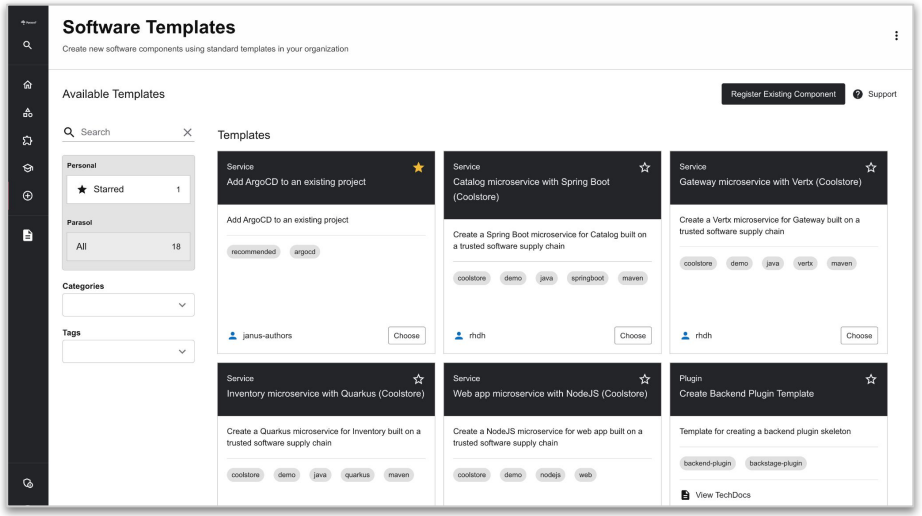
Products that allow developers to **quickly and securely deliver applications to production**

Enabling platform engineers to more easily enforce operational controls and compliance across hybrid multi-cloud environments

Red Hat

Key Technical Capabilities: Developer Productivity

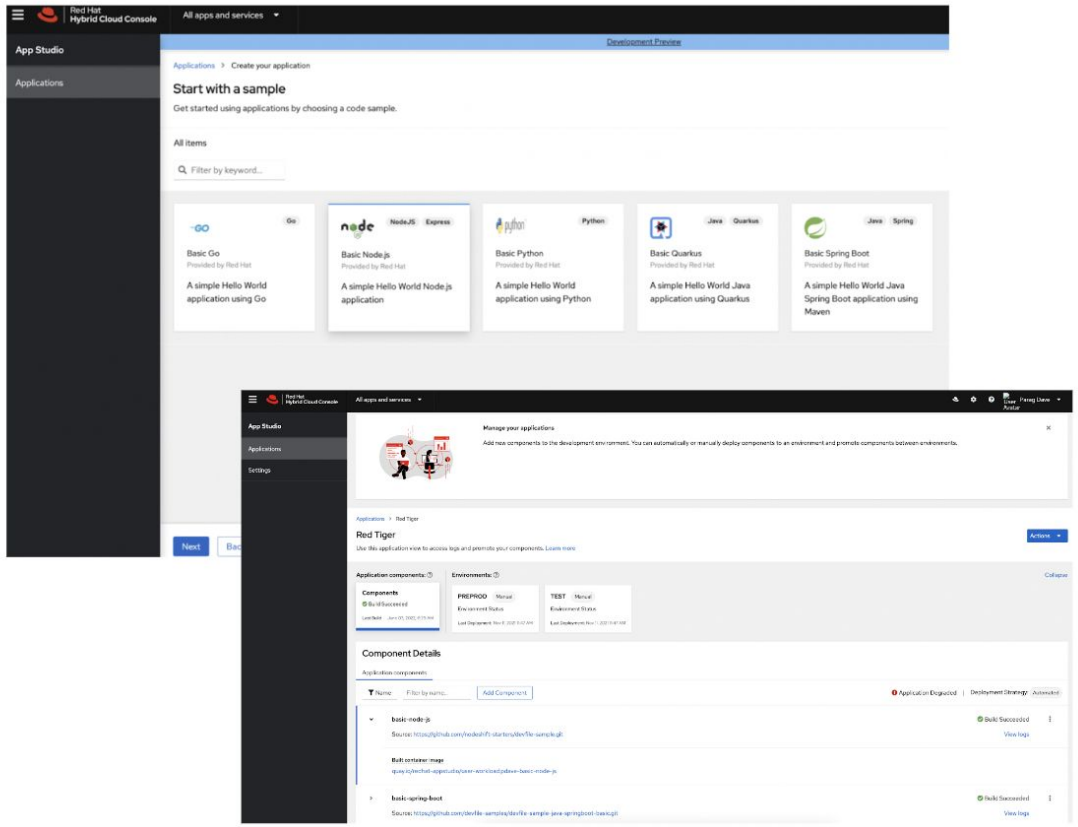




- Accelerated onboarding and frictionless self-service
- Single-pane visibility to simplify app ownership and governance
- Consistent Development-to-Production Workflows
- Automated Software Templates
- Streamlined Automation for Faster, Secure Development
- Rootless, Secure Container Execution

Red Hat

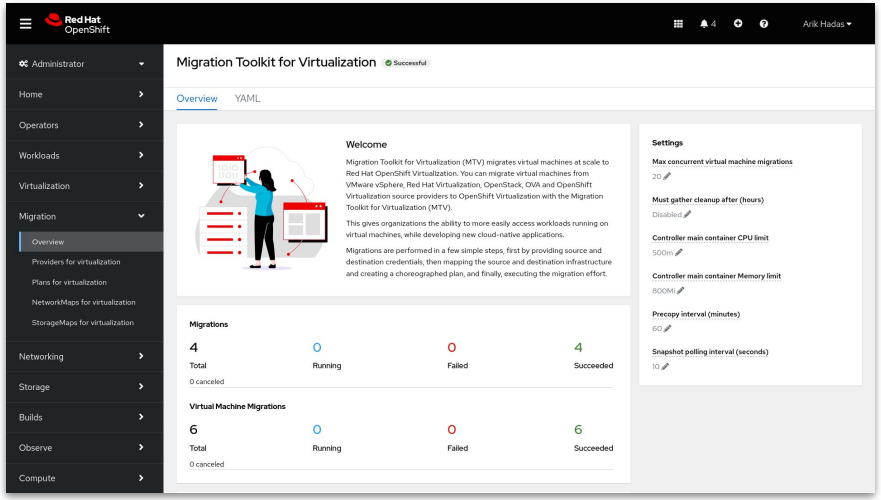
Key Technical Capabilities: Supply Chain Security

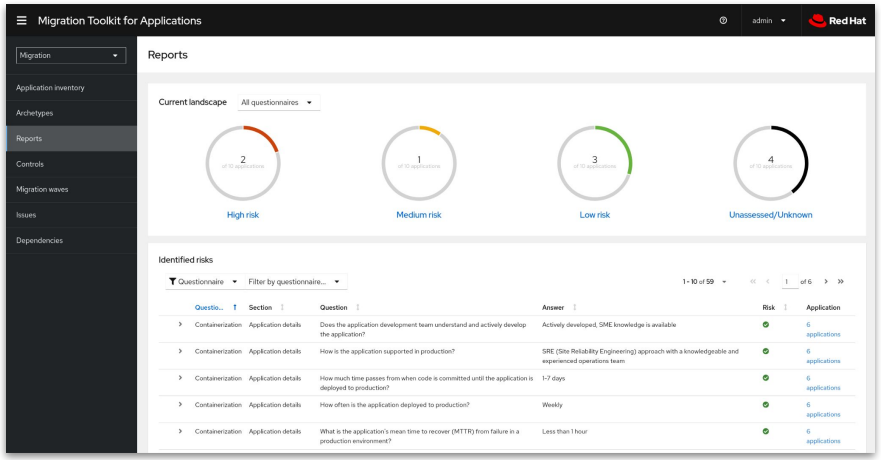


- Early artifact signing, verification, and secure builds within CI/CD system
- Automated signing & verification, provenance & attestations with immutable ledger
- Trusted Content and Policy-Based Remediation Guidance
- Risk Profile Monitoring & Blast Radius Analysis
- Software Composition Analysis (SCA)
- Software Bill of Materials (SBOMs) Generation & Management
- SLSA Framework – Secure software build pipelines with provenance guarantees
- Integration with ACS for buildtime policy checks, network policy generation, live application scanning, cluster-level security policy enforcement, automated compliance and deep insights

Red Hat

Key Technical Capabilities: App Infra Modernization

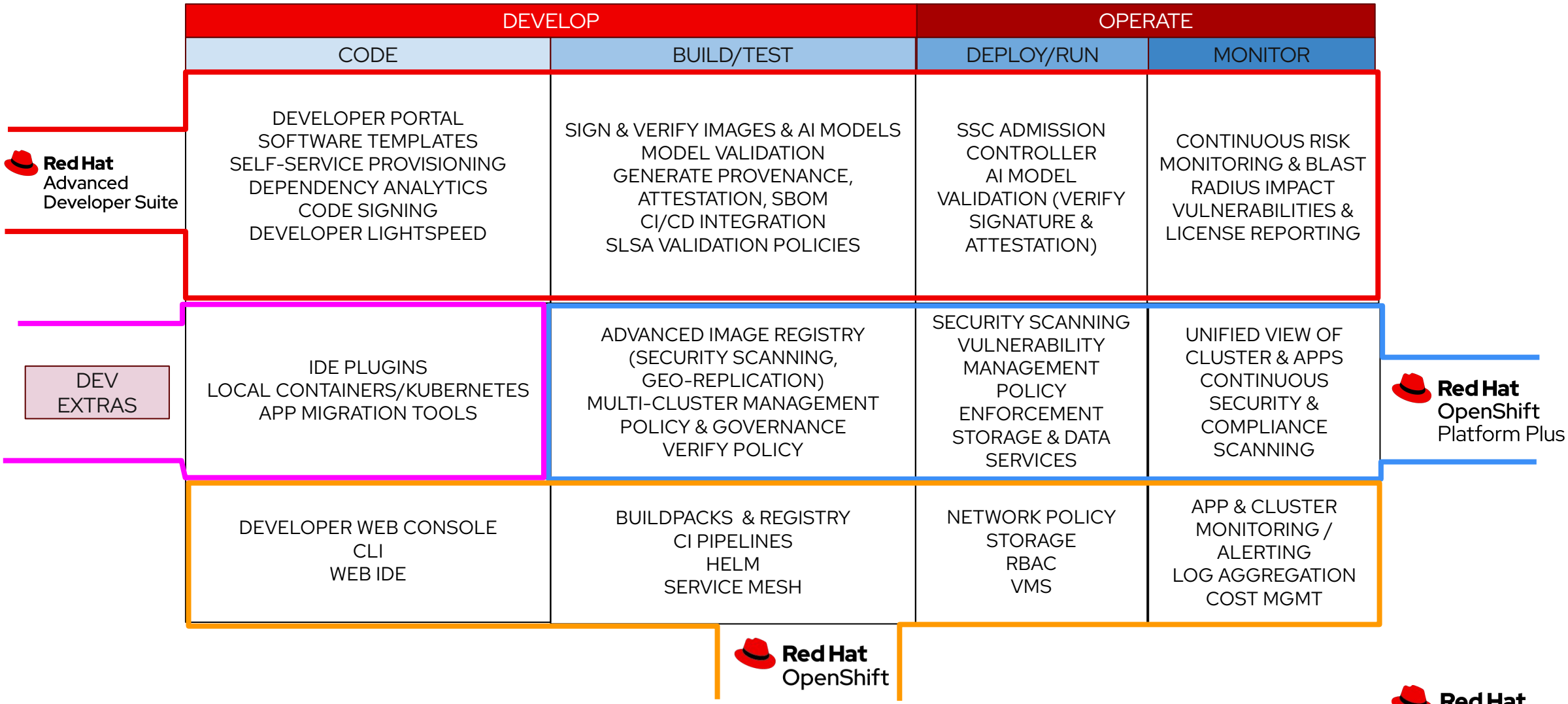




- Integration with CI/CD Pipelines
- Automated Workload Migration
- AI-assisted Workload Modernization
- Docker-Compatible, Rootless Container Management

Red Hat

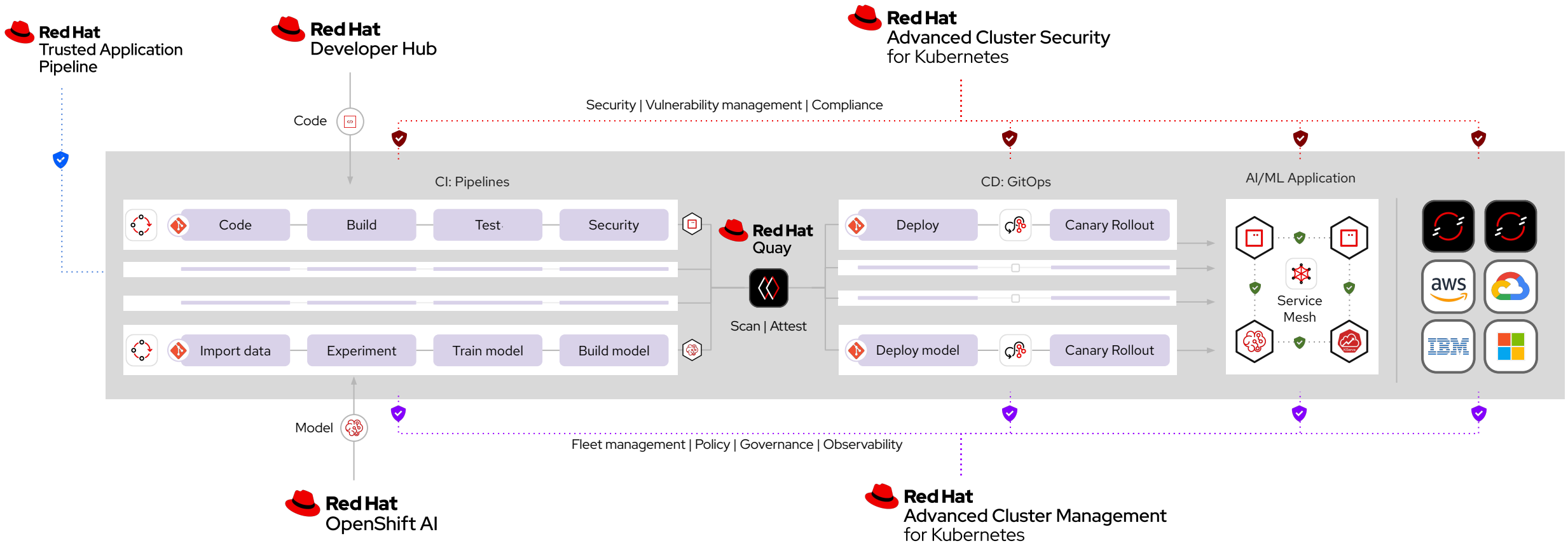
Expanded OpenShift capabilities across the lifecycle



Red Hat

Red Hat Advanced Developer Suite

Extending Openshift to build, deploy and run secure AI/ML applications at scale



Red Hat

Other Announcements

- Quarkus 3.20
- Trusted Artifact Signer 1.2
- Trusted Application Pipeline 1.5
- Red Hat Learning Subscription 90 Day Trials

Red Hat

Red Hat Summit

Thank you

linkedin.com/company/red-hat

facebook.com/redhatinc

youtube.com/user/RedHatVideos

x.com/RedHat

Red Hat