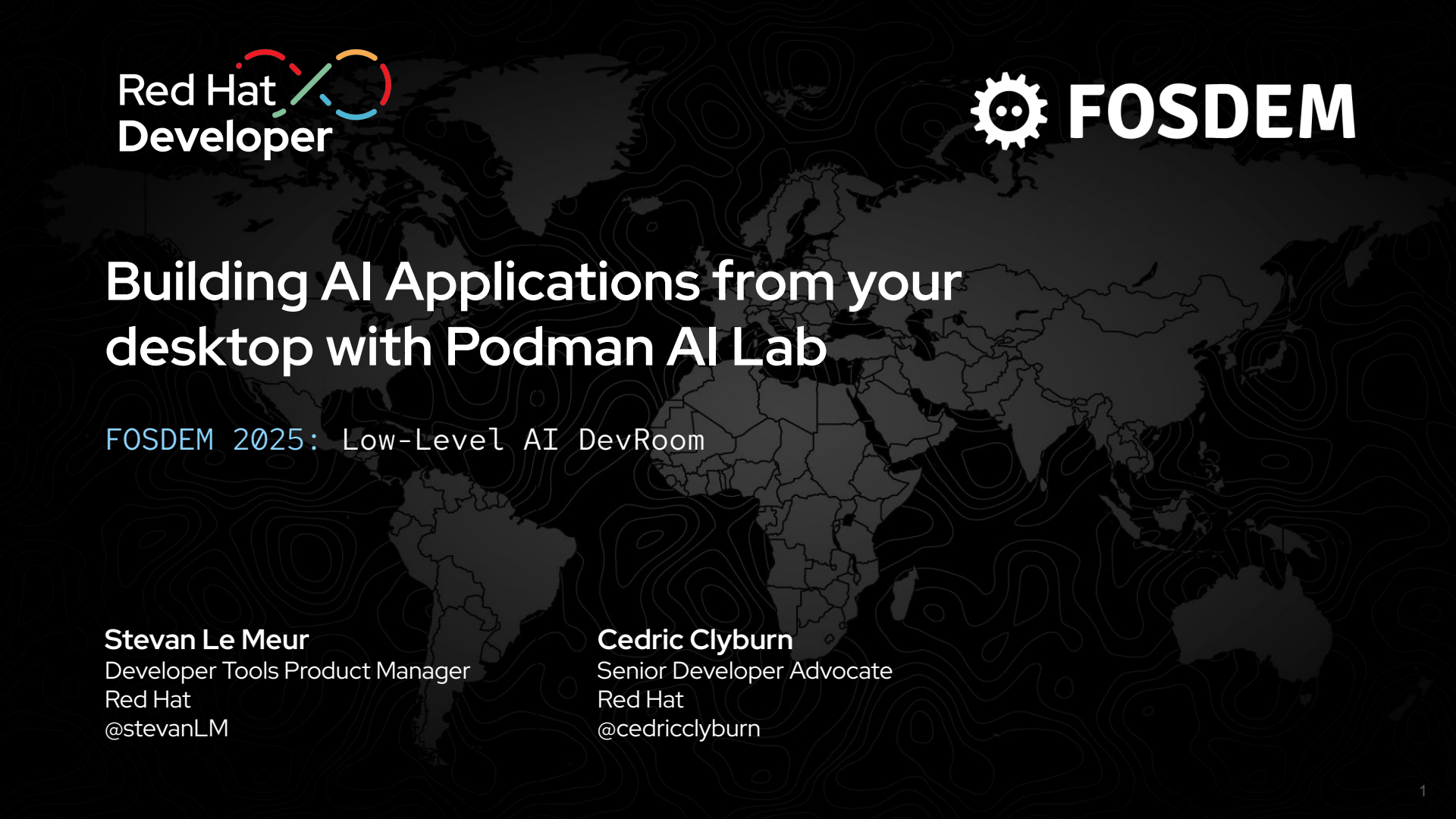


Building AI Applications from your desktop with Podman AI Lab

FOSDEM 2025: Low-Level AI DevRoom

Stevan Le Meur
Developer Tools Product Manager
Red Hat
@stevanLM

Cedric Clyburn
Senior Developer Advocate
Red Hat
@cedricclyburn

Today's Schedule

- What's **Podman Desktop**? 
- Generative AI **for Developers**
- The **Podman AI Lab**! 

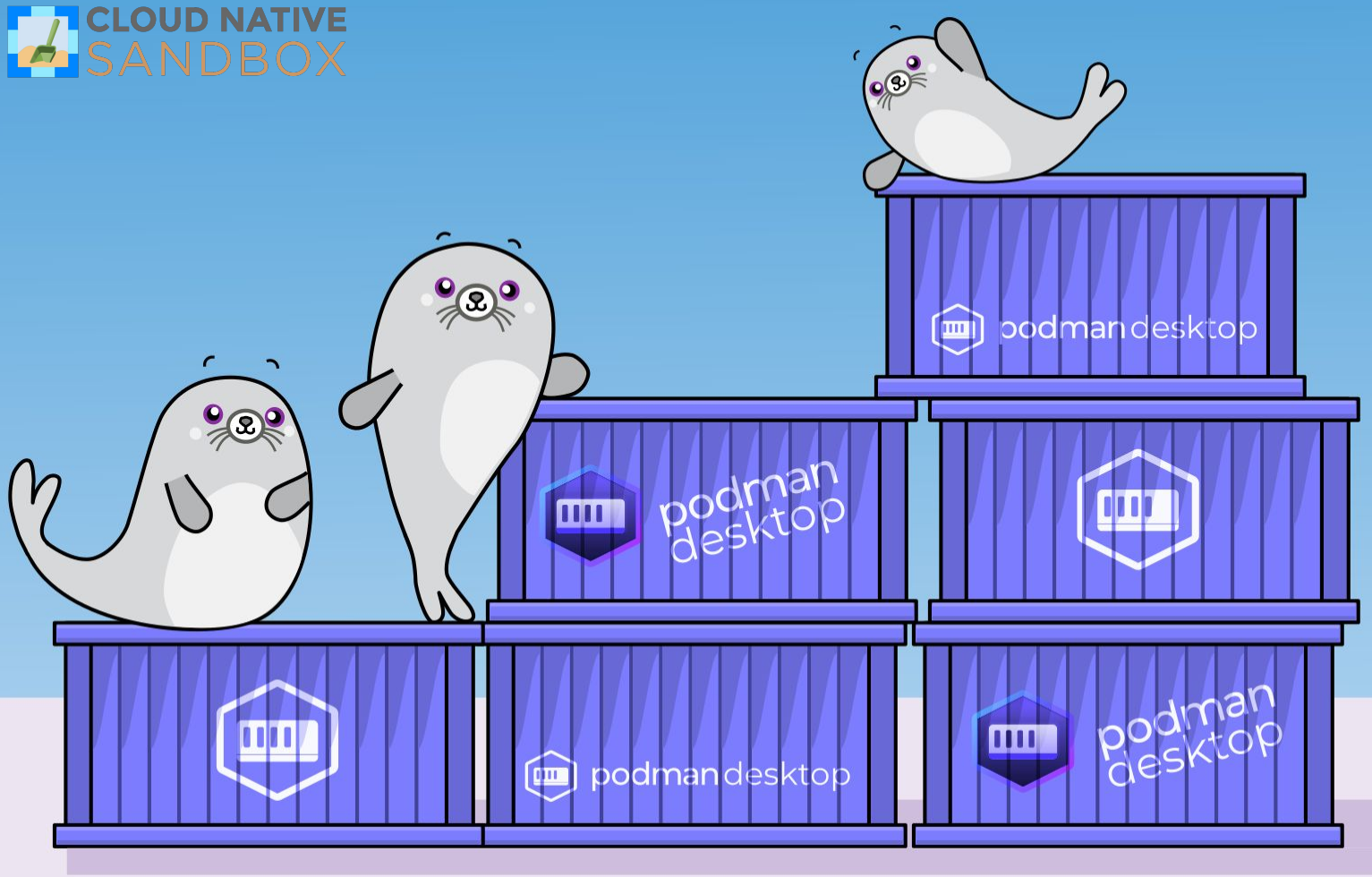
- **Demo #1:** Open source models
- **Demo #2:** RAG & Multi-Modality
- **Demo #3:** Adding AI to existing Apps!

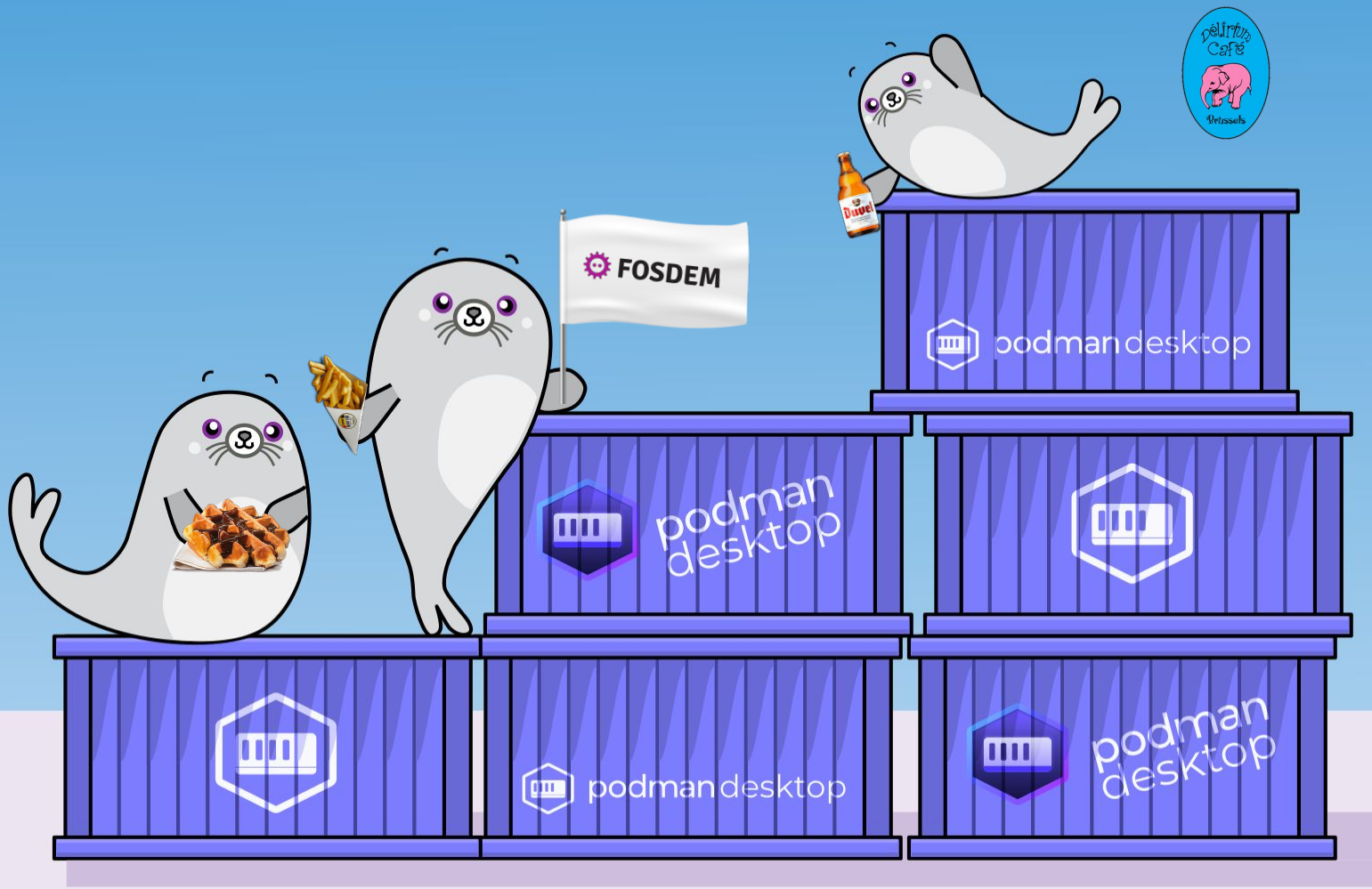
Session Slides

red.ht/ai-lab-slides

Try Podman Desktop!

podman-desktop.io


CLOUD NATIVE
SANDBOX
podman desktop
podman desktop
podman desktop
podman desktop
podman desktop


FOSDEM
podman desktop
podman desktop
podman desktop
podman desktop
Duvel
Brussels

What is Podman?

A seamless way to work with containers (& Kubernetes!)

Fast and light

Daemonless, using the fastest technologies for a snappy experience.

Secure

Rootless containers allow you to contain privileges without compromising functionality.

Open

Podman is open source first and won't lock you in. Podman Desktop even supports Docker as an engine!

Compatible

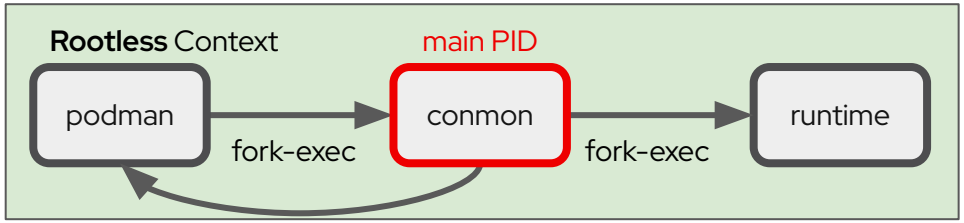
Compatible with other OCI compliant container formats including Docker, as well as docker-compose files.

Red Hat Developer

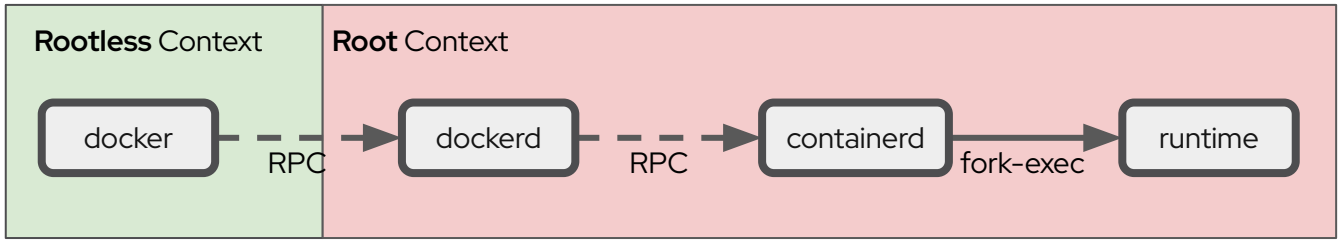
Source:
https://podman.io
https://www.redhat.com/architect/hpc-containers-scale-using-podman

Red Hat

Container Engine Architectures



Compared to...



More About Docker vs Podman

Some resources to check out... later!

Podman vs. Docker

IBM Technology
656K subscribers
Subscribe
1.5K
Share

50K views 6 months ago Kubernetes Essentials
IBM and Red Hat solutions → https://ibm.biz/BdykC2

Is it time to switch from Docker to Podman?

Christian Lempa
182K subscribers
Subscribe
4.2K
Share

140K views 4 weeks ago
In this video, we will be exploring an alternative to Docker - Podman. With its claims of being faster, more

Red Hat Developer

Source:
https://www.youtube.com/watch?v=Xx588nbshIM
https://www.youtube.com/watch?v=7-qo6tTPdTM

Red Hat

Introducing Podman Desktop!

Containers and Kubernetes for Application Developers

Podman and Kubernetes Environments

- Install and run anywhere: Windows, Mac and Linux
- Keep it up-to-date

Containers and Pods

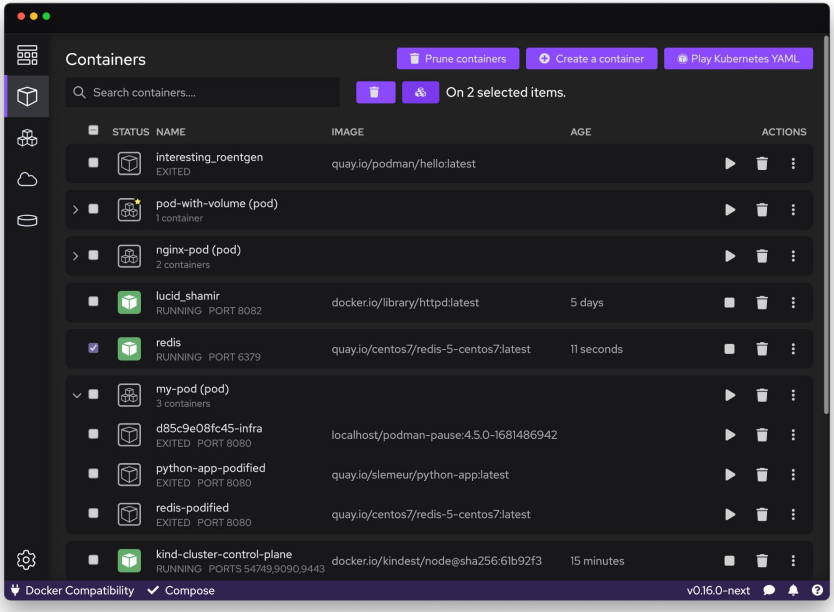
- Build, run, manage and debug Containers and Pods
- Run Pods with or without Kubernetes
- Manage multiple container Engines
- Compatibility with Docker and Compose

Kubernetes Integration

- Easily move from containers to Kubernetes through the Kubernetes UI for Deployments, Services, Ingress & Routes, and much more!

Extensibility

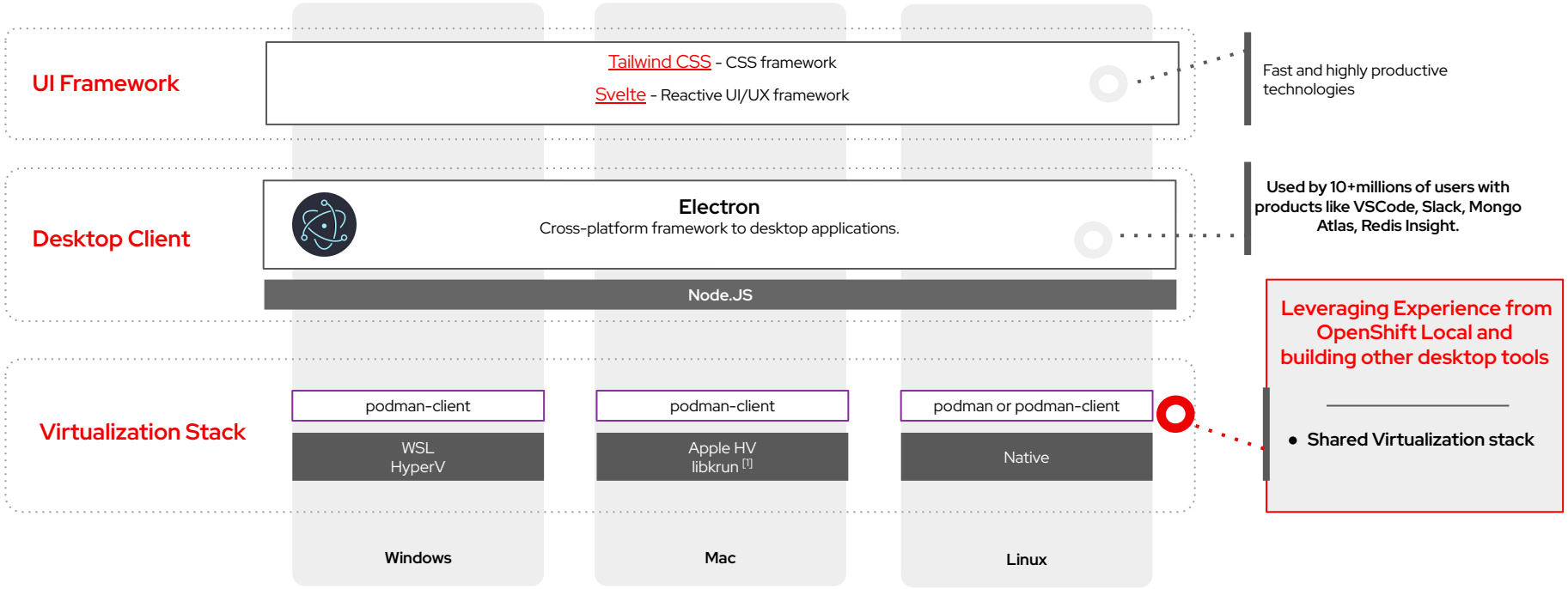
- Enabling extension points and other container/K8s technologies



Red Hat Developer

Red Hat

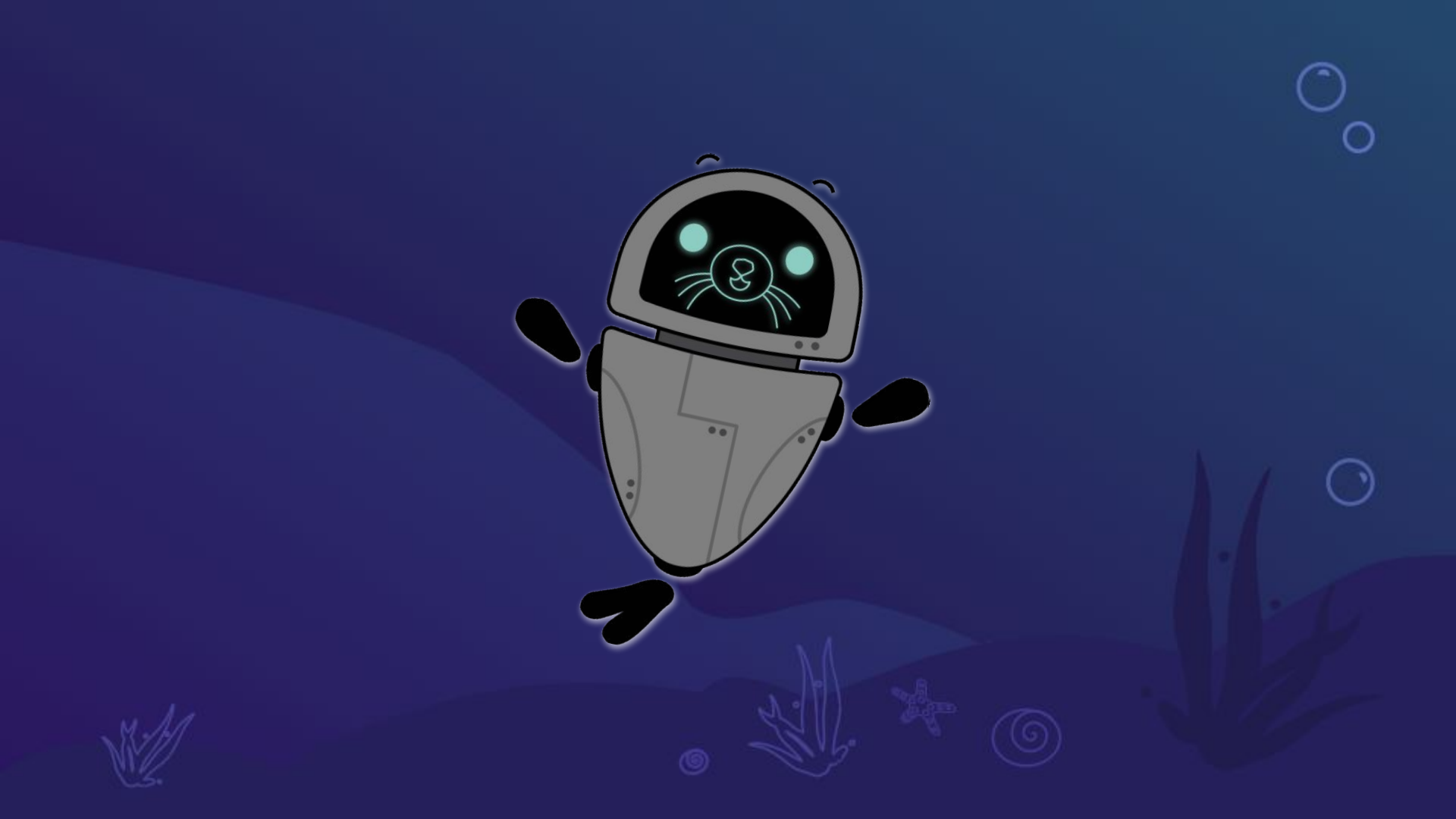
Podman Desktop: Behind the Scenes



Red Hat Developer

Red Hat

Transitioning to new era...



It all starts with **enabling** the developers to use models

Why run a model locally?

Take advantage of total AI customization and control

For Developers

Convenience & Simplicity

Familiarity with the Development Environment and adherence of the developers to their “local developer experience” in particular for testing and debugging

Ease of Integration

Simplify the integration of the model with existing systems and applications that are already running locally.

Customization & Control

Easily train or fine-tune your own model, from the convenience of the developer’s local machine.

Direct Access to Hardware

For Organizations

Data Privacy and Security

Data is the fuel for AI, and a differentiator factor (quality, quantity, qualification)

Keeping data on-premises ensures sensitive information doesn’t leave the local environment → crucial for privacy-sensitive applications

Cost Control

While there is an initial investment in hardware and setup, running locally can potentially reduce ongoing costs of cloud computing services and alleviate the vendor-locking played by Amazon, MSFT, Google

Regulatory Compliance

Some industries have strict regulations about where and how data is processed

Podman Desktop AI Lab

Your pathway to Gen AI, bridging with containers and Kubernetes

Adopting Generative AI for an Application Developer

Ideation & Prototyping

- Find LLMs
- Try prompts
- Experiment with your data
- Benchmarking

As a developer, how do I **evaluate** models and pick the best one for my use case?

Building & Refining

- Connect to data source
- Exception handling
- Limited fine tuning
- Retrieval-Augmented Generation (RAG)
- Chaining
- Evaluate flows

As a developer, how do I **build** an application with LLMs?

Operationalizing

- Model serving
- Endpoints
- Monitoring
- Integrate with apps

As a developer, how do I **deploy** my application with LLMs?

Adopting Generative AI for an Application Developer

Ideation & Prototyping

- Find LLMs
- Try prompts
- Experiment with your data
- Benchmarking

As a developer, how do I **evaluate** models and pick the best one for my use case?

Building & Refining

- Connect to data source
- Exception handling
- Limited fine tuning
- Retrieval-Augmented Generation (RAG)
- Chaining
- Evaluate flows

As a developer, how do I **build** an application with LLMs?

As a developer, how do I **deploy** my application with LLMs?

Introducing: Podman AI Lab

Discover GenAI with out-of-the box Recipes

- Get inspired by AI use cases
- Learn how to integrate AI in an optimal way
- Experiment with different compatible Models

Model Catalog

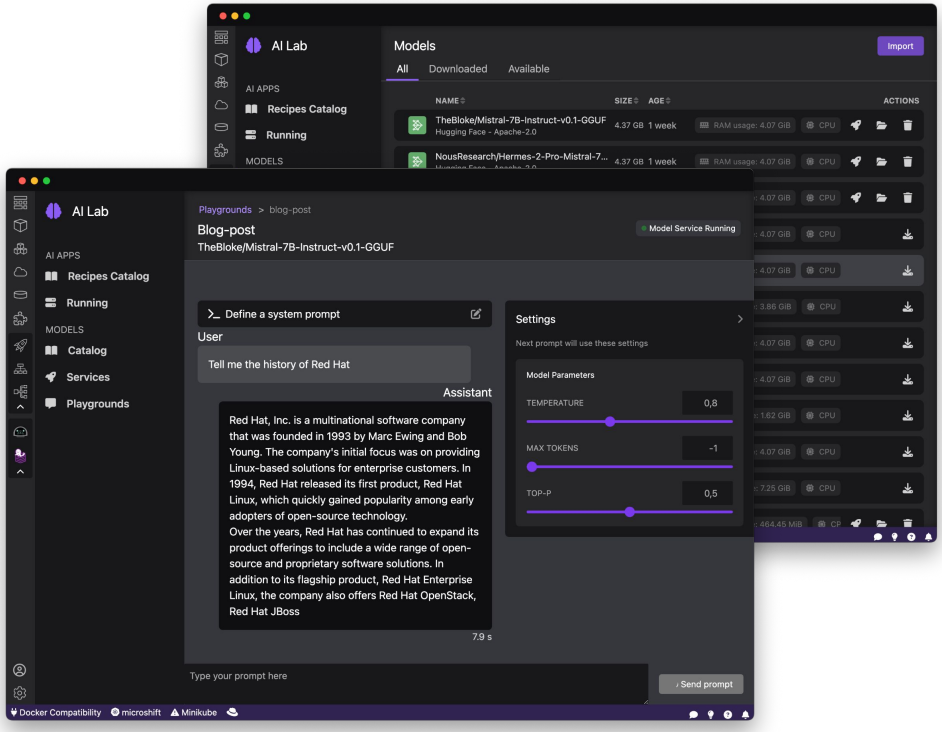
- Leverage a curated list of open source large language models available out of the box
- Import your own models

Run Models Locally

- Run models with an inference server running in UBI image
- Get OpenAI compatible API
- Use code snippets

Playground Environment

- Experiment with models and prompts
- Configure settings and system prompts
- Test and validate prompt workflows before using in your application



Red Hat Developer

Red Hat

Podman AI Lab!

Your path way to GenAI bridging with Containers and Kubernetes

Run LLM's Locally

Simplify the environment for developing and debugging AI-enabled applications in your inner loop.

Cost Efficiency

Running AI models locally reduces cloud computing costs and vendor locking.

Gen AI Education

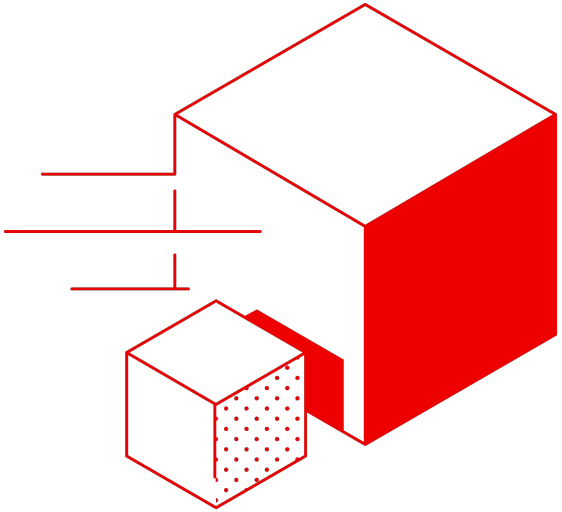
Learn how to infuse AI in your applications through sample use cases and best practices.

Data Ownership

Maintain ownership and control over your sensitive data from your local environment.

DEMO

Why containers, Kubernetes, and DevOps for AI/ML?



Agility

Respond quickly with automated compute resource management.

Flexibility

Provision AI/ML environments as and when you need them.

Portability

Develop and deploy ML models consistently across datacenter, edge, and public clouds.

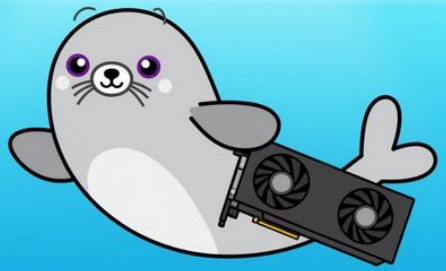
Scalability

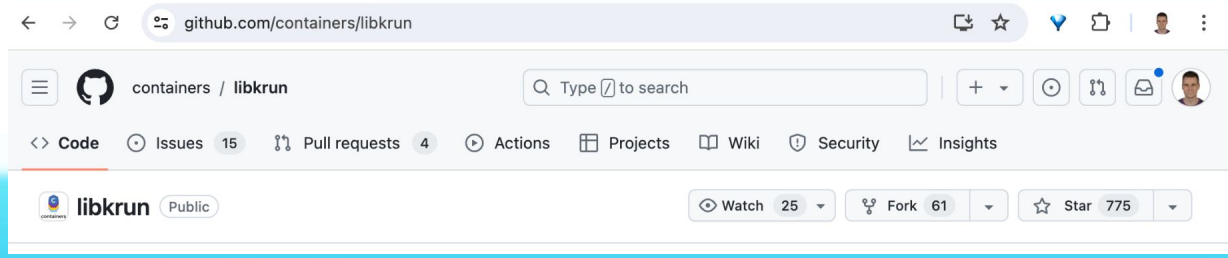
Autoscale and high availability of the AI/ML solution stack.

GPU Acceleration for Containers

- GPU are improving performances of running LLMs on a local environment or on a server.

- **Windows**: GPU passthrough with WSL2
- **Mac**: Integration **libkrun** - available today as an experimental feature
- **Linux**: GPU





Podman AI Lab

Run LLMs locally and build AI applications

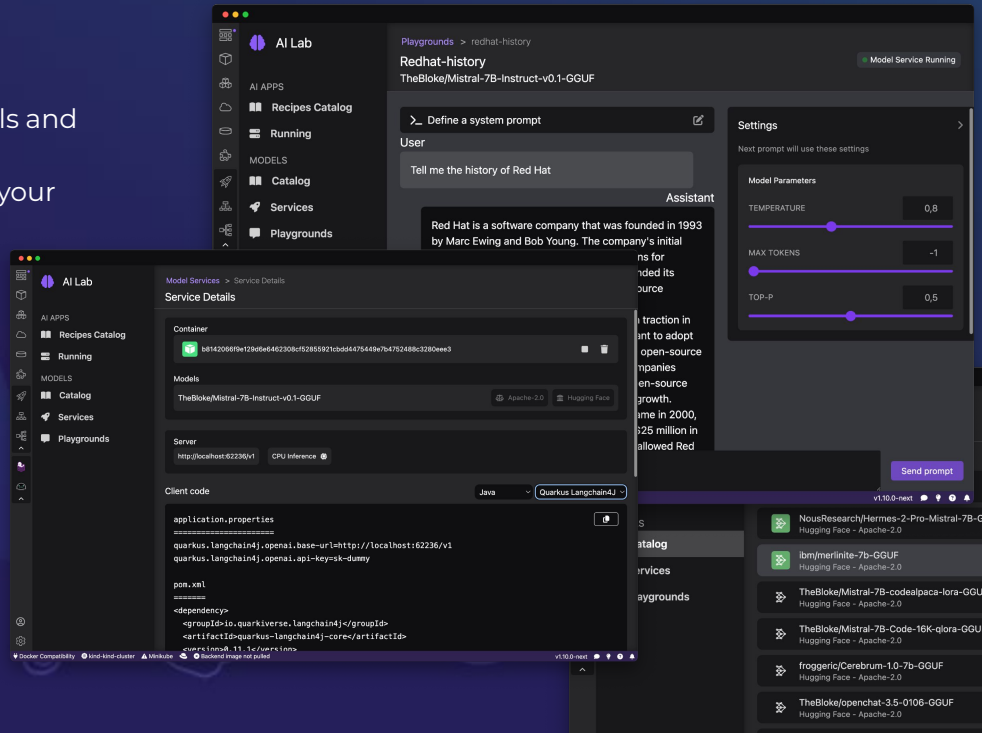
From getting started with AI, to experimenting with models and prompts, Podman AI Lab enables you to bring AI into your applications without depending on infrastructure beyond your laptop.

Supported platforms:

Download now at:

podman-desktop.io

podman desktop



Thank you

Join the DevNation

Red Hat Developer serves the builders. The problem solvers who create careers with code. Let's keep in touch!

- Join Red Hat Developer at **developers.redhat.com/register**
- Follow us on any of our social channels
- Visit **dn.dev/upcoming** for a schedule of our upcoming events

Red Hat Developer

Build here. Go anywhere.


linkedin.com/company/red-hat


youtube.com/user/RedHatVideos


facebook.com/redhatinc


twitter.com/RedHat

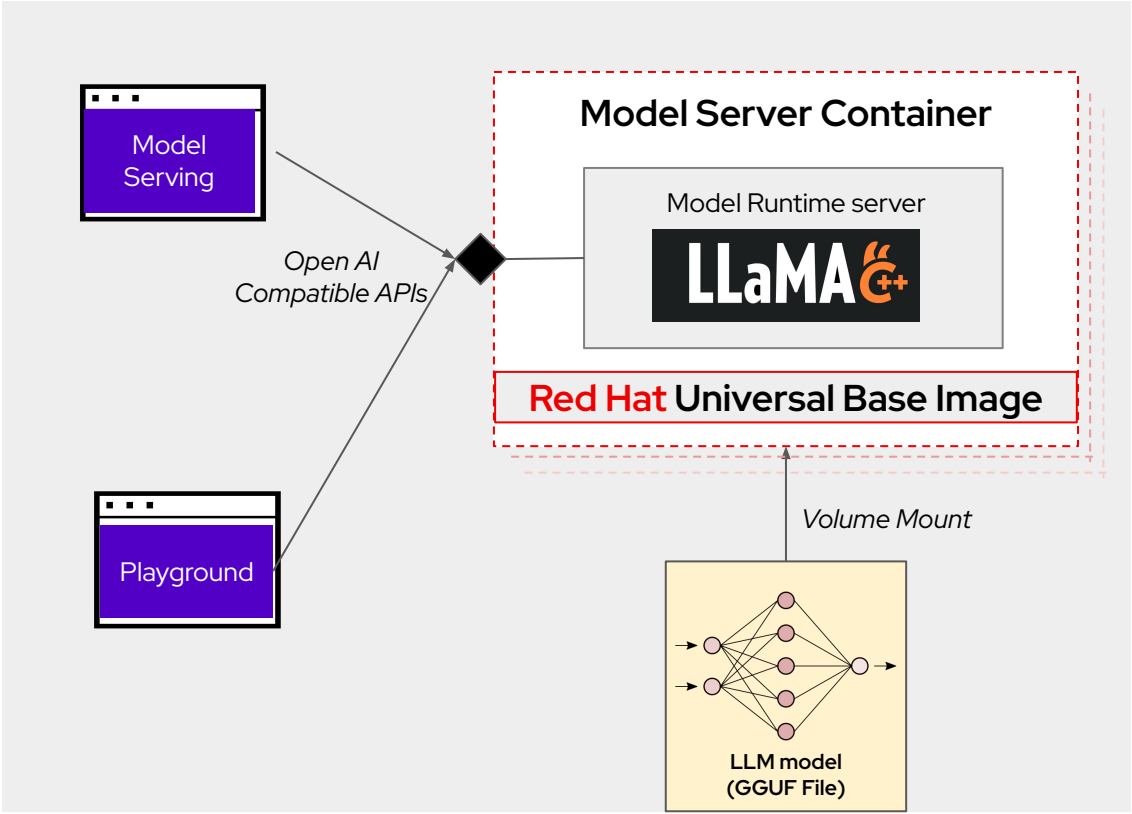
Red Hat Developer

Red Hat

Technical Overview

Enabling Model Serving and Playground Environments

- Built on top of llama.cpp
- Container built with UBI
  - Maintained by Red Hat
- Performances and GPU acceleration

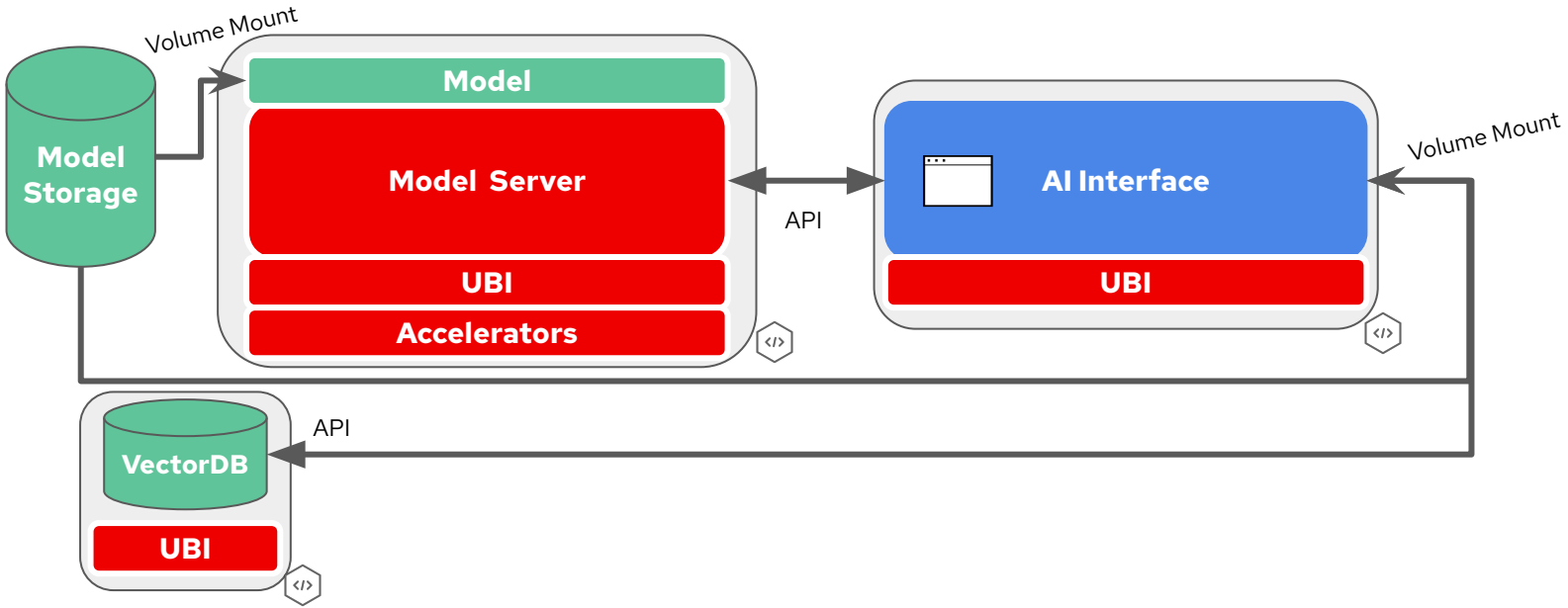


Recipes

- Default catalog for general AI use cases provided with the Podman AI Lab extension
- Developers can package and add their own recipes to the catalog
  - The extension's catalog.json provides a format for lists of categories, recipes, and models

Anatomy of a Recipe

...they're open source



Red Hat Developer

Red Hat

Anatomy of a Recipe

...they're open source