

Run LLMs Your Way

- Deploying Ollama on Azure Container Apps



Agenda

Deploy

Deploy container workloads with ease

Integrate

Integrate Ollama into your own apps for private, cost-efficient inference

Secure and scale

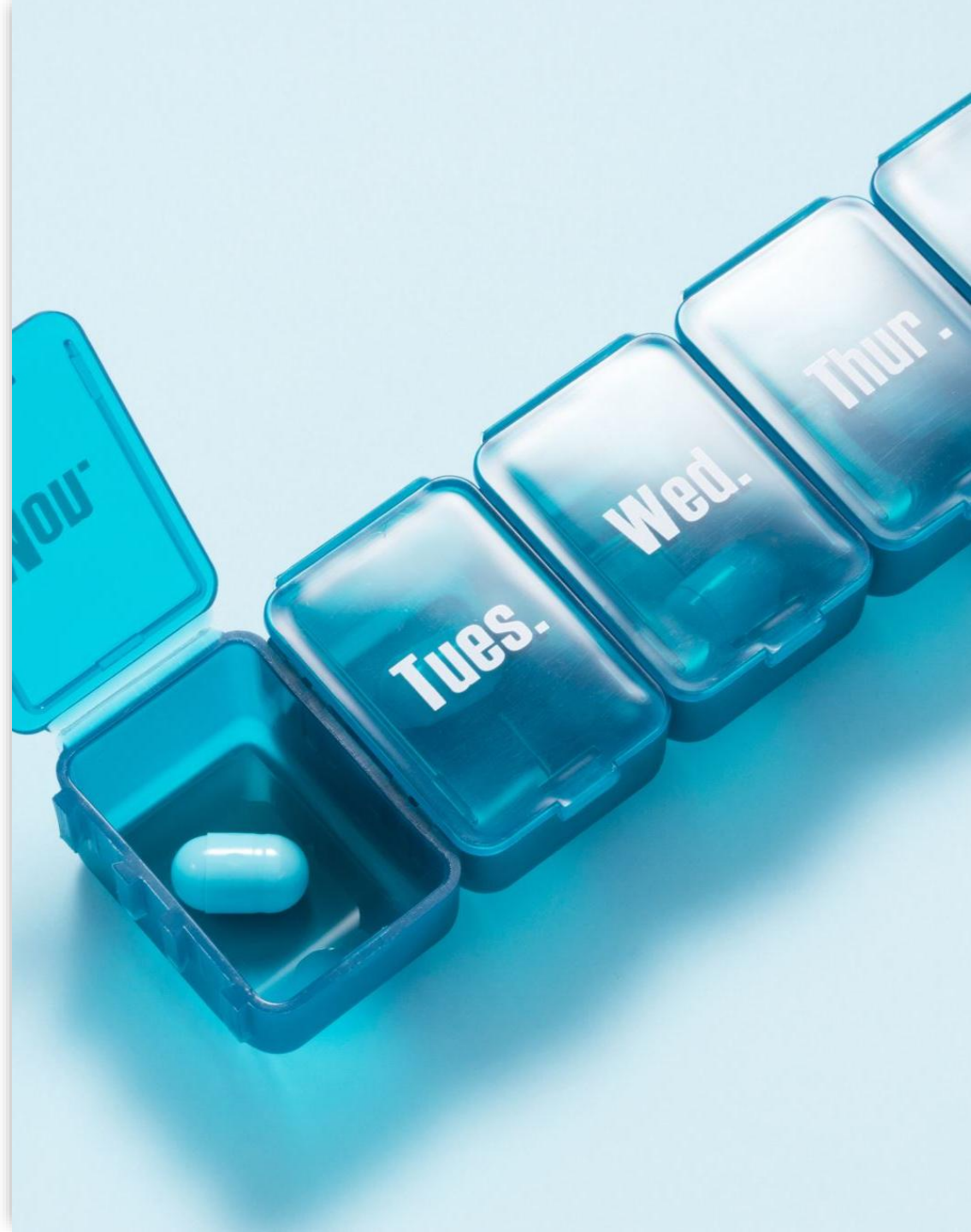
Secure and scale your AI services in a cloud-native way

See

See a live demo of an end-to-end LLM deployment

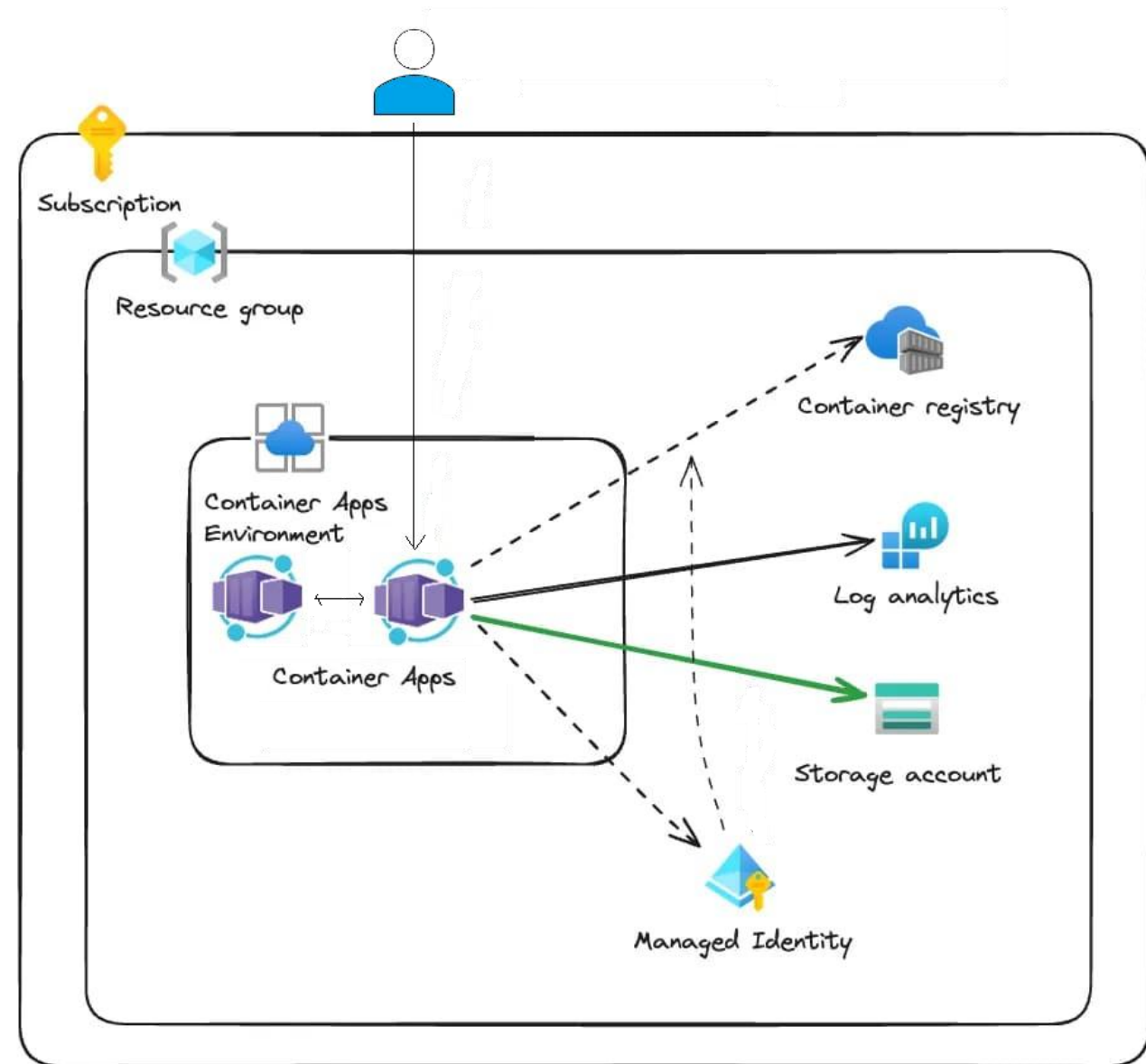
Azure Container Options

- Azure Container Apps
- Azure App Service
- Azure Container Instances
- Azure Kubernetes Service
- Azure Functions
- Azure Red Hat OpenShift



Azure Container Apps

- Create Resource Group
- Create Container Registry
- Create Storage Account
- Create Container Environment
- Create Container App



Storage mounts in Azure Container Apps

Storage type	Description	Persistence	Usage example
Container-scoped storage	Ephemeral storage available to a running container	Data is available until container shuts down	Writing a local app cache.
Replica-scoped storage	Ephemeral storage for sharing files between containers in the same replica	Data is available until replica shuts down	The main app container writing log files that a sidecar container processes.
Azure Files	Permanent storage	Data is persisted to Azure Files	Writing files to a file share to make data accessible by other systems.

Azure Files

- Container can mount multiple Azure Files volume
- Multiple containers can mount the same Azure Files volume
- Must use a vNet if you want to use NFS
- SMB can be used without a vNet

Basic Ollama CLI Commands

Command	Description
<code>ollama serve</code>	Starts Ollama on your local system.
<code>ollama create <new_model></code>	Creates a new model from an existing one for customization or training.
<code>ollama show <model></code>	Displays details about a specific model, such as its configuration and release date.
<code>ollama run <model></code>	Runs the specified model, making it ready for interaction.
<code>ollama pull <model></code>	Downloads the specified model to your system.
<code>ollama list</code>	Lists all the downloaded models. The same as <code>ollama ls</code>
<code>ollama ps</code>	Shows the currently running models.
<code>ollama stop <model></code>	Stops the specified running model.
<code>ollama rm <model></code>	Removes the specified model from your system.
<code>ollama help</code>	Provides help about any command.

A close-up, artistic shot of a camera lens aperture. The lens is dark, and the aperture blades are visible, creating a series of concentric circles. A bright, colorful light source is visible through the aperture, creating a lens flare effect with streaks of green, blue, and yellow. The word "Demo" is centered over the aperture in a white, sans-serif font.

Demo



Resources

- Azure Container Apps documentation
<https://learn.microsoft.com/en-us/azure/container-apps/>
- Quickstart: Deploy your first container app using the Azure portal
<https://learn.microsoft.com/en-us/azure/container-apps/quickstart-portal>