Self-Hosting AI LLMs

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

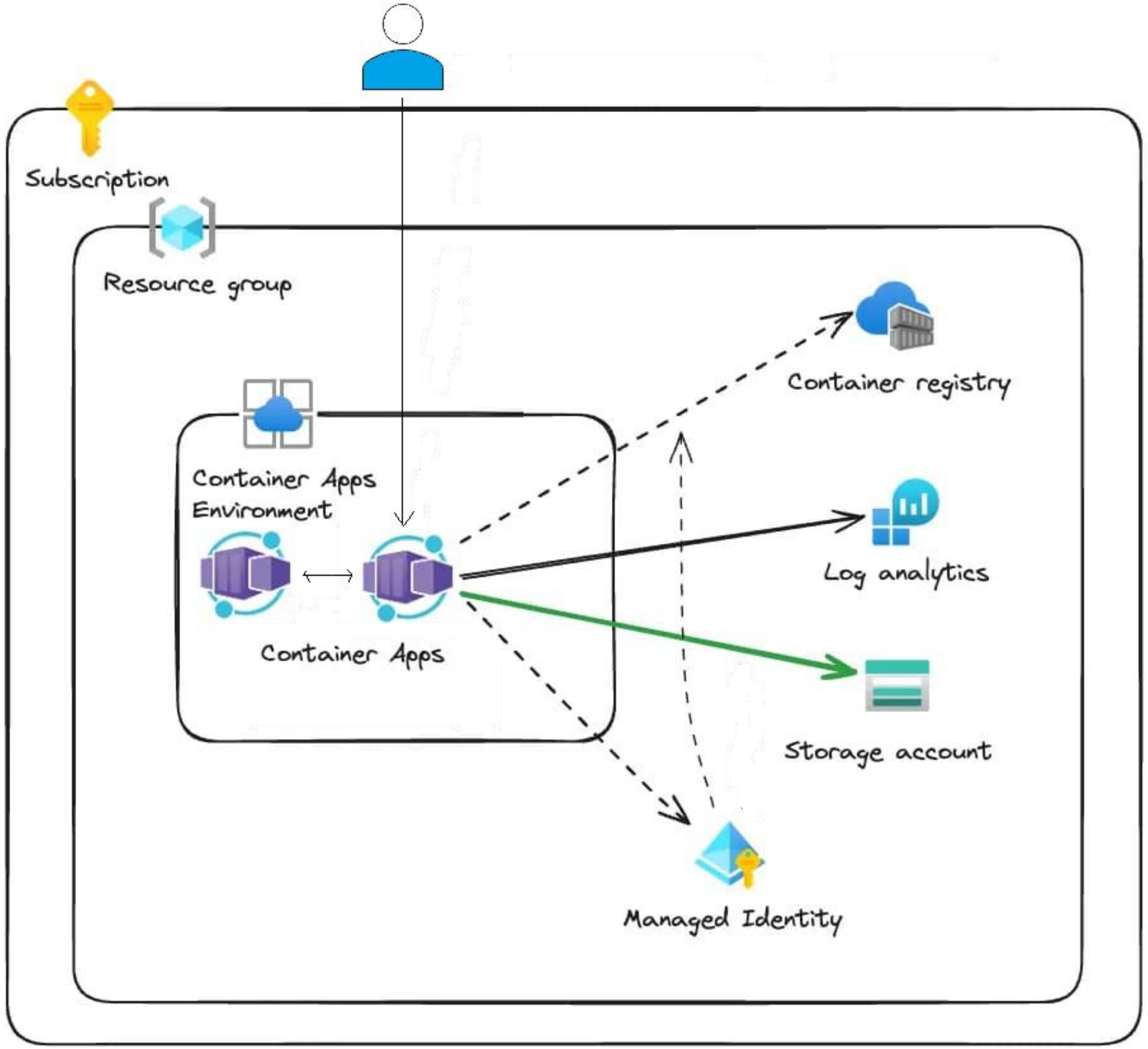
See a live demo of an end-to-end LLM deployment

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.

https://learn.microsoft.com/en-us/azure/container-apps/storage-mounts?tabs=smb&pivots=azure-cli

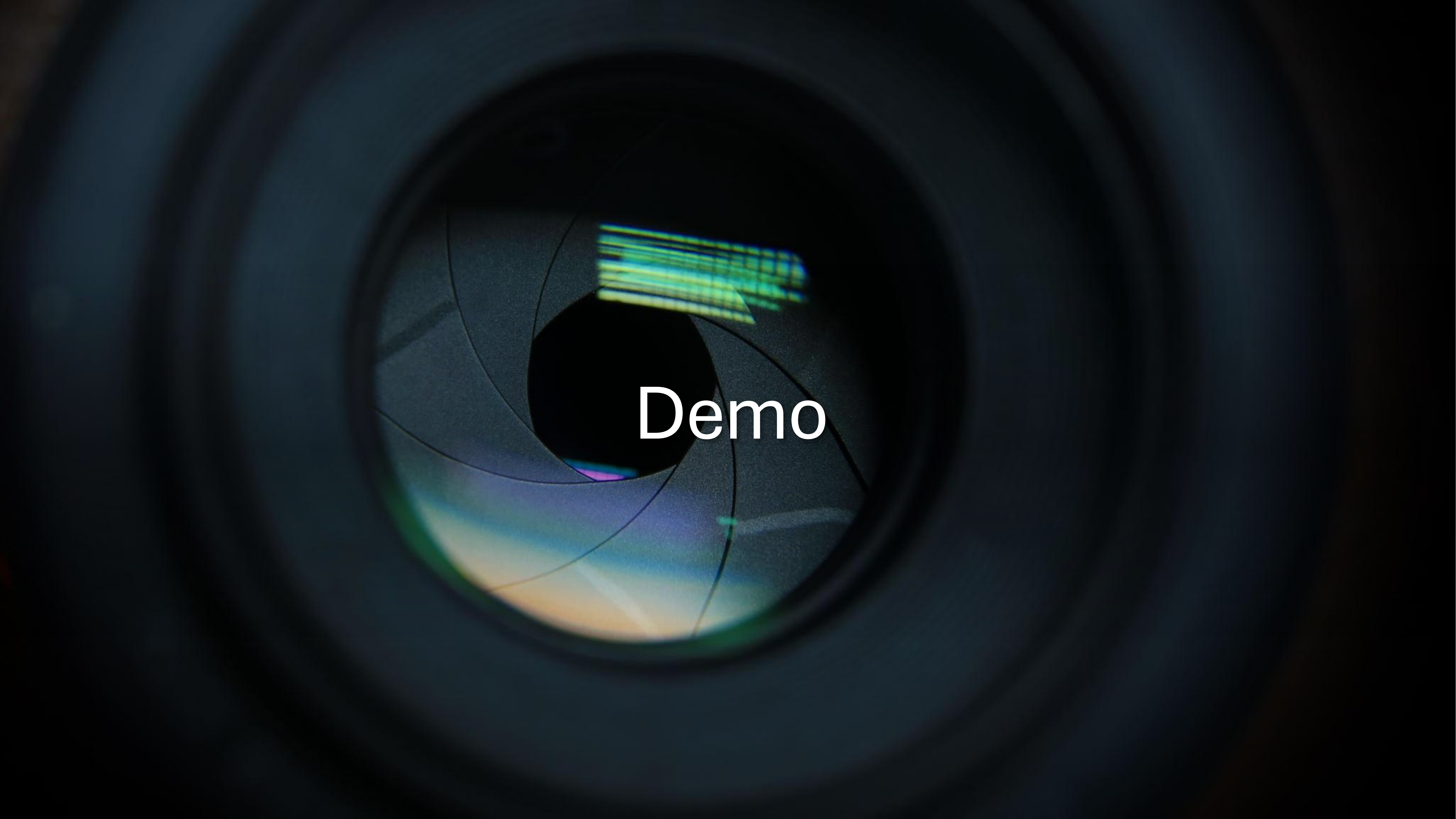
Azure Files

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

https://www.glukhov.org/post/2024/12/ollama-cheatsheet/



Demo