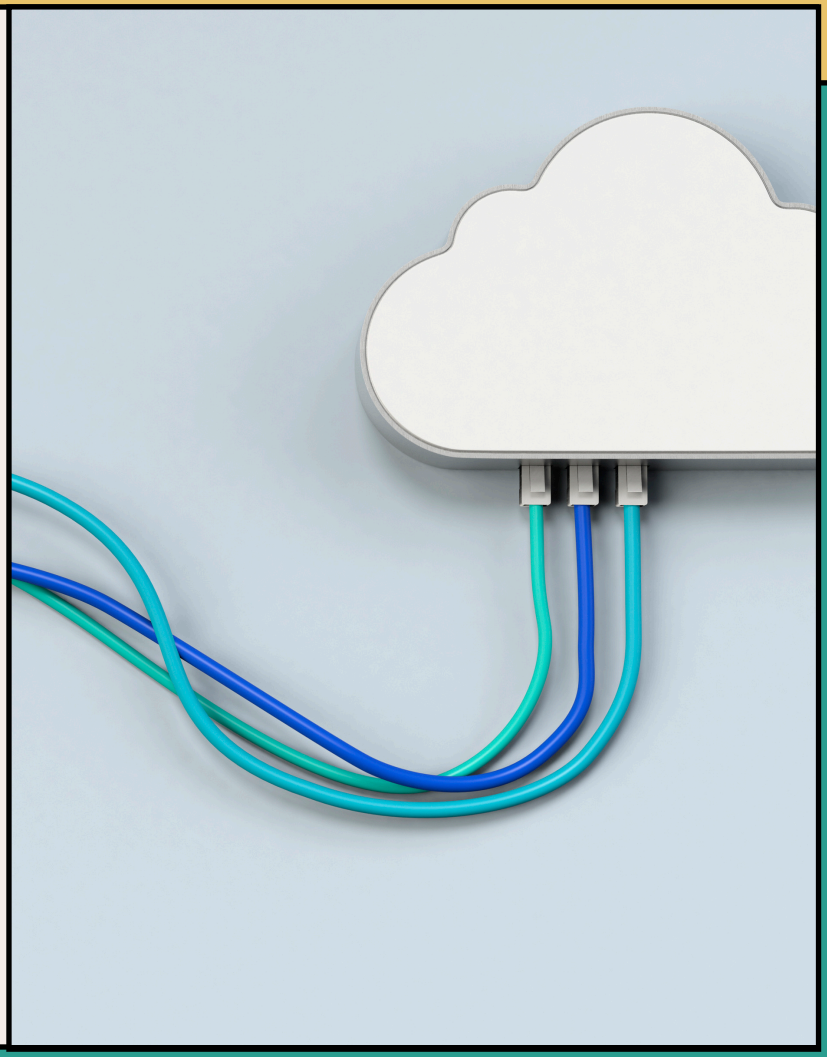
Delivering Transformational AI Apps with AKS

Maryam Tavakkoli
Senior Cloud Engineer @ RELEX Solutions

Microsoft AI Tour, Helsinki, October 2024



Who Am I?

- **Maryam Tavakkoli**
- Senior Cloud Engineer **@ RELEX Solutions**
- CNCF Ambassador
- Microsoft MVP
- Kubernetes & CNCF Meetup Co-organizer
- **LinkedIn:** maryam-tavakkoli
- **Medium:** @maryam.tavakoli.3

Agenda

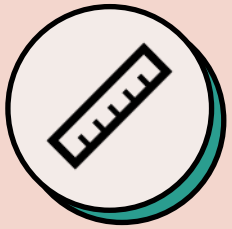
1. Introduction
2. Definitions
3. Benefits of AKS for AI Applications
4. Demo
5. Summary

Introduction

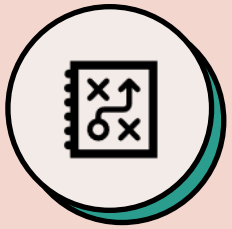
Artificial Intelligence (AI) is transforming industries by automating tasks, enhancing decision-making, and enabling innovation. However, deploying and managing AI workloads comes with significant challenges—scaling, cost, security, and ensuring high availability.
Today, we'll see how Azure Kubernetes Service (AKS) simplifies these challenges and transforms AI app deployment.

What is Azure Kubernetes Service (AKS)?
AKS is a managed Kubernetes service provided by Microsoft Azure that simplifies deploying, managing, and scaling containerized applications. Kubernetes, is an open-source platform for orchestrating containers, allowing you to automate deployment, scaling, and management of your applications.

Benefits of AKS for AI Applications



Scalability & Flexibility



High Availability and Resilience

Cost Efficiency

Security

Integration with Azure AI Services

Scalability and Flexibility

AI models can experience variable traffic. With AKS, you can easily scale your AI application.

Node Autoscaling: AKS automatically scales the number of cluster nodes, so when traffic spikes, additional nodes are added to meet the demand.

Horizontal Pod Autoscaler (HPA): Dynamically adjust the number of replicas based on CPU and memory usage, ensuring your app can handle high-demand periods while scaling down when demand decreases.

Support for GPU and Specialized Nodes

AKS allows you to create clusters with GPU-enabled nodes. This is essential for deep learning models that require intense computation.

High Availability and Resilience

Built-in Disaster Recovery and Fault Tolerance: AKS is designed to provide high availability for your AI applications. You can distribute your workloads across multiple regions or availability zones to ensure they remain operational even if parts of the infrastructure fail.

Rolling Updates and Zero-Downtime Deployments: With AKS, you can update your AI models or underlying infrastructure with zero downtime. This is critical when you need to release new versions of AI models while maintaining service availability.

Cost Efficiency

Optimized Resource Usage: With AKS, you can use **cluster autoscaling** and spot instances to optimize resource costs. For AI models that do not require constant uptime, spot instances can be used to drastically reduce compute costs.

Right-Sizing AI Workloads: AKS allows fine-grained control over how much CPU, memory, and GPU resources are allocated to different AI workloads. This ensures that resources are used optimally, reducing over-provisioning and costs.

Security

Security at Scale: AKS provides built-in security features, such as **Azure Active Directory (AAD)** integration, role-based access control (RBAC), and network policies. This is critical when deploying AI applications that may process sensitive data. You can control who has access to specific resources and ensure that models are only accessible by authorized personnel or systems.

Integration with Azure AI Services

Integration with Azure OpenAI: By combining AKS with Azure OpenAI services, you can easily deploy and manage natural language models like GPT-4. These pre-trained models can be containerized and deployed on AKS, allowing developers to leverage OpenAI's powerful capabilities without the need to train models from scratch.

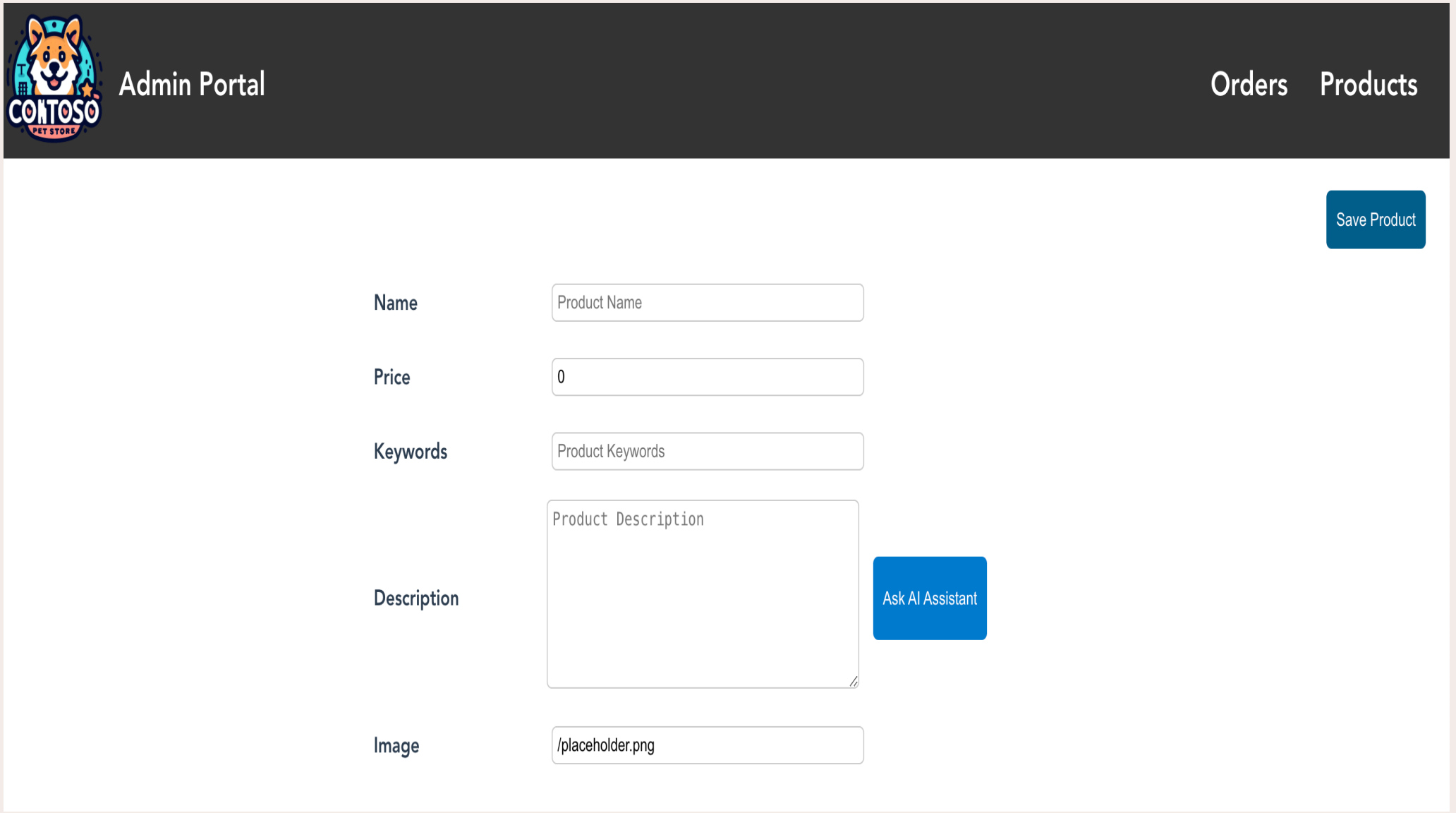
Demo Agenda

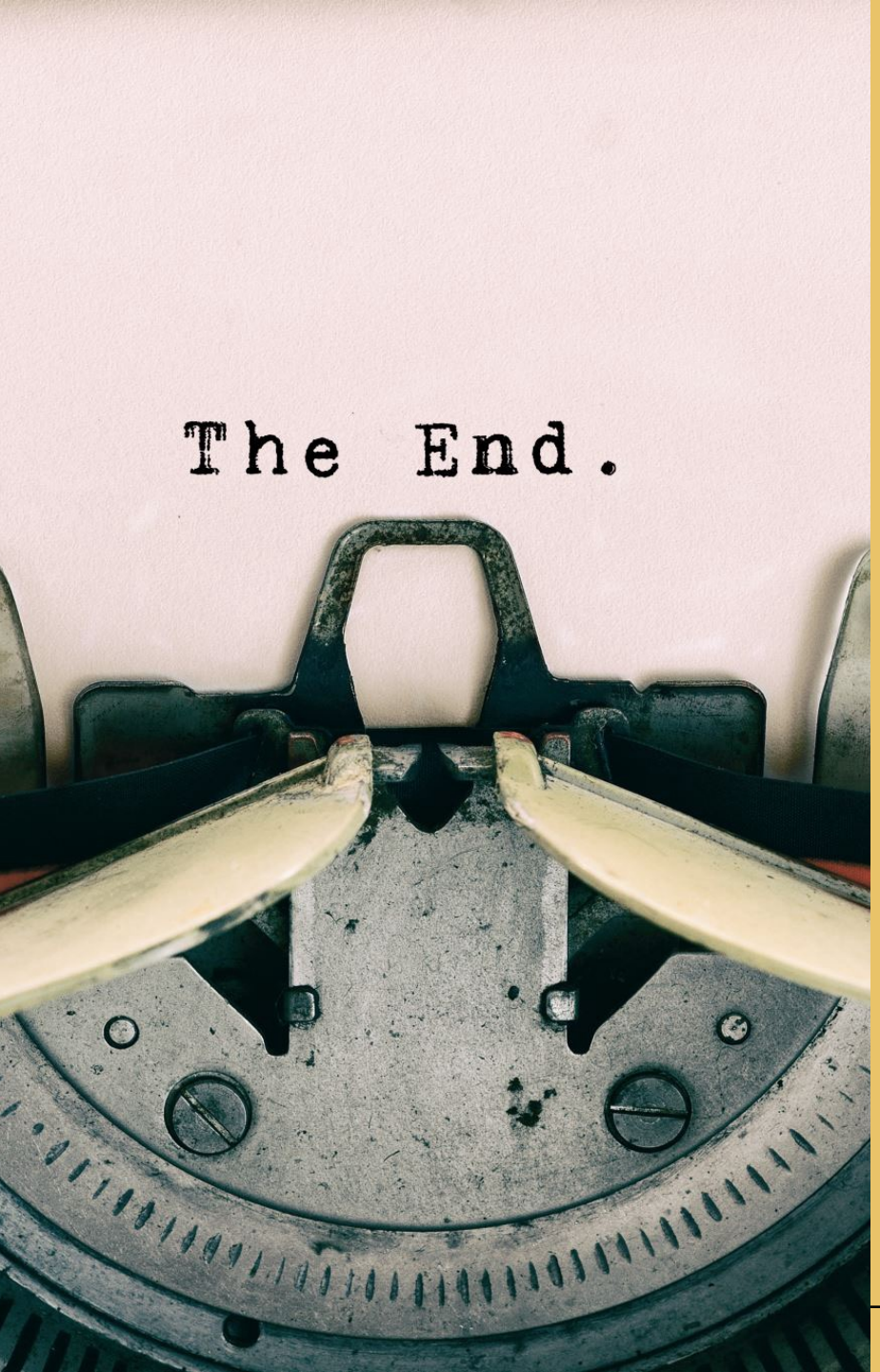
Deploying an AI Model on AKS with OpenAI

Deploy an application that uses OpenAI on Azure Kubernetes Service (AKS)

Aks Demo application: Pet supply store

The role of OpenAI is to reduce the workload for employees by automating the task of recording product details. The OpenAI model helps generate product descriptions.


Admin Portal
Orders
Products
Save Product
Name
Product Name
Price
0
Keywords
Product Keywords
Description
Product Description
Ask AI Assistant
Image
/placeholder.png



Summary

- **Scalability**: AKS automatically adjusts resources.
- **High availability:** You can distribute AI workload on multiple regions or availability zone
- **Cost Efficiency**: Aks has resource optimization features like autoscaling and spot instances.
- **Integration**: AKS integrates with Azure OpenAI.
- **Security & Reliability**: Aks has built-in security features.

Thank you

Maryam Tavakkoli

Senior Cloud Engineer @ RELEX Solutions | CNCF Ambassador | Microsoft MVP |...