Microsoft

Generative AI with JavaScript

Azure tools & services for hosting and storing AI apps

Agenda

- AI Apps
- Tooling, Azure Developer CLI, azd
- Architecture, container-based, serverless
- Manage APIs in production, Azure API Management

AI Apps

Chat apps

- Create chat apps, integrate Azure Open AI

RAG Apps

- Azure AI Search
- Azure Cosmos DB
- Azure Open AI
- Azure Blob Storage

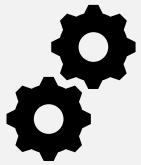
Assistant/agent apps

- RAG App + Goal oriented
- Azure AI Search, Azure Cosmos DB, Azure Blob Storage
- Azure Functions

Taking your AI to production

Deploy whole solutions to the cloud

- A tool like Azure Developer CLI helps you deploy a whole solution

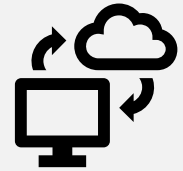


Select architecture for your app

- 2 commonly used architectures: **serverless** or **container based**
- Azure Functions
- Azure Static Apps
- Azure Container Apps

Monitor, scale and protect your APIs

- Secure, managed identity
- Scale, load balancing
- Error management, short circuit patterns and more
- Monitor, logs, token usage and more

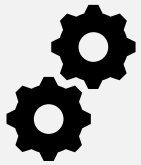


Tooling

Azure Developer CLI (azd)

Accelerate process: from local env to Azure

- Open source CLI tool
- Helps you deploy a whole solution



Provides

- Developer friendly commands
- Support in terminal, editor IDE and CI/CD

Workflow

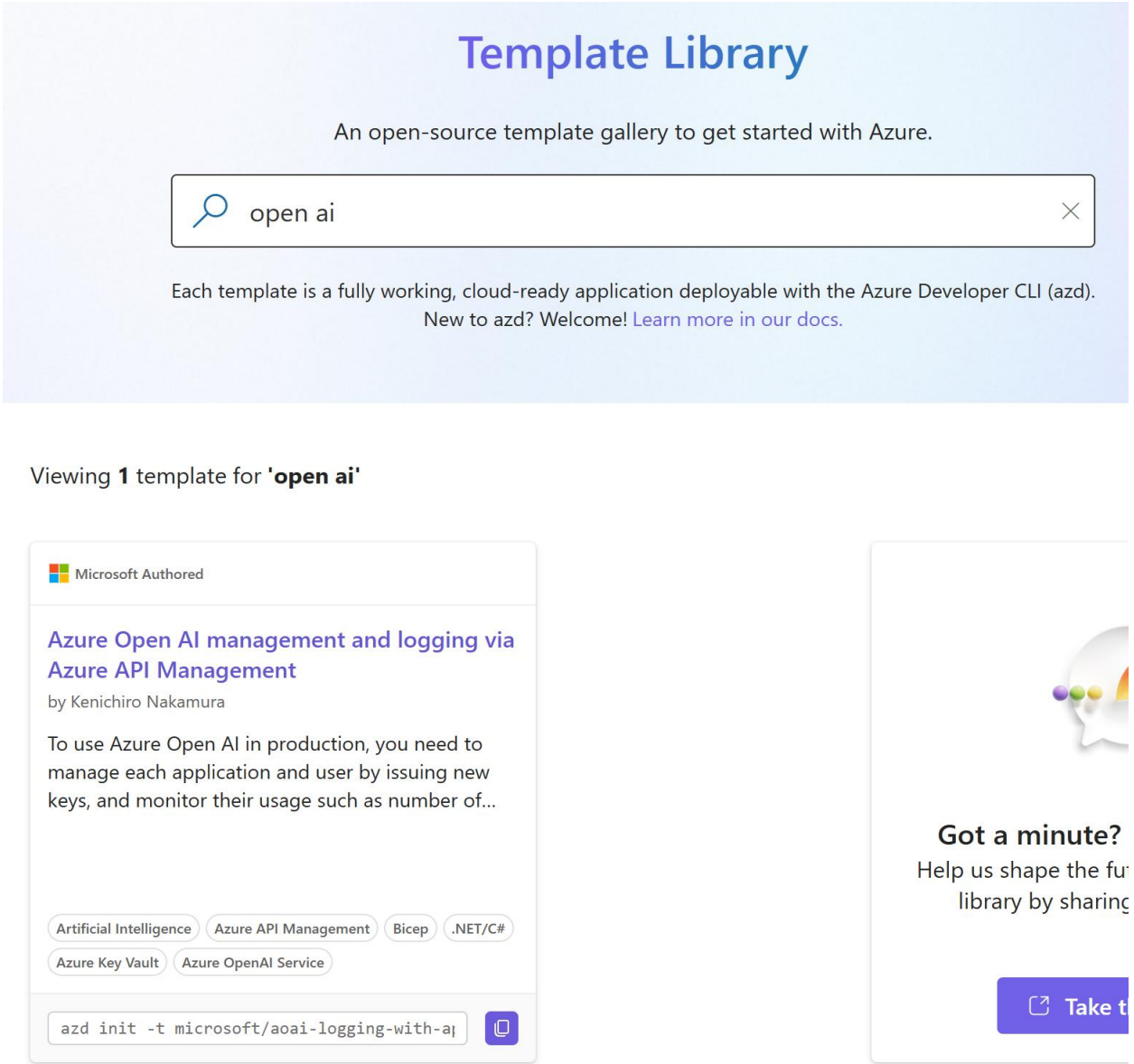
- **Select** Azure Developer CLI Template
- **azd init :** init project
- **azd up:** package, provision and deploy app
- **azd deploy:** update code or infra, deploy changes

Get started with azd

Find a template:

- aka.ms/azd/ai-templates

```
> azd init -t microsoft/aoai-logging-with-apim
```



Architecture

Architecture

Serverless

- **Event-driven:** code that runs when something important happens
- Azure Functions
- Host Static Apps with serverless backend

Container-based

- Scale from a few to millions of users in the cloud
- Containers runs the same in your local environment as in production

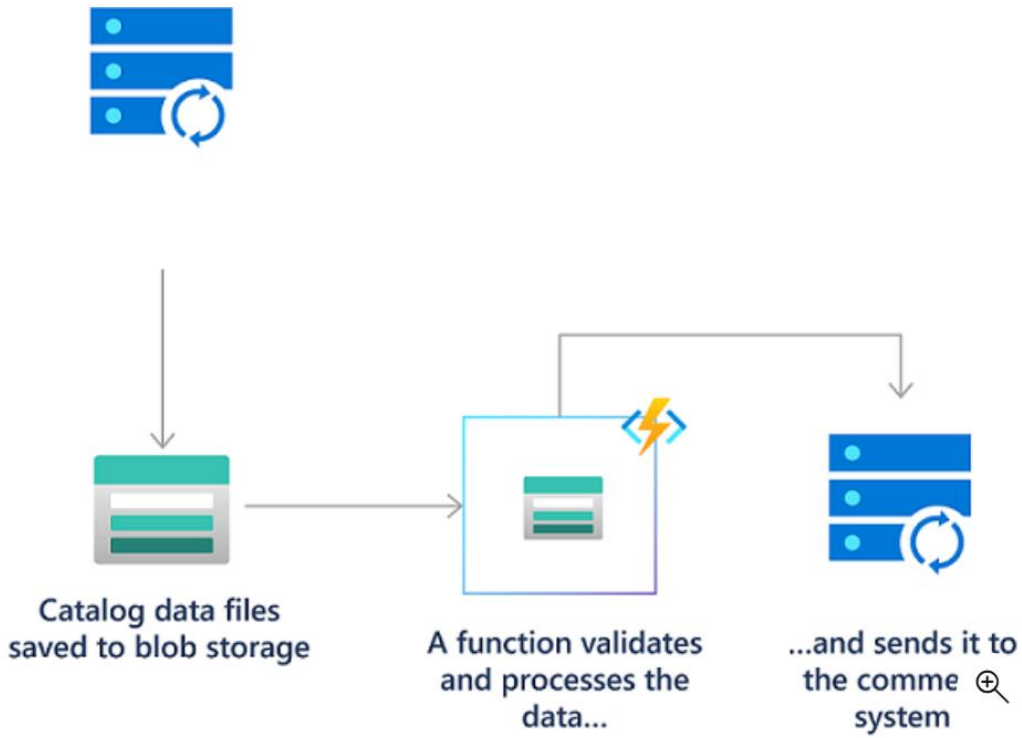
Serverless with Azure Functions

Definition

- Run event-triggered code without having to explicitly provision or manage infrastructure

Integrate

- **Triggers and bindings** allows easy integration with other services, e.g HTTP triggers, timer triggers, and many more
- **Combine** Azure products e.g Azure Functions with Azure Container Apps to deploy modern apps and microservices using serverless containers



Azure Static Apps

How it works

- **Automatically deploys** full-stack web apps to Azure from a code repository
- **Tailored** to a developer's daily workflow
- **Apps** are built and deployed based on code changes
- **Serverless backend** with Azure Functions

Features

- Global hosting
- API Functions
- Streamlined build and deployment
- Seamless staging environments, dev experience and CI/CD

Azure Functions and AI

Azure Open AI extension for Azure Functions

Action	Trigger/binding type
Use a standard text prompt for content completion	Azure OpenAI text completion input binding
Respond to an assistant request to call a function	Azure OpenAI assistant trigger
Create an assistant	Azure OpenAI assistant create output binding
Message an assistant	Azure OpenAI assistant post input binding
Get assistant history	Azure OpenAI assistant query input binding
Read text embeddings	Azure OpenAI embeddings input binding
Write to a vector database	Azure OpenAI embeddings store output binding
Read from a vector database	Azure OpenAI semantic search input binding

https://learn.microsoft.com/azure/azure-functions/functions-bindings-openai

Azure Container Apps

Serverless platform

- Maintain less infrastructure
- Save costs, while running container-based apps

Use for

- API endpoints
- Background processing jobs
- Event-driven processing
- Microservices

Platform

- Automatic scaling based on
  - HTTP traffic,
  - Event-driven processing
  - CPU or memory load
- Supports continuous deployment, code push, app deploy, image build & push, and revisions (immutable snapshot of container app)

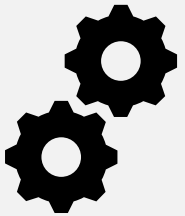
Azure Container Apps

Integrates with Dapr

- Rich microservices programming model e.g Observability, pub/sub, service-to-service invocation with mutual TLS, retries and more

Choose your runtime

- Use any runtime, programming language, or development stack of your choice with Azure Container Apps.
- *Currently, only Linux-based container images can be used*



Scale, monitor and protect app

Azure API Management

What

- **Turn-key solution** for publishing APIs to external and internal customers

API Gateway

- **Allows** you to quickly create consistent and modern API gateways for existing back-end services hosted anywhere and
- **Analyse and optimise** your APIs

E2E Management

- **Secure**, scalable, and reliable way of publishing, consuming, and managing the execution of APIs over the Azure platform
- **Provides** all the important tools essential for the end-to-end management of those APIs

Azure API Management and AI

Companion service

- Azure API Management can be a great companion service
- Managing generative APIs means paying attention to *cost, fairness of used resources, security*, and more.

Policies

- **Token limit policy**, manage the token usage across multiple applications. Ensures that a single application does not consume the entire token quota
- **Emit Token Metric Policy**, track token usage across different applications. Helps in calculating cross-charges for multiple applications or teams using Azure OpenAI Service models
- **Load balancer and Circuit Breaker**, distribute the load across multiple Azure OpenAI endpoints. Ensures that the committed capacity in Provisioned Throughput Units (PTUs) is exhausted before falling back to pay-as-you-go instances
- **Semantic Caching Policy**, caches responses from Generative AI models. Improves the performance of your applications by reducing the number of calls to the backend services

Summary

Azure Developer Tools	• **azd** for deploying whole solution
Select your architecture	• **Serverless**: Azure Static Web Apps, Azure Functions • **Container-based**: Azure Containers Apps
API Management	• Azure API Management (APIM) • Use APIM policies to secure, scale and keep track of cost and usage in production

Microsoft

Resources

- APIM with OpenAI sample aka.ms/genai-apim
- GenAI gateway with APIM aka.ms/genai/apim-ai-gateway
- Azure for JS developers aka.ms/js/azure