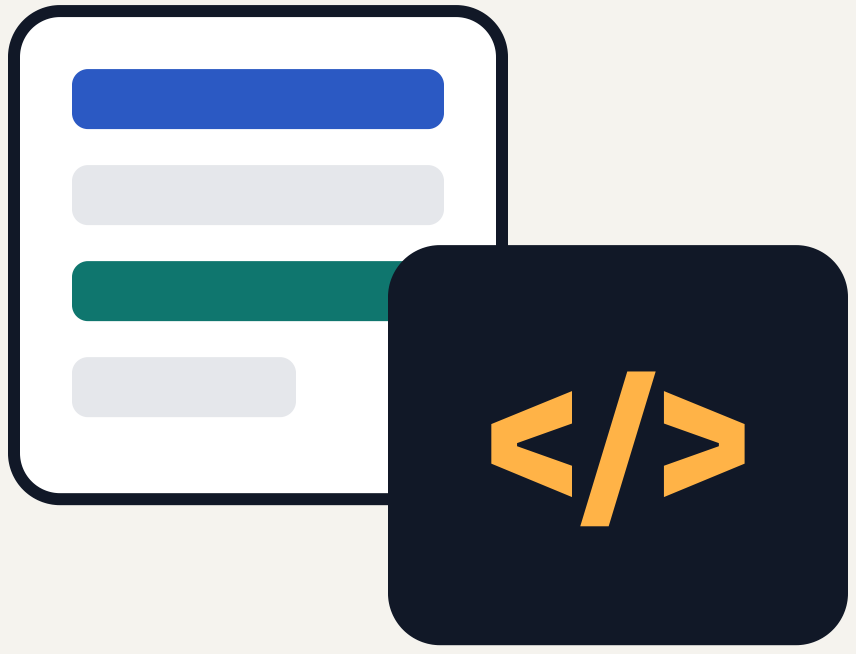
GUÍA PARA GUARDAR



¿Prompt Agent o Hosted Agent?

Los dos tipos de agente de Microsoft Foundry: configurar o traer tu propio código. Qué te da cada uno, cuánto cuesta y cuándo dar el salto.

Fernando Prada · AI Architect **fernandopradagorge.dev**

LAS DOS FORMAS

Configurar o programar

1

Prompt Agent

Eliges el modelo, escribes las instrucciones y añades herramientas. Foundry lo ejecuta: sin código propio que mantener ni contenedores.

2

Hosted Agent

Tu código y tu framework en un contenedor (o un .zip que Foundry convierte en imagen). Foundry lo ejecuta con endpoint, escalado e identidad propia.

Los dos admiten las mismas funcionalidades: herramientas (OpenAPI, MCP, A2A...), memoria e identidad. Lo que cambia es quién escribe y opera la lógica.

PROMPT AGENT

Pura configuración

Se crea en	El portal de Foundry, el SDK, la API REST o azd.
Código	Ninguno que mantener: Foundry ejecuta el agente.
Cómputo	Gestionado por Foundry, escala con las peticiones.
Coste	Llamadas al modelo y uso de herramientas.
Extra	Arnés de Copilot (preview) para flujos de varios pasos con herramientas y archivos.
Mejor para	Empezar rápido y agentes de producción sin orquestación propia.

HOSTED AGENT

Tu código, gestionado

Frameworks

Agent Framework, LangGraph, OpenAI Agents SDK, Anthropic Agent SDK o código propio, en Python o C#.

Las mismas herramientas

OpenAPI, MCP, A2A, Code Interpreter, Web Search, Azure AI Search, Browser Automation y más, conectadas a través de un Toolbox. También la memoria de Foundry (preview).

Lo que pone Foundry

Endpoint, escalado, identidad de Entra propia, estado de sesión y observabilidad. Protocolos Responses e Invocations para webhooks u otros formatos.

HOSTED AGENT

Así corre cada sesión

Aislamiento	Cada sesión en su propio sandbox aislado a nivel de máquina virtual.
Archivos	$HOME y /files se conservan entre turnos.
Inactividad	Se apaga tras 2 min a 4 h (15 min por defecto) y al volver recupera el estado.
Retención	La sesión se borra tras 30 días sin uso.
Tamaños	Por sesión: 0,5 vCPU y 1 GiB, 1 vCPU y 2 GiB, o 2 vCPU y 4 GiB.
Coste	Inferencia y herramientas, más la CPU y memoria de las sesiones activas.

Fuente: Microsoft Learn · Foundry Agent Service

PARA ARQUITECTOS

Red, regiones y tamaño

Red privada

Los Prompt Agents usan un pequeño grupo fijo de IP. Cada sesión de un Hosted Agent consume una IP de tu subred: Microsoft recomienda /24 para producción (/27 como mínimo).

Regiones

Los Hosted Agents están en 31 regiones, entre ellas Spain Central, West Europe, France Central y Sweden Central.

Dimensionar

Si los picos pasan del 70 % de lo asignado, sube la siguiente versión; si quedan muy por debajo, bájala para pagar menos.

COMPARATIVA

Cara a cara

	Prompt Agent	Hosted Agent
Se crea con	Portal, SDK, REST o azd	Tu framework o tu código
Código	Ninguno	Tu lógica
Cómputo	Gestionado	Contenedor gestionado
Herramientas	Se añaden al agente	Las mismas, vía Toolbox
Memoria	Memory store (preview)	Memory store (preview)
Identidad	Entra	Entra, una por agente
Coste	Inferencia y herramientas	Lo mismo, más el cómputo
Mejor para	Empezar y producción sin código propio	Tu código y tu orquestación

LA CHULETA

¿Cuál uso?

Prompt Agent

Por defecto. Si la configuración cubre el caso, quédate aquí.

Hosted Agent

Tu propio código, protocolos propios, control de CPU y memoria o trabajos largos con estado.

Responses API

¿El agente ya vive en tu aplicación? Llama a la Responses API directamente, sin guardarlo en Foundry.

Más guías gratis en fernandopradagorge.dev · Fernando Prada · oct. 2026
