Azure AI Foundry Architecture

Secure, Network-Isolated AI Services with Private Endpoints

Virtual Network: ai-foundry-vnet-eastus2 (192.168.0.0/16)

agent-subnet

192.168.0.0/24
Delegated to Microsoft.App/environments
AI Agent Runtime Environment
Network Injection Target

pe-subnet

192.168.1.0/24
Private Endpoints Subnet
Network isolation for services

Private Endpoints (in pe-subnet)

- AI Services (192.168.1.8-10)
- Search Service (192.168.1.6)
- Cosmos DB (192.168.1.4-5)
- Storage Account (192.168.1.7)

Private DNS Zones

- privatelink.cognitiveservices.azure.com
- privatelink.openai.azure.com
- privatelink.services.ai.azure.com
- privatelink.search.windows.net
- privatelink.documents.azure.com (Cosmos DB)
- privatelink.blob.core.windows.net

Azure AI Services Hub

ai-foundry-hub

AI Foundry Agent Service:
- Project: ai-foundry-project
- Use Any Model: GPT-4o, Llama, etc.
- Agent Capability Host
- Thread Orchestration & State Management
- Autonomous Tool Calling & Workflows
- Enterprise Identity & RBAC Integration
- Content Safety & Policy Enforcement
- Full Observability & Tracing

Connected Resources:
- Vector Store: ai-search-service
- Storage: aifoundrystorage
- Thread Storage: ai-cosmos-db
- Network: Injected to agent-subnet

Azure AI Search

ai-search-service

Configuration:
- Standard SKU (1 replica, 1 partition)
- Vector Store for RAG
- System Assigned Identity
- AAD Authentication
- Public Network Access: Disabled
- Semantic Search: Disabled

Azure Cosmos DB

ai-cosmos-db

Purpose:
- Thread Message Storage
- Agent Entity Store
- Session Consistency
- Autoscale (100-1000 RU/s)
- SQL API with Role-Based Access
- Geo-redundant Backup

Azure Storage Account

aifoundrystorage

Services:
- Blob, File, Queue, Table Services
- Zone-Redundant Storage (ZRS)
- Agent Blob Store Containers
- TLS 1.2+ Only
- No Public Blob Access
- Hot Access Tier

Enterprise-Grade AI Agent Platform: Azure AI Foundry Agent Service provides a production-ready foundation with composable agents that can use any model (GPT-4o, Llama, etc.), autonomous tool calling, thread orchestration, and full lifecycle management. All services are network-isolated with private endpoints, enterprise identity integration, content safety enforcement, and comprehensive observability for reliable agent deployment.

AI/ML Services | Search & Vector Store | Database Services | Storage Services | Network Subnets | Private Endpoints