Workshop AI (day 2) Microsoft-AGIC-UniPadova

UI di Azure AI Foundry

Azure AI Foundry è una piattaforma generativa progettata per lo **sviluppo di applicazioni basate su AI generativa**, tramite un ambiente unificato e scalabile. La piattaforma mette a disposizione degli utenti una vasta gamma di modelli, strumenti di sviluppo e risorse che consentono di costruire, testare e distribuire applicazioni AI avanzate.

Creazione del Resource Group

Il primo passo per iniziare a lavorare in Azure AI Foundry è creare un ***progetto***. Si tratta di una vera e propria ***Resource*** Azure e in quanto tale, al momento della creazione, deve essere associata a un ***Resource Group***, esistente o da creare.

È importante scegliere con attenzione la ***region*** in cui si crea il progetto, poiché questa determina la disponibilità di alcune funzionalità e servizi.

I progetti vengono creati all'interno di un ***Hub***, che è la risorsa principale per la governance e l'organizzazione del lavoro in Foundry. Tutti i progetti associati a uno stesso hub ne ereditano configurazioni di sicurezza, accessi e connessioni.

Introduzione all'interfaccia di AI Foundry

Una volta creato un progetto, si accede all'ambiente di lavoro all'interno del portale Azure AI Foundry. Vediamo le principali sezioni discusse nel Workshop:

Azure AI Foundry Project

│

├── Overview

```
│   └── Management Center
│       ├── Overview
│       ├── Quota
│       ├── Users
│       ├── Models + Endpoints
│       └── Connected Resources
│
├── Model Catalog
├── Playground
└── AI Services
```

Overview

La sezione *Overview* è il punto di ingresso principale di ogni progetto in Azure AI Foundry. **Qui si trovano le informazioni essenziali** del progetto, come il nome, la regione e la ***Project Connection String***, necessaria per collegarsi alla risorsa Foundry da codice o SDK. Dalla stessa pagina si accede al ***Management Center****,* che permette la gestione di utenti, endpoint, connections e altro.

Dal pannello laterale a sinistra nel management center è possibile navigare tra le pagine seguenti:

Quota

La sezione *Quota* permette di visualizzare e gestire la capacità disponibile per i deployment AI del progetto , in base alla subscription e alla regione. Da qui puoi monitorare l'uso delle risorse, regolare le assegnazioni e, se necessario, richiedere un aumento per supportare carichi maggiori.

Overview (hub o progetto)

Qui si trovano le **informazioni e le proprietà principali** dell'hub o del progetto.

Users (hub o progetto)

Nella sezione *Users* è possibile **visionare, modificare, aggiungere utenti** e i ruoli a questi associati.

Models + Endpoints (hub o progetto)

La sezione *Models + Endpoints* consente di **gestire tutti i modelli distribuiti all'interno del progetto (o hub)** Foundry. Per ogni modello vengono mostrati nome, versione, stato, tipo di content filter, retirement date (se prevista) e tipo di deployment. Da qui è possibile accedere agli **endpoint**, modificarne la configurazione o distribuirne di nuovi.

Connected resources (hub o progetto)

Nella sezione *Connected Resources* puoi gestire tutte le **risorse esterne collegate al progetto**. Alcuni esempi sono i collegamenti con *Azure OpenAI Service*, *AIServices, Azure AI Search e Azure Blob Storage*. Per ogni connessione vengono mostrate varie informazioni, tra cui il tipo di servizio, l'endpoint di destinazione, la chiave di autenticazione e il livello di accesso (condiviso tra progetti o specifico del progetto). Le connessioni possono usare diversi metodi di autenticazione, ad esempio ***API Key*** o ***SAS Token***.

Compute (hub)

La sezione *Compute* permette di gestire le risorse di calcolo associate all'hub. Qui è possibile creare, avviare, arrestare o eliminare **compute instances**, ovvero macchine virtuali dedicate che consentono agli utenti di eseguire operazioni di calcolo in un ambiente sicuro e containerizzato. È anche possibile accedere a risorse ***serverless*** nella scheda dedicata.

Model Catalog

Azure AI Foundry mette a disposizione un ampio *Model Catalog* che consente di **esplorare e confrontare modelli AI** in base a parole chiave, filtri e benchmark di performance. Per ogni modello è disponibile una scheda dettagliata con informazioni tecniche, versioni, tipi di dati supportati e altre informazioni. Il catalogo è un punto di partenza ideale per scegliere il modello più adatto al proprio caso d'uso.

Playground

Il *Playground* è l'**ambiente interattivo** di Azure AI Foundry dove **puoi testare e conversare con un modello** chat appena distribuito, anche senza integrare ancora i tuoi dati. Dopo aver selezionato e deployato un modello in un progetto, puoi aprirlo direttamente nel Playground per iniziare a interagirci.

Puoi personalizzare il comportamento dell'assistente impostando un **messaggio di sistema**, aggiungendo eventualmente **messaggi di sicurezza** preconfigurati, e osservare come il modello risponde alle richieste in un contesto di base. In questa modalità, il modello non è ancora "**grounded**" sui tuoi dati, quindi risponderà in modo generico.

Il Playground è ideale per test rapidi, prototipazione di prompt e valutazioni preliminari prima di passare all'integrazione con dati aziendali o alla distribuzione in ambienti più complessi.

AI Services

Azure *AI Services* offre un **insieme di modelli e API preconfigurati**, pronti all'uso, che permettono di integrare rapidamente funzionalità di intelligenza artificiale nelle proprie applicazioni. Alcuni esempi di servizi sono ***Azure OpenAI***, ***Speech***, ***Vision*** e ***Content Safety***.

Demo chat completion app (in locale)

Vediamo un primo utilizzo del SDK di Azure AI Foundry.

1. Connessione al progetto Foundry

Il primo passo è importare le librerie che gestiscono la autenticazione e la connessione al progetto che abbiamo precedentemente creato su Azure AI Foundry.

```python
from azure.ai.projects import AIProjectClient
from azure.identity import DefaultAzureCredential
```

Quindi possiamo stabilire la connessione al progetto tramite la ***connection string***, che si può reperire nella sezione *Overview* della UI di Foundry.

```python
project_connection_string = "..."
project = AIProjectClient.from_connection_string(
    conn_str=project_connection_string,
credential=DefaultAzureCredential()
)
```

2. Inizializzazione del client

Creiamo un client per inviare richieste al modello deployato nel progetto.

```python
chat = project.inference.get_chat_completions_client()
```

3. Funzione per inviare messaggi al chatbot (e ricevere le risposte)

Creiamo una funzione che costruisce il prompt e invia un messaggio al modello.

> ***messages***: è una lista di dizionari, ciascuno rappresenta **un messaggio nella conversazione tra l'utente e il modello**.
>
> ***context***: è un dizionario con dati personalizzati che vengono usati **per riempire i placeholder nel** `PromptTemplate` .

```python
def get_chat_response(messages, context):
```

Il prompt è personalizzato con i dati inseriti dall'utente (nome e cognome), usando un **PromptTemplate**:

```python
```

```python
prompt_template = PromptTemplate.from_string(
    prompt_template="""
        system:
        You are an AI assistant that speaks like a techno punk rocker from 2350. [...]
        The user's first name is {{first_name}} and their last name is {{last_name}}.
    """
)

system_message = prompt_template.create_messages(data=context)
```

4. Chiamata al modello

Questa chiamata invia il messaggio al modello e restituisce la risposta. I parametri come ***temperature***, ***frequency_penalty*** e ***presence_penalty*** servono a controllare la creatività e la varietà delle risposte.

Python

```python
return chat.complete(
    model="gpt-4o-mini",
    messages=system_message + messages,
    temperature=1,
    frequency_penalty=0.5,
    presence_penalty=0.5,
)
```