Via Nairobi 40
00144, Roma (RM)
info@zendata.it

Sommario

Via Nairobi 40
00144, Roma (RM)
info@zendata.it

Chi siamo

ZenData è una startup dinamica e innovativa, specializzata nello sviluppo di soluzioni basate sull'intelligenza artificiale, per aumentare l'efficienza e stimolare la crescita.

Le nostre soluzioni abbracciano una vasta gamma di settori e reparti, spaziando dallo sviluppo di assistenti virtuali per pubbliche amministrazioni e professionisti autonomi, fino all'analisi e all'ottimizzazione dei processi produttivi aziendali.

La nostra strategia è orientata alla progettazione di soluzioni che si integrano in maniera semplice e fluida nei processi operativi dei nostri clienti, garantendo un'adozione senza attriti e un impatto positivo immediato.

Siamo convinti che il nostro approccio all'innovazione, unito alla capacità di personalizzare le soluzioni in base alle specifiche esigenze dei nostri clienti, ci permetta di offrire un valore aggiunto significativo e di contribuire al successo dei nostri partner commerciali.

Via Nairobi 40
00144, Roma (RM)
info@zendata.it

WWW.ZENDATA.IT

1. Introduzione al progetto

1.1 Panoramica

Il presente progetto prevede la realizzazione di due assistenti virtuali, uno rivolto agli studenti e uno dedicato agli amministratori dell'Università Campus Bio Medico di Roma, entrambi basati su un'architettura comune progettata per garantire scalabilità, efficienza e sicurezza nella gestione delle informazioni.

L'**assistente virtuale per gli studenti** è pensato per facilitare l'accesso rapido a documenti e contenuti di supporto, migliorando l'esperienza utente nella ricerca di informazioni. La documentazione sarà inerente a regolamenti didattici e tematiche amministrative relative allo studente.

L'assistente virtuale sarà fruibile su sito internet UCBM.

L'**assistente virtuale per gli amministratori**, invece, offre supporto nella consultazione della documentazione di natura tecnico amministrativa già caricata sulla intranet UCBM come linee guida di processo, regolamenti e istruzioni operative.

L'assistente virtuale sarà fruibile sia sulla pagina internet UCBM che sulla intranet.

A supporto di queste funzionalità, il progetto prevede anche lo sviluppo di una **dashboard**, che consentirà agli amministratori di monitorare l'utilizzo dei due assistenti virtuali, analizzare metriche di performance e gestire i contenuti in modo semplice e intuitivo.

L'intera soluzione sarà implementata su **tecnologie cloud di Azure**, assicurando **scalabilità e sicurezza.**

1.2 Obiettivi Dashboard

1. **Gestione dei Contenuti**
    - Creazione di un'interfaccia intuitiva per il **caricamento e l'eliminazione dei documenti** della Knowledge Base (**KB**) dei due chatbot.

2. **Analisi e Monitoraggio**
    - **Visualizzazione delle metriche di utilizzo** dei due sistemi, fornendo insight sulle interazioni degli utenti e sull'efficacia del sistema.
    - **Monitoraggio delle performance del sistema RAG**, con l'obiettivo di ottimizzare il recupero delle informazioni e migliorare la qualità delle risposte generate.
3. **Sicurezza e Controllo Accessi**
    - **Integrazione di autenticazione sicura tramite Azure AD** per garantire che solo utenti autorizzati possano accedere alle funzionalità amministrative.

1.3 Obiettivi Assistenti Virtuali

1. Realizzazione di un RAG System avanzato

- Utilizzo di un **database vettoriale** per la ricerca semantica, migliorando la qualità delle risposte fornite dal chatbot.
- Mantenimento di un **database NoSQL scalabile** basato su **Azure Cosmos DB for MongoDB**, garantendo elevate prestazioni e affidabilità.
- Esecuzione della logica applicativa in un **servizio gestito** tramite **Azure App Service**, assicurando scalabilità e semplificazione della gestione infrastrutturale.

2. Sicurezza by Design

- **Gestione centralizzata** delle identità e degli accessi tramite **Azure AD**, assicurando un controllo rigoroso sugli utenti abilitati.
- **Limitazione del traffico** di rete attraverso **Virtual Network e private endpoints**, proteggendo le comunicazioni interne al sistema.
- **Protezione di segreti e chiavi** sensibili con **Azure Key Vault**, riducendo i rischi legati alla gestione delle credenziali.
- **Monitoraggio continuo** delle performance e della sicurezza tramite **Azure Monitor**, garantendo proattività nella gestione operativa.

2. Architettura degli assistenti virtuali

La seguente sezione descrive l'architettura prevista per la messa in produzione di entrambi gli assistenti virtuali. L'infrastruttura sarà implementata su Microsoft Azure, garantendo scalabilità, sicurezza e flessibilità. L'adozione di una piattaforma cloud permetterà di ottimizzare le prestazioni, semplificare la gestione e garantire un'evoluzione continua della soluzione senza impatti sull'operatività.

2.1 Panoramica dell'Architettura

L’architettura dei due assistenti virtuali si compone sostanzialmente di 4 layer funzionali:

1. **Layer di Ingestion**: rappresenta le soluzioni software e le risorse cloud che permetteranno l’ingestion semi-automatica dei documenti come knowledge base dei due sistemi.
2. **Layer di Storage**: si compone dei database a supporto del funzionamento di entrambi gli assistenti virtuali. Per questo livello verranno utilizzati un database vettoriale per la ricerca semantica tra la documentazione (Retrieval del RAG system) e un database NoSQL (Azure Cosmos DB for MongoDB) che semplifica la gestione operazionale delle conversazioni e al contempo la gestione di collection dedicate alle analytics della dashboard.
3. **Layer di AI**: Questo livello si compone dei Foundational Models forniti da Azure sia per la ricerca di documentazione attraverso modelli di Embedding che per la formulazione delle risposte tramite l’uso di Large Language Models.
4. **Layer Applicativo**: Si compone delle risorse cloud per la realizzazione delle App backend e frontend.

2.2 Autenticazione e Identità

Per garantire la protezione dei dati e il controllo degli accessi, il sistema adotterà **Azure Active Directory (Azure AD)** per la gestione centralizzata delle identità e l'autenticazione sicura alle risorse Azure.

L'uso delle **Managed Identities** permette ai servizi Azure di autenticarsi in modo sicuro senza necessità di gestire manualmente credenziali statiche. In particolare, l'**App Service** utilizzerà una **System-Assigned Managed Identity** per ottenere token da Azure AD e accedere in modo sicuro a **Key Vault, Cosmos DB e al database vettoriale**, sfruttando il **Role-Based Access Control (RBAC)** e politiche di accesso dedicate.

Per rafforzare la sicurezza della rete e ridurre la superficie di attacco, il sistema sarà isolato all'interno di una **Virtual Network (VNet) dedicata**. Questa configurazione permette di limitare l'accesso ai soli sistemi autorizzati e instradare il traffico attraverso connessioni private, evitando esposizioni pubbliche. Sarà prevista l'integrazione con la stessa Vnet per gli assistenti virtuali e la dashboard.

L'architettura prevede:

- **VNet dedicata**
- **Integrazione dell'App Service con la VNet**, per connettersi internamente a **Cosmos DB, il database vettoriale e Key Vault** tramite IP privati.
- **Private endpoint per Cosmos DB e il database vettoriale**, assegnati a una subnet dedicata per garantire che il traffico resti confinato all'interno della VNet, senza esposizione pubblica.
- **Private endpoint per Key Vault**, che limita l'accesso solo ai servizi interni alla rete privata.

Questa strategia garantisce un'elevata sicurezza dell'infrastruttura, proteggendo i dati e minimizzando i rischi legati agli accessi non autorizzati.

2.3 Layer di Ingestion

Il layer di ingestion si compone di tutte quelle risorse che hanno l'obiettivo di permettere il popolamento della Knowledge base dei due assistenti virtuali.

Questo layer prevede l'utilizzo di Azure Functions per l'estrazione del contenuto della documentazione da integrare all'interno degli assistenti virtuali. In particolare, queste funzioni si occuperanno di:

- Manipolare il testo non strutturato della documentazione:
    - Preprocessing per l'estrazione del testo e dei metadati utili all'intelligenza artificiale per l'ottimizzazione delle ricerche.
    - Suddivisione in chunk per ottimizzare le prestazioni della RAG.
    - Calcolo dei vettori di embedding per l'indicizzazione dei documenti all'interno del database vettoriale.
- Caricare i documenti pre processati all'interno del database vettoriale.
- Caricare i documenti su un Blob Storage.

2.4 Layer di storage

Il layer di storage dei due sistemi consiste sostanzialmente di due database:

- **Azure Cosmos DB for Mongodb** viene utilizzato come database scalabile per la gestione dei dati transazionali utilizzati dai due assistenti per il recupero dei messaggi di una conversazione. Contemporaneamente, questa tipologia di Database, permette l'esecuzione di query volte a mostrare le analitiche di utilizzo all'interno di una dashboard di amministrazione.
- Azure AI Search è il database vettoriale adottato per la gestione della documentazione non strutturata alla base del processo di **Information Retrieval (IR)** del RAG system.

Entrambi i database permettono l'accesso in sicurezza ai dati attraverso connessioni su rete privata utilizzando integrazione tramite Virtual Network e l'utilizzo di Private Endpoint.

2.5 Layer di AI

Questo layer si compone dei foundational models che verranno utilizzati sia per la ricerca di documentazione all'interno del database vettoriale (attraverso modelli di Embedding), che per la formulazione delle risposte degli assistenti virtuali in linguaggio naturale (attraverso l'uso di Large Language Models).

Lo stack tecnologico di questo layer prevede:

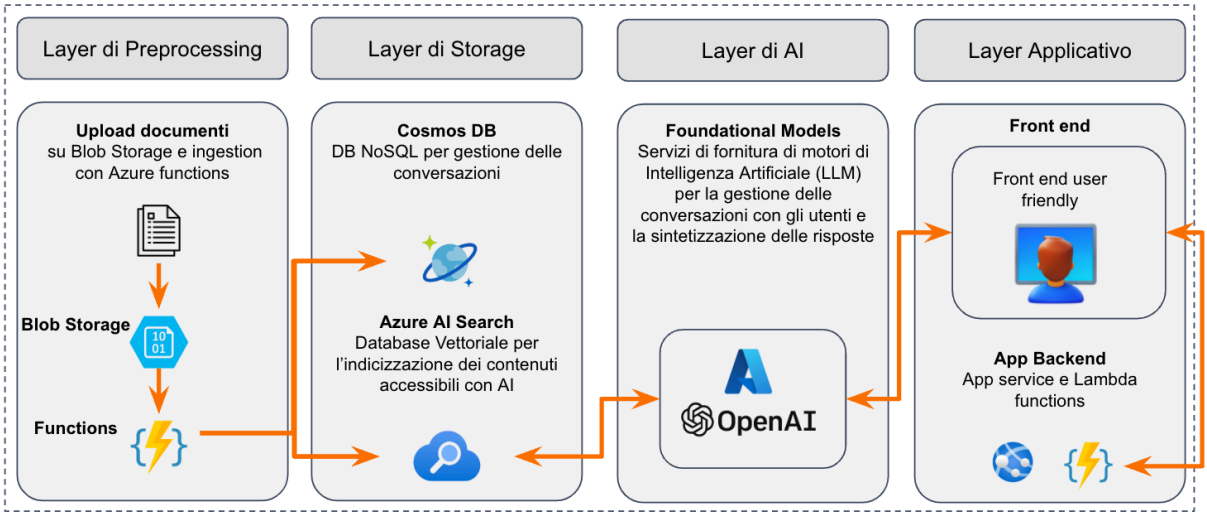
- **Motore di Embedding (IR)** che utilizza la domanda dell'utente per cercare le sezioni della documentazione più pertinenti all'interno del Database vettoriale.
- **Large Language Model** che, utilizzando i contenuti estratti dal motore IR, contestualizza la risposta fornita.

2.6 Layer Applicativo

Entrambi gli assistenti virtuali saranno distribuiti attraverso un unico backend centralizzato. Per l'hosting e la gestione della logica applicativa, verrà utilizzato **Azure App Service**, che eseguirà l'intero sistema all'interno di **container Docker**. Questa soluzione garantisce scalabilità, isolamento e una gestione semplificata delle risorse cloud. Per il deployment dell'interfaccia front-end dei due assistenti virtuali è stata individuata la risorsa **Azure Static Web App**. Questa soluzione offre un'elevata integrazione con gli altri servizi Azure e facilita l'inserimento nei siti esistenti. Tuttavia, nel corso del progetto verranno valutati ulteriori dettagli per garantire la massima compatibilità con i sistemi di destinazione. L'accesso del backend applicativo agli altri layer dell'architettura avverrà tramite **Virtual Networks (VNet) e Private Endpoints**, garantendo connessioni sicure su rete privata. Le chiavi di accesso saranno archiviate in **Azure Key Vault** e gestite attraverso **Managed Identity**, assicurando un'integrazione sicura e senza l'uso di credenziali esplicite.

Per il monitoring dello stato dell'applicazione verrà adottato Azure Application Insights.

Di seguito si riporta l'architettura ad alto livello precedentemente descritta:



Come requisito modulare sarà possibile integrare interrogazione e risposta in modalità vocale. In particolare utilizzando:

- **Speech-to-text:** utilizzo di Azure Speech Service o altri servizi per la trascrizione vocale in tempo reale
- **Text-to-speech:** utilizzo di Azure Speech Service o altri servizi per la sintesi vocale

3. Architettura della Dashboard

La seguente sezione descrive l'architettura prevista per la messa in produzione della Dashboard. L'infrastruttura sarà implementata su Microsoft Azure, garantendo perfetta integrazione con gli assistenti virtuali descritti precedentemente.

L'architettura della Dashboard è stata progettata per offrire un'interfaccia amministrativa completa che consenta la gestione efficiente dei contenuti e il monitoraggio delle performance degli assistenti virtuali.

3.1 Panoramica dell'Architettura

L'architettura della Dashboard si compone di quattro componenti principali:

1. Frontend Amministrativo:

- Offre funzionalità di gestione dei documenti e visualizzazione delle analytics
- Design responsive ottimizzato per diversi dispositivi

2. Backend API:

- Gestisce le richieste dal frontend e coordina le operazioni sui dati

3. Function App per l'elaborazione documenti:

- Componente serverless che gestisce operazioni asincrone e di lunga durata
- Si occupa dell'estrazione del testo dai documenti e della creazione di chunk ottimizzati

4. Storage condiviso:

- Utilizzo delle stesse risorse di storage degli assistenti virtuali
- Include Cosmos DB, database vettoriale e Blob Storage
- Garantisce coerenza dei dati e riduce la duplicazione

Le attività principali supportate da questa architettura includono:

- Gestione completa del ciclo di vita dei documenti (caricamento, elaborazione, indicizzazione, eliminazione)

- Monitoraggio delle performance degli assistenti virtuali attraverso metriche e analytics

- Amministrazione sicura con autenticazione Azure AD

3.2 Autenticazione e Controllo Accessi

Per garantire la protezione dei dati e il controllo degli accessi, il sistema adotterà le stesse politiche di sicurezza descritte nell'architettura degli assistenti virtuali. Azure Active Directory (Azure AD) viene utilizzato per autenticare gli amministratori e i servizi che accedono alla dashboard, garantendo un accesso centralizzato e sicuro. L'uso delle Managed Identities consente all'applicazione di autenticarsi automaticamente con i servizi Azure senza necessità di gestire manualmente credenziali.

3.3 Infrastruttura di Rete

Verrà utilizzata una subnet dedicata per i componenti della Dashboard all'interno della VNet esistente. Verranno utilizzati gli stessi private endpoint configurati per gli assistenti virtuali, assicurando che il traffico resti confinato all'interno della rete privata.

3.4 Raccolta e Analisi delle Metriche

La dashboard prevede la raccolta di metriche pensate per monitorare le performance, individuare trend e ottimizzare il servizio nel tempo. Le metriche definitive potranno essere accordate in dettaglio in una seconda fase, ma includeranno il tracciamento delle query al chatbot (volume, tipologia, distribuzione temporale) e l'analisi delle tematiche più frequenti. I dati raccolti verranno poi resi disponibili attraverso dashboard interattive.

4. Struttura dell’offerta

Di seguito si riporta il dettaglio dei costi complessivi di progetto che comprende la realizzazione e la manutenzione di entrambi gli assistenti virtuali e della dashboard amministrativa:

Costi di Progetto	
Servizi professionali Zendata	
implementazione (CapEx)	10.000 €
manutenzione (OpEx) (esclusa infrastuttura)	1.000 €/mese

Tutti i prezzi sono da considerarsi IVA esclusa.

Di seguito si riportano i costi in dettaglio dell’infrastruttura:

Nome Servizio	Descrizione	€/mese
Azure Cosmos DB	Azure Cosmos DB for MongoDB (RU), Autoscale provisioned throughput	40
Azure AI Search	Basic, 1 Unit(s), 730 hours	70
Virtual Network	Sweden Central region	10
Azure Blob Storage	Block Blob Storage, Standard, General Purpose V2, Flat Namespace, Hot, LRS, 50 GB,	2

Azure functions	Sweden Central region, Premium	150
App Service	Premium V3, 2 core, 8 GB RAM, 250 GB Storage. Backend degli Assistenti virtuali. 1 year reserved	90
App Service	Basic B2, 3.5 GB RAM 10GB Storage. Backend Dashboard amministrativa. On demand.	20
Key Vault	Sweden Central region	1
Totale		383

Di seguito sono riportate le stime dei costi per i servizi di OpenAI relativi al chatbot progettato per supportare studenti, basato sull'utilizzo di Azure OpenAI Services. Queste stime considerano due scenari: un volume minimo di messaggi (lower bound) e un volume massimo (upper bound). Per i servizi di Speech-to-Text e Text-to-Speech, si è volutamente assunto che vengano utilizzati su tutti i messaggi inviati e ricevuti, al fine di fornire una valutazione completa dei costi potenziali. I dettagli sono i seguenti:

1) Lower Bound: totale di messaggi/mese di 56k, l'equivalente di 333 messaggi all'ora (considerando 21 giorni lavorativi ed 8 ore di lavoro):

Nome Servizio	Descrizione	€/mese (lower bound)
Azure OpenAI Services	3500 studenti, 2 conversazioni mese per studente, 8 messaggi ciascuna (56k messaggi/mese). Token in input: 45k per messaggio (comprensivi di contesto recuperato dalla KB e della domanda dell'utente); token in output: 200 per messaggio.	280

Azure OpenAI Services	Embedding Models, Text-Embedding-3-Small, 12k pagine di documentazione	2
Speech-to-Text	Speech-to-Text per 28k messaggi/mese (tutti quelli inviati)	28
Text-to-Speech	Text-to-Speech per 28k messaggi/mese (tutti quelli ricevuti)	336

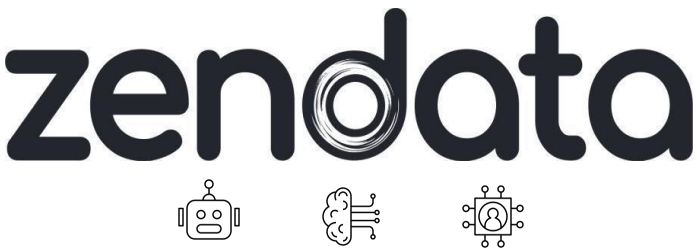
2) Upper Bound: totale di messaggi/mese di 210k, l'equivalente di 1250 messaggi all'ora (considerando 21 giorni lavorativi):

Nome Servizio	Descrizione	€/mese (upper bound)
Azure OpenAI Services	6 conversazioni mese per studente, 10 messaggi ciascuna (210k messaggi/mese). Token in input: 45k per messaggio (comprensivi di contesto recuperato dalla KB e della domanda dell'utente); token in output: 200 per messaggio.	1050
Azure OpenAI Services	Embedding Models, Text-Embedding-3-Small, 12k pagine di documentazione	2
Speech-to-Text	Speech-to-Text per 105k messaggi/mese (tutti quelli inviati)	105
Text-to-Speech	Text-to-Speech per 105k messaggi/mese (tutti quelli ricevuti)	1260

Tutte le stime di costo sono basate su ipotesi iniziali, il costo scalerà in proporzione al numero di messaggi effettivi in produzione.

La licenza è considerata per un intervallo temporale pari ad 1 anno.

I costi per l'infrastruttura rimangono i medesimi sia nello scenario di SaaS che nello scenario self- hosted UCBM utilizzando come provider Azure.



info@zendata.it

www.zendata.it