zendata

Offerta Tecnica - 27/05/2024

Fornitore: Zendata S.r.l.

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

Sommario

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

Panoramica

Il presente progetto propone la creazione di un assistente virtuale basato su intelligenza artificiale generativa. In una fase iniziale di sperimentazione, l'obiettivo è quello di fornire supporto ai dipendenti di Q8 nella ricerca di informazioni presenti all'interno dei rispettivi spazi Confluence. Questa soluzione innovativa mira a offrire un supporto rapido ed efficiente, in quanto l'assistente verrà integrato direttamente nel contesto aziendale tramite Microsoft Teams.

Obiettivi

1. Integrazione del database documentale di Confluence: collegare l'assistente virtuale alla documentazione interna di Confluence per fornire informazioni precise ed aggiornate.
2. Migliorare l'Accessibilità: creare un assistente virtuale che agevoli l'accesso alla documentazione presente su confluence, centralizzando il punto di accesso alle informazioni degli spazi. L'assistente non si limiterà a dare i riferimenti agli spazi rilevanti rispetto ad una richiesta ma dovrà sintetizzare una risposta in base alla documentazione fornitagli.
3. Estendibilità e Supporto Interattivo: avere la possibilità di interagire con la documentazione attraverso un'interfaccia di chat per consentire agli utenti di chiedere supporto ad un'assistente virtuale direttamente dentro Microsoft Teams. L'integrazione dell'assistente all'interno di Teams semplificherà anche l'estensione del perimetro di gestione del bot.

Perimetro di progetto

Durante questa fase iniziale, ci si concentrerà sulla definizione del perimetro gestito dall'assistente virtuale nei primi tre mesi di sviluppo. Per ragioni di complessità tecnica e vastità delle informazioni su Confluence, si procederà come segue:

1. Si limiterà la gestione ai **dati testuali e tabellari**.

2. Non si gestiranno direttamente immagini presenti negli spazi né documenti **allegati alle pagine** di Confluence, ma **si fornirà un link** per accedervi.
3. In collaborazione con i referenti del progetto, verranno selezionate **tra 70 e 100 pagine** Confluence per dimostrare l'utilità dell'assistente, considerando:
    - La frequenza di accesso alle pagine.
    - L'utilità nel lavoro quotidiano.
    - Il numero di persone potenzialmente interessate a ciascuna pagina.
4. L'architettura e lo sviluppo dell'applicazione saranno **progettati per consentire l'estensione del perimetro** gestito dall'assistente virtuale per future espansioni degli spazi o per altre esigenze di utilizzo.

Fasi di progetto

Il piano di progetto si struttura in **tre fasi principali** che delineano tutte le attività necessarie alla finalizzazione di un assistente integrato e aggiornato autonomamente rispetto alla documentazione presente su Confluence. In questa prima fase progettuale si procederà valutando un perimetro sufficientemente esteso da dimostrare i benefici che l'integrazione di un'assistente su Teams possa portare. Seguendo una logica di progressivo sviluppo e di costante affinamento, si procederà in modo da rendere l'applicazione facilmente estendibile non solo a nuovi spazi e pagine Confluence, ma anche ad ulteriori documentazioni interne di Q8.

Fase 1: Analisi tecnica e funzionale

Durante la fase iniziale di analisi tecnico-funzionale, si dovrà, in prima istanza, avere accesso agli spazi Confluence interni sia con accesso User che con accesso programmatico tramite API (A1). In questo modo sarà possibile condurre le analisi degli spazi in collaborazione con Q8 ed anche in autonomia per verificare le modalità di resa dei dati presenti sugli spazi quando acceduti tramite API Confluence (A2). Sarà successivamente obiettivo di questa fase finalizzare l'accesso agli ambienti di sviluppo cloud Microsoft ed Amazon al fine di configurare le risorse richieste per l'architettura del progetto (A3). L'obiettivo sarà analizzare un numero significativo di spazi e pagine Confluence per selezionare un insieme ristretto su cui concentrare gli sforzi. Durante questa fase, si validerà l'architettura tecnica e logica delle risorse necessarie per il progetto (A4).

Ci si riserva la possibilità di valutare modifiche all'architettura in corso di progetto in base a necessità emergenti o a potenziali miglioramenti in termini di sicurezza e modelli di AI presenti sul mercato. La ricerca e il confronto tra diverse soluzioni di AI saranno cruciali, considerando che rappresentano un fattore discriminante per il successo del progetto ed essendo costantemente in evoluzione.

Inoltre, durante questa fase, verranno create le risorse cloud necessarie per lo sviluppo del progetto (A5).

Verranno definite in questa fase anche la cadenza di aggiornamento delle basi dati rispetto a quanto scritto giornalmente sugli spazi confluence identificati.

ID ATTIVITA'	BREVE DESCRIZIONE
A1	Accesso a Confluence (API/User)
A2	Analisi spazi Confluence
A3	Accesso ad infrastruttura cloud e validazione architettura
A4	Selezione delle pagine Confluence
A5	Creazione risorse cloud

Fase 2: Implementazione delle pipeline di ingestion

Una volta terminata la fase di analisi e selezione degli spazi e pagine confluence da dover gestire si potrà cominciare la fase di sviluppo delle pipeline di ingestion dei dati.

In questa fase verranno implementate le logiche di:

- Caricamento iniziale delle pagine sul database NoSQL (A6)
- Caricamento iniziale delle pagine sul database Vettoriale (A7)
- Aggiornamento delle pagine sui database (A8)

Per ottenere in tempi stretti dei primi risultati visibili con l'integrazione del chatbot su teams provvederemo ad un primo caricamento manuale di un set ristretto di pagine Confluence.

Le API di confluence permettono di estrarre il contenuto di una pagina in 7 formati differenti motivo per il quale ogni formato può mancare di alcuni contenuti o fornirne di più. Di conseguenza nella fase di implementazione degli script di caricamento dei dati sul database di conoscenza del modello di Generative Ai (GenAi) dovrà essere fatta un'analisi di quale sia il formato più adatto e contestualmente realizzarne i codici di pulizia e formattazione del contenuto ottenuto dalle API (A9).

Secondo un approccio di tipo agile verranno effettuati dei test in parallelo allo sviluppo del chatbot secondo cui sceglieremo in funzione dei risultati con i diversi formati dati iniziali quale formato di partenza è il più efficace per popolare il database documentale di Questi8. Oltre che alla scelta del formato dati più adatto verrà valutato anche come la modalità di formattazione dei contenuti tramite codice impatta sull'efficacia delle risposte di Questi8 (e.g. una tabella può essere formattata in modi diversi come html/markdown , ecc..) (A10). Seguendo un approccio agile procederemo effettuando piccoli rilasci per ogni tipologia di formattazione seguiti da test per verificare l'efficacia delle pipeline.

Una volta stabilizzati i flussi di ingestion potrà essere messa in piedi l'architettura che in modo schedulato nell'arco di tempo definito nella Fase 1 si occupa dell'inserimento e aggiornamento delle basi NoSQL e vettoriale (A11).

ID ATTIVITA'	BREVE DESCRIZIONE
A6	Caricamento iniziale NoSql
A7	Caricamento iniziale Vector DB
A8	Aggiornamento automatico dei Database
A9	Analisi Formati Confluence API
A10	Rifinitura Script di caricamento delle pagine
A11	Deploy risorse dell'architettura

Fase 3: Realizzazione e test del Chatbot

L' ultima macro fase di progetto comprende la realizzazione ed ingegnerizzazione del Chatbot alle pagine Confluence definite nella Fase 1. L'integrazione della totalità delle pagine verrà seguita in maniera incrementale, ci si aspetta quindi di rilasciare una versione aggiornata della base documentale gestita da Questi8 circa ogni due settimane. Al termine di queste in base all'esito dei test verranno reiterati i processi di sviluppo realizzati o integrate nuove pagine del set prescelto.

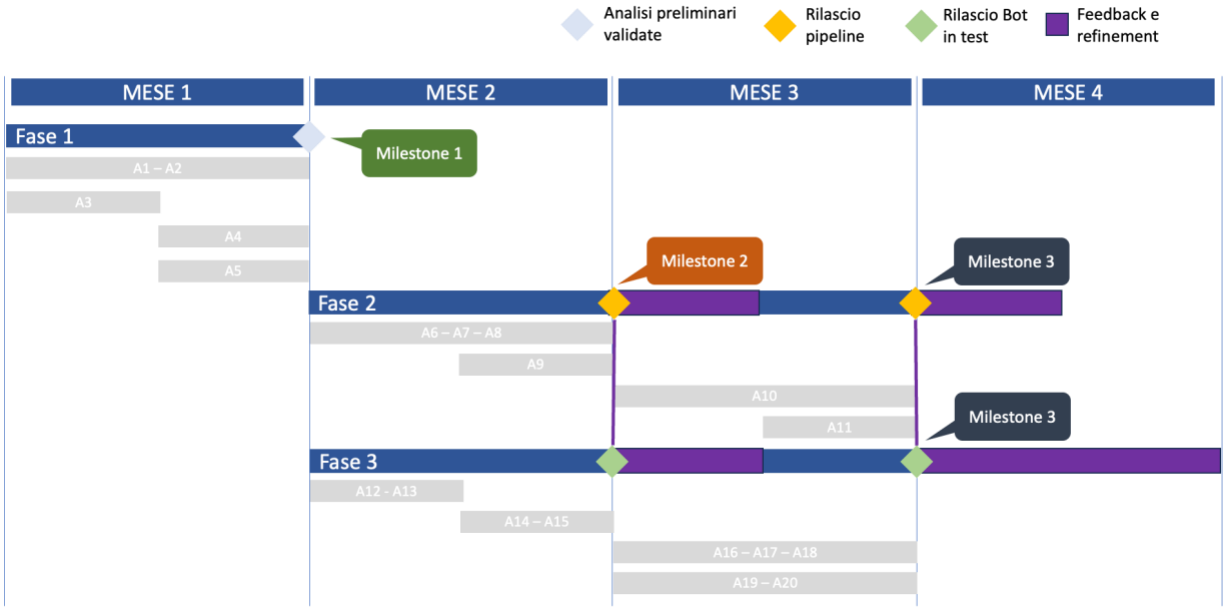
Questa frase progettuale si compone delle seguenti macro attività:

1. Deploy del servizio Microsoft Copilot Studio per avere il chatbot realizzato integrabile dentro il Teams aziendale (A12).
2. Realizzazione dei flussi su Power Automate per permettere il trigger al Backend del chatbot a partire da un messaggio inviato su teams (A13).
3. Gestione degli accessi al backend richiamato da power Automate (A14).
4. Realizzazione del software ed API per gestire le chiamate provenienti dai flussi di Power Automate:
    a. Sviluppo iniziale del chatbot (A15).
    b. Gestione dello storico delle conversazioni realizzata tramite database NoSQL, per archiviare e recuperare dinamicamente i dati delle interazioni passate (A16).
    c. Ingegnerizzazione del software per gestire l'incremento delle pagine confluence attraverso l'ottimizzazione degli algoritmi di ricerca a DB e delle logiche di ricerca del chatbot (A17).
    d. Prompt Engineering del modello fondazionale (A18).
    e. Adozione e test di differenti modelli di embeddings e logiche di chunking (A19).
    f. Adozione e test di differenti modelli Large Language Models (LLM) per la generazione di risposte (A20).

ID ATTIVITA’	BREVE DESCRIZIONE
A12	Deploy Chatbot
A13	Realizzazione Flussi
A14	Gestione Accessi
A15	Prima Release Chatbot
A16	Gestione Storico Conversazioni
A17	Ottimizzazione del Software
A18	Prompt Engineering
A19	Test Embeddings e Chunking
A20	Test LLM

Gantt di progetto

Di seguito si riporta il dettaglio delle attività definite nella sezione precedente declinate in un piano dei tempi:



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

In corrispondenza di ciascuna milestone verranno prodotti deliverable di carattere documentale e/o software. In particolare al termine del primo mese verrà prodotto un documento contenente:

1. La specifica di quali pagine Confluence gestirà l'assistente virtuale.
2. Il criterio di selezione di queste.
3. I potenziali team interessati all'uso di Questi8.

Al termine del secondo mese saranno rilasciati:

- La prima versione del chatbot .
- I flussi di data ingestion nei DB automatizzati.

A valle dei primi tre mesi di lavoro si avrà disponibile una versione funzionante del chatbot in grado di gestire tutte le pagine identificate a monte.

2. Specifiche tecniche di sviluppo

L’architettura che verrà messa in piedi per la realizzazione di Questi8 si organizza in 4 livelli principali:

Layer di Ingestion Computazionale

Questo layer si compone di **Lambda Functions schedulate** nel tempo per l’estrazione del contenuto delle pagine presenti su confluence. Queste funzioni si occuperanno di:

1. Manipolazione del testo non strutturato dei documenti:
    a. **Preprocessing automatizzato** per l’estrazione di metadati utili all’intelligenza artificiale per ottimizzare le ricerche all’interno degli spazi
    b. Suddivisione in chunk per ottimizzare le prestazioni della **Retrieval Augmented Generation (RAG)**.
    c. **Calcolo dei vettori di embedding** per indicizzare i documenti in un database ottimizzato per la ricerca basata su vettori.
2. Caricamento dei documenti pre processati all'interno del **Database Vettoriale**.
3. Caricamento dei metadati principali della pagina confluence su un **Database NoSql**

Layer di Storage

Questo layer si compone dello stack tecnologico su cui i dati estratti da confluence verranno memorizzati. Questo layer si compone di due database principali:

1. **Amazon Dynamodb come Database Documentale** per l'archiviazione dei metadati delle pagine confluence e per la memorizzazione delle conversazioni con gli utenti.
2. **Pinecone DB come Database Vettoriale** per lo storage indicizzato mediante modelli di embedding dei documenti presenti nel Cosmos DB.

Layer di AI

È il layer che si compone dei modelli foundational serviti dai principali provider di servizi cloud per ottimizzare la ricerca sul Db vettoriale e per sintetizzare la risposta in linguaggio naturale.

Questo stack si compone di:

1. **Motore di Embedding (IR)** che utilizza la domanda dell'utente per cercare le sezioni dei documenti più pertinenti nel Database Vettoriale.
2. **Large Language Model** che contestualizza la risposta fornita basandosi sui contenuti estratti dal motore di IR.

In una prima fase verranno privilegiati i modelli fondazionali proposti da AWS e valuteremo l'utilizzo di altri provider in caso di scarsi risultati a valle dei test.

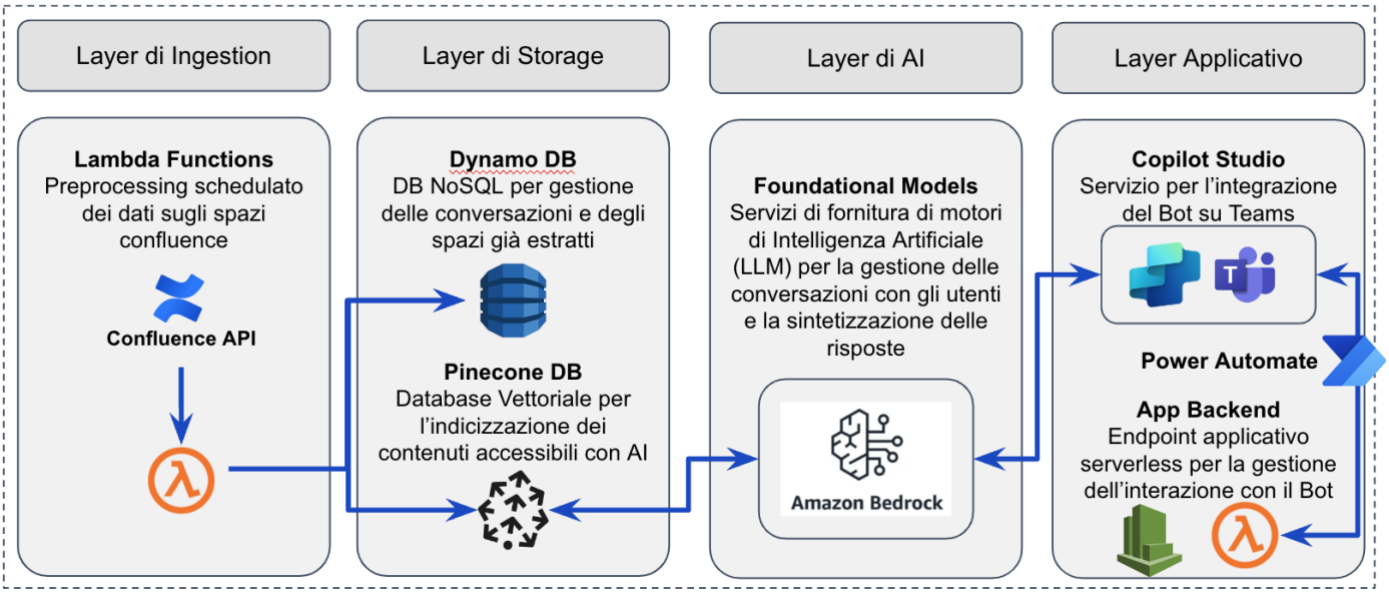
Layer Applicativo

L'applicativo backend per la gestione delle chat sarà interamente realizzato con un modello serverless per ottimizzare il livello di scalabilità unito ad un costo contenuto e dipendente unicamente dall'utilizzo.

1. Le API di backend verranno deployate attraverso **Lambda Functions**
2. Il Front end sarà **Microsoft teams**
3. Su Teams verrà integrato il chatbot realizzato su **Microsoft Copilot Studio** che attraverso **Power Automate** potrà gestire flussi per chiamare le API al momento della ricezione di un messaggio da parte del bot.

Le metriche di performance dell'applicativo e la gestione dei log verranno gestiti attraverso **AWS Cloudwatch**.

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



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

3. Gestione dei rischi

I tempi di sviluppo dell'applicazione sono previsti complessivamente in un totale di 3 mesi. Tuttavia, per gestire eventuali rischi che potrebbero emergere durante il progetto, ci rendiamo disponibili per un totale di 4 mesi. Il mese di buffer di cui si dispone sarà dedicato alla gestione delle problematiche che potrebbero sorgere a seguito dei test e a possibili rallentamenti legati alla fase iniziale di analisi di progetto in quanto fortemente dipendente dalla struttura delle pagine Confluence.

Di seguito si riporta un elenco di possibili rischi progettuali per cui si potrebbero avere dei ritardi:

1. Ritardi dovuti alla fornitura di utenze per integrarsi alle API di confluence in quanto attività sperimentale del progetto
2. Ritardi causati da possibili inefficienze nei modelli di AWS per cui si dovrà valutare l'adozione di altri provider di modelli.
3. Rischi di rallentamenti a causa dell'integrazione del chatbot all'interno del Teams aziendale per cui in una prima istanza si potrebbe ipotizzare di implementare un Front End su cui effettuare i test.
4. Rischi legati alla sperimentazione di Confluence API come sorgente da cui vengono estratti i dati:
    a. La variabilità di come le pagine sono state scritte dagli utenti potrebbero richiedere ulteriori sforzi di pulizia dei dati prima dell'indicizzazione a DB
    b. Il formato dati che ritorna confluence è molto variabile e ogni formato andrà analizzato per comprendere bene le potenzialità di ognuno
    c. Gli script di pulizia dei dati potrebbe non essere replicabile tra i diversi formati dati di confluence

4. Estensione del progetto

L'obiettivo della prima fase è quello di sviluppare una versione di chatbot che sia immediatamente funzionale, ma progettato per essere esteso ad altri spazi Confluence, al fine di servire l'intera base dipendenti di Q8.

L'estensione del progetto prevede ulteriori sviluppi software per includere tutti gli spazi Confluence di interesse. Questo processo includerà un'analisi dettagliata per definire gli algoritmi e le modalità operative necessarie a compartimentare efficacemente le pagine, garantendo una personalizzazione accurata e mirata alle esigenze individuali.

ID ATTIVITA’	BREVE DESCRIZIONE
B1	Analisi Operativa di compartimentazione delle pagine Confluence
B2	Implementazione di logiche di compartimentazione delle pagine Confluence
B3	Ottimizzazione della metadatazione dei DB
B4	Aggiornamento dell’algoritmo di Retrieval Augmented Generation
B5	Test Funzionali

5. Struttura dell’offerta

Per quanto concerne la realizzazione della prima versione di Questi8, Zendata S.r.l. ha identificato tre risorse principali necessarie allo di sviluppo di progetto.

Le figure professionali coinvolte mirano ad ottimizzare la sinergia tra competenze legate all’area dell’intelligenza artificiale e dello sviluppo di applicativi Backend.

In particolare, verranno disposte due figure aventi conoscenza ed esperienza nella realizzazione di infrastrutture software basate sull’utilizzo di Large Language Models.

Di seguito si riporta il dettaglio delle risorse con il prezzo giornaliero e l’effort che dedicheranno al progetto:

Seniority	Ruolo	Prezzo al giorno	FTE
Expert	AI Software Engineer	270	0.5
Junior	AI Software Engineer	230	1
Junior	Backend Engineer	230	1

Con riferimento all’estensione di progetto mirata alla gestione di tutta la documentazione Confluence come Base documentale si stima lo stesso effort delle risorse riportate sopra.

In aggiunta, verranno offerte un totale di 50 giornate da utilizzare entro un anno dall'inizio del progetto per la manutenzione dell'infrastruttura e la realizzazione di miglioramenti o estensioni.

Il cliente potrà riservare l'uso di queste giornate in base alle proprie necessità una volta completata la prima versione del chatbot.

Si riporta di seguito il dettaglio dei costi complessivi di progetto inclusa la parte proposta e opzionale:

Ambito	Ruolo	Effort (gg)	Prezzo totale al giorno	Costo Totale
Prima Versione Questi8	Expert Ai Software Engineer	31	270	8.370€
	Junior Ai Software Engineer	63	230	14.490€
	Junior Backend Engineer	63	230	14.490€
Totale				~~**37.350€**~~
Totale Scontato				**31.500€**
Parte Opzionale di Progetto				
Manutenzione		50	230	Ad utilizzo
Estensione Del Chatbot*				34.000€*

*Stima che può essere soggetta a variazioni a seguito delle analisi condotte nella realizzazione della prima versione di Questi8

Tutti i prezzi sono da considerarsi IVA esclusa.

Condizioni di fatturazione:

- 30% al termine dei primi due mesi di lavoro (Milestone 2).
- 30% al rilascio della Milestone 3 dopo tre mesi dall'inizio dei lavori.
- 40% A conclusione del progetto al termine della fase di test del chatbot.

info@zendata.it

www.zendata.it

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

zendata

WWW.ZENDATA.IT