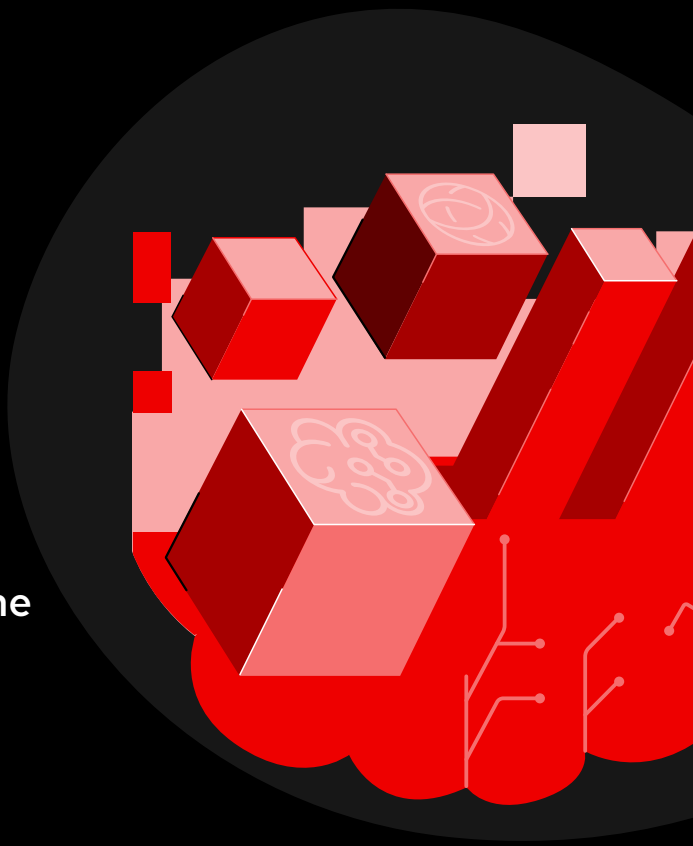
Red Hat

Considerazioni chiave per la

creazione di una base per l'IA generativa

Sommario

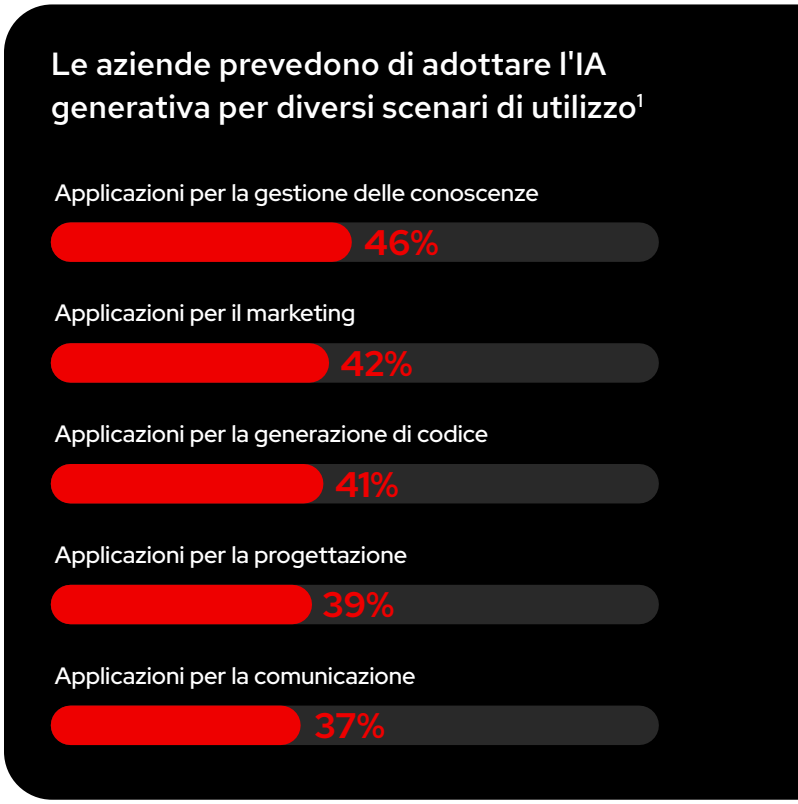


Sfrutta nuove opportunità di innovazione per il tuo business

L'intelligenza artificiale (IA) generativa è uno strumento efficace per tutte le organizzazioni che desiderano sviluppare prodotti innovativi, ottimizzare i processi e ottenere vantaggi competitivi nei mercati in rapida evoluzione. Tecnologia nata dai progressi nell'ambito del deep learning e delle reti neurali, presenta capacità che vanno oltre quelle dell'IA predittiva perché è in grado non solo di elaborare i dati ma anche di generare contenuti nuovi e originali. L'IA generativa promuove l'adozione di approcci innovativi alla risoluzione dei problemi e offre notevoli vantaggi in ogni settore, trasformando l'interazione tra l'uomo e la macchina.

Oggi, le organizzazioni di tutto il mondo realizzano applicazioni innovative servendosi delle tecnologie di IA generativa. Secondo IDC il 39% delle aziende sta investendo su tecnologie di IA generativa e un altro 37% sta valutando i possibili scenari di utilizzo.[1] Di seguito alcune finalità legate a scenari di utilizzo diffusi per l'IA generativa:

- **Elaborare previsioni per gli scenari complessi.** L'IA generativa analizza i dati storici, individua gli schemi ed elabora previsioni accurate che agevolano la pianificazione strategica e la gestione dei rischi.
- **Personalizzare i materiali pubblicitari.** Analizzando i dati l'IA generativa è in grado di individuare le preferenze e i comportamenti dei clienti e creare materiali pubblicitari personalizzati, come email, pubblicità e promozioni, allo scopo di migliorare il coinvolgimento e incrementare il tasso di conversione.
- **Automatizzare e personalizzare l'assistenza clienti.** L'IA generativa, tecnologia su cui si basano i chatbot intelligenti e gli assistenti virtuali, è in grado di rispondere automaticamente a richieste e interazioni dei clienti e permette quindi di ottimizzare e personalizzare il servizio di assistenza.



1 IDC Web Conference Proceeding, "Unlocking Business Success with Generative AI", documento n. US50789223, giugno 2023.

L'IA generativa e le sue criticità

Per quanto a oggi i vantaggi e i rischi legati all'IA generativa non siano ancora del tutto noti, molte organizzazioni scelgono comunque di investire in queste tecnologie. Consapevoli delle possibili criticità dell'IA generativa, le aziende possono definire precise linee guida per un utilizzo etico dell'IA e framework di sviluppo accessibili, garantire la conformità alle normative vigenti e ai requisiti di settore, oltre a rilevare e correggere eventuali problemi.

- **Privacy dei dati.** La possibilità che i dati per l'addestramento dei modelli di IA generativa contengano informazioni sensibili o personali è fonte di preoccupazione e solleva questioni in merito alla tutela della privacy individuale.
- **Proprietà dei dati.** L'utilizzo di modelli proprietari, o modelli addestrati con dati proprietari, introduce il problema della proprietà dei dati e il rischio di incorrere in controversie legali.
- **Pregiudizi e correttezza.** Si è appurato che gli output proposti dagli strumenti di IA generativa ripropongono i pregiudizi umani, compresi stereotipi e discorsi di incitamento all'odio.
- **Utilizzo etico.** Gli hacker possono servirsi dei modelli di IA generativa per creare contenuti fittizi e deepfake da utilizzare in attività illecite, come violazioni della privacy o campagne di disinformazione.
- **Esplicabilità e interpretabilità.** Strumenti di IA generativa poco trasparenti rendono difficile interpretare, comprendere e spiegare i risultati generati dai modelli e quindi assegnare le responsabilità per informazioni errate o inventate.
- **Conseguenze impreviste.** Per la sua natura autonoma l'IA generativa può avere conseguenze impreviste che potrebbero danneggiare le persone e le organizzazioni.
- **Sfide normative.** L'evoluzione delle tecnologie di IA generativa procede più rapidamente rispetto allo sviluppo di impianti normativi; il che rende difficile creare e applicare linee guida per un utilizzo responsabile ed etico dell'IA.
- **Consumi energetici.** L'addestramento dei modelli di IA richiede risorse con elevate capacità di elaborazione e quantità di energia; il che solleva questioni in merito all'impatto e alla sostenibilità ambientale.

Questo ebook propone alcune considerazioni chiave per creare un'infrastruttura di base affidabile a supporto delle iniziative di IA generativa.

Preparati per l'IA generativa

In "Unlocking Business Success with Generative AI", IDC consiglia alcuni interventi propedeutici all'introduzione dell'IA generativa.[2]

- **Crea un ambiente di sperimentazione agile** per gli scenari di utilizzo prioritari che supportano le esigenze aziendali.
- **Definisci delle policy aziendali** per un utilizzo responsabile dell'IA che disincentivino i comportamenti dannosi.
- **Stabilisci l'impatto dell'IA generativa** sulla forza lavoro e adotta una gestione del cambiamento proattiva.
- **Collabora con partner tecnologici** e provider di servizi affidabili per creare l'infrastruttura di IA.
- **Migliora le competenze tecniche** tramite nuove assunzioni, corsi di formazione o l'acquisto di servizi di supporto professionali.

2 IDC Web Conference Proceeding, "Unlocking Business Success with Generative AI", documento n. US50789223, giugno 2023.

Considerazioni per la realizzazione di una base per l'IA generativa

La base tecnologica scelta per supportare le iniziative di IA generativa incide notevolmente sulla riuscita dell'adozione e, in generale, di ogni aspetto coinvolto del progetto. Questo capitolo illustra alcuni aspetti chiave da considerare per la realizzazione di una base ottimale per l'IA generativa.

Considerazione 1: Adotta un set di strumenti affidabili

Sviluppare applicazioni basate su modelli di IA generativa è complesso. Il giusto set di strumenti, con linguaggi, framework e runtime basati su progetti open source e soluzioni commerciali, accelera l'ottimizzazione dei modelli e semplifica lo sviluppo e il deployment delle applicazioni.

Scegli una base che includa i tuoi strumenti preferiti per sviluppare soluzioni di IA innovative in maniera rapida ed efficiente. È importante ricordare che il supporto tramite interfacce interattive di analisi dei dati esplorativa, addestramento e fine tuning semplifica la collaborazione. Inoltre, un set di strumenti preintegrati e capacità self service aiutano a snellire le operazioni IT e assicurano la portabilità e la coerenza tra i diversi ambienti.

Considerazione 2: Accelera l'ottimizzazione dei modelli

L'addestramento dei modelli di IA generativa è un processo lungo e costoso, di conseguenza la maggior parte delle organizzazioni sviluppa le sue soluzioni di IA a partire da modelli fondativi già addestrati su dati per utilizzo generico. I data scientist si occupano di ottimizzare questi modelli fondativi con dati diversificati e specifici per un determinato dominio per far sì che eseguano attività specializzate. Anche questo processo di fine tuning richiede risorse con elevate capacità di elaborazione per cui occorre dotarsi di processori ad alta prestazione e di un'infrastruttura cloud ibrida distribuita.

Scegli piattaforme di IA con funzionalità di gestione e orchestrazione dei carichi di lavoro distribuiti che eseguano i cicli di addestramento, per qualunque tipo di modello, volume di dati o durata, in tutto l'ambiente cloud ibrido. La possibilità di ottimizzare i modelli fondativi in datacenter on premise agevola la conformità ai requisiti normativi e tecnici per i modelli limitati. Inoltre, le funzionalità di addestramento in batch consentono di anticipare l'ottimizzazione dei carichi di lavoro e semplificano la condivisione e la gestione delle risorse.

Alternative per l'ottimizzazione dei modelli

I ricercatori stanno studiando nuovi modi per accelerare e razionalizzare l'ottimizzazione dei modelli fondativi. La **retrieval-augmented generation (RAG)** è un framework di IA per il recupero di fact da sorgenti esterne, come database interni, reti intranet aziendali o Internet, per far sì che i modelli di IA generativa dispongano sempre di informazioni accurate e aggiornate.

Nel caso del **fine tuning dei prompt** i modelli di IA ricevono suggerimenti o prompt front end, come parole extra o numeri generati dall'intelligenza artificiale, che li indirizzano verso la decisione auspicata. In questo modo anche le organizzazioni con dati limitati possono personalizzare un modello fondativo perché svolga attività specifiche.

Considerazione 3: Distribuisci i modelli in maniera efficiente

Riuscire a fornire esperienze dell'utente migliorate grazie alle soluzioni di IA generativa non è un risultato scontato. Per far fronte alle applicazioni con esigenze in continua evoluzione è necessario disporre di un'infrastruttura scalabile e automatizzare la gestione. Un deployment efficiente dei modelli richiede funzionalità per il monitoraggio delle prestazioni e per il ripristino rapido alle versioni precedenti. Inoltre, considerando che le soluzioni di IA elaborano grandi quantità di dati, anche l'applicazione di rigorosi standard di sicurezza in tutti gli ambienti è fondamentale.

Scegli piattaforme che garantiscano il deployment e la scalabilità delle applicazioni e dei modelli di IA generativa in tutto l'ambiente cloud ibrido, dall'infrastruttura on premise al cloud pubblico ai dispositivi edge. La possibilità di distribuire i modelli di IA generativa da ambienti on premise o isolati assicura che i dati proprietari non vengano utilizzati per riaddestrare modelli disponibili al pubblico. Inoltre, il supporto per i rollout canary e per gli strumenti di esplicabilità contribuisce ad aumentare la coerenza e l'affidabilità dei risultati generati dai modelli.

Considerazione 4: Automatizza la gestione del ciclo di vita

Le pipeline di **integrazione e distribuzione continue (CI/CD)** permettono di automatizzare il deployment e la gestione delle soluzioni di IA generativa. Riaddestrando e aggiornando i modelli e le applicazioni attraverso modifiche rapide e incrementali è possibile accelerare lo sviluppo e migliorare le prestazioni dei modelli. Tuttavia, le pipeline di IA sono più complesse dei classici flussi di lavoro CI/CD perché includono alcuni passaggi supplementari, come estrazione dei dati, addestramento, fine tuning, convalida e riaddestramento.

Scegli una base che consenta di creare pipeline di IA, basate su strumenti CI/CD come Tekton e Jenkins, e di integrarle nei flussi di lavoro DevOps esistenti per migliorare lo sviluppo, l'addestramento, il monitoraggio e il riaddestramento dei modelli di IA generativa. Gli strumenti di distribuzione continua **GitOps**, come ArgoCD, consentono di definire e automatizzare il deployment delle soluzioni di IA complesse con un approccio "as Code" e distribuire i modelli e le applicazioni in maniera più coerente.

I container a supporto dell'IA generativa

I **container** e **Kubernetes** forniscono deployment, gestione e scalabilità agili che possono contribuire ad accelerare lo sviluppo cloud native delle soluzioni di IA generativa. Esegui il provisioning on demand degli ambienti in datacenter, cloud pubblici e su dispositivi edge. Automatizza la creazione, il deployment, la scalabilità e la gestione delle istanze dei container su infrastrutture fisiche e virtuali. Integra le soluzioni di IA generativa con componenti e datastore provenienti da un assodato ecosistema di partner open source e provider commerciali. Scopri di più sui **vantaggi dei container per l'IA**.

Considerazione 5: Monitora i modelli in modo coerente

I modelli di IA generativa possono comportare ripercussioni importanti su persone e organizzazioni. Monitorando il comportamento dei modelli, è possibile valutare le decisioni e le giustificazioni, identificare le prestazioni carenti e notificare tempestivamente i comportamenti dannosi. Una gestione efficace dei modelli basata su queste informazioni aiuta a garantire che i modelli generino risposte imparziali, oggettive e corrette negli ambienti di produzione.

Scegli una base per l'IA che offra funzionalità di monitoraggio centralizzate, metriche su pregiudizi e data drift, rilevamento delle anomalie ed esplicabilità per punto. In questo modo potrai esaminare, gestire e correggere i modelli di IA generativa più agevolmente. Il monitoraggio continuo e automatico negli ambienti di produzione migliora la conformità agli standard aziendali per la gestione dei modelli. La presenza di interfacce intuitive e report non tecnici e leggibili in chiaro promuove l'utilizzo responsabile dei modelli e ne agevola la manutenzione.

Concetti chiave per i modelli di IA generativa

- **Pregiudizi:** si tratta della presenza di schemi nel comportamento dei modelli che compromettono l'equità, l'inclusività e l'etica dei risultati generati, ad esempio favoritismi verso determinati gruppi o risposte in linea con gli stereotipi.
- **Data drift:** si verifica quando le proprietà statistiche dei dati per l'addestramento cambiano nel tempo, causando una diminuzione delle prestazioni del modello e la generazione di risposte meno accurate e pertinenti.
- **Rilevamento delle anomalie:** è la capacità di individuare e segnalare se i modelli si comportano in maniera anomala o difforme rispetto agli esempi visti durante l'addestramento.
- **Esplicabilità per punto:** è la capacità di esaminare le cause che hanno portato i modelli a generare determinati risultati. In questo modo si migliorano la visibilità sulle applicazioni e la trasparenza.

Considerazione 6: Sfrutta la competenza di partner assodati

Per offrire esperienze dell'utente innovative, occorre integrare le soluzioni di IA generativa con componenti aggiuntivi. Adottando la giusta combinazione di tecnologie sviluppate da un ecosistema collaborativo di fornitori comprovati, è possibile accelerare lo sviluppo delle applicazioni, gestire con efficacia le problematiche legate alla presenza di pregiudizi e data drift e migliorare le prestazioni dell'intera soluzione.

Affidati a fornitori che dispongono di un vasto ecosistema di partner certificati dove trovare soluzioni complete per lo sviluppo e il deployment delle applicazioni e dei modelli di IA generativa. Un'ampia scelta di componenti, con funzionalità che vanno dall'integrazione e preparazione dei dati all'addestramento e alla distribuzione dei modelli, permette di accelerare e ottimizzare lo sviluppo e il deployment delle soluzioni di IA. Inoltre, scegliendo soluzioni certificate che garantiscono elevati livelli di interoperabilità è possibile ridurre le richieste di supporto IT e aumentare la produttività.

Considerazione 7: Collabora con esperti della piattaforma

Il deployment e la gestione efficaci di soluzioni di IA generativa richiedono esperienza e conoscenze specializzate. I requisiti di scalabilità, i problemi di affidabilità e l'integrazione con i sistemi esistenti sono tutti aspetti che complicano i deployment in produzione. L'utilizzo inefficiente delle risorse di elaborazione può portare a spese inutili. Senza contare che il mancato rispetto degli standard di sicurezza, delle policy in merito alla privacy e degli impianti normativi per l'IA possono portare a conseguenze indesiderate.

Scegli fornitori che garantiscano un team di esperti in grado di offrire supporto completo e linee guida per la creazione di soluzioni di IA generativa. Ad esempio, un team di tecnici dedicato può supportare l'intera piattaforma con strumenti, risorse e conoscenze mirate e velocizzare i progetti di IA. La consulenza di esperti costituisce un valido aiuto per risolvere i problemi di deployment, migliorare l'efficienza dell'infrastruttura e garantire l'interoperabilità della soluzione di IA. Inoltre, i servizi di formazione professionale possono aiutare ad acquisire le conoscenze e le competenze necessarie per avviare più rapidamente nuovi progetti di IA generativa.

La collaborazione è un punto cardine dell'IA generativa

Creare un team interfunzionale che abbia capacità diversificate è essenziale per il successo dei progetti di IA generativa.[3]

- **Leader aziendali:** utilizzano e godono dei vantaggi della soluzione.
- **Specialisti dell'IA:** ottimizzano, gestiscono e aggiornano i modelli di IA generativa.
- **Data scientist:** elaborano e forniscono dati corretti e neutri per l'addestramento dei modelli.
- **Responsabili dell'etica e della conformità:** garantiscono che le iniziative di IA siano conformi alle normative.
- **Specialisti delle operazioni IT:** integrano le soluzioni nell'infrastruttura esistente e applicano i criteri di sicurezza.

*3 Kearney, "**Standing up tiger teams to tackle generative AI complexity**", novembre 2023.*

Accelera l'innovazione con una base flessibile e open source

Red Hat offre una gamma completa di soluzioni tecnologiche, una comprovata esperienza e partnership strategiche per aiutare i clienti a realizzare i loro obiettivi nell'ambito dell'IA generativa. Fornisce una base per lo sviluppo e il deployment delle applicazioni e dei modelli di IA generativa, nonché servizi e percorsi formativi per accelerare l'adozione.

Red Hat® OpenShift® è una piattaforma applicativa unificata ed enterprise ready per l'innovazione cloud native. Le risorse di elaborazione on demand, il supporto per l'accelerazione hardware e la coerenza tra gli ambienti (on premise, cloud pubblico ed ambienti edge) assicurano la velocità e la flessibilità necessarie per il successo delle iniziative aziendali. Red Hat OpenShift offre una piattaforma self service dove data scientist, data engineer e sviluppatori possono creare rapidamente applicazioni intelligenti. Le funzionalità che favoriscono la collaborazione permettono ai team di creare e condividere con i colleghi e gli sviluppatori i risultati dei modelli in container, in maniera coerente.

Red Hat OpenShift AI, una soluzione basata su Red Hat OpenShift, offre una piattaforma completa per la creazione, l'addestramento, il fine tuning, la distribuzione e il monitoraggio di modelli e applicazioni e supporta le prestazioni e i carichi di lavoro tipici delle moderne soluzioni di IA generativa. I team possono passare rapidamente dalla fase di sperimentazione alla produzione in un ambiente coerente e collaborativo in cui sono integrate soluzioni chiave di partner certificati, tra cui NVIDIA, Intel, Starburst, Anaconda, IBM, Run:ai e Pachyderm. Insieme all'ecosistema di partner tecnologici, Red Hat OpenShift AI fornisce componenti e funzionalità che accelerano lo sviluppo e il deployment di soluzioni di IA generativa innovative in tutto il cloud ibrido.

IBM watsonx.ai AI studio offre una selezione di modelli e opzioni di deployment con le funzionalità di IA generativa adatte alle applicazioni intelligenti. Distribuisci i modelli, compresi modelli open source, di terze parti e quelli sviluppati da IBM, nell'ambiente in cui si trova il carico di lavoro per aumentare le prestazioni e l'efficienza delle soluzioni di IA. Grazie al **modello fondativo sviluppato da IBM** su dati rilevanti per le aziende, si ottengono soluzioni di IA generativa in grado di analizzare le specificità del tuo business, rendendolo più competitivo.

Red Hat Ansible® Lightspeed with IBM watsonx Code Assistant è un servizio di IA generativa che permette ai team di automazione di creare, adottare e gestire i contenuti Ansible in modo più efficiente. Collegato a IBM watsonx Code Assistant, Red Hat Ansible Lightspeed consente di trasformare i tuoi progetti di automazione in codice Ansible tramite prompt formulati in linguaggio naturale. Il servizio aiuta a migliorare la produttività e rende l'automazione più accessibile all'interno dell'organizzazione.

Inizia subito a utilizzare l'IA generativa

L'IA generativa è uno strumento efficace per la creazione di contenuti originali che sta trasformando l'interazione tra l'uomo e la tecnologia.

Grazie alle sue soluzioni, alla comprovata esperienza e alle partnership, Red Hat è in grado di offrire una base comune per sviluppare e distribuire le applicazioni di intelligenza artificiale (IA) e i modelli di machine learning (ML) con trasparenza e controllo. Le sue piattaforme e i suoi strumenti di IA vengono utilizzati per migliorare anche le utilità di altri software open source. Inoltre, le integrazioni dei partner danno accesso a un ampio ecosistema di strumenti di IA affidabili e compatibili con le piattaforme open source, come Red Hat OpenShift AI.

Scopri di più e prova gratuitamente Red Hat OpenShift AI.

Scegli Red Hat Consulting per un'adozione più veloce

Lavora al fianco degli esperti Red Hat per velocizzare l'avvio dei progetti di AI/ML. Red Hat offre servizi di consulenza e formazione che possono aiutare le organizzazioni ad adottare l'AI/ML più rapidamente.

- Scopri di più sui servizi di AI/ML: red.ht/aiml-consulting
- Prenota una discovery session gratuita: redhat.com/consulting

1057715_0324_KVM

Red Hat