



Valutazione generativa del carico di lavoro AI

AWS Linee guida prescrittive



AWS Linee guida prescrittive: Valutazione generativa del carico di lavoro AI

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

I marchi e l'immagine commerciale di Amazon non possono essere utilizzati in relazione a prodotti o servizi che non siano di Amazon, in una qualsiasi modalità che possa causare confusione tra i clienti o in una qualsiasi modalità che denigri o discrediti Amazon. Tutti gli altri marchi non di proprietà di Amazon sono di proprietà dei rispettivi proprietari, che possono o meno essere affiliati, collegati o sponsorizzati da Amazon.

Table of Contents

Introduzione	1
Scopo di questa guida	2
Destinatari e vantaggi	2
Ambito	2
Obiettivi aziendali specifici	4
Considerazioni e prerequisiti per la valutazione	7
Inizia con casi d'uso chiari	7
Garantire l'allineamento aziendale	8
Implementa la governance e la supervisione	8
Risolvi i dati e i prerequisiti tecnici	8
Prendi in considerazione i requisiti delle risorse di elaborazione	8
Affronta le implicazioni sulla privacy e sulla sicurezza	9
Coinvolgi tempestivamente le parti	9
Iterate e imparate	9
Questionario generativo di valutazione del carico di lavoro AI	10
Prontezza	10
Casi d'uso	13
Architettura	16
Storage	17
Regolamenti e conformità	18
Integrazione	19
Test in corso	21
Implementazione e automazione	23
Strategia dei dati	25
Tradurre le informazioni di valutazione in risultati attuabili	28
Passaggi successivi	30
Domande frequenti	31
Qual è l'obiettivo principale?	31
Chi dovrebbe utilizzare questa valutazione?	31
Quali sono i componenti chiave?	31
In che modo questo aiuta a definire l'architettura?	31
Quali sono i vantaggi?	32
Come possiamo implementarlo con successo?	32
Quali sono le sfide?	32

Quali sono i requisiti normativi e di conformità? 32

Qual è il ruolo delle parti interessate? 33

Come possiamo misurare il successo? 33

In che modo l'approccio differisce in base alle dimensioni dell'organizzazione? 33

Risorse 35

Cronologia dei documenti 36

..... xxxvii

Valutazione generativa del carico di lavoro AI

Tabby Ward e Deepak Dixit, Amazon Web Services (AWS)

Novembre 2024 (cronologia [dei documenti](#))

La valutazione generativa del carico di lavoro di intelligenza artificiale è un metodo strategico volto a valutare e migliorare la preparazione di un'organizzazione a creare o aggiornare i propri carichi di lavoro generativi di intelligenza artificiale. Questa valutazione è importante perché l'integrazione dell'IA generativa nelle operazioni aziendali può cambiare notevolmente il modo in cui funzionano le cose e può fornire nuove efficienze e funzionalità. Tuttavia, per adottare con successo l'IA generativa, è essenziale comprendere a fondo i sistemi attuali e avere un piano chiaro per il futuro.

I carichi di lavoro di intelligenza artificiale generativa si riferiscono a attività computazionali che prevedono l'uso di modelli di intelligenza artificiale in grado di creare nuovi contenuti, come testo, immagini, codice o altri tipi di dati. Questi carichi di lavoro richiedono in genere una notevole potenza di calcolo, hardware specializzato come GPUs e set di dati di grandi dimensioni per l'addestramento e l'inferenza. L'integrazione dei carichi di lavoro di intelligenza artificiale generativa nelle operazioni presenta diverse sfide:

- **Requisiti dell'infrastruttura:** fornitura delle risorse computazionali significative e dell'hardware specializzato richiesti dai modelli di intelligenza artificiale generativa.
- **Gestione dei dati:** garanzia della qualità, della privacy e della conformità dei dati durante la gestione di set di dati di grandi dimensioni.
- **Divario di competenze:** mancanza di esperienza nelle tecnologie di intelligenza artificiale e nell'implementazione dei modelli.
- **Considerazioni etiche:** affrontare i pregiudizi, l'equità e la trasparenza nei contenuti generati dall'intelligenza artificiale.
- **Complessità di integrazione:** integrazione perfetta dell'IA generativa nei flussi di lavoro e nei sistemi legacy esistenti.
- **Gestione dei costi:** bilanciamento dei potenziali vantaggi con gli elevati costi di implementazione e funzionamento.

Il superamento di queste sfide richiede un'attenta pianificazione, investimenti in infrastrutture e talenti e un approccio strategico all'implementazione.

Scopo di questa guida

L'intelligenza artificiale generativa sta rapidamente diventando una componente fondamentale in molti settori. Offre opportunità di trasformazione ma pone anche sfide in termini di integrazione, conformità e scalabilità. Molte organizzazioni hanno difficoltà a sfruttare appieno l'IA a causa della debolezza delle basi tecnologiche, della resistenza al cambiamento e dei problemi di qualità dei dati. La valutazione generativa del carico di lavoro dell'IA affronta queste sfide identificando i requisiti per la modernizzazione, definendo l'ambito di implementazione e mettendo alla prova i sistemi e il modo di pensare esistenti. Inoltre aiuta a determinare i prodotti minimi praticabili (MVPs) e aiuta a sviluppare un'architettura di soluzioni mirata, garantendo un approccio strutturato e strategico all'adozione dell'IA.

Questa guida funge da approccio strutturato per aiutare le organizzazioni a superare le complessità legate all'adozione di tecnologie di intelligenza artificiale generativa. Invece di definire chiaramente i requisiti fin dall'inizio, la guida aiuta a:

- Identificazione dei potenziali casi d'uso dell'IA generativa all'interno dell'organizzazione.
- Valutare la preparazione dell'organizzazione all'adozione dell'IA generativa.
- Definizione e perfezionamento degli obiettivi dei casi d'uso e degli stretch goal.
- Determinazione dell'ambito e dei requisiti per l'implementazione dell'IA generativa.
- Sviluppo di un'architettura di soluzione mirata.

Destinatari e vantaggi

Questa valutazione è progettata specificamente per architetti di soluzioni, architetti aziendali e architetti di applicazioni che desiderano valutare gli aspetti tecnici della modernizzazione generativa dei carichi di lavoro AI. È utile anche per i responsabili dei programmi e del personale che desiderano valutare i requisiti complessivi di preparazione, allocazione delle risorse e abilitazione del proprio team. Le migliori pratiche del settore sottolineano l'importanza di una valutazione completa per garantire la preparazione all'adozione dell'IA. Ciò include la valutazione dell'architettura, dello storage, della conformità, dell'integrazione, dei test, dell'implementazione e dell'automazione.

Ambito

I seguenti argomenti rientrano nell'ambito del metodo di valutazione generativa del carico di lavoro AI:

- Tecnologie e modelli attuali di intelligenza artificiale generativa (ad esempio, modelli linguistici di grandi dimensioni, modelli di generazione di immagini)
- Applicazioni di intelligenza artificiale ristrette che utilizzano tecniche generative
- Integrazione dell'IA generativa con i sistemi e i flussi di lavoro esistenti
- Strategie di dati per la formazione e la messa a punto dei modelli di intelligenza artificiale generativa
- Considerazioni etiche e pratiche di intelligenza artificiale responsabili per le attuali applicazioni di intelligenza artificiale generativa
- Strategie di test e implementazione per l'IA generativa negli ambienti di produzione
- Considerazioni sulla sicurezza e sulla privacy per le implementazioni di intelligenza artificiale generativa
- Ottimizzazione delle prestazioni e scalabilità dei carichi di lavoro di intelligenza artificiale generativa
- Casi d'uso e applicazioni dell'IA generativa in vari settori
- Valutazione dei risultati generativi dell'IA e dei processi di garanzia della qualità

I seguenti argomenti non rientrano nell'ambito di applicazione:

- Scenari di intelligenza artificiale generale (AGI) e superintelligenza artificiale (ASI)
- Progressi speculativi futuri nell'IA oltre gli attuali modelli generativi
- Applicazioni di calcolo quantistico nell'IA
- Calcolo neuromorfico e interfacce cervello-computer
- Coscienza e consapevolezza di sé nei sistemi di intelligenza artificiale
- Impatti sociali a lungo termine dell'IA avanzata oltre alle attuali applicazioni di intelligenza artificiale generativa
- Quadri normativi per ipotetiche tecnologie di intelligenza artificiale future
- Dibattiti filosofici sulla natura dell'intelligenza e della coscienza nelle macchine
- Casi estremi o casi d'uso altamente speculativi dell'IA
- Specifiche tecniche dettagliate di modelli o architetture di intelligenza artificiale proprietarie

Obiettivi aziendali specifici

La valutazione generativa del carico di lavoro AI mira a fornire diversi risultati mirati che sono fondamentali per modernizzare con successo i carichi di lavoro di intelligenza artificiale generativa. Questi risultati garantiscono che le organizzazioni siano ben preparate a integrare le tecnologie di intelligenza artificiale in modo efficace ed efficiente.

Per ogni risultato mirato, la valutazione generativa del carico di lavoro dell'IA si concentra su:

- **Interdipendenze:** identifica e chiarisci eventuali interdipendenze tra il risultato e altri aspetti del processo di modernizzazione. Ciò include la comprensione di come un risultato possa influenzare o essere influenzato da altri, per garantire un approccio olistico alla modernizzazione.
- **Allineamento delle parti interessate:** delinea le strategie per allineare le varie parti interessate a ciascun risultato. Ciò implica la comunicazione del valore e dell'impatto di ciascun risultato ai diversi livelli organizzativi e dipartimenti, per favorire il consenso e il supporto.
- **Assegnazione delle priorità:** nei casi in cui vengono identificati più casi d'uso o risultati, fornisci un framework per assegnare loro priorità in base a fattori quali l'impatto aziendale, i requisiti di risorse e l'allineamento strategico.
- **Miglioramento continuo:** per ogni risultato, stabilisci meccanismi di valutazione e perfezionamento continui. Ciò garantisce che gli sforzi di modernizzazione rimangano adattivi e reattivi ai cambiamenti del panorama tecnologico e delle esigenze aziendali.

Ecco una discussione dettagliata di ogni risultato mirato:

Architettura di Target

- **Definizione:** la valutazione aiuta a definire un'architettura target chiara e scalabile per carichi di lavoro di intelligenza artificiale generativa.
- **Componenti:** ciò include la selezione dei servizi cloud appropriati, la progettazione di pipeline di dati e la garanzia dell'interoperabilità del sistema.
- **Vantaggi:** un'architettura ben definita supporta scalabilità, affidabilità e ottimizzazione delle prestazioni e fornisce una solida base per la modernizzazione.

Disponibilità del cliente

- **Valutazione:** valuta lo stato attuale dell'infrastruttura, dei processi e della cultura dell'organizzazione per determinare la preparazione all'adozione della modernizzazione generativa dell'IA.
- **Criteri:** ciò implica la valutazione delle capacità tecniche, della qualità dei dati e della disponibilità organizzativa ad accettare il cambiamento.
- **Risultato:** l'identificazione delle lacune e delle aree di miglioramento garantisce che l'organizzazione sia pronta per una transizione graduale verso soluzioni e tecnologie moderne.

Usa gli obiettivi dei casi e gli stretching goal

- Gli obiettivi dei casi d'uso stabiliscono obiettivi chiari per l'implementazione di soluzioni mirate, concentrandosi su problemi o opportunità aziendali specifici.

Un obiettivo di un caso d'uso nel contesto della modernizzazione generativa dell'IA si riferisce a un obiettivo specifico e misurabile che un'organizzazione intende raggiungere implementando soluzioni di intelligenza artificiale generativa. Questi obiettivi sono in genere allineati con obiettivi aziendali più ampi e si concentrano sull'affrontare sfide o opportunità particolari all'interno dell'organizzazione. Alcuni esempi di obiettivi dei casi d'uso potrebbero includere:

- Riduzione del 50% dei tempi di risposta del servizio clienti utilizzando chatbot generativi basati sull'intelligenza artificiale.
- Miglioramento dell'efficienza di revisione del codice del 30 per cento attraverso l'analisi generativa del codice assistita dall'intelligenza artificiale.
- Miglioramento della precisione del rilevamento delle frodi del 25% utilizzando il riconoscimento generativo dei pattern AI.
- Gli Stretch goals definiscono obiettivi ambiziosi che superano i limiti di ciò che la modernizzazione generativa dell'IA può raggiungere all'interno dell'organizzazione.
- **Impatto:** la definizione di obiettivi raggiungibili e ambiziosi aiuta ad allineare le iniziative di modernizzazione generativa dell'IA con gli obiettivi aziendali strategici e incoraggia l'innovazione.

Stima dello sforzo

- **Scopo:** una stima accurata degli sforzi aiuta nella pianificazione delle risorse e garantisce che i progetti vengano consegnati in tempo e nel rispetto del budget.
- **Ambito:** stima le risorse, il tempo e il budget necessari per implementare il piano di modernizzazione generativa dell'IA.
- **Fattori:** considera la complessità tecnica, le sfide di integrazione e i potenziali rischi.

Esigenze di abilitazione

- **Formazione e sviluppo:** Identifica le competenze e le conoscenze necessarie per adottare con successo la modernizzazione generativa dell'IA.
- **Risorse:** determina la necessità di programmi di formazione, workshop e altre attività di abilitazione.
- **Risultato:** garantire che il personale sia dotato delle competenze necessarie migliora l'efficacia delle iniziative di modernizzazione generativa dell'IA e favorisce il successo a lungo termine.

Piano di implementazione

- **Roadmap:** Sviluppa un piano dettagliato che delinei i passaggi necessari per raggiungere la modernizzazione dell'IA generativa.
- **Pietre miliari:** definisci le tappe fondamentali e i risultati finali per monitorare i progressi.
- **Vantaggi:** un piano di implementazione chiaro fornisce indicazioni e responsabilità e facilita un approccio strutturato alla modernizzazione dell'IA generativa.

Considerazioni e prerequisiti per la valutazione

Inizia con casi d'uso chiari

Identifica problemi o opportunità aziendali specifici che l'IA generativa può risolvere. Concentrati sui casi d'uso che siano in linea con gli obiettivi aziendali strategici e offrano vantaggi misurabili. Dai priorità ai casi d'uso che riguardano le sfide più comuni all'interno dell'organizzazione per garantire che l'architettura della soluzione possa fungere da modello per più scenari.

Avviare il processo di valutazione con una comprensione generale delle potenziali applicazioni di intelligenza artificiale generativa è vantaggioso ma non obbligatorio. Il [questionario](#) incluso in questa guida soddisfa vari livelli di preparazione, dalle organizzazioni che hanno casi d'uso ben definiti a quelle che hanno solo idee generali. Il processo di valutazione serve a:

- Perfeziona e chiarisci queste idee iniziali sui casi d'uso.
- Identifica nuovi potenziali casi d'uso.
- Sviluppa obiettivi specifici e misurabili per ogni caso d'uso.
- Valuta la fattibilità e il potenziale impatto di ogni caso d'uso.

Prendiamo in considerazione un esempio ipotetico: una società di servizi finanziari decide di esplorare la modernizzazione dell'IA generativa. Iniziano con un'idea generale di migliorare il servizio clienti e i processi di rilevamento delle frodi.

- Valutazione iniziale: il questionario li aiuta a valutare i sistemi attuali, la qualità dei dati e la preparazione organizzativa per l'adozione dell'IA generativa.
- Perfezionamento dei casi d'uso: attraverso il processo di valutazione, perfezionano le loro idee iniziali in due casi d'uso specifici:
 - Implementazione di un chatbot generativo basato sull'intelligenza artificiale per le richieste dei clienti
 - Utilizzo dell'intelligenza artificiale generativa per il rilevamento delle frodi nelle transazioni in tempo reale
- Definizione degli obiettivi: per ogni caso d'uso, definiscono obiettivi specifici:
 - Ridurre i tempi di risposta del servizio clienti del 40% entro 6 mesi
 - Migliora la precisione del rilevamento delle frodi del 20% e riduci i falsi positivi del 15%

- Obiettivi estensibili: hanno inoltre fissato questi obiettivi ambiziosi:
 - Raggiungi l'80% di soddisfazione dei clienti con risposte assistite dall'intelligenza artificiale
 - Sviluppa un modello predittivo di rilevamento delle frodi che identifichi nuovi modelli di frode
- Definizione di MVP: il questionario li aiuta a determinare un MVP per ogni caso d'uso, concentrandosi sulle funzionalità essenziali che offrono valore immediato.
- Architettura di destinazione: Infine, sviluppano un'architettura target che supporta uno o entrambi i casi d'uso e garantisce la scalabilità e l'integrazione con i sistemi esistenti.

Garantire l'allineamento aziendale

Allinea le iniziative di intelligenza artificiale generativa alla strategia e agli obiettivi aziendali generali. Per ogni caso d'uso, sviluppa una proposta di valore chiara che dimostri come l'IA generativa contribuisca alla crescita, all'efficienza o all'innovazione del business. Stabilisci metriche per misurare l'impatto delle implementazioni di intelligenza artificiale generativa sugli indicatori chiave di performance (). KPIs

Implementa la governance e la supervisione

Crea un comitato direttivo interfunzionale per supervisionare le iniziative di intelligenza artificiale generativa. Sviluppa politiche e linee guida per un uso responsabile dell'IA, affrontando considerazioni etiche e potenziali pregiudizi. Stabilisci un processo di revisione per i progetti di intelligenza artificiale generativa per garantire la conformità agli standard organizzativi e ai requisiti normativi.

Risolvi i dati e i prerequisiti tecnici

Valuta e migliora la qualità dei dati e implementa pratiche di governance dei dati per garantire input affidabili per i modelli di intelligenza artificiale generativa. Sviluppa una strategia di dati che affronti la raccolta, l'archiviazione e la gestione dei dati specifiche per le esigenze di intelligenza artificiale generativa. Valuta e migliora l'infrastruttura di dati per supportare il volume e la velocità dei dati necessari per i carichi di lavoro di intelligenza artificiale generativa.

Prendi in considerazione i requisiti delle risorse di elaborazione

Valuta l'attuale infrastruttura IT e identifica le lacune nella capacità computazionale per i carichi di lavoro generativi di intelligenza artificiale. Pianifica risorse di elaborazione scalabili, considerando

opzioni come servizi cloud o cluster di elaborazione ad alte prestazioni locali. Ottimizza l'allocazione delle risorse per bilanciare prestazioni ed economicità per i carichi di lavoro di formazione e inferenza.

Affronta le implicazioni sulla privacy e sulla sicurezza

Implementa solide misure di sicurezza per proteggere i dati sensibili utilizzati nella formazione e nelle operazioni di intelligenza artificiale generativa. Garantisci la conformità alle normative sulla protezione dei dati come il Regolamento generale sulla protezione dei dati (GDPR) o il California Consumer Privacy Act (CCPA) nella gestione delle informazioni personali. Sviluppa protocolli per l'implementazione e il monitoraggio sicuri dei modelli per prevenire l'accesso non autorizzato o l'uso improprio delle funzionalità di intelligenza artificiale generativa.

Coinvolgi tempestivamente le parti

Coinvolgi le principali parti interessate sin dall'inizio per ottenere il consenso e il supporto della leadership. Comunica chiaramente i vantaggi e il potenziale impatto delle iniziative di modernizzazione, in particolare per i carichi di lavoro generativi di intelligenza artificiale. Fornisci formazione e risorse per aiutare le parti interessate a comprendere le tecnologie di intelligenza artificiale generativa e le loro implicazioni.

Iterate e imparate

Adotta un approccio incrementale che ti consenta di perfezionare le soluzioni target. Utilizza i cicli di feedback per migliorare continuamente l'architettura e i processi dei carichi di lavoro. Valuta regolarmente le prestazioni e l'impatto delle implementazioni di intelligenza artificiale generativa e adatta le strategie secondo necessità in base ai risultati del mondo reale e alle esigenze aziendali in evoluzione.

Questionario generativo di valutazione del carico di lavoro AI

Le sezioni seguenti forniscono domande che puoi utilizzare per valutare diversi aspetti della modernizzazione generativa dei carichi di lavoro AI per la tua organizzazione. Questo questionario completo valuta la preparazione dell'organizzazione ad adottare e implementare carichi di lavoro di intelligenza artificiale generativa con domande riguardanti aree chiave, tra cui casi d'uso, architettura, storage, conformità, integrazione, test, implementazione e strategia dei dati. Affrontando gli aspetti critici dell'implementazione dell'IA generativa, dall'infrastruttura tecnica alle considerazioni normative, questo questionario ti aiuta a identificare i punti di forza, le lacune e le opportunità nel tuo percorso di modernizzazione dell'IA.

Sezioni:

- [Prontezza](#)
- [Casi d'uso](#)
- [Architettura](#)
- [Storage](#)
- [Regolamenti e conformità](#)
- [Integrazione](#)
- [Test in corso](#)
- [Implementazione e automazione](#)
- [Strategia dei dati](#)

Puoi anche scaricare il questionario in formato Microsoft Excel e utilizzarlo per registrare le tue informazioni.

 [il questionario](#)

Scarica

Prontezza

Domanda	Example response
Disponi di AWS account che possono essere utilizzati per questi carichi di lavoro?	Sì o no.

Domanda	Example response
Hai un contratto aziendale esistente con AWS?	Sì o no
Quanto è scalabile la tua attuale infrastruttura cloud per gestire carichi di lavoro di intelligenza artificiale generativa?	La nostra infrastruttura cloud è altamente scalabile, con funzionalità di scalabilità automatica per le risorse di calcolo e sistemi di archiviazione distribuiti progettati per gestire in modo efficiente carichi di lavoro di intelligenza artificiale generativa su larga scala.
Disponi di funzionalità di pipeline di dati per la preelaborazione e l'ingegneria delle funzionalità su larga scala?	Le nostre pipeline di dati utilizzano framework di elaborazione distribuiti come Apache Spark per la preelaborazione dei dati su larga scala e l'ingegneria delle funzionalità, con supporto per l'elaborazione di dati in batch e in streaming.
Disponete di funzionalità di fornitura e gestione degli account?	Sì o no.
Come descriveresti l'alfabetizzazione in materia di intelligenza artificiale e la disponibilità della tua organizzazione ad adottare tecnologie di intelligenza artificiale generativa?	La nostra organizzazione ha investito molto in programmi di formazione sull'intelligenza artificiale e la maggior parte del personale tecnico ha completato la formazione di base sull'AI/ML. L'organizzazione ha una cultura dell'innovazione che abbraccia le nuove tecnologie, inclusa l'intelligenza artificiale generativa.
Quali competenze in materia di AI/ML esistono all'interno della vostra organizzazione e come vengono distribuite?	Abbiamo un Centro di eccellenza dedicato all'intelligenza artificiale con data scientist esperti e ingegneri ML. Miglioriamo le competenze degli esperti di settore in diverse unità aziendali per diventare esperti di intelligenza artificiale e identificare i casi d'uso dell'IA generativa.

Domanda	Example response
Hai un business case di alto livello che illustra gli obiettivi, i vantaggi e i costi del programma cloud?	Sì o no.
Qual è la tempistica prevista per portare la soluzione in produzione?	Settimane, mesi e così via.
I vostri principali stakeholder (ad esempio, CFO, CIT/CTO, COO) si sono impegnati a finanziare?	Sì o no
Come garantite la conformità alle normative sulla protezione dei dati nelle vostre iniziative di intelligenza artificiale generativa?	Abbiamo un team di conformità dedicato che lavora a stretto contatto con i nostri team di intelligenza artificiale. Conduciamo regolarmente valutazioni dell'impatto sulla privacy, implementiamo i principi di protezione dei dati fin dalla progettazione e conserviamo registri dettagliati di elaborazione dei dati per tutti i progetti di intelligenza artificiale generativa.
Quanto sono maturi i sistemi esistenti che si integrano con le nuove tecnologie di intelligenza artificiale generativa?	La nostra architettura IT si basa su microservizi e consente APIs l'integrazione flessibile di nuove tecnologie di intelligenza artificiale generativa. Questi sistemi sono standardizzati su formati e protocolli di dati comuni per garantire l'interoperabilità.
Che esperienza avete nell'operatività dei modelli di machine learning e in che modo ciò potrebbe applicarsi ai sistemi di intelligenza artificiale generativa?	Disponiamo di MLOps pratiche consolidate, tra cui pipeline di implementazione automatizzata dei modelli, sistemi di monitoraggio e framework di test A/B. Queste pratiche vengono adattate per gestire i requisiti unici dei modelli di intelligenza artificiale generativa su larga scala.

Casi d'uso

Domanda	Example response
Qual è l'obiettivo principale o il criterio di successo del caso d'uso?	Per migliorare i tempi di risposta dell'assistenza clienti, aumentare le conversioni di vendita, migliorare i consigli sui prodotti. Inoltre: per migliorare la soddisfazione degli utenti, il tasso di completamento delle attività, la qualità della risposta e così via.
In che modo questo caso d'uso si allinea agli obiettivi strategici della vostra organizzazione?	Ciò è in linea con il nostro obiettivo strategico di aumentare la soddisfazione del cliente riducendo i tempi di risposta nel servizio clienti.
Qual è il volume di dati o di richieste previsto per il caso d'uso?	500 transazioni al secondo (TPS).
Quali tipi di fonti di dati sono necessarie per supportare i carichi di lavoro di intelligenza artificiale generativa?	Database strutturati interni (record dei clienti, dati di vendita e così via); dati di testo non strutturati provenienti da documenti, e-mail e social media; file audio e video per attività di riconoscimento vocale e di immagini; dati di streaming in tempo reale da dispositivi e sensori IoT; set di dati pubblici e APIs per l'arricchimento.
Con che frequenza è necessario aggiornare o aggiornare i dati provenienti da queste fonti?	Database transazionali: aggiornamenti quasi in tempo reale; archivi di documenti: aggiornamenti giornalieri in batch; feed di social media: aggiornamenti orari; dati dei sensori IoT: streaming continuo in tempo reale; set di dati pubblici: aggiornamenti mensili o trimestrali.
Quali formati di dati richiedono come input i tuoi modelli di intelligenza artificiale generativa?	Dati strutturati: tabelle di database CSV, JSON e SQL; dati di testo: testo semplice, PDF e

Domanda	Example response
	HTML; dati di immagine: JPEG, PNG e TIFF; dati audio: WAV e; dati video: e AVI MP3. MP4
Quali sono i tuoi principali problemi di qualità dei dati per i carichi di lavoro di intelligenza artificiale generativa?	Completezza: garantire che non manchino campi critici; precisione: verifica della correttezza dei dati ed eliminazione degli errori; coerenza: mantenimento di formati e valori uniformi tra le fonti; tempestività: garanzia che i dati siano aggiornati per l'inferenza in tempo reale; pertinenza: conferma che i dati siano allineati con lo specifico compito di intelligenza artificiale generativa.
Quali sono i requisiti prestazionali chiave (ad esempio, tempo di risposta, produttività, precisione)?	Precisione del 95%; tempo di risposta < 500 ms; capacità di gestire 1000 richieste/sec. Alta precisione (95% +), precisione moderata (80-90%), massimo sforzo e così via.
Ne avete altri KPIs per misurare il successo di questo caso d'uso?	KPIs I fattori chiave includono la riduzione del tasso di errore, il risparmio di tempo per transazione e i punteggi di soddisfazione dei clienti.
Quanta precisione del modello si desidera e come si bilancia con il costo?	Elevata precisione (> 90%) con costi moderati, precisione moderata (70-80%) a basso costo e così via.
Quali sono i casi d'uso o gli scenari principali per la soluzione di intelligenza artificiale generativa?	Chatbot per il servizio clienti, generazione di contenuti, consigli sui prodotti e così via.
Quali sono gli utenti o i personaggi target del sistema di intelligenza artificiale generativa?	Agenti del servizio clienti, team di marketing, dipendenti, utenti finali e così via.
Qual è il volume previsto di richieste o utenti?	1.000 richieste al giorno; 10.000 utenti attivi al mese.

Domanda	Example response
Esistono vincoli o requisiti specifici relativi ai casi d'uso?	Risposta in tempo reale, supporto multilingue, privacy dei dati e così via.
Hai un budget stanziato per lo sviluppo e la manutenzione della soluzione di intelligenza artificiale generativa?	Il costo di sviluppo iniziale è stimato in 200.000 dollari, con costi di manutenzione annuali di 50.000 dollari.
Quali sono il ritorno sull'investimento (ROI) e il periodo di ammortamento previsti per questo caso d'uso?	ROI previsto del 150% in tre anni, con un periodo di recupero dell'investimento di 18 mesi.
Ci sono costi nascosti o potenziali risparmi da prendere in considerazione?	I potenziali risparmi includono la riduzione dei costi straordinari. I costi nascosti potrebbero comportare una formazione aggiuntiva per il personale.
Quali sono la scalabilità e le possibilità di espansione future di questa soluzione di intelligenza artificiale generativa?	La soluzione è progettata per adattarsi alle nostre operazioni, con la possibilità di espandersi ad altri reparti in futuro.
Come garantite l'equità e mitigate i pregiudizi nei vostri modelli di intelligenza artificiale generativa?	Intendiamo mitigare i pregiudizi attraverso la raccolta di dati diversificati, controlli periodici dei pregiudizi e l'implementazione di tecniche di mitigazione dei pregiudizi.
Quali processi avete messo in atto per affrontare le preoccupazioni etiche o le conseguenze non intenzionali?	Gestiremo le preoccupazioni etiche attraverso un piano di risposta agli incidenti basato sull'intelligenza artificiale consolidato, valutazioni periodiche dei rischi etici, un sistema di segnalazione anonimo per i dipendenti, la collaborazione con esperti di etica esterni e il monitoraggio e l'adeguamento continui dei modelli implementati in base al feedback.

Domanda	Example response
Come affrontate l'assegnazione delle priorità e il sequenziamento delle valutazioni generative e dei carichi di lavoro basati sull'intelligenza artificiale in diversi progetti e reparti della vostra organizzazione?	Conducendo un'indagine di alto livello in tutti i reparti per identificare potenziali casi d'uso dell'IA generativa e valutarli in base a tre criteri chiave: impatto sul business, fattibilità tecnica e considerazioni etiche. Viene data priorità ai progetti con un impatto potenziale elevato, barriere tecniche inferiori e preoccupazioni etiche minime.

Architettura

Domanda	Example response
Che tipo di modello o architettura di intelligenza artificiale generativa viene preso in considerazione?	Trasformatore, rete neurale convoluzionale (CNN), rete neurale ricorrente (RNN), alberi decisionali e così via.
Qual è la scala o il volume previsto di dati e calcoli?	Milioni di utenti, petabyte di dati e così via.
Quali sono i requisiti hardware (ad esempio, CPUs o GPUs) per l'addestramento e l'inferenza?	Cluster CPU di fascia alta GPUs, istanze cloud e così via.
In che modo il modello di intelligenza artificiale generativa verrà aggiornato o riqualificato nel tempo?	Attraverso l'apprendimento continuo, la riqualificazione periodica, gli aggiornamenti manuali e così via.
Quali sono i requisiti di preelaborazione dei dati e di progettazione delle funzionalità?	Pulizia del testo, aumento delle immagini, selezione delle funzionalità e così via.
In che modo il sistema di intelligenza artificiale generativa gestirà i casi limite, i valori anomali o gli input con scarsa confidenza?	Ricorrendo alla supervisione umana, richiedi chiarimenti e così via.

Domanda	Example response
Quali sono i requisiti di latenza per l'applicazione AI generativa?	Elaborazione in batch in tempo reale, quasi in tempo reale e così via.

Storage

Domanda	Example response
Dove verranno archiviati i dati di allenamento?	Nell'archiviazione cloud (ad esempio, Amazon S3, archiviazione di file, archiviazione a blocchi o archiviazione di oggetti), nell'archiviazione locale e così via.
Quali sono i requisiti di archiviazione per i dati di addestramento e gli artefatti del modello (ad esempio, capacità, durabilità, disponibilità)?	Storage su scala petabyte, elevata durabilità (99,99999% di durabilità), alta disponibilità e così via.
Quali sono i requisiti di conservazione e backup dei dati per i dati di addestramento e gli artefatti del modello?	Conservazione dei dati per x anni, backup giornalieri, backup fuori sede e così via.
Quali formati di file vengono utilizzati principalmente per archiviare i set di dati di addestramento AI (ad esempio, CSV, JSON, Parquet,)? HDF5	File Parquet per dati strutturati e HDF5 per array multidimensionali di grandi dimensioni e dati non strutturati come immagini e testo. Utilizziamo formati specializzati, ad esempio per ottimizzare il caricamento dei dati durante l'TFRecord allenamento.
Come sono organizzati i set di dati di formazione: come singoli file, in database o utilizzando formati di dati AI specializzati?	I set di dati di piccole e medie dimensioni vengono archiviati come singoli file Parquet nello storage a oggetti per garantire la massima flessibilità. I set di dati di grandi dimensioni vengono archiviati in un database distribuito (Cassandra) per gestire la scalabilità.

Domanda	Example response
Utilizzate tecniche di compressione o codifica dei dati specifiche per i dati di addestramento generativo dell'intelligenza artificiale?	Per i dati tabulari, utilizziamo tecniche di codifica dei dizionari e di bit-packing disponibili in Parquet. Per le immagini, utilizziamo la compressione JPEG con perdita di dati con impostazioni di qualità ottimizzate per i nostri modelli.
Come gestite il controllo delle versioni e l'archiviazione di diverse iterazioni di set di dati di addestramento? Che impatto ha questo sulle vostre esigenze complessive di storage?	Utilizziamo un sistema di versione dei dati (DVC) integrato con la nostra piattaforma ML.

Regolamenti e conformità

Domanda	Example response
Quali sono le normative o i requisiti di conformità pertinenti per la soluzione di intelligenza artificiale generativa (ad esempio, GDPR, HIPAA, PCI-DSS)?	GDPR per la gestione dei dati personali, HIPAA per i dati sanitari, PCI-DSS per i dati di pagamento e così via.
Quali linee guida o framework etici di intelligenza artificiale generativa ha adottato la tua organizzazione?	Abbiamo implementato le nostre linee guida sull'intelligenza artificiale responsabile. Tutti i progetti di intelligenza artificiale generativa vengono sottoposti a una revisione etica prima dell'approvazione e dell'implementazione.
Quali sono i requisiti di sicurezza per il sistema di intelligenza artificiale generativa?	Crittografia dei dati, comunicazioni di rete sicure, controlli di sicurezza regolari.
Quali sono i requisiti per la privacy e la protezione dei dati?	Anonimizzazione dei dati, crittografia, controllo degli accessi e così via.
Quali sono i requisiti della soluzione per gestire dati sensibili o riservati?	Controlli di accesso rigorosi, mascheramento dei dati, requisiti di residenza dei dati e così via.

Domanda	Example response
Come verranno gestite l'autenticazione e l'autorizzazione degli utenti?	Utilizzando chiavi API OAuth, Single Sign-On (SSO) e controllo degli accessi basato sui ruoli (RBAC).
Come verrà monitorata e gestita la soluzione in produzione?	Utilizzando strumenti di monitoraggio come Prometheus e Datadog, strumenti di registrazione come ELK Stack, sistemi di avviso e così via.

Integrazione

Domanda	Example response
Quali sono i requisiti per integrare la soluzione di intelligenza artificiale generativa con i sistemi o le fonti di dati esistenti?	REST APIs, code di messaggi, connettori di database e così via.
In che modo verranno acquisiti e preelaborati i dati per la soluzione di intelligenza artificiale generativa?	Utilizzando l'elaborazione in batch, lo streaming di dati, le trasformazioni dei dati e l'ingegneria delle funzionalità.
Come verrà utilizzato o integrato il risultato della soluzione di intelligenza artificiale generativa con i sistemi downstream?	Tramite endpoint API, code di messaggi, aggiornamenti del database e così via.
Quali modelli di integrazione basati sugli eventi possono essere utilizzati per la soluzione di intelligenza artificiale generativa?	Code di messaggi (come Amazon SQS, Apache Kafka, RabbitMQ), sistemi pub/sub, webhook, piattaforme di streaming di eventi.
Quali approcci di integrazione basati su API possono essere utilizzati per connettere la soluzione di intelligenza artificiale generativa con altri sistemi?	RESTful APIs, GraphQL APIs, SOAP APIs (per sistemi legacy).

Domanda	Example response
Quali componenti dell'architettura dei microservizi possono essere utilizzati per l'integrazione di soluzioni di intelligenza artificiale generativa?	Service Mesh per comunicazioni tra servizi, gateway API, orchestrazione di container (ad esempio, Kubernetes).
Come può essere implementata l'integrazione ibrida per la soluzione di intelligenza artificiale generativa?	Combinando modelli basati sugli eventi per aggiornamenti in tempo reale, elaborazione in batch per dati storici e APIs integrazione di sistemi esterni.
In che modo l'output della soluzione di intelligenza artificiale generativa può essere integrato con i sistemi downstream?	Tramite endpoint API, code di messaggi, aggiornamenti del database, webhook ed esportazioni di file.
Quali misure di sicurezza devono essere prese in considerazione per l'integrazione della soluzione di intelligenza artificiale generativa?	Meccanismi di autenticazione (come OAuth o JWT), crittografia (in transito e a riposo), limitazione della velocità delle API ed elenchi di controllo degli accessi (ACLs).
Come intendete integrare framework open source come LlamaIndex o LangChain nella pipeline di dati esistente e nel flusso di lavoro di intelligenza artificiale generativa?	Abbiamo intenzione di utilizzarlo per LangChain creare applicazioni di intelligenza artificiale generativa complesse, in particolare per le sue capacità di gestione degli agenti e della memoria. Il nostro obiettivo è utilizzare il 60% dei nostri progetti di intelligenza artificiale generativa LangChain entro i prossimi 6 mesi.
Come garantirete la compatibilità tra i framework open source scelti e l'infrastruttura di dati esistente?	Stiamo creando un team di integrazione dedicato per garantire una perfetta compatibilità. Entro il terzo trimestre, il nostro obiettivo è disporre di una pipeline completamente integrata che consenta l'indicizzazione e LlamaIndex il recupero efficienti dei dati all'interno della nostra attuale struttura di data lake.

Domanda	Example response
Come intendete sfruttare i componenti modulari dei framework, ad esempio per la prototipazione e la sperimentazione rapide? LangChain	Stiamo configurando un ambiente sandbox in cui gli sviluppatori possano prototipare rapidamente utilizzando i componenti di Microsoft. LangChain
Qual è la tua strategia per stare al passo con gli aggiornamenti e le nuove funzionalità di questi framework open source in rapida evoluzione?	Abbiamo assegnato un team per monitorare gli GitHub archivi e i forum della community per e. LangChain LlamaIndex Abbiamo intenzione di valutare e integrare i principali aggiornamenti trimestralmente, con particolare attenzione al miglioramento delle prestazioni e alle nuove funzionalità.

Test in corso

Domanda	Example response
Quali sono i requisiti di test (ad esempio, test unitari, test di integrazione, test)? end-to-end	Test unitari per singoli componenti, test di integrazione con sistemi esterni, end-to-end test per scenari critici e così via.
Come garantite la qualità e la coerenza dei dati tra diverse fonti per la formazione generativa sull'intelligenza artificiale?	Manteniamo la qualità dei dati attraverso strumenti automatizzati di profilazione dei dati, controlli regolari dei dati e un catalogo di dati centralizzato. Abbiamo implementato politiche di governance dei dati per garantire la coerenza tra le fonti e mantenere la derivazione dei dati.
Come verrà valutato e convalidato il modello di intelligenza artificiale generativa?	Utilizzando un set di dati holdout, valutazione umana, test A/B e così via.
Quali sono i criteri per valutare le prestazioni e l'accuratezza del modello di intelligenza artificiale generativa?	Precisione, richiamo, punteggio F1, perplessità, valutazione umana e così via.

Domanda	Example response
Come verranno identificati e gestiti i casi limite e i casi isolati?	Utilizzando una suite di test completa, valutazione umana, test antagonistici e così via.
Come verificherete i potenziali pregiudizi nel modello di intelligenza artificiale generativa?	Utilizzando l'analisi della parità demografica, i test sulle pari opportunità, le tecniche di neutralizzazione delle controversie, i test controfattuali e così via.
Quali metriche verranno utilizzate per misurare l'equità dei risultati del modello?	Rapporto di impatto disparato, quote equalizzate, parità demografica, metriche di equità individuali e così via.
Come garantirete una rappresentazione diversificata nei set di dati di test per il rilevamento delle distorsioni?	Utilizzando il campionamento stratificato tra gruppi demografici, la collaborazione con esperti di diversità, l'uso di dati sintetici per colmare le lacune e così via.
Quale processo verrà implementato per il monitoraggio continuo dell'equità del modello dopo l'implementazione?	Controlli di equità regolari, sistemi automatici di rilevamento delle distorsioni, analisi del feedback degli utenti, riqualificazione periodica con set di dati aggiornati e così via.
Come affronterete i pregiudizi intersezionali nel modello di intelligenza artificiale generativa?	Utilizzando l'analisi dell'equità intersezionale, i test dei sottogruppi, la collaborazione con esperti di settore sull'intersezionalità e così via.
Come testerai le prestazioni del modello in diverse lingue e contesti culturali?	Utilizzando set di test multilingue, collaborazione con esperti culturali, metriche di equità localizzate, studi comparativi interculturali e così via.

Implementazione e automazione

Domanda	Example response
Quali sono i requisiti per la scalabilità e il bilanciamento del carico?	Routing intelligente delle richieste, sistema di scalabilità automatico; ottimizzazione per partenze rapide a freddo mediante l'utilizzo di tecniche come il caching dei modelli, il lazy loading e i sistemi di storage distribuiti; progettazione del sistema per gestire modelli di traffico impetuosi e imprevedibili.
Quali sono i requisiti per l'aggiornamento e il lancio di nuove versioni?	Distribuzioni blu/verdi, versioni Canary, aggiornamenti continui e così via.
Quali sono i requisiti per il disaster recovery e la continuità aziendale?	Procedure di backup e ripristino, meccanismi di failover, configurazioni ad alta disponibilità e così via.
Quali sono i requisiti per automatizzare la formazione, l'implementazione e la gestione del modello di intelligenza artificiale generativa?	Pipeline di formazione automatizzata, implementazione continua, scalabilità automatica e così via.
In che modo verrà aggiornato e riqualficato il modello di intelligenza artificiale generativa non appena saranno disponibili nuovi dati?	Attraverso la riqualficazione periodica, l'apprendimento incrementale, l'apprendimento trasferito e così via.
Quali sono i requisiti per automatizzare il monitoraggio e la gestione?	Avvisi automatici, ridimensionamento automatico, riparazione automatica e così via.
Qual è il tuo ambiente di implementazione preferito per i carichi di lavoro di intelligenza artificiale generativa?	Un approccio ibrido che utilizza AWS per la formazione dei modelli e la nostra infrastruttura locale per l'inferenza per soddisfare i requisiti di residenza dei dati.
Esistono piattaforme cloud specifiche che preferisci per le implementazioni di intelligenza artificiale generativa?	Servizi AWS, in particolare Amazon SageMaker AI per lo sviluppo e l'implementazione di

Domanda	Example response
	modelli e Amazon Bedrock per i modelli di base.
Quali tecnologie di containerizzazione state considerando per i carichi di lavoro di intelligenza artificiale generativa?	Vogliamo standardizzarci su contenitori Docker orchestrati con Kubernetes per garantire portabilità e scalabilità nel nostro ambiente ibrido.
Hai degli strumenti preferiti per CI/CD nella tua pipeline di intelligenza artificiale generativa?	GitLab per il controllo delle versioni e le pipeline CI/CD, integrato con Jenkins per test e implementazione automatizzati.
Quali strumenti di orchestrazione state considerando per la gestione dei flussi di lavoro di intelligenza artificiale generativa?	Apache Airflow per l'orchestrazione del flusso di lavoro, in particolare per la preelaborazione dei dati e le pipeline di formazione dei modelli.
Hai requisiti specifici per l'infrastruttura locale per supportare carichi di lavoro di intelligenza artificiale generativa?	Stiamo investendo in server accelerati da GPU e reti ad alta velocità per supportare i carichi di lavoro di inferenza locali.
Come intendete gestire il controllo delle versioni e l'implementazione dei modelli in ambienti diversi?	Abbiamo intenzione di utilizzarlo MLflow per il monitoraggio e il controllo delle versioni dei modelli e di integrarlo con la nostra infrastruttura Kubernetes per un'implementazione senza interruzioni in tutti gli ambienti.
Quali strumenti di monitoraggio e osservabilità state considerando per le implementazioni di intelligenza artificiale generativa?	Prometheus per la raccolta delle metriche e Grafana per la visualizzazione, con soluzioni di registrazione personalizzate aggiuntive per il monitoraggio specifico del modello.
Come state affrontando lo spostamento e la sincronizzazione dei dati in un modello di implementazione ibrido?	Lo utilizzeremo AWS DataSync per un trasferimento efficiente dei dati tra l'archiviazione locale e AWS, con processi di sincronizzazione automatizzati pianificati in base ai nostri cicli di formazione.

Domanda	Example response
Quali misure di sicurezza state implementando per le implementazioni di intelligenza artificiale generativa in diversi ambienti?	Utilizzeremo IAM per le risorse cloud, integrate con il nostro Active Directory locale per implementare la end-to-end crittografia e la segmentazione della rete per proteggere i flussi di dati.

Strategia dei dati

Domanda	Example response
Quali tipi di dati specifici sono fondamentali per i carichi di lavoro di intelligenza artificiale generativa e quale percentuale di questi è attualmente accessibile?	I registri delle chiamate dei clienti e i dati sulle recensioni dei prodotti sono fondamentali. Attualmente, l'85% di questi tipi di dati è accessibile per i nostri progetti di intelligenza artificiale generativa.
Come garantite e misurate la qualità dei vostri dati?	Abbiamo implementato metriche sulla qualità dei dati, tra cui completezza, accuratezza, coerenza e tempestività. Utilizziamo strumenti automatizzati per valutare regolarmente queste metriche e disponiamo di un team dedicato per la pulizia e l'arricchimento dei dati.
Quale percentuale dei tuoi dati soddisfa i tuoi standard di qualità per l'uso dell'IA generativa?	Attualmente, il 78% dei nostri dati soddisfa i nostri standard di qualità. Puntiamo a raggiungere il 95% entro i prossimi 12 mesi attraverso migliori processi di pulizia dei dati.
Come pensate di creare fiducia tra i vostri stakeholder in merito all'utilizzo dei dati nell'IA generativa?	Stiamo implementando un comitato etico per l'IA, fornendo spiegazioni chiare sulle decisioni in materia di intelligenza artificiale e conducendo audit trimestrali sull'IA per garantire trasparenza ed equità.

Domanda	Example response
Quanto è completa la vostra documentazione sulle fonti di dati e sulla provenienza dei dati?	Disponiamo di un catalogo di dati dettagliato che include i metadati per tutte le nostre fonti di dati, tra cui origine, frequenza di aggiornamento e utilizzo. Utilizziamo strumenti di data lineage per monitorare il flusso e la trasformazione dei dati tra i nostri sistemi.
Come garantite la diversità dei set di dati per prevenire distorsioni nei modelli di intelligenza artificiale?	Riceviamo attivamente dati da diversi dati demografici e controlliamo regolarmente i nostri set di dati per individuare eventuali distorsioni rappresentazionali. Utilizziamo anche tecniche di generazione di dati sintetici per bilanciare le categorie sottorappresentate.
Qual è la frequenza di aggiornamento dei dati per i modelli di intelligenza artificiale generativa critici e come si determina questa frequenza?	I modelli critici vengono aggiornati settimanalmente. Questa frequenza è determinata dalle metriche prestazionali dei test A/B e miriamo a una riduzione non superiore al 2% tra un aggiornamento e l'altro.
Quante versioni di set di dati critici conservate e per quanto tempo?	Conserviamo le ultime cinque versioni di ogni set di dati critico, con un periodo di conservazione di 18 mesi per ogni versione.
Quanti team interfunzionali sono coinvolti nelle vostre iniziative di intelligenza artificiale generativa e hanno accesso ai vostri dati?	Abbiamo tre team interfunzionali. Ogni team include data scientist, esperti di settore, esperti di etica e analisti aziendali.
Quali politiche e pratiche di governance dei dati avete in atto?	Abbiamo un comitato interfunzionale per la governance dei dati che supervisiona le nostre politiche sui dati. Abbiamo implementato controlli degli accessi basati sui ruoli, schemi di classificazione dei dati e audit regolari per garantire la conformità al nostro quadro di governance.

Domanda	Example response
Quali misure avete adottato per garantire la privacy dei dati, ottenere il consenso adeguato e mantenere la riservatezza?	Abbiamo implementato un quadro completo sulla privacy dei dati in linea con GDPR e CCPA. Ciò include l'ottenimento del consenso esplicito per l'utilizzo dei dati, l'implementazione di tecniche di anonimizzazione dei dati e regolari valutazioni dell'impatto sulla privacy.
Quale percentuale dei vostri set di dati di formazione sull'intelligenza artificiale è stata verificata per individuare eventuali distorsioni nell'ultimo trimestre?	Il 70% dei nostri set di dati di formazione sull'intelligenza artificiale è stato verificato per rilevare eventuali distorsioni lo scorso trimestre . Stiamo implementando strumenti automatici di rilevamento dei pregiudizi per ottenere audit trimestrali al 100%.
Qual è la tua attuale capacità di elaborazione dei dati e quanto prevedi di averne bisogno per i futuri carichi di lavoro di intelligenza artificiale generativa?	La nostra capacità attuale è del 10% TB/day. We project needing 30 TB/day entro un anno e stiamo scalando la nostra infrastruttura per soddisfare questa domanda.
Qual è la tua strategia per bilanciare la privacy dei dati con le esigenze dei dati dei modelli di intelligenza artificiale generativa?	Stiamo implementando tecniche di anonimizzazione avanzate e generazione di dati sintetici . Il nostro obiettivo è aumentare i nostri dati utilizzabili per l'IA del 40%, riducendo al contempo i rischi per la privacy del 60% nel prossimo anno.
Quale percentuale dei tuoi set di dati di machine learning (ML) è etichettata con precisione e qual è il tuo tasso di precisione obiettivo?	Attualmente, l'85% dei nostri set di dati ML è etichettato con precisione. Puntiamo a un tasso di precisione del 95% entro il prossimo trimestre utilizzando tecniche di etichettatura sia umane che automatizzate.

Tradurre le informazioni di valutazione in risultati attuabili

Questa sezione fornisce un framework per analizzare le risposte al questionario e utilizzare tali informazioni per modellare l'architettura di destinazione e altri risultati chiave dell'iniziativa di modernizzazione dell'IA generativa. Questo framework colma il divario tra raccolta e implementazione dei dati e garantisce che la valutazione informi e guidi direttamente la strategia di modernizzazione.

Definizione dell'architettura di destinazione:

- Utilizza le risposte al questionario per orientare la selezione dei servizi cloud e la progettazione di pipeline di dati.
- Assicurati che la progettazione dell'architettura supporti la scalabilità e l'interoperabilità, come evidenziato nella guida.

Valutazione della fattibilità del cliente:

- Analizza le risposte al questionario relative all'infrastruttura, ai processi e alla cultura organizzativa attuali.
- Identifica le lacune e crea un piano per colmarle. Dai priorità alle lacune fondamentali per il successo di un MVP.

Usa case e allunga gli obiettivi:

- Estrai problemi aziendali specifici dalle risposte al questionario per definire obiettivi chiari per i casi d'uso.
- Stabilisci obiettivi ambiziosi in linea con la visione a lungo termine della tua organizzazione per la modernizzazione generativa dell'IA.

Stima dello sforzo:

- Utilizza i dati del questionario per stimare risorse, tempo e budget sia per l'MVP che per l'implementazione completa.
- Crea un approccio graduale che inizi con l'MVP e delinei le fasi successive.

Esigenze di abilitazione:

- Sulla base delle risposte al questionario, individua le lacune in termini di competenze e le esigenze di formazione.
- Sviluppa un piano di formazione che supporti sia le esigenze immediate di MVP che l'adozione dell'IA generativa a lungo termine.

Piano di implementazione:

- Crea una tabella di marcia completa che inizi con l'MVP e delinei i passaggi verso la completa modernizzazione dell'IA generativa.
- Definisci traguardi e risultati finali chiari per ogni fase dell'implementazione.

Passaggi pratici:

- Matrice di prioritizzazione: crea una matrice che mappa le risposte al questionario in base ai [sei risultati](#) per aiutare a stabilire le priorità delle funzionalità e degli sforzi.
- Approccio iterativo: progetta l'MVP in modo che sia la prima iterazione di una serie di versioni pianificate, in cui ogni versione si basa sull'architettura target completa.
- Allineamento degli stakeholder: utilizza i risultati del questionario per allineare gli stakeholder sull'ambito del programma MVP e sull'approccio graduale per il raggiungimento di tutti i risultati.
- Ciclo di feedback continuo: implementa meccanismi per raccogliere feedback dopo l'implementazione dell'MVP e utilizza le informazioni per perfezionare i piani per le fasi successive.
- Implementazione agile: adotta una metodologia agile che consenta la flessibilità necessaria per affrontare tutti i risultati nel tempo, a partire dai risultati più critici dell'MVP.

Passaggi successivi

Dopo aver completato la valutazione generativa del carico di lavoro AI, segui questi passaggi:

1. Fornisci un'architettura mirata dettagliata

- **Obiettivo:** l'architetto della soluzione crea un'architettura target completa in linea con gli obiettivi dell'organizzazione e i risultati della valutazione.
- **Componenti:** questa architettura include la progettazione dell'inserimento dei dati, dei punti di integrazione e dell'interoperabilità del sistema per garantire scalabilità, affidabilità e ottimizzazione delle prestazioni.

2. Spiega in che modo specifico si adatta al caso d'uso Servizi AWS

- **Mappatura dei servizi:** identifica e mappa le caratteristiche specifiche Servizi AWS che meglio si adattano ai casi d'uso identificati.
- **Vantaggi:** evidenzia come questi servizi rispondono a esigenze aziendali specifiche, migliorano l'efficienza e forniscono scalabilità.

3. Fornisci soluzioni alternative opzionali con vantaggi e svantaggi

- **Alternative:** presentare soluzioni alternative che potrebbero soddisfare anche i requisiti dell'organizzazione.
- **Analisi:** offri un'analisi dettagliata dei vantaggi e degli svantaggi di ciascuna alternativa considerando fattori quali costi, complessità e allineamento con gli obiettivi aziendali.

4. Fornisci una stima dettagliata dei prezzi di Servizi AWS

- **Analisi dei costi:** Fornisci una stima dettagliata dei costi proposta Servizi AWS, inclusi potenziali scenari di utilizzo e modelli di prezzo.
- **Allineamento del budget:** assicurati che i costi siano in linea con i vincoli di bilancio dell'organizzazione e forniscano una chiara comprensione delle implicazioni finanziarie.

5. Ottieni feedback sull'architettura proposta

- **Coinvolgimento delle parti interessate:** interagisci con le parti interessate per presentare l'architettura proposta e raccogliere feedback.
- **Miglioramento iterativo:** utilizza il feedback per affinare e migliorare la soluzione e confermare che soddisfi le esigenze e le aspettative di tutte le parti interessate.

Domande frequenti

Qual è l'obiettivo principale della valutazione generativa del carico di lavoro dell'IA?

L'obiettivo principale della valutazione è valutare la preparazione di un'organizzazione a modernizzare i carichi di lavoro generativi di intelligenza artificiale, identificare i casi d'uso e sviluppare un'architettura di soluzioni mirata. Mira a definire i requisiti di modernizzazione, determinare l'ambito di implementazione e prepararsi per una modernizzazione dell'IA generativa di successo.

Chi dovrebbe utilizzare questa valutazione?

Questa valutazione è rivolta agli architetti di soluzioni, agli architetti aziendali e agli architetti delle applicazioni che desiderano valutare gli aspetti tecnici della modernizzazione generativa dell'IA. È utile anche per i responsabili dei programmi e i responsabili del personale per valutare le esigenze complessive di preparazione, allocazione delle risorse e abilitazione.

Quali sono i componenti chiave valutati nella valutazione?

La valutazione riguarda la fattibilità generale, il caso d'uso, l'architettura, lo storage, le normative e la conformità, l'integrazione, i test, l'automazione dell'implementazione e la strategia dei dati. Questi componenti sono fondamentali per determinare la preparazione tecnica e organizzativa per l'adozione della modernizzazione generativa dell'IA.

In che modo la valutazione aiuta a definire l'architettura di destinazione?

La valutazione fornisce un approccio strutturato per valutare i sistemi attuali e identificare i miglioramenti. Ti aiuta a selezionare le tecnologie appropriate e a progettare architetture scalabili in linea con gli obiettivi aziendali e i requisiti dei casi d'uso.

Quali sono i vantaggi di condurre una valutazione generativa del carico di lavoro AI?

I vantaggi includono una maggiore efficienza, un migliore processo decisionale, la garanzia della conformità, la promozione dell'innovazione e la preparazione alla scalabilità. La valutazione stabilisce un approccio strategico alla modernizzazione generativa dell'IA e massimizza i potenziali benefici mitigando i rischi.

In che modo le organizzazioni possono garantire un'implementazione di successo dopo la valutazione?

Le organizzazioni dovrebbero sviluppare un piano di implementazione chiaro che includa traguardi definiti, coinvolgere tempestivamente le parti interessate e adottare un approccio iterativo. Le migliori pratiche consigliate sono anche la creazione di un Centro di eccellenza (CoE) e l'attenzione allo sviluppo dei talenti.

Quali sfide potrebbero affrontare le organizzazioni durante la valutazione?

Le sfide potrebbero includere la resistenza al cambiamento, i problemi di qualità dei dati e le complessità di conformità. Affrontare queste sfide richiede la promozione di una cultura dell'innovazione, la garanzia della disponibilità dei dati e l'implementazione di solide misure di sicurezza.

In che modo la valutazione risponde ai requisiti normativi e di conformità?

La valutazione valuta le attuali misure di conformità e identifica le lacune. Garantisce che le soluzioni mirate rispettino le normative pertinenti e le leggi sulla privacy dei dati e incorporino le migliori pratiche di sicurezza per proteggere le informazioni sensibili.

Che ruolo gioca il coinvolgimento delle parti interessate nel processo di valutazione?

Il coinvolgimento delle parti interessate è fondamentale per ottenere il consenso, allineare le iniziative di modernizzazione agli obiettivi aziendali e garantire un'implementazione di successo. Il coinvolgimento precoce e la comunicazione chiara dei vantaggi sono fondamentali per superare le resistenze e promuovere il supporto.

In che modo le organizzazioni possono misurare il successo delle loro iniziative di modernizzazione generativa dell'IA dopo la valutazione?

Il successo può essere misurato utilizzando indicatori chiave di performance (KPIs) in linea con gli obiettivi aziendali. Il monitoraggio e la valutazione regolari di queste metriche aiutano a guidare il processo decisionale e a dimostrare il valore della modernizzazione generativa dell'IA agli stakeholder.

In che modo l'approccio di valutazione differisce per le organizzazioni di diverse dimensioni (piccole, medie o imprese) o settori?

Piccole organizzazioni:

- Potrebbero disporre di risorse e competenze limitate per valutazioni complete
- Probabilmente si concentrerà su casi d'uso specifici ad alto impatto anziché sull'adozione a livello aziendale
- Potrebbe affidarsi maggiormente a strumenti e servizi di terze parti per la valutazione
- Il processo di valutazione potrebbe essere meno formale e più agile

Organizzazioni di medie dimensioni:

- Spesso dispongono di team IT o di gestione dati dedicati, ma potrebbero non disporre di competenze specializzate in intelligenza artificiale

- Potrebbe adottare un approccio graduale, a partire da progetti pilota nei reparti chiave
- Necessità di bilanciare l'innovazione con i sistemi e i processi esistenti
- La valutazione probabilmente coinvolge team interfunzionali

Organizzazioni aziendali:

- In genere dispongono di AI/ML team dedicati e di più risorse per una valutazione completa
- È necessario prendere in considerazione integrazioni complesse con i sistemi aziendali esistenti
- Potrebbe avere requisiti normativi specifici del settore da tenere in considerazione
- La valutazione spesso implica processi di governance formali

Risorse

- [AI generativa attiva AWS](#)
- [AWS offre nuove guide sull'intelligenza artificiale, l'apprendimento automatico e l'intelligenza artificiale generativa per pianificare la tua strategia di intelligenza artificiale](#) (AWS post sul blog)
- [Le migliori pratiche per creare applicazioni di intelligenza artificiale generativa su AWS](#)(AWS post sul blog)
- [Generative AI Application Builder su AWS](#)(Solutions Library)AWS
- [Funzionalità di intelligenza artificiale generativa](#) (AWS Security Reference Architecture)
- [AWS framework generativo di best practice per l'intelligenza artificiale \(Guida per AWS Audit Manager l'utente\)](#)
- [Scelta di un servizio di intelligenza artificiale generativa \(guidaAWS decisionale\)](#)
- [Che cos'è Amazon Bedrock?](#) (Guida per l'utente di Amazon Bedrock)
- [Cos'è Amazon SageMaker AI?](#)(Guida per sviluppatori Amazon SageMaker AI)

Cronologia dei documenti

La tabella seguente descrive le modifiche significative apportate a questa guida. Per ricevere notifiche sugli aggiornamenti futuri, puoi abbonarti a un [feed RSS](#).

Modifica	Descrizione	Data
Pubblicazione iniziale	—	6 novembre 2024

Le traduzioni sono generate tramite traduzione automatica. In caso di conflitto tra il contenuto di una traduzione e la versione originale in Inglese, quest'ultima prevarrà.