Una risposta alle 4 domande più importanti sulla sicurezza dell'IA generativa

Adozione rapida dell'IA generativa con garanzia di sicurezza, privacy e conformità

Questo eBook si rivolge ai leader aziendali, in particolare ai responsabili delle decisioni IT e dei team addetti alla sicurezza che stanno pianificando o cercando di capire come integrare l'IA generativa nelle rispettive organizzazioni senza compromettere la sicurezza.

Indice

INTRODUZIONE

Pronti, partenza... e via verso l'IA generativa: come adottarla in modo rapido e sicuro

La corsa all'IA generativa è iniziata. Le aziende stanno spingendo sull'acceleratore per reinventare le applicazioni e le esperienze dei clienti, attratte da miglioramenti potenzialmente enormi in termini di produttività ed esperienza.

L'era dell'intelligenza artificiale (IA) generativa è appena agli inizi, ma le organizzazioni stanno già ottenendo vantaggi tangibili in quasi tutti i reparti aziendali. Tuttavia i professionisti della sicurezza consigliano cautela. Privacy dei dati, distorsioni nei modelli, creazione di contenuti dannosi (come i deepfake) e rischi di input dannosi nei modelli: questi sono alcuni dei motivi per cui prediligono un approccio prudente rispetto all'adozione dell'IA generativa.

È indispensabile che le organizzazioni si avvicinino all'IA generativa con una strategia ben chiara per proteggere i dati, gli utenti e la reputazione, consentendo al contempo una rapida adozione e migliorando l'esperienza dei clienti.

Si tratta senz'altro di una sfida dalle molte sfaccettature, ma le organizzazioni non devono dimenticare che rimangono sempre valide le best practice standard per l'IA, il machine learning (ML), la protezione dei dati e la sicurezza dei carichi di lavoro cloud. In effetti, la tua organizzazione potrebbe essere più preparata a un'adozione sicura dell'IA generativa di quanto tu creda.

L'implementazione tempestiva di protezioni adeguate per i carichi di lavoro dell'IA generativa andrà a beneficio dell'innovazione in tutta l'organizzazione, dando ai team la sicurezza necessaria per perseguire grandi idee e la libertà di concentrarsi sulla crescita del business.

Questo eBook affronta le 4 principali domande da porsi quando si dà il via al viaggio verso una maggiore sicurezza dei carichi di lavoro dell'IA generativa.

1. **Cosa è necessario proteggere?**
2. **Come è possibile risolvere i problemi di conformità?**
3. **Come è possibile garantire che i modelli funzionino nel modo previsto?**
4. **Da dove iniziare?**

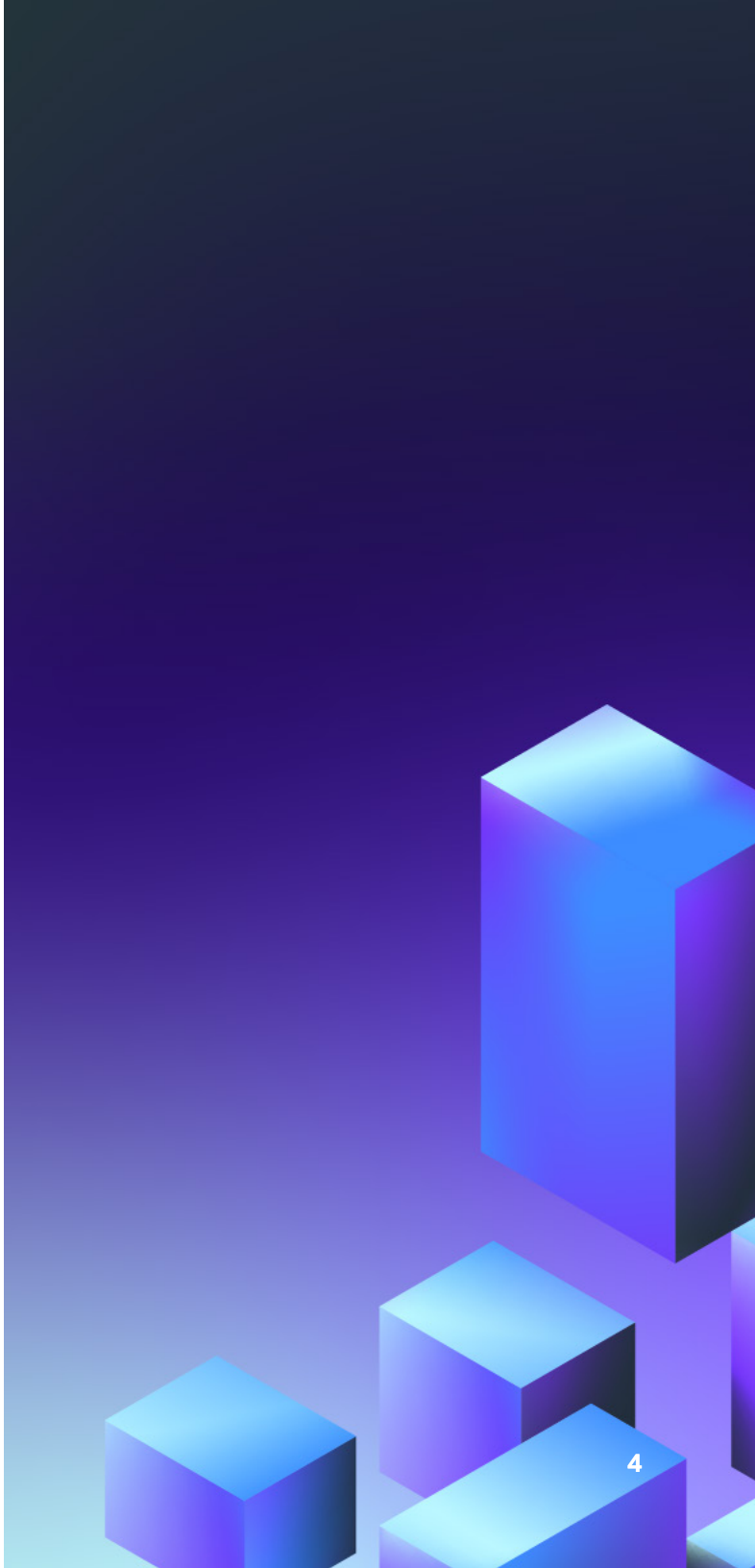
REQUISITI DI PROTEZIONE DEI DATI

Domanda 1:

Cosa è necessario proteggere?

Prima di poter sviluppare e implementare in sicurezza le applicazioni di IA generativa, è importante capire esattamente cosa occorre proteggere. Può essere utile raggruppare gli elementi in tre categorie:

- **Protezione dei carichi di lavoro cloud**
- **Protezione dei dati**
- **Protezione delle applicazioni di IA generativa**



Protezione dei carichi di lavoro cloud

Per utilizzare l'IA generativa senza compromettere gli obiettivi di sicurezza e privacy occorre iniziare proteggendo l'infrastruttura, i servizi e le configurazioni cloud nel loro complesso. A questo scopo, è necessario prima di tutto che tu distingua le tue responsabilità in materia di sicurezza da quelle del tuo provider cloud.

Per una guida in questo ambito, i clienti di Amazon Web Services (AWS) possono fare riferimento al **Modello di responsabilità condivisa**, nel quale si spiega che, in generale, AWS è responsabile di azionare, gestire e controllare l'infrastruttura su cui vengono eseguiti tutti i servizi offerti nel cloud AWS. Questa parte è denominata "Sicurezza *del* cloud".

I clienti di AWS si assumono invece la responsabilità della gestione del sistema operativo guest (inclusi gli aggiornamenti e le patch di sicurezza), di altro software applicativo, nonché della configurazione del firewall del gruppo di sicurezza fornito da AWS. L'ambito e i compiti specifici che i clienti devono svolgere variano in funzione dei servizi AWS che scelgono di utilizzare. Questa parte è denominata "Sicurezza *nel* cloud".

Sebbene la popolarità dell'IA generativa sia nuova, un utile punto di partenza è rappresentato dalle tradizionali best practice di sicurezza, incluse quelle basilari per l'igiene informatica in materia di:

- Identity and Access Management (IAM)
- Rilevamento e risposta
- Protezione dell'infrastruttura
- Protezione dei dati
- Sicurezza delle applicazioni

Protezione dei dati

Il passaggio successivo consiste nell'impegnarsi per garantire la sicurezza e la privacy dei dati utilizzati dalle applicazioni di IA generativa. Alcuni esempi di tali dati sono: informazioni proprietarie, proprietà intellettuale di valore e informazioni di identificazione personale (PII).

Le applicazioni di IA generativa si basano su modelli di fondazione (FM), che vengono addestrati su grandi quantità di dati. Gli FM analizzano i dati per identificare dei modelli e imparare a generare nuovi contenuti simili. Per creare applicazioni di IA generativa che soddisfino i requisiti aziendali specifici, in genere è necessario personalizzare un FM esistente addestrandolo sui dati dell'organizzazione.

Per favorire la protezione di questi dati, dovrai valutare l'attuazione di controlli sulla privacy dei dati e best practice per le policy IAM.

Quando personalizzi un FM, assicurati che i tuoi team lavorino con una versione del modello archiviata in modo sicuro e non utilizzata per migliorare l'FM stesso. Attraverso la configurazione di una soluzione dedicata a tenant singolo in **Amazon Bedrock**, il servizio può collegare le proprie istanze di inferenza al tuo **Amazon Virtual Private Cloud** (Amazon VPC) per poter eseguire operazioni di lettura e scrittura in **Amazon Simple Storage Service** (Amazon S3).

Una gestione IAM efficace consente di verificare che le persone e le macchine giuste abbiano accesso alle risorse giuste nelle giuste condizioni. Il **Framework AWS Well-Architected** descrive i principi di progettazione e le best practice architetturali per agevolare la gestione delle identità. Questa risorsa è uno strumento utile per sviluppare policy IAM e risolvere altri problemi di sicurezza, come il rilevamento delle minacce e la sicurezza della rete.

Protezione delle applicazioni di IA generativa

Per rendere sicura l'IA generativa a livello di applicazione, è necessario identificare, classificare, correggere e mitigare continuamente i rischi. Un primo passo consiste nell'implementazione delle best practice esistenti per proteggere gli ambienti e i dati.

Su questa base, occorre poi considerare come spostare la sicurezza a una fase precedente del processo di sviluppo. Tutto questo può semplificare il tuo lavoro e consentire ai team di sviluppo di innovare più velocemente e con maggiore libertà, evitando al contempo di trasformare la sicurezza in un collo di bottiglia.

Successivamente, è necessario riflettere su come proteggere i tre componenti critici di qualsiasi applicazione di IA: gli input, gli output e il modello.

Protezione degli input

Inizia esaminando i dati che entrano nel sistema di IA. Per ridurre il rischio di attacchi all'integrità, come manomissione, spoofing o iniezione di prompt, gli utenti non devono avere accesso diretto all'FM senza filtraggio degli input. Queste tecniche di attacco aggirano i controlli o fanno un uso illecito del modello. Altre strategie da considerare per la protezione degli input sono l'automazione della qualità dei dati, il monitoraggio continuo e la modellazione delle minacce.

Protezione degli output

I rischi legati agli output delle applicazioni di IA generativa includono la divulgazione di informazioni, gli incidenti che coinvolgono la proprietà intellettuale e l'uso improprio o illecito del modello. Tutte queste situazioni possono danneggiare la reputazione dell'organizzazione. Quando sviluppi il tuo modello di minaccia, considera il footprint delle informazioni e il contesto di utilizzo, quindi includi il rilevamento e il monitoraggio dei comportamenti complessi.

Protezione del modello

L'ultimo passaggio prevede l'analisi dei modi in cui gli antagonisti possono tentare di rimuovere dati dal modello o dai componenti ad esso associati. I rischi includono false rappresentazioni del mondo reale o dei dati nel modello e danni all'integrità o alla disponibilità del modello. Occorre quindi modellare le minacce in base agli obiettivi aziendali e implementare il monitoraggio degli scenari di minaccia.

Domanda 2:

Come è possibile risolvere i problemi di conformità?

Riducendo i rischi legati alla progettazione e allo sviluppo delle applicazioni di IA generativa, la tua organizzazione può promuovere la fiducia di partner e clienti, tenere alta la reputazione del marchio e continuare a soddisfare i requisiti di conformità.

La regolamentazione legislativa delle applicazioni di IA generativa è tuttora in fase iniziale e non c'è ancora consenso sulle best practice. Pertanto, orientarsi nel labirinto di standard e controlli in conflitto tra le diverse giurisdizioni rappresenta una sfida complessa e continua.

Collabora con i tuoi consulenti legali ed esperti di privacy per valutare i requisiti e le implicazioni della creazione di un'applicazione di IA generativa. Ad esempio, potresti dover verificare i tuoi diritti legali in relazione all'utilizzo di specifici dati e modelli, oltre a determinare l'applicabilità sia delle leggi in materia di privacy, biometria e antidiscriminazione sia di altre normative specifiche per singoli casi d'uso.

Non dimenticare che stati, province e Paesi diversi possono avere requisiti legali differenti, e che in tutto il mondo vengono proposte sempre nuove normative sull'IA. Riesamina queste considerazioni nelle fasi operative e di implementazione future.

Collaborare con colleghi, esperti di IA e organizzazioni governative può anche aiutarti a mantenere la conformità, dimostrando al contempo ai clienti che prendi in seria considerazione gli standard legali ed etici correlati all'IA. Di recente Amazon si è unita alla Casa Bianca e a sei aziende leader nel campo dell'IA nell'**assunzione di impegni volontari per uno sviluppo responsabile e sicuro dell'IA**, dimostrando il valore di tali iniziative e gettando le basi per una collaborazione futura.

Rischi intrinseci dell'intelligenza artificiale

In modo analogo alle altre soluzioni che utilizzano il machine learning, le applicazioni di IA generativa presentano rischi che vanno oltre quelli correlati ai software tradizionali. Per creare e implementare in sicurezza applicazioni di IA generativa, è necessario valutare e sviluppare strategie per mitigare questi rischi, che includono:

- Output distorti, non veritieri, fuorvianti, dannosi o offensivi
- Complessità e costi su vasta scala
- Set di dati che diventano troppo grandi, obsoleti o scollegati dal contesto previsto
- Preoccupazioni relative a una maggiore opacità e riproducibilità
- Procedure e standard di test non sufficientemente sviluppati

Nella sezione successiva vedremo alcune strategie generali volte a mitigare alcuni di questi rischi, oltre alle best practice per definire l'impatto a livello professionale, organizzativo e sociale delle applicazioni di IA generativa.

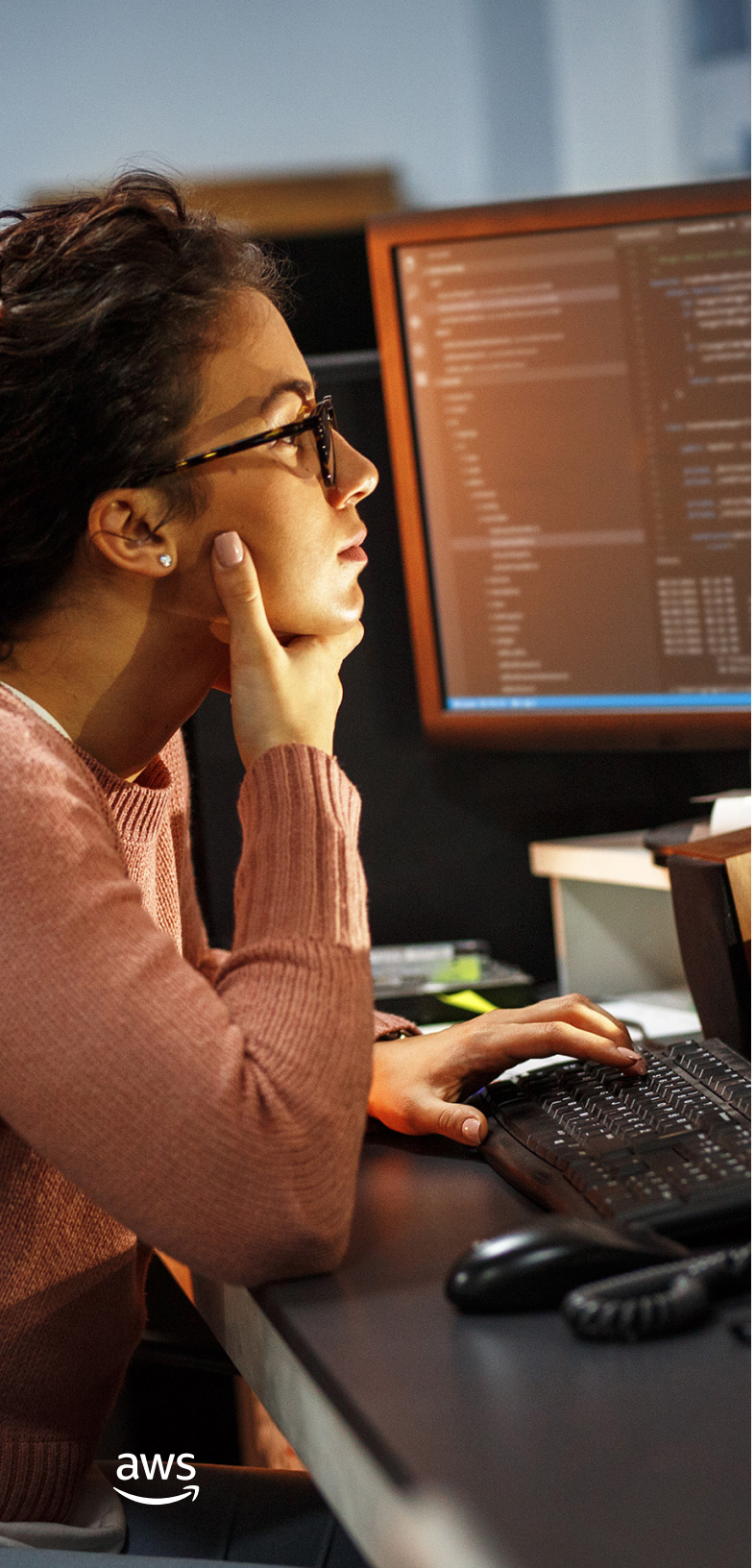
Domanda 3:

Come è possibile garantire che i modelli funzionino nel modo previsto?

Garantire l'uso responsabile dell'IA generativa si sta rivelando un'attività aziendale essenziale, oltre che un fattore fondamentale per l'innovazione continua.

Gli FM si addestrano su enormi set di dati, eseguendo analisi complesse che li aiutano a capire come generare contenuti simili. Sebbene molti FM ottengano risultati straordinari, il ben noto principio "Garbage In, Garbage Out" (GIGO) rimane sempre valido. Se un FM riceve dati imprecisi, incompleti o distorti, i relativi output potrebbero presentare difetti simili.

Se imperfetti, i dati aprono la strada a situazioni di uso improprio, azioni dannose e altri rischi. Man mano che un'applicazione di IA generativa si espande in termini di utenti, ambito e funzionalità, il potenziale impatto di questi problemi aumenta.



Sviluppo di un'IA responsabile

Impegnarsi ad adottare una strategia di IA responsabile contribuisce a mitigare i rischi. Le dimensioni dell'IA responsabile includono: spiegabilità, correttezza, governance, privacy, sicurezza, robustezza e trasparenza. Includono anche la comprensione dei modi in cui culture e situazioni demografiche diverse vengono visualizzate, trattate e influenzate dall'applicazione.

L'ideale è iniziare a considerare l'IA responsabile fin dall'inizio del viaggio verso l'IA generativa, per poi continuare a tenerne conto durante l'intero ciclo di vita dell'applicazione come una parte fondamentale della vision. Basta cominciare con azioni relativamente piccole e semplici. Quindi è possibile dimensionare l'impatto dell'IA responsabile sulla progettazione, sullo sviluppo e sulle operazioni nel corso del tempo.

Mentre elabori le policy per la governance e l'IA responsabile, considera il modo in cui la tua applicazione di IA generativa influirà sugli utenti, sui clienti, sui dipendenti e sulla società. Assicurati di prendere in considerazione fattori come l'equità algoritmica, la rappresentazione diversificata e inclusiva, il rilevamento dei pregiudizi.

Lotta alla tossicità

La tossicità nei modelli linguistici di grandi dimensioni (LLM) si riferisce alla generazione di testi volgari, irrispettosi o irragionevoli. Esistono molte strategie per contrastare la tossicità e garantire la correttezza nelle applicazioni di IA generativa. Ad esempio, potresti identificare i testi offensivi o contenenti pregiudizi e rimuoverli dai dati di addestramento. Oppure puoi eseguire test di correttezza più circoscritti incentrati sul caso d'uso specifico della tua applicazione, sui destinatari oppure sui prompt e le query che è più probabile che l'applicazione riceva.

Puoi anche addestrare modelli di controllo su set di dati annotati che identificano diversi tipi e gradi di tossicità. Tutto questo può aiutare l'FM a imparare a rilevare e filtrare i contenuti indesiderati nei dati di addestramento, nei prompt di input e negli output generati in modo automatico.

Protezione della privacy

È possibile attuare svariate misure per prevenire l'esposizione indesiderata di informazioni sensibili, segreti commerciali e proprietà intellettuale durante l'utilizzo di applicazioni di IA generativa.

Uno dei metodi con cui è possibile risolvere i problemi di privacy consiste nell'eliminazione del modello, che comporta la rimozione dei dati utilizzati in modo improprio subito dopo la relativa identificazione, contribuendo così ad annullare l'impatto di tali dati su qualsiasi componente dell'FM.

Un altro approccio possibile consiste nella realizzazione di partizioni: i dati di addestramento vengono suddivisi in porzioni più piccole, poi utilizzate per l'addestramento di sottomodelli separati, che alla fine si uniscono per formare l'FM complessivo. Questa pratica può semplificare notevolmente la correzione degli FM che contengono o rischiano di esporre informazioni private. Invece di riaddestrare l'intero modello, è sufficiente rimuovere i dati indesiderati o utilizzati in modo improprio dalla relativa partizione e quindi riaddestrare solo quel sottomodello.

Altri approcci efficaci includono operazioni di filtraggio e blocco. Questi metodi prevedono il confronto esplicito tra informazioni protette e contenuti generati, prima della visualizzazione da parte dell'utente. Se informazioni protette e contenuti generati sono troppo simili, i contenuti vengono eliminati o sostituiti per evitare eventuali esposizioni. Può essere utile anche limitare il numero di volte in cui uno specifico elemento di contenuto compare nei dati di addestramento.

Miglioramento di spiegabilità e verificabilità

Per supportare ulteriormente l'IA responsabile, è opportuno prendere in considerazione la necessità di spiegare la metodologia e i fattori chiave che influenzano l'output delle applicazioni. Un altro elemento importante per l'IA responsabile è la verificabilità. È possibile implementare meccanismi per monitorare ed esaminare lo sviluppo e il funzionamento di un'applicazione di IA generativa. In questo modo è più facile tenere traccia delle cause radice di eventuali problemi e soddisfare i requisiti di governance.

Valuta anche la possibilità di documentare gli input e le decisioni di progettazione rilevanti durante l'intero ciclo di vita dello sviluppo. La creazione di un record tracciabile può aiutare i team interni o esterni a valutare lo sviluppo e il funzionamento dell'applicazione di IA generativa.

Nessun compromesso in termini di responsabilità

Infine, rifletti su come contribuirai a garantire la continua osservanza delle policy per l'IA responsabile. Assicurati di mettere in pratica le lezioni e le esperienze acquisite per far evolvere le procedure aziendali in materia di sicurezza e privacy. Informa periodicamente tutti i dipendenti della tua organizzazione sugli obblighi in termini di sicurezza e protezione per l'implementazione dell'IA generativa. Promuovi una cultura dell'IA responsabile, usa gli strumenti giusti per agevolare il monitoraggio delle prestazioni dei modelli e la comunicazione dei rischi, consenti ai tuoi team di ispezionare i modelli e i relativi componenti in caso di necessità. Esegui e ripeti i test e, in caso di dubbi, ripetili ancora.

Domanda 4:

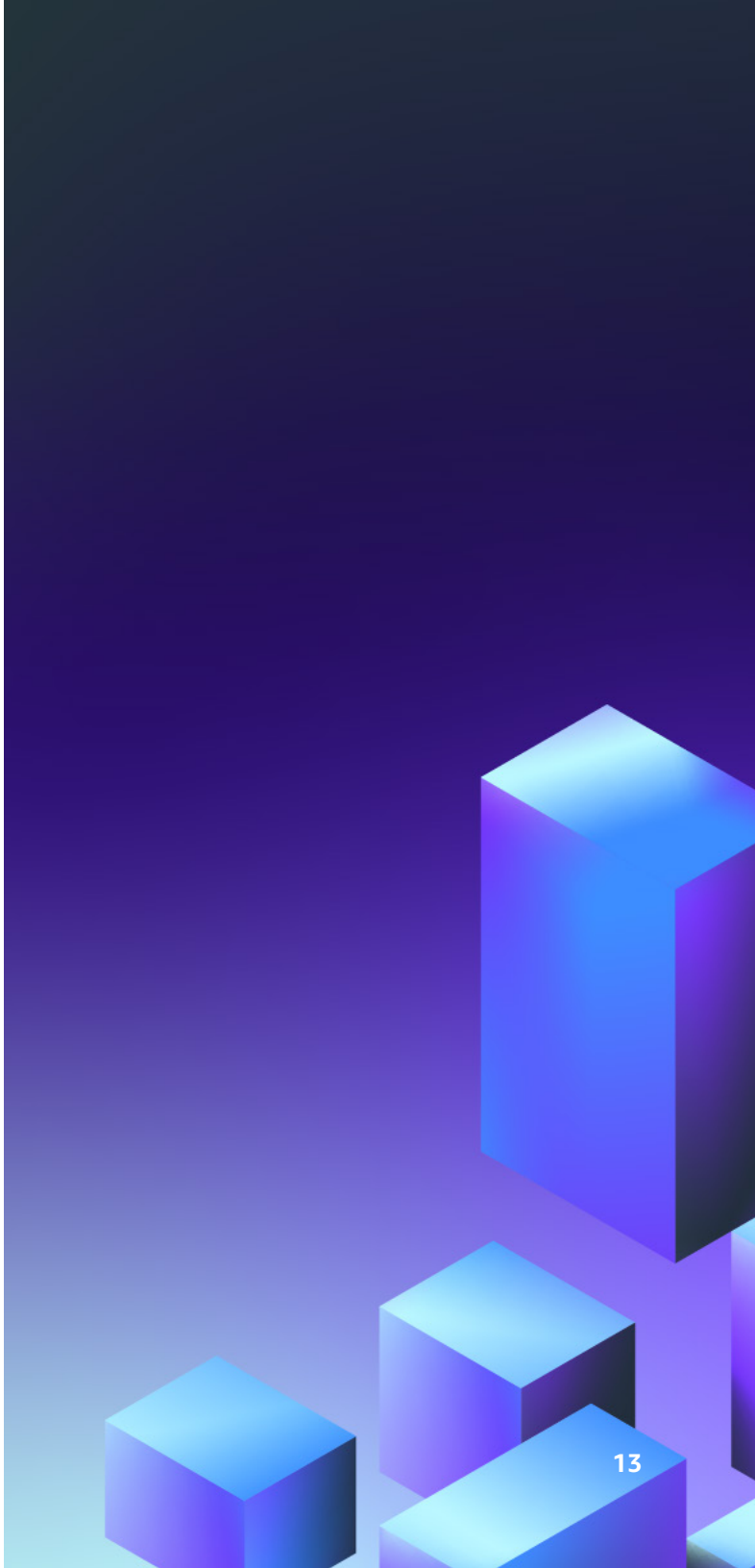
Da dove iniziare?

Proteggere le applicazioni di IA generativa non è un'impresa facile e non esiste un set universale di azioni da intraprendere per raggiungere l'obiettivo. Tuttavia, collaborando con il fornitore giusto e implementando gli strumenti adatti, la strada verso il successo diventa molto più chiara.

Ad esempio, l'uso di **Amazon Bedrock** può semplificare e accelerare notevolmente il percorso verso lo sviluppo di applicazioni di IA generativa sicure. Amazon Bedrock è un servizio completamente gestito che mette a disposizione tramite un'API gli FM di Amazon e delle principali startup nel settore dell'IA.

Quando personalizzi un modello con Amazon Bedrock, il servizio può ottimizzare il modello per una determinata attività senza che il tuo team debba annotare grandi volumi di dati. Quindi, Amazon Bedrock crea una copia separata dell'FM di base accessibile solo a te e addestra questa copia privata del modello. Nessuno dei tuoi dati viene utilizzato per addestrare i modelli di base originali, il che aiuta a mantenere privati e sicuri i tuoi dati proprietari.

Puoi anche configurare le impostazioni di **Amazon VPC** in modo da accedere alle API di Amazon Bedrock e fornire i dati per l'ottimizzazione dei modelli in totale sicurezza. I tuoi dati vengono sempre crittografati mentre sono in transito e inattivi tramite chiavi gestite dal servizio. Inoltre, con **AWS PrivateLink**, puoi trasferire i tuoi dati dal cloud AWS ad Amazon Bedrock esclusivamente tramite la rete AWS e mai tramite reti Internet pubbliche.



Ottimizzazione della privacy con AWS

Non importa se crei le tue applicazioni di IA generativa con Amazon Bedrock, un altro servizio (come **Amazon SageMaker**) o strumenti personalizzati: quando esegui e gestisci le applicazioni su AWS, puoi usufruire di protezioni e controlli della privacy leader del settore.

AWS supporta 143 standard di sicurezza e certificazioni di conformità, contribuendo quindi a soddisfare i requisiti dei nostri clienti in tutto il mondo. È possibile crittografare tutti i dati mentre sono inattivi utilizzando specifiche chiavi **del sistema AWS di gestione delle chiavi** (AWS KMS), ottenendo visibilità e controllo completi sulle modalità di archiviazione e accesso per i dati e gli FM.

CONCLUSIONI

Fasi successive

AWS si impegna a supportarti nella creazione di applicazioni di IA generativa in grado di far crescere il tuo business, aiutandoti al contempo a raggiungere i tuoi obiettivi in termini di sicurezza, privacy e conformità.

È nostra ferma convinzione che le applicazioni di IA generativa possano essere progettate, sviluppate e gestite in modo sicuro. In più, riconosciamo la fondatezza delle preoccupazioni relative a queste tecnologie in termini di sicurezza e privacy. L'**IA generativa pone nuove sfide** per quanto riguarda la definizione, la misurazione e la mitigazione delle problematiche correlate a privacy dei dati, proprietà intellettuale, regolamenti legislativi, equità e trasparenza.

Introduzione di nuovi prodotti, crescenti complessità e dimensioni delle soluzioni, nuovi parametri di addestramento e set di dati in continua crescita: la sicurezza dell'IA generativa diventerà un fattore ancora più essenziale nel prossimo futuro. Sviluppando adesso una strategia di sicurezza efficace e completa per i carichi di lavoro dell'IA generativa, è possibile massimizzare il vantaggio competitivo e prepararsi ad affrontare un futuro in rapido avvicinamento.

C'è una buona notizia: i controlli di base necessari per progettare, sviluppare e gestire le applicazioni di IA generativa in modo sicuro esistono da anni e sono allineati a principi di sicurezza del cloud affidabili e comprovati, come ad esempio quelli del **Framework AWS Well-Architected**.

Prendendo in esame le pratiche descritte in questo eBook, hai già fatto il primo passo verso la protezione dei tuoi carichi di lavoro dell'IA generativa.

Ora fai il passo successivo con AWS. Possiamo fornirti le informazioni approfondite e le istruzioni specifiche che ti servono per restare sempre al passo con le ultime novità, riflettere sulle specifiche sfide che ti trovi ad affrontare e sfruttare tutti i vantaggi dell'IA generativa. Tutto continuando a proteggere i tuoi dati, i tuoi clienti e il tuo business.

Scopri di più sull'IA generativa su AWS ›

Inizia subito con Amazon Bedrock ›

Crea e personalizza FM su Amazon SageMaker ›

Aumenta la sicurezza nel cloud con AWS ›

Passa dalla teoria alla pratica nell'ambito dell'IA responsabile ›