Valutazione generativa del carico di lavoro AI

AWS Guida prescrittiva

AWS Guida prescrittiva: Valutazione generativa del carico di lavoro AI

Table of Contents

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 percento 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:

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

Scarica il questionario

Prontezza

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

Domanda	Example response
	unità aziendali per diventare esperti di intellige nza artificiale e identificare i casi d'uso dell'IA generativa.
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 regolarme nte 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 intellige nza artificiale generativa?	La nostra architettura IT si basa su microserv izi e consente APIs l'integrazione flessibile di nuove tecnologie di intelligenza artificia le generativa. Questi sistemi sono standardi zzati su formati e protocolli di dati comuni per garantire l'interoperabilità.

Domanda	Example response
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.

Domanda	Example response
Con che frequenza è necessario aggiornare o aggiornare i dati provenienti da queste fonti?	Database transazionali: aggiornamenti quasi in tempo reale; archivi di documenti: aggiornam enti 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 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 correttez za 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.

Domanda	Example response
Quali sono i casi d'uso o gli scenari principal i 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.
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 pregiudiz i 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.

Domanda	Example response
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.
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.

Domanda	Example response
In che modo il modello di intelligenza artificiale generativa verrà aggiornato o riqualificato nel tempo?	Attraverso l'apprendimento continuo, la riqualifi cazione 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.
Quali sono i requisiti di latenza per l'applica zione 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 principal mente per archiviare i set di dati di addestram ento 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

Domanda	Example response
	per ottimizzare il caricamento dei dati durante l' TFRecord allenamento.
Come sono organizzati i set di dati di formazion e: 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.
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 disponibi li 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'archivi azione 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 conformit à 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 intellige nza artificiale generativa ha adottato la tua organizzazione?	Abbiamo implementato le nostre linee guida sull'intelligenza artificiale responsabile. Tutti i progetti di intelligenza artificiale generativa

Domanda	Example response
	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.
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 registraz ione 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.

Domanda	Example response
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).
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.

Domanda	Example response
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 compatibi lità. 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'inter no della nostra attuale struttura di data lake.
Come intendete sfruttare i componenti modulari dei framework, ad esempio per la prototipa zione 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 intenzion e di valutare e integrare i principali aggiornam enti 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

Domanda	Example response
	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 artificia le generativa?	Precisione, richiamo, punteggio F1, perplessità, valutazione umana e così via.
Come verranno identificati e gestiti i casi limite e i casi isolati?	Utilizzando una suite di test completa, valutazio ne umana, test antagonistici e così via.
Come verificherete i potenziali pregiudizi nel modello di intelligenza artificiale generativa?	Utilizzando l'analisi della parità demografi ca, 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 equalizza te, 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 automatic i 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.

Domanda	Example response
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 riqualificato il modello di intelligenza artificiale generativa non appena saranno disponibili nuovi dati?	Attraverso la riqualificazione 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.

Domanda	Example response
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 modelli e Amazon Bedrock per i modelli di base.
Quali tecnologie di containerizzazione state considerando per i carichi di lavoro di intellige nza 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 infrastru ttura Kubernetes per un'implementazione senza interruzioni in tutti gli ambienti.

Domanda	Example response
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 trasferim ento efficiente dei dati tra l'archiviazione locale e AWS, con processi di sincronizzazione automatizzati pianificati in base ai nostri cicli di formazione.
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 artificia le generativa e quale percentuale di questi è attualmente accessibile?	I registri delle chiamate dei clienti e i dati sulle recensioni dei prodotti sono fondament ali. 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.

Domanda	Example response
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.
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.

Domanda	Example response
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.
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.

Domanda	Example response
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 anonimizz azione 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 team AI/ML 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

AWS Glossario delle linee guida prescrittive

I seguenti sono termini di uso comune nelle strategie, nelle guide e nei modelli forniti da AWS Prescriptive Guidance. Per suggerire voci, utilizza il link Fornisci feedback alla fine del glossario.

Numeri

7 R

Sette strategie di migrazione comuni per trasferire le applicazioni sul cloud. Queste strategie si basano sulle 5 R identificate da Gartner nel 2011 e sono le seguenti:

- Rifattorizzare/riprogettare: trasferisci un'applicazione e modifica la sua architettura sfruttando appieno le funzionalità native del cloud per migliorare l'agilità, le prestazioni e la scalabilità. Ciò comporta in genere la portabilità del sistema operativo e del database. Esempio: migra il tuo database Oracle locale all'edizione compatibile con Amazon Aurora PostgreSQL.
- Ridefinire la piattaforma (lift and reshape): trasferisci un'applicazione nel cloud e introduci un certo livello di ottimizzazione per sfruttare le funzionalità del cloud. Esempio: migra il tuo database Oracle locale ad Amazon Relational Database Service (Amazon RDS) per Oracle in. Cloud AWS
- Riacquistare (drop and shop): passa a un prodotto diverso, in genere effettuando la transizione da una licenza tradizionale a un modello SaaS. Esempio: migra il tuo sistema di gestione delle relazioni con i clienti (CRM) su Salesforce.com.
- Eseguire il rehosting (lift and shift): trasferisci un'applicazione sul cloud senza apportare modifiche per sfruttare le funzionalità del cloud. Esempio: migra il database Oracle locale su Oracle su un'istanza in. EC2 Cloud AWS
- Trasferire (eseguire il rehosting a livello hypervisor): trasferisci l'infrastruttura sul cloud senza acquistare nuovo hardware, riscrivere le applicazioni o modificare le operazioni esistenti. Si esegue la migrazione dei server da una piattaforma locale a un servizio cloud per la stessa piattaforma. Esempio: migra un'applicazione suMicrosoft Hyper-V. AWS
- Riesaminare (mantenere): mantieni le applicazioni nell'ambiente di origine. Queste potrebbero includere applicazioni che richiedono una rifattorizzazione significativa che desideri rimandare a un momento successivo e applicazioni legacy che desideri mantenere, perché non vi è alcuna giustificazione aziendale per effettuarne la migrazione.
- Ritirare: disattiva o rimuovi le applicazioni che non sono più necessarie nell'ambiente di origine.

A

ABAC

Vedi controllo degli accessi basato sugli attributi.

servizi astratti

Vedi servizi gestiti.

ACIDO

Vedi atomicità, consistenza, isolamento, durata.

migrazione attiva-attiva

Un metodo di migrazione del database in cui i database di origine e di destinazione vengono mantenuti sincronizzati (utilizzando uno strumento di replica bidirezionale o operazioni di doppia scrittura) ed entrambi i database gestiscono le transazioni provenienti dalle applicazioni di connessione durante la migrazione. Questo metodo supporta la migrazione in piccoli batch controllati anziché richiedere una conversione una tantum. È più flessibile ma richiede più lavoro rispetto alla migrazione attiva-passiva.

migrazione attiva-passiva

Un metodo di migrazione del database in cui i database di origine e di destinazione vengono mantenuti sincronizzati, ma solo il database di origine gestisce le transazioni provenienti dalle applicazioni di connessione mentre i dati vengono replicati nel database di destinazione. Il database di destinazione non accetta alcuna transazione durante la migrazione.

funzione di aggregazione

Una funzione SQL che opera su un gruppo di righe e calcola un singolo valore restituito per il gruppo. Esempi di funzioni aggregate includono `SUM` e. `MAX`

Intelligenza artificiale

Vedi intelligenza artificiale.

AIOps

Guarda le operazioni di intelligenza artificiale.

anonimizzazione

Il processo di eliminazione permanente delle informazioni personali in un set di dati. L'anonimizzazione può aiutare a proteggere la privacy personale. I dati anonimi non sono più considerati dati personali.

anti-modello

Una soluzione utilizzata di frequente per un problema ricorrente in cui la soluzione è controproducente, inefficace o meno efficace di un'alternativa.

controllo delle applicazioni

Un approccio alla sicurezza che consente l'uso solo di applicazioni approvate per proteggere un sistema dal malware.

portfolio di applicazioni

Una raccolta di informazioni dettagliate su ogni applicazione utilizzata da un'organizzazione, compresi i costi di creazione e manutenzione dell'applicazione e il relativo valore aziendale. Queste informazioni sono fondamentali per il processo di scoperta e analisi del portfolio e aiutano a identificare e ad assegnare la priorità alle applicazioni da migrare, modernizzare e ottimizzare.

intelligenza artificiale (IA)

Il campo dell'informatica dedicato all'uso delle tecnologie informatiche per svolgere funzioni cognitive tipicamente associate agli esseri umani, come l'apprendimento, la risoluzione di problemi e il riconoscimento di schemi. Per ulteriori informazioni, consulta la sezione Che cos'è l'intelligenza artificiale?

operazioni di intelligenza artificiale (AIOps)

Il processo di utilizzo delle tecniche di machine learning per risolvere problemi operativi, ridurre gli incidenti operativi e l'intervento umano e aumentare la qualità del servizio. Per ulteriori informazioni su come AIOps viene utilizzato nella strategia di AWS migrazione, consulta la guida all'integrazione delle operazioni.

crittografia asimmetrica

Un algoritmo di crittografia che utilizza una coppia di chiavi, una chiave pubblica per la crittografia e una chiave privata per la decrittografia. Puoi condividere la chiave pubblica perché non viene utilizzata per la decrittografia, ma l'accesso alla chiave privata deve essere altamente limitato.

atomicità, consistenza, isolamento, durabilità (ACID)

Un insieme di proprietà del software che garantiscono la validità dei dati e l'affidabilità operativa di un database, anche in caso di errori, interruzioni di corrente o altri problemi.

Controllo degli accessi basato su attributi (ABAC)

La pratica di creare autorizzazioni dettagliate basate su attributi utente, come reparto, ruolo professionale e nome del team. Per ulteriori informazioni, consulta ABAC AWS nella documentazione AWS Identity and Access Management (IAM).

fonte di dati autorevole

Una posizione in cui è archiviata la versione principale dei dati, considerata la fonte di informazioni più affidabile. È possibile copiare i dati dalla fonte di dati autorevole in altre posizioni allo scopo di elaborarli o modificarli, ad esempio anonimizzandoli, oscurandoli o pseudonimizzandoli.

Zona di disponibilità

Una posizione distinta all'interno di un edificio Regione AWS che è isolata dai guasti in altre zone di disponibilità e offre una connettività di rete economica e a bassa latenza verso altre zone di disponibilità nella stessa regione.

AWS Cloud Adoption Framework (CAF)AWS

Un framework di linee guida e best practice AWS per aiutare le organizzazioni a sviluppare un piano efficiente ed efficace per passare con successo al cloud. AWS CAF organizza le linee guida in sei aree di interesse chiamate prospettive: business, persone, governance, piattaforma, sicurezza e operazioni. Le prospettive relative ad azienda, persone e governance si concentrano sulle competenze e sui processi aziendali; le prospettive relative alla piattaforma, alla sicurezza e alle operazioni si concentrano sulle competenze e sui processi tecnici. Ad esempio, la prospettiva relativa alle persone si rivolge alle parti interessate che gestiscono le risorse umane (HR), le funzioni del personale e la gestione del personale. In questa prospettiva, AWS CAF fornisce linee guida per lo sviluppo delle persone, la formazione e le comunicazioni per aiutare a preparare l'organizzazione all'adozione del cloud di successo. Per ulteriori informazioni, consulta il sito web di AWS CAF e il white paper AWS CAF.

AWS Workload Qualification Framework (WQF)AWS

Uno strumento che valuta i carichi di lavoro di migrazione dei database, consiglia strategie di migrazione e fornisce stime del lavoro. AWS WQF è incluso in (). AWS Schema Conversion Tool AWS SCT Analizza gli schemi di database e gli oggetti di codice, il codice dell'applicazione, le dipendenze e le caratteristiche delle prestazioni e fornisce report di valutazione.

B

bot difettoso

Un bot che ha lo scopo di interrompere o causare danni a individui o organizzazioni.

BCP

Vedi la pianificazione della continuità operativa.

grafico comportamentale

Una vista unificata, interattiva dei comportamenti delle risorse e delle interazioni nel tempo. Puoi utilizzare un grafico comportamentale con Amazon Detective per esaminare tentativi di accesso non riusciti, chiamate API sospette e azioni simili. Per ulteriori informazioni, consulta Dati in un grafico comportamentale nella documentazione di Detective.

sistema big-endian

Un sistema che memorizza per primo il byte più importante. Vedi anche endianness.

Classificazione binaria

Un processo che prevede un risultato binario (una delle due classi possibili). Ad esempio, il modello di machine learning potrebbe dover prevedere problemi come "Questa e-mail è spam o non è spam?" o "Questo prodotto è un libro o un'auto?"

filtro Bloom

Una struttura di dati probabilistica ed efficiente in termini di memoria che viene utilizzata per verificare se un elemento fa parte di un set.

distribuzioni blu/verdi

Una strategia di implementazione in cui si creano due ambienti separati ma identici. La versione corrente dell'applicazione viene eseguita in un ambiente (blu) e la nuova versione dell'applicazione nell'altro ambiente (verde). Questa strategia consente di ripristinare rapidamente il sistema con un impatto minimo.

bot

Un'applicazione software che esegue attività automatizzate su Internet e simula l'attività o l'interazione umana. Alcuni bot sono utili o utili, come i web crawler che indicizzano le informazioni su Internet. Alcuni altri bot, noti come bot dannosi, hanno lo scopo di disturbare o causare danni a individui o organizzazioni.

botnet

Reti di bot infettate da malware e controllate da un'unica parte, nota come bot herder o bot operator. Le botnet sono il meccanismo più noto per scalare i bot e il loro impatto.

ramo

Un'area contenuta di un repository di codice. Il primo ramo creato in un repository è il ramo principale. È possibile creare un nuovo ramo a partire da un ramo esistente e quindi sviluppare funzionalità o correggere bug al suo interno. Un ramo creato per sviluppare una funzionalità viene comunemente detto ramo di funzionalità. Quando la funzionalità è pronta per il rilascio, il ramo di funzionalità viene ricongiunto al ramo principale. Per ulteriori informazioni, consulta Informazioni sulle filiali (documentazione). GitHub

accesso break-glass

In circostanze eccezionali e tramite una procedura approvata, un mezzo rapido per consentire a un utente di accedere a un sito a Account AWS cui in genere non dispone delle autorizzazioni necessarie. Per ulteriori informazioni, vedere l'indicatore Implementate break-glass procedures nella guida Well-Architected AWS .

strategia brownfield

L'infrastruttura esistente nell'ambiente. Quando si adotta una strategia brownfield per un'architettura di sistema, si progetta l'architettura in base ai vincoli dei sistemi e dell'infrastruttura attuali. Per l'espansione dell'infrastruttura esistente, è possibile combinare strategie brownfield e greenfield.

cache del buffer

L'area di memoria in cui sono archiviati i dati a cui si accede con maggiore frequenza.

capacità di business

Azioni intraprese da un'azienda per generare valore (ad esempio vendite, assistenza clienti o marketing). Le architetture dei microservizi e le decisioni di sviluppo possono essere guidate dalle capacità aziendali. Per ulteriori informazioni, consulta la sezione Organizzazione in base alle funzionalità aziendali del whitepaper Esecuzione di microservizi containerizzati su AWS.

pianificazione della continuità operativa (BCP)

Un piano che affronta il potenziale impatto di un evento che comporta l'interruzione dell'attività, come una migrazione su larga scala, sulle operazioni e consente a un'azienda di riprendere rapidamente le operazioni.

C

CAF

Vedi AWS Cloud Adoption Framework.

implementazione canaria

Il rilascio lento e incrementale di una versione agli utenti finali. Quando sei sicuro, distribuisci la nuova versione e sostituisci la versione corrente nella sua interezza.

CCoE

Vedi Cloud Center of Excellence.

CDC

Vedi Change Data Capture.

Change Data Capture (CDC)

Il processo di tracciamento delle modifiche a un'origine dati, ad esempio una tabella di database, e di registrazione dei metadati relativi alla modifica. È possibile utilizzare CDC per vari scopi, ad esempio il controllo o la replica delle modifiche in un sistema di destinazione per mantenere la sincronizzazione.

ingegneria del caos

Introduzione intenzionale di guasti o eventi dirompenti per testare la resilienza di un sistema. Puoi usare AWS Fault Injection Service (AWS FIS) per eseguire esperimenti che stressano i tuoi AWS carichi di lavoro e valutarne la risposta.

CI/CD

Vedi integrazione continua e distribuzione continua.

classificazione

Un processo di categorizzazione che aiuta a generare previsioni. I modelli di ML per problemi di classificazione prevedono un valore discreto. I valori discreti sono sempre distinti l'uno dall'altro. Ad esempio, un modello potrebbe dover valutare se in un'immagine è presente o meno un'auto.

crittografia lato client

Crittografia dei dati a livello locale, prima che il destinatario li Servizio AWS riceva.

Centro di eccellenza cloud (CCoE)

Un team multidisciplinare che guida le iniziative di adozione del cloud in tutta l'organizzazione, tra cui lo sviluppo di best practice per il cloud, la mobilitazione delle risorse, la definizione delle tempistiche di migrazione e la guida dell'organizzazione attraverso trasformazioni su larga scala. Per ulteriori informazioni, consulta gli CCoE post sull' Cloud AWS Enterprise Strategy Blog.

cloud computing

La tecnologia cloud generalmente utilizzata per l'archiviazione remota di dati e la gestione dei dispositivi IoT. Il cloud computing è generalmente collegato alla tecnologia di edge computing.

modello operativo cloud

In un'organizzazione IT, il modello operativo utilizzato per creare, maturare e ottimizzare uno o più ambienti cloud. Per ulteriori informazioni, consulta Building your Cloud Operating Model.

fasi di adozione del cloud

Le quattro fasi che le organizzazioni in genere attraversano quando migrano verso Cloud AWS:

- Progetto: esecuzione di alcuni progetti relativi al cloud per scopi di dimostrazione e apprendimento
- Fondamento: effettuare investimenti fondamentali per scalare l'adozione del cloud (ad esempio, creazione di una landing zone, definizione di una CCo E, definizione di un modello operativo)
- Migrazione: migrazione di singole applicazioni
- Reinvenzione: ottimizzazione di prodotti e servizi e innovazione nel cloud

Queste fasi sono state definite da Stephen Orban nel post sul blog The Journey Toward Cloud-First & the Stages of Adoption on the Enterprise Strategy. Cloud AWS Per informazioni su come si relazionano alla strategia di AWS migrazione, consulta la guida alla preparazione alla migrazione.

CMDB

Vedi database di gestione della configurazione.

repository di codice

Una posizione in cui il codice di origine e altri asset, come documentazione, esempi e script, vengono archiviati e aggiornati attraverso processi di controllo delle versioni. Gli archivi cloud più comuni includono GitHub oBitbucket Cloud. Ogni versione del codice è denominata ramo. In una struttura a microservizi, ogni repository è dedicato a una singola funzionalità. Una singola pipeline CI/CD può utilizzare più repository.

cache fredda

Una cache del buffer vuota, non ben popolata o contenente dati obsoleti o irrilevanti. Ciò influisce sulle prestazioni perché l'istanza di database deve leggere dalla memoria o dal disco principale, il che richiede più tempo rispetto alla lettura dalla cache del buffer.

dati freddi

Dati a cui si accede raramente e che in genere sono storici. Quando si eseguono interrogazioni di questo tipo di dati, le interrogazioni lente sono in genere accettabili. Lo spostamento di questi dati su livelli o classi di storage meno costosi e con prestazioni inferiori può ridurre i costi.

visione artificiale (CV)

Un campo dell'intelligenza artificiale che utilizza l'apprendimento automatico per analizzare ed estrarre informazioni da formati visivi come immagini e video digitali. Ad esempio, Amazon SageMaker AI fornisce algoritmi di elaborazione delle immagini per CV.

deriva della configurazione

Per un carico di lavoro, una modifica della configurazione rispetto allo stato previsto. Potrebbe causare la non conformità del carico di lavoro e in genere è graduale e involontaria.

database di gestione della configurazione (CMDB)

Un repository che archivia e gestisce le informazioni su un database e il relativo ambiente IT, inclusi i componenti hardware e software e le relative configurazioni. In genere si utilizzano i dati di un CMDB nella fase di individuazione e analisi del portafoglio della migrazione.

Pacchetto di conformità

Una raccolta di AWS Config regole e azioni correttive che puoi assemblare per personalizzare i controlli di conformità e sicurezza. È possibile distribuire un pacchetto di conformità come singola entità in una regione Account AWS and o all'interno di un'organizzazione utilizzando un modello YAML. Per ulteriori informazioni, consulta i Conformance Pack nella documentazione. AWS Config

integrazione e distribuzione continua (continuous integration and continuous delivery, CI/CD)

Il processo di automazione delle fasi di origine, compilazione, test, gestione temporanea e produzione del processo di rilascio del software. CI/CD viene comunemente descritto come una pipeline. CI/CD può aiutarvi ad automatizzare i processi, migliorare la produttività, migliorare la qualità del codice e velocizzare le consegne. Per ulteriori informazioni, consulta Vantaggi

della distribuzione continua. CD può anche significare continuous deployment (implementazione continua). Per ulteriori informazioni, consulta Distribuzione continua e implementazione continua a confronto.

CV

Vedi visione artificiale.

D

dati a riposo

Dati stazionari nella rete, ad esempio i dati archiviati.

classificazione dei dati

Un processo per identificare e classificare i dati nella rete in base alla loro criticità e sensibilità. È un componente fondamentale di qualsiasi strategia di gestione dei rischi di sicurezza informatica perché consente di determinare i controlli di protezione e conservazione appropriati per i dati. La classificazione dei dati è un componente del pilastro della sicurezza nel AWS Well-Architected Framework. Per ulteriori informazioni, consulta Classificazione dei dati.

deriva dei dati

Una variazione significativa tra i dati di produzione e i dati utilizzati per addestrare un modello di machine learning o una modifica significativa dei dati di input nel tempo. La deriva dei dati può ridurre la qualità, l'accuratezza e l'equità complessive nelle previsioni dei modelli ML.

dati in transito

Dati che si spostano attivamente attraverso la rete, ad esempio tra le risorse di rete.

rete di dati

Un framework architettonico che fornisce la proprietà distribuita e decentralizzata dei dati con gestione e governance centralizzate.

riduzione al minimo dei dati

Il principio della raccolta e del trattamento dei soli dati strettamente necessari. Praticare la riduzione al minimo dei dati in the Cloud AWS può ridurre i rischi per la privacy, i costi e l'impronta di carbonio delle analisi.

perimetro dei dati

Una serie di barriere preventive nell' AWS ambiente che aiutano a garantire che solo le identità attendibili accedano alle risorse attendibili delle reti previste. Per ulteriori informazioni, consulta Building a data perimeter on. AWS

pre-elaborazione dei dati

Trasformare i dati grezzi in un formato che possa essere facilmente analizzato dal modello di ML. La pre-elaborazione dei dati può comportare la rimozione di determinate colonne o righe e l'eliminazione di valori mancanti, incoerenti o duplicati.

provenienza dei dati

Il processo di tracciamento dell'origine e della cronologia dei dati durante il loro ciclo di vita, ad esempio il modo in cui i dati sono stati generati, trasmessi e archiviati.

soggetto dei dati

Un individuo i cui dati vengono raccolti ed elaborati.

data warehouse

Un sistema di gestione dei dati che supporta la business intelligence, come l'analisi. I data warehouse contengono in genere grandi quantità di dati storici e vengono generalmente utilizzati per interrogazioni e analisi.

linguaggio di definizione del database (DDL)

Istruzioni o comandi per creare o modificare la struttura di tabelle e oggetti in un database.

linguaggio di manipolazione del database (DML)

Istruzioni o comandi per modificare (inserire, aggiornare ed eliminare) informazioni in un database.

DDL

Vedi linguaggio di definizione del database.

deep ensemble

Combinare più modelli di deep learning per la previsione. È possibile utilizzare i deep ensemble per ottenere una previsione più accurata o per stimare l'incertezza nelle previsioni.

deep learning

Un sottocampo del ML che utilizza più livelli di reti neurali artificiali per identificare la mappatura tra i dati di input e le variabili target di interesse.

defense-in-depth

Un approccio alla sicurezza delle informazioni in cui una serie di meccanismi e controlli di sicurezza sono accuratamente stratificati su una rete di computer per proteggere la riservatezza, l'integrità e la disponibilità della rete e dei dati al suo interno. Quando si adotta questa strategia AWS, si aggiungono più controlli a diversi livelli della AWS Organizations struttura per proteggere le risorse. Ad esempio, un defense-in-depth approccio potrebbe combinare l'autenticazione a più fattori, la segmentazione della rete e la crittografia.

amministratore delegato

In AWS Organizations, un servizio compatibile può registrare un account AWS membro per amministrare gli account dell'organizzazione e gestire le autorizzazioni per quel servizio. Questo account è denominato amministratore delegato per quel servizio specifico. Per ulteriori informazioni e un elenco di servizi compatibili, consulta Servizi che funzionano con AWS Organizations nella documentazione di AWS Organizations .

implementazione

Il processo di creazione di un'applicazione, di nuove funzionalità o di correzioni di codice disponibili nell'ambiente di destinazione. L'implementazione prevede l'applicazione di modifiche in una base di codice, seguita dalla creazione e dall'esecuzione di tale base di codice negli ambienti applicativi.

Ambiente di sviluppo

Vedi ambiente.

controllo di rilevamento

Un controllo di sicurezza progettato per rilevare, registrare e avvisare dopo che si è verificato un evento. Questi controlli rappresentano una seconda linea di difesa e avvisano l'utente in caso di eventi di sicurezza che aggirano i controlli preventivi in vigore. Per ulteriori informazioni, consulta Controlli di rilevamento in Implementazione dei controlli di sicurezza in AWS.

mappatura del flusso di valore dello sviluppo (DVSM)

Un processo utilizzato per identificare e dare priorità ai vincoli che influiscono negativamente sulla velocità e sulla qualità nel ciclo di vita dello sviluppo del software. DVSM estende il processo di

mappatura del flusso di valore originariamente progettato per pratiche di produzione snella. Si concentra sulle fasi e sui team necessari per creare e trasferire valore attraverso il processo di sviluppo del software.

gemello digitale

Una rappresentazione virtuale di un sistema reale, ad esempio un edificio, una fabbrica, un'attrezzatura industriale o una linea di produzione. I gemelli digitali supportano la manutenzione predittiva, il monitoraggio remoto e l'ottimizzazione della produzione.

tabella delle dimensioni

In uno schema a stella, una tabella più piccola che contiene gli attributi dei dati quantitativi in una tabella dei fatti. Gli attributi della tabella delle dimensioni sono in genere campi di testo o numeri discreti che si comportano come testo. Questi attributi vengono comunemente utilizzati per il vincolo delle query, il filtraggio e l'etichettatura dei set di risultati.

disastro

Un evento che impedisce a un carico di lavoro o a un sistema di raggiungere gli obiettivi aziendali nella sua sede principale di implementazione. Questi eventi possono essere disastri naturali, guasti tecnici o il risultato di azioni umane, come errori di configurazione involontari o attacchi di malware.

disaster recovery (DR)

La strategia e il processo utilizzati per ridurre al minimo i tempi di inattività e la perdita di dati causati da un disastro. Per ulteriori informazioni, consulta Disaster Recovery of Workloads su AWS: Recovery in the Cloud in the AWS Well-Architected Framework.

DML

Vedi linguaggio di manipolazione del database.

progettazione basata sul dominio

Un approccio allo sviluppo di un sistema software complesso collegandone i componenti a domini in evoluzione, o obiettivi aziendali principali, perseguiti da ciascun componente. Questo concetto è stato introdotto da Eric Evans nel suo libro,Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston: Addison-Wesley Professional, 2003). Per informazioni su come utilizzare la progettazione basata sul dominio con il modello del fico strangolatore (Strangler Fig), consulta la sezione Modernizzazione incrementale dei servizi Web Microsoft ASP.NET (ASMX) legacy utilizzando container e il Gateway Amazon API.

DOTT.

Vedi disaster recovery.

rilevamento della deriva

Tracciamento delle deviazioni da una configurazione di base. Ad esempio, è possibile AWS CloudFormation utilizzarlo per rilevare deviazioni nelle risorse di sistema oppure AWS Control Tower per rilevare cambiamenti nella landing zone che potrebbero influire sulla conformità ai requisiti di governance.

DVSM

Vedi la mappatura del flusso di valore dello sviluppo.

E

EDA

Vedi analisi esplorativa dei dati.

MODIFICA

Vedi scambio elettronico di dati.

edge computing

La tecnologia che aumenta la potenza di calcolo per i dispositivi intelligenti all'edge di una rete IoT. Rispetto al cloud computing, l'edge computing può ridurre la latenza di comunicazione e migliorare i tempi di risposta.

scambio elettronico di dati (EDI)

Lo scambio automatizzato di documenti aziendali tra organizzazioni. Per ulteriori informazioni, vedere Cos'è lo scambio elettronico di dati.

crittografia

Un processo di elaborazione che trasforma i dati in chiaro, leggibili dall'uomo, in testo cifrato.

chiave crittografica

Una stringa crittografica di bit randomizzati generata da un algoritmo di crittografia. Le chiavi possono variare di lunghezza e ogni chiave è progettata per essere imprevedibile e univoca.

endianità

L'ordine in cui i byte vengono archiviati nella memoria del computer. I sistemi big-endian memorizzano per primo il byte più importante. I sistemi little-endian memorizzano per primo il byte meno importante.

endpoint

Vedi service endpoint.

servizio endpoint

Un servizio che puoi ospitare in un cloud privato virtuale (VPC) da condividere con altri utenti. Puoi creare un servizio endpoint con AWS PrivateLink e concedere autorizzazioni ad altri Account AWS o a AWS Identity and Access Management (IAM) principali. Questi account o principali possono connettersi al servizio endpoint in privato creando endpoint VPC di interfaccia. Per ulteriori informazioni, consulta Creazione di un servizio endpoint nella documentazione di Amazon Virtual Private Cloud (Amazon VPC).

pianificazione delle risorse aziendali (ERP)

Un sistema che automatizza e gestisce i processi aziendali chiave (come contabilità, MES e gestione dei progetti) per un'azienda.

crittografia envelope

Il processo di crittografia di una chiave di crittografia con un'altra chiave di crittografia. Per ulteriori informazioni, vedete Envelope encryption nella documentazione AWS Key Management Service (AWS KMS).

ambiente

Un'istanza di un'applicazione in esecuzione. Di seguito sono riportati i tipi di ambiente più comuni nel cloud computing:

- ambiente di sviluppo: un'istanza di un'applicazione in esecuzione disponibile solo per il team principale responsabile della manutenzione dell'applicazione. Gli ambienti di sviluppo vengono utilizzati per testare le modifiche prima di promuoverle negli ambienti superiori. Questo tipo di ambiente viene talvolta definito ambiente di test.
- ambienti inferiori: tutti gli ambienti di sviluppo di un'applicazione, ad esempio quelli utilizzati per le build e i test iniziali.
- ambiente di produzione: un'istanza di un'applicazione in esecuzione a cui gli utenti finali possono accedere. In una CI/CD pipeline, l'ambiente di produzione è l'ultimo ambiente di distribuzione.

- ambienti superiori: tutti gli ambienti a cui possono accedere utenti diversi dal team di sviluppo principale. Si può trattare di un ambiente di produzione, ambienti di preproduzione e ambienti per i test di accettazione da parte degli utenti.

epica

Nelle metodologie agili, categorie funzionali che aiutano a organizzare e dare priorità al lavoro. Le epiche forniscono una descrizione di alto livello dei requisiti e delle attività di implementazione. Ad esempio, le epopee della sicurezza AWS CAF includono la gestione delle identità e degli accessi, i controlli investigativi, la sicurezza dell'infrastruttura, la protezione dei dati e la risposta agli incidenti. Per ulteriori informazioni sulle epiche, consulta la strategia di migrazione AWS , consulta la guida all'implementazione del programma.

ERP

Vedi pianificazione delle risorse aziendali.

analisi esplorativa dei dati (EDA)

Il processo di analisi di un set di dati per comprenderne le caratteristiche principali. Si raccolgono o si aggregano dati e quindi si eseguono indagini iniziali per trovare modelli, rilevare anomalie e verificare ipotesi. L'EDA viene eseguita calcolando statistiche di riepilogo e creando visualizzazioni di dati.

F

tabella dei fatti

Il tavolo centrale con schema a stella. Memorizza dati quantitativi sulle operazioni aziendali. In genere, una tabella dei fatti contiene due tipi di colonne: quelle che contengono misure e quelle che contengono una chiave esterna per una tabella di dimensioni.

fallire velocemente

Una filosofia che utilizza test frequenti e incrementali per ridurre il ciclo di vita dello sviluppo. È una parte fondamentale di un approccio agile.

limite di isolamento dei guasti

Nel Cloud AWS, un limite come una zona di disponibilità Regione AWS, un piano di controllo o un piano dati che limita l'effetto di un errore e aiuta a migliorare la resilienza dei carichi di lavoro. Per ulteriori informazioni, consulta AWS Fault Isolation Boundaries.

ramo di funzionalità

Vedi filiale.

caratteristiche

I dati di input che usi per fare una previsione. Ad esempio, in un contesto di produzione, le caratteristiche potrebbero essere immagini acquisite periodicamente dalla linea di produzione.

importanza delle caratteristiche

Quanto è importante una caratteristica per le previsioni di un modello. Di solito viene espresso come punteggio numerico che può essere calcolato con varie tecniche, come Shapley Additive Explanations (SHAP) e gradienti integrati. Per ulteriori informazioni, consulta Interpretabilità del modello di machine learning con AWS.

trasformazione delle funzionalità

Per ottimizzare i dati per il processo di machine learning, incluso l'arricchimento dei dati con fonti aggiuntive, il dimensionamento dei valori o l'estrazione di più set di informazioni da un singolo campo di dati. Ciò consente al modello di ML di trarre vantaggio dai dati. Ad esempio, se suddividi la data "2021-05-27 00:15:37" in "2021", "maggio", "giovedì" e "15", puoi aiutare l'algoritmo di apprendimento ad apprendere modelli sfumati associati a diversi componenti dei dati.

prompt con pochi scatti

Fornire a un LLM un numero limitato di esempi che dimostrino l'attività e il risultato desiderato prima di chiedergli di eseguire un'attività simile. Questa tecnica è un'applicazione dell'apprendimento contestuale, in cui i modelli imparano da esempi (immagini) incorporati nei prompt. I prompt con pochi passaggi possono essere efficaci per attività che richiedono una formattazione, un ragionamento o una conoscenza del dominio specifici. Vedi anche zero-shot prompting.

FGAC

Vedi il controllo granulare degli accessi.

controllo granulare degli accessi (FGAC)

L'uso di più condizioni per consentire o rifiutare una richiesta di accesso.

migrazione flash-cut

Un metodo di migrazione del database che utilizza la replica continua dei dati tramite l'acquisizione dei dati delle modifiche per migrare i dati nel più breve tempo possibile, anziché utilizzare un approccio graduale. L'obiettivo è ridurre al minimo i tempi di inattività.

FM

Vedi modello di base.

modello di fondazione (FM)

Una grande rete neurale di deep learning che si è addestrata su enormi set di dati generalizzati e non etichettati. FMs sono in grado di svolgere un'ampia varietà di attività generali, come comprendere il linguaggio, generare testo e immagini e conversare in linguaggio naturale. Per ulteriori informazioni, consulta Cosa sono i modelli Foundation.

G

AI generativa

Un sottoinsieme di modelli di intelligenza artificiale che sono stati addestrati su grandi quantità di dati e che possono utilizzare un semplice prompt di testo per creare nuovi contenuti e artefatti, come immagini, video, testo e audio. Per ulteriori informazioni, consulta Cos'è l'IA generativa.

blocco geografico

Vedi restrizioni geografiche.

limitazioni geografiche (blocco geografico)

In Amazon CloudFront, un'opzione per impedire agli utenti di determinati paesi di accedere alle distribuzioni di contenuti. Puoi utilizzare un elenco consentito o un elenco di blocco per specificare i paesi approvati e vietati. Per ulteriori informazioni, consulta Limitare la distribuzione geografica dei contenuti nella CloudFront documentazione.

Flusso di lavoro di GitFlow

Un approccio in cui gli ambienti inferiori e superiori utilizzano rami diversi in un repository di codice di origine. Il flusso di lavoro Gitflow è considerato obsoleto e il flusso di lavoro basato su trunk è l'approccio moderno e preferito.

immagine dorata

Un'istantanea di un sistema o di un software che viene utilizzata come modello per distribuire nuove istanze di quel sistema o software. Ad esempio, nella produzione, un'immagine dorata può essere utilizzata per fornire software su più dispositivi e contribuire a migliorare la velocità, la scalabilità e la produttività nelle operazioni di produzione dei dispositivi.

strategia greenfield

L'assenza di infrastrutture esistenti in un nuovo ambiente. Quando si adotta una strategia greenfield per un'architettura di sistema, è possibile selezionare tutte le nuove tecnologie senza il vincolo della compatibilità con l'infrastruttura esistente, nota anche come brownfield. Per l'espansione dell'infrastruttura esistente, è possibile combinare strategie brownfield e greenfield.

guardrail

Una regola di alto livello che aiuta a governare le risorse, le politiche e la conformità tra le unità organizzative (). OUs I guardrail preventivi applicano le policy per garantire l'allineamento agli standard di conformità. Vengono implementati utilizzando le policy di controllo dei servizi e i limiti delle autorizzazioni IAM. I guardrail di rilevamento rilevano le violazioni delle policy e i problemi di conformità e generano avvisi per porvi rimedio. Sono implementati utilizzando Amazon AWS Config AWS Security Hub GuardDuty AWS Trusted Advisor, Amazon Inspector e controlli personalizzati AWS Lambda .

H

AH

Vedi disponibilità elevata.

migrazione di database eterogenea

Migrazione del database di origine in un database di destinazione che utilizza un motore di database diverso (ad esempio, da Oracle ad Amazon Aurora). La migrazione eterogenea fa in genere parte di uno sforzo di riprogettazione e la conversione dello schema può essere un'attività complessa. AWS offre AWS SCT che aiuta con le conversioni dello schema.

alta disponibilità (HA)

La capacità di un carico di lavoro di funzionare in modo continuo, senza intervento, in caso di sfide o disastri. I sistemi HA sono progettati per il failover automatico, fornire costantemente prestazioni di alta qualità e gestire carichi e guasti diversi con un impatto minimo sulle prestazioni.

modernizzazione storica

Un approccio utilizzato per modernizzare e aggiornare i sistemi di tecnologia operativa (OT) per soddisfare meglio le esigenze dell'industria manifatturiera. Uno storico è un tipo di database utilizzato per raccogliere e archiviare dati da varie fonti in una fabbrica.

dati di esclusione

Una parte di dati storici etichettati che viene trattenuta da un set di dati utilizzata per addestrare un modello di apprendimento automatico. È possibile utilizzare i dati di holdout per valutare le prestazioni del modello confrontando le previsioni del modello con i dati di holdout.

migrazione di database omogenea

Migrazione del database di origine in un database di destinazione che condivide lo stesso motore di database (ad esempio, da Microsoft SQL Server ad Amazon RDS per SQL Server). La migrazione omogenea fa in genere parte di un'operazione di rehosting o ridefinizione della piattaforma. Per migrare lo schema è possibile utilizzare le utilità native del database.

dati caldi

Dati a cui si accede frequentemente, come dati in tempo reale o dati di traduzione recenti. Questi dati richiedono in genere un livello o una classe di storage ad alte prestazioni per fornire risposte rapide alle query.

hotfix

Una soluzione urgente per un problema critico in un ambiente di produzione. A causa della sua urgenza, un hotfix viene in genere creato al di fuori del tipico DevOps flusso di lavoro di rilascio.

periodo di hypercare

Subito dopo la conversione, il periodo di tempo in cui un team di migrazione gestisce e monitora le applicazioni migrate nel cloud per risolvere eventuali problemi. In genere, questo periodo dura da 1 a 4 giorni. Al termine del periodo di hypercare, il team addetto alla migrazione in genere trasferisce la responsabilità delle applicazioni al team addetto alle operazioni cloud.

I

IaC

Considera l'infrastruttura come codice.

Policy basata su identità

Una policy associata a uno o più principi IAM che definisce le relative autorizzazioni all'interno dell' Cloud AWS ambiente.

applicazione inattiva

Un'applicazione che prevede un uso di CPU e memoria medio compreso tra il 5% e il 20% in un periodo di 90 giorni. In un progetto di migrazione, è normale ritirare queste applicazioni o mantenerle on-premise.

IIoT

Vedi Industrial Internet of Things.

infrastruttura immutabile

Un modello che implementa una nuova infrastruttura per i carichi di lavoro di produzione anziché aggiornare, applicare patch o modificare l'infrastruttura esistente. Le infrastrutture immutabili sono intrinsecamente più coerenti, affidabili e prevedibili delle infrastrutture mutabili. Per ulteriori informazioni, consulta la best practice Deploy using immutable infrastructure in Well-Architected AWS Framework.

VPC in ingresso (ingress)

In un'architettura AWS multi-account, un VPC che accetta, ispeziona e indirizza le connessioni di rete dall'esterno di un'applicazione. La AWS Security Reference Architecture consiglia di configurare l'account di rete con funzionalità in entrata, in uscita e di ispezione VPCs per proteggere l'interfaccia bidirezionale tra l'applicazione e la rete Internet in generale.

migrazione incrementale

Una strategia di conversione in cui si esegue la migrazione dell'applicazione in piccole parti anziché eseguire una conversione singola e completa. Ad esempio, inizialmente potresti spostare solo alcuni microservizi o utenti nel nuovo sistema. Dopo aver verificato che tutto funzioni correttamente, puoi spostare in modo incrementale microservizi o utenti aggiuntivi fino alla disattivazione del sistema legacy. Questa strategia riduce i rischi associati alle migrazioni di grandi dimensioni.

Industria 4.0

Un termine introdotto da Klaus Schwab nel 2016 per riferirsi alla modernizzazione dei processi di produzione attraverso progressi in termini di connettività, dati in tempo reale, automazione, analisi e AI/ML.

infrastruttura

Tutte le risorse e gli asset contenuti nell'ambiente di un'applicazione.

infrastruttura come codice (IaC)

Il processo di provisioning e gestione dell'infrastruttura di un'applicazione tramite un insieme di file di configurazione. Il processo IaC è progettato per aiutarti a centralizzare la gestione dell'infrastruttura, a standardizzare le risorse e a dimensionare rapidamente, in modo che i nuovi ambienti siano ripetibili, affidabili e coerenti.

IIoInternet delle cose industriale (T)

L'uso di sensori e dispositivi connessi a Internet nei settori industriali, come quello manifatturiero, energetico, automobilistico, sanitario, delle scienze della vita e dell'agricoltura. Per ulteriori informazioni, vedere Creazione di una strategia di trasformazione digitale per l'Internet of Things (IIoT) industriale.

VPC di ispezione

In un'architettura AWS multi-account, un VPC centralizzato che gestisce le ispezioni del traffico di rete tra VPCs (nello stesso o in modo diverso Regioni AWS), Internet e le reti locali. La AWS Security Reference Architecture consiglia di configurare l'account di rete con informazioni in entrata, in uscita e di ispezione VPCs per proteggere l'interfaccia bidirezionale tra l'applicazione e Internet in generale.

Internet of Things (IoT)

La rete di oggetti fisici connessi con sensori o processori incorporati che comunicano con altri dispositivi e sistemi tramite Internet o una rete di comunicazione locale. Per ulteriori informazioni, consulta Cos'è l'IoT?

interpretabilità

Una caratteristica di un modello di machine learning che descrive il grado in cui un essere umano è in grado di comprendere in che modo le previsioni del modello dipendono dai suoi input. Per ulteriori informazioni, vedere Interpretabilità del modello di machine learning con. AWS

IoT

Vedi Internet of Things.

libreria di informazioni IT (ITIL)

Una serie di best practice per offrire servizi IT e allinearli ai requisiti aziendali. ITIL fornisce le basi per ITSM.

gestione dei servizi IT (ITSM)

Attività associate alla progettazione, implementazione, gestione e supporto dei servizi IT per un'organizzazione. Per informazioni sull'integrazione delle operazioni cloud con gli strumenti ITSM, consulta la guida all'integrazione delle operazioni.

ITIL

Vedi la libreria di informazioni IT.

ITSM

Vedi Gestione dei servizi IT.

L

controllo degli accessi basato su etichette (LBAC)

Un'implementazione del controllo di accesso obbligatorio (MAC) in cui agli utenti e ai dati stessi viene assegnato esplicitamente un valore di etichetta di sicurezza. L'intersezione tra l'etichetta di sicurezza utente e l'etichetta di sicurezza dei dati determina quali righe e colonne possono essere visualizzate dall'utente.

zona di destinazione

Una landing zone è un AWS ambiente multi-account ben progettato, scalabile e sicuro. Questo è un punto di partenza dal quale le organizzazioni possono avviare e distribuire rapidamente carichi di lavoro e applicazioni con fiducia nel loro ambiente di sicurezza e infrastruttura. Per ulteriori informazioni sulle zone di destinazione, consulta la sezione Configurazione di un ambiente AWS multi-account sicuro e scalabile.

modello linguistico di grandi dimensioni (LLM)

Un modello di intelligenza artificiale di deep learning preaddestrato su una grande quantità di dati. Un LLM può svolgere più attività, come rispondere a domande, riepilogare documenti, tradurre testo in altre lingue e completare frasi. Per ulteriori informazioni, consulta Cosa sono. LLMs

migrazione su larga scala

Una migrazione di 300 o più server.

BIANCO

Vedi controllo degli accessi basato su etichette.

Privilegio minimo

La best practice di sicurezza per la concessione delle autorizzazioni minime richieste per eseguire un'attività. Per ulteriori informazioni, consulta Applicazione delle autorizzazioni del privilegio minimo nella documentazione di IAM.

eseguire il rehosting (lift and shift)

Vedi 7 R.

sistema little-endian

Un sistema che memorizza per primo il byte meno importante. Vedi anche endianità.

LLM

Vedi modello linguistico di grandi dimensioni.

ambienti inferiori

Vedi ambiente.

M

machine learning (ML)

Un tipo di intelligenza artificiale che utilizza algoritmi e tecniche per il riconoscimento e l'apprendimento di schemi. Il machine learning analizza e apprende dai dati registrati, come i dati dell'Internet delle cose (IoT), per generare un modello statistico basato su modelli. Per ulteriori informazioni, consulta la sezione Machine learning.

ramo principale

Vedi filiale.

malware

Software progettato per compromettere la sicurezza o la privacy del computer. Il malware potrebbe interrompere i sistemi informatici, divulgare informazioni sensibili o ottenere accessi non autorizzati. Esempi di malware includono virus, worm, ransomware, trojan horse, spyware e keylogger.

servizi gestiti

Servizi AWS per cui AWS gestisce il livello di infrastruttura, il sistema operativo e le piattaforme e si accede agli endpoint per archiviare e recuperare i dati. Amazon Simple Storage Service

(Amazon S3) Simple Storage Service (Amazon S3) e Amazon DynamoDB sono esempi di servizi gestiti. Questi sono noti anche come servizi astratti.

sistema di esecuzione della produzione (MES)

Un sistema software per tracciare, monitorare, documentare e controllare i processi di produzione che convertono le materie prime in prodotti finiti in officina.

MAP

Vedi Migration Acceleration Program.

meccanismo

Un processo completo in cui si crea uno strumento, si promuove l'adozione dello strumento e quindi si esaminano i risultati per apportare le modifiche. Un meccanismo è un ciclo che si rafforza e si migliora man mano che funziona. Per ulteriori informazioni, consulta Creazione di meccanismi nel AWS Well-Architected Framework.

account membro

Tutti gli account Account AWS diversi dall'account di gestione che fanno parte di un'organizzazione in. AWS Organizations Un account può essere membro di una sola organizzazione alla volta.

MEH

Vedi sistema di esecuzione della produzione.

Message Queuing Telemetry Transport (MQTT)

Un protocollo di comunicazione machine-to-machine (M2M) leggero, basato sul modello di pubblicazione/sottoscrizione, per dispositivi IoT con risorse limitate.

microservizio

Un servizio piccolo e indipendente che comunica tramite canali ben definiti ed è in genere di proprietà di piccoli team autonomi. APIs Ad esempio, un sistema assicurativo potrebbe includere microservizi che si riferiscono a funzionalità aziendali, come vendite o marketing, o sottodomini, come acquisti, reclami o analisi. I vantaggi dei microservizi includono agilità, dimensionamento flessibile, facilità di implementazione, codice riutilizzabile e resilienza. Per ulteriori informazioni, consulta Integrazione dei microservizi utilizzando servizi serverless. AWS

architettura di microservizi

Un approccio alla creazione di un'applicazione con componenti indipendenti che eseguono ogni processo applicativo come microservizio. Questi microservizi comunicano attraverso un'interfaccia

ben definita utilizzando sistemi leggeri. APIs Ogni microservizio in questa architettura può essere aggiornato, distribuito e dimensionato per soddisfare la richiesta di funzioni specifiche di un'applicazione. Per ulteriori informazioni, vedere Implementazione dei microservizi su. AWS

Programma di accelerazione della migrazione (MAP)

Un AWS programma che fornisce consulenza, supporto, formazione e servizi per aiutare le organizzazioni a costruire una solida base operativa per il passaggio al cloud e per contribuire a compensare il costo iniziale delle migrazioni. MAP include una metodologia di migrazione per eseguire le migrazioni precedenti in modo metodico e un set di strumenti per automatizzare e accelerare gli scenari di migrazione comuni.

migrazione su larga scala

Il processo di trasferimento della maggior parte del portfolio di applicazioni sul cloud avviene a ondate, con più applicazioni trasferite a una velocità maggiore in ogni ondata. Questa fase utilizza le migliori pratiche e le lezioni apprese nelle fasi precedenti per implementare una fabbrica di migrazione di team, strumenti e processi per semplificare la migrazione dei carichi di lavoro attraverso l'automazione e la distribuzione agile. Questa è la terza fase della strategia di migrazione AWS.

fabbrica di migrazione

Team interfunzionali che semplificano la migrazione dei carichi di lavoro attraverso approcci automatizzati e agili. I team di Migration Factory in genere includono addetti alle operazioni, analisti e proprietari aziendali, ingegneri addetti alla migrazione, sviluppatori e DevOps professionisti che lavorano nell'ambito degli sprint. Tra il 20% e il 50% di un portfolio di applicazioni aziendali è costituito da schemi ripetuti che possono essere ottimizzati con un approccio di fabbrica. Per ulteriori informazioni, consulta la discussione sulle fabbriche di migrazione e la Guida alla fabbrica di migrazione al cloud in questo set di contenuti.

metadati di migrazione

Le informazioni sull'applicazione e sul server necessarie per completare la migrazione. Ogni modello di migrazione richiede un set diverso di metadati di migrazione. Esempi di metadati di migrazione includono la sottorete, il gruppo di sicurezza e l'account di destinazione. AWS

modello di migrazione

Un'attività di migrazione ripetibile che descrive in dettaglio la strategia di migrazione, la destinazione della migrazione e l'applicazione o il servizio di migrazione utilizzati. Esempio: riorganizza la migrazione su Amazon EC2 con AWS Application Migration Service.

Valutazione del portfolio di migrazione (MPA)

Uno strumento online che fornisce informazioni per la convalida del business case per la migrazione a. Cloud AWS MPA offre una valutazione dettagliata del portfolio (dimensionamento corretto dei server, prezzi, confronto del TCO, analisi dei costi di migrazione) e pianificazione della migrazione (analisi e raccolta dei dati delle applicazioni, raggruppamento delle applicazioni, prioritizzazione delle migrazioni e pianificazione delle ondate). Lo strumento MPA (richiede l'accesso) è disponibile gratuitamente per tutti i AWS consulenti e i consulenti dei partner APN.

valutazione della preparazione alla migrazione (MRA)

Il processo di acquisizione di informazioni sullo stato di preparazione al cloud di un'organizzazione, l'identificazione dei punti di forza e di debolezza e la creazione di un piano d'azione per colmare le lacune identificate, utilizzando il CAF. AWS Per ulteriori informazioni, consulta la guida di preparazione alla migrazione. MRA è la prima fase della strategia di migrazione AWS.

strategia di migrazione

L'approccio utilizzato per migrare un carico di lavoro verso. Cloud AWS Per ulteriori informazioni, consulta la voce 7 R in questo glossario e consulta Mobilita la tua organizzazione per accelerare le migrazioni su larga scala.

ML

Vedi machine learning.

modernizzazione

Trasformazione di un'applicazione obsoleta (legacy o monolitica) e della relativa infrastruttura in un sistema agile, elastico e altamente disponibile nel cloud per ridurre i costi, aumentare l'efficienza e sfruttare le innovazioni. Per ulteriori informazioni, vedere Strategia per la modernizzazione delle applicazioni in. Cloud AWS

valutazione della preparazione alla modernizzazione

Una valutazione che aiuta a determinare la preparazione alla modernizzazione delle applicazioni di un'organizzazione, identifica vantaggi, rischi e dipendenze e determina in che misura l'organizzazione può supportare lo stato futuro di tali applicazioni. Il risultato della valutazione è uno schema dell'architettura di destinazione, una tabella di marcia che descrive in dettaglio le fasi di sviluppo e le tappe fondamentali del processo di modernizzazione e un piano d'azione per colmare le lacune identificate. Per ulteriori informazioni, vedere Valutazione della preparazione alla modernizzazione per le applicazioni in. Cloud AWS

applicazioni monolitiche (monoliti)

Applicazioni eseguite come un unico servizio con processi strettamente collegati. Le applicazioni monolitiche presentano diversi inconvenienti. Se una funzionalità dell'applicazione registra un picco di domanda, l'intera architettura deve essere dimensionata. L'aggiunta o il miglioramento delle funzionalità di un'applicazione monolitica diventa inoltre più complessa man mano che la base di codice cresce. Per risolvere questi problemi, puoi utilizzare un'architettura di microservizi. Per ulteriori informazioni, consulta la sezione Scomposizione dei monoliti in microservizi.

MAPPA

Vedi Migration Portfolio Assessment.

MQTT

Vedi Message Queuing Telemetry Transport.

classificazione multiclasse

Un processo che aiuta a generare previsioni per più classi (prevedendo uno o più di due risultati). Ad esempio, un modello di machine learning potrebbe chiedere "Questo prodotto è un libro, un'auto o un telefono?" oppure "Quale categoria di prodotti è più interessante per questo cliente?"

infrastruttura mutabile

Un modello che aggiorna e modifica l'infrastruttura esistente per i carichi di lavoro di produzione. Per migliorare la coerenza, l'affidabilità e la prevedibilità, il AWS Well-Architected Framework consiglia l'uso di un'infrastruttura immutabile come best practice.

O

OAC

Vedi Origin Access Control.

QUERCIA

Vedi Origin Access Identity.

OCM

Vedi gestione delle modifiche organizzative.

migrazione offline

Un metodo di migrazione in cui il carico di lavoro di origine viene eliminato durante il processo di migrazione. Questo metodo prevede tempi di inattività prolungati e viene in genere utilizzato per carichi di lavoro piccoli e non critici.

OI

Vedi l'integrazione delle operazioni.

OLA

Vedi accordo a livello operativo.

migrazione online

Un metodo di migrazione in cui il carico di lavoro di origine viene copiato sul sistema di destinazione senza essere messo offline. Le applicazioni connesse al carico di lavoro possono continuare a funzionare durante la migrazione. Questo metodo comporta tempi di inattività pari a zero o comunque minimi e viene in genere utilizzato per carichi di lavoro di produzione critici.

OPC-UA

Vedi Open Process Communications - Unified Architecture.

Comunicazioni a processo aperto - Architettura unificata (OPC-UA)

Un protocollo di comunicazione machine-to-machine (M2M) per l'automazione industriale. OPC-UA fornisce uno standard di interoperabilità con schemi di crittografia, autenticazione e autorizzazione dei dati.

accordo a livello operativo (OLA)

Un accordo che chiarisce quali sono gli impegni reciproci tra i gruppi IT funzionali, a supporto di un accordo sul livello di servizio (SLA).

revisione della prontezza operativa (ORR)

Un elenco di domande e best practice associate che aiutano a comprendere, valutare, prevenire o ridurre la portata degli incidenti e dei possibili guasti. Per ulteriori informazioni, vedere Operational Readiness Reviews (ORR) nel Well-Architected AWS Framework.

tecnologia operativa (OT)

Sistemi hardware e software che interagiscono con l'ambiente fisico per controllare le operazioni, le apparecchiature e le infrastrutture industriali. Nella produzione, l'integrazione di sistemi OT e di tecnologia dell'informazione (IT) è un obiettivo chiave per le trasformazioni dell'Industria 4.0.

integrazione delle operazioni (OI)

Il processo di modernizzazione delle operazioni nel cloud, che prevede la pianificazione, l'automazione e l'integrazione della disponibilità. Per ulteriori informazioni, consulta la guida all'integrazione delle operazioni.

trail organizzativo

Un percorso creato da noi AWS CloudTrail che registra tutti gli eventi di un'organizzazione per tutti Account AWS . AWS Organizations Questo percorso viene creato in ogni Account AWS che fa parte dell'organizzazione e tiene traccia dell'attività in ogni account. Per ulteriori informazioni, consulta Creazione di un percorso per un'organizzazione nella CloudTrail documentazione.

gestione del cambiamento organizzativo (OCM)

Un framework per la gestione di trasformazioni aziendali importanti e che comportano l'interruzione delle attività dal punto di vista delle persone, della cultura e della leadership. OCM aiuta le organizzazioni a prepararsi e passare a nuovi sistemi e strategie accelerando l'adozione del cambiamento, affrontando i problemi di transizione e promuovendo cambiamenti culturali e organizzativi. Nella strategia di AWS migrazione, questo framework si chiama accelerazione delle persone, a causa della velocità di cambiamento richiesta nei progetti di adozione del cloud. Per ulteriori informazioni, consultare la Guida OCM.

controllo dell'accesso all'origine (OAC)

In CloudFront, un'opzione avanzata per limitare l'accesso per proteggere i contenuti di Amazon Simple Storage Service (Amazon S3). OAC supporta tutti i bucket S3 in generale Regioni AWS, la crittografia lato server con AWS KMS (SSE-KMS) e le richieste dinamiche e dirette al bucket S3. `PUT` `DELETE`

identità di accesso origine (OAI)

Nel CloudFront, un'opzione per limitare l'accesso per proteggere i tuoi contenuti Amazon S3. Quando usi OAI, CloudFront crea un principale con cui Amazon S3 può autenticarsi. I principali autenticati possono accedere ai contenuti in un bucket S3 solo tramite una distribuzione specifica. CloudFront Vedi anche OAC, che fornisce un controllo degli accessi più granulare e avanzato.

ORR

Vedi la revisione della prontezza operativa.

- NON

Vedi la tecnologia operativa.

VPC in uscita (egress)

In un'architettura AWS multi-account, un VPC che gestisce le connessioni di rete avviate dall'interno di un'applicazione. La AWS Security Reference Architecture consiglia di configurare l'account di rete con funzionalità in entrata, in uscita e di ispezione VPCs per proteggere l'interfaccia bidirezionale tra l'applicazione e Internet in generale.

P

limite delle autorizzazioni

Una policy di gestione IAM collegata ai principali IAM per impostare le autorizzazioni massime che l'utente o il ruolo possono avere. Per ulteriori informazioni, consulta Limiti delle autorizzazioni nella documentazione di IAM.

informazioni di identificazione personale (PII)

Informazioni che, se visualizzate direttamente o abbinate ad altri dati correlati, possono essere utilizzate per dedurre ragionevolmente l'identità di un individuo. Esempi di informazioni personali includono nomi, indirizzi e informazioni di contatto.

Informazioni che consentono l'identificazione personale degli utenti

Visualizza le informazioni di identificazione personale.

playbook

Una serie di passaggi predefiniti che raccolgono il lavoro associato alle migrazioni, come l'erogazione delle funzioni operative principali nel cloud. Un playbook può assumere la forma di script, runbook automatici o un riepilogo dei processi o dei passaggi necessari per gestire un ambiente modernizzato.

PLC

Vedi controllore logico programmabile.

PLM

Vedi la gestione del ciclo di vita del prodotto.

policy

Un oggetto in grado di definire le autorizzazioni (vedi politica basata sull'identità), specificare le condizioni di accesso (vedi politicabasata sulle risorse) o definire le autorizzazioni massime per tutti gli account di un'organizzazione in (vedi politica di controllo dei servizi). AWS Organizations

persistenza poliglotta

Scelta indipendente della tecnologia di archiviazione di dati di un microservizio in base ai modelli di accesso ai dati e ad altri requisiti. Se i microservizi utilizzano la stessa tecnologia di archiviazione di dati, possono incontrare problemi di implementazione o registrare prestazioni scadenti. I microservizi vengono implementati più facilmente e ottengono prestazioni e scalabilità migliori se utilizzano l'archivio dati più adatto alle loro esigenze. Per ulteriori informazioni, consulta la sezione Abilitazione della persistenza dei dati nei microservizi.

valutazione del portfolio

Un processo di scoperta, analisi e definizione delle priorità del portfolio di applicazioni per pianificare la migrazione. Per ulteriori informazioni, consulta la pagina Valutazione della preparazione alla migrazione.

predicate

Una condizione di interrogazione che restituisce o, in genere, si trova in una clausola `WHERE`. `true` `false`

predicato pushdown

Una tecnica di ottimizzazione delle query del database che filtra i dati della query prima del trasferimento. Ciò riduce la quantità di dati che devono essere recuperati ed elaborati dal database relazionale e migliora le prestazioni delle query.

controllo preventivo

Un controllo di sicurezza progettato per impedire il verificarsi di un evento. Questi controlli sono la prima linea di difesa per impedire accessi non autorizzati o modifiche indesiderate alla rete. Per ulteriori informazioni, consulta Controlli preventivi in Implementazione dei controlli di sicurezza in AWS.

principale

Un'entità in AWS grado di eseguire azioni e accedere alle risorse. Questa entità è in genere un utente root per un Account AWS ruolo IAM o un utente. Per ulteriori informazioni, consulta Principali in Termini e concetti dei ruoli nella documentazione di IAM.

privacy fin dalla progettazione

Un approccio di ingegneria dei sistemi che tiene conto della privacy durante l'intero processo di sviluppo.

zone ospitate private

Un contenitore che contiene informazioni su come desideri che Amazon Route 53 risponda alle query DNS per un dominio e i relativi sottodomini all'interno di uno o più. VPCs Per ulteriori informazioni, consulta Utilizzo delle zone ospitate private nella documentazione di Route 53.

controllo proattivo

Un controllo di sicurezza progettato per impedire l'implementazione di risorse non conformi. Questi controlli analizzano le risorse prima del loro provisioning. Se la risorsa non è conforme al controllo, non viene fornita. Per ulteriori informazioni, consulta la guida di riferimento sui controlli nella AWS Control Tower documentazione e consulta Controlli proattivi in Implementazione dei controlli di sicurezza su. AWS

gestione del ciclo di vita del prodotto (PLM)

La gestione dei dati e dei processi di un prodotto durante l'intero ciclo di vita, dalla progettazione, sviluppo e lancio, attraverso la crescita e la maturità, fino al declino e alla rimozione.

Ambiente di produzione

Vedi ambiente.

controllore logico programmabile (PLC)

Nella produzione, un computer altamente affidabile e adattabile che monitora le macchine e automatizza i processi di produzione.

concatenamento rapido

Utilizzo dell'output di un prompt LLM come input per il prompt successivo per generare risposte migliori. Questa tecnica viene utilizzata per suddividere un'attività complessa in sottoattività o per perfezionare o espandere iterativamente una risposta preliminare. Aiuta a migliorare l'accuratezza e la pertinenza delle risposte di un modello e consente risultati più granulari e personalizzati.

pseudonimizzazione

Il processo di sostituzione degli identificatori personali in un set di dati con valori segnaposto. La pseudonimizzazione può aiutare a proteggere la privacy personale. I dati pseudonimizzati sono ancora considerati dati personali.

publish/subscribe (pub/sub)

Un modello che consente comunicazioni asincrone tra microservizi per migliorare la scalabilità e la reattività. Ad esempio, in un MES basato su microservizi, un microservizio può pubblicare

messaggi di eventi su un canale a cui altri microservizi possono abbonarsi. Il sistema può aggiungere nuovi microservizi senza modificare il servizio di pubblicazione.

Q

Piano di query

Una serie di passaggi, come le istruzioni, utilizzati per accedere ai dati in un sistema di database relazionale SQL.

regressione del piano di query

Quando un ottimizzatore del servizio di database sceglie un piano non ottimale rispetto a prima di una determinata modifica all'ambiente di database. Questo può essere causato da modifiche a statistiche, vincoli, impostazioni dell'ambiente, associazioni dei parametri di query e aggiornamenti al motore di database.

R

Matrice RACI

Vedi responsabile, responsabile, consultato, informato (RACI).

STRACCIO

Vedi Retrieval Augmented Generation.

ransomware

Un software dannoso progettato per bloccare l'accesso a un sistema informatico o ai dati fino a quando non viene effettuato un pagamento.

Matrice RASCI

Vedi responsabile, responsabile, consultato, informato (RACI).

RCAC

Vedi controllo dell'accesso a righe e colonne.

replica di lettura

Una copia di un database utilizzata per scopi di sola lettura. È possibile indirizzare le query alla replica di lettura per ridurre il carico sul database principale.

riprogettare

Vedi 7 Rs.

obiettivo del punto di ripristino (RPO)

Il periodo di tempo massimo accettabile dall'ultimo punto di ripristino dei dati. Questo determina ciò che si considera una perdita di dati accettabile tra l'ultimo punto di ripristino e l'interruzione del servizio.

obiettivo del tempo di ripristino (RTO)

Il ritardo massimo accettabile tra l'interruzione del servizio e il ripristino del servizio.

rifattorizzare

Vedi 7 R.

Regione

Una raccolta di AWS risorse in un'area geografica. Ciascuna Regione AWS è isolata e indipendente dalle altre per fornire tolleranza agli errori, stabilità e resilienza. Per ulteriori informazioni, consulta Specificare cosa può usare Regioni AWS il tuo account.

regressione

Una tecnica di ML che prevede un valore numerico. Ad esempio, per risolvere il problema "A che prezzo verrà venduta questa casa?" un modello di ML potrebbe utilizzare un modello di regressione lineare per prevedere il prezzo di vendita di una casa sulla base di dati noti sulla casa (ad esempio, la metratura).

riospitare

Vedi 7 R.

rilascio

In un processo di implementazione, l'atto di promuovere modifiche a un ambiente di produzione.

trasferisco

Vedi 7 Rs.

ripiattaforma

Vedi 7 Rs.

riacquisto

Vedi 7 Rs.

resilienza

La capacità di un'applicazione di resistere o ripristinare le interruzioni. L'elevata disponibilità e il disaster recovery sono considerazioni comuni quando si pianifica la resilienza in. Cloud AWSPer ulteriori informazioni, vedere Cloud AWS Resilience.

policy basata su risorse

Una policy associata a una risorsa, ad esempio un bucket Amazon S3, un endpoint o una chiave di crittografia. Questo tipo di policy specifica a quali principali è consentito l'accesso, le azioni supportate e qualsiasi altra condizione che deve essere soddisfatta.

matrice di assegnazione di responsabilità (RACI)

Una matrice che definisce i ruoli e le responsabilità di tutte le parti coinvolte nelle attività di migrazione e nelle operazioni cloud. Il nome della matrice deriva dai tipi di responsabilità definiti nella matrice: responsabile (R), responsabile (A), consultato (C) e informato (I). Il tipo di supporto (S) è facoltativo. Se includi il supporto, la matrice viene chiamata matrice RASCI e, se la escludi, viene chiamata matrice RACI.

controllo reattivo

Un controllo di sicurezza progettato per favorire la correzione di eventi avversi o deviazioni dalla baseline di sicurezza. Per ulteriori informazioni, consulta Controlli reattivi in Implementazione dei controlli di sicurezza in AWS.

retain

Vedi 7 R.

andare in pensione

Vedi 7 Rs.

Retrieval Augmented Generation (RAG)

Una tecnologia di intelligenza artificiale generativa in cui un LLM fa riferimento a una fonte di dati autorevole esterna alle sue fonti di dati di formazione prima di generare una risposta. Ad esempio, un modello RAG potrebbe eseguire una ricerca semantica nella knowledge base o nei dati personalizzati di un'organizzazione. Per ulteriori informazioni, consulta Cos'è il RAG.

rotazione

Processo di aggiornamento periodico di un segreto per rendere più difficile l'accesso alle credenziali da parte di un utente malintenzionato.

controllo dell'accesso a righe e colonne (RCAC)

L'uso di espressioni SQL di base e flessibili con regole di accesso definite. RCAC è costituito da autorizzazioni di riga e maschere di colonna.

RPO

Vedi l'obiettivo del punto di ripristino.

RTO

Vedi l'obiettivo del tempo di ripristino.

runbook

Un insieme di procedure manuali o automatizzate necessarie per eseguire un'attività specifica. In genere sono progettati per semplificare operazioni o procedure ripetitive con tassi di errore elevati.

S

SAML 2.0

Uno standard aperto utilizzato da molti provider di identità (IdPs). Questa funzionalità abilita il single sign-on (SSO) federato, in modo che gli utenti possano accedere AWS Management Console o chiamare le operazioni AWS API senza che tu debba creare un utente in IAM per tutti i membri dell'organizzazione. Per ulteriori informazioni sulla federazione basata su SAML 2.0, consulta Informazioni sulla federazione basata su SAML 2.0 nella documentazione di IAM.

SCADA

Vedi controllo di supervisione e acquisizione dati.

SCP

Vedi la politica di controllo del servizio.

Secret

In AWS Secrets Manager, informazioni riservate o riservate, come una password o le credenziali utente, archiviate in forma crittografata. È costituito dal valore segreto e dai relativi metadati. Il

valore segreto può essere binario, una stringa singola o più stringhe. Per ulteriori informazioni, consulta Cosa c'è in un segreto di Secrets Manager? nella documentazione di Secrets Manager.

sicurezza fin dalla progettazione

Un approccio di ingegneria dei sistemi che tiene conto della sicurezza durante l'intero processo di sviluppo.

controllo di sicurezza

Un guardrail tecnico o amministrativo che impedisce, rileva o riduce la capacità di un autore di minacce di sfruttare una vulnerabilità di sicurezza. Esistono quattro tipi principali di controlli di sicurezza: preventivi, investigativi, reattivi e proattivi.

rafforzamento della sicurezza

Il processo di riduzione della superficie di attacco per renderla più resistente agli attacchi. Può includere azioni come la rimozione di risorse che non sono più necessarie, l'implementazione di best practice di sicurezza che prevedono la concessione del privilegio minimo o la disattivazione di funzionalità non necessarie nei file di configurazione.

sistema di gestione delle informazioni e degli eventi di sicurezza (SIEM)

Strumenti e servizi che combinano sistemi di gestione delle informazioni di sicurezza (SIM) e sistemi di gestione degli eventi di sicurezza (SEM). Un sistema SIEM raccoglie, monitora e analizza i dati da server, reti, dispositivi e altre fonti per rilevare minacce e violazioni della sicurezza e generare avvisi.

automazione della risposta alla sicurezza

Un'azione predefinita e programmata progettata per rispondere o porre rimedio automaticamente a un evento di sicurezza. Queste automazioni fungono da controlli di sicurezza investigativi o reattivi che aiutano a implementare le migliori pratiche di sicurezza. AWS Esempi di azioni di risposta automatizzate includono la modifica di un gruppo di sicurezza VPC, l'applicazione di patch a un'istanza EC2 Amazon o la rotazione delle credenziali.

Crittografia lato server

Crittografia dei dati a destinazione, da parte di chi li riceve. Servizio AWS

Policy di controllo dei servizi (SCP)

Una politica che fornisce il controllo centralizzato sulle autorizzazioni per tutti gli account di un'organizzazione in. AWS Organizations SCPs definire barriere o fissare limiti alle azioni

che un amministratore può delegare a utenti o ruoli. È possibile utilizzarli SCPs come elenchi consentiti o elenchi di rifiuto, per specificare quali servizi o azioni sono consentiti o proibiti. Per ulteriori informazioni, consulta le politiche di controllo del servizio nella AWS Organizations documentazione.

endpoint del servizio

L'URL del punto di ingresso per un Servizio AWS. Puoi utilizzare l'endpoint per connetterti a livello di programmazione al servizio di destinazione. Per ulteriori informazioni, consulta Endpoint del Servizio AWS nei Riferimenti generali di AWS.

accordo sul livello di servizio (SLA)

Un accordo che chiarisce ciò che un team IT promette di offrire ai propri clienti, ad esempio l'operatività e le prestazioni del servizio.

indicatore del livello di servizio (SLI)

Misurazione di un aspetto prestazionale di un servizio, ad esempio il tasso di errore, la disponibilità o la velocità effettiva.

obiettivo a livello di servizio (SLO)

Una metrica target che rappresenta lo stato di un servizio, misurato da un indicatore del livello di servizio.

Modello di responsabilità condivisa

Un modello che descrive la responsabilità condivisa AWS per la sicurezza e la conformità del cloud. AWS è responsabile della sicurezza del cloud, mentre tu sei responsabile della sicurezza nel cloud. Per ulteriori informazioni, consulta Modello di responsabilità condivisa.

SIEM

Vedi il sistema di gestione delle informazioni e degli eventi sulla sicurezza.

punto di errore singolo (SPOF)

Un guasto in un singolo componente critico di un'applicazione che può disturbare il sistema.

SLAM

Vedi il contratto sul livello di servizio.

SLI

Vedi l'indicatore del livello di servizio.

LENTA

Vedi obiettivo del livello di servizio.

split-and-seed modello

Un modello per dimensionare e accelerare i progetti di modernizzazione. Man mano che vengono definite nuove funzionalità e versioni dei prodotti, il team principale si divide per creare nuovi team di prodotto. Questo aiuta a dimensionare le capacità e i servizi dell'organizzazione, migliora la produttività degli sviluppatori e supporta una rapida innovazione. Per ulteriori informazioni, vedere Approccio graduale alla modernizzazione delle applicazioni in. Cloud AWS

SPOF

Vedi punto di errore singolo.

schema a stella

Una struttura organizzativa di database che utilizza un'unica tabella dei fatti di grandi dimensioni per archiviare i dati transazionali o misurati e utilizza una o più tabelle dimensionali più piccole per memorizzare gli attributi dei dati. Questa struttura è progettata per l'uso in un data warehouse o per scopi di business intelligence.

modello del fico strangolatore

Un approccio alla modernizzazione dei sistemi monolitici mediante la riscrittura e la sostituzione incrementali delle funzionalità del sistema fino alla disattivazione del sistema legacy. Questo modello utilizza l'analogia di una pianta di fico che cresce fino a diventare un albero robusto e alla fine annienta e sostituisce il suo ospite. Il modello è stato introdotto da Martin Fowler come metodo per gestire il rischio durante la riscrittura di sistemi monolitici. Per un esempio di come applicare questo modello, consulta Modernizzazione incrementale dei servizi Web legacy di Microsoft ASP.NET (ASMX) mediante container e Gateway Amazon API.

sottorete

Un intervallo di indirizzi IP nel VPC. Una sottorete deve risiedere in una singola zona di disponibilità.

controllo di supervisione e acquisizione dati (SCADA)

Nella produzione, un sistema che utilizza hardware e software per monitorare gli asset fisici e le operazioni di produzione.

crittografia simmetrica

Un algoritmo di crittografia che utilizza la stessa chiave per crittografare e decrittografare i dati.

test sintetici

Test di un sistema in modo da simulare le interazioni degli utenti per rilevare potenziali problemi o monitorare le prestazioni. Puoi usare Amazon CloudWatch Synthetics per creare questi test.

prompt di sistema

Una tecnica per fornire contesto, istruzioni o linee guida a un LLM per indirizzarne il comportamento. I prompt di sistema aiutano a impostare il contesto e stabilire regole per le interazioni con gli utenti.

T

tags

Coppie chiave-valore che fungono da metadati per l'organizzazione delle risorse. AWS Con i tag è possibile a gestire, identificare, organizzare, cercare e filtrare le risorse. Per ulteriori informazioni, consulta Tagging delle risorse AWS.

variabile di destinazione

Il valore che stai cercando di prevedere nel machine learning supervisionato. Questo è indicato anche come variabile di risultato. Ad esempio, in un ambiente di produzione la variabile di destinazione potrebbe essere un difetto del prodotto.

elenco di attività

Uno strumento che viene utilizzato per tenere traccia dei progressi tramite un runbook. Un elenco di attività contiene una panoramica del runbook e un elenco di attività generali da completare. Per ogni attività generale, include la quantità stimata di tempo richiesta, il proprietario e lo stato di avanzamento.

Ambiente di test

Vedi ambiente.

training

Fornire dati da cui trarre ispirazione dal modello di machine learning. I dati di training devono contenere la risposta corretta. L'algoritmo di apprendimento trova nei dati di addestramento i pattern che mappano gli attributi dei dati di input al target (la risposta che si desidera prevedere). Produce un modello di ML che acquisisce questi modelli. Puoi quindi utilizzare il modello di ML per creare previsioni su nuovi dati di cui non si conosce il target.

Transit Gateway

Un hub di transito di rete che puoi utilizzare per interconnettere le tue reti VPCs e quelle locali. Per ulteriori informazioni, consulta Cos'è un gateway di transito nella AWS Transit Gateway documentazione.

flusso di lavoro basato su trunk

Un approccio in cui gli sviluppatori creano e testano le funzionalità localmente in un ramo di funzionalità e quindi uniscono tali modifiche al ramo principale. Il ramo principale viene quindi integrato negli ambienti di sviluppo, preproduzione e produzione, in sequenza.

Accesso attendibile

Concessione delle autorizzazioni a un servizio specificato dall'utente per eseguire attività all'interno dell'organizzazione AWS Organizations e nei suoi account per conto dell'utente. Il servizio attendibile crea un ruolo collegato al servizio in ogni account, quando tale ruolo è necessario, per eseguire attività di gestione per conto dell'utente. Per ulteriori informazioni, consulta Utilizzo AWS Organizations con altri AWS servizi nella AWS Organizations documentazione.

regolazione

Modificare alcuni aspetti del processo di training per migliorare la precisione del modello di ML. Ad esempio, puoi addestrare il modello di ML generando un set di etichette, aggiungendo etichette e quindi ripetendo questi passaggi più volte con impostazioni diverse per ottimizzare il modello.

team da due pizze

Una piccola DevOps squadra che puoi sfamare con due pizze. Un team composto da due persone garantisce la migliore opportunità possibile di collaborazione nello sviluppo del software.

U

incertezza

Un concetto che si riferisce a informazioni imprecise, incomplete o sconosciute che possono minare l'affidabilità dei modelli di machine learning predittivi. Esistono due tipi di incertezza: l'incertezza epistemica, che è causata da dati limitati e incompleti, mentre l'incertezza aleatoria è causata dal rumore e dalla casualità insiti nei dati. Per ulteriori informazioni, consulta la guida Quantificazione dell'incertezza nei sistemi di deep learning.

compiti indifferenziati

Conosciuto anche come sollevamento di carichi pesanti, è un lavoro necessario per creare e far funzionare un'applicazione, ma che non apporta valore diretto all'utente finale né offre vantaggi competitivi. Esempi di attività indifferenziate includono l'approvvigionamento, la manutenzione e la pianificazione della capacità.

ambienti superiori

Vedi ambiente.

V

vacuum

Un'operazione di manutenzione del database che prevede la pulizia dopo aggiornamenti incrementali per recuperare lo spazio di archiviazione e migliorare le prestazioni.

controllo delle versioni

Processi e strumenti che tengono traccia delle modifiche, ad esempio le modifiche al codice di origine in un repository.

Peering VPC

Una connessione tra due VPCs che consente di indirizzare il traffico utilizzando indirizzi IP privati. Per ulteriori informazioni, consulta Che cos'è il peering VPC? nella documentazione di Amazon VPC.

vulnerabilità

Un difetto software o hardware che compromette la sicurezza del sistema.

W

cache calda

Una cache del buffer che contiene dati correnti e pertinenti a cui si accede frequentemente. L'istanza di database può leggere dalla cache del buffer, il che richiede meno tempo rispetto alla lettura dalla memoria dal disco principale.

dati caldi

Dati a cui si accede raramente. Quando si eseguono interrogazioni di questo tipo di dati, in genere sono accettabili query moderatamente lente.

funzione finestra

Una funzione SQL che esegue un calcolo su un gruppo di righe che si riferiscono in qualche modo al record corrente. Le funzioni della finestra sono utili per l'elaborazione di attività, come il calcolo di una media mobile o l'accesso al valore delle righe in base alla posizione relativa della riga corrente.

Carico di lavoro

Una raccolta di risorse e codice che fornisce valore aziendale, ad esempio un'applicazione rivolta ai clienti o un processo back-end.

flusso di lavoro

Gruppi funzionali in un progetto di migrazione responsabili di una serie specifica di attività. Ogni flusso di lavoro è indipendente ma supporta gli altri flussi di lavoro del progetto. Ad esempio, il flusso di lavoro del portfolio è responsabile della definizione delle priorità delle applicazioni, della pianificazione delle ondate e della raccolta dei metadati di migrazione. Il flusso di lavoro del portfolio fornisce queste risorse al flusso di lavoro di migrazione, che quindi migra i server e le applicazioni.

VERME

Vedi scrivere una volta, leggere molti.

WQF

Vedi AWS Workload Qualification Framework.

scrivi una volta, leggi molte (WORM)

Un modello di storage che scrive i dati una sola volta e ne impedisce l'eliminazione o la modifica. Gli utenti autorizzati possono leggere i dati tutte le volte che è necessario, ma non possono modificarli. Questa infrastruttura di archiviazione dei dati è considerata immutabile.

Z

exploit zero-day

Un attacco, in genere malware, che sfrutta una vulnerabilità zero-day.

vulnerabilità zero-day

Un difetto o una vulnerabilità assoluta in un sistema di produzione. Gli autori delle minacce possono utilizzare questo tipo di vulnerabilità per attaccare il sistema. Gli sviluppatori vengono spesso a conoscenza della vulnerabilità causata dall'attacco.

prompt zero-shot

Fornire a un LLM le istruzioni per eseguire un'attività ma non esempi (immagini) che possano aiutarla. Il LLM deve utilizzare le sue conoscenze pre-addestrate per gestire l'attività. L'efficacia del prompt zero-shot dipende dalla complessità dell'attività e dalla qualità del prompt. Vedi anche few-shot prompting.

applicazione zombie

Un'applicazione che prevede un utilizzo CPU e memoria inferiore al 5%. In un progetto di migrazione, è normale ritirare queste applicazioni.

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.