Strategie Model Context Protocol su AWS

AWS Guida prescrittiva

AWS Guida prescrittiva: Strategie Model Context Protocol su AWS

Table of Contents

Strategie Model Context Protocol su AWS

Amazon Web Services (collaboratori)

Marzo 2026 (cronologia dei documenti)

Questa guida può aiutarti a sviluppare e implementare strategie MCP (Model Context Protocol) in tutta l'organizzazione per supportare il tuo percorso verso l'intelligenza artificiale agentica. Poiché gli agenti e i modelli linguistici diventano sempre più centrali nelle operazioni aziendali, la definizione di una strategia MCP è fondamentale per soluzioni agentiche di successo.

Questa guida esplora tre pilastri fondamentali per la creazione di una strategia MCP: progettazione di strumenti MCP, hosting di server MCP e governance MCP. Affrontando questi componenti interconnessi, le organizzazioni possono creare sistemi scalabili, sicuri ed efficaci per la gestione del contesto del modello in tutte le loro implementazioni di intelligenza artificiale. Queste linee guida forniscono informazioni utili e linee guida strategiche per le organizzazioni in qualsiasi fase del percorso di intelligenza artificiale di un'organizzazione, dalla sperimentazione iniziale alle implementazioni di produzione su vasta scala. Questo le aiuta a sviluppare soluzioni MCP personalizzate in linea con le loro esigenze e obiettivi specifici.

Queste best practice derivano da implementazioni reali di organizzazioni che implementano MCP su scala aziendale, da un'analisi degli attuali standard di specifica MCP e dalle lezioni apprese dalle applicazioni personalizzate Large Language Model (LLM) in produzione.

I sistemi di intelligenza artificiale utilizzano sistemi sempre più sofisticati e robusti LLMs in un'ampia varietà di casi d'uso. LLMs eccellono nella comprensione del linguaggio naturale, nella generazione di risposte simili a quelle umane e nel ragionamento su informazioni complesse. Tuttavia, per passare LLMs da interfacce conversazionali a sistemi in grado di svolgere in modo autonomo attività complesse, le organizzazioni stanno adottando architetture di intelligenza artificiale agentiche, sistemi di intelligenza artificiale in grado di percepire l'ambiente, ragionare sugli obiettivi, prendere decisioni autonome, orchestrare in più fasi e intraprendere azioni per raggiungere gli obiettivi per conto degli utenti. Questo approccio agentico aiuta le organizzazioni a creare sistemi di intelligenza artificiale in grado di comprendere l'intento dell'utente attraverso il linguaggio naturale, coordinarsi autonomamente tra più fonti di dati e strumenti e fornire esperienze personalizzate su una scala che non era possibile con i tradizionali modelli di richiesta-risposta. Per rendere questi agenti più capaci, le organizzazioni devono fornire l'accesso agli strumenti e ai dati esistenti per arricchire la comprensione contestuale dell'agente e consentirgli di agire per conto dell'utente.

MCP fornisce un protocollo standardizzato per l'integrazione di strumenti di intelligenza artificiale, che consente una comunicazione coerente tra agenti e risorse esterne. Sebbene MCP stesso definisca lo standard di comunicazione, la sua implementazione efficace richiede un'attenta considerazione dei modelli architettonici, dei modelli di sicurezza, delle pratiche operative e delle strategie di ottimizzazione delle prestazioni per ottenere soluzioni scalabili, sicure e gestibili.

Questa guida sintetizza le lezioni apprese dalle implementazioni MCP aziendali, fornendo consigli pratici in linea con il Well-Architected Framework.AWS Descrive le strategie per la progettazione di strumenti MCP, l'hosting di server MCP e la governance MCP, essenziali per creare soluzioni MCP personalizzate. Le raccomandazioni contenute in questa guida si riferiscono ai seguenti cinque pilastri del AWS Well-Architected Framework:

- Sicurezza: isolamento dei token, credenziali ridotte, autorizzazione separata read/write
- Eccellenza operativa: metriche di precisione nella selezione degli strumenti, set di dati eccezionali per i test di regressione
- Affidabilità: limitazione della velocità per utente e per utensile, riduzione del carico
- Efficienza delle prestazioni: strumenti basati sul flusso di lavoro, filtraggio degli strumenti, ricerca semantica per ridurre l'utilizzo delle finestre contestuali
- Ottimizzazione dei costi: server MCP riutilizzabili tra i team, riduzione dei costi dei token per richiesta grazie al filtraggio degli strumenti

Destinatari principali

Questa guida è destinata ad architetti, sviluppatori e leader tecnologici che implementano soluzioni di intelligenza artificiale agentica nelle loro organizzazioni. Per comprendere i concetti di questa guida, è necessario comprendere come LLMs funziona e avere conoscenze di base su MCP, strumenti e progettazione rapida.

Obiettivi

Costruire sistemi di intelligenza artificiale Agentic pronti per la produzione significa risolvere insieme i problemi di governance, ottimizzazione e sicurezza per supportare le politiche dell'organizzazione. Di seguito viene spiegato in che modo questa guida affronta questi obiettivi:

- Governance: senza una governance centralizzata, non è possibile rispondere alle domande di audit sui carichi di lavoro di intelligenza artificiale, tra cui quali agenti hanno avuto accesso a quali

dati, con quali autorizzazioni e quando. Inoltre, non è possibile imporre il controllo delle versioni. La sezione sulla strategia di hosting MCP di questa guida spiega come gli utenti potrebbero utilizzare server MCP locali obsoleti con vulnerabilità note a causa della mancanza di applicazioni sistematiche.

Per i settori regolamentati, la governance è fondamentale. I revisori vogliono vedere l'applicazione delle politiche e il monitoraggio dell'utilizzo degli strumenti tra tutti gli agenti da un unico pannello. La governance MCP lo prevede.

Seguendo i consigli di questa guida, è possibile migliorare la precisione delle attività del 28-32% nei benchmark sottoposti a revisione paritaria. Per ulteriori informazioni, consulta MARCO: Multi-Agent Real-time Chat Orchestration (sito web ACL Anthology). La governance non riguarda solo la conformità, ma migliora anche le prestazioni del sistema di intelligenza artificiale agentica.

- Ottimizzazione: i tuoi team potrebbero creare le stesse integrazioni più di una volta. Ad esempio, quando cinque team diversi scrivono il proprio script di interrogazione del database per consentire alla loro applicazione di intelligenza artificiale di comunicare con i propri database, si tratta di cinque volte il costo di sviluppo e cinque serie di bug da gestire. MCP ti consente di crearlo una sola volta e condividerlo con l'intera comunità di ingegneri. I risparmi aumentano man mano che aumenta il numero di agenti.

  C'è anche un problema di costo per richiesta che la maggior parte dei team non nota all'inizio. Ogni definizione di strumento utilizza i token della finestra contestuale. Con 20 strumenti, spendi 5.000-10.000 token per invocazione solo per le descrizioni, oltre alle richieste degli utenti. Ciò aumenta la latenza e i costi di inferenza LLM e riduce la precisione poiché il modello fatica a scegliere lo strumento giusto dall'elenco degli strumenti disponibili.

  Gli agenti che utilizzano strumenti wrapper strutturati sono circa tre volte più accurati nelle attività del database rispetto agli agenti che accedono APIs direttamente (per ulteriori informazioni, vedere Middleware for LLMs: Gli strumenti sono strumentali per gli agenti linguistici in ambienti complessi). Il modo in cui si progettano e si presentano gli strumenti per un modello di intelligenza artificiale è importante. Questa guida consiglia di fornire agli strumenti schemi chiari, di applicarli ai flussi di lavoro effettivi anziché agli endpoint grezzi e di limitare le informazioni nella finestra contestuale. La sezione sulla strategia di progettazione degli strumenti MCP di questa guida approfondisce questi aspetti.
- Sicurezza e conformità: immaginate un sistema di intelligenza artificiale agentico che allucina una fase di pulizia e tenta di eliminare un database di produzione. Se l'agente ha ereditato le credenziali di amministratore complete dell'utente, l'eliminazione potrebbe andare a buon fine. Con

l'isolamento dei token e le credenziali ridotte che garantiscono solo l'accesso in lettura e creazione, l'operazione fallisce in modo sicuro.

I flussi di lavoro regolamentati migliorano ulteriormente questa situazione. La guida fornisce alcuni esempi (pipeline sanitarie che richiedono la convalida HIPAA e l'anonimizzazione delle informazioni di identificazione personale prima di elaborare i dati dei pazienti). L'integrazione di tale logica negli strumenti MCP significa che la conformità avviene in modo deterministico ogni volta.

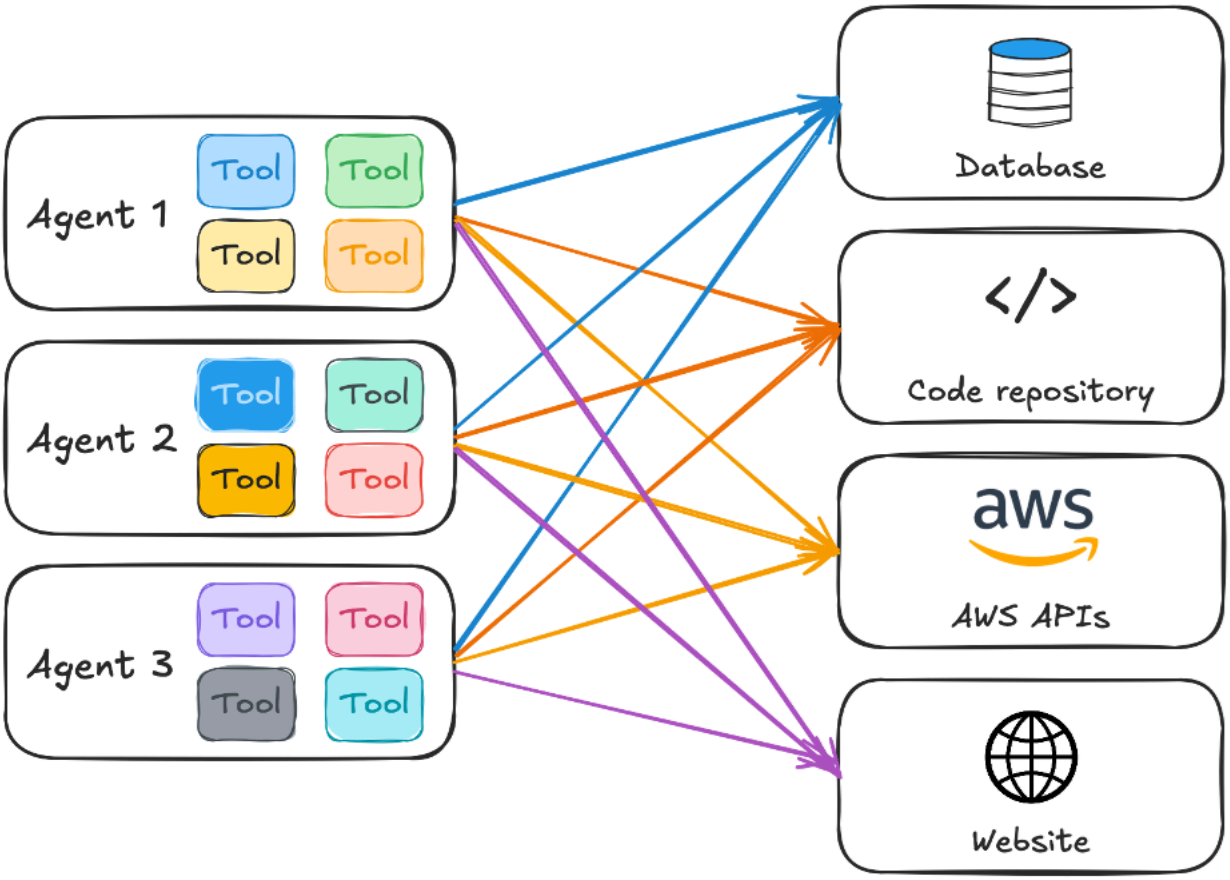
Che cos'è l'MCP?

LLMs lavorano prevedendo una risposta a un prompt in base ai dati di addestramento. Ciò significa che l'LLM può fornire solo risposte su dati ed eventi che ha già visto. Metodi come Retrieval Augmented Generation (RAG) e le knowledge base consentono di includere dati contestuali. Tuttavia, se chiedeste a un LLM quali saranno le previsioni del tempo di domani o quanti clienti ci sono nel vostro database, probabilmente vi sarebbero allucinazioni o non sarebbe in grado di fornire una risposta, perché queste non rientrano nelle conoscenze pre-addestrate del LLM. Per poter rispondere a questo tipo di domande, un agente deve accedere a funzionalità e dati esterni e al di APIs fuori del contesto nativo del LLM.

Strumenti di comprensione

Possiamo fornire all'LLM l'accesso a sistemi e contesti aggiuntivi tramite strumenti.Gli strumenti sono funzioni fornite al LLM per raggiungere un obiettivo chiaro. Uno strumento può richiamare un'API, interrogare un database, eseguire operazioni con la calcolatrice, gestire una sandbox di codice, eseguire una ricerca sul Web e persino richiamare un altro sistema di intelligenza artificiale oppure. agent-as-a-tool Ogni strumento dovrebbe includere una descrizione che indichi all'LLM cosa fa lo strumento, quando usarlo e quali parametri accetta. Ciò consente all'LLM di prendere decisioni dettagliate su quale strumento o combinazione di strumenti invocare in base all'input dell'utente. L'LLM viene informato sugli strumenti a disposizione dell'agente, consentendogli di generare risposte che indicano all'agente di richiamare lo strumento. Ad esempio, quando chiedi all'LLM quanti clienti ci sono nel tuo database, l'LLM invierà una risposta all'agente richiedendo di eseguire lo strumento con parametri di input specifici. `query_database` L'LLM determina quale strumento invocare e gli input per la chiamata allo strumento. L'agente esegue quindi lo strumento, che converte l'input in linguaggio naturale in una chiamata di funzione sintatticamente corretta ed esegue la query. L'agente richiama lo strumento o gli strumenti in base alle istruzioni del LLM e tali risultati vengono restituiti all'LLM. Ciò sfrutta la capacità dell'LLM di ragionare utilizzando input testuali e di selezionare gli strumenti appropriati per il lavoro.

L'immagine seguente mostra come ogni agente gestisce il proprio set di strumenti per ogni destinazione.



La scalabilità dell'accesso agli strumenti può presentare sfide per le soluzioni di intelligenza artificiale agentica:

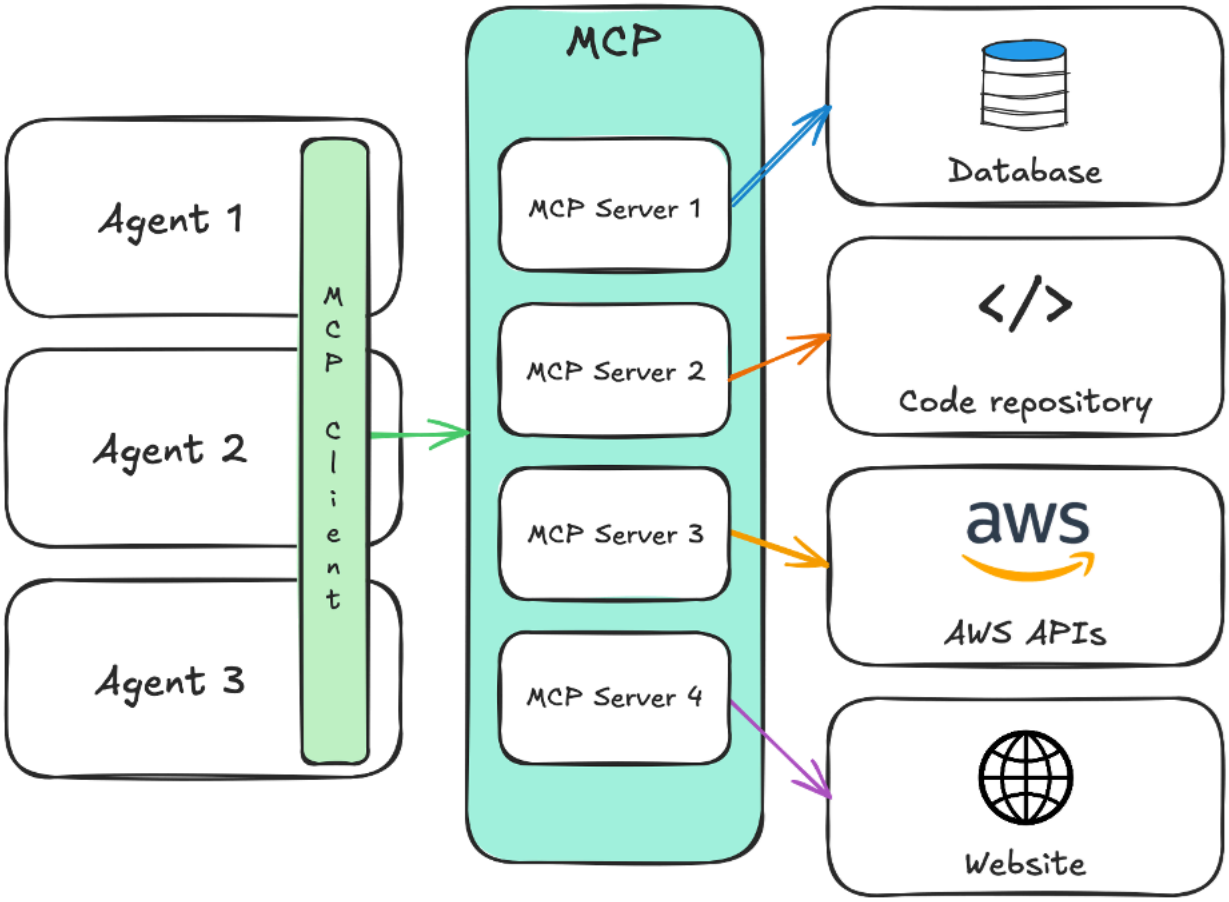
- Se ogni sviluppatore crea il proprio strumento per le stesse funzionalità esterne, si assiste a molti sforzi duplicati e a modi non standardizzati di interagire con queste funzionalità esterne. Ciò produce implementazioni incoerenti tra i vostri agenti. Sebbene sia possibile risolvere il problema sviluppando strumenti standard nelle librerie e distribuendoli, ciò manca di una governance centralizzata. Ciò rende difficile applicare le politiche di sicurezza, tenere traccia dell'utilizzo degli strumenti, gestire il controllo delle versioni tra i team o garantire la conformità agli standard organizzativi. Inoltre, quando incorpori gli strumenti direttamente con l'agente, devi ridistribuirlo ogni volta che viene creato un nuovo strumento o ne viene aggiornato uno esistente.
- La fornitura di strumenti a un LLM utilizza la relativa finestra contestuale. La finestra contestuale è il numero di token (unità di testo LLMs elaborate, che in genere rappresentano parole, parti di parole o punteggiatura) che un modello può considerare in qualsiasi momento. LLMs hanno dei limiti per una finestra contestuale. Gli strumenti e la relativa documentazione utilizzano quella finestra a contesto limitato insieme ai prompt di sistema e ai prompt degli utenti. Man mano che la finestra contestuale si riempie, le prestazioni LLMs possono peggiorare a causa di molteplici

fattori: difficoltà nell'identificare le informazioni pertinenti, maggiore complessità di elaborazione e ridotta capacità di ragionamento. La sfida si aggrava quando le definizioni degli strumenti, i prompt di sistema e la cronologia delle conversazioni competono per lo spazio limitato nella finestra contestuale, poiché vengono forniti in ogni chiamata LLM.

Pertanto, il numero di strumenti e il modo in cui vengono documentati hanno un impatto diretto sulle prestazioni del LLM, in termini di tempi di risposta e precisione.

MCP stabilisce uno standard universale per connettere gli agenti a funzionalità esterne. Viene comunemente definita «USB-C per applicazioni AI». Invece di registrare gli strumenti direttamente con gli agenti, i server MCP fungono da intermediari per ospitare strumenti che vengono scoperti e richiamati tramite JSON-RPC 2.0. Invece di aggiungere decine o centinaia di strumenti diversi all'agente e mantenerli nel tempo, MCP consente di registrare server MCP che racchiudono gli strumenti a cui l'agente può accedere. Questo approccio standardizza il modo in cui gli strumenti vengono impacchettati, presentati e richiamati. Questo può aiutare ad affrontare le sfide di scalabilità e governance legate all'utilizzo degli strumenti all'interno degli agenti. Inoltre, separa lo sviluppo e le operazioni degli agenti dagli strumenti utilizzati per le funzionalità esterne.

La figura seguente mostra gli agenti che utilizzano MCP per accedere a risorse esterne.



Tuttavia, lo standard MCP non risolve tutte le sfide di scalabilità e governance. L'implementazione dei server MCP deve essere combinata con strategie efficaci di progettazione degli strumenti, hosting e governance aziendale. Questa guida fornisce le migliori pratiche per ogni strategia per aiutarti a creare e utilizzare MCP come parte delle tue soluzioni di intelligenza artificiale agentica.

Quando usare MCP

MCP fornisce un'infrastruttura strategica per scalare le tue iniziative di intelligenza artificiale agentica. Centralizzando la gestione e la governance degli strumenti, i server MCP riducono il costo cumulativo della creazione e del mantenimento di integrazioni personalizzate tra più agenti. Ciò offre rendimenti crescenti man mano che l'ecosistema degli agenti si espande.

È probabile che MCP diventi parte della tua strategia quando:

- È necessaria una governance centralizzata per il modo in cui gli agenti accedono ai sistemi e ai servizi aziendali, come database APIs, strumenti interni e integrazioni di terze parti.

- Gli sviluppatori dedicano troppo tempo a scrivere integrazioni personalizzate che non sono coerenti tra le implementazioni.
- Disponi di strumenti duplicati che potrebbero offrire funzionalità comuni.
- Desiderate offrire i vostri strumenti o dati proprietari a consumatori esterni o sistemi di agenti di terze parti attraverso interfacce MCP standardizzate e gestite, sbloccando nuovi flussi di entrate mantenendo al contempo sicurezza e controllo.

Dopo aver deciso che i server MCP faranno parte della vostra strategia, valutate se le implementazioni dei server MCP open source esistenti soddisfano le vostre esigenze, se richiedono miglioramenti o se è necessario creare server personalizzati. Molte implementazioni predefinite di server MCP sono disponibili negli archivi pubblici e coprono funzionalità comuni come l'accesso al file system, la navigazione sul Web, le sandbox di codice, l'accesso al database e le integrazioni API.

In molti casi, i server MCP preesistenti sono sufficienti. Ad esempio, AWS fornisce il AWS MCP Server, un server MCP remoto gestito che fornisce agli assistenti e agli agenti di intelligenza artificiale un accesso sicuro e autenticato tramite interazioni in linguaggio naturale. Servizi AWS È possibile utilizzarlo AWS MCP Server per eseguire AWS attività complesse e in più fasi combinando accesso in tempo reale alla AWS documentazione, chiamate API sintatticamente corrette e flussi di lavoro predefiniti denominati Agent che seguono le migliori pratiche. SOPs AWS AWS li testa continuamente AWS MCP Server per assicurarsi che gli agenti del cliente possano utilizzarli con successo.

Dovresti testare questi server MCP esistenti con i tuoi agenti per determinare se soddisfano i tuoi casi d'uso. Se un agente non riesce a completare i flussi di lavoro, genera risposte errate o non ottimali, non riesce a gestire processi complessi in più fasi o non rispetta importanti best practice o considerazioni sulla sicurezza specifiche del dominio, dovreste prendere in considerazione miglioramenti in diverse dimensioni.

Quando i server MCP esistenti non soddisfano appieno le vostre esigenze e fanno fatica a utilizzare correttamente gli strumenti esistenti o a produrre risposte accurate, prendete in considerazione questi approcci di miglioramento prima di creare server personalizzati:

- Arricchisci il contesto degli agenti: se il tuo agente ha difficoltà a utilizzare in modo corretto o efficiente gli strumenti di un server MCP esistente, valuta la possibilità di integrare tali definizioni degli strumenti con documentazione o esempi aggiuntivi. Questo aiuta a fornire un contesto aggiuntivo all'LLM.

- Aggiungi strumenti complementari: estendi i server MCP esistenti con strumenti che accedono a dati o contesti organizzativi aggiuntivi di cui gli agenti hanno bisogno per completare con successo i flussi di lavoro.
- Migliora la funzionalità di base APIs: semplifica il servizio APIs per renderlo più adatto agli LLM riducendo la complessità dei parametri, fornendo messaggi di errore più chiari e offrendo impostazioni predefinite ragionevoli utilizzabili dagli agenti.

Sebbene l'utilizzo delle implementazioni di server MCP esistenti acceleri lo sviluppo di funzionalità comuni, la creazione di server MCP personalizzati è una necessità quando il caso d'uso richiede funzionalità specializzate. I server MCP personalizzati consentono di incapsulare le competenze del settore, applicare gli standard organizzativi, migliorare l'affidabilità degli agenti per flussi di lavoro complessi e supportare la conformità ai requisiti di sicurezza. Prendi in considerazione la creazione di un server MCP personalizzato nelle seguenti situazioni:

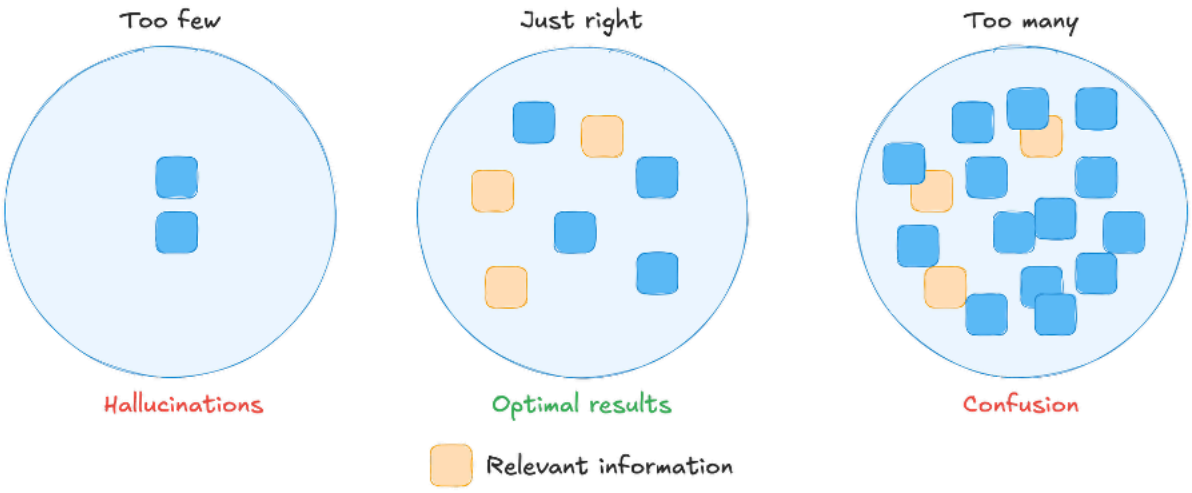
- Flussi di lavoro specifici del dominio: i flussi di lavoro in più fasi che richiedono competenze di settore devono essere incapsulati in strumenti MCP personalizzati quando le conoscenze necessarie non vengono acquisite nella documentazione delle API. Ad esempio, anziché consentire agli agenti di orchestrare complesse pipeline di dati sanitari che devono convalidare la conformità all'Health Insurance Portability and Accountability Act (HIPAA), rendere anonime le PII e passare al formato HL7 FHIR, offri uno strumento che incorpori direttamente l'esperienza del settore. `process_patient_data` Ciò elimina la dipendenza dall'LLM per orchestrare ed eseguire correttamente le fasi del flusso di lavoro, migliorando la coerenza e la conformità.
- Astrazioni del percorso privilegiato: gli agenti potrebbero avere difficoltà a implementare approcci ottimali perché mancano di contesto organizzativo e si basano su modelli di base piuttosto che sulle migliori pratiche organizzative. In questi scenari, è possibile applicare standard prescrittivi in termini di costi, prestazioni o sicurezza incapsulando questi percorsi fondamentali in strumenti MCP personalizzati. Ad esempio, anziché consentire agli agenti di implementare un'infrastruttura con impostazioni predefinite che potrebbero essere non sicure o inefficienti, fornite uno strumento che incorpori direttamente gli standard della vostra organizzazione. `deploy_secure_infrastructure`
- Orchestrazione multiservizio complessa: anziché far orchestrare all'agente flussi di lavoro complessi cercando di dedurre la sequenza e il set di servizi corretti da utilizzare in ogni fase, puoi creare tale logica in modo deterministico all'interno di uno strumento MCP. Potresti anche voler fornire competenze sui modelli ottimali di integrazione dei servizi di cui l'agente potrebbe non essere a conoscenza. Ciò può anche migliorare la precisione e l'efficienza dei tuoi agenti.

- Best practice specifiche per i servizi: si tratta di una prassi comune per gli strumenti incentrati sulla sicurezza che aiutano gli agenti a implementare politiche di crittografia, controlli di accesso e modelli di conformità specifici per il servizio a cui si accede tramite lo strumento dell'agente. Inoltre, se esistono best practice operative specifiche per un servizio che non sono ovvie, l'utilizzo di un server MCP può aiutarvi ad assicurarvi che vengano implementate e non lasciate che sia un agente a valutarle.

Strategia di progettazione degli strumenti MCP

Il compito principale del client e del server MCP è scoprire e presentare gli strumenti all'LLM in modo che possa utilizzarli per migliorare le sue risposte. Ciò rende la progettazione degli strumenti MCP una delle strategie più importanti per creare soluzioni MCP efficaci. Dal punto di vista del modello, gli strumenti sono una funzione che possono richiamare secondo necessità per fornire risposte più accurate e complete. L'interfaccia funzionale riassume l'implementazione sottostante di uno strumento, che può spaziare da un wrapper per una singola chiamata API a una logica di flusso di lavoro complessa.

Tuttavia, è necessario trovare un equilibrio con la quantità di strumenti forniti al LLM. Se gli strumenti sono troppo pochi, l'LLM potrebbe non essere in grado di raccogliere il contesto e le informazioni corretti, quindi baserà l'ipotesi migliore sulla base delle informazioni disponibili all'interno del modello. Se gli strumenti sono troppi, l'LLM potrebbe confondersi sulla scelta e sulla sequenza corrette degli strumenti, con conseguenti allucinazioni. Il tuo obiettivo è ottenere il numero giusto di strumenti. L'immagine seguente mostra le problematiche legate al numero insufficiente e al numero eccessivo di strumenti.



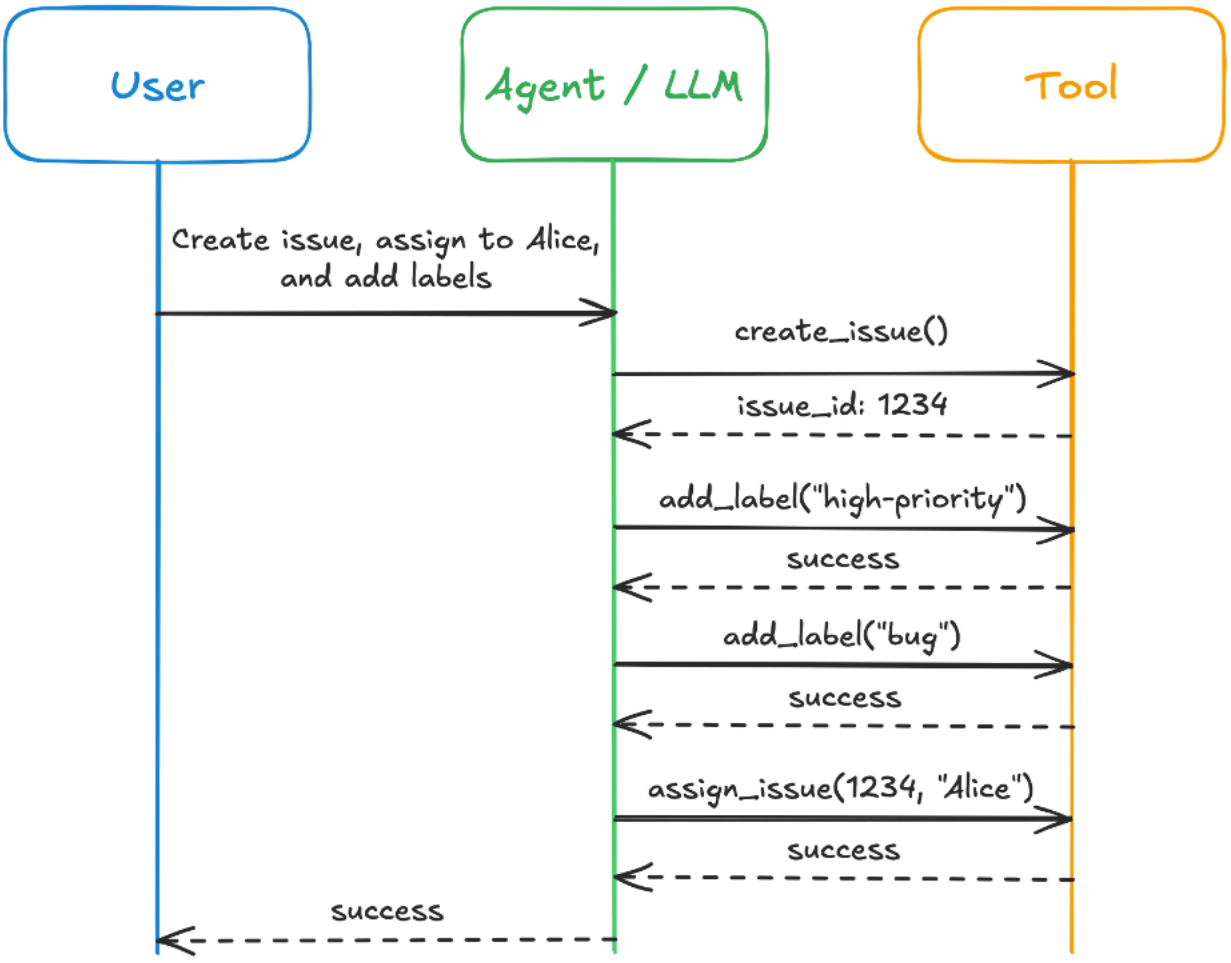
La soluzione richiede la comprensione del numero di strumenti da fornire e dell'ambito di ciascun strumento. La granularità degli strumenti, indipendentemente dal fatto che si riferiscano a singole chiamate API o a flussi di lavoro completi, influisce direttamente sul numero totale di strumenti di cui gli agenti hanno bisogno e sull'efficacia con cui possono utilizzarli. Questa sezione fornisce le migliori pratiche per la definizione degli strumenti MCP, la creazione di definizioni degli strumenti, la loro scoperta e la loro organizzazione.

Ambito dello strumento

Esistono due approcci per lo sviluppo di strumenti: granulari e a grana grossa.

Granulare

Con un approccio granulare, creeresti uno strumento per API, azione o query. Ad esempio, puoi crearecreate_issue, get_issue add_labelassign_issue, e close_issue strumenti per il tuo repository Git. Ciò consentirebbe all'LLM di effettuare chiamate granulari a ciascuna API e di orchestrarle ciascuna secondo necessità. Considerate la seguente richiesta: «Create un problema per il servizio di prodotto chiamato 'Query restituisce solo risultati parziali', etichettatelo come bug e con priorità alta e assegnatelo ad Alice». L'immagine seguente mostra come un tool-per-API approccio risponderebbe a questa richiesta.

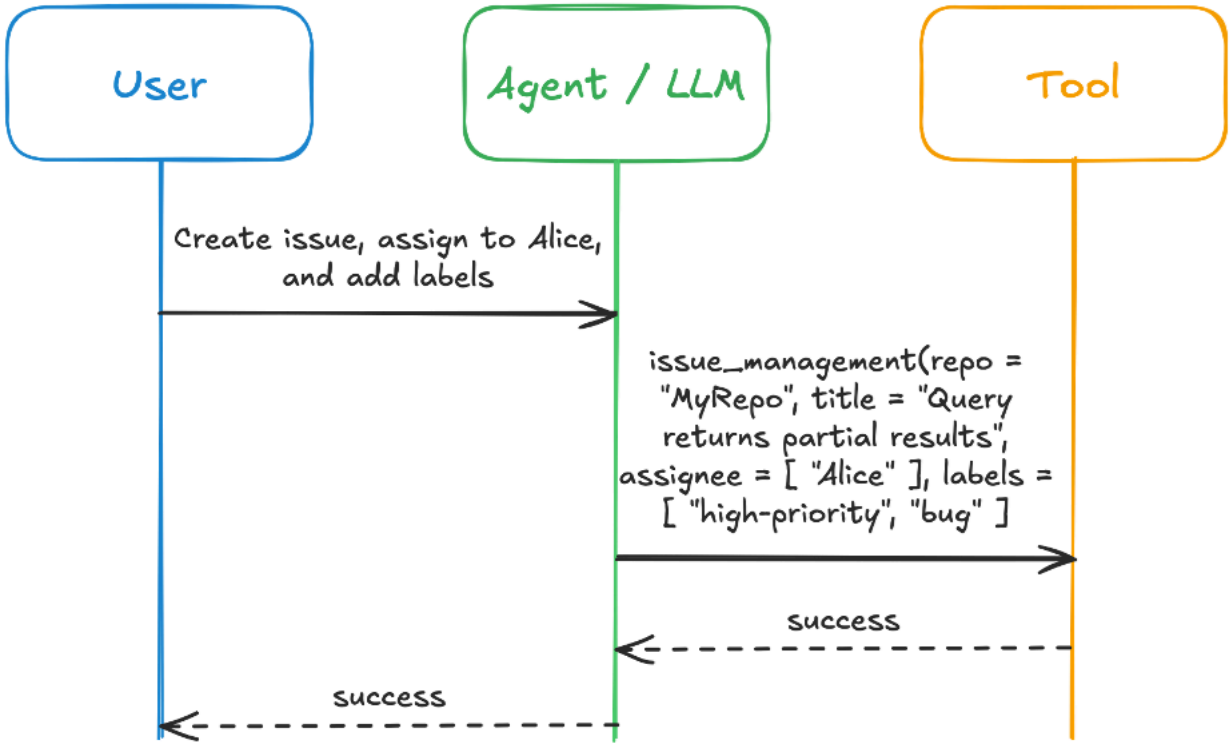


In questo approccio, il prompt di sistema e ogni definizione di strumento registrata vengono fornite all'LLM a ogni chiamata. Ciò consuma ulteriore contesto e comporta una penalità di latenza

perché ogni chiamata allo strumento rappresenta una singola chiamata al LLM. Inoltre, aumenta la complessità della gestione degli errori all'interno del flusso di lavoro.

A grana grossa

Un approccio a grana grossa, o basato sul flusso di lavoro, sarebbe costituito da strumenti orientati al flusso di lavoro. Lo strumento si concentra sull'intento dell'utente piuttosto che sulla struttura delle API. end-to-end Invece di a tool-per-API, hai uno strumento che ne chiama in modo deterministico molti. APIs Utilizzando il precedente esempio di repository Git, è possibile creare uno `create_and_setup_issue` strumento che viene chiamato una sola volta dall'agente. L'implementazione dello strumento crea il problema, aggiunge etichette e lo assegna a un utente, in base ai parametri forniti allo strumento. L'immagine seguente mostra come un approccio a grana grossa elaborerebbe lo stesso prompt.



Questo approccio mostra come tutta la complessità rimanga nascosta al livello LLM. Quando la logica di orchestrazione è incorporata nell'implementazione dello strumento, tutti i passaggi sequenziali, la registrazione, la logica dei tentativi, gli interruttori automatici e la limitazione della velocità vengono eseguiti in modo deterministico nello strumento. L'approccio basato sul flusso di lavoro rende più semplice per l'LLM richiamare lo strumento corretto con i parametri corretti. È importante notare che alcune API potrebbero già fornire l'intento del flusso di lavoro, come l'API Amazon `RunInstances` EC2. In questi casi, a tool-per-API potrebbe fornire il design orientato al flusso di lavoro che desideri.

Tuttavia, anche gli strumenti possono avere una grana troppo grossa. Se un unico strumento di flusso di lavoro tenta di fare troppe cose e ha molti parametri possibili, l'LLM può avere difficoltà a ragionare su come utilizzare correttamente lo strumento. Può anche creare problemi nella selezione dei parametri e nella gestione degli errori. Pertanto, lo sviluppo degli strumenti deve raggiungere un equilibrio che sia in linea con le intenzioni dell'utente ed evitare funzionalità insufficienti o eccessive in un singolo strumento. Ti consigliamo di progettare strumenti basati su flussi di lavoro utente completi, raggruppando le operazioni che di solito avvengono insieme (come tre o più chiamate API). Ti consigliamo inoltre di scomporre gli strumenti che superano otto o più parametri o di gestire più intenti utente distinti. Esegui il test con istruzioni reali per verificare che gli agenti siano in grado di utilizzare correttamente ogni strumento.

Se disponi di flussi di lavoro complessi e dinamici che non possono essere facilmente incapsulati come strumento deterministico, potresti prendere in considerazione l'utilizzo del modello. agent-as-tool Invece che l'agente principale cerchi di orchestrare attività complesse in un flusso di lavoro, un agente specializzato può fungere da strumento. Questi tipi di strumenti possono implementare processi decisionali e ramificazioni avanzati, gestire errori e riprovare logiche che non possono essere gestite facilmente nel codice deterministico. Si tratta di un protocollo simile ma distinto dal protocollo Agent2Agent (A2A). Il protocollo A2A è complementare e fornisce interoperabilità e collaborazione tra agenti in qualsiasi framework agentico.

Ti consigliamo di iniziare con l'analisi del flusso di lavoro mappando i flussi di lavoro degli utenti più comuni per identificare le funzionalità principali di cui ogni agente ha bisogno. Questo stabilisce il set di strumenti minimo utilizzabile. Sulla base della nostra esperienza nello sviluppo di server MCP su larga scala, consigliamo le seguenti pratiche. Quando queste pratiche sono in conflitto, dai la priorità alle intenzioni e al flusso di lavoro dell'utente.

Le migliori pratiche per la definizione degli strumenti MCP

- Pensate alle storie degli utenti e raggruppate le operazioni comuni: gli strumenti devono essere mappati direttamente per completare le interazioni degli utenti anziché richiedere l'orchestrazione di più operazioni. Se i flussi di lavoro richiedono in genere tre o più chiamate separate, combinale in un unico strumento. Ciò riduce il carico cognitivo sull'LLM, minimizza il numero di chiamate agli strumenti, riduce il consumo di contesto e la latenza necessari per completare le attività e migliora la precisione e la latenza.
- Limita i parametri a otto o meno: se uno strumento supera gli otto parametri, scomponilo in più strumenti. LLMs faticano a selezionare i parametri man mano che la complessità aumenta.

> **Note**
>
> Se le operazioni di raggruppamento richiedono più di otto parametri, dai la priorità al raggruppamento rispetto al conteggio dei parametri, perché la semplificazione del flusso di lavoro è più importante dei rigidi limiti dei parametri.

- Operazioni di lettura e scrittura separate: fornisce diversi strumenti per leggere i dati e modificarli. Questa separazione rende esplicito quando gli agenti eseguono operazioni potenzialmente distruttive, abilita politiche di autorizzazione diverse e riduce il rischio di modifiche involontarie durante la raccolta delle informazioni.
- Fornisci impostazioni predefinite ragionevoli: progetta strumenti in modo che l'LLM debba specificare solo i parametri specifici della singola richiesta. Le impostazioni predefinite riducono la complessità dei parametri e migliorano la precisione della selezione degli strumenti riducendo al minimo le informazioni su cui l'LLM deve ragionare.
- Preferisci l'esecuzione deterministica: rendi deterministici l'esecuzione degli utensili e l'output quando possibile. Gli strumenti deterministici sono più affidabili e facili da testare. Per flussi di lavoro complessi che richiedono un'orchestrazione intelligente, una logica di ramificazione o una gestione avanzata degli errori che non possono essere facilmente gestite nel codice deterministico, prendi in considerazione l'utilizzo di agenti specializzati come strumenti. Tuttavia, utilizzate questo modello in modo selettivo perché aggiunge complessità.

Definizioni degli strumenti

Quando un LLM riceve una richiesta che non può gestire direttamente, esaminerà gli strumenti disponibili per aiutarlo a completare la richiesta. L'LLM seleziona gli strumenti in base alla sua comprensione semantica dei nomi e delle descrizioni degli strumenti forniti e delle eventuali istruzioni fornite nel prompt. Quindi creerà un input basato sullo schema di input definito e si aspetterà un output basato sullo schema di output. Pertanto, la creazione di definizioni descrittive degli strumenti e schemi di input e output convalidati è fondamentale per aiutare l'LLM a selezionare gli strumenti in modo efficace. Esistono generalmente due approcci per creare questa documentazione: l'approccio alla specificazione degli strumenti e l'approccio docstring.

Approccio alla specifica degli strumenti

L'approccio consigliato consiste nel seguire direttamente le specifiche dell'utensile MCP durante la definizione dell'utensile. L'esempio seguente viene mostrato utilizzando il decoratore di strumenti Strands Agent:

```
@tool(
  name = "search_website",
  description = "This tool searches the provided website for semantic matches to the
 query provided",
  inputSchema = {
    "json": {
      "type": "object",
      "properties": {
        "url": {
          "type": "string",
          "description": "The url of the website to load and search."
        },
        "query": {
          "type": "string",
          "description": "The content you want to try and match in the website."
      }
    },
    "required": ["url", "query"]
  },
  outputSchema = {
    "json": {
      "type": "object",
      "properties": {
      "results": {
          "type": "array",
          "items": {
            "type": "string"
          }
        }
      }
    }
  }
)
def search_website:
  …
```

Utilizza campi standard, come, `name` `descriptioninputSchema`, e `outputSchema` assicura che ogni strumento abbia una documentazione coerente che sia l'LLM che gli umani possano comprendere. Ogni strumento dovrebbe definire questi campi come minimo e, facoltativamente, fornire un titolo e delle annotazioni, che sono suggerimenti opzionali sul comportamento dello strumento. Quando possibile, utilizzate le enumerazioni per i valori dei parametri per facilitare la selezione delle opzioni corrette da parte dell'LLM. Le enumerazioni funzionano meglio per insiemi finiti, come valori di status o priorità, ma non sono adatte per testo in formato libero, valori dinamici, numeri arbitrari o identificatori di risorse. In questi casi, fornisci invece descrizioni ed esempi chiari. Includi anche un valore predefinito, quando possibile, in modo che LLM non debba indovinare quale sia l'opzione corretta. Tieni presente che le definizioni degli strumenti sono incluse nel prompt LLM di ogni chiamata e occupano spazio nella finestra contestuale insieme alle istruzioni di sistema e alla cronologia delle conversazioni.

Approccio Docstring

Un altro approccio, se stai scrivendo i tuoi strumenti in Python, è usare le docstring per fornire la descrizione, l'utilizzo e l'output dello strumento. Di seguito è riportato un esempio di questo approccio:

```
def search_website(url: str, query: str) -> list:

    """
    This tool loads the specified website and then attempts to find content that
 matches the provided query through semantic search. It provides back a list of strings
 that are the sentences that match the query.
    Args:
      url: the website url to load
      query: the content you want to semantically match in the website
    """
```

Le docstring non applicano uno schema o un formato standardizzato. L'utilizzo di questo approccio potrebbe produrre risultati incoerenti in base al modo in cui gli sviluppatori di strumenti scelgono di documentare ogni strumento. La definizione e l'applicazione di uno standard a livello di organizzazione sono essenziali se si segue questo approccio.

Le migliori pratiche per le definizioni degli strumenti MCP

- Segui le specifiche dello strumento MCP: fornisci`name`, `descriptioninputSchema`, e `outputSchema` campi per ogni strumento. Per le implementazioni Python, usa i modelli Pydantic per fornire documentazione in linea tramite descrizioni dei campi, convalida automatica dei tipi

e valori vincolati tramite enumerazioni. Ciò rende gli schemi autodocumentabili e migliora la comprensione LLM delle opzioni di parametro valide.

- Scrivi le descrizioni come istruzioni: le descrizioni degli strumenti sono istruzioni che guidano il processo decisionale LLM. Includi i componenti essenziali dello scopo dello strumento (cosa fa lo strumento), quando utilizzarlo (modelli o scenari delle intenzioni dell'utente), il contesto dell'output (a cosa serve l'output), i parametri e le condizioni di errore.
- Fornisci esempi concreti: includere esempi di flusso di lavoro con valori effettivi è il modo più efficace per guidare LLMs l'utilizzo corretto degli strumenti.
- Documenta le dipendenze in modo esplicito: includi prerequisiti, sequenze numerate, modifiche di stato e azioni successive.

Scoperta degli strumenti

Esistono tre approcci per scoprire e registrare gli strumenti nell'agente con i server MCP: definizione statica, rilevamento dinamico e funzione di ricerca.

Definizione statica

Innanzitutto, è possibile definire staticamente gli strumenti disponibili direttamente nel codice dell'agente. In questo approccio si definisce uno strumento remoto (un oggetto di riferimento sul lato client in un framework come Strands Agent SDK) per ogni strumento fornito dal server MCP a cui accede un client MCP. L'esempio seguente utilizza un trasporto HTTP ottimizzato:

```
from mcp.client.streamable_http import streamablehttp_client
from strands import Agent
from strands.tools.mcp import MCPClient

streamable_http_mcp_client = MCPClient(
    lambda: streamablehttp_client("https://mcp1:8000/mcp")
)

reverse_text = RemoteTool(
    name="reverseText",
    client=streamable_http_mcp_client
)

agent = Agent(tools=[reverse_text])
```

La registrazione individuale degli strumenti ti aiuta a essere molto selettivo riguardo agli strumenti che metti a disposizione del LLM, il che riduce al minimo la quantità di finestra contestuale utilizzata. Il compromesso è che richiede la conoscenza dei nomi degli strumenti disponibili e può essere fragile se gli strumenti disponibili cambiano nel server MCP.

Scoperta dinamica

L'approccio successivo prevede l'utilizzo del rilevamento dinamico e la registrazione di tutti gli strumenti disponibili con l'agente. Questo approccio utilizza il contesto in modo lineare man mano che vengono aggiunti altri strumenti al server MCP. Di seguito è riportato un esempio di questo approccio:

```
from mcp.client.streamable_http import streamablehttp_client
from strands import Agent
from strands.tools.mcp import MCPClient

streamable_http_mcp_client = MCPClient(
    lambda: streamablehttp_client("https://mcp1:8000/mcp")
)

with streamable_http_mcp_client:
    tools = streamable_http_mcp_client.list_tools_sync()
    agent = Agent(tools=tools)
```

Consideriamo uno scenario in cui una tipica definizione di utensile utilizza circa 250-500 token (inclusi nome, descrizione e schema). La registrazione di 20 strumenti consumerebbe da 5.000 a 10.000 token della finestra contestuale. Quando si dispone di un numero limitato di server MCP e si ha il controllo sul numero di strumenti, questa opzione è la più semplice da implementare. Tuttavia, se si prevede che l'elenco degli strumenti aumenti, è possibile che si verifichino problemi di gestione silenziosa del contesto negli agenti. Una variante alternativa di questo approccio consiste nell'utilizzare un parametro di filtro degli strumenti durante la chiamata`list_tools`, come quello fornito dall'SDK di Strands Agents, per ridurre il numero di strumenti registrati con l'agente.

Funzione di ricerca

La terza opzione consiste nell'utilizzare una funzione di ricerca per trovare gli strumenti pertinenti durante l'esecuzione. Elenca tutti gli strumenti disponibili dal server MCP e quindi esegue una ricerca semantica su tali strumenti in base al prompt dell'utente. Quindi, gli strumenti risultanti vengono registrati presso il vostro agente. Amazon Bedrock AgentCore Gateway offre una funzionalità di ricerca semantica nativa che può semplificare l'implementazione di questo tipo di soluzione.

Le migliori pratiche per l'individuazione di strumenti MCP

- Conservazione della finestra contestuale: scegli un approccio di scoperta e registrazione degli strumenti che conservi la maggior parte possibile della finestra contestuale.
- Utilizza funzionalità di filtraggio degli strumenti o di ricerca semantica: fornisci dinamicamente al LLM un set di strumenti ristretto tra cui scegliere, il che ne migliora la precisione e l'efficacia nella scelta dello strumento giusto. Il filtraggio degli strumenti può operare sui nomi degli strumenti (corrispondenza o modelli esatti), sulle descrizioni degli strumenti (corrispondenza semantica) o sui tag di dominio o di categoria. La ricerca semantica è particolarmente efficace per abbinare le intenzioni degli utenti alle descrizioni degli strumenti. Entrambi gli approcci riducono l'uso delle finestre contestuali.

Organizzazione degli strumenti

Scoprire gli strumenti giusti e assicurarsi che l'LLM possa utilizzarli in modo efficace è una delle parti più importanti dello sviluppo di strumenti efficaci. Quando si inizia a sviluppare server MCP, è necessaria una strategia che determini:

- Quanti strumenti sono inclusi in un server MCP
- Quali strumenti non devono essere inseriti nello stesso server MCP
- Come assegnare un nome agli strumenti per renderli ricercabili e prevenire le collisioni tra nomi (strumenti diversi con lo stesso nome)
- Come documentare gli strumenti e il server MCP per renderli facili da usare da parte dell'LLM

L'organizzazione dei namespace è un modello di progettazione che impedisce le collisioni tra i nomi degli strumenti, raggruppa le funzionalità correlate e facilita l'identificazione efficiente degli strumenti tramite. LLMs Il modello stabilisce una categorizzazione strutturata analoga ai sistemi di storage organizzati piuttosto che all'accumulo non strutturato. Consigliamo lo domain-noun-verbschema per la denominazione degli strumenti. Ad esempio,`github_issue_create`,`github_issue_list`,`github_issue_update`, `github_pullrequest_creategithub_pullrequest_list`, o`github_pullrequest_merge`. Il vantaggio di questo modello è evidente quando si esamina il comportamento di ordinamento alfabetico. Quando gli strumenti sono elencati in ordine alfabetico, tutte le operazioni relative ai problemi si raggruppano (`create`,,`update`)`list`, seguite dalle operazioni di pull request (,,). `create` `list` `merge` Il sostantivo (tipo di risorsa) funge da confine organizzativo. Questa struttura

facilita sia la scansione degli strumenti LLM che la navigazione nella documentazione umana perché le funzionalità correlate si raggruppano naturalmente.

Il server MCP deve essere limitato a livello di dominio, ma può essere suddiviso in base alla separazione dei compiti in base alle funzionalità che fornisce. Ad esempio, potreste disporre di server MCP separati per le operazioni di scrittura e le operazioni di lettura su un database. Per applicare questa separazione, si consiglia di implementare dei guardrail a livello di agente che limitino l'accesso ai server MCP in base alle intenzioni e alle autorizzazioni dell'utente. Ciò può essere ottenuto mediante una combinazione dei seguenti elementi:

- Caricamento condizionale del server: carica il server MCP di sola lettura solo quando l'agente rileva operazioni di lettura nell'input dell'utente.
- Filtro basato sulle autorizzazioni: utilizza l'autorizzazione dell'utente per concedere l'accesso solo ai server MCP appropriati.

Infine, è necessario creare un limite superiore al numero di strumenti forniti da un server MCP. Non fate supposizioni su come gli agenti utilizzeranno il vostro server MCP. Potrebbero elencare ingenuamente tutti gli strumenti disponibili e fornirli tutti all'LLM. Se hai più di 50 strumenti in un singolo server, dovresti prendere in considerazione la possibilità di suddividerlo in più server.

Le migliori pratiche per l'organizzazione degli strumenti MPC

- Utilizza lo standard domain-noun-verb di denominazione per gli strumenti: implementa strategie per prevenire le collisioni di nomi sia nei server MCP che negli agenti.
- Imposta un limite superiore: limita il numero di strumenti in un singolo server MCP.
- Dividi i server MCP: utilizza la separazione dei compiti per dividere i server MCP in gruppi logici.

Strategia di hosting MCP

L'astrazione degli strumenti disponibili nei server MCP separa lo sviluppo degli agenti dagli strumenti disponibili. Questo introduce le sfide legate al luogo in cui si ospita il server MCP e al modo in cui gli strumenti sono organizzati all'interno di tali server.

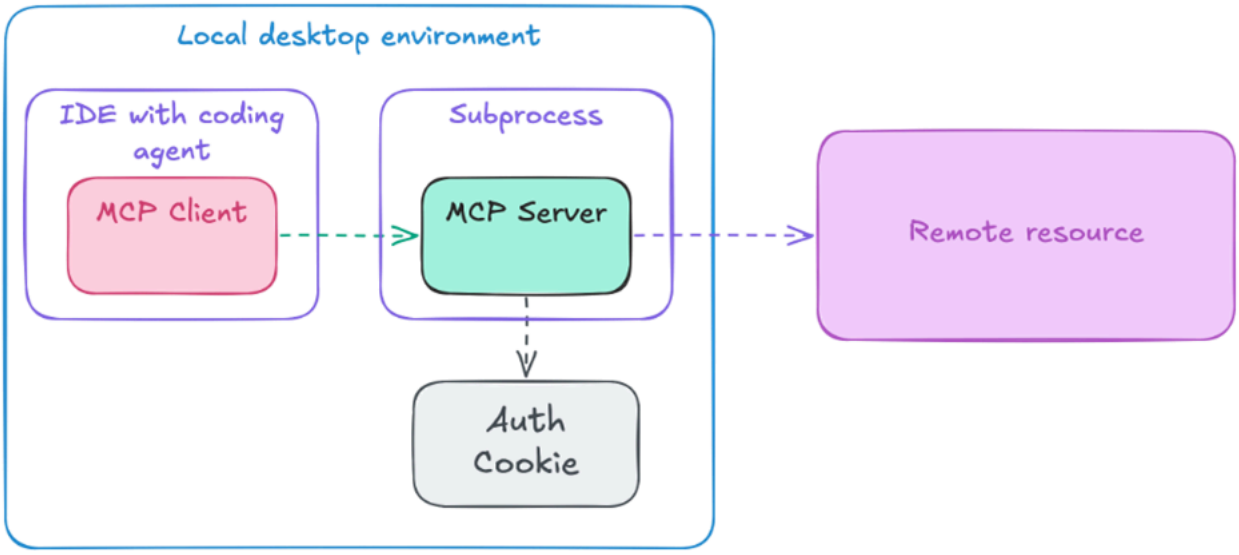
Approcci di hosting

Esistono tre opzioni per ospitare i server MCP: eseguirli localmente su un computer dell'utente finale, ospitarli in remoto o ospitarli tramite un gateway MCP. Ogni opzione presenta vantaggi e compromessi.

Hosting locale

L'hosting locale esegue il server MCP come sottoprocesso sul computer locale insieme all'agente che comunica con il server utilizzando JSON-RPC su flussi di input e output standard. Questo approccio non richiede l'autenticazione tra il client e il server. Gli strumenti possono interagire con applicazioni e file locali, utilizzare credenziali archiviate localmente ed ereditare l'accesso alla rete del computer locale dell'utente. Questo è il modello di hosting più semplice e presenta diversi vantaggi.

Molti clienti iniziano a usare MCP utilizzando server locali. Consentono agli ingegneri di iterare e risolvere rapidamente una serie di problemi dal loro ambiente locale. Prendiamo in considerazione un server MCP che si connette a un repository Git utilizzato dall'assistente di programmazione di un ingegnere. Mantenere il server MCP locale ha molto senso perché può utilizzare le credenziali univoche del tecnico per accedere al repository e non aggiunge un'ulteriore chiamata di rete a un server MCP remoto. L'immagine seguente mostra un server MCP ospitato localmente utilizzato con un agente di codifica in un IDE.



Per questi tipi di implementazioni, è necessario considerare come vengono sviluppati e distribuiti i server MCP. La maggior parte dei clienti sviluppa un registro MCP in cui i server possono essere registrati e scaricati dagli utenti finali. È molto simile a un registro di contenitori in cui un utente può cercare funzionalità specifiche e trovare i server MCP adatti alle proprie esigenze.

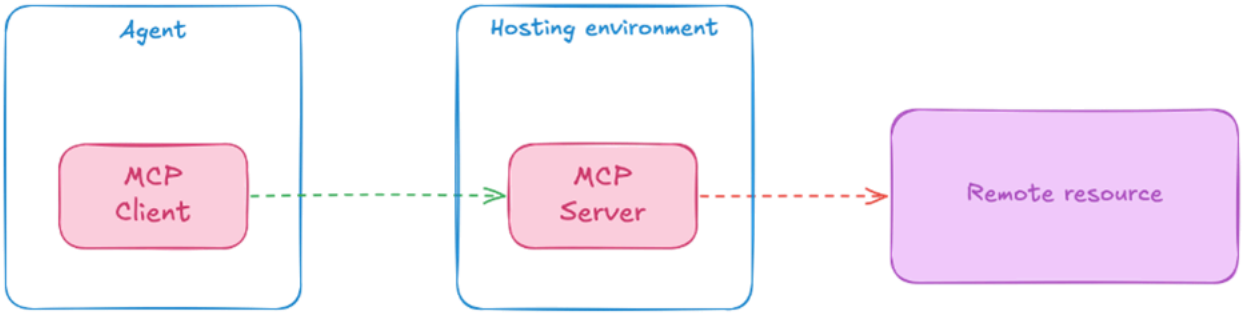
Esistono registri MCP pubblici, come il registro MCP ufficiale, e registri ospitati privatamente. Le organizzazioni in genere allineano la propria strategia di registro MCP alle politiche esistenti relative alla distribuzione del software open source, ai registri dei container e alla gestione interna dei pacchetti. È necessario prendere in considerazione fattori come la scansione di sicurezza, i flussi di lavoro di approvazione e i requisiti di conformità.

Tuttavia, l'hosting locale introduce sfide operative che le organizzazioni dovrebbero prendere in considerazione. Innanzitutto, gli utenti finali devono scoprire, scaricare e configurare i server MCP in modo indipendente. Ciò può aumentare la complessità necessaria per iniziare a utilizzare ogni singolo server MCP che utilizzano localmente. In secondo luogo, non è possibile controllare il ciclo di vita del server MCP, il che significa che gli utenti possono continuare a eseguire versioni obsolete localmente con vulnerabilità di sicurezza o funzionalità mancanti. Ciò può complicare il rispetto dei requisiti di conformità. Alcuni IDEs strumenti CLI, come Kiro, consentono alle organizzazioni di gestire e controllare gli strumenti MCP disponibili, garantendo coerenza e sicurezza tra i team.

Hosting remoto

La seconda opzione consiste nell'ospitare server MCP remoti a cui si accede tramite HTTP o HTTPS. Ciò fornisce l'accesso a qualsiasi client connesso alla rete. L'utilizzo dell'hosting remoto consente di controllare centralmente l'accesso alle risorse e alle funzionalità MCP, implementare l'autenticazione

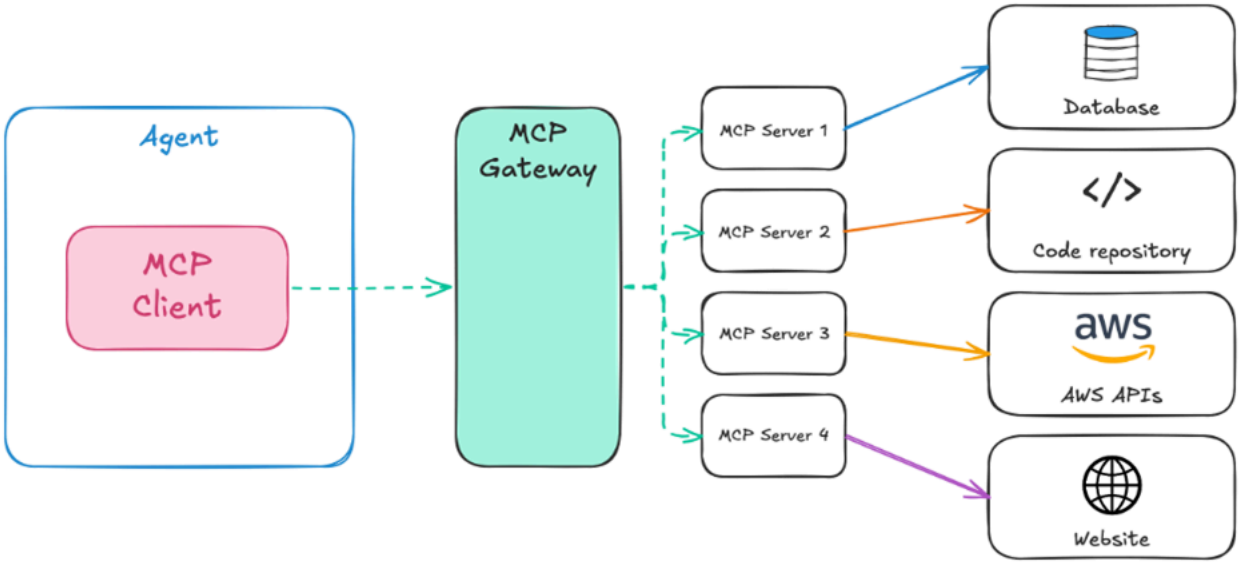
e l'autorizzazione e controllare il controllo delle versioni e gli aggiornamenti della logica del server MCP. L'hosting remoto richiede comunque l'uso di un registro MCP in modo che gli utenti finali possano scoprire i server MCP che desiderano utilizzare con il proprio agente. L'immagine seguente mostra l'approccio all'hosting remoto.



Dal punto di vista dello sviluppo degli agenti, l'esperienza è simile indipendentemente dal fatto che il server MCP sia locale o remoto. La modifica più importante riguarda l'implementazione dell'autenticazione e dell'autorizzazione, che includono sia l'accesso dell'agente al server MCP sia l'accesso del server alle risorse esterne. Le implementazioni dei server MCP remoti devono essere pianificate attentamente per prendere in considerazione l'accesso multi-tenant e la gestione dei privilegi. Il capitolo sulla strategia di governance MCP contiene ulteriori informazioni sulle considerazioni relative all'autenticazione e all'autorizzazione.

Gateway MCP

L'ultima opzione è l'utilizzo di un gateway MCP. I gateway MCP fungono da proxy centralizzato tra client e server MCP e orchestrano l'accesso ai server MCP registrati. Senza un gateway, ogni agente deve registrare ogni server MCP remoto che potrebbe voler utilizzare. Un gateway consente all'agente di connettersi a un singolo endpoint che gestisce l'autenticazione, l'autorizzazione, il routing e la traduzione del protocollo. È possibile aggiungere dinamicamente nuovi server e strumenti MCP e renderli immediatamente disponibili all'agente. L'immagine seguente mostra l'approccio del gateway MCP.



Alcune soluzioni gateway, come Docker MCP Gateway, gestiscono anche il ciclo di vita dei server MCP, avviando i server su richiesta in base alle esigenze. I gateway MCP, come Amazon Bedrock AgentCore Gateway, possono anche aiutare a gestire l'individuazione degli strumenti fornendo funzionalità di ricerca semantica native. Ciò fornisce agli agenti un unico endpoint per connettersi con un client MCP e aiuta a ottimizzare l'utilizzo della finestra contestuale. Il risultato sono agenti semplici in grado di scegliere e utilizzare gli strumenti MCP in modo efficace. Tuttavia, presenta sfide legate all'identità simili a quelle dell'approccio al server MCP remoto.

Le migliori pratiche per l'hosting di server MCP

- La gamma di opzioni di hosting non è valida per tutti. Gran parte dell'utilizzo dei server MCP oggi è locale.
- Quando iniziate a utilizzare server MCP remoti, la vostra considerazione principale è l'autenticazione e l'autorizzazione coerenti verso il server MCP e il modo in cui il server MCP esegue l'autenticazione e l'autorizzazione sulle risorse a valle.
- I gateway MCP semplificano la connettività, l'autenticazione e l'autorizzazione per l'hosting di più server MCP remoti. Forniscono inoltre funzionalità per migliorare la gestione delle finestre contestuali mediante la ricerca di strumenti applicabili.

Strategia di governance MCP

L'altra funzionalità fondamentale che MCP offre alle organizzazioni è il supporto per la governance centralizzata. La vostra strategia di governance MCP dovrebbe riguardare l'autenticazione e l'autorizzazione sia per i server MCP che per le risorse a cui accedono. Dovrebbe inoltre riguardare la limitazione della velocità per proteggere le risorse a valle, le metriche operative per il monitoraggio dell'utilizzo e delle prestazioni degli strumenti e la gestione delle implementazioni e della distribuzione dei server MCP.

Autenticazione e autorizzazione

Una delle parti più importanti della strategia di autenticazione e autorizzazione è la gestione dell'accesso alle risorse a valle dai server MCP. Quando un utente chiama un agente, vengono eseguite l'autenticazione e l'autorizzazione per garantire che l'utente disponga delle autorizzazioni per chiamare l'agente. Quindi, l'agente orchestra la chiamata di strumenti specifici nei server MCP. È necessario decidere come autorizzare l'accesso in base allo strumento.

Un'opzione è l'machine-to-machine autorizzazione, in cui non è richiesto il consenso o l'interazione dell'utente. Ad esempio, una chiamata di un agente basata sul tempo utilizza un server MCP per raccogliere i log da un'applicazione e analizzarli. In questo scenario, l'agente è preautorizzato ad accedere ai dati specificati. La seconda opzione è l'accesso delegato dall'utente, in cui un utente fornisce il proprio consenso all'accesso a dati e risorse specifici dell'utente.

La tabella seguente mostra i modelli di autenticazione e autorizzazione.

Fattore	Accesso delegato dall'utente	Machine-to-machine
Proprietà dei dati	Autorizzazione specifica dell'utente ai dati	Dati a livello di sistema o organizzazione
Interazione con l'utente	L'utente è presente e può acconsentire	Nessuna interazione con l'utente
Tempistica delle operazioni	Interattivo o in tempo reale	In background, pianificato o in batch
Ambito dell'autorizzazione	Le autorizzazioni variano in base all'utente	Autorizzazioni coerenti a livello di agente

L'accesso delegato dall'utente richiede un'implementazione attenta e deve essere sviluppato con il team di sicurezza. Gli agenti devono essere in grado di valutare quali strumenti ha selezionato un LLM e se richiedono un'autorizzazione aggiuntiva. Gli strumenti MCP devono includere descrizioni per indicare i requisiti di autenticazione e autorizzazione e dove recuperare i token di accesso. Le applicazioni client devono supportare richieste di autenticazione intermedie e il client MCP deve fornire le credenziali recuperate all'agente per ogni chiamata allo strumento.

È necessario assicurarsi che gli strumenti MCP dispongano sempre dei propri token per accedere alle funzionalità esterne e che l'accesso sia registrato e verificato. Le credenziali utente non devono essere propagate attraverso il sistema agentic. Ad esempio, i server MCP non devono utilizzare lo stesso token per accedere ai dati utilizzati per richiamare l'agente. Le chiamate downstream devono utilizzare token con ambito esplicito e generati appositamente. Questo aiuta a fornire barriere aggiuntive per impedire l'accesso involontario ai dati per conto di azioni. Può anche aiutare a prevenire che le allucinazioni producano risultati indesiderati. Immaginate che un utente con autorizzazioni amministrative complete chieda a un agente di clonare un database di produzione da utilizzare in fase di pre-produzione. A tal fine, l'utente necessita solo delle autorizzazioni e dei permessi necessari`READ`. `CREATE` Supponiamo che l'LLM abbia allucinazioni e creda di dover ripulire il vecchio database come parte di questa richiesta. Se riutilizza le credenziali dell'utente, probabilmente avrà successo perché le credenziali originali dell'utente dispongono delle autorizzazioni. `DELETE` Invece, se il server MCP utilizza un token intenzionalmente ridotto per la richiesta con solo `READ` e `CREATE` autorizzazioni, il tentativo di eliminare il database di produzione fallirebbe.

Puoi utilizzare Amazon Bedrock AgentCore Identity per contribuire all'implementazione di questi modelli. Assicurati di scegliere intenzionalmente se le autorizzazioni per elencare e richiamare gli strumenti ospitati da un server MCP implichino l'autorizzazione alle funzionalità esterne esposte dal server MCP. Questo flusso di identità dal server MCP alla risorsa e viceversa all'utente dipende dal tipo di servizio di autenticazione e autorizzazione utilizzato. È necessario decidere come gestirlo su larga scala per i server MCP.

Durante la progettazione dei modelli di autenticazione e autorizzazione, implementate meccanismi di isolamento dei token che recuperino token di accesso diversi per ogni strumento a cui si accede. Non riutilizzate i token tra strumenti e server. AgentCore L'identità fornisce questa funzionalità di isolamento dei token. Gestisce automaticamente sia i token del carico di lavoro (per machine-to-machine l'autenticazione) che i token utente (per l'accesso delegato dall'utente) per garantire una separazione adeguata e prevenire l'aumento delle autorizzazioni. Ciò è particolarmente importante quando si incorporano server MCP remoti o gateway MCP.

Le migliori pratiche per l'autenticazione e l'autorizzazione MCP

- Separazione dei token: non trasferite i token portatori dai chiamanti ai servizi a valle. Convalida che il campo aud (audience) corrisponda al server che riceve il token. L'indicazione audience specifica a quale servizio è destinato il token, impedendo il riutilizzo non autorizzato del token su diversi server MCP.
- Seleziona un approccio di accesso: scegli tra machine-to-machine l'accesso delegato dall'utente per ogni strumento fornito dai server MCP. Prendi in considerazione la possibilità di raggruppare gli strumenti nello stesso server MCP che utilizzano lo stesso modello di autenticazione.

Controllo del carico

Come con qualsiasi sistema distribuito, è necessario considerare come controllare il carico nel parco server MCP. Innanzitutto, valutate se implementare la limitazione della velocità nei server MCP e dove implementarla. Se scegliete di non implementare la limitazione della velocità, trasferite qualsiasi limitazione di velocità eseguita dalle risorse a valle. Molti sistemi scelgono di limitare la velocità in base agli attributi della richiesta, come l'ID utente o l'account. Verifica che le richieste inviate ai servizi downstream contengano gli stessi attributi in modo che più utenti non siano influenzati dal carico generato da un altro utente.

Se scegli di implementare la limitazione della velocità, l'approccio consigliato consiste nell'implementare la limitazione della velocità principale a livello del server MCP, con servizi di backend che forniscono una protezione secondaria e agenti che adattano il loro comportamento in base al feedback sul limite di velocità. Valuta se i limiti di velocità sono per server MCP o per strumento. I limiti di velocità per server MCP aiutano a proteggere la flotta e i servizi di server MCP in un ambiente multi-tenant. Tuttavia, ciò può essere molto complicato. I limiti di velocità per utensile sono stati progettati per evitare che le risorse a valle vengano sovraccaricate, che potrebbero non essere sufficientemente limitate. Se uno strumento effettua chiamate multiple APIs, è necessario impostare il limite di velocità in modo che corrisponda alla tariffa più bassa consentita da tali strumenti. APIs

L'inserimento delle informazioni sui limiti di velocità nelle intestazioni HTTP può inoltre essere una metrica utile per gli utenti e i sistemi automatizzati per gestire la propria strategia relativa alla frequenza di richieste e ai tentativi. Ad esempio, è possibile inviare queste intestazioni all'agente dal server MCP, come illustrato nell'esempio seguente:

```
X-RateLimit-Limit: 100
```

```
X-RateLimit-Remaining: 45
X-RateLimit-Reset: 1640995200
```

Inoltre, prendete in considerazione la riduzione del carico per proteggere l'intero servizio quando nessun cliente supera un limite di velocità, ma il carico influisce sulle prestazioni del sistema.

Le migliori pratiche per il controllo del carico

- Scegliete un approccio basato sulla limitazione della velocità: pianificate di limitare la velocità dei singoli utenti in base all'uso delle risorse downstream o all'uso del server e degli strumenti MCP.
- Prendi in considerazione la riduzione del carico: proteggi la tua flotta di server MCP da un sovraccarico generale che non sia causato da un singolo o da una manciata di clienti.

Parametri operativi

Le metriche chiave da acquisire per le implementazioni MCP devono concentrarsi sull'esperienza del cliente che offrono. Queste metriche includono in genere l'utilizzo dei token, l'accuratezza della selezione degli strumenti, il numero di strumenti registrati con l'agente e la latenza degli strumenti. Ad esempio, il monitoraggio dei token di output restituiti da ogni strumento consente di impostare allarmi quando gli strumenti superano una soglia per l'utilizzo della finestra contestuale. Quando uno strumento supera tale soglia, potresti voler esaminare il comportamento dello strumento. Ciò si collega anche alla strategia di progettazione degli strumenti MCP. Le metriche di precisione nella selezione degli strumenti indicano in che misura gli agenti scelgono gli strumenti appropriati per determinate attività, mentre la velocità di esecuzione e le percentuali di successo evidenziano problemi di prestazioni e affidabilità.

Ad esempio, per valutare le metriche di selezione e precisione relative all'uso degli strumenti, i team hanno creato set di dati pregiati per i test di regressione. AWS I set di dati sono stati generati sinteticamente utilizzando i log di invocazione delle API storici sulle query degli utenti. LLMs Utilizzando le metriche predefinite di selezione e utilizzo degli strumenti (come la precisione della selezione degli strumenti, l'accuratezza dei parametri degli strumenti e l'accuratezza delle chiamate alle funzioni multigiro), i AWS team potevano valutare oggettivamente la capacità dell'agente AI di identificare correttamente gli strumenti appropriati, compilare i propri parametri con valori accurati e mantenere sequenze di invocazione degli strumenti coerenti durante i turni di conversazione.

La misurazione delle metriche relative al numero di strumenti registrati con un agente può aiutarti a identificare potenziali problemi di gestione delle finestre di contesto, nonché le modifiche agli

strumenti disponibili presentati dai server MCP. È necessario rivedere regolarmente le metriche operative che indicano l'esperienza utente con il server e gli strumenti MCP.

Collaboratori

Creazione

- Alex Torres, architetto senior delle soluzioni, AWS
- Saikat Gomes, responsabile senior delle soluzioni per i clienti, AWS
- Mike Haken, responsabile delle soluzioni senior, AWS
- Sreeja Das, ingegnere principale, AWS

Revisione

- Ted Swinyar, responsabile dell'architettura delle soluzioni, AWS
- Raju Patil, esperto di dati, AWS

Scrittura tecnica

- Lilly AbouHarb, scrittrice tecnica senior, AWS

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	—	16 marzo 2026

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:

- Refactor/re-architect — Sposta un'applicazione e modificane l'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 Amazon PostgreSQL-Compatible Aurora.
- 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 tuo database Oracle locale su Oracle su un'istanza EC2 in. 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. Esegui 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

A2A () Agent-to-Agent

Un protocollo statico per la collaborazione tra agenti che supporta la delega delle attività e il trasferimento dello stato.

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.

Agente

Un sistema di intelligenza artificiale in grado di ragionare, pianificare e intraprendere azioni in modo autonomo utilizzando strumenti per raggiungere gli obiettivi.

Agente Ops

Pratiche operative per la creazione, il test, l'implementazione e l'esecuzione di agenti di intelligenza artificiale in produzione su larga scala.

funzione aggregata

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 frequentemente 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 viene utilizzato AIOps nella strategia di migrazione AWS , 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 for 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 disturbare 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.

blue/green dispiegamento

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, consulta l'indicatore Implementare le procedure break-glass nella guida. AWS Well-Architected

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.

CoE

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.

Sviluppatore cittadino

Un utente aziendale che crea applicazioni di intelligenza artificiale utilizzando piattaforme senza code/low codice senza competenze tecniche specializzate.

crittografia lato client

Crittografia dei dati localmente, prima che il bersaglio li Servizio AWS riceva.

centro di eccellenza del 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 i post di CCoE 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 dimensionare l'adozione del cloud (ad esempio, creazione di una zona di destinazione, definizione di un CCoE, definizione di un modello operativo)
- Migrazione: migrazione di singole applicazioni

- Re-invention — Ottimizzazione di prodotti e servizi e innovazione nel cloud

Queste fasi sono state definite da Stephen Orban nel post del blog The Journey Toward Cloud-First & the Stages of Adoption sul blog Enterprise Strategy. Cloud AWS Per informazioni sulla loro relazione con la 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 CI/CD pipeline 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 continue () 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.

difesa in profondità

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 approccio di difesa approfondita potrebbe combinare autenticazione a più fattori, segmentazione della rete e 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 Workload 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 lo strangler fig pattern, consulta Modernizzare i servizi Web Microsoft ASP.NET (ASMX) legacy in modo incrementale utilizzando contenitori e Amazon API Gateway.

DOTT.

Vedi disaster recovery.

rilevamento della deriva

Tracciamento delle deviazioni da una configurazione di base. Ad esempio, puoi utilizzarlo AWS CloudFormation per rilevare la deriva nelle risorse di sistema oppure puoi usarlo AWS Control Tower per rilevare cambiamenti nella tua 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. Big-endian i sistemi memorizzano per primi il byte più importante. Little-endian i sistemi memorizzano per primi 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. Few-shot i suggerimenti 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 il 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. Le FM sono in grado di eseguire 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.

Gateway FM

Un intermediario centralizzato che controlla e normalizza l'accesso ai modelli di base. Conosciuto anche come gateway LLM.

G

IA generativa

Un sottoinsieme di modelli di intelligenza artificiale che sono stati addestrati su grandi quantità di dati e che possono utilizzare un semplice messaggio 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 livello elevato che consente di governare risorse, policy e conformità tra le unità organizzative (OU). 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 CSPM GuardDuty AWS Trusted Advisor, Amazon Inspector e controlli personalizzati AWS Lambda .

guardrail (AI)

Meccanismi di sicurezza che filtrano, convalidano e limitano gli input e gli output degli agenti per contribuire a garantire un comportamento dell'IA responsabile e sicuro.

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.

human-in-the-loop (HITL)

Un modello di flusso di lavoro in cui l'esecuzione degli agenti viene sospesa per la revisione e l'approvazione umana nei punti decisionali critici.

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, ad esempio 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

Vedi 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 nel Framework. AWS Well-Architected

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. Nel documento Architettura di riferimento per la sicurezza di

AWS si consiglia di configurare l'account di rete con VPC in entrata, in uscita e di ispezione per proteggere l'interfaccia bidirezionale tra l'applicazione e 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.

Internet delle cose industriale (IIoT)

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, consulta Creazione di una strategia di trasformazione digitale dell'Internet delle cose industriale (IIoT).

VPC di ispezione

In un'architettura AWS multi-account, un VPC centralizzato che gestisce le ispezioni del traffico di rete tra VPC (uguali o diversi Regioni AWS), Internet e reti locali. Nel documento Architettura di riferimento per la sicurezza di AWS si consiglia di configurare l'account di rete con VPC in entrata, in uscita e di ispezione 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, consulta 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 gli LLM.

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.

MCP

Vedi Model Context Protocol.

Model Context Protocol (MCP)

Un protocollo stateless per la comunicazione tra agenti e strumenti.

Server MCP

Un servizio che espone uno o più strumenti tramite il Model Context Protocol.

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, vedete 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 da macchina a macchina (M2M) leggero, basato sul publish/subscribemodello, per dispositivi IoT con risorse limitate.

microservizio

Un piccolo servizio indipendente che comunica tramite API ben definite ed è in genere di proprietà di piccoli team autonomi. 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 tramite un'interfaccia ben definita utilizzando API leggere. 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

Cross-functional team che semplificano la migrazione dei carichi di lavoro attraverso approcci automatizzati e agili. I team di Migration Factory includono in genere 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 AWS con 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 da macchina a macchina (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 Framework. AWS Well-Architected

tecnologia operativa (OT)

Sistemi hardware e software che interagiscono con l'ambiente fisico per controllare operazioni, apparecchiature e 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 in tutto tutti i bucket S3 Regioni AWS, la crittografia lato server con AWS KMS (SSE-KMS) e le richieste dinamiche PUT e dirette al bucket S3. 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. Nel documento Architettura di riferimento per la sicurezza di AWS si consiglia di configurare l'account di rete con VPC in entrata, in uscita e di ispezione 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.

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` che restituisce `true` o `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 ingegneristico dei sistemi che tiene conto della privacy durante l'intero processo di sviluppo.

zone ospitate private

Un container che contiene informazioni su come si desidera che Amazon Route 53 risponda alle query DNS per un dominio e i relativi sottodomini all'interno di uno o più VPC. 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).

RAG

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.

Region

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 obiettivo del punto di ripristino.

VERSO

Vedi 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 Console di gestione AWS 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 Gestione dei segreti AWS, 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 Amazon EC2 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 policy che fornisce il controllo centralizzato sulle autorizzazioni per tutti gli account di un'organizzazione in AWS Organizations. Le SCP definiscono i guardrail o fissano i limiti alle azioni che un amministratore può delegare a utenti o ruoli. Puoi utilizzare le SCP 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.

Shadow AI

Applicazioni di intelligenza artificiale non autorizzate create o utilizzate al di fuori dei canali regolamentati all'interno di un'organizzazione.

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.

modello split-and-seed

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 Modernizzare i servizi Web Microsoft ASP.NET (ASMX) legacy in modo incrementale utilizzando contenitori e Amazon API Gateway.

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

tag

Key-value coppie 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.

strumento

Una funzione o API che un agente può richiamare per eseguire operazioni in sistemi esterni.

Transit Gateway

Un hub di transito di rete che è possibile utilizzare per collegare i VPC e le reti on-premise. 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.

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 VPC che consente di instradare il traffico tramite 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 interrogazioni 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.