Opzioni e architetture di Retrieval Augmented Generation su AWS

AWS Guida prescrittiva

AWS Guida prescrittiva: Opzioni e architetture di Retrieval Augmented Generation su AWS

Table of Contents

Opzioni e architetture di Retrieval Augmented Generation su AWS

Mithil Shah, Rajeev Muralidhar e il forte di Natacha, Amazon Web Services

Ottobre 2024 (storia del documento)

L'intelligenza artificiale generativa si riferisce a un sottoinsieme di modelli di intelligenza artificiale in grado di creare nuovi contenuti e artefatti, come immagini, video, testo e audio, da una semplice richiesta di testo. I modelli di intelligenza artificiale generativa vengono addestrati su grandi quantità di dati che comprendono un'ampia gamma di argomenti e attività. Ciò consente loro di dimostrare una notevole versatilità nell'esecuzione di varie attività, anche quelle per le quali non sono stati addestrati esplicitamente. A causa della capacità di un singolo modello di eseguire più attività, questi modelli vengono spesso definiti modelli di base ()FMs.

Una delle applicazioni più importanti dei modelli di intelligenza artificiale generativa è la loro capacità di rispondere alle domande. Tuttavia, ci sono sfide specifiche che sorgono quando questi modelli vengono utilizzati per rispondere a domande basate su documenti personalizzati. I documenti personalizzati possono includere informazioni proprietarie, siti Web interni, documentazione interna, Confluence pagine, SharePoint pagine e altro. Un'opzione è utilizzare Retrieval Augmented Generation (RAG). Con RAG, il modello di base fa riferimento a una fonte di dati autorevole che non rientra nelle sue fonti di dati di addestramento (come i documenti personalizzati) prima di generare una risposta.

Questa guida descrive le diverse opzioni di intelligenza artificiale generativa disponibili per rispondere alle domande della documentazione personalizzata, inclusi i sistemi Retrieval Augmented Generation (RAG). Fornisce inoltre una panoramica della creazione di sistemi RAG su Amazon Web Services (AWS). Esaminando le opzioni e le architetture RAG, puoi scegliere tra servizi completamente gestiti su AWS architetture RAG personalizzate.

Destinatari principali

I destinatari di questa guida sono gli architetti e i manager di intelligenza artificiale generativa che desiderano creare una soluzione RAG, esaminare le architetture disponibili e comprendere i vantaggi e gli svantaggi di ciascuna opzione.

Obiettivi

Questa guida ti consente di:

- Comprendi le opzioni di intelligenza artificiale generativa disponibili per rispondere a domande contenute nei documenti personalizzati
- Esamina le opzioni di architettura per i sistemi RAG su AWS
- Comprendi i vantaggi e gli svantaggi di ciascuna opzione RAG
- Scegliete un'architettura RAG per il vostro ambiente AWS

Opzioni di intelligenza artificiale generativa per l'interrogazione di documenti personalizzati

Organizations dispone spesso di diverse fonti di dati strutturati e non strutturati. Questa guida si concentra su come utilizzare l'intelligenza artificiale generativa per rispondere a domande basate su dati non strutturati.

I dati non strutturati dell'organizzazione possono provenire da varie fonti. Questi possono essere file di testo PDFs, wiki interni, documenti tecnici, siti Web pubblici, basi di conoscenza o altro. Se desideri un modello di base in grado di rispondere a domande sui dati non strutturati, sono disponibili le seguenti opzioni:

- Addestra un nuovo modello di base utilizzando i tuoi documenti personalizzati e altri dati di formazione
- Perfeziona un modello di base esistente utilizzando i dati dei tuoi documenti personalizzati
- Usa l'apprendimento contestuale per passare un documento al modello di base quando poni una domanda
- Utilizzate un approccio RAG (Retrieval Augmented Generation)

Addestrare da zero un nuovo modello di base che includa dati personalizzati è un'impresa ambiziosa. Alcune aziende ci sono riuscite con successo, ad esempio Bloomberg con il loro BloombergGPTmodello. Un altro esempio è il EXAONEmodello multimodale diLG AI Research, che è stato addestrato utilizzando 600 miliardi di opere d'arte e 250 milioni di immagini ad alta risoluzione, accompagnate da testo. Secondo The Cost of AI: Should You Build or Buy Your Foundation Model (LinkedIn), la formazione di un modello simile Meta Llama 2 costa circa 4,8 milioni di dollari. Esistono due prerequisiti principali per la formazione di un modello da zero: l'accesso alle risorse (finanziarie, tecniche, temporali) e un chiaro ritorno sull'investimento. Se questo non sembra la soluzione giusta, l'opzione successiva è quella di mettere a punto un modello di base esistente.

L'ottimizzazione di un modello esistente implica l'adozione di un modello, ad esempio un modello Amazon Titan, Mistral o Llama, e quindi l'adattamento del modello ai dati personalizzati. Esistono varie tecniche per la regolazione fine, la maggior parte delle quali prevede la modifica solo di alcuni parametri anziché la modifica di tutti i parametri del modello. Si tratta della cosiddetta regolazione di precisione efficiente in termini di parametri. Esistono due metodi principali per la regolazione fine:

- La messa a punto supervisionata utilizza dati etichettati e consente di addestrare il modello per un nuovo tipo di attività. Ad esempio, se desideri generare un report basato su un modulo PDF, potresti dover insegnare al modello come farlo fornendo un numero sufficiente di esempi.
- La messa a punto senza supervisione è indipendente dalle attività e adatta il modello di base ai dati personali. Addestra il modello a comprendere il contesto dei documenti. Il modello ottimizzato crea quindi contenuti, ad esempio un report, utilizzando uno stile più personalizzato dell'organizzazione.

Tuttavia, la messa a punto potrebbe non essere la soluzione ideale per i casi d'uso con domande e risposte. Per ulteriori informazioni, consultate la sezione Confronto tra RAG e messa a punto in questa guida.

Quando si pone una domanda, è possibile passare un documento come modello base e utilizzare l'apprendimento contestuale del modello per restituire le risposte contenute nel documento. Questa opzione è adatta per l'interrogazione ad hoc di un singolo documento. Tuttavia, questa soluzione non funziona bene per interrogare più documenti o per interrogare sistemi e applicazioni, come Microsoft SharePoint o Atlassian Confluence.

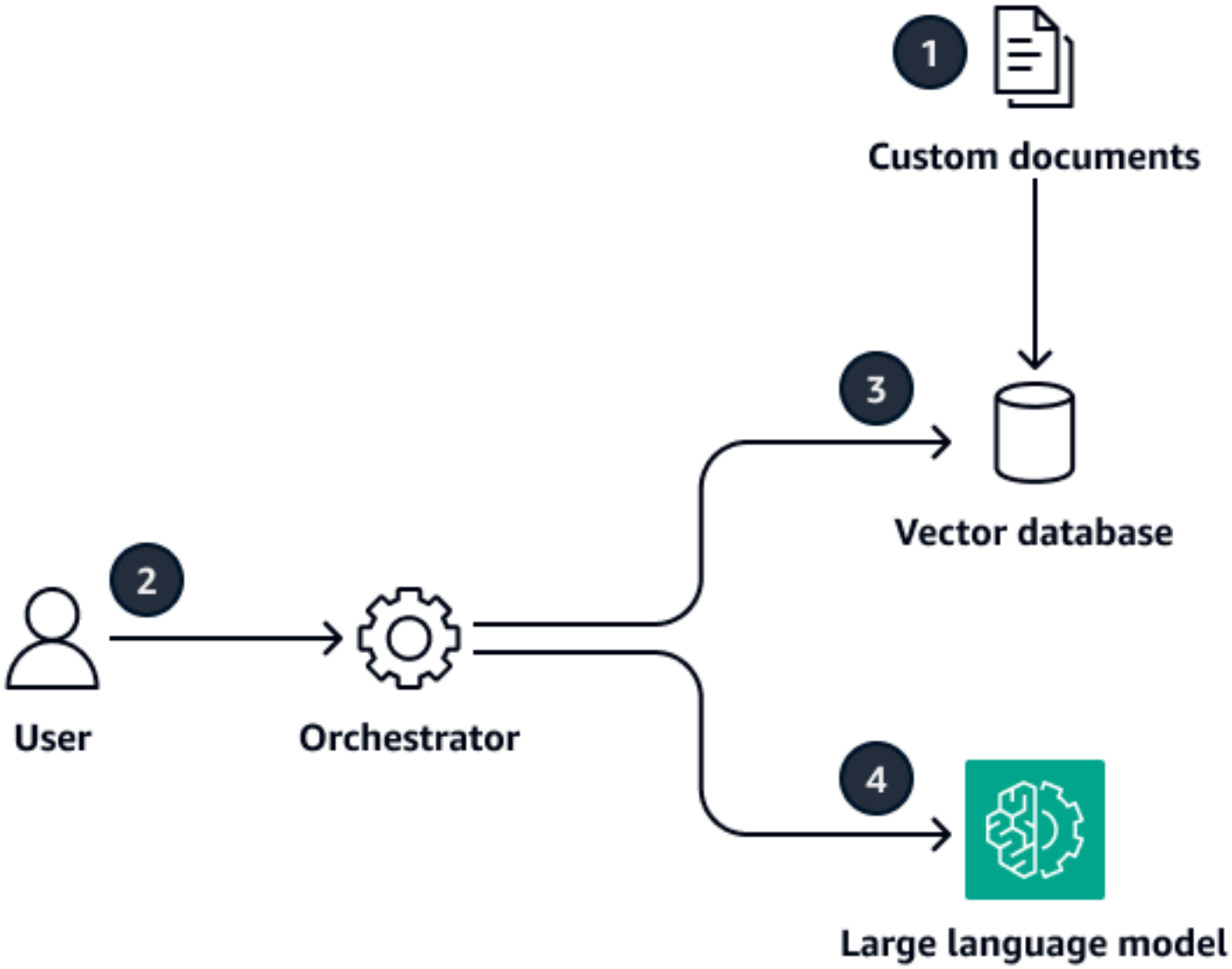
L'ultima opzione è usare RAG. Con RAG, il modello di base fa riferimento ai documenti personalizzati prima di generare una risposta. RAG estende le funzionalità del modello alla knowledge base interna dell'organizzazione, il tutto senza la necessità di riqualificare il modello. È un approccio conveniente per migliorare l'output del modello in modo che rimanga pertinente, accurato e utile in vari contesti.

Argomenti in questa sezione:

- Comprendere Retrieval Augmented Generation
- Confronto tra Retrieval Augmented Generation e messa a punto
- Casi d'uso per Retrieval Augmented Generation

Comprendere Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) è una tecnica utilizzata per potenziare un modello linguistico di grandi dimensioni (LLM) con dati esterni, come i documenti interni di un'azienda. Ciò fornisce al modello il contesto necessario per produrre risultati accurati e utili per ogni caso d'uso specifico. RAG è un approccio pragmatico ed efficace da utilizzare LLMs in un'azienda. Il diagramma seguente mostra una panoramica di alto livello del funzionamento di un approccio RAG.



In generale, il processo RAG prevede quattro fasi. Il primo passaggio viene eseguito una volta e gli altri tre passaggi vengono eseguiti tutte le volte che è necessario:

1. Si creano incorporamenti per inserire i documenti interni in un database vettoriale. Gli incorporamenti sono rappresentazioni numeriche del testo nei documenti che catturano il significato semantico o contestuale dei dati. Un database vettoriale è essenzialmente un database di questi incorporamenti e talvolta viene chiamato archivio vettoriale o indice vettoriale. Questo passaggio richiede la pulizia, la formattazione e la suddivisione in blocchi dei dati, ma si tratta di un'attività iniziale e una tantum.
2. Un essere umano invia una richiesta in linguaggio naturale.

3. Un orchestratore esegue una ricerca di similarità nel database vettoriale e recupera i dati pertinenti. L'orchestratore aggiunge i dati recuperati (noti anche come contesto) al prompt che contiene la query.
4. L'orchestratore invia la query e il contesto al LLM. L'LLM genera una risposta alla query utilizzando il contesto aggiuntivo.

Dal punto di vista dell'utente, RAG sembra interagire con qualsiasi LLM. Tuttavia, il sistema conosce molto di più sui contenuti in questione e fornisce risposte adattate alla base di conoscenze dell'organizzazione.

Per ulteriori informazioni su come funziona un approccio RAG, consultate Cos'è il RAG sul sito Web. AWS

Componenti dei sistemi RAG a livello di produzione

La creazione di un sistema RAG a livello di produzione richiede una riflessione su diversi aspetti del flusso di lavoro RAG. Concettualmente, un flusso di lavoro RAG a livello di produzione richiede le seguenti funzionalità e componenti, indipendentemente dall'implementazione specifica:

- Connettori: collegano diverse fonti di dati aziendali con il database vettoriale. Esempi di fonti di dati strutturate includono database transazionali e analitici. Esempi di fonti di dati non strutturate includono archivi di oggetti, basi di codice e piattaforme Software as a Service (SaaS). Ogni fonte di dati potrebbe richiedere modelli di connettività, licenze e configurazioni diversi.
- Elaborazione dei dati: i dati sono disponibili in molte forme, ad esempio immagini scansionate PDFs, documenti, presentazioni e file. Microsoft SharePoint È necessario utilizzare tecniche di elaborazione dei dati per estrarre, elaborare e preparare i dati per l'indicizzazione.
- Incorporamenti: per eseguire una ricerca pertinente, è necessario convertire i documenti e le query degli utenti in un formato compatibile. Utilizzando modelli linguistici incorporati, convertite i documenti in rappresentazioni numeriche. Questi sono essenzialmente input per il modello di base sottostante.
- Database vettoriale: il database vettoriale è un indice degli incorporamenti, del testo associato e dei metadati. L'indice è ottimizzato per la ricerca e il recupero.
- Retriever: per la query dell'utente, il retriever recupera il contesto pertinente dal database vettoriale e classifica le risposte in base ai requisiti aziendali.
- Modello di base: il modello di base per un sistema RAG è in genere un LLM. Elaborando il contesto e il prompt, il modello di base genera e formatta una risposta per l'utente.

- Guardrails: i Guardrail sono progettati per garantire che la richiesta, il prompt, il contesto recuperato e la risposta LLM siano accurati, responsabili, etici e privi di allucinazioni e pregiudizi.
- Orchestrator: l'orchestratore è responsabile della pianificazione e della gestione del flusso di lavoro. end-to-end
- Esperienza utente: in genere, l'utente interagisce con un'interfaccia di chat conversazionale dotata di funzionalità avanzate, tra cui la visualizzazione della cronologia chat e la raccolta del feedback degli utenti sulle risposte.
- Gestione delle identità e degli utenti: è fondamentale controllare l'accesso degli utenti all'applicazione con una granularità precisa. Nel Cloud AWS, le politiche, i ruoli e le autorizzazioni sono generalmente gestiti tramite AWS Identity and Access Management (IAM).

Chiaramente, c'è una notevole quantità di lavoro per pianificare, sviluppare, rilasciare e gestire un sistema RAG. I servizi completamente gestiti, come Amazon Bedrock o Amazon Q Business, possono aiutarti a gestire parte del sollevamento indifferenziato di carichi pesanti. Tuttavia, le architetture RAG personalizzate possono fornire un maggiore controllo sui componenti, come il retriever o il database vettoriale.

Confronto tra Retrieval Augmented Generation e messa a punto

La tabella seguente descrive i vantaggi e gli svantaggi degli approcci di fine-tuning e basati su RAG.

Approccio	Vantaggi	Svantaggi
Fine-tuning	• Se un modello perfezionato viene addestrato utilizzando l'approccio senza supervisi one, è in grado di creare contenuti che si avvicinan o maggiormente allo stile dell'organizzazione. • Un modello ottimizza to e addestrato su dati proprietari o normativi può aiutare l'organizzazione a seguire gli standard interni	• La messa a punto può richiedere da alcune ore a giorni, a seconda delle dimensioni del modello. Pertanto, non è una buona soluzione se i documenti personalizzati vengono modificati frequentemente. • La messa a punto richiede una conoscenza di tecniche come l'adattamento a basso rango (LoRa) e la regolazio ne di precisione efficiente in

Approccio	Vantaggi	Svantaggi
	o specifici del settore in materia di dati e conformità.	termini di parametri (PEFT). La messa a punto potrebbe richiedere un data scientist. • La messa a punto potrebbe non essere disponibile per tutti i modelli. • I modelli ottimizzati non forniscono un riferimento alla fonte nelle loro risposte. • Quando si utilizza un modello perfezionato per rispondere alle domande, può aumentare il rischio di allucinazioni.

Approccio	Vantaggi	Svantaggi
RAG	• RAG ti consente di creare un sistema di risposta alle domande per i tuoi documenti personalizzati senza alcuna ottimizza zione. • RAG può incorporare i documenti più recenti in pochi minuti. • AWS offre soluzioni RAG completamente gestite. Pertanto, non è richiesto alcun data scientist o conoscenze specialistiche di machine learning. • Nella sua risposta, un modello RAG fornisce un riferimento alla fonte di informazioni. • Poiché RAG utilizza il contesto della ricerca vettoriale come base della risposta generata, il rischio di allucinazioni è ridotto.	• RAG non funziona bene quando si riassumono le informazioni di interi documenti.

Se avete bisogno di creare una soluzione per rispondere a domande che faccia riferimento ai vostri documenti personalizzati, allora vi consigliamo di iniziare da un approccio basato su RAG. Utilizzate il fine-tuning se avete bisogno del modello per eseguire attività aggiuntive, come il riepilogo.

È possibile combinare gli approcci di fine-tuning e RAG in un unico modello. Nel caso, l'architettura RAG non cambia, ma anche l'LLM che genera la risposta viene perfezionato con i documenti personalizzati. Questo combina il meglio dei due mondi e potrebbe essere la soluzione ottimale per il vostro caso d'uso. Per ulteriori informazioni su come combinare il fine-tuning supervisionato con

RAG, consultate la ricerca RAG RAFT: Adapting Language Model to Domain Specific di. University of California, Berkeley

Casi d'uso per Retrieval Augmented Generation

Di seguito sono riportati i casi d'uso più comuni per l'utilizzo di un approccio RAG:

- Motori di ricerca: i motori di ricerca compatibili con RAG possono fornire frammenti più accurati e up-to-date in evidenza nei risultati di ricerca.
- Sistemi di risposta alle domande: RAG può migliorare la qualità delle risposte nei sistemi di risposta alle domande. Il modello basato sul recupero utilizza la ricerca per similarità per trovare passaggi o documenti pertinenti che contengono la risposta. Quindi, genera una risposta concisa e pertinente basata su tali informazioni.
- Vendita al dettaglio o e-commerce: RAG può migliorare l'esperienza degli utenti nell'e-commerce fornendo consigli sui prodotti più pertinenti e personalizzati. Recuperando e incorporando informazioni sulle preferenze degli utenti e sui dettagli dei prodotti, RAG può generare consigli più accurati e utili per i clienti.
- Industriale o manifatturiero: nel settore manifatturiero, RAG consente di accedere rapidamente a informazioni critiche, come le operazioni degli impianti di fabbrica. Può inoltre contribuire ai processi decisionali, alla risoluzione dei problemi e all'innovazione organizzativa. Per i produttori che operano all'interno di quadri normativi rigorosi, RAG può recuperare rapidamente normative e standard di conformità aggiornati da fonti interne ed esterne, ad esempio dagli standard di settore o dalle agenzie di regolamentazione.
- Sanità: RAG ha un potenziale nel settore sanitario, dove l'accesso a informazioni accurate e tempestive è fondamentale. Recuperando e incorporando le conoscenze mediche pertinenti da fonti esterne, RAG può fornire risposte più accurate e consapevoli del contesto nelle applicazioni sanitarie. Tali applicazioni aumentano le informazioni accessibili da un medico umano, che in ultima analisi effettua la chiamata e non il modello.
- Legale: RAG può essere applicato con efficacia in scenari legali, come fusioni e acquisizioni, in cui documenti legali complessi forniscono un contesto per le domande. Questo può aiutare i professionisti legali a risolvere rapidamente questioni normative complesse.

Opzioni Retrieval Augmented Generation completamente gestite su AWS

Per gestire i flussi di lavoro Retrieval Augmented Generation (RAG) AWS, puoi utilizzare pipeline RAG personalizzate o utilizzare alcune delle funzionalità di servizi completamente gestiti che offre. AWS Poiché includono molti dei componenti principali di un sistema basato su RAG, i servizi completamente gestiti possono aiutarvi a gestire parte del sollevamento indifferenziato di carichi pesanti. Tuttavia, questi servizi offrono minori opportunità di personalizzazione.

La versione completamente gestita Servizi AWS utilizza connettori per importare dati da fonti di dati esterne, come siti Web, Atlassian Confluence o Microsoft. SharePoint Le fonti di dati supportate variano in base. Servizio AWS

Questa sezione esplora le seguenti opzioni completamente gestite per la creazione di flussi di lavoro RAG su: AWS

- Basi di conoscenza per Amazon Bedrock
- Amazon Q Business
- Amazon SageMaker AI Canvas

Per ulteriori informazioni su come scegliere tra queste opzioni, consulta Scelta di un'opzione Retrieval Augmented Generation su AWS questa guida.

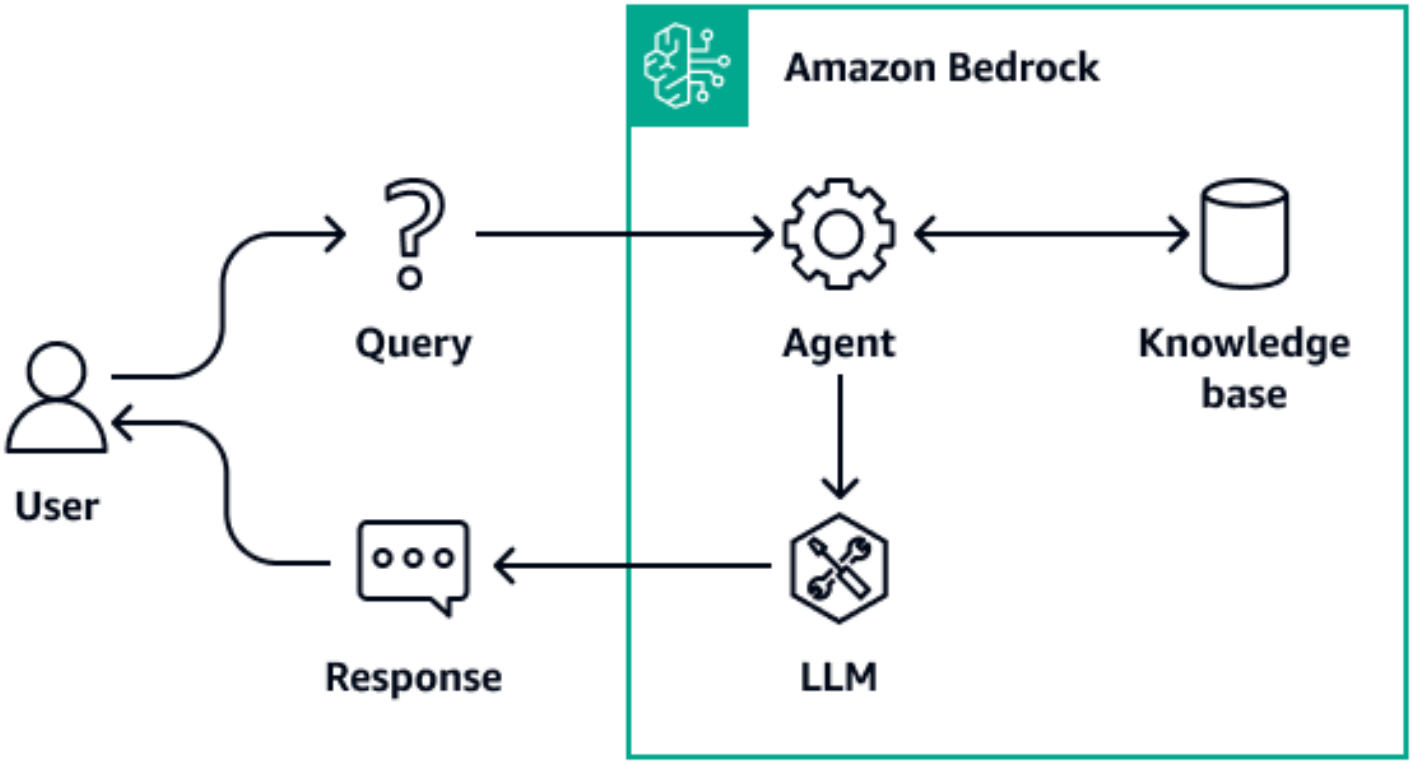
Basi di conoscenza per Amazon Bedrock

Amazon Bedrock è un servizio completamente gestito che rende disponibili per l'uso modelli di base ad alte prestazioni (FMs) delle principali startup di intelligenza artificiale e di Amazon tramite un'API unificata. Le Knowledge Bases sono una funzionalità di Amazon Bedrock che ti aiuta a implementare l'intero flusso di lavoro RAG, dall'ingestione al recupero e al rapido aumento. Non è necessario creare integrazioni personalizzate con le fonti di dati o gestire i flussi di dati. La gestione del contesto delle sessioni è integrata in modo che la tua applicazione di intelligenza artificiale generativa possa supportare prontamente conversazioni a più turni.

Dopo aver specificato la posizione dei dati, le knowledge base per Amazon Bedrock recuperano internamente i documenti, li suddividono in blocchi di testo, convertono il testo in incorporamenti e quindi archiviano gli incorporamenti nel database vettoriale di tua scelta. Amazon Bedrock gestisce

e aggiorna gli incorporamenti, mantenendo il database vettoriale sincronizzato con i dati. Per ulteriori informazioni sul funzionamento delle knowledge base, consulta Come funzionano le knowledge base Amazon Bedrock.

Se aggiungi knowledge base a un agente Amazon Bedrock, l'agente identifica la knowledge base appropriata in base all'input dell'utente. L'agente recupera le informazioni pertinenti e le aggiunge al prompt di input. Il prompt aggiornato fornisce al modello ulteriori informazioni di contesto per generare una risposta. Per migliorare la trasparenza e ridurre al minimo le allucinazioni, le informazioni recuperate dalla knowledge base sono riconducibili alla fonte.



Amazon Bedrock supporta i seguenti due sistemi APIs per RAG:

- RetrieveAndGenerate— Puoi utilizzare questa API per interrogare la tua knowledge base e generare risposte a partire dalle informazioni recuperate. Internamente, Amazon Bedrock converte le query in incorporamenti, interroga la knowledge base, amplia il prompt con i risultati della ricerca come informazioni di contesto e restituisce la risposta generata da LLM. Amazon Bedrock gestisce anche la memoria a breve termine della conversazione per fornire risultati più contestuali.
- Recupera: puoi utilizzare questa API per interrogare la tua knowledge base con informazioni recuperate direttamente dalla knowledge base. È possibile utilizzare le informazioni restituite da questa API per elaborare il testo recuperato, valutarne la pertinenza o sviluppare un flusso di

lavoro separato per la generazione di risposte. Internamente, Amazon Bedrock converte le query in incorporamenti, effettua ricerche nella knowledge base e restituisce i risultati pertinenti. Puoi creare flussi di lavoro aggiuntivi in aggiunta ai risultati di ricerca. Ad esempio, puoi utilizzare il LangChain`AmazonKnowledgeBasesRetriever`plug-in per integrare i flussi di lavoro RAG in applicazioni di intelligenza artificiale generativa.

Per esempi di modelli architettonici e step-by-step istruzioni per l'uso di APIs, consulta Knowledge Bases ora offre un'esperienza RAG completamente gestita in Amazon Bedrock (post del AWS blog). Per ulteriori informazioni su come utilizzare l'`RetrieveAndGenerate`API per creare un flusso di lavoro RAG per un'applicazione intelligente basata su chat, consulta Creare un'applicazione chatbot contestuale utilizzando Amazon Bedrock Knowledge Bases (post del blog).AWS

Origini dati per knowledge base

Puoi collegare i tuoi dati proprietari a una knowledge base. Dopo aver configurato un connettore per le sorgenti dati, puoi sincronizzare o mantenere aggiornati i dati con la tua knowledge base e renderli disponibili per l'interrogazione. Le knowledge base di Amazon Bedrock supportano le connessioni alle seguenti fonti di dati:

- Amazon Simple Storage Service (Amazon S3) — Puoi connettere un bucket Amazon S3 a una knowledge base Amazon Bedrock utilizzando la console o l'API. La knowledge base inserisce e indicizza i file nel bucket. Questo tipo di origine dati supporta le seguenti funzionalità:
  - Campi di metadati del documento: puoi includere un file separato per specificare i metadati per i file nel bucket Amazon S3. Puoi quindi utilizzare questi campi di metadati per filtrare e migliorare la pertinenza delle risposte.
  - Filtri di inclusione o esclusione: puoi includere o escludere determinati contenuti durante la scansione.
  - Sincronizzazione incrementale: le modifiche ai contenuti vengono tracciate e solo i contenuti modificati dall'ultima sincronizzazione vengono sottoposti a scansione.
- Confluence— Puoi connettere un'Atlassian Confluenceistanza a una knowledge base di Amazon Bedrock utilizzando la console o l'API. Questo tipo di origine dati supporta le seguenti funzionalità:
  - Rilevamento automatico dei campi del documento principale: i campi di metadati vengono rilevati automaticamente e sottoposti a scansione. È possibile utilizzare questi campi per filtrare.
  - Filtri di contenuto di inclusione o esclusione: puoi includere o escludere determinati contenuti utilizzando un prefisso o uno schema di espressione regolare nello spazio, nel titolo della pagina, nel titolo del blog, nel commento, nel nome dell'allegato o nell'estensione.

  - Sincronizzazione incrementale: le modifiche ai contenuti vengono tracciate e solo i contenuti modificati dall'ultima sincronizzazione vengono sottoposti a scansione.
  - OAuth autenticazione 2.0, autenticazione con token Confluence API: le credenziali di autenticazione sono archiviate in. Gestione dei segreti AWS
- Microsoft SharePoint— È possibile connettere un'SharePointistanza a una knowledge base utilizzando la console o l'API. Questo tipo di origine dati supporta le seguenti funzionalità:
  - Rilevamento automatico dei campi del documento principale: i campi di metadati vengono rilevati automaticamente e sottoposti a scansione. È possibile utilizzare questi campi per filtrare.
  - Filtri di contenuto di inclusione o esclusione: puoi includere o escludere determinati contenuti utilizzando un prefisso o un modello di espressione regolare nel titolo della pagina principale, nel nome dell'evento e nel nome del file (inclusa l'estensione).
  - Sincronizzazione incrementale: le modifiche ai contenuti vengono tracciate e solo i contenuti modificati dall'ultima sincronizzazione vengono sottoposti a scansione.
  - OAuth Autenticazione 2.0: le credenziali di autenticazione vengono archiviate in. Gestione dei segreti AWS
- Salesforce— È possibile connettere un'Salesforceistanza a una knowledge base utilizzando la console o l'API. Questo tipo di origine dati supporta le seguenti funzionalità:
  - Rilevamento automatico dei campi del documento principale: i campi di metadati vengono rilevati automaticamente e sottoposti a scansione. È possibile utilizzare questi campi per filtrare.
  - Filtri di contenuto di inclusione o esclusione: puoi includere o escludere determinati contenuti utilizzando un prefisso o un modello di espressione regolare. Per un elenco dei tipi di contenuto a cui puoi applicare filtri, consulta i filtri di inclusione/esclusione nella documentazione di Amazon Bedrock.
  - Sincronizzazione incrementale: le modifiche ai contenuti vengono tracciate e solo i contenuti modificati dall'ultima sincronizzazione vengono sottoposti a scansione.
  - OAuth Autenticazione 2.0: le credenziali di autenticazione vengono archiviate in. Gestione dei segreti AWS
- Web Crawler: un web crawler di Amazon Bedrock si connette e ne esegue la scansione. URLs Sono supportate le seguenti funzionalità:
  - Seleziona più elementi URLs da scansionare
  - Rispetta le direttive robots.txt standard, come e `Allow` `Disallow`
  - Escludi URLs che corrisponda a uno schema
  - Limita la velocità di scansione

- In Amazon CloudWatch, visualizza lo stato di ogni URL sottoposto a scansione

Per ulteriori informazioni sulle fonti di dati che puoi connettere alla tua knowledge base Amazon Bedrock, consulta Creare un connettore di origine dati per la tua knowledge base.

Database vettoriali per basi di conoscenza

Quando si imposta una connessione tra la knowledge base e l'origine dati, è necessario configurare un database vettoriale, noto anche come archivio vettoriale. Un database vettoriale è il luogo in cui Amazon Bedrock archivia, aggiorna e gestisce gli incorporamenti che rappresentano i tuoi dati. Ogni fonte di dati supporta diversi tipi di database vettoriali. Per determinare quali database vettoriali sono disponibili per la tua fonte di dati, consulta i tipi di fonti di dati.

Se preferisci che Amazon Bedrock crei automaticamente un database vettoriale in Amazon OpenSearch Serverless per te, puoi scegliere questa opzione quando crei la knowledge base. Tuttavia, puoi anche scegliere di configurare il tuo database vettoriale. Se configuri il tuo database vettoriale, consulta Prerequisiti per il tuo archivio vettoriale per una knowledge base. Ogni tipo di database vettoriale ha i propri prerequisiti.

A seconda del tipo di origine dati, le knowledge base di Amazon Bedrock supportano i seguenti database vettoriali:

- Amazon OpenSearch Serverless
- Amazon Aurora PostgreSQL-Compatible Edition
- Pinecone(Pineconedocumentazione)
- Redis Enterprise Cloud(Redisdocumentazione)
- MongoDB Atlas(MongoDBdocumentazione)

Amazon Q Business

Amazon Q Business è un assistente completamente gestito e basato sull'intelligenza artificiale generativa che puoi configurare per rispondere a domande, fornire riepiloghi, generare contenuti e completare attività in base ai tuoi dati aziendali. Consente agli utenti finali di ricevere risposte immediate e basate sulle autorizzazioni da fonti di dati aziendali con citazioni.

Funzionalità principali

Le seguenti funzionalità di Amazon Q Business possono aiutarti a creare un'applicazione AI generativa basata su RAG di livello di produzione:

- Connettori integrati: Amazon Q Business supporta più di 40 tipi di connettori, ad esempio connettori per Adobe Experience Manager (AEM)Salesforce,Jira, eMicrosoft SharePoint. Per un elenco completo, consulta Connettori supportati. Se hai bisogno di un connettore non supportato, puoi utilizzare Amazon AppFlow per estrarre dati dalla tua fonte di dati in Amazon Simple Storage Service (Amazon S3) Simple Storage Service (Amazon S3) e quindi connettere Amazon Q Business al bucket Amazon S3. Per un elenco completo delle fonti di dati AppFlow supportate da Amazon, consulta Applicazioni supportate.
- Pipeline di indicizzazione integrate: Amazon Q Business fornisce una pipeline integrata per l'indicizzazione dei dati in un database vettoriale. Puoi usare una AWS Lambda funzione per aggiungere una logica di preelaborazione per la tua pipeline di indicizzazione.
- Opzioni di indicizzazione: puoi creare e fornire un indice nativo in Amazon Q Business e utilizzare un Amazon Q Business retriever per estrarre i dati da quell'indice. In alternativa, puoi utilizzare un indice Amazon Kendra preconfigurato come retriever. Per ulteriori informazioni, consulta Creazione di un retriever per un'applicazione Amazon Q Business.
- Modelli base: Amazon Q Business utilizza i modelli di base supportati in Amazon Bedrock. Per un elenco completo, consulta Modelli di base supportati in Amazon Bedrock.
- Plugin: Amazon Q Business offre la possibilità di utilizzare i plug-in per l'integrazione con i sistemi di destinazione, ad esempio un modo automatico per riepilogare le informazioni sui ticket e creare i ticket. Jira Una volta configurati, i plugin possono supportare azioni di lettura e scrittura che possono aiutarti a incrementare la produttività degli utenti finali. Amazon Q Business supporta due tipi di plug-in: plug-in integrati e plug-in personalizzati.
- Guardrails — Amazon Q Business supporta controlli globali e controlli a livello di argomento. Ad esempio, questi controlli possono rilevare informazioni di identificazione personale (PII), abusi o informazioni sensibili nei prompt. Per ulteriori informazioni, consulta Controlli amministrativi e guardrail in Amazon Q Business.
- Gestione delle identità: con Amazon Q Business, puoi gestire gli utenti e il loro accesso all'applicazione AI generativa basata su RAG. Per ulteriori informazioni, consulta Gestione delle identità e degli accessi per Amazon Q Business. Inoltre, i connettori Amazon Q Business indicizzano le informazioni dell'elenco di controllo degli accessi (ACL) allegate a un documento insieme al documento stesso. Quindi, Amazon Q Business archivia le informazioni ACL indicizzate nell'Amazon Q Business User Store per creare mappature di utenti e gruppi e filtrare le risposte

delle chat in base all'accesso dell'utente finale ai documenti. Per ulteriori informazioni, consulta Concetti relativi ai connettori di origine dati.

- Arricchimento dei documenti: la funzionalità di arricchimento dei documenti consente di controllare sia i documenti e gli attributi dei documenti che vengono inseriti nell'indice sia il modo in cui vengono inseriti. Ciò può essere ottenuto mediante due approcci:
  - Configura le operazioni di base: utilizza le operazioni di base per aggiungere, aggiornare o eliminare gli attributi del documento dai dati. Ad esempio, puoi cancellare i dati PII scegliendo di eliminare qualsiasi attributo del documento relativo alle PII.
  - Configura le funzioni Lambda: utilizza una funzione Lambda preconfigurata per eseguire una logica di manipolazione degli attributi dei documenti più personalizzata e avanzata sui tuoi dati. Ad esempio, i dati aziendali potrebbero essere archiviati come immagini scansionate. In tal caso, puoi utilizzare una funzione Lambda per eseguire il riconoscimento ottico dei caratteri (OCR) sui documenti scansionati per estrarne il testo. Quindi, ogni documento scansionato viene trattato come un documento di testo durante l'importazione. Infine, durante la chat, Amazon Q prenderà in considerazione i dati testuali estratti dai documenti scansionati quando genera le risposte.

  Quando implementi la tua soluzione, puoi scegliere di combinare entrambi gli approcci di arricchimento dei documenti. È possibile utilizzare le operazioni di base per eseguire una prima analisi dei dati e quindi utilizzare una funzione Lambda per operazioni più complesse. Per ulteriori informazioni, consulta Arricchimento dei documenti in Amazon Q Business.
- Integrazione: dopo aver creato l'applicazione Amazon Q Business, puoi integrarla in altre applicazioni, come Slack oMicrosoft Teams. Ad esempio, consulta Implementare un Slack gateway per Amazon Q Business e Implementare un gateway Microsoft Teams per Amazon Q Business (AWS post di blog).

Personalizzazione per l'utente finale

Amazon Q Business supporta il caricamento di documenti che potrebbero non essere archiviati nelle fonti di dati e nell'indice della tua organizzazione. I documenti caricati non vengono archiviati. Sono disponibili solo per la conversazione in cui vengono caricati i documenti. Amazon Q Business supporta tipi di documenti specifici per il caricamento. Per ulteriori informazioni, consulta Caricare file e chattare in Amazon Q Business.

Amazon Q Business include una funzionalità di filtraggio per attributo del documento. Sia gli amministratori che gli utenti finali possono utilizzare questa funzionalità. Gli amministratori possono personalizzare e controllare le risposte alla chat per gli utenti finali utilizzando gli attributi. Ad

esempio, se il tipo di origine dati è un attributo allegato ai documenti, puoi specificare che le risposte alla chat vengano generate solo da una fonte di dati specifica. In alternativa, puoi consentire agli utenti finali di limitare l'ambito delle risposte alla chat utilizzando i filtri degli attributi che hai selezionato.

Gli utenti finali possono creare app Amazon Q leggere e su misura all'interno del più ampio ambiente applicativo Amazon Q Business. Le app Amazon Q consentono l'automazione delle attività per un dominio specifico, ad esempio un'app creata appositamente per il team di marketing.

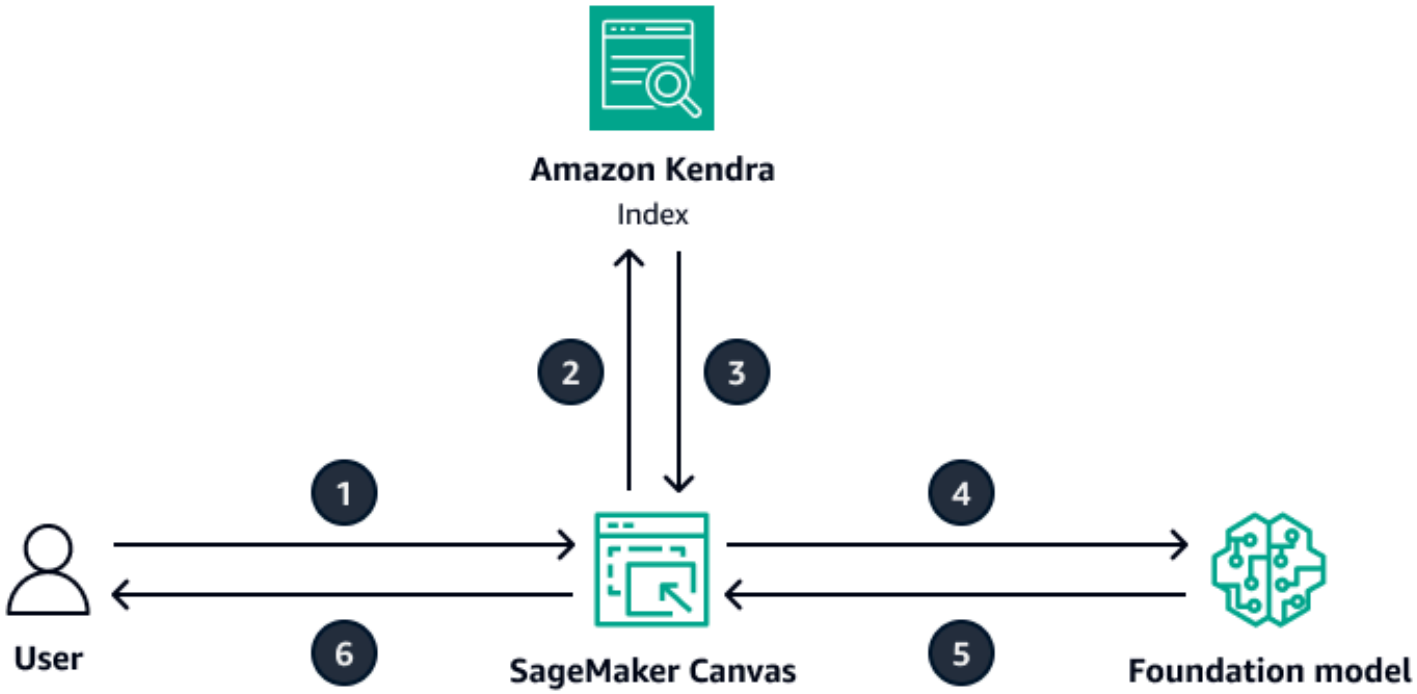
Amazon SageMaker AI Canvas

Amazon SageMaker AI Canvas ti aiuta a utilizzare l'apprendimento automatico per generare previsioni senza dover scrivere alcun codice. Fornisce un'interfaccia visiva senza codice che consente di preparare dati, creare e distribuire modelli di machine learning, semplificando il ciclo di vita del end-to-end machine learning in un ambiente unificato. Le complessità della preparazione dei dati, dello sviluppo dei modelli, del rilevamento delle distorsioni, della spiegabilità e del monitoraggio sono riassunte in un'interfaccia intuitiva. Gli utenti non devono essere esperti di SageMaker intelligenza artificiale o di operazioni di machine learning (MLOps) per sviluppare, rendere operativi e monitorare i modelli con AI Canvas. SageMaker

Con SageMaker AI Canvas, la funzionalità RAG viene fornita tramite una funzionalità di interrogazione dei documenti senza codice. Puoi arricchire l'esperienza di chat in SageMaker AI Canvas utilizzando un indice Amazon Kendra come ricerca aziendale sottostante. Per ulteriori informazioni, consulta Estrarre informazioni dai documenti con interrogazioni sui documenti.

La connessione di SageMaker AI Canvas all'indice Amazon Kendra richiede una configurazione unica. Come parte della configurazione del dominio, un amministratore cloud può scegliere uno o più indici Kendra su cui l'utente può interrogare quando interagisce con Canvas. SageMaker Per istruzioni su come abilitare la funzionalità di interrogazione dei documenti, consulta Guida introduttiva all'uso di Amazon SageMaker AI Canvas.

SageMaker AI Canvas gestisce la comunicazione di base tra Amazon Kendra e il modello di base selezionato. Per ulteriori informazioni sui modelli di base supportati da SageMaker AI Canvas, consulta Modelli di base AI generativi in SageMaker AI Canvas. Il diagramma seguente mostra come funziona la funzionalità di interrogazione dei documenti dopo che l'amministratore cloud ha collegato SageMaker AI Canvas a un indice Amazon Kendra.



Il diagramma mostra il flusso di lavoro seguente:

1. L'utente avvia una nuova chat in SageMaker AI Canvas, attiva i documenti Query, seleziona l'indice di destinazione e quindi invia una domanda.
2. SageMaker AI Canvas utilizza la query per cercare dati pertinenti nell'indice Amazon Kendra.
3. SageMaker AI Canvas recupera i dati e le relative fonti dall'indice Amazon Kendra.
4. SageMaker AI Canvas aggiorna il prompt per includere il contesto recuperato dall'indice Amazon Kendra e invia il prompt al modello di base.
5. Il modello di base utilizza la domanda originale e il contesto recuperato per generare una risposta.
6. SageMaker AI Canvas fornisce la risposta generata all'utente. Include riferimenti alle fonti di dati, come i documenti, che sono state utilizzate per generare la risposta.

Architetture Custom Retrieval Augmented Generation su AWS

La sezione precedente descrive come utilizzare un RAG (Fully managed Servizio AWS for Retrieval Augmented Generation). Tuttavia, alcuni casi d'uso richiedono un maggiore controllo sui componenti del sistema, come il retriever o l'LLM (chiamato anche generatore). Ad esempio, potrebbe essere necessaria la flessibilità necessaria per scegliere il proprio database vettoriale o accedere a una fonte di dati non supportata. Per questi casi d'uso, puoi creare un'architettura RAG personalizzata.

Questa sezione contiene i seguenti argomenti:

- Retriever per flussi di lavoro RAG
- Generatori per flussi di lavoro RAG

Per ulteriori informazioni su come scegliere tra le opzioni retriever e generator in questa sezione, consulta questa Scelta di un'opzione Retrieval Augmented Generation su AWS guida.

Retriever per flussi di lavoro RAG

Questa sezione spiega come costruire un retriever. Puoi utilizzare una soluzione di ricerca semantica completamente gestita, come Amazon Kendra, oppure puoi creare una ricerca semantica personalizzata utilizzando un database vettoriale. AWS

Prima di esaminare le opzioni di retriever, assicurati di aver compreso i tre passaggi del processo di ricerca vettoriale:

1. I documenti che devono essere indicizzati vengono separati in parti più piccole. Questa operazione si chiama suddivisione in blocchi.
2. Si utilizza un processo chiamato incorporamento per convertire ogni blocco in un vettore matematico. Quindi, indicizzate ogni vettore in un database vettoriale. L'approccio utilizzato per indicizzare i documenti influenza la velocità e la precisione della ricerca. L'approccio di indicizzazione dipende dal database vettoriale e dalle opzioni di configurazione che offre.
3. La query dell'utente viene convertita in un vettore utilizzando lo stesso processo. Il retriever cerca nel database vettoriale vettori simili al vettore di query dell'utente. La somiglianza viene calcolata utilizzando metriche come la distanza euclidea, la distanza del coseno o il prodotto scalare.

Questa guida descrive come utilizzare i seguenti servizi Servizi AWS o quelli di terze parti per creare un livello di recupero personalizzato su: AWS

- Amazon Kendra
- OpenSearch Servizio Amazon
- Amazon Aurora PostgreSQL e pgvector
- Analisi di Amazon Neptune
- Amazon MemoryDB
- Amazon DocumentDB
- Pinecone
- MongoDB Atlas
- Weaviate

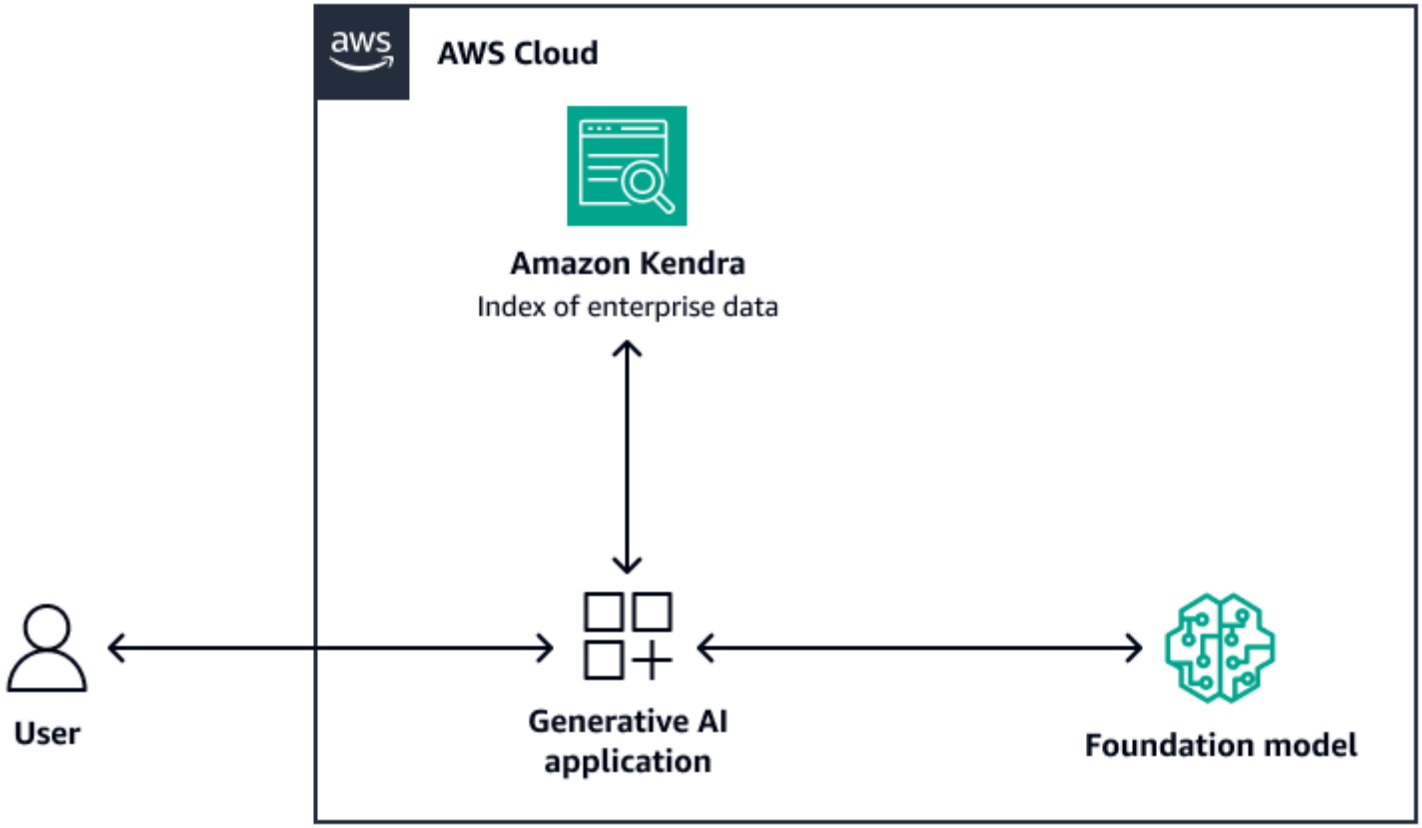
Amazon Kendra

Amazon Kendra è un servizio di ricerca intelligente e completamente gestito che utilizza l'elaborazione del linguaggio naturale e algoritmi avanzati di apprendimento automatico per restituire risposte specifiche alle domande di ricerca dai tuoi dati. Amazon Kendra ti aiuta a importare direttamente documenti da più fonti e a interrogare i documenti dopo che si sono sincronizzati correttamente. Il processo di sincronizzazione crea l'infrastruttura necessaria per creare una ricerca vettoriale sul documento importato. Pertanto, Amazon Kendra non richiede i tre passaggi tradizionali del processo di ricerca vettoriale. Dopo la sincronizzazione iniziale, puoi utilizzare una pianificazione definita per gestire l'ingestione continua.

Di seguito sono riportati i vantaggi dell'utilizzo di Amazon Kendra for RAG:

- Non è necessario mantenere un database vettoriale perché Amazon Kendra gestisce l'intero processo di ricerca vettoriale.
- Amazon Kendra contiene connettori predefiniti per le fonti di dati più diffuse, come database, crawler di siti Web, bucket Amazon S3, istanze e istanze. Microsoft SharePoint Atlassian Confluence Sono disponibili connettori sviluppati dai AWS partner, come connettori per e. Box GitLab
- Amazon Kendra fornisce un filtro per la lista di controllo degli accessi (ACL) che restituisce solo i documenti a cui l'utente finale ha accesso.
- Amazon Kendra può potenziare le risposte in base ai metadati, come la data o l'archivio di origine.

L'immagine seguente mostra un'architettura di esempio che utilizza Amazon Kendra come livello di recupero del sistema RAG. Per ulteriori informazioni, consulta Crea rapidamente applicazioni di intelligenza artificiale generativa ad alta precisione su dati aziendali utilizzando Amazon Kendra LangChain e modelli linguistici di grandi dimensioni (post di blog).AWS



Per il modello base, puoi utilizzare Amazon Bedrock o un LLM distribuito tramite Amazon AI. SageMaker JumpStart Puoi usare AWS Lambda with LangChainper orchestrare il flusso tra l'utente, Amazon Kendra e LLM. Per creare un sistema RAG che utilizzi Amazon Kendra LLMs e vari altri, consulta il repository Amazon LangChainLangChain Kendra Extensions. GitHub

OpenSearch Servizio Amazon

Amazon OpenSearch Service fornisce algoritmi ML integrati per la ricerca k-Nearest Neighbors (k-NN) al fine di eseguire una ricerca vettoriale. OpenSearch Il servizio fornisce anche un motore vettoriale per Amazon EMR Serverless. Puoi utilizzare questo motore vettoriale per creare un sistema RAG con funzionalità di archiviazione e ricerca vettoriali scalabili e ad alte prestazioni. Per ulteriori informazioni su come creare un sistema RAG utilizzando OpenSearch Serverless, consulta Creare flussi di lavoro RAG scalabili e serverless con un motore vettoriale per i modelli Amazon Serverless OpenSearch e Amazon Bedrock Claude (post di blog).AWS

Di seguito sono riportati i vantaggi dell'utilizzo di Service per la ricerca vettoriale: OpenSearch

- Fornisce il controllo completo sul database vettoriale, inclusa la creazione di una ricerca vettoriale scalabile utilizzando Serverless. OpenSearch
- Fornisce il controllo sulla strategia di suddivisione in blocchi.
- Utilizza algoritmi approssimativi più vicini (ANN) delle librerie Non-Metric Space Library (NMSLIB), Faiss e Apache Lucene per alimentare una ricerca k-NN. È possibile modificare l'algoritmo in base al caso d'uso. Per ulteriori informazioni sulle opzioni per personalizzare la ricerca vettoriale tramite OpenSearch Service, consulta la spiegazione delle funzionalità del database vettoriale di Amazon OpenSearch Service (AWS post del blog).
- OpenSearch Serverless si integra con le knowledge base di Amazon Bedrock come indice vettoriale.

Amazon Aurora PostgreSQL e pgvector

Amazon Aurora PostgreSQL Compatible Edition è un motore di database relazionale completamente gestito che ti aiuta a configurare, gestire e scalare le distribuzioni PostgreSQL. pgvector è un'estensione PostgreSQL open source che fornisce funzionalità di ricerca per somiglianze vettoriali. Questa estensione è disponibile sia per Aurora compatibile con PostgreSQL che per Amazon Relational Database Service (Amazon RDS) per PostgreSQL. Per ulteriori informazioni su come creare un sistema basato su RAG che utilizzi Aurora, compatibile con PostgreSQL e pgvector, consulta i seguenti post di blog: AWS

- Creazione di ricerche basate sull'intelligenza artificiale in PostgreSQL utilizzando Amazon AI e pgvector SageMaker
- Sfrutta pgvector e Amazon Aurora PostgreSQL per l'elaborazione del linguaggio naturale, i chatbot e l'analisi del sentiment

Di seguito sono riportati i vantaggi dell'utilizzo di pgvector e Aurora PostgreSQL compatibili:

- Supporta la ricerca esatta e approssimativa del vicino più prossimo. Supporta anche le seguenti metriche di somiglianza: distanza L2, prodotto interno e distanza del coseno.
- Supporta l'indicizzazione Inverted File with Flat Compression (IVFFlat) e Hierarchical Navigable Small Worlds (HNSW).
- È possibile combinare la ricerca vettoriale con query su dati specifici del dominio disponibili nella stessa istanza PostgreSQL.

- Aurora PostgreSQL Compatible è ottimizzata e fornisce la memorizzazione nella cache su più livelli. I/O Per carichi di lavoro che superano la memoria disponibile dell'istanza, pgvector può aumentare le query al secondo per la ricerca vettoriale fino a 8 volte.

Analisi di Amazon Neptune

Amazon Neptune Analytics è un motore di database grafico ottimizzato per la memoria per l'analisi. Supporta una libreria di algoritmi di analisi dei grafici ottimizzati, query grafiche a bassa latenza e funzionalità di ricerca vettoriale all'interno delle traversate di grafici. Dispone inoltre di una ricerca integrata per somiglianza vettoriale. Fornisce un endpoint per creare un grafico, caricare dati, richiamare query ed eseguire ricerche di somiglianza vettoriale. Per ulteriori informazioni su come creare un sistema basato su RAG che utilizza Neptune Analytics, consulta Utilizzo dei knowledge graphs per creare applicazioni GraphRag con Amazon Bedrock e Amazon Neptune (post del blog).AWS

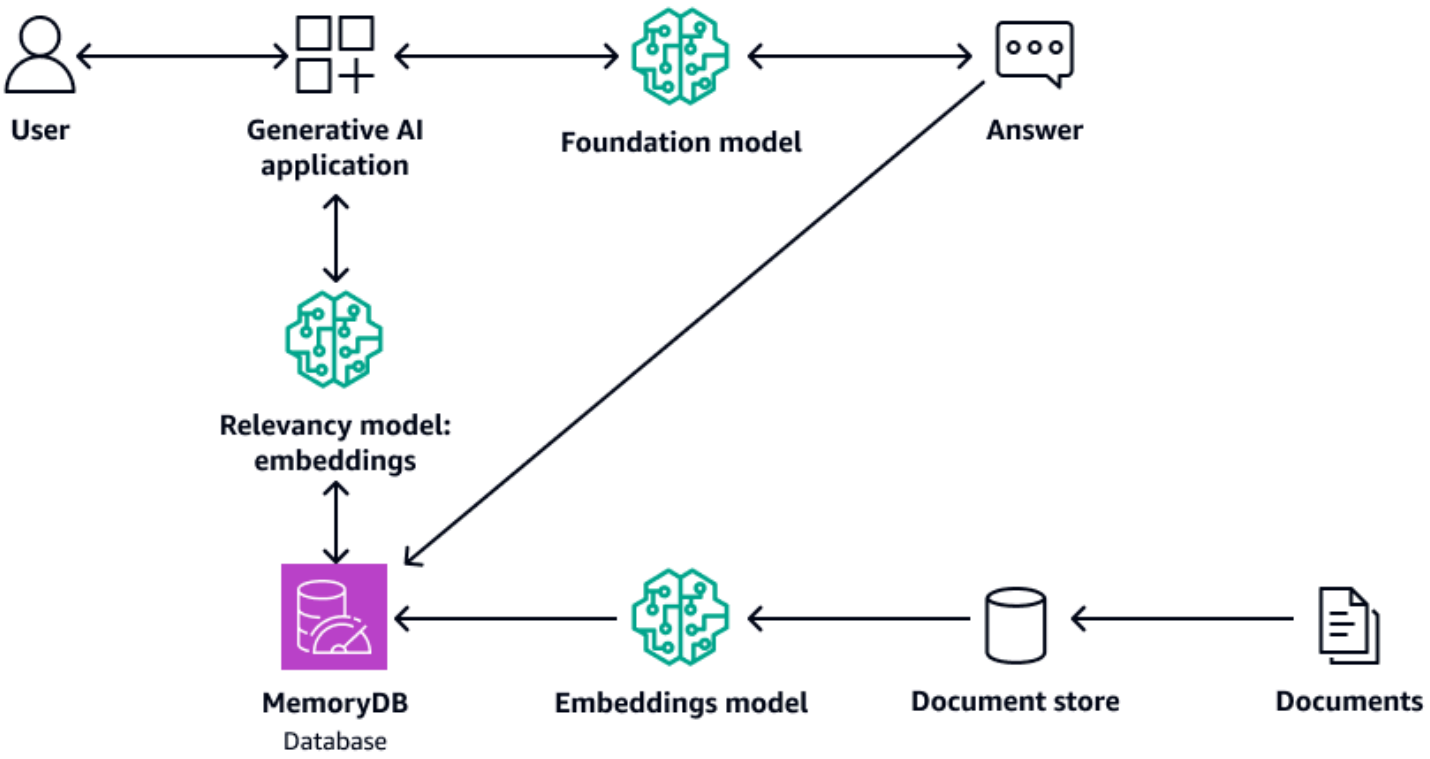
Di seguito sono riportati i vantaggi dell'utilizzo di Neptune Analytics:

- È possibile archiviare e cercare gli incorporamenti nelle query grafiche.
- Se integri Neptune Analytics LangChain con, questa architettura supporta le query grafiche in linguaggio naturale.
- Questa architettura archivia grandi set di dati grafici in memoria.

Amazon MemoryDB

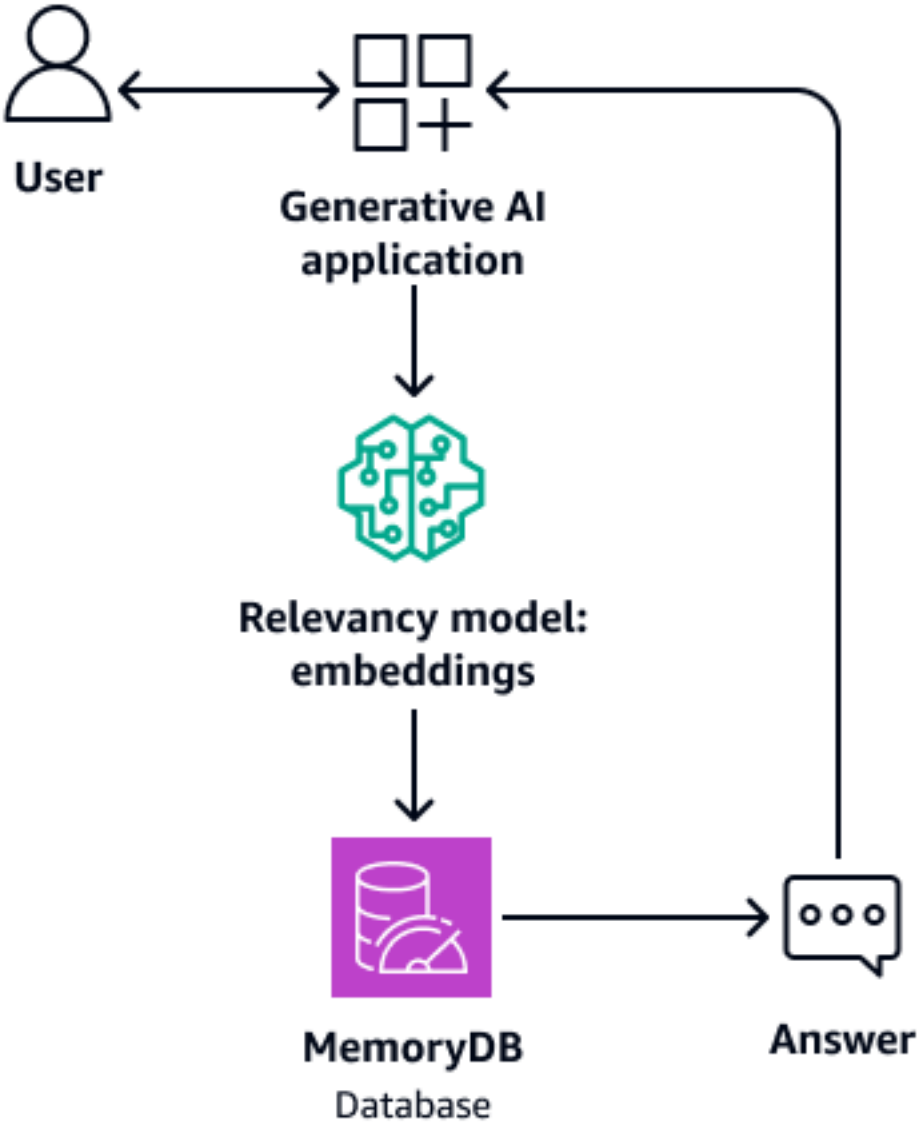
Amazon MemoryDB è un servizio di database in memoria durevole che offre prestazioni ultraveloci. Tutti i dati sono archiviati in memoria, che supporta la lettura in microsecondi, una latenza di scrittura di una cifra in millisecondi e un throughput elevato. La ricerca vettoriale per MemoryDB estende le funzionalità di MemoryDB e può essere utilizzata insieme alle funzionalità di MemoryDB esistenti. Per ulteriori informazioni, consulta la sezione Risposte alle domande con il repository LLM e RAG su. GitHub

Il diagramma seguente mostra un'architettura di esempio che utilizza MemoryDB come database vettoriale.



Di seguito sono riportati i vantaggi dell'utilizzo di MemoryDB:

- Supporta algoritmi di indicizzazione sia Flat che HNSW. Per ulteriori informazioni, consulta Vector search for Amazon MemoryDB è ora disponibile a tutti nel News Blog AWS
- Può anche fungere da memoria buffer per il modello di base. Ciò significa che le domande a cui si è risposto in precedenza vengono recuperate dal buffer anziché ripetere il processo di recupero e generazione. Il diagramma seguente mostra questo processo.



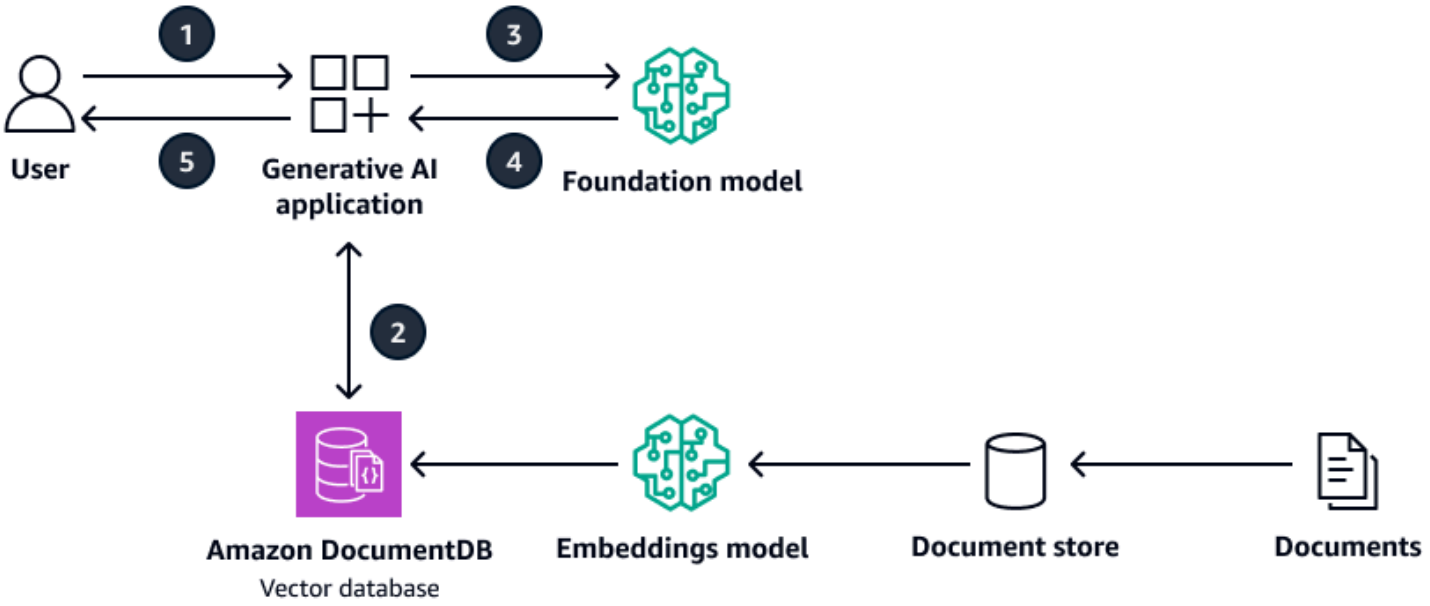
- Poiché utilizza un database in memoria, questa architettura fornisce un tempo di interrogazione di una cifra di millisecondi per la ricerca semantica.
- Fornisce fino a 33.000 query al secondo con un richiamo del 95-99% e 26.500 query al secondo con un richiamo superiore al 99%. Per ulteriori informazioni, guarda il video AWS re:Invent 2023 - Ricerca vettoriale a latenza ultra bassa per Amazon MemoryDB su. YouTube

Amazon DocumentDB

Amazon DocumentDB (con compatibilità con MongoDB) è un servizio di database veloce, affidabile e completamente gestito. Semplifica la configurazione, il funzionamento e la scalabilità di database MongoDB compatibili nel cloud. La ricerca vettoriale per Amazon DocumentDB combina la flessibilità

e la ricca capacità di interrogazione di un database di documenti basato su JSON con la potenza della ricerca vettoriale. Per ulteriori informazioni, consulta il repository Question response with LLM and RAG on. GitHub

Il diagramma seguente mostra un'architettura di esempio che utilizza Amazon DocumentDB come database vettoriale.



Il diagramma mostra il flusso di lavoro seguente:

1. L'utente invia una query all'applicazione AI generativa.
2. L'applicazione AI generativa esegue una ricerca di similarità nel database vettoriale Amazon DocumentDB e recupera gli estratti dei documenti pertinenti.
3. L'applicazione di intelligenza artificiale generativa aggiorna la query dell'utente con il contesto recuperato e invia il prompt al modello di base di destinazione.
4. Il modello di base utilizza il contesto per generare una risposta alla domanda dell'utente e restituisce la risposta.
5. L'applicazione AI generativa restituisce la risposta all'utente.

Di seguito sono riportati i vantaggi dell'utilizzo di Amazon DocumentDB:

- Supporta sia i metodi HNSW che quelli di indicizzazione IVFFlat .
- Supporta fino a 2.000 dimensioni nei dati vettoriali e supporta le metriche della distanza dei prodotti euclidei, coseni e scalari.

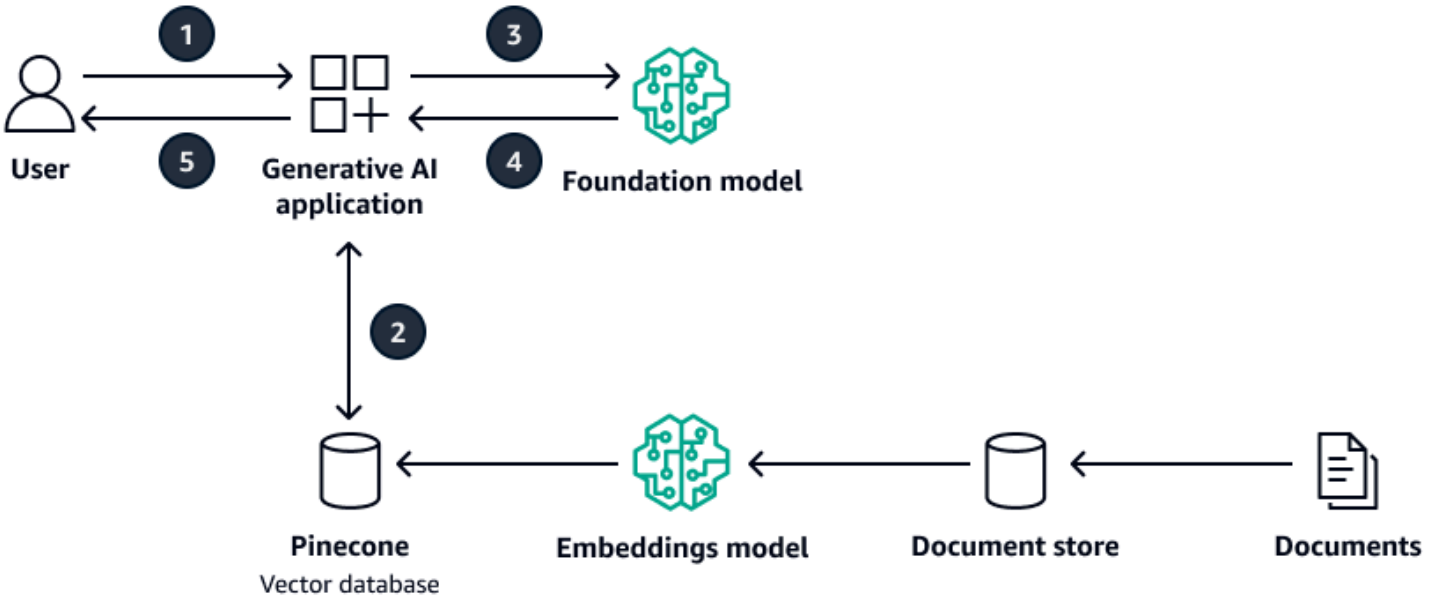
- Fornisce tempi di risposta di millisecondi.

Pinecone

Pineconeè un database vettoriale completamente gestito che consente di aggiungere la ricerca vettoriale alle applicazioni di produzione. È disponibile tramite. Marketplace AWS La fatturazione si basa sull'utilizzo e gli addebiti vengono calcolati moltiplicando il prezzo del pod per il numero di pod. Per ulteriori informazioni su come creare un sistema basato su RAG che utilizziPinecone, consulta i seguenti post del blog: AWS

- Mitiga le allucinazioni tramite RAG utilizzando il Pinecone database vettoriale e Llama-2 di Amazon AI SageMaker JumpStart
- Usa Amazon SageMaker AI Studio per creare una soluzione RAG per rispondere alle domande con Llama 2 e Pinecone per una rapida LangChain sperimentazione

Il diagramma seguente mostra un'architettura di esempio che utilizza Pinecone come database vettoriale.



Il diagramma mostra il flusso di lavoro seguente:

1. L'utente invia una query all'applicazione AI generativa.
2. L'applicazione AI generativa esegue una ricerca di somiglianza nel database Pinecone vettoriale e recupera gli estratti dei documenti pertinenti.

3. L'applicazione di intelligenza artificiale generativa aggiorna la query dell'utente con il contesto recuperato e invia il prompt al modello di base di destinazione.
4. Il modello di base utilizza il contesto per generare una risposta alla domanda dell'utente e restituisce la risposta.
5. L'applicazione AI generativa restituisce la risposta all'utente.

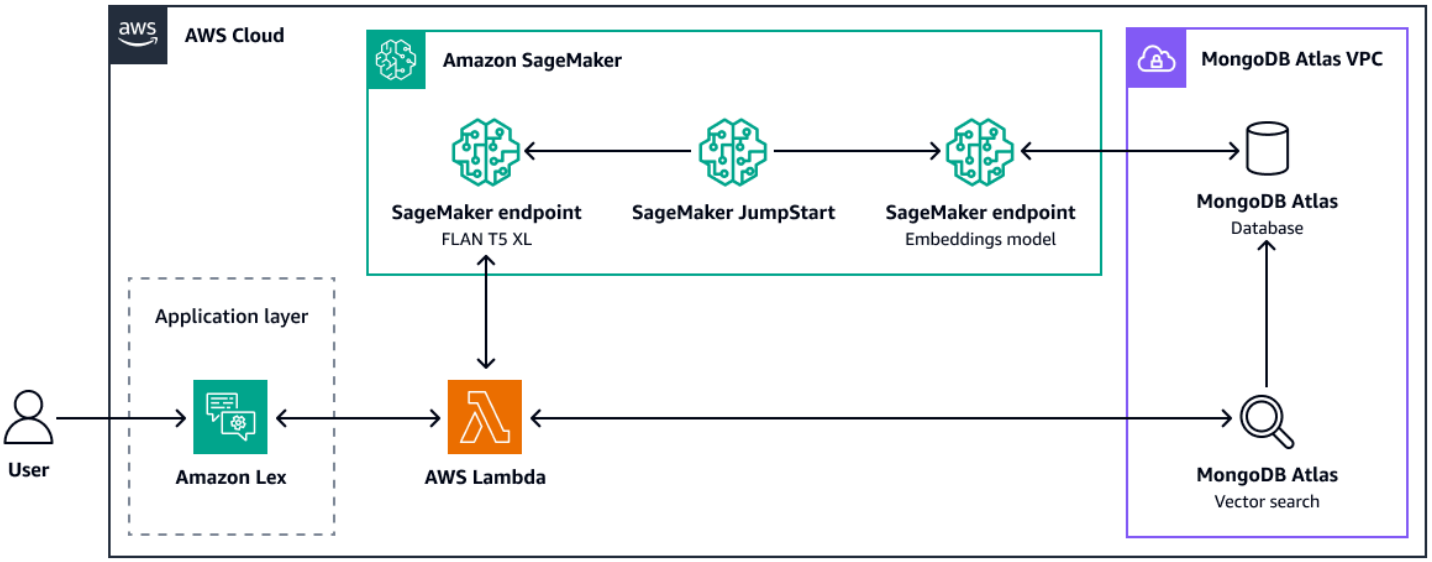
Di seguito sono riportati i vantaggi dell'utilizzoPinecone:

- È un database vettoriale completamente gestito che elimina il sovraccarico di gestione della propria infrastruttura.
- Fornisce funzionalità aggiuntive di filtraggio, aggiornamenti in tempo reale degli indici e potenziamento delle parole chiave (ricerca ibrida).

MongoDB Atlas

MongoDB Atlasè un database cloud completamente gestito che gestisce tutta la complessità dell'implementazione e della gestione delle distribuzioni su. AWS Puoi usare la ricerca vettoriale per MongoDB Atlas archiviare gli incorporamenti vettoriali nel tuo database. MongoDB Le knowledge base di Amazon Bedrock supportano MongoDB Atlas lo storage vettoriale. Per ulteriori informazioni, consulta Get Started with the Amazon Bedrock Knowledge Base Integration nella MongoDB documentazione.

Per ulteriori informazioni su come utilizzare la ricerca MongoDB Atlas vettoriale per RAG, consulta Retrieval-Augmented Generation withLangChain, SageMaker Amazon JumpStart AI MongoDB Atlas e Semantic Search (post del blog).AWS Il diagramma seguente mostra l'architettura della soluzione dettagliata in questo post del blog.



Di seguito sono riportati i vantaggi dell'utilizzo della ricerca vettoriale: MongoDB Atlas

- È possibile utilizzare l'implementazione esistente di MongoDB Atlas per archiviare e cercare incorporamenti vettoriali.
- È possibile utilizzare l'API MongoDB Query per interrogare gli incorporamenti vettoriali.
- È possibile scalare in modo indipendente la ricerca vettoriale e il database.
- Gli incorporamenti vettoriali vengono archiviati vicino ai dati di origine (documenti), il che migliora le prestazioni di indicizzazione.

Weaviate

Weaviateè un popolare database vettoriale open source a bassa latenza che supporta tipi di media multimodali, come testo e immagini. Il database memorizza sia oggetti che vettori, il che combina la ricerca vettoriale con il filtraggio strutturato. Per ulteriori informazioni sull'utilizzo Weaviate di Amazon Bedrock per creare un flusso di lavoro RAG, consulta Crea soluzioni di intelligenza artificiale generativa pronte per l'azienda con i modelli di base Cohere in Amazon Bedrock e Weaviate il database vettoriale su (post del blog). Marketplace AWSAWS

Di seguito sono riportati i vantaggi dell'utilizzo: Weaviate

- È open source e supportato da una forte comunità.
- È progettato per la ricerca ibrida (sia vettoriali che parole chiave).

- Puoi implementarlo AWS come offerta SaaS (Managed Software as a Service) o come cluster Kubernetes.

Generatori per flussi di lavoro RAG

I modelli linguistici di grandi dimensioni (LLMs) sono modelli di deep learning molto grandi preaddestrati su grandi quantità di dati. Sono incredibilmente flessibili. LLMs possono svolgere diverse attività, come rispondere a domande, riassumere documenti, tradurre lingue e completare frasi. Hanno il potenziale di interrompere la creazione di contenuti e il modo in cui le persone utilizzano i motori di ricerca e gli assistenti virtuali. Sebbene non siano perfetti, LLMs dimostrano una notevole capacità di fare previsioni sulla base di un prompt o di un numero di input relativamente piccolo.

LLMs sono un componente fondamentale di una soluzione RAG. Per le architetture RAG personalizzate, ce ne sono due Servizi AWS che fungono da opzioni principali:

- Amazon Bedrock è un servizio completamente gestito che mette LLMs a tua disposizione le principali aziende di intelligenza artificiale e Amazon tramite un'API unificata.
- Amazon SageMaker AI JumpStart è un hub ML che offre modelli di base, algoritmi integrati e soluzioni ML predefinite. Con l' SageMaker intelligenza artificiale JumpStart, puoi accedere a modelli preaddestrati, inclusi i modelli di base. Puoi anche utilizzare i tuoi dati per perfezionare i modelli preaddestrati.

Amazon Bedrock

Amazon Bedrock offre modelli leader del settore diAnthropic,,,Stability AI, Meta CohereAI21 Labs, Mistral AI e Amazon. Per un elenco completo, consulta Modelli di base supportati in Amazon Bedrock. Amazon Bedrock ti consente anche di personalizzare i modelli con i tuoi dati.

Puoi valutare le prestazioni del modello per determinare quali sono i più adatti al tuo caso d'uso RAG. È possibile testare i modelli più recenti e verificare quali funzionalità e caratteristiche offrono i migliori risultati al miglior prezzo. Il modello Anthropic Claude Sonnet è una scelta comune per le applicazioni RAG perché eccelle in un'ampia gamma di attività e offre un elevato grado di affidabilità e prevedibilità.

SageMaker INTELLIGENZA ARTIFICIALE JumpStart

SageMaker JumpStart L'intelligenza artificiale fornisce modelli open source preaddestrati per un'ampia gamma di tipi di problemi. È possibile addestrare e perfezionare questi modelli in modo incrementale prima della distribuzione. Puoi accedere ai modelli preaddestrati, ai modelli di soluzione e agli esempi tramite la JumpStart landing page dedicata all' SageMaker intelligenza artificiale in Amazon SageMaker AI Studio o utilizzare l'SDK SageMaker AI Python.

SageMaker L'intelligenza artificiale JumpStart offre modelli di state-of-the-art base per casi d'uso come scrittura di contenuti, generazione di codice, risposta a domande, copywriting, riepilogo, classificazione, recupero di informazioni e altro ancora. Usa i modelli di JumpStart base per creare le tue soluzioni di intelligenza artificiale generativa e integra soluzioni personalizzate con funzionalità di intelligenza artificiale aggiuntive. SageMaker Per ulteriori informazioni, consulta la sezione Guida introduttiva ad Amazon SageMaker AI JumpStart.

SageMaker L'intelligenza artificiale JumpStart integra e mantiene modelli di base disponibili al pubblico per consentirti di accedere, personalizzare e integrare nei tuoi cicli di vita del machine learning. Per ulteriori informazioni, consulta Modelli di base disponibili pubblicamente. SageMaker L'intelligenza artificiale include JumpStart anche modelli di base proprietari di fornitori terzi. Per ulteriori informazioni, consulta Modelli di base proprietari.

Scelta di un'opzione Retrieval Augmented Generation su AWS

Le sezioni Opzioni RAG completamente gestite e architetture RAG personalizzate di questa guida descrivono vari approcci per la creazione di una soluzione di ricerca basata su RAG. AWS Questa sezione descrive come selezionare tra queste opzioni in base al caso d'uso. In alcune situazioni, potrebbe funzionare più di un'opzione. In questo scenario, la scelta dipende dalla facilità di implementazione, dalle competenze disponibili nell'organizzazione e dalle politiche e dagli standard aziendali.

Ti consigliamo di prendere in considerazione le opzioni RAG completamente gestite e personalizzate nella sequenza seguente e di scegliere la prima opzione adatta al tuo caso d'uso:

1. Usa Amazon Q Business a meno che:
   - Questo servizio non è disponibile nella tua regione e Regione AWS i tuoi dati non possono essere spostati in una regione in cui sono disponibili
   - Hai un motivo specifico per personalizzare il flusso di lavoro RAG
   - Vuoi usare un database vettoriale esistente o un LLM specifico
2. Utilizza le knowledge base per Amazon Bedrock a meno che:
   - Hai un database vettoriale che non è supportato
   - Hai un motivo specifico per personalizzare il flusso di lavoro RAG
3. Combina Amazon Kendra con il generatore che preferisci, a meno che:
   - Vuoi scegliere il tuo database vettoriale
   - Vuoi personalizzare la strategia di suddivisione in blocchi
4. Se desideri un maggiore controllo sul retriever e desideri selezionare il tuo database vettoriale:
   - Se non disponi di un database vettoriale esistente e non hai bisogno di query grafiche o a bassa latenza, prendi in considerazione l'utilizzo di Amazon Service. OpenSearch
   - Se disponi di un database PostgreSQL vettoriale esistente, prendi in considerazione l'utilizzo di Amazon Aurora PostgreSQL pgvector and option.
   - Se hai bisogno di una bassa latenza, prendi in considerazione un'opzione in memoria, come Amazon MemoryDB o Amazon DocumentDB.
   - Se desideri combinare la ricerca vettoriale con una query grafica, prendi in considerazione Amazon Neptune Analytics.

- Se utilizzi già un database vettoriale di terze parti o ne trovi uno specifico vantaggio, prendi in considerazione Pinecone l'idea di e. MongoDB Atlas Weaviate

5. Se vuoi scegliere un LLM:
   - Se usi Amazon Q Business, non puoi scegliere il LLM.
   - Se usi Amazon Bedrock, puoi scegliere uno dei modelli di base supportati.
   - Se utilizzi Amazon Kendra o un database vettoriale personalizzato, puoi utilizzare uno dei generatori descritti in questa guida o utilizzare un LLM personalizzato.

   > **Note**
   >
   > Puoi anche utilizzare i tuoi documenti personalizzati per perfezionare un LLM esistente e aumentare la precisione delle sue risposte. Per ulteriori informazioni sul tagging, consulta Confronto tra RAG e fine-tuning in questa guida.

6. Se disponi di un'implementazione esistente di Amazon SageMaker AI Canvas che desideri utilizzare o se desideri confrontare le risposte RAG di diverse applicazioni LLMs, prendi in considerazione Amazon SageMaker AI Canvas.

Conclusioni

Questa guida descrive le varie opzioni su cui costruire un sistema Retrieval Augmented Generation (RAG). AWS Puoi iniziare con servizi completamente gestiti, come le knowledge base Amazon Q Business e Amazon Bedrock. Se desideri un maggiore controllo sul flusso di lavoro RAG, puoi scegliere un retriever personalizzato. Come generatore, puoi utilizzare un'API per chiamare un LLM supportato in Amazon Bedrock oppure puoi implementare il tuo LLM utilizzando Amazon AI. SageMaker JumpStart Consulta i consigli in Scegliere un'opzione RAG per determinare l'opzione più adatta al tuo caso d'uso. Dopo aver selezionato l'opzione migliore per il vostro caso d'uso, utilizzate i riferimenti forniti in questa guida per iniziare a creare la vostra applicazione basata su RAG.

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	—	28 ottobre 2024

AWS Glossario delle linee guida prescrittive

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

Numeri

7 R

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

- Rifattorizzare/riprogettare: trasferisci un'applicazione e modifica la sua architettura sfruttando appieno le funzionalità native del cloud per migliorare l'agilità, le prestazioni e la scalabilità. Ciò comporta in genere la portabilità del sistema operativo e del database. Esempio: migra il tuo database Oracle locale all'edizione compatibile con Amazon Aurora PostgreSQL.
- Ridefinire la piattaforma (lift and reshape): trasferisci un'applicazione nel cloud e introduci un certo livello di ottimizzazione per sfruttare le funzionalità del cloud. Esempio: migra il tuo database Oracle locale ad Amazon Relational Database Service (Amazon RDS) per Oracle in. Cloud AWS
- Riacquistare (drop and shop): passa a un prodotto diverso, in genere effettuando la transizione da una licenza tradizionale a un modello SaaS. Esempio: migra il tuo sistema di gestione delle relazioni con i clienti (CRM) su Salesforce.com.
- Eseguire il rehosting (lift and shift): trasferisci un'applicazione sul cloud senza apportare modifiche per sfruttare le funzionalità del cloud. Esempio: migra il tuo database Oracle locale a 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

ABAC

Vedi controllo degli accessi basato sugli attributi.

servizi astratti

Vedi servizi gestiti.

ACIDO

Vedi atomicità, consistenza, isolamento, durata.

migrazione attiva-attiva

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

migrazione attiva-passiva

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

funzione di aggregazione

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

Intelligenza artificiale

Vedi intelligenza artificiale.

AIOps

Guarda le operazioni di intelligenza artificiale.

anonimizzazione

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

anti-modello

Una soluzione utilizzata 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 AIOps viene utilizzata nella strategia di AWS migrazione, consulta la guida all'integrazione delle operazioni.

crittografia asimmetrica

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

atomicità, consistenza, isolamento, durabilità (ACID)

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

Controllo degli accessi basato su attributi (ABAC)

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

fonte di dati autorevole

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

Zona di disponibilità

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

AWS Cloud Adoption Framework (CAF)AWS

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

AWS Workload Qualification Framework (WQF)AWS

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

B

bot difettoso

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

BCP

Vedi la pianificazione della continuità operativa.

grafico comportamentale

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

sistema big-endian

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

Classificazione binaria

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

filtro Bloom

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

implementazione blu/verde

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

bot

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

botnet

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

ramo

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

accesso break-glass

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

strategia brownfield

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

cache del buffer

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

capacità di business

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

pianificazione della continuità operativa (BCP)

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

C

CAF

Vedi Cloud Adoption AWS Framework.

implementazione canaria

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

CCoE

Vedi Cloud Center of Excellence.

CDC

Vedi Change Data Capture.

Change Data Capture (CDC)

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

ingegneria del caos

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

CI/CD

Vedi integrazione continua e distribuzione continua.

classificazione

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

crittografia lato client

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

Centro di eccellenza cloud (CCoE)

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

cloud computing

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

modello operativo cloud

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

fasi di adozione del cloud

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

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

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

CMDB

Vedi database di gestione della configurazione.

repository di codice

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

cache fredda

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

dati freddi

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

visione artificiale (CV)

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

deriva della configurazione

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

database di gestione della configurazione (CMDB)

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

Pacchetto di conformità

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

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

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

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

CV

Vedi visione artificiale.

D

dati a riposo

Dati stazionari nella rete, ad esempio i dati archiviati.

classificazione dei dati

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

deriva dei dati

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

dati in transito

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

rete di dati

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

riduzione al minimo dei dati

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

perimetro dei dati

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

pre-elaborazione dei dati

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

provenienza dei dati

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

soggetto dei dati

Un individuo i cui dati vengono raccolti ed elaborati.

data warehouse

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

linguaggio di definizione del database (DDL)

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

linguaggio di manipolazione del database (DML)

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

DDL

Vedi linguaggio di definizione del database.

deep ensemble

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

deep learning

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

defense-in-depth

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

amministratore delegato

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

implementazione

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

Ambiente di sviluppo

Vedi ambiente.

controllo di rilevamento

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

mappatura del flusso di valore dello sviluppo (DVSM)

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

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

gemello digitale

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

tabella delle dimensioni

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

disastro

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

disaster recovery (DR)

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

DML

Vedi linguaggio di manipolazione del database.

progettazione basata sul dominio

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

DOTT.

Vedi disaster recovery.

rilevamento della deriva

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

DVSM

Vedi la mappatura del flusso di valore dello sviluppo.

E

EDA

Vedi analisi esplorativa dei dati.

MODIFICA

Vedi scambio elettronico di dati.

edge computing

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

scambio elettronico di dati (EDI)

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

crittografia

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

chiave crittografica

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

endianità

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

endpoint

Vedi service endpoint.

servizio endpoint

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

pianificazione delle risorse aziendali (ERP)

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

crittografia envelope

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

ambiente

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

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

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

epica

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

ERP

Vedi pianificazione delle risorse aziendali.

analisi esplorativa dei dati (EDA)

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

F

tabella dei fatti

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

fallire velocemente

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

limite di isolamento dei guasti

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

ramo di funzionalità

Vedi filiale.

caratteristiche

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

importanza delle caratteristiche

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

trasformazione delle funzionalità

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

prompt con pochi scatti

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

FGAC

Vedi il controllo granulare degli accessi.

controllo granulare degli accessi (FGAC)

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

migrazione flash-cut

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

FM

Vedi modello di base.

modello di fondazione (FM)

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

G

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

strategia greenfield

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

guardrail

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

H

AH

Vedi disponibilità elevata.

migrazione di database eterogenea

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

alta disponibilità (HA)

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

modernizzazione storica

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

dati di blocco

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

migrazione di database omogenea

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

dati caldi

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

hotfix

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

periodo di hypercare

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

I

IaC

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 in Well-Architected AWS Framework.

VPC in ingresso (ingress)

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

migrazione incrementale

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

Industria 4.0

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

infrastruttura

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

infrastruttura come codice (IaC)

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

IIoInternet delle cose industriale (T)

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

VPC di ispezione

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

Internet of Things (IoT)

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

interpretabilità

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

IoT

Vedi Internet of Things.

libreria di informazioni IT (ITIL)

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

gestione dei servizi IT (ITSM)

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

ITIL

Vedi la libreria di informazioni IT.

ITSM

Vedi Gestione dei servizi IT.

L

controllo degli accessi basato su etichette (LBAC)

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

zona di destinazione

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

modello linguistico di grandi dimensioni (LLM)

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

migrazione su larga scala

Una migrazione di 300 o più server.

BIANCO

Vedi controllo degli accessi basato su etichette.

Privilegio minimo

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

eseguire il rehosting (lift and shift)

Vedi 7 R.

sistema little-endian

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

LLM

Vedi modello linguistico di grandi dimensioni.

ambienti inferiori

Vedi ambiente.

M

machine learning (ML)

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

ramo principale

Vedi filiale.

malware

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

servizi gestiti

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

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

sistema di esecuzione della produzione (MES)

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

MAP

Vedi Migration Acceleration Program.

meccanismo

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

account membro

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

MEH

Vedi sistema di esecuzione della produzione.

Message Queuing Telemetry Transport (MQTT)

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

microservizio

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

architettura di microservizi

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

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

Programma di accelerazione della migrazione (MAP)

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

migrazione su larga scala

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

fabbrica di migrazione

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

metadati di migrazione

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

modello di migrazione

Un'attività di migrazione ripetibile che descrive in dettaglio la strategia di migrazione, la destinazione della migrazione e l'applicazione o il servizio di migrazione utilizzati. Esempio: riorganizza la migrazione su Amazon EC2 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 machine-to-machine (M2M) per l'automazione industriale. OPC-UA fornisce uno standard di interoperabilità con schemi di crittografia, autenticazione e autorizzazione dei dati.

accordo a livello operativo (OLA)

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

revisione della prontezza operativa (ORR)

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

tecnologia operativa (OT)

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

integrazione delle operazioni (OI)

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

trail organizzativo

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

gestione del cambiamento organizzativo (OCM)

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

controllo dell'accesso all'origine (OAC)

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

identità di accesso origine (OAI)

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

ORR

Vedi la revisione della prontezza operativa.

NON

Vedi la tecnologia operativa.

VPC in uscita (egress)

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

P

limite delle autorizzazioni

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

informazioni di identificazione personale (PII)

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

Informazioni che consentono l'identificazione personale degli utenti

Visualizza le informazioni di identificazione personale.

playbook

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

PLC

Vedi controllore logico programmabile.

PLM

Vedi la gestione del ciclo di vita del prodotto.

policy

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

persistenza poliglotta

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

valutazione del portfolio

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

predicate

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

predicato pushdown

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

controllo preventivo

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

principale

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

privacy fin dalla progettazione

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

zone ospitate private

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

controllo proattivo

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

gestione del ciclo di vita del prodotto (PLM)

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

Ambiente di produzione

Vedi ambiente.

controllore logico programmabile (PLC)

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

concatenamento rapido

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

pseudonimizzazione

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

publish/subscribe (pub/sub)

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

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

Q

Piano di query

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

regressione del piano di query

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

R

Matrice RACI

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

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

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

endpoint del servizio

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

accordo sul livello di servizio (SLA)

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

indicatore del livello di servizio (SLI)

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

obiettivo a livello di servizio (SLO)

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

Modello di responsabilità condivisa

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

SIEM

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

punto di errore singolo (SPOF)

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

SLAM

Vedi il contratto sul livello di servizio.

SLI

Vedi l'indicatore del livello di servizio.

LENTA

Vedi obiettivo del livello di servizio.

split-and-seed modello

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

SPOF

Vedi punto di errore singolo.

schema a stella

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

modello del fico strangolatore

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

sottorete

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

controllo di supervisione e acquisizione dati (SCADA)

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

crittografia simmetrica

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

test sintetici

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

prompt di sistema

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

T

tag

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

variabile di destinazione

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

elenco di attività

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

ambiente di test

Vedi ambiente.

training

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

Transit Gateway

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

flusso di lavoro basato su trunk

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

Accesso attendibile

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

regolazione

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

team da due pizze

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

U

incertezza

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

compiti indifferenziati

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

ambienti superiori

Vedi ambiente.

V

vacuum

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

controllo delle versioni

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

Peering VPC

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

vulnerabilità

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

W

cache calda

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

dati caldi

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

funzione finestra

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

Carico di lavoro

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

flusso di lavoro

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

VERME

Vedi scrivere una volta, leggere molti.

WQF

Vedi AWS Workload Qualification Framework.

scrivi una volta, leggi molte (WORM)

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

Z

exploit zero-day

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

vulnerabilità zero-day

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

prompt zero-shot

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

applicazione zombie

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

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