Utilizzo di Amazon Comprehend Medical LLMs e per il settore sanitario e delle scienze biologiche

AWS Guida prescrittiva

AWS Guida prescrittiva: Utilizzo di Amazon Comprehend Medical LLMs e per il settore sanitario e delle scienze biologiche

Table of Contents

Utilizzo di Amazon Comprehend Medical LLMs e per il settore sanitario e delle scienze biologiche

Amazon Web Services (???collaboratori)

Dicembre 2025 (cronologia dei documenti)

Panoramica di

Il volume sempre crescente di dati medici e la necessità di un'elaborazione efficiente e accurata hanno portato all'adozione dell'elaborazione del linguaggio naturale (NLP) con tecnologie di intelligenza artificiale e apprendimento automatico (AI/ML). I modelli di classificazione predefiniti e i modelli linguistici di grandi dimensioni (LLMs) si sono affermati come potenti strumenti per varie attività di PNL in ambito medico, tra cui la risposta a domande cliniche, il riepilogo di report e la generazione di approfondimenti. Tuttavia, il settore sanitario e delle scienze biologiche presenta sfide uniche a causa della complessità della terminologia medica, delle conoscenze specifiche del settore e dei requisiti normativi. L'uso efficace di classificatori già addestrati o LLMs in questo settore richiede un approccio ben progettato che combini i punti di forza di questi modelli con risorse e tecniche specifiche del dominio.

Le pratiche del settore sanitario e delle scienze della vita si sono tradizionalmente basate su sistemi basati su regole, codifica manuale e processi di revisione esperti. Questi sistemi e processi richiedono molto tempo e sono soggetti a errori. L'integrazione di tecnologie AI e NLP, come Amazon Comprehend Medical e i modelli base di Amazon Bedrock, offre soluzioni efficienti e scalabili per l'elaborazione dei dati medici, migliorando al contempo la precisione e la coerenza.

Questa guida esplora l'uso di Amazon Comprehend Medical e LLMs l'automazione intelligente nel settore sanitario. Descrive le migliori pratiche, le sfide e gli approcci pratici per semplificare i processi di codifica medica, estrazione delle informazioni sui pazienti e riepilogo dei registri. Utilizzando le funzionalità di Amazon Comprehend Medical LLMs and, le organizzazioni sanitarie possono sbloccare nuovi livelli di efficienza operativa, ridurre i costi e potenzialmente migliorare l'assistenza ai pazienti.

La guida descrive in dettaglio le considerazioni specifiche del settore sanitario, come la comprensione della terminologia medica, l'uso di domini specifici e la risoluzione LLMs dei limiti dei sistemi. AI/ML Fornisce un percorso decisionale completo per i responsabili IT, gli architetti e i responsabili tecnici

del settore sanitario per valutare la prontezza organizzativa, valutare le opzioni di implementazione e utilizzare gli strumenti appropriati Servizi AWS per un'automazione di successo.

Seguendo le linee guida e le migliori pratiche descritte in questa guida, le organizzazioni sanitarie possono sfruttare la potenza delle AI/ML tecnologie mentre affrontano le complessità del settore medico. Questo approccio supporta la conformità alle linee guida etiche e normative e promuove l'uso responsabile dei sistemi di intelligenza artificiale nel settore sanitario. È progettato per generare approfondimenti accurati e privati.

Destinatari principali

Questa guida è destinata agli stakeholder tecnologici, agli architetti, ai responsabili tecnici e ai responsabili delle decisioni che desiderano implementare soluzioni di elaborazione del linguaggio naturale basate sull'intelligenza artificiale per l'analisi e l'automazione dei dati medici.

Obiettivi

Le organizzazioni sanitarie e delle scienze della vita possono raggiungere diversi obiettivi aziendali utilizzando Amazon Comprehend Medical LLMs e. Questi risultati includono in genere l'aumento dell'efficienza operativa, la riduzione dei costi e il miglioramento dell'assistenza ai pazienti. Questa sezione descrive gli obiettivi aziendali chiave e i vantaggi associati all'implementazione delle strategie e delle migliori pratiche descritte in questa guida.

Di seguito sono riportati alcuni degli obiettivi che le organizzazioni possono raggiungere implementando le linee guida e le migliori pratiche contenute in questa guida:

- Ridurre i tempi di sviluppo: l'obiettivo finale di questa guida è ridurre i tempi di sviluppo e i costi, ridurre il debito tecnico e mitigare il potenziale fallimento del progetto dovuto al POC. Comprendendo AI/ML i servizi chiave, come Amazon Comprehend Medical, e i vantaggi e i limiti dell'utilizzo del LLM per le attività sanitarie, le aziende possono accelerare il time-to-market e accelerare il raggiungimento degli obiettivi aziendali.
- Estrai informazioni per automatizzare le attività di codifica medica: dopo le visite dei pazienti, gli specialisti della programmazione e i fornitori possono estrarre informazioni dal testo medico, come note soggettive, oggettive, di valutazione e pianificazione (SOAP). Ciò può ridurre gli sforzi di documentazione manuale e aiutare il fornitore a concentrarsi sulle esigenze del paziente. Combinando le funzionalità di riconoscimento delle entità di Amazon Comprehend Medical LLMs con, le organizzazioni possono estrarre informazioni mediche pertinenti dalle cartelle cliniche dei

pazienti, dalle note cliniche e da altre fonti di dati sanitari. Questo può ridurre al minimo gli errori umani e promuovere pratiche coerenti.

- Riepilogo delle cartelle cliniche e delle cartelle cliniche dei pazienti: il riepilogo automatico della storia del paziente, dei piani di trattamento e dei risultati medici può far risparmiare tempo prezioso agli operatori sanitari. LLMs può aiutare a generare una documentazione clinica completa e strutturata. Puoi ottenere un contesto aggiuntivo con Amazon Comprehend Medical, utilizzare un LLM di dominio medico o perfezionare un LLM con dati medici. Questi approcci possono aiutare a fornire riepiloghi accurati e a garantire che la documentazione sia conforme ai requisiti e agli standard di conformità.
- Supporta le decisioni cliniche e l'assistenza ai pazienti: utilizzando il collegamento ontologico in Amazon Comprehend Medical e LLMs utilizzando, i fornitori possono rispondere a domande mediche o chiedere consigli sull'assistenza ai pazienti. Ciò consente agli operatori sanitari di prendere decisioni informate che migliorano gli esiti dei pazienti e riducono il rischio di errori medici.

Approcci generativi di intelligenza artificiale e PNL per l'assistenza sanitaria e le scienze della vita

L'elaborazione del linguaggio naturale (NLP) è una tecnologia di apprendimento automatico che offre ai computer la capacità di interpretare, manipolare e comprendere il linguaggio umano. Le organizzazioni sanitarie e delle scienze della vita dispongono di grandi volumi di dati provenienti dalle cartelle cliniche dei pazienti. Possono utilizzare il software NLP per elaborare automaticamente questi dati. Ad esempio, possono combinare la PNL con l'intelligenza artificiale generativa per semplificare la codifica medica, estrarre informazioni sui pazienti e riepilogare i record.

A seconda dell'attività di PNL che si desidera eseguire, diverse architetture potrebbero essere più adatte al caso d'uso. Questa guida affronta le seguenti opzioni generative di intelligenza artificiale e PNL per applicazioni nel settore sanitario e delle scienze della vita su: AWS

- Utilizzo di Amazon Comprehend Medical— Scopri come usare Amazon Comprehend Medical in modo indipendente, senza integrarlo con un modello di linguaggio di grandi dimensioni (LLM).
- Combinazione di Amazon Comprehend Medical con modelli linguistici di grandi dimensioni— Scopri come combinare Amazon Comprehend Medical con un LLM in un'architettura Retrieval Augment Generation (RAG).
- Utilizzo di modelli linguistici di grandi dimensioni per casi d'uso nel settore sanitario e delle scienze della vita— Scopri come utilizzare un LLM per applicazioni sanitarie e biologiche, utilizzando un LLM ottimizzato o un'architettura RAG.

Utilizzo di Amazon Comprehend Medical

Amazon Comprehend Medical rileva e restituisce informazioni utili in testo clinico non strutturato come note mediche, riepiloghi delle dimissioni, risultati dei test e note sui casi. Servizio AWS Utilizza modelli di elaborazione del linguaggio naturale (NLP) per rilevare le entità. Le entità sono riferimenti testuali a informazioni mediche, come condizioni mediche, farmaci o informazioni sanitarie protette (PHI).

> **Important**
>
> Amazon Comprehend Medical non sostituisce consulenze, diagnosi o trattamenti medici professionali. Amazon Comprehend Medical fornisce punteggi di affidabilità che indicano

il livello di fiducia nell'accuratezza delle entità rilevate. Identificare la soglia di confidenza giusta per il caso d'uso e utilizzare soglie di confidenza elevata in situazioni che richiedono un'elevata precisione. In alcuni casi d'uso, i risultati devono essere esaminati e verificati da revisori umani adeguatamente formati. Ad esempio, Amazon Comprehend Medical deve essere usato in scenari di assistenza ai pazienti solo dopo aver verificato l'accuratezza e l'attendibilità del giudizio medico da parte di professionisti medici qualificati.

Puoi accedere ad Amazon Comprehend Medical tramite, Console di gestione AWS AWS Command Line Interface il AWS CLI() o tramite AWS SDKs. AWS SDKs Sono disponibili per vari linguaggi e piattaforme di programmazione, come Java, Python, Ruby, .NET, iOS e Android. Puoi utilizzarlo per accedere in modo programmatico SDKs ad Amazon Comprehend Medical dalla tua applicazione client.

Questa sezione esamina le funzionalità principali di Amazon Comprehend Medical. Descrive inoltre i vantaggi dell'utilizzo di questo servizio rispetto a un modello linguistico di grandi dimensioni (LLM).

Funzionalità di Amazon Comprehend Medical

Amazon Comprehend Medical APIs offre inferenze quasi in tempo reale e in batch. Questi APIs possono assimilare testo medico e fornire risultati per le attività di PNL in ambito medico utilizzando il riconoscimento delle entità mediche e l'identificazione delle relazioni tra le entità. Puoi eseguire analisi su singoli file o in batch su più file archiviati in un bucket Amazon Simple Storage Service (Amazon S3). Amazon Comprehend Medical offre le seguenti operazioni API di analisi del testo per il rilevamento sincrono di entità:

- Rileva entità: rileva categorie mediche generali come anatomia, condizione medica, categoria PHI, procedure ed espressioni temporali.
- Rileva PHI: rileva entità specifiche come età, data, nome e informazioni personali simili.

Amazon Comprehend Medical include anche diverse operazioni API che puoi utilizzare per eseguire analisi di testo in batch su documenti clinici. Per ulteriori informazioni su come utilizzare queste operazioni API, consulta Text analysis batch APIs.

Usa Amazon Comprehend Medical per rilevare entità nel testo clinico e collegarle a concetti di ontologie mediche standardizzate, tra cui RxNorm le knowledge base ICD-10-CM e SNOMED CT. Puoi eseguire analisi sia su singoli file che come analisi in batch su documenti di grandi dimensioni o

più file archiviati in un bucket Amazon S3. Amazon Comprehend Medical offre le seguenti operazioni di collegamento ontologico delle API:

- Infer ICD10 CM — L'operazione Infer ICD10 CM rileva potenziali condizioni mediche e le collega ai codici della versione 2019 della classificazione internazionale delle malattie, decima revisione, modifica clinica (ICD-10-CM). Per ogni potenziale condizione medica rilevata, Amazon Comprehend Medical elenca i codici e le descrizioni ICD-10-CM corrispondenti. Le condizioni mediche elencate nei risultati includono un punteggio di confidenza, che indica la fiducia di Amazon Comprehend Medical nell'accuratezza delle entità rispetto ai concetti corrispondenti nei risultati.
- InferRxNorm— L'InferRxNormoperazione identifica i farmaci elencati nella cartella clinica di un paziente come entità. Collega le entità agli identificatori concettuali (RxCUI) presenti RxNorm nel database della National Library of Medicine. Ogni RxCUI è unico per diversi dosaggi e dosaggi. I farmaci elencati nei risultati includono un punteggio di confidenza, che indica la fiducia di Amazon Comprehend Medical nell'accuratezza delle entità corrispondenti ai concetti RxNorm della knowledge base. Amazon Comprehend Medical elenca i migliori CUIs Rx potenzialmente corrispondenti per ogni farmaco rilevato in ordine decrescente in base al punteggio di confidenza.
- InfersnomeDCT — L'operazione InfersnomeDCT identifica i possibili concetti medici come entità e li collega ai codici della versione 2021-03 della nomenclatura sistematica della medicina, termini clinici (SNOMED CT). SNOMED CT fornisce un vocabolario completo di concetti medici, tra cui condizioni mediche e anatomia, nonché test, trattamenti e procedure mediche. Per ogni concept ID corrispondente, Amazon Comprehend Medical restituisce i cinque concetti medici principali, ciascuno con un punteggio di confidenza e informazioni contestuali come tratti e attributi. Il concetto SNOMED CT IDs può quindi essere utilizzato per strutturare i dati clinici dei pazienti per la codifica, la reportistica o l'analisi clinica se utilizzato con la poligerarchia SNOMED CT.

Per ulteriori informazioni, consulta Analisi del testo APIs e collegamento ontologico APIs nella documentazione di Amazon Comprehend Medical.

Casi d'uso per Amazon Comprehend Medical

Come servizio autonomo, Amazon Comprehend Medical potrebbe risolvere i casi d'uso della tua organizzazione. Amazon Comprehend Medical può eseguire attività come le seguenti:

- Assistenza nella codifica medica nelle cartelle cliniche dei pazienti
- Rileva i dati delle informazioni sanitarie protette (PHI)
- Convalida dei farmaci, inclusi attributi quali dosaggio, frequenza e forma

I risultati di Amazon Comprehend Medical sono comprensibili per la maggior parte degli studi medici. Tuttavia, potresti dover prendere in considerazione delle alternative se hai limitazioni come le seguenti:

- Definizioni di entità diverse: ad esempio, la tua definizione FREQUENCY di entità medicinale potrebbe essere diversa. Per quanto riguarda la frequenza, Amazon Comprehend Medical prevede in base alle necessità, ma la tua organizzazione potrebbe utilizzare il termine pro re nata (PRN).
- Enorme quantità di risultati: ad esempio, le note dei pazienti contengono spesso più sintomi e parole chiave che corrispondono a più codici ICD-10-CM. Tuttavia, molte delle parole chiave non sono applicabili alla diagnosi. In questo caso, il fornitore deve valutare numerose entità ICD-10-CM e i relativi punteggi di affidabilità, il che richiede tempi di elaborazione manuali.
- Entità personalizzate o attività di NLP: ad esempio, i fornitori potrebbero voler estrarre prove PRN, ad esempio prenderle quando necessario in caso di difficoltà. Poiché non è disponibile tramite Amazon Comprehend Medical, è garantito un modello AI/ML diverso. È necessaria una AI/ML soluzione diversa se l'attività di PNL non rientra nel campo del riconoscimento delle entità, ad esempio se si tratta di riepilogo, risposta a domande e analisi del sentiment.

Combinazione di Amazon Comprehend Medical con modelli linguistici di grandi dimensioni

Uno studio del 2024 condotto da NEJM AI ha dimostrato che l'utilizzo di un LLM, con prompt zero-shot, per attività di codifica medica generalmente porta a prestazioni scadenti. L'uso di Amazon Comprehend Medical con un LLM può aiutare a mitigare questi problemi di prestazioni. I risultati di Amazon Comprehend Medical sono un contesto utile per un LLM che esegue attività di PNL. Ad esempio, fornire un contesto da Amazon Comprehend Medical al modello linguistico di grandi dimensioni può aiutarti a:

- Migliora l'accuratezza delle selezioni delle entità utilizzando i risultati iniziali di Amazon Comprehend Medical come contesto per il LLM
- Implementa il riconoscimento personalizzato delle entità, il riepilogo, la risposta alle domande e altri casi d'uso

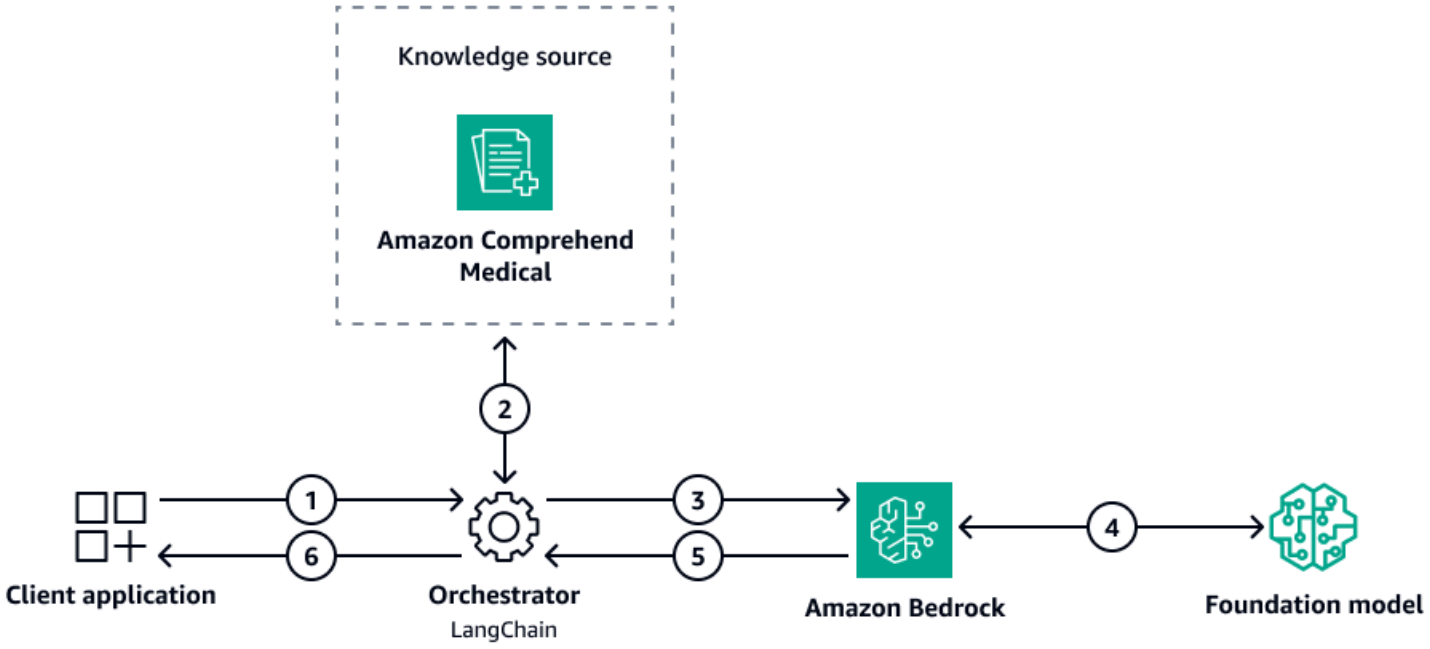
Questa sezione descrive come combinare Amazon Comprehend Medical con un LLM utilizzando un approccio Retrieval Augmented Generation (RAG). Retrieval Augmented Generation (RAG) è una tecnologia di intelligenza artificiale generativa in cui un LLM fa riferimento a una fonte di dati

autorevole esterna alle sue fonti di dati di addestramento prima di generare una risposta. Per ulteriori informazioni, consulta Cos'è il RAG.

Per illustrare questo approccio, questa sezione utilizza l'esempio di codifica medica (diagnostica) relativa all'ICD-10-CM. Include un'architettura di esempio e modelli di progettazione rapidi per accelerare l'innovazione. Include inoltre le migliori pratiche per l'utilizzo di Amazon Comprehend Medical all'interno di un flusso di lavoro RAG.

Architettura basata su RAG con Amazon Comprehend Medical

Il diagramma seguente illustra un approccio RAG per identificare i codici di diagnosi ICD-10-CM dalle note dei pazienti. Utilizza Amazon Comprehend Medical come fonte di conoscenza. In un approccio RAG, il metodo di recupero recupera in genere informazioni da un database vettoriale contenente le conoscenze applicabili. Invece di un database vettoriale, questa architettura utilizza Amazon Comprehend Medical per l'attività di recupero. L'orchestratore invia le informazioni sulla nota del paziente ad Amazon Comprehend Medical e recupera le informazioni sul codice ICD-10-CM. L'orchestratore invia questo contesto al downstream Foundation Model (LLM), tramite Amazon Bedrock. L'LLM genera una risposta utilizzando le informazioni sul codice ICD-10-CM e tale risposta viene rispedita all'applicazione client.



Il diagramma mostra il seguente flusso di lavoro RAG:

1. L'applicazione client invia le note del paziente come interrogazione all'orchestratore. Un esempio di queste annotazioni sul paziente potrebbe essere «La paziente è una paziente di 71 anni del Dr.

X. La paziente si è presentata al pronto soccorso ieri sera con una storia di dolore addominale di circa 7-8 giorni, che è stata persistente. Non ha avuto febbri o brividi precisi e nessuna storia di ittero. Il paziente nega qualsiasi significativa perdita di peso recente».

2. L'orchestratore utilizza Amazon Comprehend Medical per recuperare i codici ICD-10-CM relativi alle informazioni mediche contenute nella query. Utilizza l'API Infer CM per estrarre e dedurre i codici ICD10 ICD-10-CM dalle note del paziente.
3. L'orchestratore crea un prompt che include il modello di prompt, la query originale e i codici ICD-10-CM recuperati da Amazon Comprehend Medical. Invia questo contesto avanzato ad Amazon Bedrock.
4. Amazon Bedrock elabora l'input e utilizza un modello di base per generare una risposta che include i codici ICD-10-CM e le prove corrispondenti ricavate dalla query. La risposta generata include i codici ICD-10-CM identificati e le evidenze tratte dalle note del paziente a supporto di ciascun codice. Di seguito è riportata una risposta di esempio:

```
<response>
<icd10>
<code>R10.9</code>
<evidence>history of abdominal pain</evidence>
</icd10>
<icd10>
<code>R10.30</code>
<evidence>history of abdominal pain</evidence>
</icd10>
</response>
```

5. Amazon Bedrock invia la risposta generata all'orchestratore.
6. L'orchestratore invia la risposta all'applicazione client, dove l'utente può esaminarla.

Casi d'uso per l'utilizzo di Amazon Comprehend Medical in un flusso di lavoro RAG

Amazon Comprehend Medical può eseguire attività di PNL specifiche. Per ulteriori informazioni, consulta Casi d'uso per Amazon Comprehend Medical.

Potresti voler integrare Amazon Comprehend Medical in un flusso di lavoro RAG per casi d'uso avanzati, come i seguenti:

- Genera riepiloghi clinici dettagliati combinando entità mediche estratte con informazioni contestuali provenienti dalle cartelle cliniche dei pazienti
- Automatizza la codifica medica per casi complessi utilizzando entità estratte con informazioni collegate all'ontologia per l'assegnazione del codice
- Automatizza la creazione di note cliniche strutturate a partire da testo non strutturato utilizzando entità mediche estratte
- Analizza gli effetti collaterali dei farmaci in base ai nomi e agli attributi dei farmaci estratti
- Sviluppa sistemi di supporto clinico intelligenti che combinano le informazioni mediche estratte con la up-to-date ricerca e le linee guida

Le migliori pratiche per l'utilizzo di Amazon Comprehend Medical in un flusso di lavoro RAG

Quando si integrano i risultati di Amazon Comprehend Medical in una richiesta di LLM, è essenziale seguire le best practice. Ciò può migliorare le prestazioni e la precisione. Di seguito sono riportate le raccomandazioni principali:

- Comprendi i punteggi di confidenza di Amazon Comprehend Medical: Amazon Comprehend Medical fornisce punteggi di affidabilità per ogni entità e collegamento ontologico rilevati. È fondamentale comprendere il significato di questi punteggi e stabilire soglie appropriate per il caso d'uso specifico. I punteggi di confidenza aiutano a filtrare le entità con scarsa fiducia, riducendo il rumore e migliorando la qualità degli input del LLM.
- Usa i punteggi di confidenza nella progettazione tempestiva: quando crei istruzioni per il LLM, prendi in considerazione l'idea di incorporare i punteggi di confidenza di Amazon Comprehend Medical come contesto aggiuntivo. Questo aiuta l'LLM a stabilire le priorità o a valutare le entità in base ai loro livelli di fiducia, migliorando potenzialmente la qualità dell'output.
- Valuta i risultati di Amazon Comprehend Medical con dati fondati: i dati Ground Truth sono informazioni di cui si sa che sono vere. Possono essere usati per verificare che un' AI/ML applicazione stia producendo risultati accurati. Prima di integrare i risultati di Amazon Comprehend Medical nel tuo flusso di lavoro LLM, valuta le prestazioni del servizio su un campione rappresentativo dei tuoi dati. Confronta i risultati con annotazioni di base per identificare potenziali discrepanze o aree di miglioramento. Questa valutazione ti aiuta a comprendere i punti di forza e i limiti di Amazon Comprehend Medical per il tuo caso d'uso.

- Seleziona strategicamente le informazioni pertinenti: Amazon Comprehend Medical può fornire una grande quantità di informazioni, ma non tutte possono essere pertinenti alla tua attività. Seleziona attentamente le entità, gli attributi e i metadati più pertinenti al tuo caso d'uso. Fornire troppe informazioni irrilevanti all'LLM può causare rumore e potenzialmente ridurre le prestazioni.
- Allinea le definizioni delle entità: assicurati che le definizioni di entità e attributi utilizzate da Amazon Comprehend Medical siano in linea con la tua interpretazione. In caso di discrepanze, valuta la possibilità di fornire un contesto o un chiarimento aggiuntivi al LLM per colmare il divario tra i risultati di Amazon Comprehend Medical e i tuoi requisiti. Se l'entità Amazon Comprehend Medical non soddisfa le tue aspettative, puoi implementare il rilevamento personalizzato delle entità includendo istruzioni aggiuntive (e possibili esempi) all'interno del prompt.
- Fornisci conoscenze specifiche del dominio: sebbene Amazon Comprehend Medical fornisca preziose informazioni mediche, potrebbe non cogliere tutte le sfumature del tuo dominio specifico. Prendi in considerazione la possibilità di integrare i risultati di Amazon Comprehend Medical con ulteriori fonti di conoscenza specifiche del dominio, come ontologie, terminologie o set di dati curati da esperti. Ciò fornisce un contesto più completo al LLM.
- Rispettare le linee guida etiche e normative: quando si tratta di dati medici, è importante attenersi ai principi etici e alle linee guida normative, come quelli relativi alla privacy dei dati, alla sicurezza e all'uso responsabile dei sistemi di intelligenza artificiale nell'assistenza sanitaria. Assicurati che la tua implementazione sia conforme alle leggi pertinenti e alle migliori pratiche del settore.

Seguendo queste best practice, AI/ML i professionisti possono utilizzare efficacemente i punti di forza di Amazon Comprehend Medical e. LLMs Per le attività mediche di PNL, queste best practice aiutano a mitigare i potenziali rischi e possono migliorare le prestazioni.

Progettazione tempestiva per il contesto di Amazon Comprehend Medical

La progettazione tempestiva è il processo di progettazione e perfezionamento dei prompt per guidare una soluzione di intelligenza artificiale generativa a generare gli output desiderati. Scegli i formati, le frasi, le parole e i simboli più appropriati che guidano l'IA a interagire con i tuoi utenti in modo più significativo.

A seconda dell'operazione API eseguita, Amazon Comprehend Medical restituisce le entità rilevate, i codici e le descrizioni ontologiche e i punteggi di confidenza. Questi risultati diventano contestuali all'interno del prompt quando la soluzione richiama il LLM di destinazione. È necessario progettare il prompt per presentare il contesto all'interno del modello di prompt.

> **Note**
>
> I prompt di esempio in questa sezione seguono le indicazioni di Anthropic. Se utilizzi un provider LLM diverso, segui i consigli di quel fornitore.

In generale, inserisci sia il testo medico originale che i risultati di Amazon Comprehend Medical nel prompt. Di seguito è riportata una struttura di prompt comune:

```
<medical_text>
medical text
</medical_text>

<comprehend_medical_text_results>
comprehend medical text results
</comprehend_medical_text_results>

<prompt_instructions>
prompt instructions
</prompt_instructions>
```

Questa sezione fornisce strategie per includere i risultati di Amazon Comprehend Medical come contesto immediato per le seguenti attività mediche comuni di PNL:

- Filtra i risultati di Amazon Comprehend Medical
- Estendi le attività di PNL in ambito medico con Amazon Comprehend Medical
- Applica i guardrail con Amazon Comprehend Medical

Filtra i risultati di Amazon Comprehend Medical

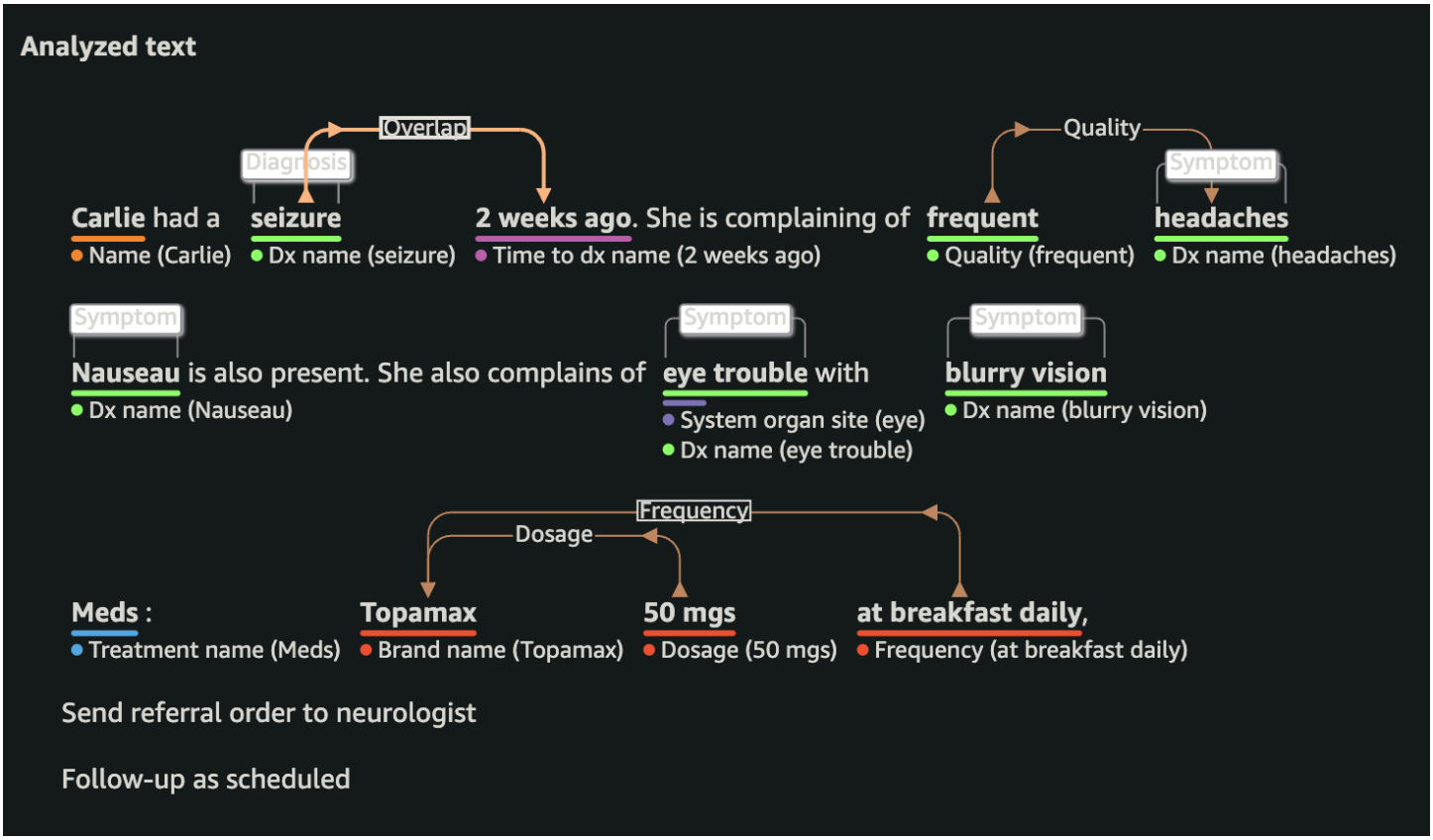
Amazon Comprehend Medical fornisce in genere una grande quantità di informazioni. Potresti voler ridurre il numero di risultati che il medico deve esaminare. In questo caso, puoi utilizzare un LLM per filtrare questi risultati. Le entità Amazon Comprehend Medical includono un punteggio di confidenza che puoi utilizzare come meccanismo di filtro durante la progettazione del prompt.

Di seguito è riportato un esempio di nota per un paziente:

```
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
```

```
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
```

In questa nota per il paziente, Amazon Comprehend Medical rileva le seguenti entità.



Le entità si collegano ai seguenti codici ICD-10-CM per convulsioni e mal di testa.

Categoria	Codice ICD-10-CM	Descrizione ICD-10-CM	Punteggio di attendibilità
Convulsioni	R56.9	Convulsioni non specificate	0,8348
Sequestro	G40.909	Epilessia, non specificata, non intrattabile, senza stato epilettico	0,5424

Convulsioni	R56,00	Semplici convulsioni febbrili	0,4937
Convulsioni	G40.09	Altre crisi epilettiche	0,4397
Convulsioni	G40.409	Altre epilessia e sindromi epilettiche generalizzate, non intrattabili, senza stato epilettico	0,4138
Mal di testa	R51	mal di testa	0,4067
Mal di testa	R51,9	Cefalea, non specifica ta	0,3844
Mal di testa	G44,52	Nuova cefalea persistente quotidiana (NDPH)	0,3005
Mal di testa	G 44	Altra sindrome di cefalea	0,2670
Mal di testa	4.8	Altre sindromi cefalee specificate	0,2542

È possibile inserire i codici ICD-10-CM nel prompt per aumentare la precisione LLM. Per ridurre il rumore, puoi filtrare i codici ICD-10-CM utilizzando il punteggio di confidenza incluso nei risultati di Amazon Comprehend Medical. Di seguito è riportato un esempio di prompt che include solo i codici ICD-10-CM con un punteggio di affidabilità superiore a 0,4:

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of frequent headaches
Nausea is also present. She also complains of eye trouble with blurry vision
Meds : Topamax 50 mgs at breakfast daily,
Send referral order to neurologist
Follow-up as scheduled
</patient_note>
```

```
<comprehend_medical_results>
<icd-10>
  <entity>
    <text>seizure</text>
    <code>
      <description>Unspecified convulsions</description>
      <code_value>R56.9</code_value>
      <score>0.8347607851028442</score>
    </code>
    <code>
      <description>Epilepsy, unspecified, not intractable, without status epilepticus</
description>
      <code_value>G40.909</code_value>
      <score>0.542376697063446</score>
    </code>
    <code>
      <description>Other seizures</description>
      <code_value>G40.89</code_value>
      <score>0.43966275453567505</score>
    </code>
    <code>
      <description>Other generalized epilepsy and epileptic syndromes, not intractable,
 without status epilepticus</description>
      <code_value>G40.409</code_value>
      <score>0.41382506489753723</score>
    </code>
  </entity>
  <entity>
    <text>headaches</text>
    <code>
      <description>Headache</description>
      <code_value>R51</code_value>
      <score>0.4066613018512726</score>
    </code>
  </entity>
  <entity>
    <text>Nausea</text>
    <code>
      <description>Nausea</description>
      <code_value>R11.0</code_value>
      <score>0.6460834741592407</score>
    </code>
  </entity>
```

```
  <entity>
    <text>eye trouble</text>
    <code>
      <description>Unspecified disorder of eye and adnexa</description>
      <code_value>H57.9</code_value>
      <score>0.6780954599380493</score>
    </code>
    <code>
      <description>Unspecified visual disturbance</description>
      <code_value>H53.9</code_value>
      <score>0.5871203541755676</score>
    </code>
    <code>
      <description>Unspecified disorder of binocular vision</description>
      <code_value>H53.30</code_value>
      <score>0.5539672374725342</score>
    </code>
  </entity>
  <entity>
    <text>blurry vision</text>
    <code>
      <description>Other visual disturbances</description>
      <code_value>H53.8</code_value>
      <score>0.9001834392547607</score>
    </code>
  </entity>
</icd-10>
</comprehend_medical_results>

<prompt>
Given the patient note and Amazon Comprehend Medical ICD-10-CM code results above,
 please select the most relevant ICD-10-CM diagnosis codes for the patient.
For each selected code, provide a brief explanation of why it is relevant based on the
 information in the patient note.
</prompt>
```

Estendi le attività di PNL in ambito medico con Amazon Comprehend Medical

Durante l'elaborazione di testi medici, il contesto di Amazon Comprehend Medical può aiutare l'LLM a selezionare token migliori. In questo esempio, vuoi abbinare i sintomi della diagnosi ai farmaci. È inoltre necessario trovare del testo che si riferisca agli esami medici, ad esempio termini che si riferiscono a un esame del sangue. Puoi usare Amazon Comprehend Medical per rilevare le entità

e i nomi dei farmaci. In questo caso, utilizzerai la versione DetectEntitiesV2 e InferRxNorm APIs per Amazon Comprehend Medical.

Di seguito è riportato un esempio di nota per un paziente:

```
Carlie had a seizure 2 weeks ago. She is complaining of increased frequent headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
Place MRI radiology order at RadNet
```

Per concentrarsi sul codice di diagnosi, nel prompt `DX_NAME` vengono utilizzate solo `MEDICAL_CONDITION` le entità correlate al tipo. Gli altri metadati sono esclusi per irrilevanza. Per le entità farmaceutiche, è incluso il nome del farmaco insieme agli attributi estratti. Altri metadati relativi alle entità farmaceutiche di Amazon Comprehend Medical sono esclusi per irrilevanza. Di seguito è riportato un prompt di esempio che utilizza risultati filtrati di Amazon Comprehend Medical. Il prompt si concentra sulle `MEDICAL_CONDITION` entità che hanno il tipo. `DX_NAME` Questo prompt è progettato per collegare in modo più preciso i codici di diagnosi con i farmaci ed estrarre con maggiore precisione i test medici:

```
<patient_note>
Carlie had a seizure 2 weeks ago. She is complaining of increased freqeunt headaches
Given lyme disease symptoms such as muscle ache and stiff neck will order
 prescription.
Meds : Topamax 50 mgs at breakfast daily. Amoxicillan 25 mg by mouth twice a day
Place MRI radiology order at RadNet
</patient_note>

<detect_entity_results>
<entity>
    <text>seizure</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>headaches</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>lyme disease</text>
    <category>MEDICAL_CONDITION</category>
```

```
    <type>DX_NAME</type>
</entity>
<entity>
    <text>muscle ache</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
<entity>
    <text>stiff neck</text>
    <category>MEDICAL_CONDITION</category>
    <type>DX_NAME</type>
</entity>
</detect_entity_results>

<rx_results>
<entity>
    <text>Topamax</text>
    <category>MEDICATION</category>
    <type>BRAND_NAME</type>
    <attributes>
        <attribute>
            <type>FREQUENCY</type>
            <text>at breakfast daily</text>
        </attribute>
        <attribute>
            <type>DOSAGE</type>
            <text>50 mgs</text>
        </attribute>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
    </attributes>
</entity>
<entity>
    <text>Amoxicillan</text>
    <category>MEDICATION</category>
    <type>GENERIC_NAME</type>
    <attributes>
        <attribute>
            <type>ROUTE_OR_MODE</type>
            <text>by mouth</text>
        </attribute>
```

```
        <attribute>
            <type>DOSAGE</type>
            <text>25 mg</text>
        </attribute>
        <attribute>
            <type>FREQUENCY</type>
            <text>twice a day</text>
        </attribute>
    </attributes>
</entity>
</rx_results>

<prompt>
Based on the patient note and the detected entities, can you please:
1. Link the diagnosis symptoms with the medications prescribed.
Provide your reasoning for the linkages.
2. Extract any entities related to medical order tests mentioned in the note.
</prompt>
```

Applica i guardrail con Amazon Comprehend Medical

Puoi utilizzare un LLM e Amazon Comprehend Medical per creare guardrail prima che venga utilizzata la risposta generata. Puoi eseguire questo flusso di lavoro su testo medico non modificato o post-elaborato. I casi d'uso includono la gestione di informazioni sanitarie protette (PHI), il rilevamento di allucinazioni o l'implementazione di politiche personalizzate per la pubblicazione dei risultati. Ad esempio, puoi utilizzare il contesto di Amazon Comprehend Medical per identificare i dati PHI e quindi utilizzare l'LLM per rimuovere tali dati PHI.

Di seguito è riportato un esempio di informazioni tratte dalla cartella clinica di un paziente che include PHI:

```
Patient name: John Doe
Patient SSN: 123-34-5678
Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
```

Di seguito è riportato un esempio di prompt che include i risultati di Amazon Comprehend Medical come contesto:

```
<original_text>
```

```
Patient name: John Doe
Patient SSN: 123-34-5678 Patient DOB: 01/01/2024
Patient address: 123 Main St, Anytown USA
Exam details: good health. Pulse is 60 bpm. needs to work on diet with BMI of 190
</original_text>

<comprehend_medical_phi_entities>
<entity>
  <text>John Doe</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9967944025993347</score>
  <type>NAME</type>
</entity>
<entity>
  <text>123-34-5678</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9998034834861755</score>
  <type>ID</type>
</entity>
<entity>
  <text>01/01/2000</text>
  <category>PROTECTED_HEALTH_INFORMATION</category>
  <score>0.9964448809623718</score>
  <type>DATE</type>
</entity>
</comprehend_medical_phi_entities>

<instructions>
Using the provided original text and the Amazon Comprehend Medical PHI entities
 detected, please analyze the text to determine if it contains any additional protected
 health information (PHI) beyond the entities already identified. If additional PHI is
 found, please list and categorize it. If no additional PHI is found, please state that
 explicitly.
In addition if PHI is found, generate updated text with the PHI removed.
</instructions>
```

Utilizzo di modelli linguistici di grandi dimensioni per casi d'uso nel settore sanitario e delle scienze della vita

Questo descrive come utilizzare modelli linguistici di grandi dimensioni (LLMs) per applicazioni nel settore sanitario e delle scienze biologiche. Alcuni casi d'uso richiedono l'uso di un modello linguistico

di grandi dimensioni per le funzionalità di intelligenza artificiale generativa. Esistono vantaggi e limiti anche per la maggior parte delle persone state-of-the-art LLMs, e i consigli contenuti in questa sezione sono progettati per aiutarti a raggiungere i risultati prefissati.

È possibile utilizzare il percorso decisionale per determinare la soluzione LLM appropriata per il proprio caso d'uso, considerando fattori quali la conoscenza del dominio e i dati di formazione disponibili. Inoltre, questa sezione illustra le più diffuse pratiche mediche preaddestrate LLMs e le migliori pratiche per la loro selezione e utilizzo. Descrive inoltre i compromessi tra soluzioni complesse e ad alte prestazioni e approcci più semplici e a basso costo.

Casi d'uso per un LLM

Amazon Comprehend Medical può eseguire attività di PNL specifiche. Per ulteriori informazioni, consulta Casi d'uso per Amazon Comprehend Medical.

Le funzionalità di intelligenza artificiale logiche e generative di un LLM potrebbero essere necessarie per i casi d'uso avanzati nel settore sanitario e delle scienze della vita, come i seguenti:

- Classificazione di entità mediche personalizzate o categorie di testo
- Rispondere a domande cliniche
- Riepilogo dei referti medici
- Generazione e rilevamento di informazioni dettagliate a partire da informazioni mediche

Approcci di personalizzazione

È fondamentale capire come LLMs vengono implementati. LLMs vengono comunemente addestrati con miliardi di parametri, inclusi i dati di addestramento provenienti da molti domini. Questa formazione consente all'LLM di affrontare le attività più generalizzate. Tuttavia, spesso sorgono sfide quando sono richieste conoscenze specifiche del dominio. Esempi di conoscenze settoriali nel settore sanitario e delle scienze della vita sono i codici clinici, la terminologia medica e le informazioni sanitarie necessarie per generare risposte accurate. Pertanto, l'utilizzo dell'LLM così com'è (senza richiedere informazioni aggiuntive senza ulteriori conoscenze di dominio) per questi casi d'uso probabilmente produce risultati imprecisi. Esistono diversi approcci popolari che è possibile utilizzare per superare questa sfida: progettazione tempestiva, Retrieval Augmented Generation (RAG) e messa a punto.

Progettazione di prompt

La progettazione tempestiva è il processo in cui si guidano le soluzioni di intelligenza artificiale generativa per creare gli output desiderati adattando gli input al LLM. Elaborando istruzioni precise con un contesto pertinente, è possibile guidare il modello verso il completamento di attività sanitarie specialistiche che richiedono un ragionamento. Un'efficace progettazione tempestiva può migliorare in modo significativo le prestazioni del modello per i casi d'uso sanitario senza richiedere modifiche al modello. Per ulteriori informazioni sulla progettazione dei prompt, consulta Implementazione della progettazione avanzata dei prompt con Amazon Bedrock (AWS post di blog). Il prompt e il prompt di Few-shot sono tecniche che chain-of-thought puoi utilizzare nella progettazione dei prompt.

Prompt few-shot

Il Few-shot prompting è una tecnica in cui si fornisce al LLM alcuni esempi dell'input-output desiderato prima di chiedergli di eseguire un'attività simile. Nei contesti sanitari, questo approccio è particolarmente utile per attività specialistiche, come il riconoscimento di entità mediche o la sintesi di note cliniche. Includendo da 3 a 5 esempi di alta qualità nel prompt, è possibile migliorare in modo significativo la comprensione da parte del modello della terminologia medica e dei modelli specifici del dominio. Per un esempio di few-shot prompt, consulta Few-shot prompt engineering and fine-tuning for in Amazon Bedrock (post del blog). LLMs AWS

Ad esempio, quando estrai i dosaggi dei farmaci dalle note cliniche, puoi fornire esempi di diversi stili di notazione che aiutano il modello a riconoscere le variazioni nel modo in cui gli operatori sanitari documentano le prescrizioni. Questo approccio è particolarmente efficace quando si lavora con formati di documentazione standardizzati o quando esistono modelli coerenti nei dati.

Chain-of-thought suggerimento

Chain-of-thought (CoT) prompting guida l'LLM attraverso un processo di ragionamento. step-by-step Ciò lo rende utile per complesse attività di supporto decisionale medico e di ragionamento diagnostico. Insegnando esplicitamente al modello a «pensare passo dopo passo» durante l'analisi degli scenari clinici, è possibile migliorarne la capacità di seguire i protocolli di ragionamento medico e ridurre gli errori diagnostici.

Questa tecnica eccelle quando il ragionamento clinico richiede più passaggi logici, come la diagnosi differenziale o la pianificazione del trattamento. Tuttavia, questo approccio presenta dei limiti quando si tratta di conoscenze mediche altamente specializzate al di fuori dei dati di formazione del modello o quando è richiesta una precisione assoluta per le decisioni di terapia intensiva.

In questi casi, la combinazione di CoT con un altro approccio può produrre risultati migliori. Un'opzione è combinare CoT con la richiesta di autocoerenza. Per ulteriori informazioni, consulta Migliorare le prestazioni dei modelli linguistici generativi con richieste di autocoerenza su Amazon Bedrock (AWS post del blog). Un'altra opzione è combinare framework di ragionamento, come il prompting, con RAG. ReAct Per ulteriori informazioni, consulta Sviluppare assistenti avanzati basati sull'intelligenza artificiale generativa basati su chat utilizzando RAG e suggerimenti (Prescriptive Guidance). ReAct AWS

Generazione potenziata da recupero dati

Retrieval Augmented Generation (RAG) è 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 addestramento prima di generare una risposta. Un sistema RAG può recuperare informazioni sull'ontologia medica (come le classificazioni internazionali delle malattie, i fascicoli nazionali sui farmaci e i titoli delle materie mediche) da una fonte di conoscenza. Ciò fornisce un contesto aggiuntivo all'LLM a supporto dell'attività di PNL medica.

Come discusso nella Combinazione di Amazon Comprehend Medical con modelli linguistici di grandi dimensioni sezione, puoi utilizzare un approccio RAG per recuperare il contesto da Amazon Comprehend Medical. Altre fonti di conoscenza comuni includono i dati del dominio medico archiviati in un servizio di database, come Amazon OpenSearch Service, Amazon Kendra o Amazon Aurora. L'estrazione di informazioni da queste fonti di conoscenza può influire sulle prestazioni di recupero, in particolare con le query semantiche che utilizzano un database vettoriale.

Un'altra opzione per archiviare e recuperare conoscenze specifiche del dominio consiste nell'utilizzare Amazon Q Business nel flusso di lavoro RAG. Amazon Q Business può indicizzare gli archivi di documenti interni o i siti Web rivolti al pubblico (come CMS.gov per i dati ICD-10). Amazon Q Business può quindi estrarre informazioni pertinenti da queste fonti prima di passare la richiesta al LLM.

Esistono diversi modi per creare un flusso di lavoro RAG personalizzato. Ad esempio, esistono molti modi per recuperare dati da una fonte di conoscenza. Per semplicità, consigliamo l'approccio di recupero comune che prevede l'utilizzo di un database vettoriale, come Amazon OpenSearch Service, per archiviare le conoscenze sotto forma di incorporamenti. Ciò richiede l'utilizzo di un modello di incorporamento, ad esempio un trasformatore di frasi, per generare incorporamenti per la query e per la conoscenza archiviata nel database vettoriale.

Per ulteriori informazioni sugli approcci RAG completamente gestiti e personalizzati, vedete Opzioni e architetture di Retrieval Augmented Generation on. AWS

Fine-tuning

La messa a punto di un modello esistente implica l'adozione di un LLM, 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. Questa operazione è denominata parameter-efficient fine-tuning (PEFT). Per ulteriori informazioni, vedi Hugging Face PEFT on. GitHub

Di seguito sono riportati due casi d'uso comuni in cui è possibile scegliere di perfezionare un LLM per un'attività di PNL medica:

- Attività generativa: i modelli basati su decoder eseguono attività di intelligenza artificiale generativa. AI/ML i professionisti utilizzano dati di base per mettere a punto un LLM esistente. Ad esempio, potresti addestrare il LLM utilizzando MedQuAD, un set di dati medici pubblico per la risposta a domande. Quando si richiama una query al LLM ottimizzato, non è necessario un approccio RAG per fornire un contesto aggiuntivo al LLM.
- Incorporamenti: i modelli basati su codificatori generano incorporamenti trasformando il testo in vettori numerici. Questi modelli basati su codificatori sono in genere chiamati modelli di incorporamento. Un modello di trasformazione delle frasi è un tipo specifico di modello di incorporamento ottimizzato per le frasi. L'obiettivo è generare incorporamenti dal testo di input. Gli incorporamenti vengono quindi utilizzati per l'analisi semantica o per attività di recupero. Per ottimizzare il modello di incorporamento, è necessario disporre di un corpus di conoscenze mediche, ad esempio documenti, da utilizzare come dati di formazione. Ciò si ottiene con coppie di testo basate sulla somiglianza o sul sentimento per mettere a punto un modello di trasformazione delle frasi. Per ulteriori informazioni, consulta Training and Finetuning Embedding Models with Sentence Transformers v3 su Hugging Face.

Puoi usare Amazon SageMaker Ground Truth per creare un set di dati di addestramento etichettato e di alta qualità. Puoi utilizzare l'output del set di dati etichettato di Ground Truth per eseguire l'addestramento dei tuoi modelli. Puoi anche utilizzare l'output come set di dati di addestramento per un modello Amazon SageMaker AI. Per ulteriori informazioni sul riconoscimento di entità denominate, sulla classificazione del testo con etichetta singola e sulla classificazione del testo multietichetta, consulta Text labeling with Ground Truth nella documentazione di Amazon SageMaker AI.

Per ulteriori informazioni sulla messa a punto, consulta questa guida. Ottimizzazione di modelli linguistici di grandi dimensioni nel settore sanitario

Scegliere un LLM

Amazon Bedrock è il punto di partenza consigliato per valutare le alte prestazioni LLMs. Per ulteriori informazioni, consulta Modelli di base supportati in Amazon Bedrock. Puoi utilizzare i processi di valutazione dei modelli in Amazon Bedrock per confrontare gli output di più output e quindi scegliere il modello più adatto al tuo caso d'uso. Per ulteriori informazioni, consulta Scegli il modello con le migliori prestazioni utilizzando le valutazioni di Amazon Bedrock nella documentazione di Amazon Bedrock.

Alcuni LLMs hanno una formazione limitata sui dati del dominio medico. Se il tuo caso d'uso richiede la messa a punto di un LLM o un LLM non supportato da Amazon Bedrock, prendi in considerazione l'utilizzo di Amazon AI. SageMaker Nell' SageMaker intelligenza artificiale, puoi utilizzare un LLM ottimizzato o scegliere un LLM personalizzato che è stato addestrato sui dati del dominio medico.

La tabella seguente elenca i più diffusi LLMs che sono stati formati sui dati del dominio medico.

LLM	Processi	Competenze	Architecture
BioBert	Recupero delle informazioni, classificazione del testo e riconoscimento di entità denominate	Riassunti PubMed, articoli a testo completo e conoscenze generali del PubMedCentral dominio	Codificatore
Clinica Albert	Recupero delle informazioni, classificazione del testo e riconoscimento delle entità denominate	Ampio set di dati multicentrico insieme a oltre 3.000.000 di cartelle cliniche elettroniche (EHR) di pazienti	Codificatore
GPT clinico	Riepilogo, risposta a domande e generazione di testo	Set di dati medici estesi e diversificati, tra cui cartelle cliniche, conoscenze specifiche del dominio	Decodificatore

		e consultazioni di dialogo a più round	
GatorTron-VAI	Riepilogo, risposta a domande, generazione di testo e recupero di informazioni	Note cliniche e letteratura biomedica	Encoder
MedBert	Recupero delle informazioni, classificazione del testo e riconoscimento di entità denominate	Ampio set di dati di testi medici, note cliniche, documenti di ricerca e documenti relativi all'assistenza sanitaria	Codificatore
Med-Palm	Risposte a domande per scopi medici	Set di dati di testo medico e biomedico	Decodificatore
MedAlpaca	Attività di risposta a domande e dialogo medico	Una varietà di testi medici, che comprendono risorse come flashcard mediche, wiki e set di dati di dialogo	Decodificatore
BioMedbert	Recupero delle informazioni, classificazione del testo e riconoscimento di entità denominate	Esclusivamente riassunti PubMed e articoli a testo completo di PubMedCentral	Codificatore
BioMedLM	Riepilogo, risposta a domande e generazione di testo	Letteratura biomedica da fonti di conoscenza PubMed	Decodificatore

Di seguito sono riportate le migliori pratiche per l'utilizzo di medici preformati: LLMs

- Comprendi i dati di formazione e la loro rilevanza per il tuo compito di PNL in ambito medico.

- Identifica l'architettura LLM e il suo scopo. Gli encoder sono appropriati per gli incorporamenti e le attività NLP. I decoder servono per attività di generazione.
- Valuta i requisiti di infrastruttura, prestazioni e costi per ospitare il LLM medico preformato.
- Se è necessaria una messa a punto precisa, assicuratevi che i dati di addestramento siano accurati e veritieri. Assicurati di mascherare o oscurare qualsiasi informazione di identificazione personale (PII) o informazione sanitaria protetta (PHI).

Le attività mediche di PNL nel mondo reale potrebbero differire da quelle già LLMs addestrate in termini di conoscenze o casi d'uso previsti. Se un LLM specifico del dominio non soddisfa i benchmark di valutazione, puoi perfezionarlo con il tuo set di dati oppure puoi addestrare un nuovo modello di base. La formazione di un nuovo modello di base è un'impresa ambiziosa e spesso costosa. Per la maggior parte dei casi d'uso, consigliamo di perfezionare un modello esistente.

Quando si utilizza o si perfeziona un LLM medico preformato, è importante occuparsi dell'infrastruttura, della sicurezza e delle barriere.

Infrastruttura

Rispetto all'utilizzo di Amazon Bedrock per l'inferenza su richiesta o in batch, l'hosting di LLM medici preformati (in genere di Hugging Face) richiede risorse significative. Per ospitare LLM medici preaddestrati, è comune utilizzare un'immagine Amazon SageMaker AI eseguita su un'istanza Amazon Elastic Compute Cloud (Amazon EC2) con una o GPUs più istanze, come le istanze ml.g5 per l'elaborazione accelerata o le istanze ml.inf2 per. AWS Inferentia Questo perché consumano una grande quantità di memoria e spazio su disco. LLMs

Sicurezza e guardrail

A seconda dei requisiti di conformità aziendale, prendi in considerazione l'utilizzo di Amazon Comprehend e Amazon Comprehend Medical per mascherare o oscurare le informazioni di identificazione personale (PII) e le informazioni sanitarie protette (PHI) dai dati di formazione. Questo aiuta a impedire che l'LLM utilizzi dati riservati quando genera risposte.

Ti consigliamo di prendere in considerazione e valutare pregiudizi, equità e allucinazioni nelle tue applicazioni di intelligenza artificiale generativa. Che tu stia utilizzando un LLM preesistente o che ne stia ottimizzando uno, implementa dei guardrail per prevenire risposte dannose. I guardrail sono protezioni personalizzabili in base ai requisiti delle applicazioni di intelligenza artificiale generativa e alle politiche di intelligenza artificiale responsabili. Ad esempio, puoi utilizzare Amazon Bedrock Guardrails.

Ottimizzazione di modelli linguistici di grandi dimensioni nel settore sanitario

L'approccio di messa a punto descritto in questa sezione supporta la conformità alle linee guida etiche e normative e promuove l'uso responsabile dei sistemi di intelligenza artificiale nel settore sanitario. È progettato per generare informazioni accurate e private. L'intelligenza artificiale generativa sta rivoluzionando l'assistenza sanitaria, ma off-the-shelf i modelli spesso non sono all'altezza negli ambienti clinici in cui la precisione è fondamentale e la conformità non è negoziabile. L'ottimizzazione dei modelli di base con dati specifici del dominio colma questa lacuna. Ti aiuta a creare sistemi di intelligenza artificiale che parlano il linguaggio della medicina rispettando al contempo rigorosi standard normativi. Tuttavia, il percorso verso una messa a punto di successo richiede un'attenta analisi delle sfide uniche dell'assistenza sanitaria: proteggere i dati sensibili, giustificare gli investimenti nell'IA con risultati misurabili e mantenere la rilevanza clinica in un panorama medico in rapida evoluzione.

Quando gli approcci più leggeri raggiungono i loro limiti, la messa a punto diventa un investimento strategico. L'aspettativa è che i miglioramenti in termini di precisione, latenza o efficienza operativa compenseranno i significativi costi di calcolo e progettazione richiesti. È importante ricordare che il ritmo di avanzamento dei modelli di base è rapido, quindi il vantaggio di un modello ottimizzato potrebbe durare solo fino alla prossima release principale del modello.

Questa sezione analizza la discussione sui seguenti due casi d'uso ad alto impatto da parte di clienti del settore sanitario: AWS

- Sistemi di supporto alle decisioni cliniche: migliorano l'accuratezza diagnostica attraverso modelli che comprendono le storie complesse dei pazienti e le linee guida in evoluzione. La messa a punto può aiutare i modelli a comprendere a fondo le storie complesse dei pazienti e a integrare linee guida specializzate. Ciò può potenzialmente ridurre gli errori di previsione dei modelli. Tuttavia, è necessario soppesare questi vantaggi rispetto al costo della formazione su set di dati sensibili di grandi dimensioni e all'infrastruttura necessaria per applicazioni cliniche ad alto rischio. La maggiore precisione e consapevolezza del contesto giustificheranno l'investimento, soprattutto quando nuovi modelli vengono rilasciati frequentemente?
- Analisi dei documenti medici: automatizza l'elaborazione di note cliniche, report di imaging e documenti assicurativi mantenendo la conformità all'Health Insurance Portability and Accountability Act (HIPAA). In questo caso, la messa a punto può consentire al modello di gestire in modo più efficace formati unici, abbreviazioni specializzate e requisiti normativi. I vantaggi si ottengono spesso grazie alla riduzione dei tempi di revisione manuale e al miglioramento della conformità.

Tuttavia, è essenziale valutare se questi miglioramenti sono sufficientemente sostanziali da giustificare le risorse necessarie per la messa a punto. Determina se la progettazione tempestiva e l'orchestrazione del flusso di lavoro sono in grado di soddisfare le tue esigenze.

Questi scenari reali illustrano il percorso di perfezionamento, dalla sperimentazione iniziale all'implementazione del modello, rispondendo al contempo ai requisiti unici dell'assistenza sanitaria in ogni fase.

Stima dei costi e del ritorno sull'investimento

Di seguito sono riportati i fattori di costo da considerare quando si perfeziona un LLM:

- Dimensioni del modello: i modelli più grandi costano di più per la messa a punto
- Dimensioni del set di dati: i costi e i tempi di elaborazione aumentano con la dimensione del set di dati per la messa a punto
- Strategia di ottimizzazione: i metodi efficienti in termini di parametri possono ridurre i costi rispetto agli aggiornamenti completi dei parametri

Nel calcolare il ritorno sull'investimento (ROI), considerate il miglioramento delle metriche scelte (ad esempio la precisione) moltiplicato per il volume delle richieste (con quale frequenza verrà utilizzato il modello) e la durata prevista prima che il modello venga superato dalle versioni più recenti.

Inoltre, considera la durata del tuo LLM di base. Nuovi modelli base emergono ogni 6-12 mesi. Se il tuo rilevatore di malattie rare impiega 8 mesi per perfezionare e convalidare, potresti ottenere solo 4 mesi di prestazioni superiori prima che i modelli più recenti colmino il divario.

Calcolando i costi, il ROI e la potenziale durata di vita per il tuo caso d'uso, puoi prendere una decisione basata sui dati. Ad esempio, se l'ottimizzazione del modello di supporto alle decisioni cliniche porta a una riduzione misurabile degli errori diagnostici in migliaia di casi all'anno, l'investimento potrebbe ripagare rapidamente. Al contrario, se la sola progettazione tempestiva consente di avvicinare il flusso di lavoro per l'analisi dei documenti alla precisione prefissata, potrebbe essere saggio rimandare la messa a punto fino all'arrivo della prossima generazione di modelli.

one-size-fits-allLa messa a punto non lo è. Se decidi di perfezionare, l'approccio giusto dipende dal caso d'uso, dai dati e dalle risorse.

Scelta di una strategia di ottimizzazione

Dopo aver stabilito che la messa a punto è l'approccio giusto per il vostro caso d'uso nel settore sanitario, il passo successivo consiste nella selezione della strategia di messa a punto più appropriata. Sono disponibili diversi approcci. Ciascuno presenta vantaggi e compromessi distinti per le applicazioni sanitarie. La scelta tra questi metodi dipende dagli obiettivi specifici, dai dati disponibili e dai limiti delle risorse.

Obiettivi di formazione

Il pre-training adattivo al dominio (DAPT) è un metodo senza supervisione che prevede la formazione preliminare del modello su un ampio corpus di testo specifico del dominio e senza etichetta (come milioni di documenti medici). Questo approccio è ideale per migliorare la capacità dei modelli di comprendere le abbreviazioni delle specialità mediche e la terminologia utilizzata da radiologi, neurologi e altri fornitori specializzati. Tuttavia, DAPT richiede grandi quantità di dati e non affronta attività specifiche.

Il Supervised Fine-Tuning (SFT) insegna al modello a seguire istruzioni esplicite utilizzando esempi strutturati di input-output. Questo approccio eccelle per i flussi di lavoro di analisi dei documenti medici, come il riepilogo dei documenti o la codifica clinica. L'ottimizzazione delle istruzioni è una forma comune di SFT in cui il modello viene addestrato sulla base di esempi che includono istruzioni esplicite abbinate agli output desiderati. Ciò migliora la capacità del modello di comprendere e seguire le diverse istruzioni dell'utente. Questa tecnica è particolarmente utile in ambito sanitario perché addestra il modello con esempi clinici specifici. Lo svantaggio principale è che richiede esempi accuratamente etichettati. Inoltre, il modello perfezionato potrebbe avere problemi con casi limite in cui non ci sono esempi. Per istruzioni sulla messa a punto con Amazon SageMaker Jumpstart, consulta Istruzioni di ottimizzazione per FLAN T5 XL con Amazon Jumpstart (post di blog). SageMaker AWS

L'apprendimento per rinforzo dal feedback umano (RLHF) ottimizza il comportamento del modello in base al feedback e alle preferenze degli esperti. Utilizza un modello di ricompensa basato sulle preferenze e sui metodi umani, come l'ottimizzazione delle politiche prossimali (PPO) o l'ottimizzazione delle preferenze dirette (DPO), per ottimizzare il modello evitando aggiornamenti distruttivi. RLHF è ideale per allineare i risultati alle linee guida cliniche e assicurarsi che le raccomandazioni rientrino nei protocolli approvati. Questo approccio richiede molto tempo da parte del medico per il feedback e prevede una pipeline di formazione complessa. Tuttavia, RLHF è particolarmente utile nel settore sanitario perché aiuta gli esperti medici a modellare il modo in cui i sistemi di intelligenza artificiale comunicano e formulano raccomandazioni. Ad esempio, i medici

possono fornire feedback per assicurarsi che il modello mantenga un atteggiamento appropriato al paziente, sappia quando esprimere incertezze e rispetti le linee guida cliniche. Tecniche come il PPO ottimizzano iterativamente il comportamento del modello sulla base del feedback degli esperti, limitando al contempo gli aggiornamenti dei parametri per preservare le conoscenze mediche di base. Ciò consente ai modelli di formulare diagnosi complesse in un linguaggio adatto al paziente, pur segnalando condizioni gravi da sottoporre a cure mediche immediate. Questo è fondamentale per l'assistenza sanitaria, dove sia la precisione che lo stile di comunicazione sono importanti. Per ulteriori informazioni su RLHF, consulta Ottimizzazione di modelli linguistici di grandi dimensioni con l'apprendimento per rinforzo basato sul feedback umano o basato sull'intelligenza artificiale (post sul blog).AWS

Metodi di implementazione

Un aggiornamento completo dei parametri comporta l'aggiornamento di tutti i parametri del modello durante l'addestramento. Questo approccio funziona meglio per i sistemi di supporto alle decisioni cliniche che richiedono una profonda integrazione delle storie dei pazienti, dei risultati di laboratorio e delle linee guida in evoluzione. Gli svantaggi includono costi di elaborazione elevati e rischio di sovraadattamento se il set di dati non è ampio e diversificato.

I metodi PEFT (Parameter-Efficient Fine-Tuning) aggiornano solo un sottoinsieme di parametri per evitare un sovraadattamento o una perdita catastrofica delle funzionalità linguistiche. I tipi includono l'adattamento a basso rango (LoRa), gli adattatori e l'ottimizzazione dei prefissi. I metodi PEFT offrono costi computazionali inferiori, una formazione più rapida e sono ideali per esperimenti come l'adattamento di un modello di supporto decisionale clinico ai nuovi protocolli o alla terminologia di un nuovo ospedale. La limitazione principale è rappresentata dalla potenziale riduzione delle prestazioni rispetto agli aggiornamenti completi dei parametri.

Per ulteriori informazioni sui metodi di fine-tuning, consulta Advanced fine-tuning methods on SageMaker Amazon AI (post del blog).AWS

Creazione di un set di dati di ottimizzazione

La qualità e la diversità del set di dati di ottimizzazione sono fondamentali per le prestazioni del modello, la sicurezza e la prevenzione delle distorsioni. Di seguito sono riportate tre aree critiche da considerare durante la creazione di questo set di dati:

- Volume basato su un approccio di ottimizzazione
- Annotazione dei dati fornita da un esperto del settore

- Diversità del set di dati

Come illustrato nella tabella seguente, i requisiti relativi alle dimensioni del set di dati per la regolazione fine variano in base al tipo di ottimizzazione eseguita.

Strategia di messa a punto	Dimensioni del set di dati
Formazione preliminare adattata al dominio	Oltre 100.000 testi di dominio
Ottimizzazione supervisionata	Oltre 10.000 paia etichettate
Apprendimento per rinforzo basato sul feedback umano	Oltre 1.000 coppie di preferenze di esperti

Puoi utilizzare AWS GlueAmazon EMR e Amazon SageMaker Data Wrangler per automatizzare il processo di estrazione e trasformazione dei dati per curare un set di dati di tua proprietà. Se non sei in grado di curare un set di dati sufficientemente grande, puoi scoprire e scaricare i set di dati direttamente nel tuo sito. Account AWS AWS Data Exchange Consultate il vostro consulente legale prima di utilizzare set di dati di terze parti.

Annotatori esperti con conoscenze di settore, come medici, biologi e chimici, dovrebbero far parte del processo di cura dei dati per incorporare le sfumature dei dati medici e biologici nell'output del modello. Amazon SageMaker Ground Truth fornisce un'interfaccia utente a basso codice per consentire agli esperti di annotare il set di dati.

Un set di dati che rappresenti la popolazione umana è essenziale per ottimizzare i casi d'uso nel settore sanitario e delle scienze biologiche per evitare distorsioni e riflettere i risultati del mondo reale. AWS Glue le sessioni interattive o le istanze di SageMaker notebook Amazon offrono un modo efficace per esplorare in modo iterativo i set di dati e ottimizzare le trasformazioni utilizzando notebook compatibili con Jupyter. Le sessioni interattive ti consentono di lavorare con una scelta di ambienti di sviluppo integrati più diffusi () nel tuo ambiente locale. IDEs In alternativa, puoi lavorare con AWS Glue i nostri notebook Amazon SageMaker Studio tramite. Console di gestione AWS

Ottimizzazione del modello

AWS fornisce servizi come Amazon SageMaker AI e Amazon Bedrock che sono fondamentali per una messa a punto di successo.

SageMaker L'intelligenza artificiale è un servizio di machine learning completamente gestito che aiuta gli sviluppatori e i data scientist a creare, addestrare e implementare rapidamente modelli di machine learning. Tre funzioni utili dell' SageMaker intelligenza artificiale per la messa a punto includono:

- SageMakerFormazione: una funzionalità di machine learning completamente gestita che consente di addestrare in modo efficiente un'ampia gamma di modelli su larga scala
- SageMaker JumpStart— Una funzionalità che si basa sui lavori di SageMaker formazione per fornire modelli preaddestrati, algoritmi integrati e modelli di soluzioni per le attività di machine learning
- SageMaker HyperPod— Una soluzione di infrastruttura appositamente progettata per la formazione distribuita dei modelli di base e LLMs

Amazon Bedrock è un servizio completamente gestito che fornisce l'accesso a modelli di base ad alte prestazioni tramite un'API, con funzionalità integrate di sicurezza, privacy e scalabilità. Il servizio offre la possibilità di perfezionare diversi modelli di base disponibili. Per ulteriori informazioni, consulta Modelli e regioni supportati per la messa a punto e la formazione preliminare continua nella documentazione di Amazon Bedrock.

Quando affronti il processo di messa a punto con uno dei due servizi, prendi in considerazione il modello base, la strategia di messa a punto e l'infrastruttura.

Scelta del modello base

I modelli closed-source, come Anthropic Claude, Meta Llama e Amazon Nova, offrono out-of-the-box prestazioni elevate con conformità gestita, ma limitano la flessibilità di ottimizzazione alle opzioni supportate dal provider, ad esempio gestite come Amazon Bedrock. APIs Ciò limita la personalizzabilità, in particolare per i casi d'uso sanitari regolamentati. Al contrario, i modelli open source, come Meta Llama, offrono controllo e flessibilità completi su tutti i servizi di SageMaker intelligenza artificiale di Amazon, rendendoli ideali quando devi personalizzare, controllare o adattare profondamente un modello ai tuoi requisiti specifici di dati o flussi di lavoro.

Strategia di perfezionamento

La semplice regolazione delle istruzioni può essere gestita tramite la personalizzazione del modello Amazon Bedrock o Amazon. SageMaker JumpStart Approcci PEFT complessi, come LoRa o adattatori, richiedono lavori di SageMaker formazione o funzionalità di ottimizzazione personalizzate in Amazon Bedrock. La formazione distribuita per modelli molto grandi è supportata da. SageMaker HyperPod

Scalabilità e controllo dell'infrastruttura

I servizi completamente gestiti, come Amazon Bedrock, riducono al minimo la gestione dell'infrastruttura e sono ideali per le organizzazioni che danno priorità alla facilità d'uso e alla conformità. Le opzioni semi-gestite, ad esempio SageMaker JumpStart, offrono una certa flessibilità con una minore complessità. Queste opzioni sono adatte per la prototipazione rapida o per l'utilizzo di flussi di lavoro predefiniti. Il pieno controllo e la personalizzazione derivano dai lavori di SageMaker formazione, che HyperPod, sebbene richiedano maggiore esperienza, sono ideali quando è necessario scalare per set di dati di grandi dimensioni o richiedere pipeline personalizzate.

Monitoraggio di modelli ottimizzati

Nel settore sanitario e delle scienze della vita, il monitoraggio della messa a punto del LLM richiede il monitoraggio di più indicatori chiave di performance. L'accuratezza fornisce una misurazione di base, ma questa deve essere bilanciata con la precisione e il richiamo, in particolare nelle applicazioni in cui le classificazioni errate comportano conseguenze significative. Il punteggio F1 aiuta a risolvere i problemi di squilibrio di classe che possono essere comuni nei set di dati medici. Per ulteriori informazioni sul tagging, consulta Valutazione LLMs per applicazioni nel settore sanitario e delle scienze della vitain questa guida.

Le metriche di calibrazione aiutano a garantire che i livelli di confidenza del modello corrispondano alle probabilità del mondo reale. Le metriche di equità possono aiutarti a rilevare potenziali pregiudizi nei diversi dati demografici dei pazienti.

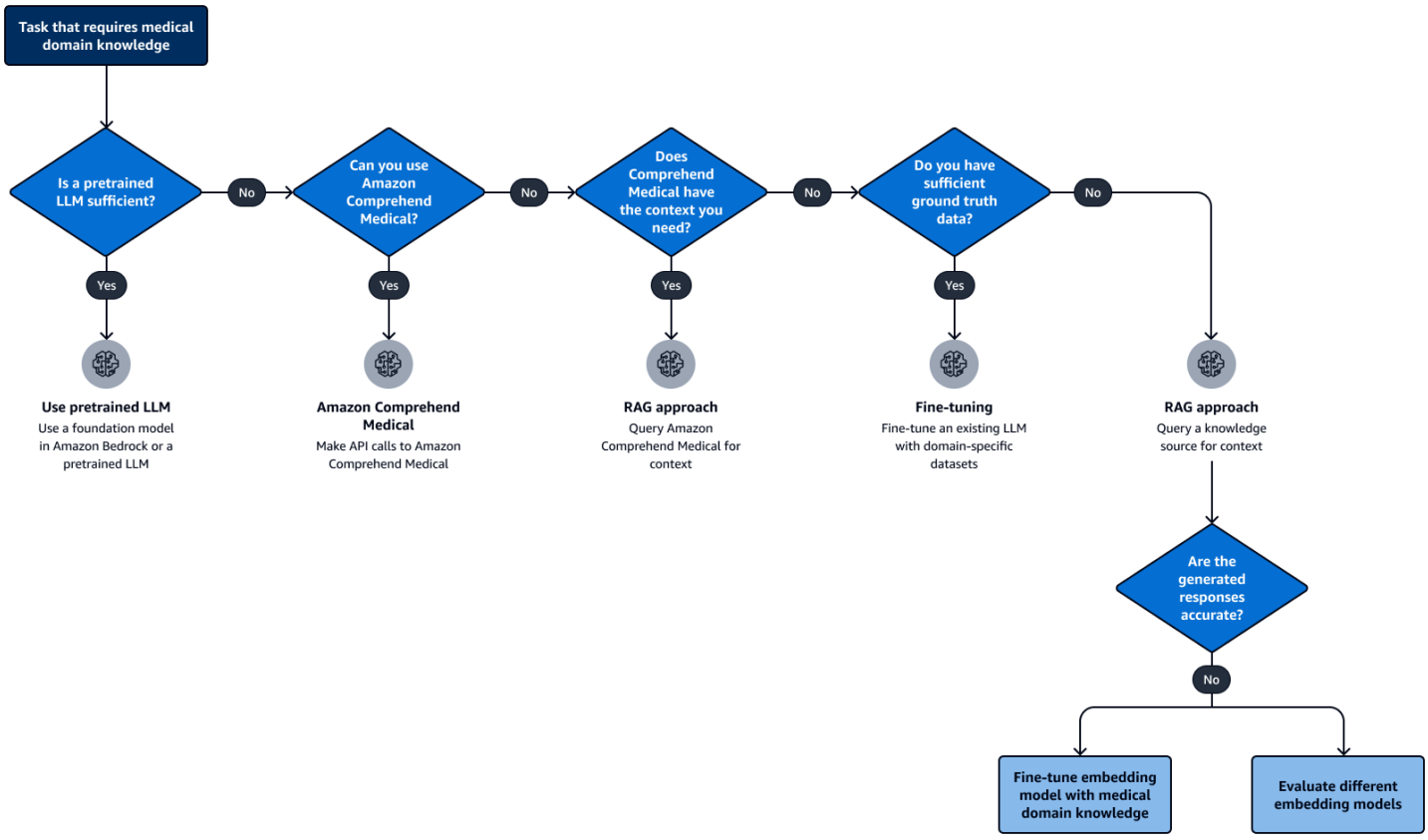
MLflowè una soluzione open source che può aiutarti a tenere traccia degli esperimenti di messa a punto. MLflow è supportato nativamente all'interno di Amazon SageMaker AI, il che ti aiuta a confrontare visivamente le metriche dei corsi di formazione. Per i lavori di ottimizzazione su Amazon Bedrock, le metriche vengono trasmesse in streaming ad Amazon CloudWatch in modo da poterle visualizzare nella console. CloudWatch

Scegliere un approccio PNL per l'assistenza sanitaria e le scienze della vita

La Approcci generativi di intelligenza artificiale e PNL per l'assistenza sanitaria e le scienze della vita sezione descrive i seguenti approcci per affrontare le attività di elaborazione del linguaggio naturale (NLP) per applicazioni nel settore sanitario e delle scienze della vita:

- Utilizzo di Amazon Comprehend Medical
- Combinazione di Amazon Comprehend Medical con un LLM in un flusso di lavoro di Retrieval Augment Generation (RAG)
- Utilizzo di un LLM ottimizzato
- Utilizzo di un flusso di lavoro RAG

Valutando i limiti noti delle LLMs attività del settore medico e il vostro caso d'uso, potete scegliere l'approccio più adatto alla vostra attività. Il seguente schema decisionale può aiutarti a scegliere un approccio LLM per il tuo compito di PNL in ambito medico:



Il diagramma mostra il flusso di lavoro seguente:

1. Per i casi d'uso nel settore sanitario e delle scienze della vita, identifica se l'attività di PNL richiede conoscenze di dominio specifiche. Se necessario, coordinatevi con esperti in materia ()SMEs.
2. Se puoi utilizzare un LLM generico o un modello che è stato addestrato su set di dati medici, utilizza un modello base disponibile in Amazon Bedrock o il LLM preformato. Per ulteriori informazioni sul tagging, consulta Scegliere un LLMin questa guida.
3. Se le funzionalità di rilevamento delle entità e di collegamento ontologico di Amazon Comprehend Medical soddisfano il tuo caso d'uso, utilizza Amazon Comprehend Medical. APIs Per ulteriori informazioni sul tagging, consulta Utilizzo di Amazon Comprehend Medicalin questa guida.
4. A volte, Amazon Comprehend Medical ha il contesto richiesto ma non supporta il tuo caso d'uso. Ad esempio, potresti aver bisogno di definizioni di entità diverse, ricevere un numero enorme di risultati, avere bisogno di entità personalizzate o avere bisogno di un'attività di PNL personalizzata. In tal caso, utilizza un approccio RAG per interrogare Amazon Comprehend Medical per conoscere il contesto. Per ulteriori informazioni sul tagging, consulta Combinazione di Amazon Comprehend Medical con modelli linguistici di grandi dimensioniin questa guida.
5. Se disponi di una quantità sufficiente di dati di base, perfeziona un LLM esistente. Per ulteriori informazioni sul tagging, consulta Approcci di personalizzazionein questa guida.
6. Se gli altri approcci non soddisfano gli obiettivi delle vostre attività di PNL dal punto di vista medico, implementate una soluzione RAG. Per ulteriori informazioni sul tagging, consulta Approcci di personalizzazionein questa guida.
7. Dopo aver implementato la soluzione RAG, valuta se le risposte generate sono accurate. Per ulteriori informazioni sul tagging, consulta Valutazione LLMs per applicazioni nel settore sanitario e delle scienze della vitain questa guida. È normale iniziare con un modello Amazon Titan Text Embeddings o un modello generico di trasformazione delle frasi, come ALL-MiniLM-L6-v2. Tuttavia, a causa della mancanza di un contesto di dominio, questi modelli potrebbero non rispecchiare la terminologia medica del testo. Se necessario, prendete in considerazione le seguenti modifiche:
   a. Valuta altri modelli di incorporamento
   b. Perfeziona il modello di incorporamento con set di dati specifici del dominio

Considerazioni sulla maturità aziendale

La maturità aziendale è fondamentale quando si adattano le soluzioni LLM per applicazioni sanitarie e scientifiche. Queste organizzazioni devono affrontare diversi livelli di complessità durante l'implementazione LLMs, a seconda dei criteri di accettazione. Spesso, le organizzazioni che non dispongono di AI/ML risorse investono nel supporto degli appaltatori per creare soluzioni LLM. In queste situazioni, è importante comprendere i seguenti compromessi:

- Prestazioni elevate per costi e manutenzione elevati: potrebbe essere necessaria una soluzione complessa che richieda una messa a punto o personalizzata LLMs per soddisfare rigorosi standard prestazionali. Tuttavia, ciò comporta costi e requisiti di manutenzione più elevati. Potrebbe essere necessario assumere risorse specializzate o collaborare con appaltatori per mantenere queste soluzioni sofisticate. Ciò può potenzialmente rallentare lo sviluppo.
- Buone prestazioni per costi e manutenzione ridotti: in alternativa, potresti scoprire che servizi come Amazon Bedrock o Amazon Comprehend Medical offrono prestazioni accettabili. Sebbene questi LLMs o questi approcci possano fornire risultati perfetti, queste soluzioni possono spesso fornire risultati coerenti e di alta qualità. Queste soluzioni hanno un costo inferiore e riducono gli oneri di manutenzione. Questo può accelerare lo sviluppo.

Se un approccio più semplice e a basso costo fornisce costantemente risultati di alta qualità che soddisfano i criteri di accettazione, valuta se l'aumento delle prestazioni valga i compromessi in termini di costi, manutenzione e tempo. Tuttavia, se la soluzione più semplice è notevolmente inferiore alle prestazioni previste e se l'organizzazione non ha la capacità di investimento per soluzioni complesse e i relativi requisiti di manutenzione, è consigliabile posticipare AI/ML lo sviluppo fino a quando non saranno disponibili più risorse o soluzioni alternative.

Inoltre, per qualsiasi soluzione di PNL medica che si basa su un LLM, si consiglia di eseguire un monitoraggio e una valutazione continui. Valuta il feedback degli utenti nel tempo e implementa valutazioni periodiche per assicurarti che la soluzione continui a soddisfare i tuoi obiettivi aziendali.

Valutazione LLMs per applicazioni nel settore sanitario e delle scienze della vita

Questa sezione fornisce una panoramica completa dei requisiti e delle considerazioni per la valutazione di modelli linguistici di grandi dimensioni (LLMs) nei casi d'uso nel settore sanitario e delle scienze della vita.

È importante utilizzare dati fondati attendibili e il feedback delle PMI per mitigare i pregiudizi e convalidare l'accuratezza della risposta generata dal LLM. Questa sezione descrive le migliori pratiche per la raccolta e la cura dei dati di formazione e test. Inoltre, consente di implementare barriere e misurare la distorsione e l'equità dei dati. Vengono inoltre illustrate le comuni attività mediche di elaborazione del linguaggio naturale (NLP), come la classificazione del testo, il riconoscimento di entità denominate e la generazione di testo, e le relative metriche di valutazione.

Presenta inoltre flussi di lavoro per eseguire la valutazione LLM durante la fase di sperimentazione della formazione e la fase di post-produzione. Il monitoraggio dei modelli e le operazioni LLM sono elementi importanti di questo processo di valutazione.

Dati di formazione e test per attività mediche di PNL

Le attività di PNL in ambito medico utilizzano in genere corpora medici (ad esempio PubMed) o informazioni sui pazienti (come gli appunti sulle visite dei pazienti in clinica) per classificare, riepilogare e generare approfondimenti. Il personale medico, ad esempio medici, amministratori sanitari o tecnici, varia in termini di competenze e punti di vista. A causa della soggettività tra questo personale medico, set di dati di formazione e test più piccoli rappresentano un rischio di parzialità. Per mitigare questo rischio, consigliamo le seguenti best practice:

- Quando utilizzi una soluzione LLM preaddestrata, assicurati di disporre di una quantità adeguata di dati di test. I dati del test dovrebbero assomigliare molto ai dati medici effettivi. A seconda dell'attività, questo può variare da 20 a più di 100 record.
- Quando perfezionate un LLM, raccogliete un numero sufficiente di record etichettati (di base) da una varietà SMEs di settori medici interessati. Un punto di partenza generale è costituito da almeno 100 documenti di alta qualità. Tuttavia, data la complessità dell'attività e i criteri di accettazione della precisione, potrebbero essere necessari più record.
- Se necessario per il tuo caso d'uso medico, implementa delle barriere e misura la distorsione e l'equità dei dati. Ad esempio, assicuratevi che l'LLM prevenga diagnosi errate dovute ai profili

razziali dei pazienti. Per ulteriori informazioni, consulta la Sicurezza e guardrail sezione di questa guida.

Molte società di ricerca e sviluppo di intelligenza artificiale, come Anthropic, hanno già implementato dei guardrail nei loro modelli di base per evitare la tossicità. È possibile utilizzare il rilevamento della tossicità per controllare i prompt di input e le risposte di output. LLMs Per ulteriori informazioni, consulta Rilevamento della tossicità nella documentazione di Amazon Comprehend e Guardrails nella documentazione di Amazon Bedrock.

In qualsiasi attività di intelligenza artificiale generativa, esiste il rischio di allucinazioni. È possibile mitigare questo rischio eseguendo attività di PNL, come la classificazione. Puoi anche utilizzare tecniche più avanzate, come le metriche di somiglianza del testo. BertScoreè una metrica di somiglianza del testo comunemente adottata. Per ulteriori informazioni sulle tecniche che è possibile utilizzare per mitigare le allucinazioni, vedere A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models.

Metriche per le attività mediche di PNL

È possibile creare metriche quantificabili dopo aver stabilito dati attendibili ed etichette fornite dalle PMI per la formazione e i test. Il controllo della qualità attraverso processi qualitativi, come lo stress test e la revisione dei risultati del LLM, è utile per uno sviluppo rapido. Tuttavia, le metriche fungono da benchmark quantitativi che supportano le future operazioni LLM e fungono da benchmark delle prestazioni per ogni versione di produzione.

Comprendere il compito medico è fondamentale. Le metriche in genere si riferiscono a una delle seguenti attività generali di PNL:

- Classificazione del testo: l'LLM classifica il testo in una o più categorie predefinite, in base alla richiesta di input e al contesto fornito. Un esempio è la classificazione di una categoria di dolore utilizzando una scala del dolore. Alcuni esempi di metriche di classificazione del testo includono:
  - Precisione
  - Precisione, nota anche come precisione macro
  - Richiamo, noto anche come richiamo di macro
  - Punteggio F1, noto anche come punteggio macro F1
  - Perdita di Hamming

- Riconoscimento di entità denominate (NER): noto anche come estrazione di testo, il riconoscimento delle entità denominate è il processo di localizzazione e classificazione delle entità denominate menzionate nel testo non strutturato in categorie predefinite. Un esempio è l'estrazione dei nomi dei farmaci dalle cartelle cliniche dei pazienti. Alcuni esempi di metriche NER includono:
  - Precisione
  - Precisione
  - Richiama
  - Punteggio F1
  - Perdita di Hamming
- Generazione: l'LLM genera nuovo testo elaborando il prompt e il contesto fornito. La generazione include attività di riepilogo o attività di risposta a domande. Alcuni esempi di metriche di generazione includono:
  - Sostituto orientato al richiamo per la valutazione del personale (ROUGE)
  - Metrica per la valutazione della traduzione con Explicit (METEOR) ORdering
  - Sostituto di valutazione bilingue (BLEU) (per le traduzioni)
  - Distanza tra le stringhe, nota anche come somiglianza del coseno

Domande frequenti sui casi d'uso nel settore sanitario e delle scienze della vita

Di seguito sono riportate le domande frequenti relative all'uso di Amazon Comprehend Medical LLMs o per attività di PNL in ambito medico.

Come faccio a scegliere tra Amazon Comprehend Medical e un LLM?

Se il tuo compito è individuare entità mediche all'interno del tuo testo medico, consulta la documentazione di Amazon Comprehend Medical per capire quali entità mediche possono essere estratte e se una qualsiasi delle ontologie si adatta al tuo caso d'uso. In caso contrario, valuta la possibilità di utilizzare un LLM. Per ulteriori informazioni, consulta Casi d'uso per Amazon Comprehend Medical e Casi d'uso per un LLM in questa guida.

Come posso fornire i risultati di Amazon Comprehend Medical a un LLM?

Puoi incorporare i risultati di Amazon Comprehend Medical come contesto nei tuoi prompt LLM. Ciò fornisce ulteriori conoscenze e terminologia mediche al LLM. Il contesto fornito può migliorare le prestazioni del LLM in attività quali il riconoscimento delle entità, la sintesi o la risposta a domande. La guida fornisce diversi esempi di come strutturare i prompt con i risultati di Amazon Comprehend Medical. Per ulteriori informazioni sul tagging, consulta Combinazione di Amazon Comprehend Medical con modelli linguistici di grandi dimensioniin questa guida.

Quali sono alcune best practice per l'utilizzo di Amazon Comprehend Medical LLMs con?

Ti consigliamo di utilizzare i punteggi di confidenza di Amazon Comprehend Medical per filtrare o assegnare priorità alle entità all'interno dei prompt. È anche importante valutarne le prestazioni sulla base di dati specifici e verificare che le definizioni delle entità siano in linea con i tuoi requisiti. La combinazione di Amazon Comprehend Medical con fonti di conoscenza specifiche del dominio può migliorare ulteriormente le prestazioni del LLM. Per ulteriori informazioni sul tagging, consulta Le

migliori pratiche per l'utilizzo di Amazon Comprehend Medical in un flusso di lavoro RAGin questa guida.

Devo utilizzare un LLM medico preformato o perfezionare un LLM generico per il mio caso d'uso sanitario?

La decisione dipende dalle vostre esigenze specifiche e dalla disponibilità di dati di formazione di alta qualità. Un medico preformato LLMs può fornire un buon punto di partenza. Tuttavia, potrebbe comunque essere necessario perfezionarli con i dati specifici del dominio. Se disponi di un numero sufficiente di dati etichettati, la messa a punto di un LLM generale può essere un'opzione valida. Per ulteriori informazioni, consulta e in questa guida. Scegliere un LLM Scegliere un approccio PNL per l'assistenza sanitaria e le scienze della vita

Come posso valutare le prestazioni delle attività LLMs di PNL in ambito medico?

Consigliamo di utilizzare metriche quantitative, come l'accuratezza, la precisione, il richiamo e il punteggio F1 per la classificazione del testo e le attività di riconoscimento delle entità nominate. Puoi usare ROUGE e METEOR per le attività di generazione di testo. È importante disporre di dati di base affidabili etichettati da esperti in materia e implementare processi per il monitoraggio delle prestazioni del modello nel tempo. Per ulteriori informazioni sul tagging, consulta Valutazione LLMs per applicazioni nel settore sanitario e delle scienze della vitain questa guida.

Quali sono i compromessi tra soluzioni LLM ad alta e bassa complessità?

La messa a punto di un LLM o la creazione di un LLM personalizzato sono soluzioni estremamente complesse. Questi approcci possono migliorare le prestazioni ma comportano costi e requisiti di manutenzione più elevati. Soluzioni più semplici, come l'utilizzo di soluzioni preaddestrate LLMs o Amazon Comprehend Medical, potrebbero fornire prestazioni accettabili con costi inferiori e cicli di sviluppo più rapidi. Tuttavia, questi approcci potrebbero non soddisfare requisiti di precisione rigorosi per alcuni casi d'uso. Per ulteriori informazioni sul tagging, consulta Considerazioni sulla maturità aziendalein questa guida.

Risorse e passaggi successivi

Questa guida ti aiuta ad Servizi AWS automatizzare la PNL medica e le attività di intelligenza artificiale generativa per applicazioni del mondo reale negli ambienti di produzione. Descrive come utilizzare Amazon Comprehend Medical, LLMs supportato in Amazon Bedrock, medico preformato o LLMs ottimizzato per raggiungere i tuoi obiettivi aziendali nel LLMs settore sanitario e delle scienze biologiche. Questa guida descrive i vantaggi e i limiti dei seguenti approcci:

- Utilizzo indipendente di Amazon Comprehend Medical
- Fornire i risultati di Amazon Comprehend Medical a un LLM
- Utilizzo di un LLM generale preformato o di un LLM medico secondo un approccio RAG (Retrieval Augmented Generation)
- Perfezionamento di un LLM generale o di un LLM medico

Utilizza l'albero decisionale e le considerazioni sulla maturità aziendale di questa guida per scegliere tra questi approcci in base al livello di maturità della tua organizzazione. AI/ML Sebbene Amazon Comprehend Medical e Amazon LLMs Bedrock offrano funzionalità potenti, hanno successo solo se vengono implementate e valutate correttamente. Utilizza le informazioni e le metriche di valutazione descritte in questa guida per convalidare le prestazioni della tua soluzione.

Per le fasi successive, consigliamo ai responsabili IT, agli architetti e ai responsabili tecnici del settore sanitario di collaborare con AI/ML i professionisti per identificare le loro attività mediche legate alla PNL. Utilizza questa guida per scegliere un percorso di sviluppo, quindi utilizza le funzionalità e le funzionalità appropriate Servizi AWS su cui implementare con successo una soluzione automatizzata. AWS

AWS risorse

- Documentazione di Amazon Comprehend Medical:
  - Guida per gli sviluppatori
  - Documentazione di riferimento delle API
- Documentazione Amazon Bedrock
  - Valutazione del modello Amazon Bedrock
  - Ottimizzazione in Amazon Bedrock

- Perfeziona un modello in Amazon AI SageMaker
- Amazon SageMaker Ground Truth
- Rilevamento della tossicità con Amazon Comprehend
- AWS Partner con competenze sanitarie

Altre risorse

- Apri la classifica Medical-LLM
- Un'indagine sui grandi modelli linguistici per l'assistenza sanitaria: dai dati, alla tecnologia e alle applicazioni alla responsabilità e all'etica
- I modelli linguistici di grandi dimensioni sono programmatori medici scadenti: analisi comparativa dell'interrogazione dei codici medici
- Dal principiante all'esperto: modellare le conoscenze mediche in conoscenze generali LLMs

Collaboratori

Creazione di testi

- Joe King, AWS esperto di dati
- Ankith Ede, architetto di AWS soluzioni
- Clement Perrot, stratega AWS senior dell'intelligenza artificiale generativa
- Jillian Forde, architetto senior delle soluzioni AWS
- Rajesh Sitaraman, consulente senior per le consegne AWS
- Ross Claytor, principale scienziato applicato AWS
- Shivesh Ummat, architetto di soluzioni AWS

Revisione

- Dilshad Raihan Akkam Veettil, Senior Data Scientist AWS
- Joseph Cottingham, AWS architetto del deep learning

Scrittura tecnica

- Lilly AbouHarb, AWS scrittrice tecnica senior

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
Nuove sezioni	Abbiamo aggiunto la messa a punto di modelli linguistici di grandi dimensioni nella sezione sanitaria e la sezione Prompt engineering.	5 dicembre 2025
Pubblicazione iniziale	—	16 dicembre 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 includono in genere operazioni, analisti e

proprietari aziendali, ingegneri addetti alla migrazione, sviluppatori e DevOps professionisti che lavorano nell'ambito degli sprint. Tra il 20% e il 50% di un portfolio di applicazioni aziendali è costituito da schemi ripetuti che possono essere ottimizzati con un approccio di fabbrica. Per ulteriori informazioni, consulta la discussione sulle fabbriche di migrazione e la Guida alla fabbrica di migrazione al cloud in questo set di contenuti.

metadati di migrazione

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

modello di migrazione

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

Valutazione del portfolio di migrazione (MPA)

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

valutazione della preparazione alla migrazione (MRA)

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

strategia di migrazione

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

ML

Vedi machine learning.

modernizzazione

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

valutazione della preparazione alla modernizzazione

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

applicazioni monolitiche (monoliti)

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

MAPPA

Vedi Migration Portfolio Assessment.

MQTT

Vedi Message Queuing Telemetry Transport.

classificazione multiclasse

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

infrastruttura mutabile

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

O

OAC

Vedi Origin Access Control.

QUERCIA

Vedi Origin Access Identity.

OCM

Vedi gestione delle modifiche organizzative.

migrazione offline

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

OI

Vedi l'integrazione delle operazioni.

OLA

Vedi accordo a livello operativo.

migrazione online

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

OPC-UA

Vedi Open Process Communications - Unified Architecture.

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

Un protocollo di comunicazione 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.

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.