



Valutazione del progetto di machine learning con la checklist MLOps

AWS Guida prescrittiva



AWS Guida prescrittiva: Valutazione del progetto di machine learning con la checklist MLOps

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

I marchi e l'immagine commerciale di Amazon non possono essere utilizzati in relazione a prodotti o servizi che non siano di Amazon, in una qualsiasi modalità che possa causare confusione tra i clienti o in una qualsiasi modalità che denigri o discrediti Amazon. Tutti gli altri marchi non di proprietà di Amazon sono di proprietà delle rispettive aziende, che possono o meno essere associate, collegate o sponsorizzate da Amazon.

Table of Contents

Introduzione	1
MLOps componenti della lista di controllo	2
1. Gestione incentrata sui dati	2
2. Sperimentazione	4
3. Osservabilità e gestione dei modelli	6
4. Promozione e pipeline solide	9
5. Integrazione continua	12
6. Monitoraggio continuo	15
7. Implementazione continua	19
8. Formazione continua	20
9. Governance	22
Esempio di utilizzo	24
Risorse	25
Cronologia dei documenti	26
Glossario	27
#	27
A	28
B	31
C	33
D	36
E	40
F	42
G	44
H	45
I	47
L	49
M	50
O	55
P	57
Q	60
R	61
S	64
T	68
U	69

V	70
W	70
Z	72
.....	lxxiii

Valutazione del tuo progetto di machine learning con la checklist MLOps

Charles Frenzel, Sharath Nagaraja e Spencer Romo Amazon Web Services (AWS)

Luglio [2023](#) (cronologia dei documenti)

La [MLOps checklist](#) è una checklist funzionale che puoi utilizzare in qualsiasi fase del tuo progetto di machine learning (ML). La checklist è uno strumento per valutare la preparazione generale, esaminare la copertura del sistema e identificare nuove aree di opportunità nei sistemi ML distribuiti. MLOps è la combinazione di persone, tecnologia e processi per la fornitura di soluzioni ML. Well-Architected MLOps aiuta le aziende a implementare i modelli di machine learning nella produzione in modo efficace e coerente e può offrire valore aziendale.

L'utilizzo della MLOps lista di controllo consente di eseguire le seguenti operazioni:

- Valuta il tuo MLOps sistema.
- Trova aree di opportunità.
- Individuate le aree di miglioramento.
- Valuta e aggiorna la tua roadmap strategica su AWS.
- Genera elementi arretrati.

Ti consigliamo di utilizzare la MLOps lista di controllo all'inizio del MLOps progetto, ma è possibile utilizzarne alcune parti durante qualsiasi fase.

MLOps componenti della lista di controllo

La [MLOps lista di controllo](#) è composta da nove diversi componenti. Ogni componente rappresenta un aspetto di un MLOps sistema maturo. Questa sezione descrive in dettaglio i componenti della lista di controllo. Nota che c'è qualche sovrapposizione in questa sezione perché alcuni componenti dipendono da altri componenti.

La lista di controllo copre le seguenti aree:

- [1. Gestione incentrata sui dati](#)
- [2. Sperimentazione](#)
- [3. Osservabilità e gestione dei modelli](#)
- [4. Promozione e pipeline solide](#)
- [5. Integrazione continua](#)
- [6. Monitoraggio continuo](#)
- [7. Implementazione continua](#)
- [8. Formazione continua](#)
- [9. Governance](#)

Nota: la sicurezza è già ben coperta da AWS. [Ti consigliamo di consultare le guide e i modelli di AWS Prescriptive Guidance relativi alla sicurezza e di esaminare le best practice di sicurezza di AWS.](#)

1. Gestione incentrata sui dati

La gestione dei dati è la pratica volta a garantire che i dati utilizzati nella formazione, nei test e nell'inferenza siano gestiti, protetti e convalidati correttamente. Quando si creano modelli su larga scala, i dati sono la risorsa principale che consente prestazioni elevate dei modelli.

1.1 Archivio di dati

Un archivio di dati richiede la capacità di tracciare i dati e vederne il punto di origine. Quando vengono aggiunti o rimossi nuovi dati, il data repository registra le modifiche in fase di ripristino. point-in-time L'archivio di dati deve tenere conto del modo in cui i dati

delle etichette vengono tracciati ed elaborati e di come vengono tracciati gli artefatti dei dati intermedi.

1.2 Integrazione di diverse fonti di dati

A seconda dell'applicazione, l'addestramento del modello potrebbe richiedere dati provenienti da diverse fonti. La progettazione e la manutenzione di un manifesto che informi i professionisti del machine learning sulle fonti di dati disponibili e sul modo in cui interagiscono è fondamentale per la creazione di modelli.

1.3 Convalida dello schema dei dati

Per alimentare i dati dei modelli, è importante che i dati di addestramento siano omogenei. Potrebbero essere necessarie trasformazioni o altre analisi esplorative per i dati archiviati in soluzioni data lake come Amazon Simple Storage Service (Amazon S3) o in archivi di dati documentali.

1.4 Versionamento e derivazione dei dati

Quando si addestrano modelli che potrebbero essere utilizzati in produzione, è necessario o essere in grado di riprodurre i risultati e disporre di un modo affidabile per eseguire [studi di ablazione](#) per comprendere meglio le prestazioni complessive del modello. Il monitoraggio dello stato dei dati di addestramento è fondamentale per questa riproducibilità. Strumenti come [Data Version Control \(DVC\)](#) possono aiutarvi in questo senso.

1.5 Flusso di lavoro di etichettatura

Nei casi in cui i dati etichettati non sono disponibili all'inizio del progetto, la creazione di dati etichettati è spesso un passaggio necessario. Strumenti come Amazon SageMaker Ground Truth richiedono che i dati di input siano strutturati in modo appropriato e richiedono un processo di etichettatura definito e testato. È necessario utilizzare una forza lavoro composta da etichettatori interni o esterni. I dati devono quindi essere convalidati, utilizzando approcci di etichettatura ridondanti o di apprendimento automatico per identificare valori anomali o errori nel set di dati di formazione.

1.6 Archiviazione delle funzionalità online e offline

Il sistema ML dispone di un [Feature Store](#) o di un archivio centralizzato per le funzionalità e i metadati associati in modo che sia possibile riutilizzare le funzionalità o gli input del modello. Puoi creare un archivio online o offline. Utilizza un negozio online per casi d'uso di inferenza in tempo reale e a bassa latenza. Utilizza un negozio offline per la formazione e l'inferenza in batch.

2. Sperimentazione

La sperimentazione include la registrazione, il tracciamento e le metriche degli esperimenti. Ciò si traduce nell'integrazione dei metadati sperimentali sulla piattaforma, nel controllo del codice sorgente e negli ambienti di sviluppo. La sperimentazione include anche la possibilità di ottimizzare le prestazioni e la precisione del modello tramite il debug.

2.1 Ambienti di sviluppo integrati

Un ambiente di sviluppo integrato (IDE) è integrato direttamente con il cloud. L'IDE può interagire e inviare comandi al sistema più grande. Idealmente, supporta quanto segue:

- Sviluppo locale
- Integrazione del controllo delle versioni
- Il debug è in atto, con tutti i log e gli artefatti generati che entrano nel controllo della versione

2.2 Controllo della versione del codice

Per garantire la riproducibilità e la riusabilità, tutto il codice viene salvato nell'archivio dei sorgenti con un adeguato controllo della versione. Ciò include il codice dell'infrastruttura, il codice dell'applicazione, il codice del modello e persino i notebook (se si sceglie di utilizzarli).

2.3 Tracciamento

Un progetto di machine learning richiede uno strumento in grado di tracciare e analizzare esperimenti di machine learning. Questo strumento dovrebbe registrare tutte le metriche, i parametri e gli artefatti durante l'esecuzione di un esperimento di machine learning, registrando tutti i metadati in una posizione centrale. La posizione centrale offrirà la possibilità di analizzare, visualizzare e controllare tutti gli esperimenti eseguiti.

2.4 Integrazione multiplatforma

I risultati storici degli esperimenti e tutti i relativi metadati sono accessibili in altre parti del sistema. Ad esempio, le pipeline di orchestrazione esistenti possono accedere a questi dati, così come gli strumenti di monitoraggio.

2.5 Debug: precisione e prestazioni del sistema

È disponibile un framework completo per il debug dei modelli che consente di esaminare le esecuzioni relative a quanto segue:

- Trova i colli di bottiglia
- Avviso sulle anomalie
- Massimizza l'utilizzo delle risorse
- Aiuto nell'analisi degli esperimenti

Quando la formazione è intensiva, la capacità di massimizzare la produttività è fondamentale e lo rende uno strumento necessario per l'ottimizzazione dei costi.

3. Osservabilità e gestione dei modelli

La sezione sull'osservabilità e la gestione dei modelli della checklist comprende il controllo della versione del modello e il tracciamento del lineage nell'intero sistema ML. Il controllo delle versioni del modello aiuta a tenere traccia e controllare tutte le modifiche applicate a un modello in modo da poter ripristinare una versione precedente quando necessario. Il tracciamento del lignaggio fornisce una panoramica dei flussi in entrata e in uscita dal modello. Un altro vantaggio chiave del tracciamento del lignaggio è il point-in-time ripristino (PITR), che automatizza l'implementazione e il ripristino del sistema.

3.1 Registro dei modelli con versioni

In generale, un registro dei modelli supporta il controllo delle versioni e il tracciamento della derivazione dei componenti del modello. Un buon registro può associare i metadati al modello con versione, tra cui:

- I dati utilizzati
- Informazioni sul modello
- Risultati delle metriche di valutazione
- Codice modello associato

3.2 Pregiudizi, equità e spiegabilità

Come minimo, un sistema di machine learning dovrebbe avere un processo in base al quale le previsioni di un modello siano spiegabili ad altre parti. Gli utenti dovrebbero essere in grado di verificare l'eventuale presenza di distorsioni nei risultati di ciascuna funzionalità. Idealment e, misura la distorsione dei dati prima di inserirli nel modello ML e registra queste metriche per le schede modello e il controllo.

3.3 Tracciamento del lignaggio: input e output dei dati

Il tracciamento viene utilizzato per seguire il flusso di dati in entrata e in uscita dal sistema (ad esempio, va dal data lake alla pipeline di formazione). Questo tracciamento funge da record da cui è possibile ricreare tutti i processi di sistema e fornisce una pista di controllo per l'analisi.

3.4 Tracciamento del lignaggio: informazioni sull'ambiente

Questo tracciamento acquisisce informazioni sulla configurazione dell'ambiente di runtime, come le immagini dei contenitori per tutto il codice del modello e le dipendenze associate ai contenitori.

3.5 Tracciamento del lignaggio: modello

Questo tracciamento acquisisce informazioni sul modello. Include tutto, dalle informazioni sull'algoritmo del modello ai parametri e agli iperparametri che compongono il modello.

3.6 Integrazione con distribuzione e monitoraggio

Il sistema dovrebbe essere collegato direttamente ai sottosistemi di monitoraggio e implementazione del PITR. Per quanto riguarda il monitoraggio, ciò significa testare le prestazioni del modello rispetto alle sessioni di addestramento per rilevare il deterioramento della qualità del modello. Per l'implementazione, questo supporta PITR e la possibilità di tornare a una versione precedente del modello, se necessario.

3.7 Configurazione dei parametri della pipeline

Tecnicamente, la configurazione dei parametri della pipeline rientra sia nel tracciamento del lignaggio che nel tracciamento dell'esperimento, poiché la configurazione della pipeline deve essere versionata e associata direttamente a un modello. La configurazione dei parametri della pipeline è elencata in questa sezione perché è fondamentale tenere traccia di tutte le configurazioni di orchestrazione del sistema e delle relative versioni.

3.8 I problemi sono tracciabili, debuggibili e riproducibili.

Un tecnico può tracciare, eseguire il debug e riprodurre tutti i problemi all'interno del sistema senza troppi sforzi. Ciò implica che esiste un livello sufficiente di osservabilità. Questo controllo deriva principalmente dall'adempimento degli altri elementi della sezione Osservabilità e gestione dei modelli.

3.9 Visualizzazione delle prestazioni

Il sistema è in grado di acquisire e raccogliere i log in un formato tipo database di serie temporali e inserirli direttamente nella dashboard. La dashboard offre una visione olistica delle metriche dei modelli e dei computer, con la possibilità di approfondire e interrogare.

4. Promozione e pipeline solide

Le pipeline offrono molte opzioni per l'ottimizzazione degli iperparametri, l'AutoML e le routine di elaborazione. Le pipeline vengono registrate da un capo all'altro. Le solide pipeline possono eseguire il training in parallelo su più istanze e framework, scalando le dimensioni del carico in base alle esigenze. Le pipeline robuste possono promuovere i modelli in produzione, implementandoli in tempo reale, in streaming e in batch. Queste implementazioni possono supportare l'inferenza a modello singolo o multimodello.

4.1 Formazione distribuita e su larga scala

Un sistema di machine learning maturo supporta la capacità di eseguire il training su istanze di grandi dimensioni ottimizzate per il calcolo in parallelo. Dispone degli strumenti necessari per garantire che queste risorse vengano utilizzate appieno e che la formazione si adatti in modo uniforme all'interno del cluster di elaborazione.

4.2 Support per più framework

Gli sviluppatori possono eseguire il porting di diversi framework di piattaforme, come PyTorch o Flax, per eseguire lavori di formazione e inferenza. Allo stesso modo, sono supportate e utilizzabili diverse lingue e versioni. Il passaggio a un altro framework non danneggerà il sistema.

4.3 Ottimizzazione degli iperparametri

Una fase di regolazione degli iperparametri fa parte della pipeline di formazione. I

modelli implementati hanno i loro iperparametri ottimizzati. Sono disponibili diverse opzioni per ottimizzare gli iperparametri. Per migliorare e la precisione, almeno una delle opzioni di regolazione deve avere un'inferenza o un approccio bayesiano.

4.4 Opzione AutoML

Per ridurre la sperimentazione e il confronto manuali, un sistema ML maturo supporta l'esecuzione di AutoML, che seleziona automaticamente la pipeline di funzionalità, gli iperparametri e il modello migliori. Nota che AutoML è una funzionalità da usare in modo pragmatico, ma non è una panacea.

4.5 Supporto all'inferenza: in tempo reale

Questo è comunemente chiamato Model as a Service (MaaS). Il sistema supporta l'inferenza in tempo reale con le operazioni dell'API REST, per richieste di inferenza su richiesta. È in grado di fornire un'infrastruttura MaaS su cui il modello può scalare sia orizzontalmente che verticalmente come API autonoma o come endpoint associato ad altre applicazioni. In alternativa, è possibile eseguire l'implementazione utilizzando la tecnologia serverless.

4.6 Supporto per l'inferenza: streaming

I modelli possono essere promossi a un formato di inferenza in tempo reale come Amazon Kinesis o Amazon Managed Streaming for Apache Kafka, in base al quale l'inferenza viene eseguita in modalità streaming sul modello. Ciò richiede che almeno il 90% della lista di controllo sia completa, poiché i guardrail, l'osservabilità e il monitoraggio sono essenziali per l'inferenza in tempo reale.

4.7 Supporto per l'inferenza: batch

Il sistema supporta la distribuzione in batch di modelli come processi pianificati o avviati. Il sistema può eseguire modelli come parte di un processo di estrazione, trasformazione e caricamento (ETL) o in modo isolato. I processi Batch registrano lo stato di ogni passaggio e vengono eseguiti secondo uno schema ordinato, ad esempio un grafo aciclico diretto. In alternativa, i job possono scrivere su un database, che funge da server per l'inferenza dei modelli.

4.8 Routine di preelaborazione e post-elaborazione

Quando necessario, i dati vengono visualizzati come parte del processo di acquisizione del modello o dei processi in batch. Se sono in gioco più modelli o più fasi, le routine di post-elaborazione si occupano della caratterizzazione dei dati.

4.9 Capacità di richiamare modelli gerarchici o simultanei

Il sistema ML può implementare più modelli insieme o eseguirli in modo sequenziale. Il primo significa l'hosting su un unico modello di endpoint su una flotta di risorse. Quest'ultimo implica che più modelli devono funzionare in modo concatenato uno dopo l'altro. Il sistema è in grado di gestire entrambi questi tipi di complessità in modo resiliente.

4.10 Strategie di scalabilità orizzontale e verticale

Una pipeline dovrebbe avere la capacità di supportare entrambi i tipi di strategie di scalabilità per la formazione e l'inferenza. Il sistema ML può aumentare le proprie dimensioni e distribuire il traffico su più macchine quando la latenza o il throughput aumentano. Le politiche per questo tipo di comportamento sono stabilite e tengono conto dell'allocazione ottimale delle risorse.

4.11 registrazione End-to-end

Il team di sviluppo dovrebbe avere la registrazione impostata all'interno di tutto il codice della pipeline in modo che la registrazione possa acquisire input, output e fasi intermedie del sistema. La registrazione dovrebbe supportare il tracciamento delle esecuzioni nella pipeline e il debug degli errori.

5. Integrazione continua

Il sistema ML esegue dei test per verificare che il sistema funzioni dall'inizio alla fine, verificando eventuali punti di errore. I test vengono eseguiti automaticamente al momento del commit e i test più lunghi vengono eseguiti secondo una pianificazione fissa. I test controllano le aree di ingegneria del software tradizionali, ad esempio a livello di unità e sistema. Inoltre, i test acquisiscono i dettagli del machine learning controllando i dati, le funzionalità e il modello.

5.1 Controlli del codice locale

Prima di inserire il codice in un repository di codice centralizzato, gli sviluppatori eseguono localmente controlli come i test unitari di base e l'analisi statica. L'esecuzione di questi controlli prima del commit aumenta la qualità complessiva del codice e rileva i problemi prima che entrino nel controllo della versione.

5.2 Analisi statica del codice

L'archivio centrale del codice dispone di strumenti statici di analisi del codice che

vengono eseguiti rapidamente su commit. Questo strumento dovrebbe migliorare lo stile e la formattazione del codice. Dovrebbe inoltre verificare le vulnerabilità di sicurezza comuni nel codice sorgente e nell'infrastruttura, i bug comuni e altri punti deboli del codice.

5.3 Test di qualità dei dati

I test di qualità dei dati dovrebbero, come minimo, verificare che i dati non abbiano violato uno schema fisso. Un approccio più completo consiste nel calcolare le statistiche dei dati al momento dell'acquisizione, stabilire vincoli sui dati ed eseguire test sulla base di tali vincoli.

È possibile impostare test di qualità dei dati indipendentemente o come parte della pipeline. Le statistiche e i vincoli vengono riutilizzati per il monitoraggio.

5.4 Test delle funzionalità

Come parte di una pipeline completa, viene generata l'importanza delle funzionalità. I feature test affermano che l'importanza delle feature, o il modo in cui il modello attribuisce i valori delle feature, non cambia. I feature test possono contribuire al monitoraggio perché possono avvisare e tenere traccia delle violazioni negli input di un modello.

5.5 Test unitari

I test unitari per tutto il codice, modello, applicazione e infrastruttura, vengono eseguiti prima del commit e al momento del commit. Ogni unit test verifica un pezzo di codice importante per confermare che funzioni come previsto. Nel caso del codice ML, è possibile eseguire test di correttezza algoritmica.

5.6 Test di integrazione

Un test di integrazione verifica che la pipeline funzioni correttamente dall'inizio alla fine, compresa la predisposizione dell'infrastruttura associata alla pipeline. Questo test verifica che il sistema funzioni e stia registrando i dati come previsto. Se la distribuzione è separata, dovrebbe essere eseguito anche un end-to-end test per assicurarsi che la distribuzione funzioni.

5.7 Test del fumo

Il sistema dispone di test di fumo che vengono eseguiti in modalità mini e rapida regressione di ogni funzionalità. Gli smoke test fanno parte dell'integrazione continua e possono essere eseguiti in un ambiente containerizzato per simulare le funzionalità del cloud.

5.8 Test di carico

Sono in corso test di carico su richiesta. Oltre a registrare il comportamento del sistema ML in presenza di carichi elevati e bassi, i test di carico forniscono statistiche sulla velocità effettiva o sulla latenza a livello di sistema. I dati raccolti tramite i test di carico forniscono informazioni sulle dimensioni delle risorse e sulle politiche di scalabilità.

5.9 Test funzionali dei modelli

Gli output e gli input del modello vengono eseguiti tramite test funzionali automatizzati. Per verificare un comportamento all'interno di una capacità, sia gli output che gli input del modello vengono testati su dati reali o falsi con esempi di base.

5.10 Test di inferenza dei modelli con casi estremi

Come parte dei test minimi di funzionalità, i test dei modelli dovrebbero verificare l'eventuale presenza di comportamenti estremi in base a determinati input prima della promozione del modello. Ciò crea un guardrail aggiuntivo che aiuta a prevenire comportamenti imprevisti.

6. Monitoraggio continuo

Nel monitoraggio continuo, i processi automatizzati osservano e rilevano problemi di prestazioni e problemi di modello. I proprietari possono quindi identificare potenziali problemi e minacce in tempo reale per risolverli rapidamente.

Il monitoraggio continuo evidenzia possibili problemi del modello, come la qualità dei dati, il cambiamento della distribuzione, il cambiamento del concetto del modello e il degrado della qualità del modello. Il monitoraggio continuo include anche la registrazione completa delle misure di sistema tradizionali come saturazione, latenza, traffico ed errori. Viene impostata una pratica strategia di notifica e avviso per avvisare i proprietari in caso di problemi.

6.1 Monitoraggio del modello: rilevamento della qualità dei dati

Il monitoraggio basato su regole consente di sapere quando i dati in entrata si discostano dai dati di addestramento del modello. Questo tipo di monitoraggio crea uno schema a partire dai dati di addestramento, imposta i vincoli in base a tale schema e quindi esegue eccezioni in caso di violazione.

6.2 Monitoraggio del modello: spostamento della distribuzione

Il monitoraggio è impostato per esaminare la distribuzione dei dati in entrata e verificare che non si sia discostata dalla distribuzione dei dati di addestramento del modello. Ad esempio, i dati in entrata vengono campionati [come una finestra mobile](#) sui dati di inferenza. Viene quindi eseguito un processo per testare la

6.3 Monitoraggio del modello: deriva del concetto di modello

distribuzione campionata e la distribuzione della formazione per verificare se sono uguali.

Un controllo della deriva concettuale verifica che la relazione tra gli input di un modello e la variabile target rimanga invariata rispetto ai dati di addestramento. Un ulteriore controllo serve a confermare che le caratteristiche relative e la loro importanza non cambino.

6.4 Monitoraggio del modello: controllo della valutazione del modello

Si tratta di un controllo di monitoraggio che valuta se la qualità del modello è peggiorata. Il controllo di valutazione del modello confronta le metriche di valutazione di base relative al periodo di formazione con i risultati in arrivo per valutare se il livello di precisione del modello è diminuito rispetto ai nuovi dati. Poiché calcola metriche di precisione, questo controllo richiede che la veridicità di base dei nuovi dati sia disponibile dopo l'inferenza.

6.5 Acquisizioni di sistema: schemi di input

Il sistema ML acquisisce lo schema dei dati di addestramento, test e convalida. Oltre a fornire informazioni sugli input, gli schemi forniscono statistiche sulla loro inclinazione e completezza. Gli schemi vengono utilizzati per test immediati e controlli di monitoraggio della qualità dei dati in produzione.

6.6 Acquisizioni di sistema: risultati di valutazione e statistiche

Il sistema ML fornisce informazioni di precisione e sui dati di convalida e formazione. È in grado di produrre previsioni ed etichette veritiere derivanti da sessioni di convalida e formazione. Questi vengono utilizzati come vincoli di monitoraggio per il modello di produzione live.

6.7 Acquisizioni di sistema: anomalie

Esiste un meccanismo di tracciamento per segnalare le anomalie nei flussi di dati in entrata. Se si verificano valori anomali nei dati in ingresso o se durante un periodo di tempo specificato la distribuzione delle funzionalità chiave cambia, il sistema riconosce che si tratta di un'anomalia e la segnala.

6.8 Registrazione: saturazione e risorse

È in atto la registrazione del livello di riempimento del sistema. Le metriche relative alle risorse e alla saturazione devono concentrarsi sull'utilizzo della CPU, sull'utilizzo delle unità di elaborazione grafica (GPU), sull'utilizzo della memoria e sull'utilizzo del disco. Queste metriche dovrebbero essere disponibili in formato di serie temporali con la possibilità di misurarle in percentili. Per i processi in batch, questo fornisce informazioni sulla velocità effettiva, che mostra quante unità di informazioni il sistema può elaborare in ogni periodo di tempo.

6.9 Registrazione: latenza

La registrazione deve essere utilizzata per misurare il ritardo nella comunicazione di rete o il tempo necessario per soddisfare una richiesta. Un tecnico dovrebbe essere in grado di valutare quanto tempo impiegano i modelli di inferenza a fornire le previsioni e quanto tempo impiega il modello a caricarsi.

6.10 Registrazione: traffico

La configurazione di registrazione per il traffico misura il volume di traffico su ciascuna istanza. Il traffico viene misurato in base al numero di richieste HTTP e di byte o pacchetti inviati o ricevuti in un determinato periodo di tempo. La registrazione del traffico fornisce informazioni sul carico di lavoro totale assegnato a un sistema.

6.11 Registrazione: errori

La configurazione di registrazione degli errori registra il numero di richieste non riuscite. Gli errori sono dei seguenti tipi:

- Espliciti (ad esempio, errori HTTP 500)
- Implicito (ad esempio, una risposta di successo HTTP 200 associata a un contenuto errato)
- Criteri (ad esempio, se ti impegni a rispettare i tempi di risposta di un secondo, qualsiasi richiesta superiore a un secondo è un errore)

Laddove i codici di risposta del protocollo non sono sufficienti a esprimere tutte le condizioni di errore, potrebbero essere necessari protocolli secondari (interni) per tenere traccia delle modalità di errore parziale.

6.12 Notifiche e avvisi

Le notifiche e gli avvisi vengono impostati dal monitoraggio. Le notifiche includono la possibilità di ricevere Slack, notifiche via email, pagine e messaggi SMS (Short Message Service). Avvisare non significa inviare notifiche per tutte le possibili violazioni. Significa invece impostare avvisi su eccezioni specifiche che sono significative e importanti per il team di sviluppo. In questo modo si evita l'affaticamento da allerta.

7. Implementazione continua

Affinché un sistema ML possa essere distribuito continuamente, deve essere in grado di deviare il traffico da o tra modelli live. Un sistema in uso continuo prevede almeno un modo per promuovere i modelli alla produzione: canary, shadow, blue/green, or A/B. Verifica che nel sistema ML sia disponibile almeno un modo per ripristinare i modelli.

7.1 Cambio di modello

Il sistema può passare da un modello all'altro in fase di allestimento e produzione. Può deviare il traffico tutto in una volta o in modo incrementale verso nuove varianti di produzione.

7.2 Processi di promozione dei modelli

È in atto un processo di convalida graduale per la promozione del modello. Il processo utilizza test offline che non influiscono sul sistema di produzione, ad esempio l'esecuzione su dati di convalida in un ambiente di staging. Vengono impostati un runbook e delle metriche per la promozione dei modelli. La promozione segue una delle strategie di lancio.

7.3 Strategie di rollback

Esiste una strategia di rollback in modo che quando si verifica un errore o il modello si discosta dal comportamento previsto, si verifichi un rollback, un fallback o un rollthrough. In caso di rollback, il modello torna a una versione di distribuzione precedente. In alternativa, il modello viene sostituito da una solida euristica. Roll through promuoverà il passaggio alla produzione del modello successivo, sostituendo il modello precedente. Esistono dei runbook per tutte queste strategie.

7.4 Implementazione di Canary

Il sistema può essere implementato utilizzando un canarino. Inizialmente una piccola parte del traffico viene inviata al nuovo modello. Nel tempo, tutto il traffico passa al nuovo

modello. Questo cambiamento viene monitorato attentamente perché i test avvengono nell'ambiente di produzione.

7.5 Distribuzione del modello shadow

Il sistema può eseguire un'implementazione ombra in cui il nuovo modello si affianca al modello esistente. Entrambi i modelli ricevono traffico, ma solo il modello precedente genera inferenze. Le valutazioni vengono eseguite sul nuovo modello rispetto al modello esistente, quindi il nuovo modello viene promosso manualmente.

7.6 Implementazione blu/verde

Il sistema può essere implementato con un nuovo modello (verde, che rappresenta l'allestimento temporaneo) e il modello precedente (blu, che rappresenta la produzione), entrambi funzionanti contemporaneamente. Al termine del test, il traffico viene deviato dall'ambiente blu a quello verde. Questa strategia previene i tempi di inattività perché sono presenti ambienti identici.

7.7 Support per test A/B o altro

Il sistema supporta l'utilizzo di versioni del modello nell'ambiente distribuito per eseguire test A/B sul traffico in entrata. Ciò può includere la possibilità di effettuare la promozione automaticamente in base al modello più recente che vince nei test. Le configurazioni più avanzate utilizzeranno un processo bandito a [più braccia](#).

8. Formazione continua

La formazione continua significa che il sistema ML riqualifica automaticamente e continuamente i modelli di machine learning per adattarli alle modifiche dei dati prima che vengano ridistribuiti. I

possibili fattori scatenanti per la ricostruzione includono modifiche ai dati, modifiche al modello o modifiche al codice.

8.1 Controlli: convalida dell'input del modello

Sono in atto controlli per verificare che l'input di un modello non si discosti da un determinato standard. La convalida degli input significa eseguire test funzionali durante la promozione del modello. Significa anche avere una verifica immediata delle richieste di input, ad esempio utilizzando asserzioni e tipi enumerati.

8.2 Attivazione della riqualificazione: lavori pianificati

Questa è la forma più elementare di automazione della formazione. La riqualificazione del modello è impostata su un programma (ad esempio, ogni settimana). In questo scenario, l'automazione è probabilmente bassa, con una revisione manuale e un controllo a campione dei risultati prima della promozione del modello.

8.3 Retrain Trigger: nuovi dati di allenamento

La riqualificazione viene avviata da una soglia di dati in entrata. Il modello può riqualificarsi da zero o eseguire aggiornamenti in modo incrementale. Data una determinata quantità di dati disponibile, inizia un processo di formazione.

8.4 Attivazione della ritrasmissione: degrado delle prestazioni del modello

Questa tecnica utilizza il monitoraggio e l'osservabilità per eseguire la riqualificazione del modello e richiede un livello di automazione maturo. Ad esempio, la precisione diminuisce rispetto a un determinato intervallo, il che funge da fattore scatenante per la riqualificazione di un modello su tutti o parte dei dati.

8.5 Attivazione della ritrasmissione: spostamento della distribuzione dei dati

Il monitoraggio dello spostamento della distribuzione dei dati fornisce un modo per impostare i trigger per riqualificare il modello

quando i dati sottostanti cambiano. Una violazione causata da un [cambio di concetto o di distribuzione dei dati](#) avvia un processo di riqualificazione del modello.

9. Governance

La governance del machine learning comprende una serie di processi e framework che aiutano nella distribuzione di modelli ML. Include la spiegabilità dei modelli, la verificabilità, la tracciabilità e altri requisiti più astratti ma essenziali per un ciclo di vita ML di successo. end-to-end

9.1 Qualità e conformità dei dati

Il sistema ML tiene conto delle considerazioni relative alle informazioni di identificazione personale (PII), inclusa l'anonimizzazione. Ha documentato e analizzato la discendenza a livello di colonna per comprendere la fonte, la qualità e l'adeguatezza dei dati. Dispone inoltre di controlli automatici della qualità dei dati per individuare eventuali anomalie.

9.2 Audit e documentazione

Il sistema ML dispone di un registro completo di tutte le modifiche apportate durante lo sviluppo, compresi gli esperimenti eseguiti e i motivi delle scelte effettuate per la conformità normativa.

9.3 Riproducibilità e tracciabilità

Il sistema ML include un'istantanea completa dei dati per una reistanziatura precisa e rapida del modello, oppure ha la capacità di ricreare l'ambiente e riaddestrarlo con un campione di dati.

9.4 Autorizzazione Human-in-the-loop

Il sistema ML prevede la verifica e l'autorizzazione manuali per la conformità normativa. Il sistema richiede l'approvazione per ogni spostamento dell'ambiente (ad esempio, Dev, QA, Pre-Prod e Prod).

9.5 Test di pregiudizi e attacchi avversi

Il sistema ML prevede test antagonisti del Red Team che utilizzano più strumenti e vettori di attacco e un controllo automatico dei pregiudizi su sottopopolazioni specifiche. Questo componente si ricollega alla sezione Osservabilità e gestione dei modelli.

Esempio di utilizzo della lista di MLOps controllo

[Non esiste un solo modo per utilizzare la lista di controllo. MLOps](#) Uno schema comune, tuttavia, è quello di ripetere la procedura fino a raggiungere i livelli desiderati in termini di qualità e prestazioni del sistema.

Un ciclo di iterazione può essere simile al seguente:

1. Esamina la lista di controllo elemento per elemento.
2. Rivedi i risultati della lista di controllo esistente.
3. Identifica eventuali lacune.
4. Dai priorità alle lacune per porvi rimedio.
5. Sviluppa implementazioni per le nuove funzionalità che desideri.
6. Ripeti il ciclo di iterazione.

Risorse

Riferimenti

- David Nigenda, Zohar Karnin, Muhammad Bilal Zafar, Raghu Ramesha, Alan Tan, Michele Donini e Krishnaram Kenthapadi. 2022. «Amazon SageMaker Model Monitor: un sistema per approfondimenti in tempo reale sui modelli di Machine Learning distribuiti»: [arXiv:2111.13657](https://arxiv.org/abs/2111.13657).
- Richard Meyes, Melanie Lu, Constantin Waubert de Puiseau e Tobias Meisen. 2019. «[Studi di ablazione in reti neurali artificiali](https://arxiv.org/abs/1901.08644)»: [arXiv:1901.08644](https://arxiv.org/abs/1901.08644).
- Brian Babcock, Mayur Datar e Rajeev Motwani. «[Campionamento da una finestra mobile su dati in streaming](https://www.princeton.edu/~babcock/campionamento)»: [Princeton University](https://www.princeton.edu/~babcock/campionamento).
- Chip Huyen. [Progettazione di sistemi di Machine Learning: un processo iterativo per applicazioni pronte per la produzione](https://www.oreilly.com/catalog/errata/csp/bookerrata.php?isbn=9781492095270). Sebastopoli, California: O'Reilly Media, 2022.
- Cathy Chen, Niall Murphy, Kranti Parisa, D. Sculley e Todd Underwood. [Machine Learning affidabile: applicazione dei principi SRE al machine learning in produzione](https://www.oreilly.com/catalog/errata/csp/bookerrata.php?isbn=9781492095270). Sebastopoli, California: O'Reilly Media, 2022.

Lettura aggiuntiva

- [AWS Well-Architected Framework: obiettivo per il Machine Learning](#)
- [Rileva la deriva dei dati NLP utilizzando Amazon SageMaker Model Monitor personalizzato](#) (post sul blog)
- [Test A/B dinamici per modelli di machine learning con SageMaker MLOps progetti Amazon](#) (post sul blog)

Strumenti

- [Amazon SageMaker](#)
- [Controllo della versione dei dati \(DVC\)](#)

Consulenza

- [Servizi professionali AWS](#)

Cronologia dei documenti

La tabella seguente descrive le modifiche significative apportate a questa guida. Per ricevere notifiche sugli aggiornamenti futuri, puoi abbonarti a un [feed RSS](#).

Modifica	Descrizione	Data
Pubblicazione iniziale	—	26 luglio 2023

AWS Glossario delle linee guida prescrittive

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

Numeri

7 R

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

- **Rifattorizzare/riprogettare:** trasferisci un'applicazione e modifica la sua architettura sfruttando appieno le funzionalità native del cloud per migliorare l'agilità, le prestazioni e la scalabilità. Ciò comporta in genere la portabilità del sistema operativo e del database. Esempio: migra il tuo database Oracle locale all'edizione compatibile con Amazon Aurora PostgreSQL.
- **Ridefinire la piattaforma (lift and reshape):** trasferisci un'applicazione nel cloud e introduci un certo livello di ottimizzazione per sfruttare le funzionalità del cloud. Esempio: migra il tuo database Oracle locale ad Amazon Relational Database Service (Amazon RDS) per Oracle in Cloud AWS
- **Riacquistare (drop and shop):** passa a un prodotto diverso, in genere effettuando la transizione da una licenza tradizionale a un modello SaaS. Esempio: migra il tuo sistema di gestione delle relazioni con i clienti (CRM) su Salesforce.com.
- **Eseguire il rehosting (lift and shift):** trasferisci un'applicazione sul cloud senza apportare modifiche per sfruttare le funzionalità del cloud. Esempio: migra il database Oracle locale su Oracle su un'istanza in EC2 Cloud AWS
- **Trasferire (eseguire il rehosting a livello hypervisor):** trasferisci l'infrastruttura sul cloud senza acquistare nuovo hardware, riscrivere le applicazioni o modificare le operazioni esistenti. Si esegue la migrazione dei server da una piattaforma locale a un servizio cloud per la stessa piattaforma. Esempio: migra un'applicazione su Microsoft 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 di 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 aggregata

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

Intelligenza artificiale

Vedi [intelligenza artificiale](#).

AIOps

Guarda le [operazioni di intelligenza artificiale](#).

anonimizzazione

Il processo di eliminazione permanente delle informazioni personali in un set di dati.

L'anonimizzazione può aiutare a proteggere la privacy personale. I dati anonimi non sono più considerati dati personali.

anti-modello

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

controllo delle applicazioni

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

portfolio di applicazioni

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

intelligenza artificiale (IA)

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

operazioni di intelligenza artificiale (AIOps)

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

crittografia asimmetrica

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

atomicità, consistenza, isolamento, durabilità (ACID)

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

Controllo degli accessi basato su attributi (ABAC)

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

fonte di dati autorevole

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

Zona di disponibilità

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

AWS Cloud Adoption Framework (CAF)AWS

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

AWS Workload Qualification Framework (WQF)AWS

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

B

bot difettoso

Un [bot](#) che ha lo scopo di interrompere o causare danni a individui o organizzazioni.

BCP

Vedi la [pianificazione della continuità operativa](#).

grafico comportamentale

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

sistema big-endian

Un sistema che memorizza per primo il byte più importante. Vedi anche [endianness](#).

Classificazione binaria

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

filtro Bloom

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

distribuzioni blu/verdi

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

bot

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

botnet

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

ramo

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

accesso break-glass

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

strategia brownfield

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

cache del buffer

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

capacità di business

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

pianificazione della continuità operativa (BCP)

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

C

CAF

Vedi [AWS Cloud Adoption Framework](#).

implementazione canaria

Il rilascio lento e incrementale di una versione agli utenti finali. Quando sei sicuro, distribuisce 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 o Bitbucket 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 is commonly described as a 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, puoi utilizzarlo AWS CloudFormation per [rilevare la deriva nelle risorse di sistema](#) oppure puoi usarlo AWS Control Tower per [rilevare cambiamenti nella tua landing zone](#) che potrebbero influire sulla conformità ai requisiti di governance.

DVSM

Vedi la [mappatura del flusso di valore dello sviluppo](#).

E

EDA

Vedi [analisi esplorativa dei dati](#).

MODIFICA

Vedi [scambio elettronico di dati](#).

edge computing

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

scambio elettronico di dati (EDI)

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

crittografia

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

chiave crittografica

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

endianità

L'ordine in cui i byte vengono archiviati nella memoria del computer. 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 pipeline CI/CD, l'ambiente di produzione è l'ultimo ambiente di implementazione.
- 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 epiche della sicurezza AWS CAF includono la gestione delle identità e degli accessi, i controlli investigativi, la sicurezza dell'infrastruttura, la protezione dei dati e la risposta agli incidenti. Per ulteriori informazioni sulle epiche, consulta la strategia di migrazione AWS , consulta la [guida all'implementazione del programma](#).

ERP

Vedi [pianificazione delle risorse aziendali](#).

analisi esplorativa dei dati (EDA)

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

F

tabella dei fatti

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

fallire velocemente

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

limite di isolamento dei guasti

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

ramo di funzionalità

Vedi [filiale](#).

caratteristiche

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

importanza delle caratteristiche

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

trasformazione delle funzionalità

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

prompt con pochi scatti

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

FGAC

Vedi il controllo [granulare degli accessi](#).

controllo granulare degli accessi (FGAC)

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

migrazione flash-cut

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

FM

[Vedi il modello di base.](#)

modello di fondazione (FM)

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

G

AI generativa

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

blocco geografico

Vedi [restrizioni geografiche](#).

limitazioni geografiche (blocco geografico)

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

Flusso di lavoro di GitFlow

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

immagine dorata

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

strategia greenfield

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

guardrail

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

H

AH

Vedi [disponibilità elevata](#).

migrazione di database eterogenea

Migrazione del database di origine in un database di destinazione che utilizza un motore di database diverso (ad esempio, da Oracle ad Amazon Aurora). La migrazione eterogenea fa in

genere parte di uno sforzo di riprogettazione e la conversione dello schema può essere un'attività complessa. [AWS offre AWS SCT](#) che aiuta con le conversioni dello schema.

alta disponibilità (HA)

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

modernizzazione storica

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

dati di esclusione

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

migrazione di database omogenea

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

dati caldi

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

hotfix

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

periodo di hypercare

Subito dopo la conversione, il periodo di tempo in cui un team di migrazione gestisce e monitora le applicazioni migrate nel cloud per risolvere eventuali problemi. In genere, questo periodo dura

da 1 a 4 giorni. Al termine del periodo di hypercare, il team addetto alla migrazione in genere trasferisce la responsabilità delle applicazioni al team addetto alle operazioni cloud.

I

IaC

Considera [l'infrastruttura come codice](#).

Policy basata su identità

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

applicazione inattiva

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

IIoT

Vedi [Industrial Internet of Things](#).

infrastruttura immutabile

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

VPC in ingresso (ingress)

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

migrazione incrementale

Una strategia di conversione in cui si esegue la migrazione dell'applicazione in piccole parti anziché eseguire una conversione singola e completa. Ad esempio, inizialmente potresti spostare

I

solo alcuni microservizi o utenti nel nuovo sistema. Dopo aver verificato che tutto funzioni correttamente, puoi spostare in modo incrementale microservizi o utenti aggiuntivi fino alla disattivazione del sistema legacy. Questa strategia riduce i rischi associati alle migrazioni di grandi dimensioni.

Industria 4.0

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

infrastruttura

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

infrastruttura come codice (IaC)

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

IIoInternet delle cose industriale (T)

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

VPC di ispezione

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

Internet of Things (IoT)

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

interpretabilità

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

IoT

Vedi [Internet of Things](#).

libreria di informazioni IT (ITIL)

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

gestione dei servizi IT (ITSM)

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

ITIL

Vedi la [libreria di informazioni IT](#).

ITSM

Vedi [Gestione dei servizi IT](#).

L

controllo degli accessi basato su etichette (LBAC)

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

zona di destinazione

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

informazioni sulle zone di destinazione, consulta la sezione [Configurazione di un ambiente AWS multi-account sicuro e scalabile](#).

modello linguistico di grandi dimensioni (LLM)

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

migrazione su larga scala

Una migrazione di 300 o più server.

BIANCO

Vedi controllo degli accessi [basato su etichette](#).

Privilegio minimo

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

eseguire il rehosting (lift and shift)

Vedi [7 R](#).

sistema little-endian

Un sistema che memorizza per primo il byte meno importante. Vedi anche [endianità](#).

LLM

Vedi [modello linguistico di grandi dimensioni](#).

ambienti inferiori

Vedi [ambiente](#).

M

machine learning (ML)

Un tipo di intelligenza artificiale che utilizza algoritmi e tecniche per il riconoscimento e l'apprendimento di schemi. Il machine learning analizza e apprende dai dati registrati, come i dati

dell'Internet delle cose (IoT), per generare un modello statistico basato su modelli. Per ulteriori informazioni, consulta la sezione [Machine learning](#).

ramo principale

Vedi [filiale](#).

malware

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

servizi gestiti

Servizi AWS per cui AWS gestisce il livello di infrastruttura, il sistema operativo e le piattaforme e si accede agli endpoint per archiviare e recuperare i dati. Amazon Simple Storage Service (Amazon S3) Simple Storage Service (Amazon S3) e Amazon DynamoDB sono esempi di servizi gestiti. Questi sono noti anche come servizi astratti.

sistema di esecuzione della produzione (MES)

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

MAP

Vedi [Migration Acceleration Program](#).

meccanismo

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

account membro

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

MEH.

Vedi [sistema di esecuzione della produzione](#).

Message Queuing Telemetry Transport (MQTT)

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

microservizio

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

architettura di microservizi

Un approccio alla creazione di un'applicazione con componenti indipendenti che eseguono ogni processo applicativo come microservizio. Questi microservizi comunicano attraverso un'interfaccia ben definita utilizzando sistemi leggeri. APIs Ogni microservizio in questa architettura può essere aggiornato, distribuito e dimensionato per soddisfare la richiesta di funzioni specifiche di un'applicazione. Per ulteriori informazioni, vedere [Implementazione dei microservizi](#) su. AWS

Programma di accelerazione della migrazione (MAP)

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

migrazione su larga scala

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

fabbrica di migrazione

Team interfunzionali che semplificano la migrazione dei carichi di lavoro attraverso approcci automatizzati e agili. I team di Migration Factory in genere includono addetti alle operazioni,

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

metadati di migrazione

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

modello di migrazione

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

Valutazione del portfolio di migrazione (MPA)

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

valutazione della preparazione alla migrazione (MRA)

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

strategia di migrazione

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

ML

[Vedi machine learning.](#)

modernizzazione

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

valutazione della preparazione alla modernizzazione

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

applicazioni monolitiche (monoliti)

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

MAPPA

Vedi [Migration Portfolio Assessment](#).

MQTT

Vedi [Message Queuing Telemetry Transport](#).

classificazione multiclasse

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

infrastruttura mutabile

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

O

OAC

Vedi [Origin Access Control](#).

QUERCIA

Vedi [Origin Access Identity](#).

OCM

Vedi [gestione delle modifiche organizzative](#).

migrazione offline

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

OI

Vedi [l'integrazione delle operazioni](#).

OLA

Vedi accordo a [livello operativo](#).

migrazione online

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

OPC-UA

Vedi [Open Process Communications - Unified Architecture](#).

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

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

accordo a livello operativo (OLA)

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

revisione della prontezza operativa (ORR)

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

tecnologia operativa (OT)

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

integrazione delle operazioni (OI)

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

trail organizzativo

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

gestione del cambiamento organizzativo (OCM)

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

persone, a causa della velocità di cambiamento richiesta nei progetti di adozione del cloud. Per ulteriori informazioni, consultare la [Guida OCM](#).

controllo dell'accesso all'origine (OAC)

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

PUT DELETE

identità di accesso origine (OAI)

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

ORR

[Vedi la revisione della prontezza operativa.](#)

- NON

Vedi la [tecnologia operativa](#).

VPC in uscita (egress)

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

P

limite delle autorizzazioni

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

informazioni di identificazione personale (PII)

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

Informazioni che consentono l'identificazione personale degli utenti

Visualizza le [informazioni di identificazione personale](#).

playbook

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

PLC

Vedi [controllore logico programmabile](#).

PLM

Vedi la gestione [del ciclo di vita del prodotto](#).

policy

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

persistenza poliglotta

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

valutazione del portfolio

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

predicate

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

predicato pushdown

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

controllo preventivo

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

principale

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

privacy fin dalla progettazione

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

zone ospitate private

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

controllo proattivo

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

gestione del ciclo di vita del prodotto (PLM)

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

Ambiente di produzione

[Vedi ambiente.](#)

controllore logico programmabile (PLC)

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

concatenamento rapido

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

pseudonimizzazione

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

publish/subscribe (pub/sub)

Un modello che consente comunicazioni asincrone tra microservizi per migliorare la scalabilità e la reattività. Ad esempio, in un [MES](#) basato su microservizi, un microservizio può pubblicare messaggi di eventi su un canale a cui altri microservizi possono abbonarsi. Il sistema può aggiungere nuovi microservizi senza modificare il servizio di pubblicazione.

Q

Piano di query

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

regressione del piano di query

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

R

Matrice RACI

Vedi [responsabile, responsabile, consultato, informato \(RACI\)](#).

STRACCIO

Vedi [Retrieval](#) Augmented Generation.

ransomware

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

Matrice RASCI

Vedi [responsabile, responsabile, consultato, informato \(RACI\)](#).

RCAC

Vedi controllo dell'[accesso a righe e colonne](#).

replica di lettura

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

riprogettare

Vedi [7 Rs](#).

obiettivo del punto di ripristino (RPO)

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

obiettivo del tempo di ripristino (RTO)

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

rifattorizzare

Vedi [7 R.](#)

Regione

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

regressione

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

riospitare

Vedi [7 R.](#)

rilascio

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

trasferisco

Vedi [7 Rs.](#)

ripiattaforma

Vedi [7 Rs.](#)

riacquisto

Vedi [7 Rs.](#)

resilienza

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

policy basata su risorse

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

matrice di assegnazione di responsabilità (RACI)

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

controllo reattivo

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

retain

Vedi [7 R](#).

andare in pensione

Vedi [7 Rs](#).

Retrieval Augmented Generation (RAG)

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

rotazione

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

controllo dell'accesso a righe e colonne (RCAC)

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

RPO

Vedi l'obiettivo del punto [di ripristino](#).

RTO

Vedi l'[obiettivo del tempo di ripristino](#).

runbook

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

S

SAML 2.0

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

SCADA

Vedi [controllo di supervisione e acquisizione dati](#).

SCP

Vedi la [politica di controllo del servizio](#).

Secret

In AWS Secrets Manager, informazioni riservate o riservate, come una password o le credenziali utente, archiviate in forma crittografata. È costituito dal valore segreto e dai relativi metadati. Il valore segreto può essere binario, una stringa singola o più stringhe. Per ulteriori informazioni, consulta [Cosa c'è in un segreto di Secrets Manager?](#) nella documentazione di Secrets Manager.

sicurezza fin dalla progettazione

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

controllo di sicurezza

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

rafforzamento della sicurezza

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

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

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

automazione della risposta alla sicurezza

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

Crittografia lato server

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

Policy di controllo dei servizi (SCP)

Una politica che fornisce il controllo centralizzato sulle autorizzazioni per tutti gli account di un'organizzazione in. AWS Organizations SCPs definire barriere o fissare limiti alle azioni che un amministratore può delegare a utenti o ruoli. È possibile utilizzarli SCPs come elenchi consentiti o elenchi di rifiuto, per specificare quali servizi o azioni sono consentiti o proibiti. Per ulteriori informazioni, consulta [le politiche di controllo del servizio](#) nella AWS Organizations documentazione.

endpoint del servizio

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

accordo sul livello di servizio (SLA)

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

indicatore del livello di servizio (SLI)

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

obiettivo a livello di servizio (SLO)

[Una metrica target che rappresenta lo stato di un servizio, misurato da un indicatore del livello di servizio.](#)

Modello di responsabilità condivisa

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

SIEM

Vedi il [sistema di gestione delle informazioni e degli eventi sulla sicurezza](#).

punto di errore singolo (SPOF)

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

SLAM

Vedi il contratto sul [livello di servizio](#).

SLI

Vedi l'indicatore del [livello di servizio](#).

LENTA

Vedi obiettivo del [livello di servizio](#).

split-and-seed modello

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

SPOF

Vedi [punto di errore singolo](#).

schema a stella

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

modello del fico strangolatore

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

sottorete

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

controllo di supervisione e acquisizione dati (SCADA)

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

crittografia simmetrica

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

test sintetici

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

prompt di sistema

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

T

tags

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

variabile di destinazione

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

elenco di attività

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

Ambiente di test

[Vedi ambiente.](#)

training

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

Transit Gateway

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

flusso di lavoro basato su trunk

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

Accesso attendibile

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

regolazione

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

team da due pizze

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

U

incertezza

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

compiti indifferenziati

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

ambienti superiori

[Vedi ambiente.](#)

V

vacuum

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

controllo delle versioni

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

Peering VPC

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

vulnerabilità

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

W

cache calda

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

dati caldi

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

funzione finestra

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

Carico di lavoro

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

flusso di lavoro

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

VERME

Vedi [scrivere una volta, leggere molti](#).

WQF

Vedi [AWS Workload Qualification Framework](#).

scrivi una volta, leggi molte (WORM)

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

Z

exploit zero-day

[Un attacco, in genere malware, che sfrutta una vulnerabilità zero-day.](#)

vulnerabilità zero-day

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

prompt zero-shot

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

applicazione zombie

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

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