AWS Nozioni di base su più regioni

# AWS Guida prescrittiva

# AWS Guida prescrittiva: AWS Nozioni di base su più regioni

# Table of Contents

# AWS Nozioni di base su più regioni

John Formento, Amazon Web Services ()AWS

Dicembre 2024 (cronologia dei documenti)

Questa guida avanzata di 300 livelli è destinata agli architetti del cloud e ai dirigenti senior che creano carichi di lavoro AWS e sono interessati a utilizzare un'architettura multiregionale per migliorare la resilienza dei propri carichi di lavoro. Questa guida presuppone una conoscenza di base dell'infrastruttura e dei servizi. AWS Descrive i casi d'uso comuni in più regioni, condivide i concetti e le implicazioni fondamentali relativi alla progettazione, allo sviluppo e all'implementazione e fornisce linee guida prescrittive per aiutarti a determinare meglio se un'architettura multiregionale è adatta ai tuoi carichi di lavoro.

## Sei Well-Architected?

Il AWS Well-Architected Framework ti aiuta a comprendere i pro e i contro delle decisioni che prendi quando crei sistemi nel cloud. I sei pilastri del Framework forniscono le migliori pratiche architettoniche per progettare e gestire sistemi affidabili, sicuri, efficienti, convenienti e sostenibili. Puoi utilizzare il AWS Well-Architected Tool, disponibile gratuitamente su AWS Management Console, per esaminare i tuoi carichi di lavoro rispetto a queste best practice rispondendo a una serie di domande per ogni pilastro.

Per ulteriori indicazioni degli esperti e best practice per la tua architettura cloud, tra cui implementazioni dell'architettura di riferimento, diagrammi e guide tecniche, consulta l'Architecture Center.AWS

## Introduzione

Ciascuna Regione AWSè composta da più zone di disponibilità indipendenti e fisicamente separate all'interno di un'area geografica. Viene mantenuta una rigorosa separazione logica tra i servizi software di ciascuna regione. Questa progettazione mirata garantisce che un guasto dell'infrastruttura o del servizio in una regione non si traduca in un guasto correlato in un'altra regione.

La maggior parte AWS degli utenti può raggiungere i propri obiettivi di resilienza per un carico di lavoro in una singola regione utilizzando più zone di disponibilità o aree regionali. Servizi AWS Tuttavia, un sottoinsieme di utenti utilizza architetture multiregionali per tre motivi:

- Hanno requisiti di elevata disponibilità e continuità delle operazioni per i carichi di lavoro di livello più elevato e desiderano stabilire un tempo di ripristino limitato in caso di danni che influiscono sulle risorse in una singola regione.
- Devono soddisfare i requisiti di sovranità dei dati (come il rispetto delle leggi, dei regolamenti e della conformità locali) che richiedono che i carichi di lavoro operino all'interno di una determinata giurisdizione.
- Devono migliorare le prestazioni e l'esperienza del cliente per il carico di lavoro eseguendo i carichi di lavoro nelle sedi più vicine agli utenti finali.

Questa guida si concentra sui requisiti di elevata disponibilità e continuità delle operazioni e aiuta a orientarsi tra le considerazioni relative all'adozione di un'architettura multiregionale per un carico di lavoro. Descrive i concetti fondamentali che si applicano alla progettazione, allo sviluppo e alla distribuzione di un carico di lavoro multiregionale e fornisce un framework prescrittivo per aiutarti a determinare se un'architettura multiregionale è la scelta giusta per un particolare carico di lavoro. Devi assicurarti che un'architettura multiregionale sia la scelta giusta per il tuo carico di lavoro perché queste architetture sono impegnative e, se l'architettura multiregionale non è costruita correttamente, è possibile che la disponibilità complessiva del carico di lavoro diminuisca.

# Progettazione e gestione per la resilienza in un'unica regione

Prima di approfondire i concetti relativi a più regioni, inizia confermando che il carico di lavoro è già il più resiliente possibile in una singola regione. Per raggiungere questo obiettivo, valuta il tuo carico di lavoro rispetto al pilastro dell'affidabilità e dell'eccellenza operativa del AWS Well-Architected Framework e apporta le modifiche necessarie in base ai compromessi e alla valutazione del rischio. I seguenti concetti sono trattati nel AWS Well-Architected Framework:

- Segmentazione del carico di lavoro basata sui confini del dominio
- Contratti di assistenza ben definiti
- Gestione e accoppiamento delle dipendenze
- Gestione degli errori, dei nuovi tentativi e delle strategie di back-off
- Operazioni idempotenti e transazioni stateful o stateless
- Prontezza operativa e gestione delle modifiche
- Comprendere lo stato del carico di lavoro
- Rispondere agli eventi

Per approfondire ulteriormente la resilienza a regione singola, rivedi e applica i concetti discussi nel paper Advanced Multi-AZ Resilience Patterns: Detecting and Mitigating Gray Failures. Questo paper fornisce le best practice per l'utilizzo delle repliche in ogni zona di disponibilità per contenere gli errori e amplia i concetti Multi-AZ introdotti nel Well Architected Framework. AWS Sebbene un'architettura multiregionale possa mitigare le modalità di errore legate alle zone di disponibilità, è opportuno prendere in considerazione alcuni compromessi associati a un approccio multiregionale. Ecco perché ti consigliamo di iniziare con un approccio Multi-AZ e quindi di valutare un carico di lavoro specifico rispetto ai fondamenti delle architetture multiregionali per determinare se un approccio multiregionale può aumentare la resilienza del carico di lavoro.

# Principi fondamentali per più regioni 1: Comprensione dei requisiti

Come accennato in precedenza, l'elevata disponibilità e la continuità delle operazioni sono ragioni comuni per perseguire architetture multiregionali. Le metriche di disponibilità misurano la percentuale di tempo in cui un carico di lavoro è disponibile per l'uso in un periodo definito, mentre le metriche di continuità delle operazioni misurano il tempo di ripristino per eventi su larga scala, e in genere di durata maggiore.

La misurazione della disponibilità è un processo quasi continuo. Le misurazioni specifiche possono variare, ma in genere si fondono attorno a una metrica di disponibilità target, spesso denominata nove (ad esempio una disponibilità del 99,99 percento). Con gli obiettivi di disponibilità, un'unica soluzione non va bene per tutti. È necessario stabilire obiettivi di disponibilità a livello di carico di lavoro e separare i componenti non critici dai componenti critici, anziché applicare un unico obiettivo a tutti i carichi di lavoro.

Per la continuità delle operazioni, in genere vengono utilizzate le seguenti point-in-time misurazioni:

- Recovery Time Objective (RTO): RTO è il ritardo massimo accettabile tra l'interruzione del servizio e il ripristino del servizio. Questo valore determina una durata accettabile per la quale il servizio è compromesso.
- Recovery Point Objective (RPO): l'RPO è il periodo di tempo massimo accettabile dall'ultimo punto di ripristino dei dati. Ciò determina quella che viene considerata una perdita di dati accettabile tra l'ultimo punto di ripristino e un'interruzione del servizio.

Analogamente alla definizione degli obiettivi di disponibilità, anche RTO e RPO dovrebbero essere definiti a livello di carico di lavoro. Una continuità operativa più aggressiva o un'elevata disponibilità richiedono maggiori investimenti. Detto questo, non tutte le applicazioni possono richiedere o richiedono lo stesso livello di resilienza. Allineare i titolari di aziende e sistemi IT per valutare la criticità delle applicazioni in base all'impatto sul business e poi suddividerle di conseguenza su più livelli può contribuire a fornire un punto di partenza. Le tabelle seguenti forniscono esempi di suddivisione in più livelli.

Questa tabella mostra un esempio di resilienza su più livelli per gli accordi sui livelli di servizio (). SLAs

| Livello di resilienza | SLA di disponibilità | Tempo di inattività accettabile/anno |
|---|---|---|
| Platino | 99,99% | 52,60 minuti |
| Oro | 99,90% | 8,77 ore |
| Argento | 99,5% | 1,83 giorni |

La tabella seguente mostra un esempio di resilienza su più livelli per RTO e RPO.

| Livello di resilienza | RTO massimo | RPO massimo | Criteri | Costo |
|---|---|---|---|---|
| Platino | 15 minuti | 5 minuti | Carichi di lavoro mission-critical | $$$ |
| Oro | 15 minuti — 6 ore | 2 ore | Carichi di lavoro importanti ma non cruciali | $$ |
| Argento | 6 ore — pochi giorni | 24 ore | Carichi di lavoro non critici | $ |

Quando progetti carichi di lavoro per la resilienza, considera la relazione tra alta disponibilità e continuità delle operazioni. Ad esempio, se un carico di lavoro richiede una disponibilità del 99,99 percento, non sono tollerabili più di 53 minuti di inattività all'anno. Possono essere necessari almeno 5 minuti per rilevare un guasto e altri 10 minuti prima che un operatore interagisca, prenda decisioni sulle fasi di ripristino ed esegua queste operazioni. Non è insolito impiegare dai 30 ai 45 minuti per il ripristino di un singolo problema. In questo caso, è utile disporre di una strategia multiregionale per fornire un'istanza isolata che rimuova l'impatto correlato. In questo modo è possibile garantire la continuità delle operazioni grazie al failover entro un periodo di tempo limitato, mentre si procede alla valutazione del danno iniziale in modo indipendente. È qui che è necessario definire il tempo di ripristino limitato appropriato e garantire l'allineamento.

Un approccio multiregionale potrebbe essere appropriato per carichi di lavoro mission-critical che hanno esigenze di disponibilità estreme (ad esempio, disponibilità del 99,99% o superiore) o requisiti

rigorosi di continuità delle operazioni che possono essere soddisfatti solo eseguendo il failover in un'altra regione. Tuttavia, questi requisiti sono in genere applicabili solo a un piccolo sottoinsieme del portafoglio di carichi di lavoro di un'azienda con un tempo di ripristino limitato, misurato in minuti o ore. A meno che un'applicazione non richieda un tempo di ripristino di pochi minuti o poche ore, potrebbe essere un approccio migliore attendere che un'interruzione regionale dell'applicazione venga risolta nella regione interessata. Questo approccio è in genere allineato ai carichi di lavoro di livello inferiore.

Prima di implementare un'architettura multiregionale, i responsabili delle decisioni aziendali e i team tecnici devono essere allineati sulle implicazioni in termini di costi, compresi i fattori di costo operativi e infrastrutturali. Una tipica architettura multiregionale può comportare un costo doppio rispetto a un approccio a regione singola. Sebbene esistano diversi modelli multiregionali per la continuità aziendale, ad esempio l'utilizzo di hot standby, warm standby o luce pilota, il modello con il rischio più basso di raggiungere gli obiettivi di ripristino comporterà l'utilizzo di hot standby e raddoppierà il costo del carico di lavoro.

## Linee guida chiave

- Gli obiettivi di disponibilità e continuità delle operazioni, come RTO e RPO, devono essere stabiliti per carico di lavoro e allineati agli stakeholder aziendali e IT.
- La maggior parte degli obiettivi di disponibilità e continuità delle operazioni può essere raggiunta all'interno di una singola regione. Per quanto riguarda gli obiettivi che non possono essere raggiunti all'interno di una singola regione, è consigliabile prendere in considerazione più aree geografiche con una visione chiara dei compromessi tra costi, complessità e vantaggi.

# Principi fondamentali per più regioni 2: comprensione dei dati

La gestione dei dati è un problema non banale quando si adottano architetture multiregionali. La distanza geografica tra le regioni impone una latenza inevitabile che si manifesta con il tempo necessario per replicare i dati tra le regioni. Saranno necessari compromessi tra disponibilità, coerenza dei dati e introduzione di una maggiore latenza in un carico di lavoro che utilizza un'architettura multiregionale. Sia che si utilizzi la replica asincrona o sincrona, sarà necessario modificare l'applicazione per gestire i cambiamenti comportamentali imposti dalla tecnologia di replica. Le sfide relative alla coerenza e alla latenza dei dati rendono molto difficile convertire un'applicazione esistente progettata per una singola regione e renderla multiregionale. Comprendere i requisiti di coerenza dei dati e i modelli di accesso ai dati per carichi di lavoro particolari è fondamentale per valutare i compromessi.

## 2.a: Comprendere i requisiti di coerenza dei dati

Il teorema CAP fornisce un riferimento per ragionare sui compromessi tra coerenza dei dati, disponibilità e partizioni di rete. Solo due di questi requisiti possono essere soddisfatti contemporaneamente per un carico di lavoro. Per definizione, un'architettura multiregionale include partizioni di rete tra regioni, quindi è necessario scegliere tra disponibilità e coerenza.

Se si seleziona la disponibilità dei dati tra le regioni, non si verificherà una latenza significativa durante le operazioni di scrittura transazionale, poiché la dipendenza dalla replica asincrona dei dati impegnati tra le regioni si traduce in una riduzione della coerenza tra le regioni fino al completamento della replica. Con la replica asincrona, in caso di errore nella regione principale, è molto probabile che le operazioni di scrittura siano in attesa della replica dalla regione principale. Ciò porta a uno scenario in cui i dati più recenti non sono disponibili fino alla ripresa della replica ed è necessario un processo di riconciliazione per gestire le transazioni in corso che non sono state replicate dalla regione che ha subito l'interruzione. Questo scenario richiede la comprensione della logica aziendale e la creazione di un processo specifico per riprodurre la transazione o confrontare gli archivi di dati tra le regioni.

Per i carichi di lavoro in cui è preferita la replica asincrona, puoi utilizzare servizi come Amazon Aurora e Amazon DynamoDB per la replica asincrona tra regioni. Sia i database globali di Amazon Aurora che le tabelle globali di Amazon DynamoDB dispongono di parametri CloudWatchAmazon predefiniti per facilitare il monitoraggio del ritardo di replica. Un database globale Aurora è composto da una regione principale in cui vengono scritti i dati e da un massimo di cinque regioni secondarie

di sola lettura. Le tabelle globali di DynamoDB sono costituite da tabelle di replica multiattive in un numero qualsiasi di regioni in cui i dati vengono scritti e letti.

Progettare il carico di lavoro per sfruttare le architetture basate sugli eventi è un vantaggio per una strategia multiregionale, perché significa che il carico di lavoro può includere la replica asincrona dei dati e consente la ricostruzione dello stato mediante la riproduzione degli eventi. Poiché i servizi di streaming e messaggistica memorizzano nel buffer i dati del payload dei messaggi in una singola regione, un processo di failover o failback regionale deve includere un meccanismo per reindirizzare i flussi di dati di input dei client. Il processo deve inoltre riconciliare i payload in volo o non consegnati immagazzinati nella regione che ha subito l'interruzione.

Se si sceglie il requisito di coerenza CAP e si utilizza un database replicato in modo sincrono tra le regioni per supportare le applicazioni eseguite contemporaneamente da più regioni, si elimina il rischio di perdita di dati e si mantengono i dati sincronizzati tra le regioni. Tuttavia, ciò introduce caratteristiche di latenza più elevate, poiché le scritture devono essere trasferite in più di una regione e le regioni possono trovarsi a centinaia o migliaia di miglia l'una dall'altra. È necessario tenere conto di questa caratteristica di latenza nella progettazione dell'applicazione. Inoltre, la replica sincrona può comportare la possibilità di errori correlati perché, per avere successo, le scritture dovranno essere eseguite su più di una regione. Se si verifica un problema all'interno di una regione, sarà necessario creare un quorum affinché le scritture abbiano esito positivo. Ciò comporta in genere la configurazione del database in tre regioni e la creazione di un quorum di due regioni su tre. Tecnologie come Paxos possono aiutare a replicare e salvare i dati in modo sincrono, ma richiedono un investimento significativo da parte degli sviluppatori.

Quando le scritture prevedono la replica sincrona su più regioni per soddisfare elevati requisiti di coerenza, la latenza di scrittura aumenta di un ordine di grandezza. Una latenza di scrittura più elevata in genere non è possibile adattarla a un'applicazione senza modifiche significative, come rivisitare la strategia di timeout e riprovare per l'applicazione. Idealmente, deve essere presa in considerazione quando l'applicazione viene progettata per la prima volta. Per i carichi di lavoro multiregionali in cui la replica sincrona è una priorità,AWS Partner le soluzioni possono essere utili.

## 2.b: Comprensione dei modelli di accesso ai dati

I modelli di accesso ai dati dei carichi di lavoro richiedono un'intensa attività di lettura o scrittura. La comprensione di questa caratteristica per un particolare carico di lavoro ti aiuterà a selezionare un'architettura multiregionale appropriata.

Per carichi di lavoro ad alta intensità di lettura, come il contenuto statico completamente di sola lettura, è possibile ottenere un'architettura multiregionale attiva e attiva che presenta una minore

complessità ingegneristica rispetto a un carico di lavoro ad alta intensità di scrittura. La distribuzione di contenuti statici all'edge utilizzando una rete di distribuzione dei contenuti (CDN) garantisce la disponibilità memorizzando nella cache i contenuti più vicini all'utente finale; l'utilizzo di set di funzionalità come il failover di origine all'interno di Amazon CloudFront può contribuire a raggiungere questo obiettivo. Un'altra opzione consiste nell'implementare l'elaborazione stateless in più regioni e utilizzare il DNS per indirizzare gli utenti alla regione più vicina per leggere il contenuto. A tal fine, puoi utilizzare Amazon Route 53 con una politica di routing di geolocalizzazione.

Per carichi di lavoro ad alta intensità di lettura che hanno una percentuale di traffico di lettura maggiore rispetto al traffico di scrittura, puoi utilizzare una strategia globale di lettura locale e scrittura. Ciò significa che tutte le richieste di scrittura vengono inviate a un database in una regione specifica, i dati vengono replicati in modo asincrono in tutte le altre regioni e le letture possono essere eseguite in qualsiasi regione. Questo approccio richiede un carico di lavoro tale da garantire la coerenza finale, poiché le letture locali potrebbero diventare obsolete a causa dell'aumento della latenza per la replica delle scritture tra regioni diverse.

I database globali Aurora possono aiutare a fornire repliche di lettura in una regione di standby in grado di gestire esclusivamente tutto il traffico di lettura a livello locale e fornire un singolo archivio dati primario in una regione specifica per gestire il traffico di scrittura. I dati vengono replicati in modo asincrono dal database primario ai database di standby (repliche di lettura) e i database di standby possono essere promossi a primari se è necessario eseguire il failover delle operazioni nella regione di standby. È inoltre possibile utilizzare DynamoDB in questo approccio. Le tabelle globali DynamoDB possono fornire tabelle di replica su più regioni, ciascuna scalabile per supportare qualsiasi volume di traffico locale di lettura o scrittura. Quando un'applicazione scrive dati in una tabella di replica in una regione, DynamoDB propaga automaticamente la scrittura alle altre tabelle di replica nelle altre regioni . Con questa configurazione, i dati vengono replicati in modo asincrono da una regione primaria definita alle tabelle di replica nelle regioni di standby. Le tabelle di replica in qualsiasi regione possono sempre accettare scritture, quindi la promozione di una regione di standby a principale viene gestita a livello di applicazione. Anche in questo caso, il carico di lavoro deve garantire la coerenza finale, il che potrebbe richiedere la riscrittura se non fosse stato progettato per questo scopo fin dall'inizio.

Per i carichi di lavoro ad alta intensità di scrittura, è necessario selezionare una regione principale e incorporare nel carico di lavoro la capacità di eseguire il failover in una regione di standby. Rispetto a un approccio attivo-attivo, un approccio di standby primario presenta ulteriori compromessi. Questo perché per un'architettura active-active, il carico di lavoro deve essere riscritto per gestire il routing intelligente verso le regioni, stabilire l'affinità delle sessioni, garantire transazioni idempotenti e gestire potenziali conflitti.

La maggior parte dei carichi di lavoro che utilizzano un approccio multiregionale per la resilienza non richiederà un approccio attivo-attivo. È possibile utilizzare una strategia di sharding per fornire una maggiore resilienza limitando l'ambito di impatto di una menomazione sulla base di clienti. Se riesci a condividere efficacemente una base di clienti, puoi selezionare diverse regioni primarie per ogni shard. Ad esempio, puoi condividere i client in modo che metà dei client siano allineati alla Regione uno e l'altra metà sia allineata alla Regione due. Trattando le regioni come celle, è possibile creare un approccio cellulare multiregionale, che si traduce in una riduzione della portata dell'impatto sul carico di lavoro. Per ulteriori informazioni, consultate la presentazione di AWS re:Invent su questo approccio.

È possibile combinare l'approccio di sharding con un approccio di standby primario per fornire funzionalità di failover per gli shard. Dovrai progettare un processo di failover collaudato nel carico di lavoro e anche un processo per la riconciliazione dei dati, per garantire la coerenza transazionale degli archivi di dati dopo il failover. Questi aspetti sono trattati più dettagliatamente più avanti in questa guida.

## Linee guida chiave

- È molto probabile che le scritture in sospeso per la replica non vengano salvate nella regione di standby in caso di errore. I dati non saranno disponibili fino alla ripresa della replica (presupponendo la replica asincrona).
- Come parte del failover, sarà necessario un processo di riconciliazione dei dati per garantire il mantenimento di uno stato transazionale coerente per gli archivi di dati che utilizzano la replica asincrona. Ciò richiede una logica aziendale specifica e non è gestita dall'archivio dati stesso.
- Quando è richiesta una forte coerenza, i carichi di lavoro dovranno essere modificati per tollerare la latenza richiesta di un data store che si replica in modo sincrono.

# Nozioni fondamentali su più regioni 3: Comprendere le dipendenze del carico di lavoro

Un carico di lavoro specifico può avere diverse dipendenze in una regione, ad esempio dipendenze Servizi AWS utilizzate, interne, dipendenze di terze parti, dipendenze di rete, certificati, chiavi, segreti e parametri. Per garantire il funzionamento del carico di lavoro durante uno scenario di errore, non dovrebbero esserci dipendenze tra la regione principale e la regione di standby; ciascuna dovrebbe essere in grado di funzionare indipendentemente dall'altra. A tal fine, esamina tutte le dipendenze del carico di lavoro per assicurarti che siano disponibili all'interno di ciascuna regione. Ciò è necessario perché un errore nella regione principale non dovrebbe influire sulla regione di standby. Inoltre, è necessario comprendere come funziona il carico di lavoro quando una dipendenza si trova in uno stato degradato o completamente non disponibile, in modo da poter progettare soluzioni per gestirla in modo appropriato.

## 3.a: Servizi AWS

Quando si progetta un'architettura multiregionale, è importante comprendere cosa verrà utilizzata, le funzionalità multiregionali di tali servizi e quali soluzioni sarà necessario progettare per raggiungere gli obiettivi multiregionali. Servizi AWS Ad esempio, Amazon Aurora e Amazon DynamoDB possono replicare in modo asincrono i dati in una regione di standby. Tutte le Servizio AWS dipendenze dovranno essere disponibili in tutte le regioni da cui verrà eseguito un carico di lavoro. Per confermare che i servizi che utilizzi sono disponibili nelle regioni desiderate, consulta l'elenco Servizi AWS per regione.

## 3.b: Dipendenze interne e di terze parti

Assicurati che le dipendenze interne di ogni carico di lavoro siano disponibili nelle regioni da cui operano. Ad esempio, se il carico di lavoro è composto da molti microservizi, identifica tutti i microservizi che comprendono una funzionalità aziendale e verifica che tutti quei microservizi siano distribuiti in ogni regione da cui opera il carico di lavoro. In alternativa, definisci una strategia per gestire in modo corretto i microservizi che diventano non disponibili.

Le chiamate interregionali tra microservizi all'interno di un carico di lavoro non sono consigliate e l'isolamento regionale deve essere mantenuto. Questo perché la creazione di dipendenze tra regioni aumenta il rischio di errori correlati, il che compensa i vantaggi delle implementazioni regionali isolate

del carico di lavoro. Anche le dipendenze locali potrebbero far parte del carico di lavoro, quindi è importante capire come le caratteristiche di queste integrazioni potrebbero cambiare se la regione principale dovesse cambiare. Ad esempio, se la regione di standby si trova più lontana dall'ambiente locale, l'aumento della latenza potrebbe avere un impatto negativo.

Comprendere le soluzioni SaaS (Software as a Service), i kit di sviluppo software (SDKs) e altre dipendenze da prodotti di terze parti e la possibilità di utilizzare scenari in cui tali dipendenze sono degradate o non disponibili forniranno maggiori informazioni su come la catena di sistemi opera e si comporta in diverse modalità di errore. Queste dipendenze potrebbero risiedere all'interno del codice dell'applicazione, ad esempio gestire i segreti esternamente tramite l'utilizzo AWS Secrets Manager, oppure potrebbero coinvolgere una soluzione di vault di terze parti (ad esempio HashiCorp) o sistemi di autenticazione che dipendono dagli accessi federati. AWS IAM Identity Center

La ridondanza quando si tratta di dipendenze può aumentare la resilienza. Se una soluzione SaaS o una dipendenza da terze parti utilizza lo stesso carico di lavoro primario Regione AWS , collabora con il fornitore per determinare se la sua posizione di resilienza corrisponde ai tuoi requisiti per il carico di lavoro.

Inoltre, tieni presente il destino condiviso tra il carico di lavoro e le sue dipendenze, come le applicazioni di terze parti. Se le dipendenze non sono disponibili in (o da) una regione secondaria dopo un failover, il carico di lavoro potrebbe non essere ripristinato completamente.

## 3.c: meccanismo di failover

Il DNS è comunemente usato come meccanismo di failover per spostare il traffico dalla regione principale a una regione di standby. Esamina e analizza in modo critico tutte le dipendenze assunte dal meccanismo di failover. Ad esempio, se il tuo carico di lavoro utilizza Amazon Route 53, sapere che il piano di controllo è ospitato in `us-east-1` significa dipendere dal piano di controllo in quella regione specifica. Questa operazione non è consigliata come parte di un meccanismo di failover se la regione primaria lo è anche `us-east-1` perché crea un singolo punto di errore. Se si utilizza un altro meccanismo di failover, è necessario avere una conoscenza approfondita degli scenari in cui il failover non funzionerebbe come previsto e quindi pianificare eventuali situazioni di emergenza o sviluppare un nuovo meccanismo, se necessario. Leggi il post sul blog Creazione di meccanismi di disaster recovery con Amazon Route 53 per scoprire gli approcci che puoi utilizzare per eseguire correttamente il failover.

Come discusso nella sezione precedente, tutti i microservizi che fanno parte di una funzionalità aziendale devono essere disponibili in ogni regione in cui viene distribuito il carico di lavoro.

Nell'ambito della strategia di failover, tutti i microservizi che fanno parte della funzionalità aziendale devono eseguire il failover contemporaneamente per eliminare la possibilità di chiamate tra regioni. In alternativa, se i microservizi eseguono il failover in modo indipendente, è possibile che si verifichino comportamenti indesiderati, ad esempio che i microservizi effettuino chiamate tra regioni diverse. Ciò introduce la latenza e potrebbe rendere il carico di lavoro non disponibile durante i timeout del client.

## 3.d: dipendenze dalla configurazione

Certificati, chiavi, segreti, Amazon Machine Images (AMIs), immagini dei container e parametri fanno parte dell'analisi delle dipendenze necessaria per progettare un'architettura multiregionale. Quando possibile, è meglio localizzare questi componenti all'interno di ciascuna regione in modo che non abbiano un destino condiviso tra le regioni per queste dipendenze. Ad esempio, è necessario variare le date di scadenza dei certificati per evitare che si verifichi uno scenario in cui un certificato in scadenza (con allarmi impostati su «notifica in anticipo») influisca su più regioni.

Anche le chiavi e i segreti di crittografia devono essere specifici della regione. In questo modo, se si verifica un errore nella rotazione di una chiave o di un segreto, l'impatto è limitato a una regione specifica.

Infine, tutti i parametri del carico di lavoro devono essere archiviati localmente affinché il carico di lavoro possa essere recuperato nella regione specifica.

## Linee guida chiave

- Un'architettura multiregionale trae vantaggio dalla separazione fisica e logica tra le regioni. L'introduzione di dipendenze interregionali a livello di applicazione annulla questo vantaggio. Evita tali dipendenze.
- I controlli di failover dovrebbero funzionare senza dipendenze dalla regione principale.
- Il failover deve essere coordinato lungo tutto il percorso dell'utente per eliminare la possibilità di un aumento della latenza e della dipendenza delle chiamate interregionali.

# Principi fondamentali per più regioni 4: prontezza operativa

La gestione di un carico di lavoro multiregionale è un'attività complessa che comporta sfide operative specifiche di un'architettura multiregionale. Queste includono la Account AWS gestione, la riorganizzazione dei processi di implementazione, la creazione di una strategia di osservabilità multiregionale, la creazione e il test dei processi di ripristino e quindi la gestione dei costi. Un Operational Readiness Review (ORR) può aiutare i team a preparare un carico di lavoro per la produzione, indipendentemente dal fatto che venga eseguito in una singola regione o in più regioni.

## 4.a: gestione Account AWS

Per distribuire un carico di lavoro in tutte le regioni Regioni AWS, assicurati che vi sia parità tra tutte le Servizio AWS quote all'interno di un account. Innanzitutto, identifica tutti gli Servizi AWS elementi che fanno parte dell'architettura, esamina l'utilizzo pianificato nelle regioni di standby e quindi confronta l'utilizzo pianificato con l'utilizzo corrente. In alcuni casi, se la regione di standby non è mai stata utilizzata in precedenza, puoi fare riferimento alle quote di servizio predefinite per comprendere il punto di partenza. Quindi, per tutti i servizi che verranno utilizzati, richiedi un aumento della quota utilizzando la console Service Quotas (è richiesto il login) oppure. APIs

Configura i ruoli AWS Identity and Access Management (IAM) in ciascuna regione per fornire agli operatori, agli strumenti di automazione e Servizi AWS le autorizzazioni appropriate alle risorse all'interno della regione di standby. Per ottenere l'isolamento regionale per le architetture multiregionali, isola i ruoli per regione. Assicurati che le autorizzazioni siano disponibili prima di passare alla modalità live con una regione in standby.

## 4.b: Pratiche di implementazione

Le funzionalità multiregionali possono complicare la distribuzione di un carico di lavoro in più regioni. È necessario assicurarsi di eseguire la distribuzione in una regione alla volta. Ad esempio, se si utilizza un approccio attivo-passivo, è necessario eseguire la distribuzione prima nella regione principale e poi nella regione di standby. AWS CloudFormationti aiuta a distribuire l'infrastruttura in una o più regioni e può essere personalizzato in base alle tue esigenze. AWS CodePipelineti aiuta a creare una pipeline integration/continuous delivery (CI/CD (continua), che prevede azioni interregionali che consentono la distribuzione in regioni diverse dalla regione in cui si trova la pipeline. Questo, combinato con solide strategie di implementazione come blu/green, consente un'implementazione con tempi di inattività minimi o pari a zero.

Tuttavia, l'implementazione delle funzionalità stateful può diventare più complessa quando lo stato dell'applicazione o dei dati non viene esternalizzato in un archivio persistente. In queste situazioni, personalizza attentamente il processo di implementazione in base alle tue esigenze. Progetta la pipeline e il processo di distribuzione in modo da distribuirli in una regione alla volta anziché distribuirli in più regioni contemporaneamente. Ciò riduce la possibilità di guasti correlati tra le regioni. Per conoscere le tecniche utilizzate da Amazon per automatizzare le distribuzioni di software, consulta l'articolo della AWS Builders' Library Automating safe hands-off deployments.

## 4.c: Osservabilità

Quando progetti per più regioni, considera in che modo monitorerai lo stato di tutti i componenti di ciascuna regione per ottenere una visione olistica dello stato di salute regionale. Ciò potrebbe includere il monitoraggio delle metriche per il ritardo di replica, che non è una considerazione per un carico di lavoro a regione singola.

Quando crei un'architettura multiregionale, prendi in considerazione l'osservazione delle prestazioni del carico di lavoro anche nelle regioni di standby. Ciò include il controllo dello stato di salute e l'esecuzione di canarini (test sintetici) dalla regione di standby per fornire una visione esterna dello stato di salute della regione primaria. Inoltre, puoi utilizzare Amazon CloudWatch Internet Monitor per comprendere lo stato della rete esterna e le prestazioni dei tuoi carichi di lavoro dal punto di vista dell'utente finale. La regione principale deve avere la stessa osservabilità per monitorare la regione di standby.

I canarini della regione di standby dovrebbero monitorare le metriche relative all'esperienza del cliente per determinare lo stato generale del carico di lavoro. Ciò è necessario perché se si verifica un problema nella regione principale, l'osservabilità nella regione primaria potrebbe essere compromessa e influirebbe sulla capacità di valutare lo stato del carico di lavoro.

In tal caso, osservare al di fuori di quella regione può fornire informazioni. Queste metriche devono essere raggruppate in dashboard disponibili in ogni regione e negli allarmi creati in ciascuna regione. Poiché CloudWatchsi tratta di un servizio regionale, è necessario disporre di allarmi in entrambe le regioni. Questi dati di monitoraggio verranno utilizzati per effettuare il failover della chiamata da una regione principale a una di standby.

## 4.d: Processi e procedure

Il momento migliore per rispondere alla domanda «Quando devo effettuare il failover?» è molto prima che sia necessario. Definisci piani di ripristino che includano persone, processi e tecnologia con largo

anticipo rispetto al problema e testali regolarmente. Decidi un quadro decisionale di recupero. Se esiste un processo di ripristino ben collaudato e i tempi necessari per il ripristino sono ben compresi, è possibile scegliere di avviare il processo di ripristino utilizzando un failover che soddisfi l'obiettivo RTO. Questo momento può avvenire immediatamente dopo l'identificazione di un problema con l'applicazione nella regione principale, oppure potrebbe avvenire in un momento successivo in cui le opzioni di ripristino all'interno dell'applicazione nella regione sono state esaurite.

L'azione di failover stessa dovrebbe essere automatizzata al 100%, ma la decisione di attivare il failover dovrebbe essere presa dagli esseri umani, in genere da un piccolo numero di individui predeterminati all'interno dell'organizzazione. Queste persone dovrebbero prendere in considerazione la perdita di dati e le informazioni sull'evento. Inoltre, i criteri per un failover devono essere chiaramente definiti e compresi a livello globale all'interno dell'organizzazione. Per definire e completare questi processi, è possibile utilizzare i AWS Systems Manager runbook, che consentono l' end-to-endautomazione completa e garantiscono la coerenza dei processi in esecuzione durante i test e il failover.

Questi runbook devono essere disponibili nelle regioni primarie e in standby per avviare i processi di failover o failback. Quando questa automazione è attiva, definisci e segui una cadenza di test regolare. Ciò garantisce che, quando si verifica un evento reale, la risposta segua un processo ben definito e pratico in cui l'organizzazione ha fiducia. È anche importante considerare le tolleranze stabilite per i processi di riconciliazione dei dati. Verifica che il processo proposto soddisfi i requisiti RPO/RTO stabiliti.

## 4.e: Test

Avere un approccio di ripristino non testato equivale a non avere un approccio di ripristino. Un livello base di test consisterebbe nell'eseguire una procedura di ripristino per cambiare la regione operativa dell'applicazione. A volte si parla di approccio di rotazione delle applicazioni. Si consiglia di sviluppare la capacità di passare da una regione all'altra mantenendo la normale postura operativa; tuttavia, questo test da solo non è sufficiente.

Il test di resilienza è fondamentale anche per convalidare l'approccio di ripristino di un'applicazione. Ciò comporta l'introduzione di particolari scenari di errore, la comprensione di come reagiscono l'applicazione e il processo di ripristino e quindi l'implementazione delle eventuali mitigazioni necessarie se il test non è andato come previsto. Il test della procedura di ripristino in assenza di errori non vi dirà come si comporta l'intera applicazione in caso di guasti. È necessario sviluppare un piano per testare il ripristino rispetto agli scenari di errore previsti. AWS Fault Injection Servicefornisce un elenco crescente di scenari per iniziare.

Ciò è particolarmente importante per le applicazioni ad alta disponibilità, in cui sono necessari test rigorosi per garantire il raggiungimento degli obiettivi di continuità aziendale. Il test proattivo delle funzionalità di ripristino riduce il rischio di guasti nella produzione, il che aumenta la fiducia che l'applicazione possa raggiungere il tempo di ripristino limitato desiderato. I test regolari rafforzano anche le competenze operative, che consentono al team di riprendersi in modo rapido e affidabile dalle interruzioni quando si verificano. Esercitare l'elemento umano, o processo, dell'approccio di ripristino è fondamentale tanto quanto gli aspetti tecnici.

## 4.f: Costo e complessità

Le implicazioni in termini di costi di un'architettura multiregionale sono determinate da un maggiore utilizzo dell'infrastruttura, dal sovraccarico operativo e dal tempo impiegato per le risorse. Come accennato in precedenza, il costo dell'infrastruttura in una regione di standby è simile al costo dell'infrastruttura in una regione principale durante il pre-provisioning, quindi raddoppia il costo totale. Fornisci capacità in modo che sia sufficiente per le operazioni quotidiane, riservando comunque una capacità di buffer sufficiente per tollerare i picchi di domanda. Quindi configura gli stessi limiti in ogni regione.

Inoltre, se state adottando un'architettura active-active, potrebbe essere necessario apportare modifiche a livello di applicazione per eseguire correttamente l'applicazione in un'architettura multiregionale. Queste modifiche possono richiedere molto tempo e risorse per la progettazione e il funzionamento. Come minimo, le organizzazioni devono dedicare del tempo alla comprensione delle dipendenze tecniche e commerciali in ciascuna regione e alla progettazione di processi di failover e failback.

I team devono inoltre eseguire i normali esercizi di failover e failback per sentirsi a proprio agio con i runbook da utilizzare durante un evento. Sebbene questi esercizi siano fondamentali per ottenere i risultati attesi da un investimento in più regioni, rappresentano un costo-opportunità e sottraggono tempo e risorse ad altre attività.

## 4.g: Strategia di failover organizzativa multiregionale

Regioni AWS forniscono limiti di isolamento dei guasti che impediscano guasti correlati e contengano l'impatto dei Servizio AWS danni, quando si verificano, su una singola regione. È possibile utilizzare questi limiti di errore per creare applicazioni multiregionali costituite da repliche indipendenti e isolate dai guasti in ciascuna regione per limitare gli scenari di destino condiviso. Ciò consente di creare applicazioni multiregionali e utilizzare una gamma di approcci, dal backup e ripristino, alla versione pilota, alla modalità active-active, per implementare l'architettura multiregionale. Tuttavia,

le applicazioni in genere non funzionano in modo isolato, quindi considera sia i componenti che utilizzerai sia le loro dipendenze come parte della tua strategia di failover. In genere, più applicazioni collaborano per supportare una storia utente, ovvero una funzionalità specifica offerta a un utente finale, ad esempio la pubblicazione di una foto e di una didascalia su un'app di social media o il check-out su un sito di e-commerce. Per questo motivo, è necessario sviluppare una strategia di failover organizzativa multiregionale che fornisca il coordinamento e la coerenza necessari per il successo del proprio approccio.

Esistono quattro strategie di alto livello tra cui le organizzazioni possono scegliere per guidare un approccio multiregionale. Queste sono elencate dall'approccio più granulare a quello più ampio:

- Failover a livello di componente
- Failover delle singole applicazioni
- Failover del grafico delle dipendenze
- Failover dell'intero portafoglio di applicazioni

Ogni strategia presenta dei compromessi e affronta diverse sfide, tra cui la flessibilità del processo decisionale in materia di failover, la capacità di testare le combinazioni di failover, la presenza di comportamenti modali e gli investimenti organizzativi nella pianificazione e nell'implementazione. Per approfondire ogni strategia in modo più dettagliato, consulta il post AWS sul blog Creazione di una strategia di failover organizzativa multiregionale.

## Linee guida chiave

- Rivedi tutte le Servizio AWS quote per assicurarti che siano uguali in tutte le regioni in cui verrà eseguito il carico di lavoro.
- Il processo di implementazione dovrebbe riguardare una regione alla volta anziché coinvolgere più regioni contemporaneamente.
- Le metriche aggiuntive, come il ritardo di replica, sono specifiche degli scenari multiregionali e devono essere monitorate.
- Estendi il monitoraggio del carico di lavoro oltre la regione principale. Monitora le metriche relative all'esperienza del cliente per ogni regione e misura questi dati dall'esterno di ciascuna regione in cui è in esecuzione un carico di lavoro.
- Verifica regolarmente il failover e il failback. Implementa un singolo runbook per i processi di failover e failback e utilizzalo sia per i test che per gli eventi live. I runbook per i test e gli eventi live non dovrebbero essere diversi.

- Comprendi i compromessi delle strategie di failover. Implementa un grafico delle dipendenze o una strategia per l'intero portafoglio di applicazioni.

# Conclusioni e risorse

Questa guida illustra i casi d'uso più comuni per le architetture multiregionali, i fondamenti dell'implementazione di tali architetture e le implicazioni di questo approccio. Puoi applicare questi principi fondamentali a qualsiasi carico di lavoro e utilizzare le informazioni come framework per decidere se un'architettura multiregionale è l'approccio giusto per la tua azienda.

Per ulteriori informazioni, consulta le seguenti risorse:

- AWS Centro di architettura
- AWS Well-Architected Framework
- AWS Well-Architected Tool
- Creazione di una strategia di failover organizzativa multiregionale (post sul blog)AWS
- AWS Funzionalità multiregionali (articolo Re:Post)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 |
| --- | --- | --- |
| Aggiornamenti | Aggiornamenti in tutta la guida. | 27 dicembre 2024 |
| Pubblicazione iniziale | — | 20 dicembre 2022 |

# AWS Glossario delle linee guida prescrittive

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

## Numeri

7 R

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

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

# A

ABAC

Vedi controllo degli accessi basato sugli attributi.

servizi astratti

Vedi servizi gestiti.

ACIDO

Vedi atomicità, consistenza, isolamento, durata.

migrazione attiva-attiva

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

migrazione attiva-passiva

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

funzione di aggregazione

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

Intelligenza artificiale

Vedi intelligenza artificiale.

AIOps

Guarda le operazioni di intelligenza artificiale.

anonimizzazione

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

anti-modello

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

controllo delle applicazioni

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

portfolio di applicazioni

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

intelligenza artificiale (IA)

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

operazioni di intelligenza artificiale (AIOps)

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

crittografia asimmetrica

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

atomicità, consistenza, isolamento, durabilità (ACID)

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

Controllo degli accessi basato su attributi (ABAC)

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

fonte di dati autorevole

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

Zona di disponibilità

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

AWS Cloud Adoption Framework (CAF)AWS

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

AWS Workload Qualification Framework (WQF)AWS

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

# B

bot difettoso

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

BCP

Vedi la pianificazione della continuità operativa.

grafico comportamentale

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

sistema big-endian

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

Classificazione binaria

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

filtro Bloom

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

distribuzioni blu/verdi

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

bot

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

botnet

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

ramo

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

accesso break-glass

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

strategia brownfield

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

cache del buffer

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

capacità di business

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

pianificazione della continuità operativa (BCP)

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

# C

CAF

Vedi AWS Cloud Adoption Framework.

implementazione canaria

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

CCoE

Vedi Cloud Center of Excellence.

CDC

Vedi Change Data Capture.

Change Data Capture (CDC)

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

ingegneria del caos

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

CI/CD

Vedi integrazione continua e distribuzione continua.

classificazione

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

crittografia lato client

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

Centro di eccellenza cloud (CCoE)

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

cloud computing

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

modello operativo cloud

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

fasi di adozione del cloud

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

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

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

CMDB

Vedi database di gestione della configurazione.

repository di codice

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

cache fredda

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

dati freddi

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

visione artificiale (CV)

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

deriva della configurazione

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

database di gestione della configurazione (CMDB)

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

Pacchetto di conformità

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

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

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

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

CV

Vedi visione artificiale.

# D

dati a riposo

Dati stazionari nella rete, ad esempio i dati archiviati.

classificazione dei dati

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

deriva dei dati

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

dati in transito

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

rete di dati

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

riduzione al minimo dei dati

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

perimetro dei dati

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

pre-elaborazione dei dati

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

provenienza dei dati

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

soggetto dei dati

Un individuo i cui dati vengono raccolti ed elaborati.

data warehouse

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

linguaggio di definizione del database (DDL)

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

linguaggio di manipolazione del database (DML)

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

DDL

Vedi linguaggio di definizione del database.

deep ensemble

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

deep learning

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

defense-in-depth

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

amministratore delegato

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

implementazione

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

Ambiente di sviluppo

Vedi ambiente.

controllo di rilevamento

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

mappatura del flusso di valore dello sviluppo (DVSM)

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

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

gemello digitale

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

tabella delle dimensioni

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

disastro

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

disaster recovery (DR)

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

DML

Vedi linguaggio di manipolazione del database.

progettazione basata sul dominio

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

DOTT.

Vedi disaster recovery.

rilevamento della deriva

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

DVSM

Vedi la mappatura del flusso di valore dello sviluppo.

# E

EDA

Vedi analisi esplorativa dei dati.

MODIFICA

Vedi scambio elettronico di dati.

edge computing

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

scambio elettronico di dati (EDI)

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

crittografia

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

chiave crittografica

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

endianità

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

endpoint

Vedi service endpoint.

servizio endpoint

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

pianificazione delle risorse aziendali (ERP)

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

crittografia envelope

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

ambiente

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

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

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

epica

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

ERP

Vedi pianificazione delle risorse aziendali.

analisi esplorativa dei dati (EDA)

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

# F

tabella dei fatti

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

fallire velocemente

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

limite di isolamento dei guasti

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

ramo di funzionalità

Vedi filiale.

caratteristiche

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

importanza delle caratteristiche

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

trasformazione delle funzionalità

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

prompt con pochi scatti

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

FGAC

Vedi il controllo granulare degli accessi.

controllo granulare degli accessi (FGAC)

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

migrazione flash-cut

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

FM

Vedi modello di base.

modello di fondazione (FM)

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

# G

AI generativa

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

blocco geografico

Vedi restrizioni geografiche.

limitazioni geografiche (blocco geografico)

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

Flusso di lavoro di GitFlow

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

immagine dorata

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

strategia greenfield

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

guardrail

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

# H

AH

Vedi disponibilità elevata.

migrazione di database eterogenea

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

alta disponibilità (HA)

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

modernizzazione storica

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

dati di esclusione

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

migrazione di database omogenea

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

dati caldi

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

hotfix

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

periodo di hypercare

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

## I

IaC

Considera l'infrastruttura come codice.

Policy basata su identità

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

applicazione inattiva

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

IIoT

Vedi Industrial Internet of Things.

infrastruttura immutabile

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

VPC in ingresso (ingress)

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

migrazione incrementale

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

Industria 4.0

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

infrastruttura

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

infrastruttura come codice (IaC)

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

IIoInternet delle cose industriale (T)

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

VPC di ispezione

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

Internet of Things (IoT)

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

interpretabilità

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

IoT

Vedi Internet of Things.

libreria di informazioni IT (ITIL)

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

gestione dei servizi IT (ITSM)

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

ITIL

Vedi la libreria di informazioni IT.

ITSM

Vedi Gestione dei servizi IT.

# L

controllo degli accessi basato su etichette (LBAC)

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

zona di destinazione

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

modello linguistico di grandi dimensioni (LLM)

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

migrazione su larga scala

Una migrazione di 300 o più server.

BIANCO

Vedi controllo degli accessi basato su etichette.

Privilegio minimo

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

eseguire il rehosting (lift and shift)

Vedi 7 R.

sistema little-endian

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

LLM

Vedi modello linguistico di grandi dimensioni.

ambienti inferiori

Vedi ambiente.

# M

machine learning (ML)

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

ramo principale

Vedi filiale.

malware

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

servizi gestiti

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

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

sistema di esecuzione della produzione (MES)

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

MAP

Vedi Migration Acceleration Program.

meccanismo

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

account membro

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

MEH

Vedi sistema di esecuzione della produzione.

Message Queuing Telemetry Transport (MQTT)

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

microservizio

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

architettura di microservizi

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

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

Programma di accelerazione della migrazione (MAP)

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

migrazione su larga scala

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

fabbrica di migrazione

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

metadati di migrazione

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

modello di migrazione

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

Valutazione del portfolio di migrazione (MPA)

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

valutazione della preparazione alla migrazione (MRA)

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

strategia di migrazione

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

ML

Vedi machine learning.

modernizzazione

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

valutazione della preparazione alla modernizzazione

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

applicazioni monolitiche (monoliti)

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

MAPPA

Vedi Migration Portfolio Assessment.

MQTT

Vedi Message Queuing Telemetry Transport.

classificazione multiclasse

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

infrastruttura mutabile

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

# O

OAC

Vedi Origin Access Control.

QUERCIA

Vedi Origin Access Identity.

OCM

Vedi gestione delle modifiche organizzative.

migrazione offline

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

OI

Vedi l'integrazione delle operazioni.

OLA

Vedi accordo a livello operativo.

migrazione online

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

OPC-UA

Vedi Open Process Communications - Unified Architecture.

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

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

accordo a livello operativo (OLA)

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

revisione della prontezza operativa (ORR)

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

tecnologia operativa (OT)

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

integrazione delle operazioni (OI)

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

trail organizzativo

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

gestione del cambiamento organizzativo (OCM)

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

controllo dell'accesso all'origine (OAC)

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

identità di accesso origine (OAI)

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

ORR

Vedi la revisione della prontezza operativa.

- NON

Vedi la tecnologia operativa.

VPC in uscita (egress)

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

# P

limite delle autorizzazioni

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

informazioni di identificazione personale (PII)

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

Informazioni che consentono l'identificazione personale degli utenti

Visualizza le informazioni di identificazione personale.

playbook

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

PLC

Vedi controllore logico programmabile.

PLM

Vedi la gestione del ciclo di vita del prodotto.

policy

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

persistenza poliglotta

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

valutazione del portfolio

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

predicate

Una condizione di interrogazione che restituisce o, in genere, si trova in una clausola `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 AWSPer ulteriori informazioni, vedere Cloud AWS Resilience.

policy basata su risorse

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

matrice di assegnazione di responsabilità (RACI)

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

controllo reattivo

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

retain

Vedi 7 R.

andare in pensione

Vedi 7 Rs.

Retrieval Augmented Generation (RAG)

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

rotazione

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

controllo dell'accesso a righe e colonne (RCAC)

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

RPO

Vedi l'obiettivo del punto di ripristino.

RTO

Vedi l'obiettivo del tempo di ripristino.

runbook

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

# S

SAML 2.0

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

SCADA

Vedi controllo di supervisione e acquisizione dati.

SCP

Vedi la politica di controllo del servizio.

Secret

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

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

sicurezza fin dalla progettazione

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

controllo di sicurezza

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

rafforzamento della sicurezza

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

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

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

automazione della risposta alla sicurezza

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

Crittografia lato server

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

Policy di controllo dei servizi (SCP)

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

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

endpoint del servizio

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

accordo sul livello di servizio (SLA)

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

indicatore del livello di servizio (SLI)

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

obiettivo a livello di servizio (SLO)

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

Modello di responsabilità condivisa

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

SIEM

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

punto di errore singolo (SPOF)

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

SLAM

Vedi il contratto sul livello di servizio.

SLI

Vedi l'indicatore del livello di servizio.

LENTA

Vedi obiettivo del livello di servizio.

split-and-seed modello

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

SPOF

Vedi punto di errore singolo.

schema a stella

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

modello del fico strangolatore

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

sottorete

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

controllo di supervisione e acquisizione dati (SCADA)

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

crittografia simmetrica

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

test sintetici

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

prompt di sistema

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

# T

tags

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

variabile di destinazione

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

elenco di attività

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

Ambiente di test

Vedi ambiente.

training

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

Transit Gateway

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

flusso di lavoro basato su trunk

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

Accesso attendibile

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

regolazione

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

team da due pizze

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

# U

incertezza

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

compiti indifferenziati

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

ambienti superiori

Vedi ambiente.

# V

vacuum

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

controllo delle versioni

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

Peering VPC

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

vulnerabilità

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

# W

cache calda

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

dati caldi

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

funzione finestra

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

Carico di lavoro

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

flusso di lavoro

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

VERME

Vedi scrivere una volta, leggere molti.

WQF

Vedi AWS Workload Qualification Framework.

scrivi una volta, leggi molte (WORM)

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

# Z

exploit zero-day

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

vulnerabilità zero-day

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

prompt zero-shot

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

applicazione zombie

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

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