Applicazione del AWS Well-Architected Framework per Amazon Neptune

AWS Guida prescrittiva

AWS Guida prescrittiva: Applicazione del AWS Well-Architected Framework per Amazon Neptune

Table of Contents

Applicazione del AWS Well-Architected Framework per Amazon Neptune

Amazon Web Services (collaboratori)

Gennaio 2026 (cronologia del documento)

Puoi creare soluzioni basate su grafici su Amazon Web Services (AWS) utilizzando Amazon Neptune. Questa guida fornisce linee guida prescrittive per applicare i principi di AWS Well-Architected Framework quando pianifichi la distribuzione di Neptune.

AWS Well-Architected Framework ti aiuta a creare infrastrutture sicure, ad alte prestazioni, resilienti ed efficienti per una varietà di applicazioni e carichi di lavoro. Fornisce inoltre un approccio coerente per valutare le architetture e implementare progetti scalabili.

Il AWS Well-Architected Framework si basa sui seguenti sei pilastri:

- Eccellenza operativa
- Sicurezza
- Affidabilità
- Efficienza delle prestazioni
- Ottimizzazione dei costi
- Sostenibilità

Questa guida fornisce informazioni sui pilastri di progettazione e sulle migliori pratiche di progettazione di AWS Well-Architected Framework e considerazioni da tenere a mente quando si implementa Neptune su. AWS

Destinatari principali

Questa guida è destinata agli ingegneri dei dati, agli architetti di soluzioni e agli analisti di dati che progettano e implementano soluzioni che utilizzano grafici. AWS

Obiettivi

Questa guida può aiutare te e la tua organizzazione a fare quanto segue:

- Scegliete tra le opzioni di distribuzione e i linguaggi di query supportati, in base al vostro caso d'uso e ai modelli di query.
- Segui i modelli di progettazione AWS Well-Architected che ti aiuteranno a migliorare la resilienza e la sicurezza.
- Progetta le tue query per prestazioni ottimali e risparmi sui costi.
- Scopri come essere efficiente dal punto di vista operativo nella gestione del cluster Neptune in produzione.

Il pilastro dell'eccellenza operativa

Il pilastro dell'eccellenza operativa del AWS Well-Architected Framework si concentra sull'esecuzione e il monitoraggio dei sistemi e sul miglioramento continuo di processi e procedure. Include la capacità di supportare lo sviluppo ed eseguire i carichi di lavoro in modo efficace, ottenere informazioni dettagliate sul loro funzionamento e migliorare continuamente i processi e le procedure di supporto per offrire valore aziendale. È possibile ridurre la complessità operativa attraverso carichi di lavoro con riparazione automatica, che rilevano e risolvono la maggior parte dei problemi senza l'intervento umano. È possibile raggiungere questo obiettivo seguendo le best practice descritte in questa sezione. Utilizza i parametri APIs e i meccanismi di Amazon Neptune per rispondere correttamente quando il carico di lavoro si discosta dal comportamento previsto.

Questa discussione sul pilastro dell'eccellenza operativa si concentra sulle seguenti aree chiave:

- Infrastructure as code (IaC)
- Gestione delle modifiche
- Strategie di resilienza
- Gestione degli incidenti
- Reportistica di audit per la conformità
- Registrazione di log e monitoraggio

Automatizza l'implementazione utilizzando un approccio IaC

Le migliori pratiche per automatizzare l'implementazione su Neptune utilizzando IAc includono quanto segue:

- Applica l'infrastruttura come codice (IaC) per distribuire i cluster Neptune ogni volta che è possibile. Per una configurazione coerente dell'ambiente, usa un AWS CloudFormationmodello o HashiCorp Terraform per creare tutte le risorse necessarie per il tuo cluster. AWS Cloud Development Kit (AWS CDK)
- Automatizza le procedure operative di Neptune, come il ridimensionamento delle istanze, l'aggiunta o la rimozione di repliche di lettura o l'esecuzione di failover manuali su tabelle globali, quando possibile.
- Archivia le stringhe di connessione esternamente al tuo client. Utilizza i processi di estrazione, trasformazione e caricamento (ETL) per facilitare le strategie di blue/green implementazione,

il disaster recovery (DR) e le migrazioni con tempi di inattività prossimi allo zero verso nuovi cluster. Le stringhe di connessione possono essere archiviate in Gestione dei segreti AWSAmazon DynamoDB o in qualsiasi posizione in cui possono essere modificate dinamicamente.

- Usa i tag per aggiungere metadati alle tue risorse Neptune e monitora l'utilizzo in base ai tag. Per ulteriori informazioni, consulta Tagging Amazon Neptune Resources.

Apporta modifiche frequenti, piccole e reversibili

Le seguenti raccomandazioni si concentrano su modifiche piccole e reversibili per ridurre al minimo la complessità e ridurre la probabilità di interruzione del carico di lavoro:

- Archivia modelli e script IAc in un servizio di controllo del codice sorgente, ad esempio o. GitHub GitLab

> **Important**
>
> Non memorizzate AWS le credenziali nel controllo del codice sorgente.

- Richiedi che le implementazioni IaC utilizzino un servizio di integrazione e distribuzione continua (CI/CD), come o. AWS CodePipelineAWS CodeBuild Questi servizi compilano, testano e distribuiscono codice in un ambiente non di produzione contenente un cluster Neptune temporaneo prima di influire sul cluster Amazon Neptune di produzione.
- Testa le query sull'infrastruttura e sulle applicazioni in un ambiente inferiore prima di distribuirle in produzione. Ciò ridurrà al minimo la probabilità di interruzioni e contribuirà a garantire che funzionino bene con il carico di lavoro e la scalabilità.

Anticipa il fallimento

Un'infrastruttura con riparazione automatica esemplifica l'eccellenza operativa anticipando i guasti e tentando di risolvere eventuali problemi senza intervento. I seguenti consigli ti aiutano a raggiungere tale maturità con Neptune:

- Crea un piano di monitoraggio che utilizzi i CloudWatch parametri di Amazon per monitorare l'utilizzo della CPU e della memoria dell'istanza DB e comprendere i modelli di utilizzo. Crea CloudWatch dashboard e allarmi per le metriche chiave e le risposte del client Neptune presenti nei

registri delle applicazioni. Per ulteriori informazioni sugli indicatori di utilizzo elevato o basso della CPU, vedere Utilizzo per CloudWatch monitorare le prestazioni delle istanze DB in Neptune nella documentazione di Neptune.

Se riscontri spesso out-of-memory eccezioni nelle tue query, valuta la possibilità di ridurre il numero totale di nodi attraversati dalla query o prova a utilizzare un'istanza della famiglia, che ha un rapporto più elevato. X2 RAM-to-CPU

- Imposta notifiche per monitorare lo stato del cluster Neptune. Ad esempio, `BufferCacheHitRatio` dovrebbe essere costantemente alto (superiore al 99,9 percento), mentre `MainRequestQueuePendingRequests` dovrebbe essere costantemente basso (idealmente 0 ma dipende dai requisiti e dalla tolleranza di latenza).
- Prendi in considerazione l'utilizzo di repliche di lettura per ottenere un'elevata disponibilità all'interno di Neptune. È necessario disporre di almeno due repliche di lettura in zone di disponibilità diverse dall'istanza writer per garantire che un'istanza sia sempre disponibile per fornire query di lettura durante un evento di failover.
- Ridimensiona automaticamente le repliche di lettura in base alle metriche di utilizzo. Per ulteriori informazioni, consulta Ridimensionamento automatico del numero di repliche in un cluster Amazon Neptune DB.
- Prova il failover per l'istanza database per capire quanto tempo impiega il processo per il tuo caso d'uso.
- Se la tua applicazione richiede di resistere a un' Regione AWS interruzione completa, prendi in considerazione l'utilizzo di database globali come parte dei tuoi piani di disaster recovery.

Imparate da tutti i fallimenti operativi

Un'infrastruttura che si ripara automaticamente è uno sforzo a lungo termine che si sviluppa in iterazioni quando si verificano problemi rari o le risposte non sono così efficaci come desiderato. L'adozione delle seguenti pratiche favorisce l'attenzione verso tale obiettivo:

- Promuovi il miglioramento imparando da tutti i fallimenti.
- Condividi ciò che viene appreso tra i team e l'organizzazione. Se più team all'interno di un'organizzazione utilizzano Neptune, crea una chat room o un gruppo di utenti comune per condividere le conoscenze e le migliori pratiche.

Utilizza le funzionalità di registrazione per monitorare attività non autorizzate o anomale

Per osservare modelli anomali di prestazioni e attività, archivia i log in Amazon Logs. CloudWatch Prendi in considerazione le seguenti best practice:

- Abilita la registrazione lenta delle query. Esamina regolarmente il registro e diagnostica il motivo per cui alcune query sono lente. Usa gli endpoint di spiegazione e profilazione di Neptune per Gremlin, SPARQL o OpenCypher per ottenere informazioni sul motivo per cui queste query sono lente.
- Abilita i log di controllo di Neptune e rivedi regolarmente i log per verificare eventuali accessi non autorizzati o anomalie.
- Se utilizzi la registrazione lenta delle query o la registrazione di controllo, abilita la pubblicazione su Logs. CloudWatch Questo vi aiuterà a evitare l'esaurimento dello spazio su disco sulle istanze. Le istanze Neptune hanno una capacità di archiviazione dei log limitata e sovrascrivono i file di registro più vecchi quando viene superato lo spazio di registro. CloudWatch Logs supporta la conservazione a lungo termine dei log. Le funzionalità di monitoraggio avanzate di CloudWatch Logs miglioreranno la capacità di interrogare i log e diagnosticare i problemi.
- Per facilitare strumenti di analisi migliori per i registri di controllo, è possibile configurare un cluster Neptune DB per pubblicare i dati dei log di controllo in un gruppo di log in Logs. CloudWatch Con CloudWatch Logs, è possibile eseguire analisi in tempo reale dei dati di registro, utilizzarli CloudWatch per creare allarmi e visualizzare metriche e utilizzare CloudWatch Logs per archiviare i record di registro in un archivio altamente durevole. Per ulteriori informazioni, consulta Pubblicazione dei log di Neptune su Amazon Logs. CloudWatch
- Neptune supporta la registrazione delle azioni del piano di controllo utilizzando. AWS CloudTrail Per ulteriori informazioni, consulta Registrazione delle chiamate API Amazon Neptune con. AWS CloudTrail

Pilastro della sicurezza

La sicurezza del cloud AWS è la massima priorità. In qualità di AWS cliente, puoi beneficiare di un data center e di un'architettura di rete progettati per soddisfare i requisiti delle organizzazioni più sensibili alla sicurezza.

La sicurezza è una responsabilità condivisa tra AWS te e te. Il modello di responsabilità condivisa descrive questo come sicurezza del cloud e sicurezza nel cloud:

- Sicurezza del cloud: AWS è responsabile della protezione dell'infrastruttura che gira Servizi AWS su Cloud AWS. AWS fornisce inoltre servizi che è possibile utilizzare in modo sicuro. I revisori esterni testano e verificano regolarmente l'efficacia della AWS sicurezza nell'ambito dei programmi di AWS conformità. Per ulteriori informazioni sui programmi di conformità che si applicano ad Amazon Neptune, consulta Servizi AWS coperti dal programma di conformità.
- Sicurezza nel cloud: la tua responsabilità è determinata dal materiale Servizio AWS che utilizzi. L'utente è anche responsabile di altri fattori, tra cui la riservatezza dei dati, i requisiti dell'azienda e le leggi e le normative applicabili. Per ulteriori informazioni sulla privacy dei dati, consulta Domande frequenti sulla privacy dei dati. Per informazioni sulla protezione dei dati in Europa, consulta il modello di responsabilitàAWS condivisa e il post sul blog sul GDPR.

Il pilastro della sicurezza ti aiuta a capire come applicare il modello di responsabilità condivisa quando usi Neptune. I seguenti argomenti illustrano come configurare Neptune per soddisfare gli obiettivi di sicurezza e conformità. Imparerai anche a usarne altri Servizi AWS che ti aiutano a monitorare e proteggere le tue risorse di Neptune.

Il pilastro della sicurezza include le seguenti aree di interesse chiave:

- Sicurezza dei dati
- Sicurezza di rete
- Autenticazione e autorizzazione

Implementare la sicurezza dei

La fuga e le violazioni dei dati mettono a rischio i tuoi clienti e possono avere un impatto negativo sostanziale sulla tua azienda. Le seguenti best practice aiutano a proteggere i dati dei clienti dall'esposizione involontaria e dolosa:

- I nomi dei cluster, i tag, i gruppi di parametri, i ruoli AWS Identity and Access Management (IAM) e altri metadati non devono contenere informazioni riservate o sensibili, poiché tali dati potrebbero apparire nei registri di fatturazione o diagnostica.
- URIs o i collegamenti a server esterni archiviati come dati in Neptune non devono contenere informazioni sulle credenziali per convalidare le richieste.
- Le istanze crittografate di Neptune offrono un livello aggiuntivo di protezione dei dati proteggendoli dagli accessi non autorizzati all'archiviazione sottostante. Puoi usare la crittografia Neptune per aumentare la protezione dei dati delle applicazioni implementate nel cloud. È inoltre possibile utilizzare la crittografia Neptune per soddisfare i requisiti di conformità per i dati archiviati.

  Per abilitare la crittografia per una nuova istanza di Neptune DB, scegli Sì nella sezione Abilita crittografia sulla console Neptune (selezionata per impostazione predefinita) o impostando la proprietà in. AWS::Neptune::DBCluster::StorageEncrypted CloudFormationSe la crittografia è abilitata, Neptune utilizzerà la chiave gestita AWS di Amazon Relational Database Service (Amazon RDS) per impostazione predefinita, oppure puoi creare una chiave gestita dal cliente. Per informazioni sulla creazione di un'istanza DB di Neptune, vedere Creazione di un nuovo cluster Neptune DB. Per maggiori dettagli, consulta Encrypting Neptune Resources at Rest. Le istantanee automatiche e manuali utilizzano la stessa crittografia selezionata per il cluster Neptune.
- Quando utilizzi i linguaggi SPARQL e OpenCypher, pratica tecniche di parametrizzazione e convalida degli input appropriate per prevenire l'iniezione di SQL e altre forme di attacchi. Evita di creare query che utilizzano la concatenazione di stringhe con input forniti dall'utente. Utilizzate interrogazioni con parametri o istruzioni preparate per passare in modo sicuro i parametri di input al database grafico. Per ulteriori informazioni, consulta Esempi di query parametrizzate OpenCypher e SPARQL Injection Defence.
- Per il linguaggio Gremlin, usa le varianti del linguaggio Gremlin invece di passare direttamente script Gremlin basati su stringhe per evitare potenziali problemi di iniezione.

Proteggi le tue reti

Un cluster Amazon Neptune DB può essere creato solo in un cloud privato virtuale (VPC) su. AWS Fino a Neptune 1.4.6.0, gli endpoint del cluster Neptune DB erano accessibili solo all'interno di quel VPC. A partire dalla versione 1.4.6.0 di Neptune e successive, le istanze di Neptune possono essere configurate per essere accessibili pubblicamente su Internet. È consigliabile utilizzare questa funzionalità solo in ambienti non di produzione per consentire l'accesso semplificato a Neptune per gli sviluppatori (sebbene l'autenticazione IAM sia sempre richiesta per consentire l'accessibilità

pubblica). Se hai abilitato l'accessibilità pubblica, valuta la possibilità di impostare le regole del gruppo di sicurezza in entrata per la porta del database solo sul traffico di indirizzi IP noto. In ambienti di produzione o con cluster contenenti dati sensibili, proteggi i dati di Neptune impedendo l'accessibilità pubblica e limitando l'accesso al VPC in cui si trova il cluster Neptune DB. Per ulteriori informazioni, consulta Connessione al grafico di Amazon Neptune.

Per proteggere i dati in transito, Neptune applica le connessioni SSL tramite HTTPS a qualsiasi istanza o endpoint del cluster utilizzando protocolli e cifrari sicuri. Neptune fornisce certificati SSL per le istanze DB di Neptune. I certificati SSL Neptune supportano solo i nomi host degli endpoint del cluster, degli endpoint di lettura e degli endpoint delle istanze.

Se si utilizza un sistema di bilanciamento del carico o un server proxy (ad esempio HAProxy), è necessario utilizzare la terminazione SSL e disporre del proprio certificato SSL sul server proxy. Il passthrough SSL non funziona perché i certificati SSL forniti non corrispondono al nome host del server proxy. Per ulteriori informazioni sulla connessione agli endpoint Neptune con SSL, consulta Utilizzo dell'endpoint HTTP REST per connettersi a un'istanza DB di Neptune.

Implementa l'autenticazione e l'autorizzazione

Per controllare chi può eseguire azioni di gestione di Neptune su cluster e istanze DB Neptune, abilita l'autenticazione del database IAM e utilizza le credenziali IAM. Quando ti connetti AWS utilizzando le credenziali IAM, il tuo ruolo IAM deve disporre di policy IAM che concedano le autorizzazioni necessarie per eseguire le operazioni di gestione di Neptune. Assicurati di seguire il principio del privilegio minimo, concedendo solo le autorizzazioni necessarie per completare un'attività. Per ulteriori informazioni, consulta Utilizzo di diversi tipi di policy IAM per il controllo dell'accesso a Neptune e Autenticazione IAM tramite credenziali temporanee.

Per controllare chi può connettersi a un cluster Neptune e interrogare i dati, puoi utilizzare IAM per autenticarti sull'istanza o sul cluster DB di Neptune. Se abiliti l'autenticazione IAM in un cluster Neptune DB, chiunque acceda al cluster DB deve prima essere autenticato. Per ulteriori informazioni, consulta Abilitazione dell'autenticazione del database IAM in Neptune per i passaggi per abilitare l'autenticazione IAM.

Quando è abilitata l’autenticazione database IAM, ogni richiesta deve essere firmata utilizzando AWS Signature Version 4. Per capire come inviare richieste firmate a tutti gli endpoint Neptune con l'autenticazione IAM abilitata, consulta Connecting and Signature with AWS Signature Version 4. Molte librerie e strumenti, come awscurl, supportano già la versione 4 di Signature. AWS

Per interagire con altri Servizi AWS, Amazon Neptune utilizza ruoli collegati ai servizi IAM. Un ruolo collegato ai servizi è un tipo di ruolo IAM univoco collegato direttamente a Neptune. I ruoli collegati ai servizi sono predefiniti da Neptune e includono tutte le autorizzazioni richieste dal servizio per chiamare altri utenti per conto dell'utente. Servizi AWS Per ulteriori informazioni, vedere Utilizzo dei ruoli collegati ai servizi per Neptune.

Pilastro dell'affidabilità

Il pilastro dell'affidabilità comprende la capacità di un carico di lavoro di svolgere la funzione prevista in modo corretto e coerente quando previsto. Ciò comprende la possibilità di utilizzare e testare il carico di lavoro per tutto il ciclo di vita.

Un carico di lavoro affidabile comincia con decisioni iniziali di progettazione sia per il software sia per l'infrastruttura. Le tue scelte di architettura influiranno sul comportamento del carico di lavoro in tutti i pilastri di AWS Well-Architected. Per l'affidabilità, è necessario seguire modelli specifici.

Il pilastro dell'affidabilità si concentra sulle seguenti aree chiave:

- Architettura del carico di lavoro, comprese le quote di servizio e i modelli di implementazione
- Gestione delle modifiche
- Gestione dei guasti

Comprendi le quote di servizio di Neptune

Un volume cluster Neptune può crescere fino a una dimensione massima di 128 tebibyte (TiB) in tutte le aree supportate Regioni AWS tranne la Cina e GovCloud, dove la quota è di 64 TiB.

La quota di 128 TiB è sufficiente per memorizzare circa 200-400 miliardi di oggetti nel grafico. In un grafico delle proprietà etichettate (LPG), un oggetto è un nodo, un bordo o una proprietà su un nodo o un bordo. In un grafico RDF (Resource Description Framework), un oggetto è un quadrilatero.

Per qualsiasi cluster Neptune Serverless, è possibile impostare sia il numero minimo che il numero massimo di unità di capacità Neptune (). NCUs Ogni NCU è composta da 2 gibibyte (GiB) di memoria e dalla vCPU e rete associate. I valori NCU minimo e massimo si applicano a tutte le istanze serverless del cluster. Il valore NCU massimo che è possibile impostare è 128,0 e il valore minimo è 1,0 NCUs. NCUs Ottimizza la gamma NCU più adatta alla tua applicazione osservando le CloudWatch metriche di Amazon e acquisendo l'intervallo in cui lavori di solito `ServerlessDatabaseCapacity` e `NCUUtilization` correlando comportamenti o costi indesiderati all'interno di tale intervallo. In molti carichi di lavoro, 1,0 NCU è un punto di partenza troppo basso e comporta un comportamento inaffidabile dopo periodi di inattività. Se ritieni che il tuo carico di lavoro non sia sufficientemente scalabile, aumenta il valore minimo NCUs per fornire un'elaborazione sufficiente per il picco iniziale durante la scalabilità.

Account AWS Ciascuna di esse prevede quote per ogni regione sul numero di risorse di database che è possibile creare. Queste risorse includono le istanze database e i cluster di database. Dopo aver raggiunto il limite per una risorsa, le ulteriori richieste di creazione di tale risorsa restituiranno un errore con un'eccezione. Alcune quote sono quote flessibili che possono essere aumentate su richiesta. Per un elenco delle quote condivise tra Amazon Neptune e Amazon RDS, Amazon Aurora e Amazon DocumentDB (con compatibilità con MongoDB), insieme ai link per richiedere aumenti delle quote quando disponibili, consulta Quotas in Amazon RDS.

Comprendi i modelli di implementazione di Neptune

Nei cluster Neptune DB, esiste un'istanza DB principale e fino a 15 repliche Neptune. L'istanza DB principale supporta le operazioni di lettura e scrittura ed esegue tutte le modifiche ai dati sul volume del cluster. Le repliche Neptune si connettono allo stesso volume di storage dell'istanza DB principale e supportano solo operazioni di lettura. Le repliche Neptune possono scaricare i carichi di lavoro di lettura dall'istanza DB principale.

Per ottenere un'elevata disponibilità, utilizzate le repliche di lettura. La disponibilità di una o più istanze di replica di lettura in diverse zone di disponibilità può aumentare la disponibilità perché le repliche di lettura fungono da obiettivi di failover per l'istanza principale. Se l'istanza writer fallisce, Neptune promuove un'istanza di replica in lettura facendola diventare l'istanza principale. Quando ciò accade, si verifica una breve interruzione (generalmente inferiore a 30 secondi) durante il riavvio dell'istanza promossa, durante la quale le richieste di lettura e scrittura fatte all'istanza principale hanno esito negativo con un'eccezione. Per la massima affidabilità, prendi in considerazione due repliche di lettura in zone di disponibilità diverse. Se l'istanza principale nella Zona di disponibilità 1 va offline, l'istanza nella Zona di disponibilità 2 viene promossa a principale, ma non può gestire le query mentre ciò accade. Pertanto è necessaria un'istanza nella Zona di disponibilità 3 per gestire le query di lettura durante la transizione.

Se utilizzi Neptune Serverless, le istanze di lettura e scrittura in tutte le zone di disponibilità verranno scalate verso l'alto e verso il basso, indipendentemente l'una dall'altra, a seconda del carico del database. È possibile impostare il livello di promozione di un'istanza Reader su 0 o 1 in modo che venga scalato verso l'alto e verso il basso in base alla capacità dell'istanza writer. In questo modo è pronta ad assumersi il carico di lavoro corrente in qualsiasi momento.

Se la tua applicazione ha un'impronta globale o richiede un failover multiregionale, prendi in considerazione l'utilizzo di un database globale Neptune. Un database globale di Amazon Neptune si estende su Regioni AWS più pagine, consente letture globali a bassa latenza e fornisce un

ripristino rapido nei rari casi in cui un'interruzione si ripercuota su un intero sistema. Regione AWS Un database globale Neptune è costituito da un cluster DB primario in una regione e da un massimo di cinque cluster DB secondari in regioni diverse.

Gestisci e ridimensiona i cluster Neptune

È possibile utilizzare l'auto-scaling di Neptune per regolare automaticamente il numero di repliche Neptune in un cluster DB per soddisfare i requisiti di connettività e carico di lavoro in base alle soglie di utilizzo della CPU. Con l'auto-scaling, il cluster Neptune DB è in grado di gestire aumenti improvvisi del carico di lavoro. Quando il carico di lavoro diminuisce, l'auto-scaling rimuove le repliche non necessarie in modo da non dover pagare per la capacità inutilizzata. Tieni presente che l'avvio di una nuova istanza può richiedere fino a 15 minuti, quindi l'auto-scaling da solo non è una soluzione sufficiente per i rapidi cambiamenti della domanda.

Puoi utilizzare l'auto-scaling solo con un cluster Neptune DB che ha già un'istanza writer principale e almeno un'istanza di lettura-replica (vedi Amazon Neptune DB Clusters and Instances). Inoltre, tutte le istanze di replica di lettura nel cluster devono essere in stato disponibile. Se una replica di lettura si trova in uno stato diverso da quello disponibile, l'auto-scaling di Neptune non fa nulla finché non sono disponibili tutte le repliche di lettura nel cluster.

Se riscontri rapidi cambiamenti nella domanda, prendi in considerazione l'utilizzo di istanze serverless. Le istanze serverless possono essere scalate verticalmente per brevi periodi, mentre l'auto-scaling è scalabile orizzontalmente per periodi più lunghi. Questa configurazione offre una scalabilità ottimale perché le istanze serverless scalano verticalmente, mentre l'auto-scaling crea istanze di nuove repliche di lettura per gestire il carico di lavoro oltre la capacità massima di una singola istanza serverless. Per ulteriori informazioni sulla scalabilità della capacità di Amazon Neptune Serverless, consulta Scalabilità della capacità in un cluster DB Neptune Serverless.

Se le tue esigenze di scalabilità cambiano in tempi prevedibili, puoi pianificare le modifiche alle istanze minime, alle istanze massime e alle soglie per gestire meglio queste mutevoli esigenze. Ricorda di pianificare gli eventi di scalabilità orizzontale con almeno 15 minuti di anticipo per consentire a tali istanze di essere online quando necessario.

Puoi gestire la configurazione del database in Amazon Neptune utilizzando i parametri in un gruppo di parametri. I gruppi di parametri fungono da container per i valori di configurazione di un motore applicati a una o più istanze database. Quando modificate i parametri del cluster in gruppi di parametri, comprendete la differenza tra parametri statici e dinamici e come e quando vengono applicati. Usa l'endpoint di stato per vedere la configurazione attualmente applicata.

Gestisci i backup e gli eventi di failover

Neptune esegue automaticamente il backup del volume del cluster e conserva i dati di backup per tutta la durata del periodo di conservazione del backup. I backup di Neptune sono continui e incrementali, in modo da poter eseguire rapidamente un ripristino a qualsiasi momento nel tempo del periodo di conservazione del backup. È possibile specificare un periodo di conservazione dei backup da 1 a 35 giorni quando si crea o si modifica un cluster DB.

Per conservare un backup oltre il periodo di conservazione del backup, è anche possibile scattare un'istantanea dei dati nel volume del cluster. All'archiviazione degli snapshot verranno applicati i costi di archiviazione standard per Neptune.

Quando crei uno snapshot Amazon Neptune di un cluster DB, Neptune crea uno snapshot del volume di storage del cluster, eseguendo il backup di tutti i dati, non solo delle singole istanze. È possibile creare un nuovo cluster database ripristinandolo da questo snapshot. Quando ripristini il cluster DB, fornisci il nome dello snapshot del cluster DB da cui eseguire il ripristino, quindi fornisci un nome per il nuovo cluster DB creato dal ripristino.

Verifica la risposta del sistema agli eventi di failover. Usa l'API Neptune per forzare un evento di failover. Il riavvio con failover è utile quando si desidera simulare un guasto di un'istanza DB per eseguire test o ripristinare le operazioni nella zona di disponibilità originale dopo un failover. Per ulteriori informazioni, consulta Configurazione e gestione di un'implementazione Multi-AZ. Quando si riavvia un'istanza di DB Writer, questa passa alla replica in standby. Il riavvio di una replica Neptune non causa un failover.

Progetta i tuoi clienti in modo da renderli affidabili. Verifica il loro comportamento durante gli eventi di failover. Implementa la logica di ripetizione dei tentativi nel tuo client con una logica di backoff esponenziale. Gli esempi di codice che implementano questa logica sono disponibili nella documentazione sotto gli esempi di AWS Lambda funzioni per Amazon Neptune.

Prendi in considerazione l'utilizzo AWS Backupse disponi di un set comune di requisiti di backup da applicare a più motori di database.

Pilastro dell'efficienza delle prestazioni

Il pilastro dell'efficienza delle prestazioni del AWS Well-Architected Framework si concentra su come ottimizzare le prestazioni durante l'acquisizione o l'interrogazione dei dati. L'ottimizzazione delle prestazioni è un processo incrementale e continuo che prevede quanto segue:

- Conferma dei requisiti aziendali
- Misurazione delle prestazioni del carico di lavoro
- Identificazione dei componenti poco performanti
- Ottimizzazione dei componenti per soddisfare le esigenze aziendali

Il pilastro relativo all'efficienza delle prestazioni fornisce linee guida specifiche per i casi d'uso che possono aiutare a identificare il modello di dati grafici e i linguaggi di interrogazione corretti da utilizzare. Include inoltre le best practice da seguire per l'acquisizione e l'utilizzo di dati da Amazon Neptune.

Il pilastro dell'efficienza prestazionale si concentra sulle seguenti aree chiave:

- Modellazione grafica
- Ottimizzazione delle query
- Ridimensionamento corretto del cluster
- Ottimizzazione della scrittura

Comprendi la modellazione dei grafici

Comprendi la differenza tra i modelli Labeled Property Graph (LPG) e Resource Description Framework (RDF). Nella maggior parte dei casi, è una questione di preferenza. Esistono tuttavia diversi casi d'uso in cui un modello è più adatto dell'altro. Se hai bisogno di conoscere il percorso che collega due nodi nel tuo grafico, scegli GPL. Se desideri federare i dati tra cluster Neptune o altri archivi Graph Triple, scegli RDF.

Se stai creando un'applicazione SaaS (Software as a Service) o un'applicazione che richiede la multi-tenancy, valuta la possibilità di incorporare la separazione logica dei tenant nel tuo modello di dati anziché avere un tenant per ogni cluster. Per ottenere questo tipo di progettazione, è

possibile utilizzare grafici denominati SPARQL e strategie di etichettatura, ad esempio anteporre gli identificatori del cliente alle etichette o aggiungere coppie chiave-valore di proprietà che rappresentano gli identificatori dei tenant. Assicurati che il tuo livello client inserisca questi valori per mantenere quella separazione logica. Per ulteriori informazioni sui consigli di multi-tenancy, consulta la guida multi-tenancy per l'esecuzione di database Amazon ISVs Neptune.

Le prestazioni delle query dipendono dal numero di oggetti grafici (nodi, bordi, proprietà) che devono essere valutati durante l'elaborazione della query. Pertanto, il modello grafico può avere un impatto significativo sulle prestazioni dell'applicazione. Utilizzate etichette granulari quando possibile e memorizzate solo le proprietà necessarie per determinare il percorso o filtrare. Per ottenere prestazioni più elevate, prendi in considerazione la possibilità di precalcolare alcune parti del grafico, ad esempio creando nodi di riepilogo o bordi più diretti che collegano percorsi comuni.

Cerca di evitare di navigare tra nodi con un numero anormalmente elevato di bordi con la stessa etichetta. Tali nodi hanno spesso migliaia di spigoli (mentre la maggior parte dei nodi ha un numero di spigoli dell'ordine di decine). Il risultato è una complessità di elaborazione e dati molto più elevata. Questi nodi potrebbero non essere problematici in alcuni modelli di query, ma consigliamo di modellare i dati in modo diverso per evitarli, soprattutto se si naviga attraverso il nodo come passaggio intermedio. È possibile utilizzare i log di interrogazione lenta per identificare le query che navigano tra questi nodi. Probabilmente osserverai metriche di latenza e accesso ai dati molto più elevate rispetto ai modelli di query medi, specialmente se utilizzi la modalità di debug.

Usa il nodo deterministico IDs per nodi e bordi se il tuo caso d'uso lo supporta invece di usare Neptune per assegnare valori GUID casuali per. IDs L'accesso ai nodi tramite ID è il metodo più efficiente.

Ottimizzazione delle query

I linguaggi OpenCypher e Gremlin possono essere usati in modo intercambiabile sui modelli GPL. Se le prestazioni sono una delle principali preoccupazioni, valuta la possibilità di utilizzare i due linguaggi in modo intercambiabile, perché uno potrebbe funzionare meglio dell'altro per modelli di query specifici.

Neptune è in fase di conversione al suo motore di interrogazione alternativo (DFE). OpenCypher funziona solo sul DFE, ma sia le query Gremlin che SPARQL possono essere impostate opzionalmente per l'esecuzione sul DFE utilizzando le annotazioni di query. Prendi in considerazione la possibilità di testare le tue query con il DFE attivato e di confrontare le prestazioni del tuo modello di query quando non usi il DFE.

Neptune è ottimizzato per le query di tipo transazionale che partono da un singolo nodo o set di nodi e partono a ventaglio da lì, anziché per le query analitiche che valutano l'intero grafico. Per i tuoi carichi di lavoro di query analitiche, usa Neptune Analytics. Neptune Analytics è la scelta ideale per carichi di lavoro investigativi, esplorativi o di data science che richiedono un'iterazione rapida per l'elaborazione di dati, analitica e algoritmica. Può anche eseguire una ricerca vettoriale su dati grafici e caricare i dati direttamente dall'istanza del database Neptune. Se Neptune Analytics non soddisfa le tue esigenze, puoi anche AWS prendere in considerazione l'SDK per Pandas o l'utilizzo di neptune-export in combinazione con Amazon EMR. AWS Glue

Per identificare inefficienze e rallentamenti nei modelli e nelle query, utilizza il linguaggio e for each query per ottenere spiegazioni dettagliate del `profile` piano di query e delle metriche di query. `explain` APIs Per ulteriori informazioni, consultate Gremlin profile, OpenCypher explain e SPARQL explain.

Comprendi i tuoi schemi di interrogazione. Se il numero di spigoli distinti in un grafico aumenta, la strategia di accesso predefinita a Neptune può diventare inefficiente. Le seguenti interrogazioni potrebbero diventare piuttosto inefficienti:

- Query che navigano all'indietro tra i bordi quando non vengono fornite etichette sui bordi.
- Clausole che utilizzano lo stesso schema internamente, come `.both()` in Gremlin, o clausole che eliminano nodi in qualsiasi lingua (il che richiede l'eliminazione dei bordi in entrata senza conoscere le etichette).
- Interrogazioni che accedono ai valori delle proprietà senza specificare le etichette delle proprietà. Queste interrogazioni potrebbero diventare piuttosto inefficienti. Se questo corrisponde al tuo modello di utilizzo, valuta la possibilità di abilitare l'indice OSGP (oggetto, soggetto, grafico, predicato).

Utilizzate la registrazione delle query lente per identificare le query lente. Le query lente possono essere causate da piani di query non ottimizzati o da un numero inutilmente elevato di ricerche nell'indice, che possono aumentare i costi. I/O Gli endpoint di spiegazione e profilazione di Neptune per Gremlin, SPARQL o OpenCypher possono aiutarti a capire perché queste query sono lente. Le cause potrebbero includere le seguenti:

- I nodi con un numero anormalmente elevato di bordi rispetto al nodo medio nel grafico (ad esempio, migliaia rispetto a decine) possono aggiungere complessità computazionale e quindi una maggiore latenza e un maggiore consumo di risorse. Determinate se questi nodi sono modellati

correttamente o se i modelli di accesso possono essere migliorati per ridurre il numero di spigoli da attraversare.

- Le interrogazioni non ottimizzate conterranno un avviso che indica che passaggi specifici non sono ottimizzati. La riscrittura di queste query per utilizzare passaggi ottimizzati potrebbe migliorare le prestazioni.
- I filtri ridondanti potrebbero causare ricerche nell'indice non necessarie. Allo stesso modo, i modelli ridondanti potrebbero causare ricerche negli indici duplicate che possono essere ottimizzate migliorando la query (vedi nell'output del profilo). `Index Operations - Duplication ratio`
- Alcuni linguaggi come Gremlin non hanno valori numerici fortemente tipizzati e utilizzano invece la promozione dei tipi. Ad esempio, se il valore è 55, Neptune cerca valori che siano numeri interi, lunghi, float e altri tipi numerici equivalenti a 55. Ciò comporta operazioni aggiuntive. Se sai in anticipo che i tuoi tipi corrispondono, puoi evitarlo utilizzando un suggerimento di interrogazione.
- Il tuo modello grafico può influire notevolmente sulle prestazioni. Prendi in considerazione la possibilità di ridurre il numero di oggetti che devono essere valutati utilizzando etichette più granulari o precalcolando scorciatoie per percorsi lineari a più hop.

Se l'ottimizzazione delle query da sola non vi consente di raggiungere i vostri requisiti di prestazioni, prendete in considerazione l'utilizzo di una varietà di tecniche di caching con Neptune per soddisfare tali requisiti.

Le prestazioni di Neptune migliorano continuamente con ogni versione. Consulta le note di rilascio per vedere i dettagli dei miglioramenti apportati a ciascuna versione. Prendi in considerazione la pianificazione di aggiornamenti regolari del tuo cluster Neptune DB per ottenere prestazioni ottimali. Le versioni più recenti supportano anche le istanze più recenti. Prendi in considerazione l'aggiornamento alla versione 1.4.5.0 o successiva per poter utilizzare le istanze. `r8g` Per ulteriori informazioni su come ciò può migliorare le prestazioni del carico di lavoro, consulta il rapporto prezzo/prestazioni delle query di scrittura 4,7 volte migliore con le istanze AWS Graviton4 R8g che utilizzano Amazon Neptune v1.4.5.

Cluster di dimensioni corrette

Dimensiona il cluster in base ai tuoi requisiti di concorrenza e velocità effettiva. Il numero di query simultanee che possono essere gestite da ciascuna istanza del cluster è pari a due volte il numero di CPUs (vCPUs) virtuali su quell'istanza. Le interrogazioni aggiuntive che arrivano mentre tutti i thread di lavoro sono occupati vengono inserite in una coda sul lato server. Queste domande vengono

gestite su base first-in-first-out (FIFO) quando i thread di lavoro diventano disponibili. Il CloudWatch parametro `MainRequestQueuePendingRequests` Amazon mostra la profondità attuale della coda per ogni istanza. Se questo valore è spesso superiore a zero, valuta la possibilità di scegliere un'istanza con più v. CPUs Se la profondità della coda supera 8.192, Neptune restituirà un errore. `ThrottlingException`

Circa il 65 percento della RAM per ogni istanza è riservato alla cache buffer. La cache buffer contiene il set di dati di lavoro (non l'intero grafico, solo i dati che vengono interrogati). Per determinare la percentuale di dati che viene recuperata dalla cache buffer anziché dallo storage, monitora la metrica. CloudWatch `BufferCacheHitRatio` Se questa metrica scende spesso al di sotto del 99,9%, prendi in considerazione l'idea di provare un'istanza con più memoria per determinare se ciò riduce la latenza e i costi. I/O

Le repliche di lettura non devono necessariamente avere le stesse dimensioni dell'istanza di Writer. Tuttavia, carichi di lavoro di scrittura pesanti possono far sì che le repliche più piccole rimangano indietro e si riavviino perché non riescono a tenere il passo con la replica. Pertanto, si consiglia di creare repliche uguali o più grandi dell'istanza di writer.

Quando utilizzi l'auto-scaling per le tue repliche di lettura, ricorda che potrebbero essere necessari fino a 15 minuti per portare online una nuova replica di lettura. Quando il traffico del client aumenta rapidamente ma in modo prevedibile, prendi in considerazione l'utilizzo della scalabilità pianificata per impostare un numero minimo di repliche di lettura più elevato in modo da tenere conto del tempo di inizializzazione.

Le istanze serverless supportano diversi casi d'uso e carichi di lavoro. Prendi in considerazione le istanze serverless anziché le istanze con provisioning per i seguenti scenari:

- Il carico di lavoro varia spesso nel corso della giornata.
- Hai creato una nuova applicazione e non sei sicuro delle dimensioni del carico di lavoro.
- Stai eseguendo attività di sviluppo e test.

È importante notare che le istanze serverless sono più costose delle istanze equivalenti con provisioning in base a un dollaro per GB di RAM. Ogni istanza serverless è composta da 2 GB di RAM con vCPU e rete associate. Esegui un'analisi dei costi tra le opzioni disponibili per evitare fatture a sorpresa. In generale, con serverless è possibile ottenere risparmi sui costi solo quando il carico di lavoro è molto intenso solo per poche ore al giorno e quasi zero il resto della giornata o se il carico di lavoro oscilla notevolmente nel corso della giornata.

Utilizza il calcolatore dei prezzi di Amazon Neptune per valutare la configurazione corretta per il tuo cluster in base a queries-per-second fattori come i requisiti (QPS).

Ottimizza le scritture

Per ottimizzare le scritture, considerate quanto segue:

- Il Neptune Bulk Loader è il modo ottimale per caricare inizialmente il database o aggiungerlo ai dati esistenti. Il loader Neptune non è transazionale e non può eliminare dati, quindi non utilizzarlo se questi sono i tuoi requisiti.
- Gli aggiornamenti transazionali possono essere effettuati utilizzando i linguaggi di interrogazione supportati. Per ottimizzare le I/O operazioni di scrittura, scrivi i dati in batch di 50-100 oggetti per commit. Un oggetto è un nodo, un bordo o una proprietà su un nodo o un bordo in GPL, oppure un triplo archivio o un quadrilatero in RDF.
- Tutte le operazioni di scrittura transazionali di Neptune sono a thread singolo per ogni connessione. Quando invii una grande quantità di dati a Neptune, prendi in considerazione la possibilità di avere più connessioni parallele, ognuna delle quali consiste nella scrittura di dati. Quando si sceglie un'istanza con provisioning di Neptune, la dimensione dell'istanza è associata a un numero di v. CPUs Neptune crea due thread di database per ogni vCPU sull'istanza, quindi inizia con il doppio di v quando esegui il test per una parallelizzazione CPUs ottimale. Le istanze serverless ridimensionano il numero di v CPUs a una velocità di circa uno per 4. NCUs

  > **Note**
  >
  > Questo non si applica all'API di caricamento in blocco, ma solo alle connessioni dirette.

- Pianifica e gestisci ConcurrentModificationExceptionsin modo efficiente tutti i processi di scrittura, anche se solo una singola connessione sta scrivendo dati in qualsiasi momento. Progetta i tuoi clienti in modo da renderli affidabili quando `ConcurrentModificationExceptions` si verificano.
- Se desideri eliminare tutti i tuoi dati, valuta la possibilità di utilizzare l'API di ripristino rapido anziché inviare query di eliminazione simultanee. Quest'ultima richiederà molto più tempo e comporterà I/O costi sostanziali rispetto alla prima.
- Se desideri eliminare la maggior parte dei tuoi dati, valuta la possibilità di esportare i dati che desideri conservare utilizzando neptune-export per caricare i dati in un nuovo cluster. Quindi elimina il cluster originale.

Pilastro dell'ottimizzazione dei costi

Il pilastro dell'ottimizzazione dei costi del AWS Well-Architected Framework si concentra sull'evitare costi inutili. I seguenti consigli possono aiutarti a soddisfare i principi di progettazione per l'ottimizzazione dei costi e le best practice architettoniche per Amazon Neptune.

Il pilastro dell'ottimizzazione dei costi si concentra sulle seguenti aree chiave:

- Comprendere la spesa nel tempo e controllare l'allocazione dei fondi
- Selezione delle risorse del tipo e della quantità corretti
- Scalabilità per soddisfare le esigenze aziendali senza spese eccessive

Comprendi i modelli di utilizzo e i servizi necessari

Neptune è la soluzione ideale per il tuo carico di lavoro se il tuo modello di dati ha una struttura a grafi riconoscibile e le tue query devono esplorare le relazioni e attraversare più hop. Un database grafico non è adatto ai seguenti modelli:

- Principalmente query a hop singolo (valuta se i tuoi dati potrebbero essere meglio rappresentati come attributi di un oggetto)
- Dati JSON o BLOB archiviati come proprietà
- Query che si aggregano in un set di dati, ad esempio il calcolo della somma di una proprietà numerica su un gran numero di nodi

Valuta se l'utilizzo congiunto di più database creati appositamente per modelli di accesso specifici possa soddisfare tutte le tue esigenze. Esempio:

- Un'API che richiede navigazioni grafiche complesse meno frequenti insieme al recupero altamente simultaneo di proprietà per un singolo nodo potrebbe essere presentata al meglio utilizzando uno o più Neptune, DynamoDB o Amazon DocumentDB.
- I database relazionali possono coesistere con Neptune per mantenere le funzionalità esistenti, ma usa Neptune solo per attraversamenti multiple-hop che non offrono prestazioni e scalabilità adeguate nei database relazionali.

Comprendi i costi associati ai servizi che interagiscono e completano Neptune, inclusi i seguenti:

- Costi di storage di Amazon Simple Storage Service (Amazon S3) per i file di dati caricati in blocco su Neptune
- Funzioni Lambda utilizzate per inserire o alterare le interrogazioni, le query di lettura e l'elaborazione dei flussi di Neptune
- Il livello API basato su Neptune per interagire con l'applicazione client (anziché avere connessioni dirette al database) in Amazon API Gateway o AWS AppSync
- AWS Glue lavori utilizzati per trasferire dati da e verso Neptune
- Le istanze Amazon Kinesis o Amazon Managed Streaming for Apache Kafka (Amazon MSK) ricevono dati in streaming per l'ingestione quasi in tempo reale in Neptune.
- AWS Database Migration Service per la migrazione di dati relazionali su Neptune
- Costi SageMaker di Amazon Runtime per i notebook Jupyter e i modelli di machine learning delle librerie Deep Graph

Seleziona le risorse prestando attenzione ai costi

I prezzi di Neptune si basano sul costo orario dell'istanza (o sulle unità di calcolo Neptune utilizzate per la versione serverless), sull'I/O dei dati e sull'utilizzo dello storage. Le istanze rappresentano, in media, l'85% del costo complessivo, quindi un corretto dimensionamento può avere implicazioni significative in termini di costi. Il modo migliore per dimensionare correttamente le istanze è testare le prestazioni delle applicazioni su una varietà di istanze e confrontare i seguenti fattori:

- La `MainRequestQueuePendingRequests` CloudWatch metrica si attesta a un numero costantemente basso vicino allo zero?
- La `BufferCacheHitRatio` CloudWatch metrica rimane pari o superiore al 99,9 percento per la maggior parte del tempo?
- Quali sono le curve dei costi e delle prestazioni, ad esempio i costi e i costi associati ai dati? I/O I costi di lettura dei dati potrebbero aumentare in modo significativo con un'istanza sottodimensionata che richiede lo scambio frequente della cache del buffer con lo storage. `BufferCacheHitRatio`diminuirà frequentemente in questi scenari.

I costi delle istanze variano in modo lineare in base alle dimensioni all'interno della stessa famiglia di istanze. Il costo orario dell'`db.r6i.2xlarge`istanza è il doppio di quello dell'`db.r6i.xlarge`istanza e ha anche il doppio dell'allocazione delle risorse. L'`db.r6i.24xlarge`istanza è 24 volte il costo orario dell'istanza. `db.r6i.xlarge`

Stima il numero di query simultanee che devi supportare. È possibile disporre di un numero compreso tra zero e quindici repliche di lettura per l'elaborazione di interrogazioni di sola lettura. Se i requisiti variano in base all'ora del giorno, della settimana o del mese, puoi utilizzare più istanze più piccole per scalare in base a una pianificazione. Ogni vCPU su un'istanza fornisce due thread per la gestione delle query simultanee. Tre repliche di `db.r6i.xlarge` lettura, con 4 vCPU ciascuna, possono gestire 24 query simultanee.

Se il volume di traffico viene invece misurato in query al secondo (QPS), è necessario sperimentare per determinare la latenza media delle query. Il numero di query al secondo che un cluster Neptune può supportare è pari a. `vCPU × 2 × (1 second/average query latency)` Ad esempio, se si dispone di 4 vCPU e una latenza di query di 100 millisecondi (0,1 secondi),. `QPS = 4 × 2 × (1s/0.1s) = 80 queries per second`

Le istanze con provisioning sono più economiche di quelle senza server per carichi di lavoro continui, stabili e prevedibili. Il serverless offre opportunità di ottimizzazione dei costi quando si dispone di un carico di lavoro che richiede un utilizzo molto elevato per poche ore al giorno (ad esempio`db.r6i.4xlarge`) e quindi quasi nessun traffico per il resto della giornata (ad esempio, 1 Neptune Compute Unit). Un'istanza serverless con scalabilità verso l'alto per alcune ore e poi ridotta sarà meno costosa rispetto all'utilizzo di un'istanza con provisioning per tutto il giorno. `db.r6i.4xlarge`

Prendi in considerazione l'aggiornamento a Neptune 1.4.5.0 o versione successiva e l'utilizzo di `r8g` istanze per ottenere una migliore velocità di lettura e scrittura a un costo inferiore rispetto alle istanze di vecchia generazione, come o. `r7g` `r6g` Per ulteriori informazioni, consulta il rapporto prezzo/prestazioni di scrittura delle query 4,7 volte migliore con le istanze AWS Graviton4 R8g che utilizzano Amazon Neptune v1.4.5 (post sul blog).AWS

I cluster Neptune vengono creati per impostazione predefinita con lo storage standard (se crei utilizzando la console, per impostazione predefinita I/O-optimized storage). With I/O-optimized storage, you pay a slightly higher cost for storage and instances, but there are no I/O costs. This leads to more predictable recurring costs, but if your I/O usage is generally low, it may be more cost efficient to utilize standard storage. If you intend to load a lot of data initially, you can optimize cost by choosing I/O selezionerà l'archiviazione ottimizzata, eseguirà il caricamento iniziale dei dati e quindi passerà allo storage standard. Il tipo di archiviazione influisce solo sul modello di fatturazione e non presenta differenze tecniche nella configurazione del cluster o dell'istanza di Neptune DB. È possibile modificare il tipo di archiviazione una volta ogni 30 giorni. Dopo 30 giorni, controlla i costi dettagliati di Neptune e utilizza la pagina dei prezzi di Neptune per calcolare se i costi sarebbero stati più elevati

utilizzando -optimized. I/O-optimized storage. If they would have been, continue to use standard storage, otherwise switch back to I/O

Scegli la configurazione dell'istanza Neptune migliore per il tuo carico di lavoro

Se l'hai creato Account AWS prima del 15 luglio 2025, puoi utilizzare il piano AWS gratuito per sperimentare Neptune come principiante. Le 750 ore gratuite di `db.t3.medium` utilizzo dell'`db.t4g.medium`istanza sono sufficienti per acquisire una buona comprensione di Neptune su bassa scala. Il tuo cluster rimarrà attivo dopo la fine del periodo di prova gratuito, anche se da quel momento in poi ti verranno addebitati i costi per l'utilizzo.

Le `db.t4g.medium` istanze `db.t3.medium` e sono ideali per ambienti di sviluppo a basso costo in cui non si utilizzano OpenCypher, Graph Explorer o varie integrazioni di intelligenza artificiale generativa. Queste istanze hanno un RAM-to-vCPU rapporto inferiore (2:1) rispetto alle istanze familiari (8:1) o alle istanze R familiari (16:1). X Questa riduzione del rapporto impedisce l'uso delle statistiche del motore DFE che abilitano le prestazioni di OpenCypher, le integrazioni GenAI (per informare l'LLM dello schema del grafico) e Graph Explorer. I profili prestazionali potrebbero differire in modo significativo quando si utilizzano istanze T familiari, in particolare per i carichi di lavoro menzionati in precedenza. Queste istanze possono anche aumentare il numero di casi in `OutOfMemoryExceptions` cui le query navigano su una parte significativa del grafico. Per determinare se quest'ultima condizione potrebbe essere influenzata, controlla la `BufferCacheHitRatio` CloudWatch metrica.

Sconsigliamo vivamente di eseguire test delle prestazioni o del carico con le istanze T familiari perché potrebbero verificarsi risultati incoerenti che non sono indicativi di un ambiente di produzione.

Le istanze con provisioning offrono la migliore combinazione di costi e prestazioni quando il carico di lavoro è abbastanza stabile e prevedibile. Scegliete la dimensione dell'istanza in base alla contemporaneità della richiesta e alla complessità della query. Una maggiore concorrenza richiede più v. CPUs Una maggiore complessità delle query richiede più RAM. Utilizza la `MainRequestQueuePendingRequests` CloudWatch metrica per determinare l'impatto della prima (un valore maggiore di zero indica un numero di richieste simultanee superiore a quello che è possibile gestire). Utilizza la `BufferCacheHitRatio` CloudWatch metrica per determinare l'impatto della seconda. Un rapporto che scende spesso al di sotto del 99,9 percento indica che non c'è abbastanza RAM per contenere la parte operativa del grafico da valutare, il che si traduce in uno

scambio di cache più frequente. Se la famiglia di istanze R fornisce una concorrenza sufficiente ma non abbastanza RAM, prova a provare la X famiglia di istanze.

I casi d'uso ideali per le istanze serverless sono descritti nella documentazione di Neptune. Se non siete sicuri se il provisioning o il serverless siano la soluzione migliore per voi e il costo è la vostra preoccupazione principale, testate il vostro carico di lavoro in modalità serverless per determinare il numero di applicazioni NCUs utilizzate e confrontate il costo di provisioned () con serverless (). `N hours × hourly provisioned cost sum of NCUs × hourly cost per NCU` Se non si è certi dell'istanza di provisioning di dimensioni equivalenti, una NCU equivale a circa 2 GB di RAM e vCPU e rete associate. Se l'istanza fornita appartiene alla `r6i` famiglia, il rapporto è 1 vCPU per 8 GB di RAM, o NCUs 4, oltre alla rete associata. Il calcolatore dei prezzi di Amazon Neptune fornisce anche un confronto per aiutarti a decidere la configurazione dei costi ottimale.

Quando utilizzi la modalità serverless per le istanze primarie e di replica, ricorda che le repliche di lettura nei livelli di promozione 0 e 1 le ridimensioneranno in base all'istanza di writer, NCUs in modo che vengano ridimensionate correttamente in caso di evento di failover. Imposta i limiti NCU per queste istanze in base a quale delle tue istanze, writer o reader, riceve la maggior parte del traffico.

In ambienti in cui il cluster non è necessario 24 ore al giorno, 7 giorni alla settimana, prendete in considerazione la possibilità di scrivere script che disattivino le istanze di Neptune quando non sono in uso e le riavviino prima di essere utilizzate. Le istanze Neptune verranno riavviate automaticamente ogni 7 giorni per garantire l'applicazione degli aggiornamenti di manutenzione richiesti. Se intendi lasciare le istanze disattivate per lunghi periodi, usa uno script settimanale per chiuderle nuovamente.

Archiviazione e trasferimento dei dati delle dimensioni corrette

Le query più efficienti (ad esempio, le query che devono toccare un minor numero di nodi, bordi e proprietà nel grafico) richiedono meno I/O trasferimenti e possono potenzialmente utilizzare istanze più piccole perché è necessaria una minore cache di buffer. Utilizzate gli endpoint Profile or Explore del vostro linguaggio di query per ottimizzare la query e valutate la possibilità di ottimizzare il modello grafico per le prestazioni delle query.

Neptune utilizza la codifica del dizionario su stringhe di grandi dimensioni e tale dizionario è ottimizzato per le prestazioni, non per l'efficienza. Se disponi di stringhe JSON di grandi dimensioni BLOBs o che cambiano frequentemente per le proprietà, valuta la possibilità di archiviarle all'esterno di Neptune in Amazon S3, Amazon DynamoDB o Amazon DocumentDB e memorizza solo un riferimento all'interno del nodo Neptune.

In alcuni casi, la scelta di un'istanza di dimensioni maggiori può essere più economica. Se i I/O costi sono molto elevati perché bassi`BufferCacheHitRatio`, è possibile che una cache buffer più grande riduca in modo significativo tale costo. Questo perché tutti i dati entrerebbero nella cache anziché essere spesso scambiati dallo storage e incorrere in una velocità di trasferimento elevata. I/O

copy-on-writeNeptune usa la clonazione. Quando si esegue la clonazione per dividere un grafico in più frammenti, potrebbe essere più efficiente non eliminare i dati indesiderati dal cluster clonato, poiché ciò comporterebbe la creazione di nuove pagine di dati, con conseguente aumento dei costi di archiviazione. I dati che non sono stati modificati prima dell'evento di clonazione saranno presenti in un'unica pagina di dati condivisa tra i due cluster e verranno addebitati solo per quella singola copia.

Non attivate l'indice OSGP o utilizzate le istanze R5d a meno che non abbiate effettuato dei test per confermare che fanno una differenza sostanziale nel carico di lavoro. Entrambi sono progettati per scenari che si verificano raramente e potrebbero aumentare i costi con guadagni minimi o nulli.

Pilastro della sostenibilità

Il pilastro della sostenibilità si concentra sulla riduzione al minimo degli impatti ambientali dell'esecuzione di carichi di lavoro cloud. Gli argomenti chiave includono un modello di responsabilità condivisa per la sostenibilità, la comprensione dell'impatto e la massimizzazione dell'uso per ridurre al minimo le risorse necessarie e ridurre gli impatti a valle.

Il pilastro della sostenibilità contiene le seguenti aree di interesse chiave:

- Il tuo impatto
- Obiettivi di sostenibilità
- Utilizzo massimizzato
- Anticipazione e adozione di nuove offerte hardware e software più efficienti
- Utilizzo di servizi gestiti
- Riduzione dell'impatto a valle

Questa guida si concentra sul tuo impatto. Per ulteriori informazioni sugli altri principi di progettazione della sostenibilità, vedere il AWS Well-Architected Framework.

Le tue scelte e i tuoi requisiti hanno un impatto sull'ambiente. Se è possibile scegliere soluzioni Regioni AWS con un'intensità di carbonio inferiore e se i requisiti riflettono le esigenze effettive del carico di lavoro anziché limitarsi a massimizzare l'operatività e la durata, la sostenibilità del carico di lavoro aumenta. Le sezioni successive illustrano le migliori pratiche e considerazioni ponderate che avranno un impatto ambientale positivo se adottate nella progettazione del carico di lavoro e nelle operazioni correnti.

Regione AWS selezione

Alcuni Regioni AWS sono vicini a progetti di energia rinnovabile di Amazon o si trovano dove la rete ha un'intensità di carbonio pubblicata inferiore rispetto ad altre. Considera l'impatto sulla sostenibilità per le regioni che potrebbe essere fattibile per il tuo carico di lavoro e confronta il tuo elenco con le regioni in cui è disponibile Neptune.

Consumo basato sui modelli di comportamento degli utenti

Il corretto dimensionamento dei consumi in base al traffico e al comportamento degli utenti aiuta a AWS ridurre al minimo l'impatto dei servizi sull'ambiente. Prendi in considerazione le seguenti best practice durante la progettazione della tua soluzione:

- Monitora i CloudWatch parametri di Amazon come `CPUUtilization` `MainRequestQueuePendingRequests`, e `TotalRequestsPerSec` per determinare quando la tua domanda è più alta e più bassa, e assicurati che le risorse del cluster siano delle dimensioni corrette in quei periodi.
- Automatizza l'arresto degli ambienti non di produzione durante le ore in cui non vengono utilizzati. Per ulteriori informazioni, consulta il post sul blog Automatizza l'arresto e l'avvio delle risorse dell'ambiente Amazon Neptune utilizzando i tag di risorsa.
- Se i tuoi modelli di traffico variano frequentemente e in modo imprevedibile, prendi in considerazione l'utilizzo di istanze Neptune Serverless che si scalano verso l'alto e verso il basso in base alla domanda invece di utilizzare un'istanza predisposta per i picchi di traffico.
- Valuta la possibilità di allineare i tuoi accordi sui livelli di servizio agli obiettivi di sostenibilità oltre agli obiettivi di continuità aziendale. La semplificazione di requisiti come il disaster recovery in più regioni, l'elevata disponibilità o la conservazione dei backup a lungo termine, in particolare per ambienti non di produzione o carichi di lavoro non mission critical, può ridurre la quantità di risorse necessarie per raggiungere tali obiettivi.

Ottimizza lo sviluppo del software e i modelli di architettura

Per prevenire gli sprechi, ottimizza i modelli e le query e condividi le risorse di calcolo in modo da utilizzare tutte le risorse disponibili nelle istanze e nei cluster Neptune. Le migliori pratiche specifiche includono:

- Chiedi agli sviluppatori di condividere le istanze di Neptune e le istanze dell'applicazione Jupyter Notebook anziché crearne di proprie. Assegna a ogni sviluppatore la propria partizione logica in un singolo cluster Neptune attraverso l'uso di strategie di partizionamento multi-tenancy e crea cartelle notebook separate per ogni sviluppatore su una singola istanza di Jupyter.
- Implementa modelli che massimizzano l'uso delle risorse e riducono al minimo i tempi di inattività, ad esempio thread paralleli per caricare dati e raggruppare i record in una transazione più ampia.

- Ottimizza le query e il modello grafico per ridurre al minimo le risorse necessarie per calcolare i risultati.
- Per i risultati delle query Gremlin, utilizzate la funzione di cache dei risultati per ridurre al minimo le risorse impiegate per ricalcolare le query impaginate o ricorrenti di frequente.
- Mantieni aggiornati i tuoi ambienti Neptune. Le versioni più recenti di Neptune supportano le più recenti istanze Amazon EC2, come Graviton, che sono più efficienti. Sono inoltre stati apportati miglioramenti all'ottimizzazione delle query e correzioni di bug che riducono la quantità di risorse necessarie per calcolare le query.

Resources

Riferimenti

- AWS Well-Architected
- AWS Documentazione Well-Architected Framework
- Ultimi aggiornamenti di Neptune
- Migliori pratiche: ottenere il massimo da Neptune
- Calcolatore dei prezzi di Amazon Neptune

Post del blog

- Test automatizzati dell'accesso ai dati di Amazon Neptune con Apache Gremlin TinkerPop
- Automatizza l'arresto e l'avvio delle risorse dell'ambiente Amazon Neptune utilizzando i tag di risorsa
- Controllo granulare degli accessi per le azioni del piano dati di Amazon Neptune
- Rapporto prezzo/prestazioni di scrittura delle query 4,7 volte migliore con le istanze AWS Graviton4 R8g che utilizzano Amazon Neptune v1.4.5
- In che modo Orca Security ha ottimizzato le prestazioni del database Amazon Neptune
- Crea applicazioni grafiche più velocemente con gli endpoint pubblici di Amazon Neptune
- La nuova versione del motore Amazon Neptune offre un throughput fino a 9 volte più veloce e 10 volte superiore per le prestazioni delle query OpenCypher

AWS Corsi Skill Builder gratuiti

- Guida introduttiva ad Amazon Neptune
- Creazione di applicazioni su Amazon Neptune
- Modellazione dei dati per Amazon Neptune

Collaboratori

I collaboratori di questa guida includono:

- Brian O'Keefe, architetto principale di Neptune Solutions, AWS
- Abhishek Mishra, architetto senior di Neptune Solutions, AWS
- Ganesh Sawhney, responsabile del team - Strategic Partner Success Solutions Architect, AWS
- Michael Havey, architetto senior delle soluzioni Neptune, AWS
- Kevin Phillips, architetto di soluzioni per Neptune, AWS
- Melissa Kwok, architetto di soluzioni per Neptune, AWS
- Sakti Mishra, architetto principale delle soluzioni AWS
- Javed Ali, architetto senior delle soluzioni, 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 della versione di Neptune	Abbiamo aggiornato la documentazione per includere informazioni su Amazon Neptune 1.4.6.0 e versioni successive.	2 gennaio 2026
Pubblicazione iniziale	—	27 settembre 2023

AWS Glossario delle linee guida prescrittive

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

Numeri

7 R

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

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

A

ABAC

Vedi controllo degli accessi basato sugli attributi.

servizi astratti

Vedi servizi gestiti.

ACIDO

Vedi atomicità, consistenza, isolamento, durata.

migrazione attiva-attiva

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

migrazione attiva-passiva

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

funzione di aggregazione

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

Intelligenza artificiale

Vedi intelligenza artificiale.

AIOps

Guarda le operazioni di intelligenza artificiale.

anonimizzazione

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

anti-modello

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

controllo delle applicazioni

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

portfolio di applicazioni

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

intelligenza artificiale (IA)

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

operazioni di intelligenza artificiale (AIOps)

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

crittografia asimmetrica

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

atomicità, consistenza, isolamento, durabilità (ACID)

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

Controllo degli accessi basato su attributi (ABAC)

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

fonte di dati autorevole

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

Zona di disponibilità

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

AWS Cloud Adoption Framework (CAF)AWS

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

AWS Workload Qualification Framework (WQF)AWS

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

B

bot difettoso

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

BCP

Vedi la pianificazione della continuità operativa.

grafico comportamentale

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

sistema big-endian

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

Classificazione binaria

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

filtro Bloom

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

implementazione blu/verde

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

bot

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

botnet

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

ramo

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

accesso break-glass

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

strategia brownfield

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

cache del buffer

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

capacità di business

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

pianificazione della continuità operativa (BCP)

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

C

CAF

Vedi Cloud Adoption AWS Framework.

implementazione canaria

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

CCoE

Vedi Cloud Center of Excellence.

CDC

Vedi Change Data Capture.

Change Data Capture (CDC)

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

ingegneria del caos

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

CI/CD

Vedi integrazione continua e distribuzione continua.

classificazione

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

crittografia lato client

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

Centro di eccellenza cloud (CCoE)

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

cloud computing

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

modello operativo cloud

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

fasi di adozione del cloud

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

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

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

CMDB

Vedi database di gestione della configurazione.

repository di codice

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

cache fredda

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

dati freddi

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

visione artificiale (CV)

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

deriva della configurazione

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

database di gestione della configurazione (CMDB)

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

Pacchetto di conformità

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

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

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

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

CV

Vedi visione artificiale.

D

dati a riposo

Dati stazionari nella rete, ad esempio i dati archiviati.

classificazione dei dati

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

deriva dei dati

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

dati in transito

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

rete di dati

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

riduzione al minimo dei dati

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

perimetro dei dati

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

pre-elaborazione dei dati

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

provenienza dei dati

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

soggetto dei dati

Un individuo i cui dati vengono raccolti ed elaborati.

data warehouse

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

linguaggio di definizione del database (DDL)

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

linguaggio di manipolazione del database (DML)

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

DDL

Vedi linguaggio di definizione del database.

deep ensemble

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

deep learning

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

defense-in-depth

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

amministratore delegato

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

implementazione

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

Ambiente di sviluppo

Vedi ambiente.

controllo di rilevamento

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

mappatura del flusso di valore dello sviluppo (DVSM)

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

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

gemello digitale

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

tabella delle dimensioni

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

disastro

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

disaster recovery (DR)

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

DML

Vedi linguaggio di manipolazione del database.

progettazione basata sul dominio

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

DOTT.

Vedi disaster recovery.

rilevamento della deriva

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

DVSM

Vedi la mappatura del flusso di valore dello sviluppo.

E

EDA

Vedi analisi esplorativa dei dati.

MODIFICA

Vedi scambio elettronico di dati.

edge computing

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

scambio elettronico di dati (EDI)

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

crittografia

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

chiave crittografica

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

endianità

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

endpoint

Vedi service endpoint.

servizio endpoint

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

pianificazione delle risorse aziendali (ERP)

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

crittografia envelope

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

ambiente

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

- ambiente di sviluppo: un'istanza di un'applicazione in esecuzione disponibile solo per il team principale responsabile della manutenzione dell'applicazione. Gli ambienti di sviluppo vengono utilizzati per testare le modifiche prima di promuoverle negli ambienti superiori. Questo tipo di ambiente viene talvolta definito ambiente di test.
- ambienti inferiori: tutti gli ambienti di sviluppo di un'applicazione, ad esempio quelli utilizzati per le build e i test iniziali.

- ambiente di produzione: un'istanza di un'applicazione in esecuzione a cui gli utenti finali possono accedere. In una CI/CD pipeline, l'ambiente di produzione è l'ultimo ambiente di distribuzione.
- ambienti superiori: tutti gli ambienti a cui possono accedere utenti diversi dal team di sviluppo principale. Si può trattare di un ambiente di produzione, ambienti di preproduzione e ambienti per i test di accettazione da parte degli utenti.

epica

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

ERP

Vedi pianificazione delle risorse aziendali.

analisi esplorativa dei dati (EDA)

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

F

tabella dei fatti

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

fallire velocemente

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

limite di isolamento dei guasti

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

ramo di funzionalità

Vedi filiale.

caratteristiche

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

importanza delle caratteristiche

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

trasformazione delle funzionalità

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

prompt con pochi scatti

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

FGAC

Vedi il controllo granulare degli accessi.

controllo granulare degli accessi (FGAC)

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

migrazione flash-cut

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

FM

Vedi modello di base.

modello di fondazione (FM)

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

G

IA generativa

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

blocco geografico

Vedi restrizioni geografiche.

limitazioni geografiche (blocco geografico)

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

Flusso di lavoro di GitFlow

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

immagine dorata

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

strategia greenfield

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

guardrail

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

H

AH

Vedi disponibilità elevata.

migrazione di database eterogenea

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

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

alta disponibilità (HA)

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

modernizzazione storica

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

dati di blocco

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

migrazione di database omogenea

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

dati caldi

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

hotfix

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

periodo di hypercare

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

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

I

IaC

Vedi l'infrastruttura come codice.

Policy basata su identità

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

applicazione inattiva

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

IIoT

Vedi Industrial Internet of Things.

infrastruttura immutabile

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

VPC in ingresso (ingress)

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

migrazione incrementale

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

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

Industria 4.0

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

infrastruttura

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

infrastruttura come codice (IaC)

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

IIoInternet delle cose industriale (T)

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

VPC di ispezione

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

Internet of Things (IoT)

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

interpretabilità

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

IoT

Vedi Internet of Things.

libreria di informazioni IT (ITIL)

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

gestione dei servizi IT (ITSM)

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

ITIL

Vedi la libreria di informazioni IT.

ITSM

Vedi Gestione dei servizi IT.

L

controllo degli accessi basato su etichette (LBAC)

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

zona di destinazione

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

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

modello linguistico di grandi dimensioni (LLM)

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

migrazione su larga scala

Una migrazione di 300 o più server.

BIANCO

Vedi controllo degli accessi basato su etichette.

Privilegio minimo

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

eseguire il rehosting (lift and shift)

Vedi 7 R.

sistema little-endian

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

LLM

Vedi modello linguistico di grandi dimensioni.

ambienti inferiori

Vedi ambiente.

M

machine learning (ML)

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

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

ramo principale

Vedi filiale.

malware

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

servizi gestiti

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

sistema di esecuzione della produzione (MES)

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

MAP

Vedi Migration Acceleration Program.

meccanismo

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

account membro

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

MEH

Vedi sistema di esecuzione della produzione.

Message Queuing Telemetry Transport (MQTT)

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

microservizio

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

architettura di microservizi

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

Programma di accelerazione della migrazione (MAP)

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

migrazione su larga scala

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

fabbrica di migrazione

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

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

metadati di migrazione

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

modello di migrazione

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

Valutazione del portfolio di migrazione (MPA)

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

valutazione della preparazione alla migrazione (MRA)

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

strategia di migrazione

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

ML

Vedi machine learning.

modernizzazione

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

valutazione della preparazione alla modernizzazione

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

applicazioni monolitiche (monoliti)

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

MAPPA

Vedi Migration Portfolio Assessment.

MQTT

Vedi Message Queuing Telemetry Transport.

classificazione multiclasse

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

infrastruttura mutabile

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

O

OAC

Vedi Origin Access Control.

QUERCIA

Vedi Origin Access Identity.

OCM

Vedi gestione delle modifiche organizzative.

migrazione offline

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

OI

Vedi l'integrazione delle operazioni.

OLA

Vedi accordo a livello operativo.

migrazione online

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

OPC-UA

Vedi Open Process Communications - Unified Architecture.

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

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

accordo a livello operativo (OLA)

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

revisione della prontezza operativa (ORR)

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

tecnologia operativa (OT)

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

integrazione delle operazioni (OI)

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

trail organizzativo

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

gestione del cambiamento organizzativo (OCM)

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

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

controllo dell'accesso all'origine (OAC)

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

identità di accesso origine (OAI)

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

ORR

Vedi la revisione della prontezza operativa.

NON

Vedi la tecnologia operativa.

VPC in uscita (egress)

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

P

limite delle autorizzazioni

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

informazioni di identificazione personale (PII)

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

Informazioni che consentono l'identificazione personale degli utenti

Visualizza le informazioni di identificazione personale.

playbook

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

PLC

Vedi controllore logico programmabile.

PLM

Vedi la gestione del ciclo di vita del prodotto.

policy

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

persistenza poliglotta

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

valutazione del portfolio

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

predicate

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

predicato pushdown

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

controllo preventivo

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

principale

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

privacy fin dalla progettazione

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

zone ospitate private

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

controllo proattivo

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

gestione del ciclo di vita del prodotto (PLM)

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

Ambiente di produzione

Vedi ambiente.

controllore logico programmabile (PLC)

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

concatenamento rapido

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

pseudonimizzazione

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

publish/subscribe (pub/sub)

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

Q

Piano di query

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

regressione del piano di query

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

R

Matrice RACI

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

RAG

Vedi Retrieval Augmented Generation.

ransomware

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

Matrice RASCI

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

RCAC

Vedi controllo dell'accesso a righe e colonne.

replica di lettura

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

riprogettare

Vedi 7 Rs.

obiettivo del punto di ripristino (RPO)

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

obiettivo del tempo di ripristino (RTO)

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

rifattorizzare

Vedi 7 R.

Region

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

regressione

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

riospitare

Vedi 7 R.

rilascio

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

trasferisco

Vedi 7 Rs.

ripiattaforma

Vedi 7 Rs.

riacquisto

Vedi 7 Rs.

resilienza

La capacità di un'applicazione di resistere alle interruzioni o di ripristinarle. L'elevata disponibilità e il disaster recovery sono considerazioni comuni quando si pianifica la resilienza in. Cloud AWSPer ulteriori informazioni, vedere Cloud AWS Resilience.

policy basata su risorse

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

matrice di assegnazione di responsabilità (RACI)

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

controllo reattivo

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

retain

Vedi 7 R.

andare in pensione

Vedi 7 Rs.

Retrieval Augmented Generation (RAG)

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

rotazione

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

controllo dell'accesso a righe e colonne (RCAC)

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

RPO

Vedi obiettivo del punto di ripristino.

VERSO

Vedi obiettivo del tempo di ripristino.

runbook

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

S

SAML 2.0

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

SCADA

Vedi controllo di supervisione e acquisizione dati.

SCP

Vedi la politica di controllo del servizio.

Secret

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

sicurezza fin dalla progettazione

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

controllo di sicurezza

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

rafforzamento della sicurezza

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

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

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

automazione della risposta alla sicurezza

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

Crittografia lato server

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

Policy di controllo dei servizi (SCP)

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

endpoint del servizio

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

accordo sul livello di servizio (SLA)

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

indicatore del livello di servizio (SLI)

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

obiettivo a livello di servizio (SLO)

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

Modello di responsabilità condivisa

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

SIEM

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

punto di errore singolo (SPOF)

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

SLAM

Vedi il contratto sul livello di servizio.

SLI

Vedi l'indicatore del livello di servizio.

LENTA

Vedi obiettivo del livello di servizio.

split-and-seed modello

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

SPOF

Vedi punto di errore singolo.

schema a stella

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

modello del fico strangolatore

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

sottorete

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

controllo di supervisione e acquisizione dati (SCADA)

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

crittografia simmetrica

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

test sintetici

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

prompt di sistema

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

T

tag

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

variabile di destinazione

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

elenco di attività

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

ambiente di test

Vedi ambiente.

training

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

Transit Gateway

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

flusso di lavoro basato su trunk

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

Accesso attendibile

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

regolazione

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

team da due pizze

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

U

incertezza

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

compiti indifferenziati

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

ambienti superiori

Vedi ambiente.

V

vacuum

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

controllo delle versioni

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

Peering VPC

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

vulnerabilità

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

W

cache calda

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

dati caldi

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

funzione finestra

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

Carico di lavoro

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

flusso di lavoro

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

VERME

Vedi scrivere una volta, leggere molti.

WQF

Vedi AWS Workload Qualification Framework.

scrivi una volta, leggi molte (WORM)

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

Z

exploit zero-day

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

vulnerabilità zero-day

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

prompt zero-shot

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

applicazione zombie

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

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