Unable to locate subtitle

Framework AWS Well-Architected

Framework AWS Well-Architected: ***Unable to locate subtitle***

Table of Contents

Framework AWS Well-Architected

Data di pubblicazione: 20 ottobre 2022 (Revisioni del documento)

Il Framework AWS Well-Architected aiuta a comprendere i pro e i contro delle decisioni prese durante la progettazione di sistemi in AWS. Utilizzando il Canone, scoprirai le best practice architetturali per progettare e gestire sistemi affidabili, sicuri, efficienti e convenienti nel cloud.

Introduzione

Il Framework AWS Well-Architected aiuta a comprendere i pro e i contro delle decisioni prese durante la progettazione di sistemi in AWS. Utilizzando il Framework, scoprirai le best practice architetturali per progettare e gestire carichi di lavoro affidabili, sicuri, efficienti, convenienti e sostenibili nel Cloud AWS. Permette di misurare in modo coerente le architetture rispetto alle best practice e identificare le aree da migliorare. Il processo di revisione di un'architettura consiste in una conversazione costruttiva sulle decisioni relative all'architettura e non è un meccanismo di audit. Disporre di sistemi ben architettati aumenta notevolmente la probabilità di successo aziendale.

AWS Solutions Architects vanta anni di esperienza nell'architettura di soluzioni in un'ampia gamma di business e casi di utilizzo. Abbiamo supportato migliaia di clienti nella progettazione e revisione delle loro architetture su AWS. Grazie a questa esperienza, abbiamo identificato best practice e strategie principali per i sistemi di architettura nel cloud.

Il Framework AWS Well-Architected documenta un insieme di domande fondamentali per capire se un'architettura specifica si allinea bene con le best practice del cloud. Il canone fornisce un approccio coerente per la valutazione dei sistemi rispetto alle qualità che ti aspetti da sistemi basati sul cloud moderni e i rimedi necessari per raggiungere tali qualità. Man mano che AWS continua a evolversi e noi continuiamo a imparare di più dal lavoro che svolgiamo con i nostri clienti, continueremo a ridefinire la definizione di canone di architettura.

Questo canone è rivolto a chi svolge ruoli tecnologici, ad esempio ai Chief Technology Officer (CTO), ai progettisti, agli sviluppatori e ai membri dei team operativi. Descrive le best practice e le strategie AWS da usare per la progettazione e il funzionamento di un carico di lavoro cloud, e fornisce collegamenti a ulteriori dettagli di implementazione e pattern architetturali. Per ulteriori informazioni, consulta la Homepage del Canone di architettura AWS.

AWS offre anche un servizio gratuito di revisione dei carichi di lavoro. Il AWS Well-Architected Tool (AWS WA Tool) è un servizio cloud che fornisce un approccio coerente per la revisione

e la valutazione della tua architettura secondo il Framework AWS Well-Architected. AWS WA Tool fornisce raccomandazioni per rendere i tuoi carichi di lavoro più affidabili, sicuri, efficienti e convenienti.

Per aiutarti ad applicare le best practice, abbiamo creato AWS Well-Architected Labs, che fornisce un repository di codice e documentazione per un'esperienza concreta di implementazione delle best practice. Abbiamo anche collaborato con partner APN (AWS Partner Network) selezionati, che sono membri del Programma Partner AWS Well-Architected. Tali partner AWS vantano una conoscenza approfondita di AWS e possono aiutarti nella revisione e nel miglioramento dei tuoi carichi di lavoro.

Definizioni

Tutti i giorni, gli esperti AWS supportano i clienti nella progettazione di sistemi di architettura per sfruttare le best practice nel cloud. Ti aiutiamo a trovare i compromessi relativi all'architettura nel processo di evoluzione dei tuoi progetti. Quando distribuisci questi sistemi in ambienti live, analizziamo le prestazioni di questi sistemi e le conseguenze dei suddetti compromessi.

Sulla base di quello che abbiamo imparato, abbiamo creato il Framework AWS Well-Architected, che fornisce a clienti e partner un insieme coerente di best practice per valutare le architetture, e comprende un insieme di domande che puoi utilizzare per valutare se la tua architettura è ben allineata alle best practice AWS.

Il Framework AWS Well-Architected si basa su sei pilastri: eccellenza operativa, sicurezza, affidabilità, efficienza delle prestazioni, ottimizzazione dei costi e sostenibilità.

Tabella 1. I pilastri del Framework AWS Well-Architected

Nome	Descrizione
Eccellenza operativa	Comprende la capacità di supportare lo sviluppo ed eseguire carichi di lavoro in modo efficace, ottenere informazioni approfondite sulle loro operazioni e migliorare continuam ente i processi e le procedure di supporto per offrire valore aggiunto.
Sicurezza	Il pilastro della sicurezza descrive come sfruttare le tecnologie cloud per proteggere

Nome	Descrizione
	dati, sistemi e asset in modo da migliorare la sicurezza.
Affidabilità	Il principio dell'affidabilità comprende la capacità di un carico di lavoro di eseguire la funzione attesa in modo corretto e coerente quando previsto. Include la possibilità di utilizzare e testare il carico di lavoro per tutto il ciclo di vita. Questo documento fornisce linee guida dettagliate sulle best practice per l'implementazione di carichi di lavoro affidabili in AWS.
Efficienza delle prestazioni	L'abilità di utilizzare in modo efficiente le risorse di elaborazione per soddisfare i requisiti del sistema e conservare tale efficienza a seconda dei cambiamenti della domanda e dell'evol uzione delle tecnologie.
Ottimizzazione dei costi	La capacità di eseguire sistemi per fornire valore aziendale al minor prezzo possibile.
Sostenibilità	La capacità di migliorare continuamente l'impatto sulla sostenibilità riducendo il consumo energetico e aumentando l'efficie nza di tutti i componenti di un carico di lavoro, massimizzando i benefici delle risorse di cui è stato eseguito il provisioning e riducendo al minimo le risorse totali richieste.

Nel Framework AWS Well-Architected, si utilizzano i seguenti termini:

- Un componente è un codice, una configurazione e delle risorse AWS che insieme riescono a soddisfare un requisito di un carico di lavoro. Spesso un componente è l'unità di proprietà tecnica ed è disaccoppiato da altri componenti.

- Con il termine carico di lavoro ci riferiamo all'insieme di componenti che forniscono valore aziendale. Un carico di lavoro, normalmente, è il livello di dettaglio comunicato dai leader aziendali e della tecnologia.
- Pensiamo a un' architettura come al modo in cui in componenti operano insieme in un carico di lavoro. Il modo di comunicare e di interagire dei componenti è spesso l'aspetto principale dei diagrammi architetturali.
- Le tappe fondamentali indicano cambiamenti chiave della tua architettura man mano che si evolve nel corso del ciclo di vita del prodotto (progettazione, test, messa online e produzione).
- Nell'ambito di un'organizzazione il portfolio delle tecnologie rappresenta l'insieme di carichi di lavoro necessari affinché l'azienda possa essere operativa.
- Il livello di impegno è la categorizzazione della quantità di tempo, sforzo e complessità che un'attività richiede per la sua realizzazione. Ogni organizzazione deve considerare le dimensioni e le competenze del team e la complessità del carico di lavoro per ottenere un contesto aggiuntivo che consenta di classificare correttamente il livello di impegno.
  - Alto: Il lavoro potrebbe richiedere più settimane o più mesi. Potrebbe essere suddiviso in molteplici fasi, rilasci e attività.
  - Medio: Il lavoro potrebbe richiedere più giorni o settimane. Potrebbe essere suddiviso in molteplici rilasci e attività.
  - Basso: Il lavoro potrebbe richiedere più ore o giorni. Potrebbe essere suddiviso in molteplici attività.

Quando progetti l'architettura dei carichi di lavoro, devi trovare dei compromessi tra i pilastri su cui si regge il tuo contesto aziendale. Questo tipo di decisioni aziendali deve essere alla base delle tue priorità ingegneristiche. Potresti ottimizzare per migliorare la sostenibilità e ridurre i costi a spese dell'affidabilità in ambienti di sviluppo oppure, per quanto riguarda le soluzioni mission-critical, potresti ottimizzare l'affidabilità a fronte di costi più elevati e di un impatto ambientale maggiore. Nelle soluzioni di e-commerce, le prestazioni possono avere un impatto sui profitti e sulla propensione all'acquisto da parte dei clienti. L'eccellenza in ambito di sicurezza e operatività generalmente non viene sacrificata rispetto agli altri pilastri.

Architettura

Negli ambienti in locale, i clienti spesso hanno un team centrale per l'architettura delle tecnologie che funziona da livello superiore per altri team di prodotto o funzionalità, al fine di garantire che i team rispettino le best practice. I team dell'architettura delle tecnologie spesso sono composti da diversi

ruoli come il Technical Architect (infrastruttura), il Solutions Architect (software), il Data Architect, il Networking Architect e il Security Architect. Spesso i team usano TOGAF o framework di Zachman come parte delle competenze architetturali aziendali.

Noi di AWS preferiamo distribuire le competenze tra i team, invece di centralizzarle in un unico team. Quando si sceglie di distribuire il potere decisionale si corrono dei rischi, ad esempio il rischio di garantire che i team interni rispettino gli standard. Noi mitighiamo questo rischi in due modi. Innanzitutto, abbiamo le pratiche (Modalità per eseguire attività, processi, standard e norme accettate) che hanno lo scopo di permettere a ogni team di possedere tali competenze e ci serviamo di esperti che garantiscano che i team adottino standard più severi di quelli che devono rispettare. In secondo luogo, implementiamo meccanismi che eseguono controlli automatizzati per garantire che gli standard vengano rispettati.

> "Le buone intenzioni non bastano mai, per avere successo servono buoni meccanismi", Jeff Bezos.

Questo significa sostituire gli sforzi di una persona con meccanismi (spesso automatizzati) che verificano la conformità alle regole e ai processi. L'approccio distribuito è supportato dai principi di leadership di Amazone stabilisce una cultura tra tutti i ruoli che lavora a ritroso dal cliente. Il lavoro a ritroso è una parte fondamentale del nostro processo di innovazione. Partiamo dal cliente e da quello che vuole e sulla base di questo definiamo e indirizziamo i nostri sforzi. I team che mettono il cliente al centro sviluppano prodotti sulla base delle necessità del cliente.

Per l'architettura questo significa che ci aspettiamo che ogni team sia in grado di creare architetture e di seguire le best practice. Per aiutare i nuovi team ad acquisire queste competenze o i team esistenti ad alzare il livello, abilitiamo l'accesso a una community virtuale di ingegneri responsabili che possono eseguire la revisione dei loro progetti e aiutarli a comprendere le best practice di AWS. La community di ingegneri responsabili lavora per rendere visibili e accessibili le best practice. Uno dei modi per fare ciò, ad esempio, è servirsi delle lunchtime talk che si concentrano sull'applicazione di best practice a esempi reali. Le lunchtime talk sono registrate e possono essere utilizzate come materiale di onboarding per i nuovi membri del team.

Le best practice AWS sono il risultato della nostra esperienza nell'esecuzione di migliaia di sistemi su Internet. Preferiamo utilizzare i dati per definire le best practice, ma ci serviamo anche di esperti in materia, come i capo ingegneri. Quando i capo ingegneri vedono emergere nuove best practice, lavorano con la community per garantire che i team le rispettino. Con il tempo, queste best practice

vengono formalizzate nei nostri processi di revisione interna e nei meccanismi che rafforzano la compliance. Il Canone di architettura è l'implementazione del nostro processo di revisione interno rivolta ai clienti, in cui abbiamo codificato la nostra idea di ingegneria responsabile attraverso ruoli di campo come Solutions Architect e i team di ingegneria interni. Il canone di architettura è un meccanismo scalabile che consente di trarre vantaggio da questi insegnamenti.

Seguendo l'approccio della community di ingegneri responsabili con la proprietà distribuita dell'architettura, riteniamo che si possa ottenere un'architettura aziendale Well-Architected che si basa sulle necessità del cliente. I leader della tecnologia (come i CTO o i manager dello sviluppo) che eseguono revisioni Well-Architected tra tutti i carichi di lavoro ti permettono di comprendere più a fondo i rischi relativi al portfolio delle tecnologie. Tramite questo approccio puoi identificare dei temi tra i team che la tua organizzazione può affrontare tramite meccanismi, formazione o dialoghi informali in cui i capo ingegneri possono condividere le loro idee su aree specifiche con diversi team.

Principi generali di progettazione

Il Canone di architettura identifica una serie di principi generali per facilitare la corretta progettazione nel cloud:

- Smetti di ipotizzare quali siano le tue esigenze di capacità: quando prendi decisioni relative alla capacità prima della distribuzione di un sistema, potresti ritrovarti con risorse inattive o ad affrontare le conseguenze della capacità limitata. Con il cloud computing, questi problemi vengono risolti. Puoi utilizzare la capacità di cui hai bisogno e ridimensionare il sistema automaticamente.
- Esegui test dei sistemi su scala produttiva: nel cloud, puoi creare un ambiente di test su scala produttiva on demand, completare i test e ritirare le risorse. Poiché paghi per l'ambiente di test solo quando è in esecuzione, puoi simulare un ambiente live a un costo notevolmente inferiore rispetti ai test in locale.
- Automatizza per facilitare la sperimentazione dell'architettura: l'automazione ti permette di creare e replicare i tuoi carichi di lavoro a basso costo e di evitare le spese del lavoro manuale. Puoi tenere traccia delle modifiche all'automazione, effettuare l'audit dell'impatto e tornare ai parametri precedenti, se necessario.
- Consenti le architetture evoluzionistiche: in un ambiente tradizionale, le decisioni relative all'architettura spesso sono implementate come eventi singoli e statici, con poche versioni principali di un sistema durante il ciclo di vita. Alla luce del continuo cambiamento di un'azienda e del suo contesto, le decisioni iniziali potrebbero ostacolare la capacità del sistema di soddisfare i requisiti aziendali in evoluzione. All'interno del cloud, la capacità di automatizzare e testare on demand diminuisce il rischio di impatto dovuto alle modifiche della progettazione. Questo permette ai

sistemi di evolversi nel tempo, in modo che le aziende possano trarre vantaggio dalle innovazioni come pratica standard.

- Promuovi le architetture servendoti dei dati: nel cloud puoi raccogliere dati relativi all'impatto delle tue scelte architetturali sul comportamento del tuo carico di lavoro. Questo ti permette di prendere decisioni basate sui fatti su come migliorare il carico di lavoro. La tua infrastruttura cloud è un codice, quindi, puoi usare tali dati a vantaggio delle scelte e dei miglioramenti relativi all'architettura nel tempo.
- Migliora con le giornate di gioco: testa le prestazioni dell'architettura e dei processi pianificando regolarmente game day per simulare eventi della produzione. Questi ti aiuta a capire dove puoi apportare dei miglioramenti e ti può aiutare a sviluppare un'esperienza organizzativa nella gestione degli eventi.

I pilastri del framework

La creazione di un sistema software è molto simile alla costruzione di un edificio. Se le fondamenta non sono solide, possono emergere problemi strutturali che minano l'integrità e la funzionalità dell'edificio. Se nella creazione dell'architettura per soluzioni tecnologiche trascuri i sei principi di eccellenza operativa, sicurezza, affidabilità, efficienza delle prestazioni, ottimizzazione dei costi e sostenibilità, può diventare complicato sviluppare un sistema che soddisfi le tue aspettative e i tuoi requisiti. L'aggiunta di questi pilastri alla tua architettura ti aiuterà a produrre sistemi efficienti e stabili. Questo ti permetterà di concentrarti su altri aspetti della progettazione, come i requisiti funzionali.

Pilastri

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

Eccellenza operativa

Il principio dell'eccellenza operativa comprende la capacità di supportare lo sviluppo ed eseguire carichi di lavoro in modo efficace, ottenere informazioni approfondite sulle loro operazioni e migliorare continuamente i processi e le procedure di supporto per offrire valore aggiunto.

Il principio dell'eccellenza operativa offre una panoramica dei principi di progettazione, delle best practice e delle domande. È possibile trovare linee guida prescrittive sull'implementazione nel Whitepaper sul principio dell'eccellenza operativa.

Argomenti

- Principi di progettazione
- Definizione
- Best practice
- Risorse

Principi di progettazione

Esistono cinque principi di progettazione per l'eccellenza operativa nel cloud:

- Esegui le operazioni come codice: nel cloud ti è possibile applicare la medesima disciplina di progettazione che utilizzi per il codice dell'applicazione a tutto il tuo ambiente. Puoi definire l'intero carico di lavoro (applicazioni, infrastruttura) come codice e aggiornarlo con il codice. Puoi implementare le tue procedure operative come codice e automatizzarne l'esecuzione attivandole in risposta agli eventi. Eseguendo le operazioni come codice, limiti gli errori umani e attivi risposte coerenti agli eventi.
- Applica modifiche frequenti, minime e reversibili: progetta i carichi di lavoro per fare in modo che i componenti siano aggiornati regolarmente. Apporta modifiche in incrementi ridotti che possono essere annullati se presentano errori (senza comportare conseguenze per i clienti, ove possibile).
- Perfeziona frequentemente le procedure operative: se usi procedure operative, cerca delle opportunità per migliorarle. Man mano che il tuo carico di lavoro si evolve, garantisci anche l'evoluzione adeguata delle tue procedure. Organizza delle simulazioni regolari per verificare e accertarti che tutte le procedure siano efficaci e che i team le conoscano adeguatamente.
- Prevedi gli insuccessi: esegui un'analisi prefallimentare per individuare le potenziali cause di errore in modo da eliminarle o mitigarle. Testa gli scenari di errore e convalida la tua comprensione relativamente al loro impatto. Testa le tue procedure di risposta per assicurarti che siano efficaci e che i team ne conoscano l'esecuzione. Organizza delle simulazioni regolari per testare i carichi di lavoro e le risposte dei team agli eventi simulati.
- Impara da tutti gli insuccessi operativi: favorisci il miglioramento tramite le lezioni apprese da tutti gli eventi e gli errori operativi. Condividi ciò che hai imparato con i vari team e con tutta l'organizzazione.

Definizione

Esistono quattro aree di best practice per l'eccellenza operativa nel cloud:

- Organizzazione
- Preparazione
- Operatività
- Evoluzione

La leadership dell'organizzazione definisce gli obiettivi aziendali. La tua organizzazione deve comprendere i requisiti e le priorità e utilizzarli per organizzare e condurre attività a supporto del raggiungimento dei risultati aziendali. Il carico di lavoro deve generare le informazioni necessarie per supportarlo. L'implementazione di servizi per consentire l'integrazione, la distribuzione e la consegna del carico di lavoro consentirà un flusso maggiore di modifiche vantaggiose in fase di produzione attraverso l'automazione dei processi ripetitivi.

Potrebbero esserci rischi inerenti al funzionamento del carico di lavoro. Devi comprendere questi rischi e prendere una decisione consapevole prima di passare alla fase di produzione. I team devono essere in grado di supportare il carico di lavoro. I parametri aziendali e operativi derivati dai risultati aziendali desiderati ti permetteranno di comprendere lo stato del carico di lavoro e le attività operative e di rispondere agli incidenti. Le priorità cambieranno di pari passo con l'evoluzione delle esigenze aziendali e dell'ambiente aziendale. Utilizza questi aspetti come ciclo di feedback per apportare continui miglioramenti all'organizzazione e alle operazioni legate al carico di lavoro.

Best practice

Argomenti

Organizzazione

È necessario che i team abbiano una comprensione condivisa dell'intero carico di lavoro, del ruolo che vi svolgono, nonché degli obiettivi aziendali condivisi. In questo modo potranno stabilire le priorità che possono favorire il successo aziendale. Un'adeguata definizione delle priorità massimizzerà i risultati dei tuoi sforzi. Valuta le esigenze dei clienti interni ed esterni coinvolgendo i principali stakeholder, compresi i team aziendali, di sviluppo e operativi, per stabilire dove concentrare le attività operative. Valutando le esigenze dei clienti otterrai una conoscenza approfondita del supporto necessario per raggiungere i risultati aziendali. Assicurati di essere a conoscenza delle linee guida o degli obblighi definiti dalla governance organizzativa e di fattori esterni, come i requisiti di conformità normativa e gli standard di settore, che possono imporre o sottolineare un'attenzione specifica. Accertati di disporre di meccanismi per identificare le modifiche ai requisiti di governance interna e di conformità esterni. Se non vengono identificati requisiti, assicurati che sia stata applicata la dovuta

diligenza per giungere a questa conclusione. Rivedi regolarmente le tue priorità in modo che possano essere aggiornate al mutare delle esigenze.

Valuta le minacce per il business (ad esempio rischi e responsabilità aziendali e minacce alla sicurezza delle informazioni) e conserva queste informazioni in un registro dei rischi. Valuta l'impatto dei rischi e dei compromessi tra interessi concorrenti o approcci alternativi. Ad esempio, accelerare l'introduzione sul mercato di nuove funzionalità può essere preferibile all'ottimizzazione dei costi. Oppure, è possibile scegliere un database relazionale per i dati non relazionali per semplificare l'iniziativa di migrazione di un sistema senza refactoring. Gestisci i vantaggi e i rischi per prendere decisioni informate nel determinare dove concentrare gli sforzi. Alcuni rischi o scelte possono essere accettabili per un certo periodo di tempo, potrebbe essere possibile ridurre i rischi associati o la presenza di un rischio potrebbe diventare inaccettabile, nel qual caso si intraprenderà un'azione per risolverlo.

I tuoi team devono comprendere quale contributo offrono nel raggiungimento dei risultati aziendali. I team devono avere obiettivi condivisi e devono comprendere il proprio ruolo nel successo degli altri team. Comprendere la responsabilità, la proprietà, il modo in cui vengono prese le decisioni e chi ha l'autorità decisionale aiuterà a concentrare gli sforzi e a ottimizzare i contributi dei team. Le esigenze di un team sono influenzate dal cliente supportato, dall'organizzazione, dalla composizione del team e dalle caratteristiche del carico di lavoro. Non è ragionevole aspettarsi che un singolo modello operativo sia in grado di supportare tutti i team e i relativi carichi di lavoro dell'organizzazione.

Assicurati che siano identificati i proprietari di ogni applicazione, carico di lavoro, piattaforma e componente dell'infrastruttura e che per ogni processo e procedura sia identificato un proprietario responsabile della sua definizione e dei proprietari responsabili delle loro prestazioni.

La comprensione del valore aziendale di ogni componente, processo e procedura, del motivo per cui tali risorse sono presenti o le attività vengono eseguite e del perché tale proprietà esiste indirizzerà le azioni dei membri del team. Definisci chiaramente le responsabilità dei membri del team in modo che possano agire in modo appropriato e disporre di meccanismi per identificare responsabilità e proprietà. Implementa meccanismi per richiedere aggiunte, modifiche ed eccezioni in modo da non porre limiti all'innovazione. Definisci gli accordi tra i team che descrivono il modo in cui collaborano per supportarsi reciprocamente e contribuire ai risultati aziendali.

Fornisci supporto ai membri del team in modo che possano essere più efficaci nell'azione e nel supporto dei risultati aziendali. La leadership aziendale di alto livello deve stabilire le aspettative e misurare il successo. Gli alti dirigenti sono promotori, sostenitori e motori per l'adozione delle best practice e l'evoluzione dell'organizzazione. Consenti ai membri del team di intervenire quando i risultati sono a rischio per ridurre al minimo l'impatto e incoraggiali a rivolgersi ai responsabili

decisionali e alle parti interessate quando ritengono che esista un rischio, in modo da poterlo risolvere e prevenire gli incidenti. Fornisci comunicazioni tempestive, chiare e concrete dei rischi noti e degli eventi pianificati in modo che i membri del team possano agire in modo tempestivo e appropriato.

Incoraggia la sperimentazione per accelerare l'apprendimento e mantenere i membri del team interessati e coinvolti. I team devono aumentare le proprie competenze per adottare nuove tecnologie e supportare i cambiamenti della domanda e delle responsabilità. Fornisci il tuo supporto e incoraggiamento offrendo tempo strutturato dedicato per l'apprendimento. Assicurati che i membri del team dispongano delle risorse, in termini sia di strumenti sia di membri del team, per avere successo e adattarsi, sostenendo i risultati aziendali. Sfrutta la diversità tra organizzazioni per cercare più prospettive uniche. Usa questa prospettiva per incrementare l'innovazione, mettere in discussione le tue ipotesi e ridurre il rischio di conferme parziali. Aumenta l'inclusione, la diversità e l'accessibilità all'interno dei team per ottenere prospettive vantaggiose.

Se esistono requisiti normativi o di conformità esterni che si applicano alla tua organizzazione, devi utilizzare le risorse fornite da Conformità di AWS Cloud per aiutarti a istruire i tuoi team in modo che possano determinare l'impatto sulle tue priorità. Il Canone di architettura enfatizza l'apprendimento, la misurazione e il miglioramento. Fornisce una strategia coerente per la valutazione delle architetture e l'implementazione di progetti in grado di ridimensionarsi nel corso del tempo. AWS mette a disposizione AWS Well-Architected Tool per aiutarti ad analizzare il tuo approccio prima dello sviluppo e lo stato dei tuoi carichi di lavoro prima e durante la fase di produzione. Puoi confrontare i carichi di lavoro con le best practice architetturali AWS più recenti, monitorarne lo stato generale e ottenere informazioni sui potenziali rischi. AWS Trusted Advisor è uno strumento che fornisce l'accesso a una serie di controlli di base che propongono ottimizzazioni utili per la definizione delle tue priorità. I clienti del supporto Business ed Enterprise hanno accesso a ulteriori controlli a livello di sicurezza, affidabilità, prestazioni e ottimizzazione dei costi che possono essere utili per definire le loro priorità.

AWS può aiutarti a istruire i tuoi team su AWS e i suoi servizi, affinché comprendano meglio in che modo le loro scelte possono influire sul carico di lavoro. Per istruire i tuoi team, è consigliabile utilizzare le risorse fornite da AWS Support (AWS Knowledge Center, AWS Discussion Forms e AWS Support Center) e la documentazione AWS. Raggiungi AWS Support attraverso il AWS Support Center per assistenza sulle tue domande su AWS. AWS condivide inoltre le best practice e i modelli appresi attraverso la gestione di AWS nella Amazon Builders' Library. Un'ampia gamma di altre informazioni utili è disponibile tramite il blog AWS e il podcast ufficiale di AWS. AWS Training and Certification offre risorse di formazione gratuite tramite corsi digitali gestiti dall'utente sulle nozioni di base di AWS. Per supportare ulteriormente lo sviluppo delle competenze AWS del tuo team, è anche possibile iscriversi a corsi di formazione con istruttore.

Per facilitare la gestione dei modelli operativi, è consigliabile utilizzare strumenti o servizi che consentano di gestire centralmente gli ambienti su più account, ad esempio AWS Organizations. Servizi come AWS Control Tower ampliano questa funzionalità di gestione consentendoti di definire piani (a supporto dei tuoi modelli operativi) per configurare gli account, applicare la governance continua tramite AWS Organizations e automatizzare il provisioning di nuovi account. I fornitori di servizi gestiti, come AWS Managed Services, AWS Managed Services Partners o i fornitori di servizi gestiti della AWS Partner Network offrono esperienza nell'implementazione di ambienti cloud e supportano i requisiti di sicurezza e conformità e gli obiettivi aziendali. L'aggiunta di servizi gestiti al tuo modello operativo ti consente di risparmiare tempo e risorse e ti permette di mantenere i team interni snelli e focalizzati sui risultati strategici che differenzieranno la tua attività, anziché sullo sviluppo di nuove competenze e funzionalità.

Le seguenti domande si concentrano su queste considerazioni relative all'eccellenza operativa. (Per l'elenco completo delle domande e delle best practice relative all'eccellenza operativa, consulta l' Appendice.).

OPS 1 In che modo stabilisci quali sono le tue priorità?

È necessario che ognuno capisca il proprio ruolo per rendere possibile il successo aziendale. Devi disporre di obiettivi comuni al fine di stabilire le priorità per le risorse. Ciò massimizzerà i risultati dei tuoi sforzi.

OPS 2 In che modo strutturi la tua organizzazione per supportare i risultati aziendali?

I tuoi team devono comprendere quale contributo offrono nel raggiungimento dei risultati aziendali . I team devono avere obiettivi condivisi e devono comprendere il proprio ruolo nel successo degli altri team. Comprendere la responsabilità, la proprietà, il modo in cui vengono prese le decisioni e chi ha l'autorità decisionale aiuterà a concentrare gli sforzi e a ottimizzare i contributi dei team.

OPS 3 In che modo la cultura aziendale supporta i risultati aziendali?

Fornisci supporto ai membri del team in modo che possano essere più efficaci nell'azione e nel supporto dei risultati aziendali.

Ad esempio, a un certo punto potresti realizzare che desideri dare maggiore risalto a un piccolo sottoinsieme delle tue priorità. Utilizza un approccio equilibrato nel lungo termine per garantire lo sviluppo delle capacità necessarie e la gestione del rischio. Rivedi regolarmente le tue priorità e aggiornale al mutare delle esigenze. Quando la responsabilità e la proprietà sono indefinite o sconosciute, rischi sia di non affrontare tempestivamente le attività necessarie sia di adoperarti in modo ridondante e potenzialmente conflittuale per rispondere a tali esigenze. La cultura organizzativa influisce direttamente sulla soddisfazione sul lavoro e sulla conservazione dei membri del team. Sostieni il coinvolgimento e le capacità dei membri del tuo team per ottenere il successo della tua attività. La sperimentazione è necessaria per realizzare l'innovazione e trasformare le idee in risultati. Un risultato indesiderato è un esperimento riuscito che ha identificato un percorso che non porterà al successo.

Preparazione

Per prepararti all'eccellenza operativa devi comprendere i carichi di lavoro e i loro comportamenti previsti. Sarai dunque in grado di progettare i carichi di lavoro in modo tale che forniscano informazioni sul loro stato e di creare le procedure per supportarli adeguatamente.

Progetta il tuo carico di lavoro affinché ti fornisca le informazioni necessarie a comprenderne lo stato interno (ad esempio, parametri, log, eventi e tracce) in tutti i componenti a supporto dell'osservabilità e dell'analisi dei problemi. Ripeti le operazioni per sviluppare la telemetria necessaria per monitorare lo stato del carico di lavoro, identificare quando i risultati sono a rischio e abilitare risposte efficaci. Mentre attivi il carico di lavoro, acquisisci un ampio spettro di informazioni per consentire la consapevolezza situazionale (ad esempio cambiamenti di stato, attività utente, accesso con privilegi, contatori di utilizzo), sapendo che hai la possibilità di applicare filtri per selezionare le informazioni più utili nel corso del tempo.

Adotta strategie che migliorino il flusso delle modifiche in produzione e che consentano la rifattorizzazione, il feedback veloce sulla qualità e la correzione di errori. Tali prassi accelerano l'ingresso in produzione delle modifiche vantaggiose, limitano i problemi distribuiti e consentono una rapida identificazione e risoluzione dei problemi introdotti attraverso le attività di distribuzione o scoperti negli ambienti.

Adotta prassi che consentano di fornire un feedback rapido sulla qualità e permettano un ripristino veloce dalle modifiche che non hanno i risultati previsti. L'uso di queste prassi consente di mitigare l'impatto dei problemi introdotti attraverso la distribuzione delle modifiche. Prepara un piano in caso di esito negativo delle modifiche in modo da poter rispondere più rapidamente se necessario, testando e convalidando le modifiche apportate. Sii consapevole delle attività pianificate nei tuoi ambienti

in modo da poter gestire il rischio di modifiche che influiscono sulle attività pianificate. Privilegia le modifiche frequenti, piccole e reversibili per limitarne l'ambito. Semplificherai così la risoluzione dei problemi, accelerando la correzione e mantenendo la possibilità di rollback delle modifiche. In tal modo, è anche possibile ottenere più frequentemente i vantaggi offerti dalle modifiche importanti.

Valuta la prontezza operativa del carico di lavoro, dei processi e delle procedure, nonché del personale, per comprendere i rischi operativi correlati al carico di lavoro. È consigliabile utilizzare un processo omogeneo (inclusi elenchi di controllo manuali o automatici) per sapere quando puoi rilasciare un carico di lavoro o una modifica. Questo inoltre ti consentirà di trovare eventuali aree che per essere affrontate necessitano di pianificazioni. Predisponi istruzioni che documentano le tue attività di routine e manuali che guidano i processi per la risoluzione dei problemi. Analizza i vantaggi e i rischi per prendere decisioni informate e consentire l'adozione delle modifiche nella produzione.

In AWS, puoi vedere il tuo carico di lavoro completo (applicazioni, infrastruttura, policy, governance e operazioni) in forma di codice. In tal modo è possibile applicare la stessa disciplina ingegneristica utilizzata per il codice dell'applicazione a ogni elemento dello stack, condividendoli tra team o organizzazioni per sfruttare al massimo i vantaggi delle attività di sviluppo. Utilizza le operazioni come codice nel cloud e sfrutta la possibilità di sperimentare per sviluppare il tuo carico di lavoro e le procedure operative ed esercitarti con gli errori in modo sicuro. AWS CloudFormation ti consente di avere ambienti di sviluppo, di prova e di produzione sandbox, omogenei e basati su modelli, con livelli crescenti di controllo operativo.

Le seguenti domande si concentrano su queste considerazioni relative all'eccellenza operativa.

OPS 4 In che modo progetti il carico di lavoro al fine di comprenderne lo stato?

Progetta il tuo carico di lavoro in modo da ottenere le informazioni necessarie tra i componenti (ad esempio, parametri, log e tracce) per comprenderne lo stato interno. Ciò ti consente di fornire risposte efficaci in base alle esigenze.

OPS 5 In che modo riduci i difetti, favorisci la correzione e migliori il flusso nella produzione?

Adotta prassi che migliorino il flusso delle modifiche nella produzione, che consentano il refactori ng e il feedback veloce su qualità e correzione di errori. Tali prassi accelerano l'ingresso in produzione delle modifiche vantaggiose, limitano i problemi distribuiti e consentono una rapida identificazione e risoluzione dei problemi introdotti attraverso le attività di distribuzione.

OPS 6 In che modo mitighi i rischi della distribuzione?

Adotta prassi che consentano di fornire un feedback rapido sulla qualità e permettano un ripristino veloce dalle modifiche che non hanno i risultati previsti. L'uso di queste prassi consente di mitigare l'impatto dei problemi introdotti attraverso la distribuzione delle modifiche.

OPS 7 Come fai a sapere che sei pronto a supportare un carico di lavoro?

Valuta la disponibilità operativa del carico di lavoro, dei processi e delle procedure, nonché del personale per comprendere i rischi operativi correlati al carico di lavoro.

Investi nell'implementazione di attività operative come codice per aumentare al massimo la produttività del personale operativo, ridurre al minimo la frequenza degli errori e consentire risposte automatizzate. Utilizza l'analisi prefallimentare per prevedere errori e creare procedure ove opportuno. Applica i metadati utilizzando i tag delle risorse e i AWS Resource Groups seguendo una strategia di applicazione dei tag coerente per consentire l'identificazione delle risorse. Applica tag alle risorse per organizzare, monitorare i costi e controllare gli accessi e ottimizza l'esecuzione delle attività operative automatizzate. Adotta procedure di distribuzione che sfruttino l'elasticità del cloud per facilitare le attività di sviluppo e la pre-distribuzione dei sistemi e avere implementazioni più rapide. Quando apporti modifiche agli elenchi di controllo che utilizzi per valutare i tuoi carichi di lavoro, pianifica quello che farai con i sistemi live che non risultano più conformi.

Operatività

La corretta operatività di un carico di lavoro è misurata dal raggiungimento di risultati per l'azienda e per i clienti. Definisci i risultati desiderati, determina in che modo verrà misurato il successo e individua i parametri che saranno usati nei calcoli per determinare se il carico di lavoro e le operazioni sono efficaci. L'integrità delle operazioni include sia lo stato del carico di lavoro sia lo stato e il successo delle operazioni a supporto del carico di lavoro (ad esempio, la distribuzione e la risposta agli incidenti). Stabilisci le basi dei parametri per migliorare, eseguire indagini e intervenire, raccogliere e analizzare i parametri, quindi conferma la tua comprensione del successo operativo e della sua evoluzione nel corso del tempo. Usa i parametri raccolti per determinare il grado di soddisfazione dei clienti, capire se stai rispondendo alle esigenze aziendali e individuare gli aspetti da migliorare.

La gestione efficiente ed efficace degli eventi operativi è fondamentale per raggiungere l'eccellenza operativa. Ciò si applica agli eventi operativi sia pianificati che non. Usa istruzioni precise per gli eventi chiari e ricorri ai manuali per favorire l'analisi e la risoluzione degli altri eventi. Attribuisci la priorità alle risposte agli eventi in base al loro impatto sull'azienda e sui clienti. Assicurati che, in caso di avvisi in risposta a un evento, vi sia una procedura associata da seguire, con un proprietario ben preciso. Definisci in anticipo il personale richiesto per risolvere un evento e includi dei trigger di escalation per coinvolgere altro personale, ove necessario, in base all'urgenza e all'impatto. Individua e coinvolgi le persone che hanno l'autorità per prendere decisioni in merito alle linee d'azione laddove vi sia un impatto aziendale dovuto a una risposta a un evento non gestito precedentemente.

Comunica lo stato operativo dei carichi di lavoro tramite pannelli di controllo e notifiche personalizzati in base al pubblico di destinazione (ad esempio cliente, azienda, sviluppatori, addetti alle operazioni), in modo che gli interessati possano agire in maniera adeguata, che le loro aspettative vengano soddisfatte e che siano informati sulla ripresa delle normali operazioni.

In AWS puoi generare panoramiche di pannelli di controllo per i parametri raccolti dai carichi di lavoro e in modo nativo da AWS. Puoi sfruttare CloudWatch o applicazioni di terze parti per aggregare e presentare panoramiche a livello di business, di carico di lavoro e di operazioni delle attività operative. AWS fornisce approfondimenti sui carichi di lavoro attraverso funzionalità di registrazione, tra cui AWS X-Ray, CloudWatch, CloudTrail e VPC Flow Logs, che consentono di identificare i problemi del carico di lavoro a supporto dell'analisi delle cause principali e della risoluzione dei problemi.

Le seguenti domande si concentrano su queste considerazioni relative all'eccellenza operativa.

OPS 8 Come fai a comprendere lo stato del tuo carico di lavoro?

Definisci, acquisisci e analizza i parametri del carico di lavoro per ottenere visibilità sugli eventi del carico di lavoro, in modo da intraprendere le azioni appropriate.

OPS 9 Come fai a comprendere lo stato delle operazioni?

Definisci, acquisisci e analizza i parametri delle operazioni per ottenere visibilità sugli eventi delle operazioni, in modo da intraprendere le azioni appropriate.

OPS 10 In che modo gestisci gli eventi del carico di lavoro e delle operazioni?

Prepara e convalida le procedure in risposta agli eventi per ridurre al minimo il loro impatto sul tuo carico di lavoro.

Tutti i parametri raccolti devono essere allineati alle esigenze aziendali e ai risultati che supportano. Sviluppa risposte con script per eventi ben compresi e automatizza le prestazioni in risposta al riconoscimento dell'evento.

Evoluzione

Devi imparare, condividere e migliorare continuamente per sostenere l'eccellenza operativa. Dedica dei cicli di lavoro al raggiungimento di miglioramenti incrementali continui. Esegui l'analisi post-incidente di tutti gli eventi che influiscono sul cliente. Identifica i fattori che contribuiscono e le azioni preventive per limitare o prevenire la ricorrenza. Comunica i fattori che contribuiscono alle comunità interessate, nel modo più adeguato. Valuta regolarmente e assegna le priorità alle opportunità di miglioramento (ad esempio, richieste di funzionalità, risoluzione dei problemi e requisiti di conformità), includendo sia il carico di lavoro sia le procedure operative.

Includi i loop di feedback nelle tue procedure per individuare rapidamente gli aspetti che devono essere migliorati e per acquisire conoscenze dall'esecuzione delle operazioni.

Condividi le lezioni apprese con i vari team per condividerne anche i vantaggi. Analizza le tendenze all'interno delle lezioni apprese ed esegui analisi trasversali retrospettive dei parametri operativi per individuare le opportunità e i metodi di miglioramento. Implementa le modifiche previste per garantire il miglioramento e valuta i risultati per favorire il successo.

In AWS, è possibile esportare i dati di log in Amazon S3 o inviare log direttamente ad Amazon S3 per lo storage a lungo termine. Utilizzando AWS Glue, è possibile individuare e preparare i dati di log in Amazon S3 per l'analisi, archiviando i metadati associati in AWS Glue Data Catalog. Amazon Athena, grazie all'integrazione nativa con AWS Glue, può essere quindi utilizzato per analizzare i dati di log, eseguendo query tramite SQL standard. Utilizzando uno strumento di business intelligence come Amazon QuickSight puoi visualizzare, esplorare e analizzare i tuoi dati. Rilevamento di tendenze ed eventi di interesse che possono portare a miglioramenti.

La seguente domanda si concentra su queste considerazioni relative all'eccellenza operativa.

OPS 11 In che modo fai evolvere le operazioni?

Dedica tempo e risorse al miglioramento incrementale continuo, per far evolvere l'efficacia e l'efficienza delle tue operazioni.

L'evoluzione efficace delle operazioni si basa sugli elementi seguenti: miglioramenti piccoli ma frequenti; creazione di ambienti sicuri e tempo per sperimentare, sviluppare e testare i miglioramenti; ambienti in cui le persone siano incoraggiate a imparare dagli errori. Il supporto alle operazioni per ambienti sandbox, di sviluppo, di prova e di produzione, con un crescente livello di controlli operativi, facilita lo sviluppo e aumenta la prevedibilità dei risultati positivi dalle modifiche passate in produzione.

Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice relative all'eccellenza operativa.

Documentazione

- DevOps e AWS

Whitepaper

- Principio dell'eccellenza operativa

Video

- DevOps di Amazon

Sicurezza

Il pilastro della sicurezza contempla la capacità di proteggere dati, sistemi e asset per sfruttare le tecnologie cloud in modo da migliorare la sicurezza.

Il principio della sicurezza offre una panoramica dei principi di progettazione, delle best practice e delle domande. È possibile trovare linee guida prescrittive sull'implementazione nel Whitepaper sul principio della sicurezza.

Argomenti

Principi di progettazione

Esistono sette principi di progettazione per la sicurezza nel cloud:

- Implementa una solida base identitariaimplementa il principio del privilegio minimo e attua la separazione dei compiti con la corretta autorizzazione per ciascuna interazione con le risorse AWS. Centralizza la gestione delle identità e mira a eliminare la dipendenza dalle credenziali statiche a lungo termine.
- Abilita la tracciabilità: monitora, avvisa e verifica le azioni e le modifiche al tuo ambiente in tempo reale. Integra la raccolta di log e parametri con i sistemi per analizzare e intervenire automaticamente.
- Applica la sicurezza a tutti i livelli: applica un approccio di difesa avanzata con più controlli di sicurezza. Applicalo a tutti i livelli (ad esempio, edge di rete, VPC, bilanciamento del carico, ogni istanza e servizio di elaborazione, sistema operativo, applicazione e codice).
- Automatizza le best practice per la sicurezza: meccanismi di sicurezza automatici basati sul software migliorano la capacità di ricalibrare il sistema in modo sicuro, più rapido e conveniente. Crea architetture sicure, compresa l'implementazione dei controlli, che sono definite e gestite come codice nei modelli controllati dalle versioni.
- Proteggi i dati in transito e a riposo: classifica i dati secondo livelli di sensibilità e meccanismi d'uso, come crittografia, tokenizzazione e controllo di accesso, ove opportuno.
- Tieni le persone a distanza dai dati: utilizza meccanismi e strumenti per ridurre o eliminare l'esigenza di accesso diretto o di elaborazione manuale dei dati. Ciò riduce il rischio di perdita, modifica e altri errori umani durante la gestione dei dati sensibili.
- Preparati per gli eventi di sicurezza: preparati per un incidente ipotetico creando policy e processi di gestione degli incidenti allineati ai requisiti dell'organizzazione. Esegui simulazioni di risposta

agli incidenti e utilizza strumenti dotati di automazione per aumentare la velocità nel rilevamento, nell'indagine e nel ripristino.

Definizione

Esistono sei aree di best practice per la sicurezza nel cloud:

- Sicurezza
- Gestione di identità e accessi (Identity and access management)
- Rilevamento
- Protezione dell'infrastruttura
- Protezione dei dati
- Risposta agli incidenti

Prima di progettare qualsiasi carico di lavoro, è necessario implementare pratiche che influenzano la sicurezza. Dovrai controllare chi può fare cosa. Inoltre, devi essere in grado di identificare gli incidenti di sicurezza, proteggere i tuoi sistemi e i tuoi servizi e mantenere la riservatezza e l'integrità dei dati attraverso la loro protezione. Dovresti avere dei processi ben definiti e rodati per rispondere a eventuali problemi di sicurezza. Questi strumenti e tecniche sono importanti perché supportano obiettivi come la prevenzione delle perdite finanziarie o la conformità agli obblighi normativi.

Il modello di responsabilità condivisa di AWS permette alle organizzazioni che adottano il cloud di raggiungere i loro obiettivi in termini di sicurezza e conformità. Dato che AWS mette fisicamente in sicurezza l'infrastruttura che supporta i nostri servizi cloud, come cliente AWS puoi concentrarti sull'utilizzo dei servizi per raggiungere gli obiettivi. Il cloud AWS fornisce, inoltre, l'accesso ai dati sulla sicurezza e offre un approccio automatico per rispondere agli eventi di sicurezza.

Best practice

Argomenti

- Risposta agli imprevisti

Sicurezza

Per gestire il carico di lavoro in modo sicuro, è necessario applicare le best practice globali a ogni area di sicurezza. Segui i requisiti e i processi definiti in termini di eccellenza operativa a livello organizzativo e di carico di lavoro e applicali a tutte le aree.

Rimanere aggiornati con le raccomandazioni di AWS e del settore nonché con l'intelligence sulle minacce aiuta a sviluppare il modello di rischio e gli obiettivi di controllo. L'automazione dei processi di sicurezza, i test e la convalida consentono di ricalibrare le operazioni di sicurezza.

La seguente domanda si concentra su queste considerazioni relative alla sicurezza (per l'elenco completo delle domande e delle best practice, consulta l' Appendice.).

SEC 1 Come gestisci in modo sicuro un carico di lavoro?
Per gestire il carico di lavoro in modo sicuro, è necessario applicare le best practice globali a ogni area di sicurezza. Segui i requisiti e i processi definiti in termini di eccellenza operativa a livello organizzativo e di carico di lavoro e applicali a tutte le aree. Rimanere aggiornati con le raccomand azioni di AWS, le fonti di settore nonché con l'intelligence sulle minacce aiuta a sviluppare il modello di rischio e gli obiettivi di controllo. L'automazione dei processi di sicurezza, i test e la convalida consentono di ricalibrare le operazioni di sicurezza.

In AWS, è consigliabile separare i diversi carichi di lavoro per account, in base alla loro funzione e ai requisiti di conformità o di sensibilità dei dati.

Gestione di identità e accessi

La gestione delle identità e degli accessi è una parte principale di un programma di sicurezza delle informazioni e garantisce che solo gli utenti e i componenti autorizzati e autenticati possano accedere alle tue risorse e solo nella modalità che hai stabilito. Ad esempio, è necessario definire i principali (ovvero account, utenti, ruoli e servizi che possono eseguire operazioni nel tuo account), creare policy allineate a tali principali e implementare una forte gestione delle credenziali. Questi elementi a gestione privilegiata formano i concetti chiave dell'autenticazione e dell'autorizzazione.

In AWS, la gestione dei privilegi è principalmente supportata dal servizio AWS Identity and Access Management (IAM), che consente di controllare l'accesso utente e l'accesso programmatico ai servizi

e alle risorse AWS. È necessario applicare criteri granulari che assegnano autorizzazioni a un utente, gruppo, ruolo o risorsa. Hai anche la possibilità di richiedere pratiche di password complesse, come il livello di complessità, evitare il riutilizzo e applicare l'autenticazione a più fattori (MFA). È possibile utilizzare la federazione con il servizio di directory esistente. Per i carichi di lavoro che richiedono che i sistemi abbiano accesso ad AWS, IAM consente l'accesso sicuro tramite ruoli, profili dell'istanza, federazione delle identità e credenziali temporanee.

Le seguenti domande si concentrano su queste considerazioni relative alla sicurezza.

SEC 2 Come gestisci l'autenticazione per persone e macchine?

Esistono due tipi di identità che è necessario gestire quando si utilizzano carichi di lavoro AWS sicuri. Comprendere il tipo di identità necessaria per gestire e concedere l'accesso ti aiuta a garantire che le identità corrette abbiano accesso alle risorse giuste nelle condizioni adeguate.

Identità umane: amministratori, sviluppatori, operatori e utenti finali necessitano di un'identità per accedere agli ambienti e alle applicazioni AWS. Si tratta di membri dell'organizzazione o utenti esterni con cui collabori e che interagiscono con le tue risorse AWS tramite browser Web, applicazioni client o strumenti a riga di comando interattivi.

Identità di macchine: le applicazioni di servizio, gli strumenti operativi e i carichi di lavoro necessita no di un'identità per effettuare richieste ai servizi AWS, ad esempio per leggere i dati. Queste identità includono macchine in esecuzione nell'ambiente AWS, ad esempio istanze Amazon EC2 o funzioni AWS Lambda. Puoi gestire le identità di macchine anche per soggetti esterni che necessitano dell'accesso. Inoltre, potresti disporre di macchine al di fuori di AWS che devono accedere al tuo ambiente AWS.

SEC 3 Come gestisci le autorizzazioni per persone e macchine?

Gestisci le autorizzazioni per controllare l'accesso alle identità di persone e macchine che richiedono l'accesso ad AWS e al tuo carico di lavoro. Le autorizzazioni controllano chi può accedere a cosa e a quali condizioni.

Le credenziali non devono essere condivise tra nessun utente o sistema. L'accesso degli utenti dovrebbe essere concesso utilizzando un approccio con privilegi minimi con le migliori pratiche, inclusi i requisiti di password e l'applicazione del MFA. L'accesso programmatico, comprese le

chiamate API ai servizi AWS, deve essere eseguito utilizzando credenziali temporanee e con privilegi limitati come quelle emesse da AWS Security Token Service.

AWS offre risorse che possono aiutarti nella gestione dell'identità e degli accessi. Per apprendere le best practice, esplora i nostri corsi pratici sulla gestione delle credenziali e dell'autenticazione, sul controllo dell'accesso umanoe sul controllo dell'accesso programmatico.

Rilevamento

Puoi utilizzare i controlli di rilevamento per identificare una potenziale minaccia o un potenziale incidente di sicurezza. Questi controlli sono una parte essenziale dei framework di governance e possono essere utilizzati per supportare il processo di qualità o un obbligo legale o di conformità e per l'identificazione delle minacce e gli sforzi nelle risposte. Ci sono diversi tipi di controlli di rilevamento. Ad esempio, la realizzazione di un inventario di risorse e dei loro attributi dettagliati promuove le decisioni più efficienti (e i controlli del ciclo di vita) per stabilire delle baseline operative. Puoi anche utilizzare audit interni, una verifica dei controlli relativi ai sistemi di informazioni, per assicurarti che le practice rispettino le policy e i requisiti e che tu abbia un set corretto di notifiche di avviso automatiche basate sulle condizioni definite. Questi controlli sono fattori di reazione importanti che possono aiutare la tua organizzazione a identificare e capire la portata dell'attività anomala.

In AWS, puoi implementare controlli investigativi elaborando registri, eventi e monitoraggio che consentono audit, analisi automatizzate e notifiche. I registri CloudTrail, le chiamate API AWS e CloudWatch forniscono il monitoraggio di parametri con notifiche, mentre AWS Config fornisce la cronologia delle configurazioni. Amazon GuardDuty è un servizio di rilevazione delle minacce che monitora costantemente possibili comportamenti dannosi o non autorizzati, così da proteggere i tuoi account e i tuoi carichi di lavoro su AWS. Sono inoltre disponibili log a livello di servizio, ad esempio puoi utilizzare Amazon Simple Storage Service (Amazon S3) per registrare le richieste di accesso.

La seguente domanda si concentra su queste considerazioni relative alla sicurezza

SEC 4 In che modo individui ed esamini gli eventi di sicurezza?

Acquisisci ed analizza gli eventi a partire da log e parametri per acquistare visibilità. Agisci su eventi di sicurezza e potenziali minacce per contribuire a rendere sicuro il carico di lavoro.

La gestione dei log è una parte importante di un carico di lavoro Well-Architected per ragioni che vanno da requisiti di sicurezza o forensi a disposizioni normative o legali. È fondamentale

analizzare i registri e rispondere in modo da identificare potenziali incidenti di sicurezza. AWS offre funzionalità che semplificano l'implementazione della gestione dei registri, offrendo la possibilità di definire un ciclo di vita di conservazione dei dati o di definire dove verranno conservati, archiviati o eventualmente eliminati. Ciò rende la gestione dei dati prevedibile e affidabile, più semplice ed economica.

Protezione dell’infrastruttura

La protezione dell'infrastruttura comprende delle metodologie di controllo, come la difesa approfondita, necessarie per rispettare le best practice e gli obblighi organizzativi e normativi. L'utilizzo di queste metodologie è fondamentale per ottenere operazioni continuative e di successo sia nel cloud che in locale.

In AWS, è possibile implementare l'ispezione di pacchetti con stato e senza stato, sia utilizzando tecnologie native di AWS, sia utilizzando prodotti e servizi dei partner disponibili attraverso Marketplace AWS. È necessario utilizzare Amazon Virtual Private Cloud (Amazon VPC) per creare un ambiente privato, protetto e scalabile in cui è possibile definire la propria topologia, inclusi gateway, tabelle di indirizzamento e sottoreti pubbliche e private.

Le seguenti domande si concentrano su queste considerazioni relative alla sicurezza.

SEC 5 In che modo proteggi le risorse di rete?

Qualsiasi carico di lavoro che abbia una qualche forma di connettività di rete, che si tratti di Internet o di una rete privata, richiede più livelli di difesa per proteggere da minacce esterne e interne basate sulla rete.

SEC 6 In che modo proteggi le risorse di calcolo?

Le risorse di calcolo nel carico di lavoro richiedono più livelli di difesa per contribuire alla protezion e da minacce esterne ed interne. Le risorse di calcolo includono istanze EC2, container, funzioni di AWS Lambda, servizi di database, dispositivi IoT e altro.

Si consigliano più livelli di difesa in qualsiasi tipo di ambiente. Nel caso della protezione dell'infrastruttura, molti concetti e metodi sono validi sia per modelli cloud che in locale. L'applicazione della protezione dei confini, il monitoraggio dei punti di ingresso e di uscita e la registrazione, il

monitoraggio e le notifiche completi sono tutti elementi essenziali per un efficace piano di sicurezza delle informazioni.

I clienti AWS sono in grado di adattare o rafforzare la configurazione di Amazon Elastic Compute Cloud (Amazon EC2), di un container Amazon Elastic Container Service (Amazon ECS) o di un'istanza AWS Elastic Beanstalk e mantenere questa configurazione su una Amazon Machine Image (AMI) immutabile. Quindi, che siano attivati da Auto Scaling o lanciati manualmente, tutti i nuovi server virtuali (istanze) lanciati con questa AMI utilizzeranno la configurazione avanzata.

Protezione dei dati

Prima di progettare qualsiasi sistema, devono essere stabiliti i requisiti fondamentali che influenzano la sicurezza. Ad esempio, la classificazione dei dati fornisce un modo per categorizzare i dati organizzativi basati sui livelli di sensibilità, mentre la crittografia protegge i dati evitandone l'intelligibilità per gli accessi non autorizzati. Questi strumenti e tecniche sono importanti perché supportano obiettivi come la prevenzione delle perdite finanziarie o la conformità agli obblighi normativi.

In AWS, le seguenti pratiche facilitano la protezione dei dati:

- Come cliente AWS mantieni il pieno controllo sui tuoi dati.
- AWS semplifica la crittografia dei dati e la gestione delle chiavi, inclusa la rotazione regolare delle chiavi, che può essere facilmente automatizzata da AWS o gestita da te.
- È disponibile la registrazione dettagliata che contiene contenuti importanti, come l'accesso ai file e le modifiche.
- AWS ha progettato sistemi di storage con una resilienza eccezionale. Ad esempio, Amazon S3 Standard, S3 Standard-IA, One Zone-IA S3 e Amazon Glacier sono tutti progettati per offrire una resistenza degli oggetti del 99,999999999% in un determinato anno. Questo livello di durabilità corrisponde a una perdita media annua prevista dello 0,000000001% di oggetti.
- Il controllo delle versioni, che può far parte di un più ampio processo di gestione del ciclo di vita dei dati, può proteggere da sovrascritture accidentali, eliminazioni e danni simili.
- AWS non avvia mai il trasferimento di dati tra Regioni. Il contenuto inserito in una regione rimarrà in quella regione a meno che tu non abiliti esplicitamente una funzione o utilizzi un servizio che fornisce tale funzionalità.

Le seguenti domande si concentrano su queste considerazioni relative alla sicurezza.

SEC 7 In che modo classifichi i dati?

La classificazione fornisce un modo per categorizzare i dati in base ai livelli di criticità e sensibilità, in modo da aiutarti a determinare i controlli di protezione e conservazione appropriati.

SEC 8 In che modo proteggi i dati inattivi?

Proteggi i dati inattivi implementando più controlli, per ridurre il rischio di accessi non autorizzati o altri comportamenti impropri.

SEC 9 In che modo proteggi i dati in transito?

Proteggi i dati in transito implementando più controlli, per ridurre il rischio di accessi non autorizzati o perdita.

AWS offre molteplici mezzi per crittografare i dati a riposo e in transito. Nei nostri servizi integriamo funzionalità che semplificano la crittografia dei dati. Ad esempio, abbiamo implementato la crittografia lato server (SSE) per Amazon S3 per semplificare l'archiviazione dei dati in forma crittografata. È inoltre possibile disporre che l'intero processo di crittografia e decrittografia HTTPS (generalmente noto come terminazione SSL) sia gestito da Elastic Load Balancing (ELB).

Risposta agli imprevisti

Anche con controlli preventivi e investigativi estremamente maturi, la tua organizzazione dovrebbe comunque attuare processi per rispondere e mitigare il potenziale impatto di incidenti di sicurezza. L'architettura del carico di lavoro influisce fortemente sulla capacità dei team di operare efficacemente durante un incidente, isolare o contenere sistemi e ripristinare le operazioni a uno stato ottimale noto. La messa in atto degli strumenti e l'accesso prima di un incidente di sicurezza e la pratica sistematica della risposta agli incidenti durante i giorni di attività ti aiuterà a garantire che la tua architettura sia in grado di supportare indagini e ripristini tempestivi.

In AWS, le seguenti pratiche facilitano una risposta efficace agli incidenti:

- Sono disponibili registrazioni dettagliate che contengono contenuti importanti, come l'accesso ai file e le modifiche.

- Gli eventi possono essere elaborati automaticamente e possono attivare strumenti che automatizzano le risposte mediante l'uso delle API di AWS.
- Puoi effettuare il pre-provisioning degli strumenti e una "clean room" utilizzando AWS CloudFormation. Questo permette di effettuare indagini forensi in un ambiente sicuro e isolato.

La seguente domanda si concentra su queste considerazioni relative alla sicurezza

SEC 10 In che modo prevedi, reagisci a e risolvi gli incidenti?, rispondi e risolvi gli eventi?
La preparazione è cruciale per un esame tempestivo ed efficace degli incidenti di sicurezza, nonché per la risposta e il ripristino, così da ridurre al minimo potenziali interruzioni dell'orga nizzazione.

Assicurati di poter garantire rapidamente l'accesso al tuo team addetto alla sicurezza e automatizzare l'isolamento delle istanze, oltre che acquisire i dati e lo stato per le indagini forensi.

Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice relative alla sicurezza.

Documentazione

- Sicurezza del cloud AWS
- Conformità di AWS
- Blog sulla sicurezza di AWS

Whitepaper

- Pilastro della sicurezza
- Panoramica sulla sicurezza di AWS
- Rischio e conformità di AWS

Video

- AWS Security State of the Union

- Panoramica sulla responsabilità condivisa

Affidabilità

Il principio dell'affidabilità comprende la capacità di un carico di lavoro di eseguire la funzione attesa in modo corretto e coerente quando previsto. Include la possibilità di utilizzare e testare il carico di lavoro per tutto il ciclo di vita. Questo documento fornisce linee guida dettagliate sulle best practice per l'implementazione di carichi di lavoro affidabili in AWS.

Il principio dell'affidabilità offre una panoramica dei principi di progettazione, delle best practice e delle domande. È possibile trovare linee guida prescrittive sull'implementazione nel Whitepaper sul principio dell'affidabilità.

Argomenti

- Principi di progettazione
- Definizione
- Best practice
- Risorse

Principi di progettazione

Esistono cinque principi di progettazione per l'affidabilità nel cloud:

- Adotta un approccio di ripristino automatico dagli errori: monitorando gli indicatori chiave di prestazione (KPI) di un carico di lavoro, è possibile attivare l'automazione in caso di superamento di una soglia. Questi KPI dovrebbero essere una misura del valore aziendale, non degli aspetti tecnici del funzionamento del servizio. Ciò consente la notifica e il tracciamento automatici degli errori e i processi di recupero automatizzati che aggirano o riparano l'errore. Con un'automazione più sofisticata è possibile anticipare e correggere gli errori prima che si verifichino.
- Collauda le procedure di ripristino: in un ambiente in locale, spesso vengono eseguiti test per dimostrare che il carico di lavoro funziona in uno scenario specifico. I test non vengono generalmente utilizzati per convalidare le strategie di recupero. Nel cloud, puoi testare il modo in cui il carico di lavoro incorre nell'errore e convalidare le procedure di ripristino. È possibile utilizzare l'automazione per simulare diversi errori o per ricreare scenari che in precedenza hanno portato a errori. Questo approccio presenta percorsi di errore che è possibile testare e correggere prima che si verifichi uno scenario di errore reale, riducendo così il rischio.

- Dimensiona orizzontalmente per aumentare la disponibilità dei carichi di lavoro aggregati: sostituisci una risorsa grande con più risorse piccole per ridurre l'impatto di un singolo guasto sul carico di lavoro complessivo. Distribuisci le richieste su molteplici risorse più piccole per garantire che non condividano un punto di errore comune.
- Smetti di fare congetture sulla capacità: una causa comune di guasti nei carichi di lavoro in locale è la saturazione delle risorse, quando le richieste assegnate a un carico di lavoro superano la capacità di quel carico di lavoro (questo è spesso l'obiettivo di attacchi di tipo Denial of Service). Nel cloud, è possibile monitorare la domanda e l'utilizzo dei carichi di lavoro, nonché automatizzare l'aggiunta o la rimozione di risorse per mantenere il livello ottimale, al fine di soddisfare la domanda senza un provisioning eccessivo o inferiore. Esistono ancora dei limiti, ma alcune quote possono essere controllate e altre possono essere gestite (consulta Gestisci vincoli e Service Quotas).
- Gestisci il cambiamento nell'automazione: le modifiche all'infrastruttura dovrebbero essere apportate utilizzando l'automazione. Le modifiche che devono essere gestite includono le modifiche all'automazione, che possono quindi essere monitorate e revisionate.

Definizione

Esistono quattro aree di best practice per l'affidabilità nel cloud:

- Fondamenti
- Architettura del carico di lavoro
- Gestione delle modifiche
- Gestione degli errori

Per ottenere affidabilità, è necessario iniziare dalle basi: un ambiente in cui le quote di servizio e la topologia di rete sono in grado di supportare il carico di lavoro. L'architettura del carico di lavoro del sistema distribuito deve essere progettata per prevenire e mitigare gli errori. Il carico di lavoro deve gestire le variazioni nella domanda o nei requisiti e deve essere progettato per rilevare l'errore e correggersi automaticamente.

Best practice

Argomenti

- Gestione delle modifiche
- Gestione degli errori

Fondamenti

I requisiti di base sono quelli il cui ambito si estende oltre un singolo carico di lavoro o progetto. Prima di progettare qualsiasi sistema, devono essere stabiliti i requisiti fondamentali che influenzano l'affidabilità. Ad esempio, è necessario disporre di una larghezza di banda di rete sufficiente verso il data center.

Con AWS, la maggior parte di questi requisiti di base è già incorporata o può essere affrontata in base alle esigenze. Il cloud è progettato per essere quasi illimitato, perciò è responsabilità di AWS soddisfare i requisiti di capacità di rete e di elaborazione sufficienti, lasciandoti libero di modificare le dimensioni delle risorse e le allocazioni on demand.

Le seguenti domande si concentrano su queste considerazioni relative all'affidabilità. (Per l'elenco completo delle domande e delle best practice relative all'affidabilità, consulta l' Appendice.).

REL 1 In che modo gestisci quote e vincoli di servizio?

Per le architetture di carichi di lavoro basate sul cloud, esistono quote di servizio (definite anche come restrizioni dei servizi). Queste quote sono presenti per evitare di effettuare accidentalmente il provisioning di più risorse di quelle necessarie e limitare i tassi di richiesta sulle operazioni API in modo da proteggere i servizi da un uso illecito. Esistono anche vincoli di risorse, ad esempio la velocità con cui è possibile trasferire i bit su un cavo in fibra ottica o la quantità di storage su un disco fisico.

REL 2 In che modo pianifichi la topologia di rete?

I carichi di lavoro sono spesso presenti in più ambienti. Questi includono più ambienti cloud (sia pubblicamente accessibili sia privati) e, possibilmente, l'infrastruttura del data center esistente. I piani devono includere considerazioni di rete, ad esempio connettività intrasistema e intersistema, gestione di indirizzi IP pubblici, gestione di indirizzi IP privati e risoluzione dei nomi di dominio.

Per le architetture di carichi di lavoro basate sul cloud, esistono quote di servizio (definite anche come restrizioni dei servizi). Queste quote sono presenti per evitare di effettuare accidentalmente

il provisioning di più risorse di quelle necessarie e limitare i tassi di richiesta sulle operazioni API in modo da proteggere i servizi da un uso illecito. I carichi di lavoro sono spesso presenti in più ambienti. È necessario monitorare e gestire queste quote per tutti gli ambienti dei carichi di lavoro. Questi includono più ambienti cloud (sia pubblicamente accessibili sia privati) e possono includere l'infrastruttura del data center esistente. I piani devono includere considerazioni di rete, ad esempio connettività intrasistema e intersistema, gestione di indirizzi IP pubblici, gestione di indirizzi IP privati e risoluzione dei nomi di dominio.

Architettura del carico di lavoro

Un carico di lavoro affidabile comincia con decisioni iniziali di progettazione sia per il software sia per l'infrastruttura. Le tue scelte architetturali avranno un impatto sul comportamento del carico di lavoro su tutti e cinque i pilastri del Well-Architected Framework. Per l'affidabilità, è necessario seguire modelli specifici.

Con AWS, gli sviluppatori di carichi di lavoro possono scegliere i linguaggi e le tecnologie da utilizzare. Gli SDK AWS semplificano la scrittura di codici fornendo API specifiche dei linguaggi per i servizi AWS. Questi SDK, oltre alla scelta dei linguaggi, consentono agli sviluppatori di implementare le best practice di affidabilità elencate qui. Gli sviluppatori possono anche leggere e scoprire come Amazon crea e gestisce software nella Amazon Builders' Library.

Le seguenti domande si concentrano su queste considerazioni relative all'affidabilità.

REL 3 In che modo progetti l'architettura del servizio di carico di lavoro?

Creazione di carichi di lavoro altamente scalabili e affidabili utilizzando un'architettura orientata ai servizi (SOA) o un'architettura di microservizi. L'architettura orientata ai servizi (SOA) è la pratica di rendere i componenti software riutilizzabili tramite interfacce di servizio. L'architettura dei microservizi va oltre, per rendere i componenti più piccoli e semplici.

REL 4 In che modo progetti le interazioni in un sistema distribuito per evitare errori?

I sistemi distribuiti si basano sulle reti di comunicazione per interconnettere i componenti (ad esempio server o servizi). Il carico di lavoro deve funzionare in modo affidabile nonostante la perdita o la latenza dei dati in queste reti. I componenti del sistema distribuito devono funzionare in modo da non influire negativamente su altri componenti o sul carico di lavoro. Queste best practice prevengono gli errori e migliorano il tempo medio tra errori (MTBF).

REL 5 In che modo progetti le interazioni in un sistema distribuito per mitigare o affrontare gli errori?

I sistemi distribuiti si basano sulle reti di comunicazione per interconnettere i componenti (ad esempio server o servizi). Il carico di lavoro deve funzionare in modo affidabile nonostante la perdita o la latenza dei dati su queste reti. I componenti del sistema distribuito devono funzionar e in modo da non influire negativamente su altri componenti o sul carico di lavoro. Queste best practice consentono ai carichi di lavoro di affrontare stress o guasti, recuperare più rapidamen te e mitigare l'impatto di tali problemi. Il risultato è un miglioramento del tempo medio di ripristino (MTTR).

Gestione delle modifiche

Le modifiche apportate al carico di lavoro o al relativo ambiente devono essere previste e gestite affinché il carico di lavoro funzioni in modo affidabile. Certe modifiche al carico di lavoro sono imposte da fattori esterni, quali i picchi di domanda, altre modifiche dipendono da fattori interni, quali le distribuzioni delle funzionalità e le patch di sicurezza.

Utilizzando AWS, puoi monitorare il comportamento di un carico di lavoro e automatizzare la risposta ai KPI. Ad esempio, il carico di lavoro può aggiungere ulteriori server man mano che il carico di lavoro acquisisce più utenti. È possibile controllare chi dispone dell'autorizzazione per apportare modifiche al carico di lavoro e controllare la cronologia di tali modifiche.

Le seguenti domande si concentrano su queste considerazioni relative all'affidabilità.

REL 6 In che modo monitori le risorse del carico di lavoro?

I log e i parametri sono strumenti molto efficaci per ottenere informazioni sullo stato del tuo carico di lavoro. È possibile configurare il carico di lavoro in modo da monitorare i log e i parametri e inviare notifiche quando vengono superate le soglie o si verificano eventi significativi. Il monitorag gio consente al carico di lavoro di riconoscere quando vengono superate le soglie di prestazioni basse o si verificano errori, in modo che possa essere ripristinato automaticamente di rimando.

REL 7 In che modo progetti il carico di lavoro per adattarti ai cambiamenti della domanda?

Un carico di lavoro scalabile fornisce elasticità per aggiungere o rimuovere risorse automatic amente, in modo che vi sia una stretta corrispondenza con la domanda attuale in un dato momento.

REL 8 In che modo implementi le modifiche?

Per distribuire nuove funzionalità e garantire che i carichi di lavoro e l'ambiente operativo eseguano software noti e che sia possibile applicare patch o sostituirli in modo prevedibile, sono necessarie modifiche controllate. Se invece non sono controllate, risulta difficile prevederne l'effetto o risolvere eventuali problemi che causano.

Progettando un carico di lavoro in grado di aggiungere e rimuovere automaticamente le risorse in risposta ai cambiamenti della domanda, non solo si aumenta l'affidabilità, ma ci si assicura anche che il successo aziendale non diventi un peso. Con il monitoraggio attivo, il tuo team verrà avvisato automaticamente quando gli indicatori KPI si discostano dalle norme previste. La registrazione automatica delle modifiche al proprio ambiente consente di controllare e identificare rapidamente le azioni che potrebbero avere influito sull'affidabilità. I controlli sulla gestione delle modifiche assicurano la possibilità di applicare le regole che garantiscono l'affidabilità di cui hai bisogno.

Gestione degli errori

In qualsiasi sistema di ragionevole complessità è previsto che si verifichino errori. L'affidabilità richiede che il carico di lavoro venga a conoscenza degli errori nel momento in cui si verificano e intervenga per evitare conseguenze sulla disponibilità. I carichi di lavoro devono essere in grado di affrontare errori e risolvere automaticamente i problemi.

Con AWS, puoi sfruttare l'automazione per reagire ai dati di monitoraggio. Ad esempio, quando un determinato parametro supera una soglia, è possibile attivare un'azione automatica per risolvere il problema. Inoltre, anziché tentare di diagnosticare e correggere una risorsa guasta che fa parte del tuo ambiente di produzione, puoi sostituirla con una nuova ed eseguire l'analisi sulla risorsa guasta fuori banda. Poiché il cloud consente di creare versioni temporanee di un intero sistema a basso costo, è possibile utilizzare i test automatizzati per verificare i processi di recupero completi.

Le seguenti domande si concentrano su queste considerazioni relative all'affidabilità.

REL 9 In che modo esegui il backup dei dati?

Esegui il backup dei dati, delle applicazioni e della configurazione per soddisfare i tuoi requisiti relativi agli obiettivi di tempo di ripristino (recovery time objective, RTO) e agli obiettivi di punto di ripristino (recovery point objective, RPO).

REL 10 In che modo utilizzi l'isolamento dei guasti per proteggere il carico di lavoro?

Le barriere per l'isolamento dei guasti limitano l'effetto di un errore all'interno di un carico di lavoro a un numero limitato di componenti. I componenti al di fuori della barriera non subiscono gli effetti del guasto. Utilizzando più barriere per l'isolamento dei guasti, puoi limitare l'impatto sul carico di lavoro.

REL 11 In che modo progetti il carico di lavoro affinché resista ai guasti dei componenti?

I carichi di lavoro con requisiti di disponibilità elevata e MTTR (Mean Time To Recovery) basso devono essere progettati per garantire la resilienza.

REL 12 In che modo testi l'affidabilità?

Dopo aver progettato il carico di lavoro in modo da essere resiliente alle sollecitazioni della produzione, i test sono l'unico modo per garantire il funzionamento corretto e offrire la resilienza prevista.

REL 13 Come pianifichi il disaster recovery (DR)?

Avere backup e componenti del carico di lavoro ridondanti in loco è l'inizio della strategia di disaster recovery. RTO e RPO sono i tuoi obiettivi per il ripristino del carico di lavoro. Imposta questi valori in base alle esigenze aziendali. Implementa una strategia per raggiungere questi obiettivi, prendendo in considerazione le posizioni e la funzione delle risorse e dei dati del carico di lavoro. La probabilità di interruzione e il costo del ripristino sono fattori chiave che aiutano a comunicare il valore aziendale che può avere il ripristino di emergenza per un carico di lavoro.

Esegui regolarmente il backup dei dati e testa i file di backup per assicurarti di poter effettuare il ripristino dopo errori sia logici che fisici. Una chiave per la gestione dei guasti è il test frequente e automatico dei carichi di lavoro che causano gli errori e quindi osservare come si ripristinano. Esegui questa operazione regolarmente e assicurati che tali test vengano attivati anche dopo importanti cambiamenti del carico di lavoro. Traccia attivamente i KPI, oltre a Obiettivo del tempo di ripristino (RTO) e Obiettivo del punto di ripristino (RPO), per valutare la resilienza di un carico di lavoro (specialmente in scenari di test degli errori). Il monitoraggio dei KPI ti aiuterà a identificare e mitigare i singoli punti di errore. L'obiettivo è testare a fondo i processi di ripristino del carico di lavoro in modo da avere la certezza di poter recuperare tutti i dati e continuare a servire i propri clienti, anche di fronte a problemi prolungati. I processi di recupero dovrebbero essere testati tanto quanto i normali processi di produzione.

Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice per l'affidabilità.

Documentazione

- Documentazione di AWS
- Infrastruttura globale di AWS
- AWS Auto Scaling: come funzionano i piani di dimensionamento
- Che cos'è AWS Backup?

Whitepaper

- Pilastro dell'affidabilità: Well-Architected AWS
- Implementazione di microservizi in AWS

Efficienza delle prestazioni

Il principio dell'efficienza delle prestazioni comprende l'abilità di utilizzare in modo efficiente le risorse di elaborazione per soddisfare i requisiti del sistema e conservare tale efficienza a seconda dei cambiamenti della domanda e dell'evoluzione delle tecnologie.

Il principio dell'efficienza delle prestazioni offre una panoramica dei principi di progettazione, delle best practice e delle domande. È possibile trovare linee guida prescrittive sull'implementazione nel Whitepaper sul principio dell'efficienza delle prestazioni.

Argomenti

Principi di progettazione

Esistono cinque principi di progettazione per l'efficienza delle prestazioni nel cloud:

- Estendi a tutti le tecnologie avanzate: facilita l'implementazione di tecnologie avanzate da parte del tuo team delegando le attività complesse al tuo fornitore di cloud. Anziché chiedere al team IT di imparare come adottare e gestire una nuova tecnologia, valuta l'opportunità di utilizzare la tecnologia come servizio. Ad esempio, i database NoSQL, la transcodifica multimediale e il machine learning sono tutte tecnologie che richiedono competenze specialistiche. Nel cloud, tali tecnologie diventano servizi che il tuo team può semplicemente utilizzare mentre si concentra sullo sviluppo di un prodotto invece che sul provisioning e sulla gestione delle risorse.
- Raggiungi una disponibilità globale in pochi minuti: distribuire il carico di lavoro in più regioni AWS in tutto il mondo ti consente di ridurre la latenza e fornire un'esperienza migliore ai tuoi clienti a costi minimi.
- Utilizza le architetture serverless: scegliendo le architetture serverless, non avrai più bisogno di gestire e mantenere server fisici per portare a termine le attività di elaborazione tradizionali. Ad esempio, i servizi di storage possono agire da siti web statici, eliminando la necessità di server web, mentre i servizi di eventi possono ospitare il codice. Questo elimina l'onere operativo della gestione dei server fisici, con una riduzione dei costi delle transazioni, dal momento che servizi gestiti di questo tipo funzionano a livello di cloud.
- Sperimenta con più frequenza: le risorse virtuali e automatizzabili ti permettono di portare a termine velocemente i test comparativi utilizzando diversi tipi di istanze, storage e configurazioni.
- Acquisisci un senso per la meccanica: scopri come vengono consumati i servizi cloud e utilizza sempre l'approccio tecnologico più adatto ai tuoi obiettivi di carico di lavoro. Ad esempio, prendi in considerazione gli schemi di accesso ai dati quando selezioni una strategia basata su database o archiviazione.

Definizione

Esistono quattro aree di best practice per l'efficienza delle prestazioni nel cloud:

- Selezione
- Revisione
- Monitoraggio
- Compromessi

Utilizza un approccio basato sui dati per la creazione di un'architettura a prestazioni elevate. Raccogli dati su tutti gli aspetti dell'architettura, dalla progettazione di alto livello alla selezione e alla configurazione dei tipi di risorse.

Rivedendo le tue decisioni a intervalli regolari, avrai la certezza di sfruttare le capacità in continua evoluzione di AWS Cloud. Il monitoraggio ti assicurerà di essere consapevole di qualsiasi divergenza rispetto alle prestazioni previste. Infine, puoi raggiungere dei compromessi nella tua architettura per migliorare le prestazioni, per esempio utilizzando la compressione o la memorizzazione nella cache oppure allentando i requisiti di coerenza.

Best practice

Argomenti

Selezione

La soluzione ottimale per un determinato carico di lavoro può variare e le soluzioni spesso combinano molteplici approcci. I carichi di lavoro Well-Architected utilizzano soluzioni multiple e impiegano funzionalità diverse per migliorare le prestazioni.

Le risorse AWS sono disponibili in numerose tipologie e configurazioni, il che semplifica la ricerca di un approccio che soddisfi appieno le tue esigenze. Inoltre, puoi trovare opzioni che non sono facili da trovare nelle infrastrutture in locale. Ad esempio, un servizio gestito come Amazon

DynamoDB offre un database NoSQL interamente gestito, con una latenza di pochissimi millisecondi, indipendentemente dalle dimensioni.

Le seguenti domande si concentrano su queste considerazioni relative all'efficienza delle prestazioni. (Per l'elenco completo delle domande e delle best practice relative all'efficienza delle prestazioni, consulta l' Appendice.).

PERF 1 In che modo selezioni l'architettura più performante?

Spesso sono necessari molteplici approcci per ottenere prestazioni ottimali in un carico di lavoro. I sistemi Well-Architected utilizzano soluzioni multiple e funzionalità diverse per migliorare le prestazioni.

Quando selezioni i modelli e l'implementazione per la tua architettura, utilizza un approccio basato sui dati per individuare la soluzione ottimale. I solutions architect AWS, le architetture di riferimento AWS e i partner AWS Partner Network (APN) possono aiutarti a selezionare un'architettura in base alla conoscenza del settore, ma per ottimizzare la tua architettura saranno necessari i dati ottenuti da benchmark o test di carico.

La tua architettura può riunire vari approcci architetturali (ad esempio basati sugli eventi, ETL o pipeline). L'implementazione della tua architettura sfrutterà i servizi AWS in grado di ottimizzarne le prestazioni. Nelle sezioni seguenti, osserveremo quattro tipi di risorse principali da prendere in considerazione: elaborazione, storage, database e rete.

Calcolo

La selezione delle risorse di calcolo in grado di soddisfare i tuoi requisiti e le tue esigenze di prestazioni e offrire grande efficienza in termini di costi e impegno ti consentirà di ottenere di più con lo stesso numero di risorse. Durante la valutazione delle opzioni di elaborazione, tieni presente i requisiti per le prestazioni del carico di lavoro e i requisiti di costo e utilizzali per prendere decisioni informate.

In AWS, l'elaborazione è disponibile in tre forme: istanze, container e funzioni.

- Istanze sono server virtualizzati che consentono di modificare le loro funzionalità con un pulsante o una chiamata API. Poiché nel cloud le decisioni relative alle risorse non sono cristallizzate nel tempo, è possibile sperimentare vari tipi di server. In AWS, tali istanze di server virtuali sono disponibili in famiglie e dimensioni diverse e offrono un'ampia gamma di funzionalità, tra cui unità a stato solido (SSD) e unità di elaborazione grafica (GPU).

- Container sono un metodo di virtualizzazione del sistema operativo che consente di eseguire un'applicazione e le relative dipendenze in processi isolati dalle risorse. Puoi scegliere AWS Fargate, un servizio di elaborazione serverless per container, oppure Amazon EC2, se hai bisogno di controllare l'installazione, la configurazione e la gestione del tuo ambiente di elaborazione. Puoi anche scegliere tra diverse piattaforme di orchestrazione di container: Amazon Elastic Container Service (ECS) o Amazon Elastic Kubernetes Service (EKS).
- Funzioni astraggono l'ambiente di esecuzione dal codice che desideri eseguire. Ad esempio, AWS Lambda ti permette di eseguire del codice senza avviare un'istanza.

Le seguenti domande si concentrano su queste considerazioni relative all'efficienza delle prestazioni.

PERF 2 In che modo selezioni la tua soluzione di calcolo?

La soluzione di calcolo ottimale per un determinato carico di lavoro varia in base alla progettaz ione dell'applicazione, ai modelli di utilizzo e alle impostazioni di configurazione. Le architetture possono utilizzare diverse soluzioni di elaborazione per vari componenti e consentire funzioni diverse per migliorare le prestazioni. Selezionare la soluzione di calcolo sbagliata per un'archit ettura può portare a una riduzione dell'efficienza delle prestazioni.

Quando pianifichi l'utilizzo della capacità di elaborazione, devi sfruttare i meccanismi di elasticità per garantirti una capacità sufficiente a fornire le giuste prestazioni al variare delle esigenze.

Storage

Lo storage sul cloud è un componente fondamentale del cloud computing, poiché predisposto all'archiviazione delle informazioni utilizzate dal carico di lavoro. Lo storage sul cloud è generalmente più affidabile, scalabile e sicuro dei tradizionali sistemi di storage locali. Scegli tra servizi di storage di oggetti, blocchi e file, nonché opzioni di migrazione dei dati nel cloud per il tuo carico di lavoro.

In AWS, lo storage è disponibile in tre forme: oggetto, blocco e file:

- Archiviazione di oggetti fornisce una piattaforma scalabile e durevole per rendere i dati accessibili da qualsiasi posizione Internet per contenuti generati dagli utenti, archivi attivi, computing serverless, storage di Big Data o backup e ripristino. Amazon Simple Storage Service (Amazon S3) è un servizio di storage di oggetti che offre scalabilità, disponibilità dei dati, sicurezza e prestazioni leader di settore. Amazon S3 è progettato per garantire una durabilità del 99,999999999% (11 nove) e memorizza i dati per milioni di applicazioni per aziende in tutto il mondo.

- Lo storage a blocchi fornisce storage a blocchi a disponibilità elevata, costante e a bassa latenza per ogni host virtuale ed è analogo allo storage collegato direttamente (DAS) o a una rete SAN (Storage Area Network). Amazon Elastic Block Store (Amazon EBS) è stato progettato per carichi di lavoro che richiedono storage persistente accessibile dalle istanze EC2 e consente di ottimizzare le applicazioni con capacità di storage, prestazioni e costi ottimali.
- Lo storage di file fornisce accesso a un file system condiviso tra più sistemi. Le soluzioni di storage di file come Amazon Elastic File System (EFS) sono ideali per casi d'uso come repository di contenuti di grandi dimensioni, ambienti di sviluppo, store multimediali o home directory. Amazon FSx rende più semplice e conveniente l'avvio e l'esecuzione di file system diffusi in modo da sfruttare le funzionalità avanzate e le prestazioni rapide dei file system open source più utilizzati e con licenza commerciale.

Le seguenti domande si concentrano su queste considerazioni relative all'efficienza delle prestazioni.

PERF 3 In che modo selezioni la soluzione di storage?

La soluzione di storage ottimale per un sistema varia in base a fattori quali: tipo di metodo di accesso (blocco, file od oggetto), schemi di accesso (casuali o sequenziali), throughput necessari o, frequenza di accesso (online, offline, archivio), frequenza di aggiornamento (WORM, dinamico) e vincoli di disponibilità e durata. I sistemi Well-Architected utilizzano più soluzioni di storage e consentono funzionalità diverse per migliorare le prestazioni e utilizzare le risorse in modo efficient e.

Nella scelta di una soluzione di storage, accertarsi che sia in linea con gli schemi di accesso sarà cruciale per raggiungere le prestazioni desiderate.

Database

Il cloud offre servizi di database dedicati che risolvono i diversi problemi presentati dal carico di lavoro. Puoi scegliere tra diversi motori di database dedicati, tra cui database relazionali, chiave-valore, documento, in memoria, grafi, serie temporali e libri mastri. Scegliendo il database migliore per risolvere un problema specifico o una serie di problematiche, potrai finalmente abbandonare i database monolitici, restrittivi e indifferenziati e concentrarti sulla creazione di applicazioni in grado di rispondere alle esigenze di prestazioni dei tuoi clienti.

In AWS puoi scegliere tra più motori di database dedicati, tra cui database relazionali, chiave-valore, documento, in memoria, grafi, serie temporali e libri mastri. Con i database AWS, non

devi preoccuparti di attività di gestione dei database come il provisioning dei server, il patching, l'impostazione, la configurazione, i backup o il ripristino. AWS monitora continuamente i cluster per mantenere i carichi di lavoro attivi e funzionanti grazie allo storage auto-riparante e allo scaling automatico, in modo che tu possa concentrarti sullo sviluppo di applicazioni di maggior valore.

Le seguenti domande si concentrano su queste considerazioni relative all'efficienza delle prestazioni.

PERF 4 In che modo selezioni la soluzione di database?

La soluzione di database ottimale per un determinato sistema può variare in base ai requisiti di disponibilità, coerenza, tolleranza della partizione, latenza, durata, scalabilità e capacità di query. Molti sistemi utilizzano diverse soluzioni di database per vari sottosistemi e consentono funzional ità differenti per migliorare le prestazioni. Selezionare la soluzione e le funzionalità del database sbagliate per un sistema può ridurre l'efficienza delle prestazioni.

L'approccio al database del carico di lavoro ha un impatto significativo sull'efficienza delle prestazioni. Spesso è un'area scelta in base alle impostazioni predefinite dell'organizzazione anziché tramite un approccio basato sui dati. E a proposito di storage, è fondamentale prendere in considerazione gli schemi di accesso del tuo carico di lavoro, nonché valutare se altre soluzioni non basate su database potrebbero risolvere il problema in modo più efficiente (ad esempio utilizzare grafici, serie temporali o un database di storage in memoria).

Rete

Poiché la rete si trova tra tutti i componenti del carico di lavoro, può avere notevoli ripercussioni positive o negative sulle prestazioni e sul comportamento del carico di lavoro. Esistono anche carichi di lavoro che dipendono in larga misura dalle prestazioni di rete, come nel caso dello High Performance Computing (HPC), dove la comprensione approfondita della rete è importante per migliorare le prestazioni del cluster. È necessario determinare i requisiti del carico di lavoro per larghezza di banda, latenza, jitter e throughput.

In AWS, le reti sono virtualizzate e vengono fornite in molti diversi tipi e configurazioni. Ciò semplifica la scelta delle metodologie di rete più adatte alle tue esigenze. AWS offre caratteristiche di prodotto (ad esempio reti avanzate, istanze Amazon EBS ottimizzate per la rete, Amazon S3 Transfer Acceleration e Amazon CloudFront dinamico) pensate per l'ottimizzazione del traffico di rete. AWS offre anche funzionalità di rete (ad esempio instradamento in base alla latenza di Amazon Route 53, endpoint Amazon VPC, AWS Direct Connect e AWS Global Accelerator) per ridurre la distanza di rete o il jitter.

Le seguenti domande si concentrano su queste considerazioni relative all'efficienza delle prestazioni.

PERF 5 In che modo configuri la soluzione di rete?

La soluzione di rete ottimale per un carico di lavoro varia in base a latenza, requisiti di throughput, jitter e larghezza di banda. I vincoli fisici, ad esempio le risorse utente o in locale, determinano le opzioni di posizione. Questi vincoli possono essere compensati con le edge location o la collocazione delle risorse.

È necessario considerare la posizione quando si distribuisce la rete. Puoi decidere di collocare le risorse vicino al punto in cui saranno utilizzate per ridurre la distanza. Utilizza i parametri di rete per apportare modifiche alla configurazione di rete a mano a mano che il carico di lavoro si evolve. Sfruttando elementi quali regioni, gruppi di collocamento e servizi edge, avrai modo di incrementare le prestazioni in maniera significativa. Le reti basate sul cloud possono essere ricostruite o modificate rapidamente, perciò, per mantenere l'efficienza delle prestazioni, l'architettura di rete deve evolvere nel tempo.

Revisione

Le tecnologie cloud sono in rapida evoluzione e devi assicurarti che i componenti del carico di lavoro utilizzino nuove tecnologie e approcci per migliorare continuamente le prestazioni. Devi continuamente valutare e prendere in considerazione le modifiche apportate ai componenti del carico di lavoro per assicurarti di raggiungere gli obiettivi di prestazioni e costi. Le nuove tecnologie, come il machine learning e l'intelligenza artificiale (IA), ti permettono di ridefinire le esperienze dei clienti e di innovare tutti i tuoi carichi di lavoro aziendali.

Sfrutta l'innovazione continua di AWS, orientata alle esigenze dei clienti. Rilasciamo nuove regioni, edge location, servizi e funzionalità a intervalli regolari. Le nuove versioni possono migliorare sensibilmente l'efficienza delle prestazioni della tua architettura.

Le seguenti domande si concentrano su queste considerazioni relative all'efficienza delle prestazioni.

PERF 6 In che modo fai evolvere il carico di lavoro per sfruttare le nuove versioni?

Quando si progettano carichi di lavoro, le opzioni tra cui scegliere sono limitate. Tuttavia, nel tempo diventano disponibili nuove tecnologie e nuovi approcci che potrebbero migliorare le prestazioni.

Le prestazioni scarse delle architetture sono in genere il risultato di un processo di revisione delle prestazioni inesistente o incompleto. Se le prestazioni dell'architettura sono insufficienti, implementare un processo di revisione delle prestazioni ti consentirà di applicare un ciclo PDCA (plan-do-check-act) di Deming per favorire un miglioramento iterativo.

Monitoraggio

Dopo avere implementato il carico di lavoro, è necessario monitorarne le prestazioni in modo da risolvere eventuali problemi prima che influiscano sui clienti. Occorre utilizzare i parametri di monitoraggio per attivare gli allarmi in caso di superamento delle soglie.

Amazon CloudWatch è un servizio di monitoraggio e osservazione che fornisce dati e informazioni utili per monitorare il carico di lavoro, rispondere alle variazioni delle prestazioni a livello di sistema, ottimizzare l'utilizzo delle risorse e ottenere una visione unificata dello stato operativo. CloudWatch raccoglie dati operativi e di monitoraggio sotto forma di log, parametri ed eventi da carichi di lavoro eseguiti su AWS e server on-premise. AWS X-Ray aiuta gli sviluppatori ad analizzare ed eseguire il debug della produzione e delle applicazioni distribuite. Con AWS X-Ray, puoi ottenere informazioni approfondite sulle prestazioni dell'applicazione, individuare le cause principali e identificare i colli di bottiglia delle prestazioni. Puoi utilizzare le informazioni ottenute per correggere rapidamente il funzionamento e mantenere le prestazioni del carico di lavoro sempre ottimali.

Le seguenti domande si concentrano su queste considerazioni relative all'efficienza delle prestazioni.

PERF 7 In che modo monitori le risorse per garantirne le prestazioni?

Le prestazioni del sistema possono peggiorare nel tempo. Monitora le prestazioni del sistema per identificare l'eventuale riduzione delle prestazioni e rimediare a fattori interni o esterni, come il sistema operativo o il carico dell'applicazione.

Garantire che non vengano visualizzati falsi positivi è fondamentale per una soluzione di monitoraggio efficace. Le attivazioni automatiche prevengono l'errore umano e possono ridurre il tempo necessario per la risoluzione dei problemi. Pianifica giornate di gioco in cui vengono eseguite simulazioni nell'ambiente di produzione, per testare la soluzione di allarme e verificare che riconosca correttamente i problemi.

Compromessi

Quando progetti le soluzioni, pondera i compromessi per garantire una strategia ottimale. A seconda della situazione, puoi accettare dei compromessi in termini di coerenza, durabilità e spazio e favorire il tempo o la latenza allo scopo di garantire prestazioni migliori.

AWS ti consente di raggiungere la disponibilità globale in pochi minuti e distribuire le risorse in più destinazioni nel mondo, al fine di operare a più stretto contatto con gli utenti finali. Inoltre, puoi aggiungere in modo dinamico repliche di sola lettura alle destinazioni di storage delle informazioni, come i sistemi di database, per ridurre il carico sul database principale.

Le seguenti domande si concentrano su queste considerazioni relative all'efficienza delle prestazioni.

PERF 8 A quali compromessi ricorri per migliorare le prestazioni?

Quando si progettano soluzioni, determinare i compromessi ti consente di selezionare un approccio ottimale. Spesso è possibile migliorare le prestazioni accettando compromessi in termini di coerenza, durata e spazio a favore di tempo e latenza.

Man mano che apporti modifiche al carico di lavoro, raccogli e valuta i parametri per stabilire l'impatto dei cambiamenti. Misura gli impatti sul sistema e sugli utenti finali per capire in che modo i compromessi adottati influiscono sul carico di lavoro. Adotta un approccio sistematico, come il test del carico, per valutare se i compromessi migliorano le prestazioni.

Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice relative all'efficienza delle prestazioni.

Documentazione

- Ottimizzazione delle prestazioni di Amazon S3
- Prestazioni dei volumi di Amazon EBS

Whitepaper

- Il principio dell'efficienza delle prestazioni

Video

- AWS re:Invent 2019: Amazon EC2 foundations (CMP211-R2)
- AWS re:Invent 2019: Leadership session: Storage state of the union (STG201-L)
- AWS re:Invent 2019: Leadership session: AWS purpose-built databases (Database dedicati AWS) (DAT209-L)
- AWS re:Invent 2019: Connectivity to AWS and hybrid AWS network architectures (NET317-R1)
- AWS re:Invent 2019: Powering next-gen Amazon EC2: Deep dive into the Nitro system (CMP303-R2)
- AWS re:Invent 2019: Scaling up to your first 10 million users (Dimensionare le risorse per i primi 10 milioni di utenti) (ARC211-R)

Ottimizzazione dei costi

Il principio dell'ottimizzazione dei costi include la possibilità di eseguire sistemi per offrire valore aggiunto al prezzo più basso.

Il principio dell'ottimizzazione dei costi offre una panoramica dei principi di progettazione, delle best practice e delle domande. È possibile trovare linee guida prescrittive sull'implementazione nel Whitepaper sul principio dell'ottimizzazione dei costi.

Argomenti

Principi di progettazione

I principi di progettazione per l'ottimizzazione dei costi nel cloud sono cinque:

- Implementa la gestione finanziaria del cloud: per migliorare i risultati finanziari e accelerare la realizzazione del valore aziendale nel cloud, devi investire nella gestione finanziaria e nell'ottimizzazione dei costi sul cloud. L'organizzazione deve dedicare tempo e risorse per creare capacità in questo nuovo dominio di gestione della tecnologia e dell'utilizzo. Come per le

tue funzionalità di sicurezza o eccellenza operativa, devi creare capacità tramite lo sviluppo di competenze, programmi, risorse e processi, per diventare un'organizzazione efficiente in termini di costi.

- Adotta un modello a consumo: paga solo le risorse di calcolo che richiedi e incrementa o riduci l'utilizzo a seconda dei requisiti aziendali, e non attraverso il ricorso a una previsione elaborata. Ad esempio, gli ambienti di test e di sviluppo sono generalmente usati solo per otto ore al giorno durante la settimana lavorativa. Puoi interrompere queste risorse quando non le utilizzi, risparmiando potenzialmente il 75% dei costi (40 ore anziché 168).
- Misura l'efficienza complessiva: misura il risultato aziendale del carico di lavoro e i costi associati alla sua produzione. Usa questi dati per conoscere i ricavi che ottieni grazie all'aumento della produttività e alla riduzione dei costi.
- Smetti di spendere denaro per onerose attività indifferenziate: AWS si occupa delle attività onerose dei data center come il racking, lo stacking e l'alimentazione dei server. Inoltre, elimina l'onere operativo della gestione di sistemi operativi e applicazioni con servizi gestiti. In questo modo, potrai concentrarti sui tuoi clienti e sui progetti aziendali anziché sull'infrastruttura IT.
- Analizza e attribuisci la spesa: il cloud ti aiuta a individuare con facilità e precisione l'utilizzo e il costo dei sistemi, il che consente quindi l'attribuzione trasparente dei costi IT per i singoli proprietari del carico di lavoro. Questo ti aiuta a misurare il ritorno sull'investimento (ROI) e offre ai proprietari del carico di lavoro la possibilità di ottimizzare le proprie risorse e ridurre i costi.

Definizione

Esistono cinque aree di best practice per l'ottimizzazione dei costi nel cloud:

- Implementazione della gestione finanziaria del cloud
- Consapevolezza delle spese e dell'utilizzo
- Risorse convenienti
- Gestione delle risorse di domanda e offerta
- Ottimizzazione nel tempo

Come per gli altri principi di base all'interno del Canone di architettura, occorre considerare alcuni compromessi; ad esempio, è meglio ottimizzare la velocità di commercializzazione o i costi? In alcuni casi, è meglio ottimizzare la velocità: entrare nel mercato rapidamente, distribuire nuove caratteristiche o semplicemente rispettare una scadenza piuttosto che investire nell'ottimizzazione

anticipata dei costi. Talvolta le decisioni di progettazione sono guidate dalla rapidità invece che dai dati, ed esiste sempre la tentazione di sovrascrivere piuttosto che dedicare tempo all'esecuzione di benchmark per la distribuzione più conveniente. Questo potrebbe portare a distribuzioni sovra-assegnate e sotto-ottimizzate. Tuttavia, si tratta di una scelta ragionevole quando devi trasferire le risorse dal tuo ambiente locale al cloud ed eseguire l'ottimizzazione di conseguenza. Investire in anticipo la giusta quantità di energia in una strategia di ottimizzazione dei costi consente di realizzare i vantaggi economici del cloud in modo più rapido, assicurando il rispetto costante delle best practice ed evitando un provisioning superfluo. Le sezioni seguenti forniscono tecniche e best practice per l'implementazione iniziale e continua della gestione finanziaria del cloud e l'ottimizzazione dei costi dei carichi di lavoro.

Best practice

Argomenti

Implementazione della gestione finanziaria del cloud

Con l'adozione del cloud, i team tecnologici innovano più rapidamente grazie a cicli di approvazione, approvvigionamento e distribuzione dell'infrastruttura più brevi. Per ottenere valore aggiunto e migliorare gli affari è necessario un nuovo approccio alla gestione finanziaria nel cloud. Questo approccio è la gestione finanziaria del cloud e crea capacità in tutta l'organizzazione implementando competenze, programmi, risorse e processi a livello organizzativo.

Molte organizzazioni sono composte da tante unità con priorità diverse. La capacità di allineare un'organizzazione a un insieme concordato di obiettivi finanziari e di fornire all'organizzazione i meccanismi per raggiungerli permette di creare un'organizzazione più efficiente. Un'organizzazione capace innova e crea più rapidamente, è più agile e si adatta a qualsiasi fattore interno o esterno.

In AWS puoi utilizzare Cost Explorer e, facoltativamente, Amazon Athena e Amazon QuickSight con il report costi e utilizzo (CUR) per fornire consapevolezza su costi e utilizzo in tutta l'organizzazione. Budget AWS fornisce notifiche proattive relative a costi e utilizzo. I blog AWS forniscono informazioni

su nuovi servizi e caratteristiche per consentirti di essere sempre aggiornato sulle nuove versioni dei servizi.

La seguente domanda si concentra su queste considerazioni relative all'ottimizzazione dei costi. (Per l'elenco completo delle domande e delle best practice relative all'ottimizzazione dei costi, consulta l' Appendice.).

COST 1 In che modo implementi la gestione finanziaria nel cloud?

L'implementazione della gestione finanziaria del cloud consente alle organizzazioni di conseguire un valore aggiunto e il successo finanziario ottimizzando i costi e l'utilizzo e ricalibrando le risorse in AWS.

Quando crei una funzione di ottimizzazione dei costi, puoi utilizzare i membri e integrare il team con esperti di gestione finanziaria del cloud e ottimizzazione dei costi. I membri già presenti nel team conoscono il funzionamento dell'organizzazione e sono in grado di implementare rapidamente i miglioramenti. Valuta anche la possibilità di includere persone con competenze aggiuntive o specialistiche, ad esempio di analisi e gestione dei progetti.

Quando implementi la consapevolezza dei costi nella tua organizzazione, prova a migliorare o sviluppare i programmi e i processi esistenti. È molto più veloce sviluppare i processi e programmi esistenti, piuttosto che crearne di nuovi. In questo modo puoi ottenere risultati molto più rapidamente.

Consapevolezza delle spese e dell'utilizzo

La maggiore flessibilità e agilità consentite dal cloud incoraggiano l'innovazione, lo sviluppo e la distribuzione rapidi. Elimina i processi manuali e il tempo associati al provisioning dell'infrastruttura locale, tra cui l'identificazione delle specifiche hardware, la negoziazione delle quotazioni dei prezzi, la gestione degli ordini di acquisto, la pianificazione delle spedizioni e la distribuzione delle risorse. Tuttavia, la facilità d'uso e la capacità on demand virtualmente illimitata richiedono un nuovo tipo di mentalità in merito alle spese.

Molte aziende sono caratterizzate da più sistemi gestiti da vari team. La capacità di attribuire i costi delle risorse ai singoli proprietari dell'organizzazione o del prodotto incoraggia un comportamento di utilizzo efficiente e contribuisce a ridurre gli sprechi. L'attribuzione precisa dei costi consente di capire quali prodotti sono effettivamente redditizi e permette anche di prendere decisioni più consapevoli in merito alle destinazioni del budget.

Con AWS puoi creare una struttura di account con AWS Organizations o AWS Control Tower per garantire la separazione e semplificare l'allocazione di costi e utilizzo. Puoi anche utilizzare l'applicazione di tag alle risorse per associare informazioni aziendali e organizzative a utilizzo e costi. Utilizza AWS Cost Explorer per osservare costi e utilizzo, oppure crea analisi e pannelli di controllo personalizzati con Amazon Athena e Amazon QuickSight. Puoi verificare costi e utilizzo con le notifiche di Budget AWS e controllarli usando AWS Identity and Access Management (IAM) e Service Quotas.

Le seguenti domande si concentrano su queste considerazioni relative all'ottimizzazione dei costi.

COST 2 In che modo gestisci l'utilizzo?

Stabilisci policy e meccanismi per assicurarti di sostenere costi adeguati mentre raggiungi gli obiettivi. Utilizzando un approccio di controllo e bilanciamento reciproco, è possibile innovare senza spendere troppo.

COST 3 In che modo monitori l'utilizzo e il costo?

Stabilisci policy e procedure per monitorare e allocare i costi in modo appropriato. Ciò ti consente di misurare e migliorare l'efficienza in termini di costi del carico di lavoro.

COST 4 In che modo ritiri le risorse?

Implementa il controllo del cambiamento e la gestione delle risorse dall'inizio del progetto alla fine del ciclo di vita. In questo modo, puoi chiudere o interrompere le risorse non utilizzate per ridurre gli sprechi.

Puoi usare i tag di allocazione dei costi per categorizzare e monitorare il tuo utilizzo di AWS e i costi. Quando applichi dei tag alle tue risorse AWS (come le istanze EC2 o i bucket S3), AWS genera un report su costi e utilizzo con i tuoi tag e i dati sul tuo utilizzo. Puoi applicare tag che rappresentano le categorie di un'organizzazione (come i centri di costo, i nomi dei carichi di lavoro o i proprietari) per organizzare i tuoi costi tra i vari servizi.

Assicurati di utilizzare il giusto livello di dettaglio e granularità quando crei report e monitori costi e utilizzo. Per informazioni e tendenze generali, utilizza i dati giornalieri di AWS Cost Explorer. Per

analisi e ispezioni più specifiche, utilizza i dati orari di AWS Cost Explorer, oppure Amazon Athena e Amazon QuickSight impostando un livello di granularità oraria nel Report costi e utilizzo.

Associando le risorse taggate al monitoraggio del ciclo di vita dell'entità (dipendenti, progetti), puoi individuare le risorse accantonate o i progetti che non generano più valore per l'organizzazione e devono quindi essere dismessi. Puoi impostare avvisi di fatturazione per ricevere notifiche relative a spese eccessive previste.

Risorse convenienti

Utilizzare risorse e istanze adeguate al tuo carico di lavoro è fondamentale per ridurre i costi. Ad esempio, un processo di reporting potrebbe impiegare cinque ore su un server più piccolo, ma un'ora su un server più grande che costa il doppio. Entrambi i server ti offrono lo stesso risultato, ma quello più piccolo comporta un costo più elevato nel tempo.

Un carico di lavoro basato sul Canone di architettura AWS si basa sulle risorse più convenienti, il che può avere un impatto economico positivo e notevole. Hai anche la possibilità di usare i servizi gestiti per ridurre i costi. Ad esempio, invece di mantenere dei server per recapitare le e-mail, puoi usare un servizio che ti invia gli addebiti in base ai messaggi inviati.

AWS offre un'ampia gamma di offerte flessibili e convenienti per acquisire istanze da Amazon EC2 e altri servizi per soddisfare al meglio le tue necessità. On demand Istanze on demand ti consentono di pagare la capacità di elaborazione a ore e non richiedono impegni minimi. Savings Plans e istanze riservate offrono risparmi fino al 75% rispetto ai prezzi on demand. Con le istanze Spot, puoi sfruttare la capacità inutilizzata di Amazon EC2 e risparmiare fino al 90% sui prezzi on demand. Istanze Spot risultano adeguate quando il sistema può tollerare l'utilizzo di un parco server in cui i singoli server possano andare e venire dinamicamente, come server Web stateless, elaborazioni batch o quando si usano HPC e Big Data.

Anche la scelta del servizio appropriato può ridurre l'utilizzo e i costi; ad esempio, CloudFront può ridurre al minimo il trasferimento dei dati o eliminare del tutto i costi, mentre l'utilizzo di Amazon Aurora su RDS può rimuovere gli elevati costi di licenza dei database.

Le seguenti domande si concentrano su queste considerazioni relative all'ottimizzazione dei costi.

COST 5 In che modo valuti i costi quando selezioni i servizi?

Amazon EC2, Amazon EBS e Amazon S3 sono servizi AWS di base. I servizi gestiti, come Amazon RDS e Amazon DynamoDB, sono servizi AWS di livello superiore o applicativo. Seleziona

COST 5 In che modo valuti i costi quando selezioni i servizi?

ndo i blocchi predefiniti e i servizi gestiti appropriati, è possibile ottimizzare questo carico di lavoro per i costi. Ad esempio, utilizzando i servizi gestiti, puoi ridurre o eliminare gran parte dei costi generali amministrativi e operativi, liberandotene per lavorare su applicazioni e attività correlate al tuo business.

COST 6 In che modo raggiungi gli obiettivi di costo quando selezioni il tipo, le dimensioni e il numero delle risorse?

Assicurati di scegliere la dimensione e il numero delle risorse appropriati per l'attività in questione. Selezionando il tipo, le dimensioni e il numero più convenienti, riduci al minimo gli sprechi.

COST 7 In che modo impieghi i modelli di prezzo per ridurre i costi?

Utilizza il modello di prezzo più appropriato per le tue risorse per ridurre al minimo le spese.

COST 8 In che modo pianifichi i costi per il trasferimento dei dati?

Assicurati di pianificare e monitorare i costi di trasferimento dei dati in modo da poter prendere decisioni sull'architettura per ridurre al minimo i costi. Una modifica piccola ma efficace dell'arch itettura può ridurre drasticamente i costi operativi nel tempo.

Scomponendo i costi durante la selezione del servizio e usando strumenti come Cost Explorer e AWS Trusted Advisor per esaminare con regolarità l'utilizzo di AWS, puoi monitorare attivamente il tuo utilizzo e modificare le implementazioni di conseguenza.

Gestione delle risorse di domanda e offerta

Quando passi al cloud, paghi solo ciò che ti occorre. Puoi fornire risorse in base alla domanda del carico di lavoro nel momento in cui sono necessarie, eliminando così la necessità di un provisioning superfluo costoso e dispendioso. Puoi anche gestire la domanda utilizzando tecniche come throttling, buffering o queuing per allentare la domanda e soddisfarla con meno risorse. In questo modo diminuirai i costi o li posticiperai con un servizio batch.

In AWS puoi predisporre automaticamente le risorse da associare alla domanda di carico di lavoro. Auto Scaling con strategie basate su domanda o tempo ti consente di aggiungere e rimuovere le risorse in base alle esigenze. Se riesci a prevedere le variazioni nella domanda, puoi risparmiare di più e assicurarti che le risorse corrispondano alle esigenze del tuo carico di lavoro. Puoi utilizzare Amazon API Gateway per implementare il throttling o Amazon SQS per implementare una coda nel carico di lavoro. Entrambi consentono di modificare la richiesta nei componenti del carico di lavoro.

La seguente domanda si concentra su queste considerazioni relative all'ottimizzazione dei costi.

COST 9 Come gestisci la domanda e fornisci le risorse?

Per avere un carico di lavoro con costo e prestazioni bilanciate, assicurati che venga utilizzato tutto ciò per cui paghi ed evita le istanze molto sottoutilizzate. Un parametro di utilizzo distorto, in qualsiasi delle suddette direzioni, ha un impatto negativo sull'organizzazione, sia per i costi operativi (basse prestazioni a causa di un utilizzo eccessivo) che per le spese AWS sprecate (a causa di un provisioning eccessivo).

Quando progetti di modificare le risorse di domanda e offerta, pensa attentamente ai modelli di utilizzo, al tempo necessario per effettuare il provisioning delle nuove risorse e alla prevedibilità del modello di domanda. Quando gestisci la domanda, assicurati di disporre di una coda o di un buffer di dimensioni corrette e di rispondere alla domanda del carico di lavoro nel periodo di tempo richiesto.

Ottimizzazione nel tempo

Poiché AWS rilascia nuovi servizi e caratteristiche, è buona prassi rivedere le decisioni correnti sull'architettura per garantire che continuino a essere le più convenienti. Man mano che le tue esigenze cambiano, disattiva tempestivamente risorse, interi servizi e sistemi non appena smettono di essere necessari.

L'implementazione di nuove caratteristiche o tipi di risorse può ottimizzare il carico di lavoro in modo incrementale e con uno sforzo minimo. In questo modo puoi migliorare continuamente l'efficienza nel tempo e essere sicuro di utilizzare le tecnologie più aggiornate per ridurre i costi operativi. Puoi anche sostituire o aggiungere nuovi componenti al carico di lavoro con nuovi servizi. In questo modo puoi aumentare in modo significativo l'efficienza, perciò è essenziale rivedere regolarmente il carico di lavoro e implementare nuovi servizi e caratteristiche.

La seguente domanda si concentra su queste considerazioni relative all'ottimizzazione dei costi.

COST 10 In che modo valuti i nuovi servizi?

Poiché AWS rilascia nuovi servizi e caratteristiche, è buona prassi rivedere le decisioni correnti sull'architettura per garantire che continuino a essere le più convenienti.

Quando esamini regolarmente le tue distribuzioni, valuta in che modo i servizi più recenti possono aiutarti a risparmiare. Ad esempio, Amazon Aurora su RDS può ridurre i costi dei database relazionali. L'utilizzo di serverless come Lambda consente di eliminare la necessità di utilizzare e gestire le istanze per eseguire il codice.

Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle nostre best practice per l'ottimizzazione dei costi.

Documentazione

- Documentazione di AWS

Whitepaper

- Principio dell'ottimizzazione dei costi

Sostenibilità

Alla base del concetto di Sostenibilità c'è l'attenzione all'impatto ambientale, soprattutto in termini di uso ed efficienza delle fonti energetiche, leve importanti che gli architetti usano per definire interventi diretti mirati a ridurre lo sfruttamento delle risorse. È possibile trovare linee guida prescrittive sull'implementazione nel Whitepaper sul principio della sostenibilità.

Argomenti

- Principi di progettazione
- Definizione
- Best practice

Principi di progettazione

Esistono sei principi di progettazione per la sostenibilità nel cloud:

- Comprendi il tuo impatto: Misura l'impatto del tuo carico di lavoro cloud e definisci il suo impatto futuro. Nella tua analisi includi ogni fonte di impatto: quelle derivanti dall'uso dei prodotti da parte dei tuoi clienti e quelle derivanti dalla rimozione e dal ritiro finali dal mercato. Confronta l'output di produzione e l'impatto totale dei tuoi carichi di lavoro cloud, partendo dall'analisi di risorse ed emissioni richieste per unità di lavoro. Usa questi dati per definire indicatori chiave di prestazione (KPI), capire come migliorare la produttività, riducendo al tempo stesso l'impatto, e stimare l'impatto delle modifiche proposte nel tempo.
- Stabilisci obiettivi di sostenibilità: Per ogni carico di lavoro cloud stabilisci obiettivi di sostenibilità a lungo termine, come, ad esempio, ridurre le risorse di calcolo e di archiviazione richieste per ciascuna transazione. Modella il ritorno sugli investimenti finalizzati alle migliorie in materia di sostenibilità per i carichi di lavoro esistenti e offri ai proprietari le risorse di cui hanno bisogno per investire negli obiettivi di sostenibilità. Pianifica lo sviluppo e progetta i tuoi carichi di lavoro in modo che la crescita comporti un impatto meno intenso se misurato rispetto a un'unità appropriata, come l'utente o la transazione. Gli obiettivi ti aiutano ad avvalorare un progetto più ampio di sostenibilità che coinvolge la tua azienda o la tua organizzazione, a identificare le regressioni e a dare la priorità a quelle aree che offrono un maggiore potenziale di miglioramento.
- Ottimizza l'utilizzo: Dimensiona correttamente i carichi di lavoro e implementa un progetto efficiente in grado di garantire un utilizzo elevato e ottimizzare l'efficienza energetica dell'hardware sottostante. Due host in esecuzione con una percentuale di utilizzo pari al 30% sono meno efficienti di un host in esecuzione al 60%, se consideriamo il consumo di base per host. Allo stesso tempo, elimina o riduci le risorse, le elaborazioni e le archiviazioni inattive per ridurre l'energia totale richiesta per alimentare il tuo carico di lavoro.
- Anticipa e adotta offerte hardware e software nuove e più efficienti: Promuovi le migliorie a monte di partner e fornitori finalizzate a ridurre l'impatto dei carichi di lavoro cloud. Monitora costantemente il mercato e valuta nuove offerte hardware e software più efficienti. Adotta la flessibilità nei tuoi progetti per consentire una rapida adozione di tecnologie nuove ed efficienti.
- Utilizza servizi gestiti: La condivisione dei servizi con un'ampia base clienti consente di ottimizzare l'uso delle risorse e ridurre al tempo stesso l'infrastruttura necessaria per supportare i carichi di lavoro nel cloud. I clienti possono ad esempio condividere l'impatto di componenti comuni di data center, come reti ed energia, migrando i carichi di lavoro su Cloud AWS e adottando servizi gestiti, come AWS Fargate per i container serverless, in cui AWS opera su vasta scala ed è responsabile della loro efficienza operativa. Utilizza i servizi gestiti per contribuire alla riduzione dell'impatto,

trasferendo automaticamente dati con accesso poco frequente all'archiviazione dei dati inattivi con le configurazioni di Amazon S3 Lifecycle o di Amazon EC2 Auto Scaling per adeguare le capacità alla domanda.

- Riduci l'impatto a valle dei carichi di lavoro cloud: Diminuisci la quantità di energia o di risorse richieste per l'utilizzo dei tuoi servizi. Riduci o elimina la necessità di eseguire upgrade dei dispositivi per consentire ai clienti di usare i tuoi servizi. Esegui test usando device farm per analizzare l'impatto atteso e conduci altri test con i clienti per capire l'impatto reale derivante dall'uso dei tuoi servizi.

Definizione

Esistono sei aree di best practice per la sostenibilità nel cloud:

- Selezione delle regioni
- Modelli di comportamento degli utenti
- Modelli di software e architetture
- Modelli di dati
- Modelli hardware
- Processo di sviluppo e implementazione

Sostenibilità nel cloud significa impegnarsi continuamente per ridurre principalmente il consumo di energia e garantire una maggiore efficienza di tutti i componenti di un carico di lavoro, ottenendo il massimo vantaggio dalle risorse fornite e riducendo al minimo le quantità richieste. Tale impegno va dalla selezione iniziale di un linguaggio di programmazione efficace, dall'adozione di algoritmi moderni e dall'uso di tecniche di archiviazione di dati efficienti alla distribuzione in infrastrutture di calcolo valide e correttamente dimensionate e alla riduzione dei requisiti per l'hardware degli utenti finali a potenza elevata.

Best practice

Argomenti

- Modelli hardware
- Modelli di sviluppo e implementazione
- Risorse

Selezione delle regioni

Scegli le Regioni in cui implementerai i tuoi carichi di lavoro, tenendo presenti sia i requisiti aziendali sia gli obiettivi di sostenibilità.

La seguente domanda si concentra su queste considerazioni relative alla sostenibilità. (Per l'elenco completo delle domande e delle best practice relative all'affidabilità, consulta l' Appendice.)

SUS 1: In che modo selezioni le Regioni per sostenere i tuoi obiettivi di sostenibilità?

Scegli le Regioni vicino ai progetti di energia rinnovabile di Amazon e le Regioni in cui la griglia presenta un'intensità di emissione di anidride carbonica nota inferiore a quella di altre sedi (o Regioni).

Modelli di comportamento degli utenti

Il modo in cui gli utenti utilizzano i tuoi carichi di lavoro e altre risorse può aiutarti a identificare i miglioramenti da implementare per raggiungere gli obiettivi di sostenibilità. Dimensiona l'infrastruttura in modo che si adegui continuamente al carico degli utenti e implementa solo le risorse minime richieste per supportare gli utenti. Allinea i livelli di servizio alle esigenze dei clienti. Posiziona le risorse in modo da limitare la rete richiesta per il consumo da parte degli utenti. Elimina risorse esistenti non utilizzate. Identifica le risorse create non utilizzate e smetti di generarle. Offri ai membri del tuo team dispositivi in grado di soddisfare le loro esigenze con un impatto ridotto in termini di sostenibilità.

La seguente domanda si concentra su queste considerazioni relative alla sostenibilità:

SUS 2: In che modo sfrutti i modelli di comportamento degli utenti per sostenere i tuoi obiettivi di sostenibilità?

Il modo in cui gli utenti utilizzano i tuoi carichi di lavoro e altre risorse può aiutarti a identific
are i miglioramenti da implementare per raggiungere gli obiettivi di sostenibilità. Dimensiona l'infrastruttura in modo che si adegui continuamente al carico degli utenti e implementa solo

SUS 2: In che modo sfrutti i modelli di comportamento degli utenti per sostenere i tuoi obiettivi di sostenibilità?

le risorse minime richieste per supportare gli utenti. Allinea i livelli di servizio alle esigenze dei clienti. Posiziona le risorse in modo da limitare la rete richiesta per il consumo da parte degli utenti. Elimina risorse esistenti non utilizzate. Identifica le risorse create non utilizzate e smetti di generarle. Offri ai membri del tuo team dispositivi in grado di soddisfare le loro esigenze con un impatto ridotto in termini di sostenibilità.

Dimensiona l'infrastruttura in base al carico degli utenti: identifica i periodi di utilizzo assente o ridotto e riduci le risorse per evitare capacità in eccesso e migliorare il livello di efficienza.

Allinea gli SLA agli obiettivi di sostenibilità: definisci e aggiorna gli Accordi sul Livello di Servizio (SLA), come la disponibilità di periodi di conservazione dei dati, per ridurre il numero di risorse richieste a supporto dei carichi di lavoro, senza per questo venire meno ai requisiti di business.

Elimina la creazione e la manutenzione di asset inutilizzati: analizza le risorse delle applicazioni (come report precompilati, set di dati e immagini statiche) e i modelli di accesso alle risorse per identificare ridondanze, sottoutilizzi e obiettivi potenziali di disattivazione. Consolida le risorse generate con contenuti ridondanti (come, ad esempio, report mensili con set di dati e output comuni o in sovrapposizione) per eliminare le risorse utilizzate per la duplicazione degli output. Disattiva le risorse non utilizzate (come, ad esempio, immagini di prodotto non più in vendita) per liberare le risorse usate e ridurre il numero di risorse sfruttate per supportare il carico di lavoro.

Ottimizza il posizionamento geografico dei carichi di lavoro in base alle posizioni degli utenti: analizza i modelli di accesso alla rete per capire da quali aree geografiche si connettono i tuoi clienti. Seleziona le Regioni e i servizi per ridurre la distanza che il traffico di rete deve percorrere e diminuire così le risorse totali di rete richieste per supportare il tuo carico di lavoro.

Ottimizza le risorse dei membri del team in base alle attività eseguite: ottimizza le risorse fornite ai membri del team per ridurre al minimo l'impatto sulla sostenibilità e supportare al tempo stesso le loro esigenze. Esegui ad esempio operazioni complesse, come rendering e compilazione, su desktop cloud condivisi altamente utilizzati invece che su sistemi per utenti singoli, sottoutilizzati e con un alto dispendio energetico.

Modelli di software e architetture

Implementa modelli per eseguire lo smoothing del carico e garantire un utilizzo elevato e coerente delle risorse implementate per ridurre al minimo il loro consumo. In seguito alle modifiche nei

comportamenti degli utenti nel tempo, alcuni componenti potrebbero diventare inattivi per mancanza di utilizzo. Rivedi modelli e architetture per consolidare i componenti sottoutilizzati e aumentare l'uso complessivo. Ritira i componenti che non sono più necessari. Analizza le prestazioni dei componenti dei tuoi carichi di lavoro e ottimizza quelli che usano la maggior quantità di risorse. Identifica i dispositivi che i clienti utilizzano per accedere ai servizi e implementa modelli in grado di ridurre al minimo la necessità di aggiornamenti dei dispositivi.

Le seguenti domande si concentrano su queste considerazioni relative alla sostenibilità:

SUS 3: In che modo sfrutti i modelli di software e architetture per sostenere i tuoi obiettivi di sostenibilità?

Implementa modelli per eseguire lo smoothing del carico e garantire un utilizzo elevato e coerente delle risorse implementate per ridurre al minimo il loro consumo. In seguito alle modifiche nei comportamenti degli utenti nel tempo, alcuni componenti potrebbero diventare inattivi per mancanza di utilizzo. Rivedi modelli e architetture per consolidare i componenti sottoutilizzati e aumentare l'uso complessivo. Ritira i componenti che non sono più necessari. Analizza le prestazioni dei componenti dei tuoi carichi di lavoro e ottimizza quelli che usano la maggior quantità di risorse. Identifica i dispositivi che i clienti utilizzano per accedere ai servizi e implement a modelli in grado di ridurre al minimo la necessità di aggiornamenti dei dispositivi.

Ottimizza software e architetture per processi asincroni e pianificati: utilizza progettazioni e architetture software efficienti per ridurre al minimo le risorse medie richieste per unità di lavoro. Implementa meccanismi che generano un utilizzo uniforme dei componenti per ridurre le risorse inattive tra le attività e diminuire l'impatto di picchi di carico.

Rimuovi o rifattorizza i componenti dei carichi di lavoro con un utilizzo ridotto o assente: monitora l'attività dei carichi di lavoro per individuare i cambiamenti che si verificano nel tempo nell'utilizzo dei singoli componenti. Elimina i componenti non utilizzati e non più necessari e rifattorizza quelli con scarso utilizzo per limitare lo spreco di risorse.

Ottimizza le aree di codice che consumano la maggior parte del tempo o delle risorse: monitora l'attività dei carichi di lavoro per individuare i componenti delle applicazioni che usano la maggior parte delle risorse. Ottimizza il codice eseguito all'interno di questi componenti per ridurre l'utilizzo delle risorse e massimizzare al tempo stesso le prestazioni.

Ottimizza l'impatto su dispositivi e apparecchiature dei clienti: identifica i dispositivi e le attrezzature che i tuoi clienti usano per accedere ai tuoi servizi, il loro ciclo di vita atteso e l'impatto finanziario

e di sostenibilità che deriva dalla loro sostituzione. Implementa modelli e architetture software per ridurre al minimo la necessità dei clienti di sostituire dispositivi e aggiornare attrezzature. Implementa ad esempio nuove caratteristiche usando un codice compatibile con versioni di hardware e sistemi operativi precedenti o gestisci la dimensione dei payload in modo che non superino la capacità di archiviazione del dispositivo target.

Usa i modelli e le architetture software che meglio supportano l'accesso ai dati e i modelli di archiviazione: scopri come i dati vengono utilizzati all'interno del tuo carico di lavoro, consumati dagli utenti, trasferiti e archiviati. Seleziona tecnologie che ti consentono di ridurre l'elaborazione dei dati e i requisiti di archiviazione.

Modelli di dati

Implementa modelli per eseguire lo smoothing del carico e garantire un utilizzo elevato e coerente delle risorse implementate per ridurre al minimo il loro consumo. In seguito alle modifiche nei comportamenti degli utenti nel tempo, alcuni componenti potrebbero diventare inattivi per mancanza di utilizzo. Rivedi modelli e architetture per consolidare i componenti sottoutilizzati e aumentare l'uso complessivo. Ritira i componenti che non sono più necessari. Analizza le prestazioni dei componenti dei tuoi carichi di lavoro e ottimizza quelli che usano la maggior quantità di risorse. Identifica i dispositivi che i clienti utilizzano per accedere ai servizi e implementa modelli in grado di ridurre al minimo la necessità di aggiornamenti dei dispositivi.

La seguente domanda si concentra su queste considerazioni relative alla sostenibilità:

SUS 4: In che modo sfrutti i modelli di accesso e di utilizzo dei dati per sostenere i tuoi obiettivi di sostenibilità?

Implementa procedure di gestione dei dati per ridurre l'archiviazione assegnata richiesta per supportare il carico di lavoro e le risorse necessarie per l'uso correlato. Analizza i tuoi dati e usa tecnologie e configurazioni di archiviazione che meglio supportano il valore aziendale dei dati e il modo in cui vengono utilizzati. Esegui il ciclo di vita dei dati su un'archiviazione più efficiente e meno performante al diminuire dei requisiti ed elimina i dati che non sono più necessari.

Implementa una policy di classificazione dei dati: classifica i dati per comprenderne il significato in favore dei risultati aziendali. Usa queste informazioni per stabilire quando trasferire i dati in un'archiviazione più efficiente dal punto di vista energetico o eliminarli in totale sicurezza.

Utilizza tecnologie che supportano l'accesso ai dati e i modelli di archiviazione: usa l'archiviazione in grado di supportare al meglio il modo in cui viene effettuato l'accesso ai dati e come vengono archiviati per ridurre la quantità di risorse assegnate e supportare al tempo stesso il tuo carico di lavoro. I dispositivi allo stato solido (SSD) utilizzano ad esempio l'energia in modo più intensivo rispetto ai drive magnetici e dovrebbero essere usati solo per casi d'uso di dati attivi. Usa storage di classe di archiviazione ad alta efficienza energetica per i dati ad accesso infrequente.

Utilizza le policy del ciclo di vita per eliminare i dati non necessari: gestisci il ciclo di vita di tutti i tuoi dati e applica in automatico cronologie di eliminazione per ridurre i requisiti totali di archiviazione del tuo carico di lavoro.

Riduci il provisioning eccessivo nell'archiviazione a blocchi: per ridurre la quantità totale di archiviazione assegnata, crea un'archiviazione a blocchi con l'allocazione di dimensioni in base al carico di lavoro. Usa i volumi elastici per espandere l'archiviazione all'aumentare dei dati senza dover ridimensionare l'archiviazione collegata alle risorse di calcolo. Esamina regolarmente i volumi elastici e riduci i volumi con un provisioning eccessivo per adattarli alla dimensione corrente dei dati.

Elimina i dati ridondanti o non necessari: duplica i dati solo quando è necessario per ridurre la quantità totale di archiviazione utilizzata. Utilizza tecnologie di backup che deduplicano i dati a livello di file e blocco. Limita l'uso di configurazioni Redundant Array of Independent Drives (RAID), ad eccezione dei casi in cui sono richieste per soddisfare gli SLA.

Utilizza file system condivisi o archiviazione di oggetti per accedere a dati comuni: adotta l'archiviazione condivisa e singole fonti di verità per evitare la duplicazione dei dati e ridurre i requisiti di archiviazione complessiva del tuo carico di lavoro. Recupera i dati dall'archiviazione condivisa solo in base alle esigenze. Distacca volumi non utilizzati per liberare le risorse. Riduci al minimo gli spostamenti dei dati tra le reti: usa un'archiviazione condivisa e accedi ai dati da archivi regionali per contenere le risorse di rete totali necessarie per supportare i trasferimenti dei dati per il carico di lavoro.

Esegui il backup dei dati solo quando sono difficili da ricreare: per ridurre al minimo l'uso delle risorse di archiviazione, esegui il backup solo dei dati che abbiano un valore aziendale o siano considerati necessari per soddisfare requisiti di conformità. Esamina le policy di backup ed escludi l'archiviazione temporanea che non offre valore in uno scenario di ripristino.

Modelli hardware

Cerca opportunità per ridurre l'impatto dei carichi di lavoro in termini di sostenibilità apportando modifiche alle tue prassi di gestione hardware. Riduci la quantità di hardware necessaria per il provisioning e l'implementazione e seleziona l'hardware più efficiente per il singolo carico di lavoro.

La seguente domanda si concentra su queste considerazioni relative alla sostenibilità:

SUS 5: In che modo la gestione dell'hardware e le procedure di utilizzo sostengono i tuoi obiettivi di sostenibilità?

Cerca opportunità per ridurre l'impatto dei carichi di lavoro in termini di sostenibilità apportando modifiche alle tue prassi di gestione hardware. Riduci la quantità di hardware necessaria per il provisioning e l'implementazione e seleziona l'hardware più efficiente per il singolo carico di lavoro.

Utilizza la quantità minima di hardware per soddisfare le tue esigenze: le funzionalità del cloud consentono di apportare modifiche frequenti alle implementazioni dei carichi di lavoro. Aggiorna i componenti distribuiti man mano che le tue esigenze cambiano.

Usa tipi di istanze con il minimo impatto: monitora costantemente il rilascio di nuovi tipi di istanza e sfrutta le migliorie in tema di efficienza energetica, inclusi i tipi di istanza progettati per supportare carichi di lavoro specifici, come la formazione del machine learning, le inferenze e la transcodifica dei video.

Utilizza servizi gestiti: i servizi gestiti consentono di affidare ad AWS la responsabilità di mantenere un utilizzo medio alto e un'ottimizzazione della sostenibilità dell'hardware implementato. Utilizza i servizi gestiti per distribuire l'impatto della sostenibilità dei servizi su tutti i tenant relativi, riducendo così il singolo contributo.

Ottimizza l'utilizzo delle GPU: le Graphics Processing Unit (GPU) possono comportare un uso energetico intensivo e molti carichi di lavoro delle GPU sono altamente variabili, come il rendering, la transcodifica e la formazione e la modellazione del machine learning. Esegui le istanze GPU solo per il tempo necessario e disattivale automaticamente quando non occorrono per ridurre la quantità di risorse utilizzate.

Modelli di sviluppo e implementazione

Cerca opportunità per ridurre l'impatto di sostenibilità apportando modifiche alle tue prassi di sviluppo, test e implementazione.

La seguente domanda si concentra su queste considerazioni relative alla sostenibilità:

SUS 6: In che modo i processi di sviluppo e implementazione adottati supportano i tuoi obiettivi di sostenibilità?

Cerca opportunità per ridurre l'impatto di sostenibilità apportando modifiche alle tue prassi di sviluppo, test e implementazione.

Adotta metodi che consentono di integrare rapidamente i miglioramenti orientati alla sostenibilità: testa e convalida potenziali modifiche di miglioramento prima di distribuirle in produzione. Tieni in considerazione il costo dei test quando calcoli il potenziale vantaggio futuro di un miglioramento. Sviluppa metodi di test a basso costo per consentire la distribuzione di piccoli miglioramenti.

Mantieni aggiornato il tuo carico di lavoro: sistemi operativi, librerie e applicazioni aggiornati possono incidere sull'efficienza dei carichi di lavoro e facilitano l'adozione di tecnologie più efficienti. Il software aggiornato potrebbe anche includere funzionalità per misurare in modo più accurato l'impatto in termini di sostenibilità del carico di lavoro, poiché i fornitori offrono caratteristiche per raggiungere i propri obiettivi di sostenibilità.

Incrementa l'utilizzo degli ambienti di sviluppo: utilizza l'automazione e l'infrastruttura come codice per rendere operativi gli ambienti di preproduzione quando necessario e dismetterli quando non vengono utilizzati. Un modello comune consiste nel pianificare periodi di disponibilità che coincidano con l'orario di lavoro dei membri del team incaricati dello sviluppo. L'ibernazione è uno strumento utile per preservare lo stato e portare rapidamente le istanze online solo quando necessario. Utilizza tipi di istanze espandibili, istanze Spot, servizi di database elastici, container e altre tecnologie per allineare la capacità di sviluppo e test all'uso.

Utilizza device farm gestite per i test: le device farm gestiste distribuiscono l'impatto di sostenibilità della produzione di hardware e dell'utilizzo delle risorse su più tenant. Le device farm gestite offrono diversi tipi di dispositivi e consentono di supportare hardware meno diffusi e di generazioni precedenti e di evitare l'impatto sulla sostenibilità dei clienti dovuti ad aggiornamenti dei dispositivi non necessari.

Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice per la sostenibilità.

Whitepaper

- Principio della sostenibilità

Video

- The Climate Pledge

Il processo di revisione

La revisione delle architetture deve essere eseguita in modo coerente, con un approccio che non colpevolizza nessuno, ma che incoraggia ad approfondire gli argomenti. Dovrebbe essere un processo leggero (di ore, non di giorni) più simile a una conversazione che non a un audit. Lo scopo della revisione di un'architettura è identificare dei problemi critici da affrontare o aree di miglioramento. Il risultato della revisione è un insieme di azioni volte a migliorare l'esperienza di utilizzo del carico di lavoro del cliente.

Come discusso nella sezione "Architettura", ogni membro del team deve prendersi la responsabilità della qualità della sua architettura. Consigliamo che i membri del team che hanno sviluppato l'architettura usino il Canone di architettura per eseguire costantemente la revisione della loro architettura, piuttosto che fare una riunione di revisione formale. Un approccio continuo permette ai membri del team di aggiornare le risposte man mano che l'architettura evolve e migliorare l'architettura di pari passo alle funzionalità.

Il Framework AWS Well-Architected è allineato alla modalità interna di revisione dei sistemi e dei servizi di AWS. Si basa su un insieme di principi di progettazione che influenzano l'approccio architetturale e su domande che garantiscano che le persone non trascurino aree che spesso figurano nell'Analisi della causa principale (RCA). Ogni volta che si presenta un problema significativo con un sistema interno, un servizio AWS o un cliente, ci serviamo della RCA per vedere se possiamo migliorare il processo di revisione utilizzato.

Le revisioni devono essere applicate a tappe fondamentali nel ciclo di vita del prodotto, all'inizio della fase di progettazione per evitare decisioni unidirezionali che sono difficili da modificare prima della data di implementazione. (Molte decisioni sono reversibili e quindi bidirezionali. Queste decisioni possono usare un processo leggero. Le decisioni unidirezionali sono difficili o impossibile da annullare e richiedono un maggiore sforzo di indagine prima di essere adottate). Una volta entrato in produzione, il carico di lavoro continuerà ad evolversi man mano che si aggiungono nuove caratteristiche e si modificano le implementazioni tecnologiche. L'architettura del carico di lavoro cambia nel tempo. Devi seguire le best practice di igiene informatica per interrompere il degrado delle caratteristiche man mano che fai evolvere l'architettura. Man mano che l'architettura cambia, dovresti seguire un insieme di processi di igiene informatica tra cui la revisione Well-Architected.

Se vuoi utilizzare la revisione come snapshot una tantum o misura indipendente, dovrai assicurarti che alla conversazione partecipino tutte le persone appropriate. Spesso ci rendiamo conto che le revisioni sono il primo momento in cui il team comprende per davvero quello che ha implementato.

Un approccio che funziona bene per la revisione dei carichi di lavoro di un altro team consiste in una serie di conversazioni informali sull'architettura in cui ottenere le risposte alla maggior parte delle domande. Quindi puoi fare una o due riunioni di follow up in cui puoi fare chiarezza o approfondire le aree ambigue e il rischio percepito.

Ecco alcuni elementi suggeriti per le tue riunioni:

- Una sala riunioni con una lavagna
- Le stampe di tutti i grafici o delle note di progettazione
- Lista di azioni delle domande che richiedono risposte a ricerche fuori banda (ad esempio, "abbiamo abilitato la crittografia o no?")

Dopo avere completato la revisione, dovresti avere un elenco di problemi a cui assegnare delle priorità sulla base del contesto aziendale. Dovrai anche prendere in considerazione l'impatto di tali problemi sul lavoro quotidiano del tuo team. Se affronti questi problemi in anticipo puoi liberare del tempo per lavorare sulla creazione di valore aziendale anziché dedicarlo a risolvere i problemi ricorrenti. Man mano che affronti i problemi, puoi aggiornare la revisione per vedere in che modo l'architettura sta migliorando.

Il valore di una revisione è evidente dopo averne eseguita una, ma all'inizio un nuovo team potrebbe essere contrario. Ecco alcune obiezioni da gestire per istruire il team sui vantaggi di una revisione:

- "Siamo troppo occupati!" (spesso si sente questa frase quando il team si sta preparando a un grande lancio.)
  - Se ti stai preparando per un grande lancio, desidererai che tutto vada bene. La revisione ti aiuta a comprendere qualsiasi problema che potresti esserti perso.
  - Ti raccomandiamo di eseguire le revisioni all'inizio del ciclo di vita del prodotto per scoprire i rischi e sviluppare un piano di mitigazione allineato con la roadmap delle funzionalità.
- "Non abbiamo tempo per utilizzare i risultati!" (Spesso questo viene detto quando c'è un evento fisso, come il Super Bowl, di cui si sta occupando il team.)
  - Questi eventi non possono essere spostati. Vuoi davvero affrontare l'evento senza conoscere i rischi della tua architettura? Anche se non ti occupi di tutti i problemi in questione, puoi comunque disporre di playbook per affrontarli se si dovessero presentare.
- "Non vogliamo che altri scoprano i segreti della nostra implementazione di soluzioni!"
  - Se poni le domande del Framework Well-Architected, il team noterà che nessuna di esse rivela informazioni proprietarie commerciali o tecniche.

Eseguendo più revisioni con i team della tua organizzazione, potresti identificare delle aree tematiche. Ad esempio, potresti notare che un gruppo di team ha gruppi di problemi in un pilastro o un argomento specifico. Puoi gestire tutte le tue revisioni in modo olistico e identificare tutti i meccanismi, la formazione o le riunioni con gli ingegneri responsabili che possono aiutare a risolvere i problemi tematici.

Conclusione

Il Framework AWS Well-Architected best fornisce practice architetturali relative a sei pilastri per la progettazione e la gestione di sistemi affidabili, sicuri, efficienti, a costi contenuti e sostenibili nel cloud. Il canone fornisce un insieme di domande che ti permettono di eseguire la revisione di un'architettura esistente o proposta. Il canone fornisce anche un insieme di best practice AWS per ogni principio. L'utilizzo del canone nella tua architettura ti aiuta a produrre sistemi stabili ed efficienti, che ti permettono di concentrarti sui tuoi requisiti funzionali.

Collaboratori

Hanno contribuito alla stesura di questo documento:

- Brian Carlson, Operations Lead Well-Architected, Amazon Web Services
- Ben Potter, Security Lead Well-Architected, Amazon Web Services
- Seth Eliot, Reliability Lead Well-Architected, Amazon Web Services
- Eric Pullen, Sr. Solutions Architect, Amazon Web Services
- Rodney Lester, Principal Solutions Architect, Amazon Web Services
- Jon Steele, Senior Technical Account Manager, Amazon Web Services
- Max Ramsay, Principal Security Solutions Architect, Amazon Web Services
- Callum Hughes, Solutions Architect, Amazon Web Services
- Aden Leirer, Content Program Manager Well-Architected, Amazon Web Services

Approfondimenti

Centro di progettazione AWS

Conformità di AWS Cloud

Programma Partner AWS Well-Architected

AWS Well-Architected Tool

Homepage di Well-Architected AWS

Whitepaper sul principio dell'eccellenza operativa

Whitepaper sul principio della sicurezza

Whitepaper sul principio dell'affidabilità

Whitepaper sul principio dell'efficienza delle prestazioni

Whitepaper sul principio dell'ottimizzazione dei costi

Whitepaper sul principio della sostenibilità

Amazon Builders' Library

Revisioni del documento

Per ricevere una notifica sugli aggiornamenti di questo whitepaper, iscriviti al feed RSS.

Modifica	Descrizione	Data
Aggiornamento di minore entità	Aggiunta della definizione di livello di impegno e aggiornamento delle best practice nell'appendice.	October 20, 2022
Whitepaper aggiornato	Aggiunta del Principio della sostenibilità e collegamenti aggiornati.	December 2, 2021
Aggiornamento principale	Aggiunta al framework del principio della sostenibilità.	November 20, 2021
Aggiornamento di minore entità	Rimozione del linguaggio non inclusivo.	April 22, 2021
Aggiornamento di minore entità	Correzione di diversi collegamenti.	March 10, 2021
Aggiornamento di minore entità	Modifiche editoriali di minore entità in varie parti del documento.	July 15, 2020
Aggiornamenti per il nuovo canone	Revisione e riscrittura della maggior parte delle domande e delle risposte.	July 8, 2020
Whitepaper aggiornato	Aggiunta di AWS Well-Architected Tool, collegamenti ai corsi AWS Well-Architected Labs e ai partner AWS Well-Architected, correzioni minori	July 1, 2019

	per abilitare la versione in più lingue del canone.	
Whitepaper aggiornato	Revisione e riscrittura di molte domande e risposte per garantire che le domande si concentrino su un argomento alla volta. Per questo motivo, alcune delle domande precedenti sono state divise in più domande. Aggiunta di termini comuni alle definizioni (carichi di lavoro, componenti, ecc.). Presentazione delle domande modificata per includere il testo descrittivo.	November 1, 2018
Whitepaper aggiornato	Aggiornamenti volti a semplificare il testo delle domande e a migliorare la leggibilità.	June 1, 2018
Whitepaper aggiornato	Eccellenza operativa spostata all'inizio della sezione sui pilastri e riscritta in modo che inquadri gli altri pilastri. Aggiornamenti degli altri principi per riflettere l'evoluzione di AWS.	November 1, 2017
Whitepaper aggiornato	Framework aggiornato per includere i pilastri dell'eccellenza operativa; altri pilastri rivisti e aggiornati per ridurre la duplicazione e incorporare le nozioni apprese grazie alle revisioni eseguite con migliaia di clienti.	November 1, 2016

Aggiornamenti di minore entità	Aggiornamento dell'Appendice con informazioni aggiornate su Amazon CloudWatch Logs.	November 1, 2015
Pubblicazione originale	Pubblicazione del Framework AWS Well-Architected.	October 1, 2015

Appendice: domande e best practice

Argomenti

Eccellenza operativa

Argomenti

Organizzazione

Domande

OPS 1 In che modo stabilisci quali sono le tue priorità?

È necessario che ognuno capisca il proprio ruolo per rendere possibile il successo aziendale. Devi disporre di obiettivi comuni al fine di stabilire le priorità per le risorse. Ciò massimizzerà i risultati dei tuoi sforzi.

Best practice

OPS01-BP01 Valutazione delle esigenze dei clienti esterni

Coinvolgi i principali stakeholder, compresi i team aziendali, di sviluppo e operativi, per determinare dove concentrare gli sforzi in base alle esigenze dei clienti esterni. Questo ti garantirà una conoscenza approfondita del supporto operativo necessario per raggiungere i risultati aziendali desiderati.

Anti-pattern comuni:

- Hai deciso di non fornire il servizio clienti al di fuori degli orari di attività principali, ma non hai esaminato i dati cronologici riguardanti le richieste di supporto. Non sai se questo determinerà un impatto sui tuoi clienti.
- Stai sviluppando una nuova funzionalità, ma non hai coinvolto i tuoi clienti per capire se è desiderata e, se sì, in quale forma; inoltre non hai condotto attività di sperimentazione per convalidarne la necessità e il metodo di distribuzione.

Vantaggi dell'adozione di questa best practice: I clienti le cui esigenze sono soddisfatte hanno maggiori probabilità di rimanere clienti. Valutando e comprendendo le esigenze dei clienti esterni sarà possibile organizzare le attività in base a priorità e offrire valore aggiunto.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Comprendi le esigenze aziendali: il successo dell'azienda è reso possibile dalla condivisione di obiettivi e dalla comprensione tra gli stakeholder, compresi i team aziendali, di sviluppo e operativi.
  - Esamina gli obiettivi aziendali, le esigenze e le priorità dei clienti esterni: coinvolgi i principali stakeholder, compresi i team aziendali, di sviluppo e operativi, per discutere obiettivi, esigenze

e priorità dei clienti esterni. In tal modo otterrai una conoscenza approfondita del supporto operativo necessario per raggiungere i risultati aziendali e del cliente.

- Stabilisci una comprensione condivisa: stabilisci una comprensione condivisa delle funzioni aziendali del carico di lavoro, dei ruoli di ciascuno dei team nel gestire il carico di lavoro e di come questi supportino i tuoi obiettivi aziendali condivisi tra clienti interni ed esterni.

Risorse

Documenti correlati:

- Concetti del Pilastro del Framework AWS Well-Architected – Ciclo di feedback

OPS01-BP02 Valutazione delle esigenze dei clienti interni

Coinvolgi i principali stakeholder, compresi i team aziendali, di sviluppo e operativi, nel determinare dove concentrare le attività in base alle esigenze dei clienti interni. Questo ti garantirà una conoscenza approfondita del supporto operativo necessario per raggiungere i risultati aziendali.

Utilizza le priorità così definite per concentrare le tue iniziative di miglioramento delle operazioni laddove avranno il maggiore impatto (ad esempio, sviluppare le competenze dei team, migliorare le prestazioni del carico di lavoro, ridurre i costi, automatizzare le istruzioni o potenziare il monitoraggio). Aggiorna le tue priorità al mutare delle esigenze.

Anti-pattern comuni:

- Per semplificare la gestione della rete hai deciso di modificare l'assegnazione degli indirizzi IP per i team di prodotto senza consultarli. Non conosci l'impatto che questo avrà sui tuoi team di prodotto.
- Stai implementando un nuovo strumento di sviluppo, ma non hai coinvolto i clienti interni per scoprire se è necessario o se è compatibile con le loro pratiche esistenti.
- Stai implementando un nuovo sistema di monitoraggio, ma non hai contattato i clienti interni per scoprire se hanno esigenze di monitoraggio o reporting da tenere in considerazione.

Vantaggi dell'adozione di questa best practice: Valutando e comprendendo le esigenze dei clienti interni sarà possibile organizzare le attività in base a priorità e offrire valore aggiunto.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Comprendi le esigenze aziendali: il successo dell'azienda è reso possibile dalla condivisione di obiettivi e dalla comprensione tra gli stakeholder, compresi i team aziendali, di sviluppo e operativi.
  - Esamina gli obiettivi aziendali, le esigenze e le priorità dei clienti interni: coinvolgi i principali stakeholder, compresi i team aziendali, di sviluppo e operativi, per discutere obiettivi, esigenze e priorità dei clienti interni. In tal modo otterrai una conoscenza approfondita del supporto operativo necessario per raggiungere i risultati aziendali e del cliente.
  - Stabilisci una comprensione condivisa: stabilisci una comprensione condivisa delle funzioni aziendali del carico di lavoro, dei ruoli di ciascuno dei team nel gestire il carico di lavoro e di come questi supportino gli obiettivi aziendali condivisi tra clienti interni ed esterni.

Risorse

Documenti correlati:

- Concetti del Pilastro del Framework AWS Well-Architected – Ciclo di feedback

OPS01-BP03 Valutazione dei requisiti di governance

Assicurati di conoscere le linee guida o gli obblighi definiti dalla tua organizzazione che possono imporre o enfatizzare l'attenzione ad aspetti specifici. Valuta i fattori interni, come policy, standard e requisiti dell'organizzazione. Accertati di disporre di meccanismi per identificare le modifiche alla governance. Se non vengono identificati requisiti di governance, assicurati che sia stata applicata la dovuta diligenza per giungere a questa conclusione.

Anti-pattern comuni:

- La tua attività è sottoposta ad audit e ti viene chiesto di fornire una prova di conformità alla governance interna. Non hai idea se sei conforme perché non hai mai valutato quali sono i tuoi requisiti di conformità.
- La tua attività è stata fortemente danneggiata e ciò ha causato una perdita finanziaria. Scopri che l'assicurazione che avrebbe coperto la perdita finanziaria era subordinata all'implementazione di controlli di sicurezza specifici che non sono stati implementati e richiesti dalla governance.
- Il tuo account amministrativo è stato compromesso e, di conseguenza, il sito Web aziendale è stato danneggiato, con conseguenze negative sulla fiducia dei clienti. La governance interna richiede

l'uso dell'autenticazione a più fattori (MFA) per proteggere gli account amministrativi. Non hai protetto il tuo account amministrativo con MFA e rischi un'azione disciplinare.

Vantaggi dell'adozione di questa best practice: Valutando e comprendendo i requisiti di governance che la tua organizzazione applica al carico di lavoro capirai come dare priorità alle tue attività e offrire valore aggiunto.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Comprendi i requisiti di governance: valuta i fattori di governance interni, come le policy organizzative o di programma e le policy riguardanti problemi o sistemi specifici, gli standard, le procedure, le baseline e le linee guida. Accertati di disporre di meccanismi per identificare le modifiche alla governance. Se non vengono identificati requisiti di governance, assicurati che sia stata applicata la dovuta diligenza per giungere a questa conclusione.

Risorse

Documenti correlati:

- Conformità di Cloud AWS

OPS01-BP04 Valutazione dei requisiti di conformità

Valuta i fattori esterni, come i requisiti di conformità normativa e gli standard di settore, per assicurarti di conoscere le linee guida o gli obblighi che potrebbero imporre o sottolineare attenzione ad aspetti specifici. Se non vengono identificati requisiti di conformità, assicurati di applicare la dovuta diligenza a questa determinazione.

Anti-pattern comuni:

- La tua attività è sottoposta a audit e ti viene chiesto di fornire una prova di conformità alle normative di settore. Non hai idea se sei conforme perché non hai mai valutato quali sono i tuoi requisiti di conformità.
- Il tuo account amministrativo è stato compromesso e, di conseguenza, si ciò ha causato il download dei dati dei clienti, con conseguenze negative sulla fiducia da parte di questi ultimi. Le best practice di settore richiedono l'uso di MFA per proteggere gli account amministrativi. Non hai protetto il tuo account amministrativo con MFA e rischi che si verifichino contenziosi con i clienti.

Vantaggi dell'adozione di questa best practice: Valutando e comprendendo i requisiti di conformità che si applicano carico di lavoro sarà possibile organizzare le attività in base a priorità e offrire valore aggiunto.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Comprendi i requisiti di conformità: valuta i fattori esterni, come i requisiti di conformità normativa e gli standard di settore, per assicurarti di conoscere le linee guida o gli obblighi che potrebbero imporre o sottolineare attenzione ad aspetti specifici. Se non vengono identificati requisiti di conformità, assicurati che sia stata applicata la dovuta diligenza alla determinazione.
  - Comprendi i requisiti di conformità normativa: individua i requisiti di conformità normativa che sei legalmente obbligato a soddisfare. Usa questi requisiti per focalizzare le tue attività. Tra gli esempi vi sono gli obblighi derivanti da leggi sulla privacy e sulla protezione dei dati.
    - Conformità di AWS
    - Programmi per la conformità di AWS
    - Ultime novità sulla conformità di AWS
  - Comprendi gli standard e le best practice di settore: identifica gli standard di settore e i requisiti di best practice applicabili al carico di lavoro, ad esempio lo standard PCI DSS (Payment Card Industry Data Security Standard). Usa questi requisiti per focalizzare le tue attività.
    - Programmi per la conformità di AWS
  - Comprendi i requisiti di conformità interna: identifica i requisiti di conformità e le best practice stabilite dalla tua organizzazione. Usa questi requisiti per focalizzare le tue attività. Alcuni esempi includono policy di sicurezza delle informazioni e standard di classificazione dei dati.

Risorse

Documenti correlati:

- Conformità di Cloud AWS
- Conformità di AWS
- Ultime novità sulla conformità di AWS
- Programmi per la conformità di AWS

OPS01-BP05 Valutazione del panorama delle minacce

Valuta le minacce per l'azienda (ad esempio, concorrenza, rischi e responsabilità aziendali, rischi operativi e minacce per la sicurezza delle informazioni) e conserva le informazioni aggiornate in un registro dei rischi. Quando stabilisci dove concentrare gli sforzi, tieni in considerazione l'impatto dei rischi.

Il Canone di architettura AWS enfatizza l'apprendimento, la misurazione e il miglioramento. Fornisce una strategia coerente per la valutazione delle architetture e l'implementazione di progetti in grado di ridimensionarsi nel corso del tempo. AWS mette a disposizione AWS Well-Architected Tool per aiutarti a rivedere il tuo approccio prima dello sviluppo e lo stato dei tuoi carichi di lavoro prima e durante la fase di produzione. Puoi confrontare il tuo approccio con le best practice architetturali AWS più recenti, monitorare lo stato complessivo dei carichi di lavoro e ottenere informazioni sui potenziali rischi.

I clienti AWS possono usufruire della revisione Well-Architected dei carichi di lavoro mission-critical per valutare le loro architetture rispetto alle best practice di AWS. I clienti con supporto Enterprise hanno diritto alla revisione delle operazioni, ideata per agevolare l'identificazione delle eventuali lacune nel loro approccio al cloud.

Il coinvolgimento trasversale dei team per tali controlli aiuta a comprendere a livello comune i carichi di lavoro e come i ruoli del team contribuiscano al successo. Le esigenze identificate nel corso dell'analisi possono aiutarti a definire le tue priorità.

AWS Trusted Advisor è uno strumento che fornisce l'accesso a una serie di controlli di base che propongono ottimizzazioni utili per la definizione delle tue priorità. I clienti del supporto Business ed Enterprise hanno accesso a ulteriori controlli a livello di sicurezza, affidabilità, prestazioni e ottimizzazione dei costi che possono essere utili per definire le loro priorità.

Anti-pattern comuni:

- Stai utilizzando una versione precedente di una libreria software nel tuo prodotto. Non sei a conoscenza di aggiornamenti di sicurezza alla libreria per problemi che potrebbero avere un impatto imprevisto sul carico di lavoro.
- Il tuo concorrente ha appena rilasciato una versione del proprio prodotto che risolve i reclami di molti dei tuoi clienti relativi al tuo prodotto. Non hai dato priorità alla risoluzione di questi problemi noti.

- Le autorità di regolamentazione hanno perseguito aziende come la tua che non sono conformi ai requisiti di conformità alla normativa legale. Non hai dato priorità ai requisiti di conformità in sospeso.

Vantaggi dell'adozione di questa best practice: Identificando e comprendendo le minacce alla tua organizzazione e al tuo carico di lavoro potrai determinare quali minacce affrontare, la loro priorità e le risorse necessarie per farlo.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Valuta il panorama delle minacce: valute le minacce per l'azienda (ad esempio, concorrenza, rischi e responsabilità aziendali, rischi operativi e minacce alla sicurezza delle informazioni), in modo da poterne includere l'impatto nel determinare dove concentrare le attività.
  - Ultimi bollettini di sicurezza AWS
  - AWS Trusted Advisor
  - Mantieni un modello delle minacce: definisci e mantieni un modello delle minacce che identifichi le potenziali minacce, le mitigazioni pianificate e predisposte e la loro priorità. Esamina la probabilità che le minacce si manifestino come incidenti, il costo del recupero da tali incidenti, il danno previsto causato e il costo per prevenire tali incidenti. Modifica le priorità man mano che i contenuti del modello di minaccia cambiano.

Risorse

Documenti correlati:

- Conformità di Cloud AWS
- Ultimi bollettini di sicurezza AWS
- AWS Trusted Advisor

OPS01-BP06 Valutazione dei compromessi

Valuta l'impatto dei compromessi tra interessi concorrenti o approcci alternativi, per aiutare a prendere decisioni informate quando si stabilisce dove concentrare le attività o scegliere una linea di azione. Ad esempio, accelerare l'introduzione sul mercato di nuove funzionalità può essere preferibile all'ottimizzazione dei costi. Oppure, è possibile scegliere un database relazionale per i dati non

relazionali per semplificare la migrazione di un sistema, anziché migrare a un database ottimizzato per il tuo tipo di dati e aggiornare l'applicazione.

AWS può aiutarti a istruire i tuoi team su AWS e i suoi servizi, affinché comprendano meglio in che modo le loro scelte possono influire sul carico di lavoro. Per istruire i tuoi team, è consigliabile utilizzare le risorse fornite da AWS Support (Knowledge Center di AWS, forum di discussione di AWSe AWS Support Center) e la Documentazione di AWS . Se hai domande riguardanti AWS, contatta AWS Support tramite AWS Support Center.

AWS condivide inoltre le best practice e i modelli appresi attraverso la gestione di AWS nella Amazon Builders' Library. Un'ampia gamma di altre informazioni utili è disponibile tramite il Blog AWS e il podcast ufficiale di AWS.

Anti-pattern comuni:

- Stai utilizzando un database relazionale per gestire serie temporali e dati non relazionali. Esistono opzioni di database ottimizzate per supportare i tipi di dati che stai utilizzando, ma non ne conosci i vantaggi perché non hai valutato i compromessi tra le soluzioni.
- I tuoi investitori richiedono di dimostrare la conformità agli standard PCI DSS (Payment Card Industry Data Security Standard). Non prendi in considerazione i compromessi tra soddisfare la loro richiesta e continuare con le attività di sviluppo già in corso. Al contrario, prosegui con il lavoro di sviluppo senza dimostrare la conformità. I tuoi investitori interrompono il supporto della tua azienda per i dubbi relativi alla sicurezza della tua piattaforma e ai loro investimenti.

Vantaggi dell'adozione di questa best practice: Comprendere le implicazioni e le conseguenze delle tue scelte ti consente di dare priorità alle tue opzioni.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Valuta i compromessi: valuta l'impatto dei compromessi tra interessi concorrenti, per aiutare a prendere decisioni informate nel determinare dove concentrare le attività. Ad esempio, è possibile accelerare la velocità di introduzione sul mercato di nuove funzionalità rispetto all'ottimizzazione dei costi.
- AWS può aiutarti a istruire i tuoi team su AWS e i suoi servizi, affinché comprendano meglio in che modo le loro scelte possono influire sul carico di lavoro. Per istruire i tuoi team, è consigliabile utilizzare le risorse fornite da AWS Support (AWS Knowledge Center, AWS Discussion Forms e

AWS Support Center) e la documentazione AWS. Se hai domande riguardanti AWS, contatta AWS Support tramite AWS Support Center.

- AWS condivide inoltre le best practice e i modelli appresi attraverso la gestione di AWS nella Amazon Builders' Library. Un'ampia gamma di altre informazioni utili è disponibile tramite il blog AWS e il podcast ufficiale di AWS.

Risorse

Documenti correlati:

- Blog AWS
- Conformità di Cloud AWS
- forum di discussione di AWS
- Documentazione di AWS
- Knowledge Center di AWS
- AWS Support
- AWS Support Center
- Amazon Builders' Library
- il podcast ufficiale di AWS

OPS01-BP07 Gestione dei vantaggi e dei rischi

Gestisci i vantaggi e i rischi per prendere decisioni informate nel determinare dove concentrare gli sforzi. Ad esempio, può essere vantaggioso distribuire un sistema con problemi irrisolti, in modo da mettere a disposizione dei clienti nuove funzionalità importanti. Può essere possibile ridurre i rischi associati o la presenza di un rischio potrebbe diventare inaccettabile, nel qual caso si intraprenderà un'azione per risolverlo.

Ad esempio, a un certo punto potresti realizzare che desideri dare maggiore risalto a un piccolo sottoinsieme delle tue priorità. Utilizza un approccio equilibrato nel lungo termine per garantire lo sviluppo delle capacità necessarie e la gestione del rischio. Aggiorna le tue priorità al mutare delle esigenze

Anti-pattern comuni:

- Hai deciso di includere una libreria che fa "tutto quello che ti serve" che uno dei tuoi sviluppatori "ha trovato su Internet". Non hai valutato i rischi di adottare questa libreria da un'origine sconosciuta e non sai se contiene vulnerabilità o codice dannoso.
- Hai deciso di sviluppare e distribuire una nuova funzionalità anziché risolvere un problema esistente. Non hai valutato i rischi posti dal fatto che il problema persiste finché la funzionalità non viene distribuita e non sai quale impatto avrà sui tuoi clienti.
- Hai deciso di non distribuire una funzionalità richiesta frequentemente dai clienti a causa di dubbi non specificati dal team di conformità.

Vantaggi dell'adozione di questa best practice: Identificare i vantaggi offerti dalle tue scelte e conoscere i rischi per la tua organizzazione ti consente di prendere decisioni informate.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Gestione dei vantaggi e dei rischi: bilancia i vantaggi delle decisioni rispetto ai rischi connessi.
  - Identificazione dei vantaggi: identifica i vantaggi in base agli obiettivi aziendali, alle esigenze e alle priorità. Gli esempi includono il time-to-market, la sicurezza, l'affidabilità, le prestazioni e il costo.
  - Identificazione dei rischi: identifica i rischi in base agli obiettivi aziendali, alle esigenze e alle priorità. Gli esempi includono il time-to-market, la sicurezza, l'affidabilità, le prestazioni e il costo.
  - Valutazione dei vantaggi rispetto ai rischi e decisioni informate: determina l'impatto dei vantaggi e dei rischi in base agli obiettivi, alle esigenze e alle priorità dei tuoi principali stakeholder, inclusi business, sviluppo e operazioni. Valuta il valore del vantaggio rispetto alla probabilità di realizzazione del rischio e al costo del suo impatto. Ad esempio, enfatizzare la velocità di accesso al mercato rispetto all'affidabilità potrebbe offrire un vantaggio competitivo. Tuttavia, potrebbe causare tempi di attività ridotti in presenza di problemi di affidabilità.

OPS 2 In che modo strutturi la tua organizzazione per supportare i risultati aziendali?

I tuoi team devono comprendere quale contributo offrono nel raggiungimento dei risultati aziendali. I team devono avere obiettivi condivisi e devono comprendere il proprio ruolo nel successo degli altri team. Comprendere la responsabilità, la proprietà, il modo in cui vengono prese le decisioni e chi ha l'autorità decisionale aiuterà a concentrare gli sforzi e a ottimizzare i contributi dei team.

Best practice

- OPS02-BP01 Associazione di proprietari identificati alle risorse
- OPS02-BP02 Assegnazione di proprietari identificati a processi e procedure
- OPS02-BP03 Assegnazione di proprietari identificati alle operazioni che siano responsabili delle relative prestazioni
- OPS02-BP04 Conoscenza della propria responsabilità da parte dei membri del team
- OPS02-BP05 Definizione di meccanismi per identificare responsabilità e proprietà
- OPS02-BP06 Definizione di meccanismi per richiedere aggiunte, modifiche ed eccezioni
- OPS02-BP07 Predefinizione o negoziazione delle responsabilità tra i team

OPS02-BP01 Associazione di proprietari identificati alle risorse

È utile comprendere chi ha la proprietà di ogni applicazione, carico di lavoro, piattaforma e componente dell'infrastruttura, qual è il valore aziendale fornito da tale componente e perché tale proprietà esiste. Comprendere il valore aziendale di questi singoli componenti e il modo in cui supportano i risultati aziendali fornisce indicazioni sui processi e le procedure applicati.

Vantaggi dell'adozione di questa best practice: Capire a chi spetta la proprietà permette di identificare chi può approvare e/o implementare i miglioramenti.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Associazione di proprietari identificati alle risorse: definisci qual è il significato della proprietà per i casi d'uso delle risorse nel tuo ambiente. Specifica e registra i proprietari delle risorse, specificando come minimo nome, informazioni di contatto, organizzazione e team. Archivia le informazioni sulla proprietà delle risorse insieme alle risorse utilizzando metadati come tag o gruppi di risorse. Utilizza AWS Organizations per strutturare gli account e implementare policy, in modo da assicurare che le informazioni sulla proprietà e le informazioni di contatto vengano acquisite.
  - Definizione delle forme di proprietà e delle modalità di assegnazione: la proprietà può avere più definizioni nella tua organizzazione con diversi casi d'uso. Il "proprietario del carico di lavoro" può essere definito come l'individuo a cui spettano il rischio e la responsabilità per il funzionamento di un carico di lavoro e che, in ultima analisi, ha l'autorità di prendere decisioni su di esso. È possibile definire la proprietà in termini di responsabilità finanziaria o amministrativa quando la proprietà viene trasferita a un'organizzazione padre. Uno sviluppatore può essere il proprietario del proprio ambiente di sviluppo ed essere responsabile degli incidenti provocati

dal suo funzionamento. Il responsabile del prodotto può essere responsabile dei costi finanziari associati al funzionamento dei propri ambienti di sviluppo.

- Definizione del proprietario di un'organizzazione, un account, una raccolta di risorse o singoli componenti: definisci e registra la proprietà in un luogo adeguatamente accessibile e organizzato per supportare la ricerca. Aggiornare le definizioni e i dettagli riguardanti la proprietà man mano che cambiano.
- Inclusione della proprietà nei metadati in relazione alle risorse: acquisisci la proprietà delle risorse utilizzando metadati come tag o gruppi di risorse, specificando proprietà e informazioni di contatto. Utilizza AWS Organizations per strutturare gli account e garantire che vengano acquisite le informazioni relative a contatto e proprietà.

OPS02-BP02 Assegnazione di proprietari identificati a processi e procedure

È utile comprendere chi ha la proprietà della definizione di singoli processi e procedure, perché tali processi e procedure specifici vengono utilizzati e perché tale proprietà esiste. Comprendere i motivi per cui vengono utilizzati processi e procedure specifici consente di identificare le opportunità di miglioramento.

Vantaggi dell'adozione di questa best practice: Capire a chi spetta la proprietà permette di identificare chi può approvare e/o implementare i miglioramenti.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Assegnazione di proprietari identificati a processi e procedure: acquisisci i processi e le procedure utilizzati nel tuo ambiente e il singolo o il team responsabile della loro definizione.
  - Identificazione di processi e procedure: identifica le attività operative eseguite a supporto dei carichi di lavoro. Documenta queste attività in un percorso individuabile.
  - Definizione del proprietario della determinazione di un processo o di una procedura: identifica in modo univoco la persona o il team responsabile della specifica di un'attività. Questo soggetto deve assicurare che essa possa essere eseguita correttamente dal componente di un team adeguatamente qualificato, che disponga di autorizzazioni, accesso e strumenti adeguati. In caso di problemi nello svolgimento di tale attività, i membri del team che la eseguono sono responsabili di fornire il feedback dettagliato necessario per migliorarla.
  - Inclusione della proprietà nei metadati dell'artefatto dell'attività: le procedure automatizzate di servizi quali AWS Systems Manager, tramite documenti, e AWS Lambda, come funzioni,

supportano l'acquisizione di informazioni sui metadati sotto forma di tag. Acquisisci la proprietà delle risorse utilizzando tag o gruppi di risorse, specificando proprietà e informazioni di contatto. Utilizza AWS Organizations per creare policy di tagging e garantire che vengano acquisite proprietà e informazioni di contatto.

OPS02-BP03 Assegnazione di proprietari identificati alle operazioni che siano responsabili delle relative prestazioni

È utile comprendere chi ha la responsabilità di eseguire attività specifiche su carichi di lavoro definiti e perché tale responsabilità esiste. Comprendere chi ha la responsabilità di eseguire le attività fornisce indicazioni su eseguirà l'attività, su chi convaliderà il risultato e su chi fornirà feedback al proprietario dell'attività.

Vantaggi dell'adozione di questa best practice: Comprendere chi è responsabile di eseguire un'attività fornisce indicazioni su chi notificare quando è necessaria un'azione, su chi la eseguirà, su chi convaliderà il risultato e su chi fornirà un feedback al proprietario dell'attività.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Assegnazione di proprietari identificati alle operazioni che siano responsabili delle relative prestazioni: acquisisci la responsabilità dell'esecuzione dei processi e delle procedure utilizzati nel tuo ambiente
  - Identificazione di processi e procedure: identifica le attività operative eseguite a supporto dei carichi di lavoro. Documenta queste attività in un percorso individuabile.
  - Definizione di chi è responsabile dell'esecuzione di ciascuna attività: identifica il team responsabile di un'attività. Assicurati che disponga dei dettagli dell'attività, delle competenze necessarie e di autorizzazioni, accesso e strumenti appropriati per svolgerla. Il team deve comprendere la condizione in cui deve essere eseguita (ad esempio, in un evento o in una pianificazione). Rendi queste informazioni individuabili in modo che i membri della tua organizzazione possano identificare chi contattare, team o individuale, per esigenze specifiche.

OPS02-BP04 Conoscenza della propria responsabilità da parte dei membri del team

Comprendere le responsabilità del tuo ruolo e il modo in cui contribuisci ai risultati aziendali fornisce indicazioni sulle priorità delle tue attività e sul perché il tuo ruolo è importante. In questo modo i membri del team possono riconoscere le esigenze e rispondere in modo appropriato.

Vantaggi dell'adozione di questa best practice: Comprendere le tue responsabilità fornisce indicazioni sulle decisioni che prendi, le azioni che intraprendi e le tue attività di distribuzione ai proprietari appropriati.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Comprensione da parte dei membri del team dei propri ruoli e responsabilità: identifica i ruoli e le responsabilità dei membri del team e assicurati che comprendano le aspettative del loro ruolo. Rendi queste informazioni individuabili in modo che i membri della tua organizzazione possano identificare chi contattare, team o individuale, per esigenze specifiche.

OPS02-BP05 Definizione di meccanismi per identificare responsabilità e proprietà

Quando non viene identificato alcun individuo o team, esistono percorsi di escalation definiti nei confronti di soggetti dotati dell'autorità per assegnare la proprietà o la pianificazione connesse al soddisfacimento dell'esigenza in questione.

Vantaggi dell'adozione di questa best practice: Comprendere chi ha la responsabilità o la proprietà ti permette di contattare il team o il componente del team appropriati per presentare una richiesta o trasferire un'attività. Avere una persona identificata che ha l'autorità di assegnare la responsabilità o la proprietà o che può pianificare il soddisfacimento delle esigenze riduce il rischio di inerzia e il pericolo che le esigenze non vengano affrontate.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Definizione di meccanismi per l'identificazione di responsabilità e proprietà: fornisci ai membri della tua organizzazione meccanismi accessibili per scoprire e identificare proprietà e responsabilità. Questo consentirà loro di identificare il team o l'individuo da contattare per esigenze specifiche.

OPS02-BP06 Definizione di meccanismi per richiedere aggiunte, modifiche ed eccezioni

È possibile effettuare richieste ai proprietari di processi, procedure e risorse. Prendi decisioni informate per approvare le richieste quando vengono ritenute fattibili e appropriate dopo una valutazione dei vantaggi e dei rischi.

Vantaggi dell'adozione di questa best practice: È fondamentale che esistano meccanismi per richiedere aggiunte, modifiche ed eccezioni a supporto delle attività dei team. Senza questa opzione, lo stato attuale diventa un vincolo per l'innovazione.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Definisci meccanismi per richiedere aggiunte, modifiche ed eccezioni: quando gli standard sono rigidi, l'innovazione è vincolata. Fornisci ai membri dell'organizzazione meccanismi per effettuare richieste ai proprietari di processi, procedure e risorse a supporto delle loro esigenze aziendali.

OPS02-BP07 Predefinizione o negoziazione delle responsabilità tra i team

Fai in modo che esistano accordi definiti o negoziati tra i team che descrivono come funzionano e si supportano reciprocamente (ad esempio, tempi di risposta, obiettivi o contratti relativi al livello di servizio). Comprendere l'impatto del lavoro dei team sui risultati aziendali e sui risultati di altri team e organizzazioni fornisce indicazioni in merito alla priorità dei loro compiti e consente loro di rispondere in modo appropriato.

Quando la responsabilità e la proprietà sono indefinite o sconosciute, rischi sia di non affrontare le attività necessarie in modo tempestivo sia di impiegare sforzi ridondanti e potenzialmente conflittuali per rispondere a tali esigenze.

Vantaggi dell'adozione di questa best practice: Stabilire le responsabilità tra i team, gli obiettivi e i metodi per comunicare le esigenze, facilita il flusso di richieste e garantisce che vengano fornite le informazioni necessarie. Questo riduce il ritardo introdotto dalle attività di transizione tra i team e aiuta a supportare il raggiungimento dei risultati aziendali.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Predefinisci o negozia le responsabilità tra i team: specifica i modi con cui i team interagiscono e le informazioni necessarie per supportarsi reciprocamente al fine di ridurre al minimo il ritardo introdotto man mano che le richieste vengono esaminate e chiarite iterativamente. Avere accordi specifici che definiscono le aspettative (ad esempio, il tempo di risposta o il tempo di realizzazione) consente ai team di elaborare piani e risorse efficaci in modo appropriato.

OPS 3 In che modo la cultura aziendale supporta i risultati aziendali?

Fornisci supporto ai membri del team in modo che possano essere più efficaci nell'azione e nel supporto dei risultati aziendali.

Best practice

- OPS03-BP01 Sponsorizzazione esecutiva
- OPS03-BP02 Potere di intervento dei membri del team quando i risultati sono a rischio
- OPS03-BP03 Incoraggiamento all'escalation
- OPS03-BP04 Comunicazioni tempestive, chiare e fruibili
- OPS03-BP05 Incoraggiamento alla sperimentazione
- OPS03-BP06 Autorizzazione e incoraggiamento ai membri del team a mantenere e ampliare le proprie competenze
- OPS03-BP07 Fornitura di risorse appropriate ai team
- OPS03-BP08 Incoraggiamento e ricerca di opinioni diverse all'interno e tra i team

OPS03-BP01 Sponsorizzazione esecutiva

Gli alti dirigenti stabiliscono chiaramente le aspettative per l'organizzazione e valutano il successo. Gli alti dirigenti sono promotori, sostenitori e motori per l'adozione delle best practice e l'evoluzione dell'organizzazione.

Vantaggi dell'adozione di questa best practice: Dirigenti coinvolti, aspettative chiare e obiettivi condivisi sono gli elementi necessari per far sì che i membri del team sappiano cosa ci si aspetta da loro. La valutazione del successo consente di identificare gli ostacoli che ne impediscono la riuscita e di superarli tramite l'intervento dei sostenitori promotori dell'iniziativa in questione o dei loro delegati.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Sponsorizzazione esecutiva: gli alti dirigenti stabiliscono chiaramente le aspettative per l'organizzazione e valutano il successo. Gli alti dirigenti sono promotori, sostenitori e motori per l'adozione delle best practice e l'evoluzione dell'organizzazione.
  - Definizione delle aspettative: definisci e pubblica gli obiettivi per le tue organizzazioni, incluso il modo in cui verranno misurati.

- Monitoraggio del raggiungimento degli obiettivi: misura regolarmente il raggiungimento incrementale degli obiettivi e condividi i risultati in modo che si possa intraprendere un'azione appropriata se i risultati sono a rischio.
- Disponibilità delle risorse necessarie per raggiungere gli obiettivi stabiliti: verifica regolarmente se le risorse sono ancora appropriate, se sono necessarie risorse aggiuntive in base a nuove informazioni o cambiamenti degli obiettivi, delle responsabilità o dell'ambiente aziendale.
- Sostegno ai team: mantieni un coinvolgimento attivo con i tuoi team in modo da comprendere come stanno e se ci sono fattori esterni che li influenzano. Quando i team sono influenzati da fattori esterni, rivaluta gli obiettivi e modifica i target in base alle esigenze. Individua gli ostacoli che impediscono l'avanzamento dei team. Agisci per conto dei tuoi team per superare gli ostacoli e rimuovere gli oneri superflui.
- Adozione delle best practice: riconosci le best practice che offrono vantaggi quantificabili e identifica creatori e destinatari. Incoraggia ulteriormente l'adozione per amplificare i vantaggi ottenuti.
- Evoluzione dei team: crea una cultura di costante miglioramento. Incoraggia la crescita e lo sviluppo sia personale sia organizzativo. Fornisci validi obiettivi a lungo termine da raggiungere in modo incrementale nel tempo. Adatta questa visione per soddisfare le tue esigenze, gli obiettivi aziendali e l'ambiente aziendale man mano che cambiano.

OPS03-BP02 Potere di intervento dei membri del team quando i risultati sono a rischio

Il proprietario del carico di lavoro definisce le linee guida e l'ambito consentendo ai membri del team di rispondere quando i risultati sono a rischio. I meccanismi di escalation vengono utilizzati ai fini dell'orientamento quando gli eventi sono al di fuori dell'ambito definito.

Vantaggi dell'adozione di questa best practice: Testando e convalidando le modifiche in anticipo, puoi risolvere i problemi con costi ridotti al minimo e limitare l'impatto sui clienti. Eseguendo il test prima della distribuzione, riduci al minimo la possibilità di errore.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Garantisci il potere di Intervento dei membri del team quando i risultati sono a rischio: fornisci ai membri del team le autorizzazioni, gli strumenti e l'opportunità per mettere in pratica le competenze necessarie per rispondere in modo efficace.

- Offri ai membri del team l'opportunità di mettere in pratica le competenze necessarie per rispondere: fornisci ambienti sicuri alternativi in cui testare i processi e sottoporre i membri del team alla dovuta formazione in modo sicuro. Esegui le giornate di simulazione per consentire ai membri del team di acquisire esperienza nel rispondere agli incidenti del mondo reale in ambienti simulati e sicuri.
- Definisci e riconosci l'autorità di intervento dei membri del team: definisci in modo specifico l'autorità di intervento dei membri del team assegnando le autorizzazioni e l'accesso ai carichi di lavoro e ai componenti supportati. Riconosci che i membri del team sono autorizzati a intervenire quando i risultati sono a rischio.

OPS03-BP03 Incoraggiamento all'escalation

I membri del team dispongono di meccanismi e sono incoraggiati a segnalare le preoccupazioni ai responsabili delle decisioni e agli stakeholder se ritengono che i risultati sono a rischio. L'escalation deve essere eseguita in anticipo e di frequente, in modo che i rischi possano essere identificati e limitati prima che provochino incidenti.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Incoraggia l'escalation anticipata e frequente: riconosci a livello organizzativo che l'escalation anticipata e frequente è la best practice. Riconosci a livello organizzativo e accetta che le escalation possono rivelarsi infondate e che è meglio avere l'opportunità di prevenire un incidente piuttosto che privarsi di quell'opportunità senza escalation.
  - Predisponi un meccanismo per l'escalation: è opportuno disporre di procedure documentate che definiscano quando e come deve verificarsi l'escalation. Documenta la serie di persone in ordine di autorità, cui è consentito intervenire o approvare un'azione e le relative informazioni di contatto. L'escalation deve continuare finché il membro del team non è soddisfatto di aver trasferito il rischio a una persona in grado di risolverlo o di aver contattato la persona che possiede il rischio e la responsabilità per il funzionamento del carico di lavoro. È quella persona che alla fine prende tutte le decisioni in relazione al carico di lavoro. Le escalation devono includere la natura del rischio, la criticità del carico di lavoro, le persone interessate, il grado di impatto e di urgenza, ovvero quando è previsto l'impatto.
  - Proteggi i dipendenti coinvolti nell'escalation: è necessario disporre di una policy che protegga i membri del team da eventuali penalizzazioni qualora si trovassero nelle condizioni di scavalcare

un decisore non reattivo o uno stakeholder. Metti in atto dei meccanismi per identificare se ciò si verifica e rispondere in modo appropriato.

OPS03-BP04 Comunicazioni tempestive, chiare e fruibili

Esistono meccanismi che vengono utilizzati per fornire tempestivamente notifiche ai membri del team in merito a rischi noti ed eventi pianificati. Laddove è possibile, vengono forniti contesto, dettagli e tempo per determinare se è necessario intervenire, in che modo e con quali tempistiche. Ad esempio, si può essere emettere un avviso di vulnerabilità del software in modo che le patch vengano applicate rapidamente, oppure si può fornire un avviso sulle promozioni di vendita pianificate al fine di bloccare le modifiche per evitare il rischio di interruzione del servizio.

Gli eventi pianificati possono essere registrati in un calendario delle modifiche o in un programma di manutenzione, in modo che i membri del team possano identificare quali attività sono in sospeso.

Su AWS, puoi utilizzare il calendario delle modifiche di AWS Systems Manager per registrare questi dettagli. Supporta controlli programmatici dello stato del calendario per determinare se il calendario è aperto o chiuso all'attività in un determinato momento. Le attività operative possono essere pianificate in base a specifiche finestre temporali approvate riservate alle attività potenzialmente causa di interferenze. AWS Systems Manager Maintenance Windows consente di pianificare le attività su istanze e altre risorse supportate per automatizzare le attività e renderle individuabili.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Comunicazioni tempestive, chiare e fruibili: sono stati attivati meccanismi per fornire tempestivamente notifiche in merito ai rischi o agli eventi pianificati in modo chiaro e fruibile al fine di consentire risposte appropriate.
  - Attività pianificate documentate in un calendario delle modifiche e notifiche correlate: fornisci un'origine di informazioni accessibile dove è possibile consultare gli eventi pianificati. Fornisci notifiche di eventi pianificati dallo stesso sistema.
  - Monitoraggio di eventi e attività impattanti il carico di lavoro: il monitoraggio delle notifiche di vulnerabilità e delle informazioni sulle patch consente di comprendere le vulnerabilità in circolazione e i potenziali rischi associati ai componenti del carico di lavoro. Invia notifiche ai membri del team in modo che possano intervenire.

Risorse

Documenti correlati:

- Il calendario delle modifiche di AWS Systems Manager
- finestre di manutenzione di AWS Systems Manager

OPS03-BP05 Incoraggiamento alla sperimentazione

La sperimentazione accelera l'apprendimento e mantiene acceso l'interesse e il coinvolgimento dei membri del team. Un risultato indesiderato è un esperimento riuscito tramite il quale viene identificato un percorso che non porterà al successo. I membri del team non vengono puniti per gli esperimenti riusciti con risultati indesiderati. La sperimentazione è necessaria per realizzare l'innovazione e trasformare le idee in risultati.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Incoraggia la sperimentazione: incoraggia la sperimentazione per supportare l'apprendimento e l'innovazione.
  - Sperimenta con diverse tecnologie: incoraggia la sperimentazione con tecnologie che potrebbero essere applicabili ora o in futuro per raggiungere i risultati desiderati. Le conoscenze derivanti possono essere utilizzate per l'innovazione futura.
  - Sperimenta in modo mirato: incoraggia la sperimentazione volta a obiettivi specifici che i membri del team possono raggiungere per o con tecnologie che potrebbero essere applicabili nel prossimo futuro. Le conoscenze derivanti possono essere utilizzate per l'innovazione.
  - Garantisci del tempo dedicato alla sperimentazione in modo strutturato: stabilisci orari specifici durante i quali i membri del team possono essere sgravati dalle normali responsabilità, in modo che possano concentrarsi sui loro esperimenti.
  - Rendi disponibili risorse per supportare la sperimentazione: finanzia le risorse necessarie per condurre esperimenti (ad esempio, software o risorse cloud).
  - Riconosci il successo: riconosci il valore generato dalla sperimentazione. Gli esperimenti con risultati indesiderati sono considerati riusciti e consentono di identificare un percorso che non porterà al successo. I membri del team non vengono puniti per risultati indesiderati degli esperimenti.

OPS03-BP06 Autorizzazione e incoraggiamento ai membri del team a mantenere e ampliare le proprie competenze

I team devono aumentare le proprie competenze per adottare nuove tecnologie e supportare i cambiamenti di domanda e responsabilità a supporto dei carichi di lavoro. L'ampliamento delle competenze nelle nuove tecnologie è spesso fonte di soddisfazione per i membri del team e supporta l'innovazione. Incoraggia i membri del team a perseguire e mantenere le certificazioni di settore in modo da convalidare e riconoscere le loro crescenti competenze. Pratica la formazione trasversale per promuovere il trasferimento di conoscenze e ridurre il rischio di impatto significativo in caso di perdita di membri del team qualificati ed esperti con competenze a livello istituzionale. Fornisci tempo strutturato dedicato per l'apprendimento.

AWS fornisce delle risorse, tra cui il Centro risorse per le nozioni di base di AWS, i Blog AWS, gli AWS OnlineTech Talks, Eventi e webinar AWSe gli AWS Well-Architected Labs, che forniscono indicazioni, esempi e procedure guidate dettagliate per formare i team.

AWS condivide inoltre le best practice e i modelli appresi attraverso la gestione di AWS nella Amazon Builders' Library e un'ampia gamma di altri utili materiali didattici tramite il Blog AWS e il podcast ufficiale di AWS.

Per formare i team, è consigliabile utilizzare le risorse fornite da AWS, ad esempio i corsi Well-Architected, AWS Support (Knowledge Center di AWS, forum di discussione AWSe AWS Support Center) e la Documentazione di AWS . Se hai domande riguardanti AWS, contatta AWS Support tramite AWS Support Center.

AWS Training and Certification offre risorse di formazione gratuite tramite corsi digitali gestiti dall'utente sulle nozioni di base di AWS. Per supportare ulteriormente lo sviluppo delle competenze AWS del tuo team, è anche possibile iscriversi a corsi di formazione con istruttore.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- I membri del team sono autorizzati e incoraggiati a mantenere e ampliare le proprie competenze: la formazione continua è indispensabile per adottare nuove tecnologie, favorire l'innovazione e supportare i cambiamenti della domanda e delle nuove responsabilità a supporto dei carichi di lavoro.
  - Metti a disposizione le risorse per la formazione: metti a disposizione del tempo in modo strutturato, accesso ai materiali di formazione, risorse di laboratorio e supporto alla partecipazione a conferenze e organizzazioni professionali che offrono opportunità di

apprendimento da docenti e colleghi. Fornisci ai membri del team di primo livello l'accesso ai membri del team senior affinché questi fungano da mentori o possano mostrare loro come lavorano trasmettendo metodi e competenze consolidati. Incoraggia l'apprendimento dei contenuti non direttamente correlati al lavoro per avere una prospettiva più ampia.

- Formazione del team e coinvolgimento tra team: pianifica le esigenze di formazione continua dei membri del tuo team. Offri ai membri del team l'opportunità di unirsi ad altri team (temporaneamente o definitivamente) per condividere competenze e best practice a beneficio dell'intera organizzazione.
- Supporta il perseguimento e il mantenimento delle certificazioni di settore: favorisci l'acquisizione e il mantenimento da parte dei membri del tuo team di certificazioni di settore che convalidano le loro conoscenze e riconoscono i loro risultati.

Risorse

Documenti correlati:

- Centro risorse per le nozioni di base di AWS
- i Blog AWS
- Conformità di Cloud AWS
- forum di discussione AWS
- Documentazione di AWS
- gli AWS OnlineTech Talks
- Eventi e webinar AWS
- Knowledge Center di AWS
- AWS Support
- AWS Training and Certification
- AWS Well-Architected Labs,
- Amazon Builders' Library
- il podcast ufficiale di AWS.

OPS03-BP07 Fornitura di risorse appropriate ai team

Mantieni la capacità dei membri del team e fornisci strumenti e risorse per supportare le esigenze del carico di lavoro. I membri del team con troppe mansioni aumentano il rischio di incidenti causati da

errori umani. Gli investimenti in strumenti e risorse (ad esempio, fornendo automazione per le attività eseguite di frequente) possono ricalibrare l'efficacia del team, consentendogli di supportare attività aggiuntive.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Risorse appropriate ai team: assicurati di comprendere i risultati dei tuoi team e i fattori che contribuiscono al loro successo o insuccesso. Agisci in modo da supportare i team con risorse appropriate.
  - Comprensione delle prestazioni del team: misura i risultati operativi raggiunti e lo sviluppo degli asset da parte dei tuoi team. Monitora le modifiche nell'output e nella percentuale di errori nel corso del tempo. Interagisci con i team per comprendere le sfide correlate al lavoro che li riguardano, come l'aumento delle responsabilità, i cambiamenti tecnologici, la perdita di personale o l'aumento dei clienti supportati.
  - Comprensione degli effetti sulle prestazioni del team: mantieni un coinvolgimento attivo con i tuoi team in modo da comprendere come stanno e se ci sono fattori esterni che li influenzano. Quando i team sono influenzati da fattori esterni, rivaluta gli obiettivi e modifica i target in base alle esigenze. Individua gli ostacoli che impediscono l'avanzamento dei team. Agisci per conto dei tuoi team per superare gli ostacoli e rimuovere gli oneri superflui.
  - Disponibilità delle risorse necessarie per il successo dei team: verifica regolarmente se le risorse sono ancora appropriate, o se sono necessarie risorse aggiuntive, e modifica di conseguenza i team di supporto.

OPS03-BP08 Incoraggiamento e ricerca di opinioni diverse all'interno e tra i team

Sfrutta la diversità tra organizzazioni per cercare più prospettive uniche. Usa questa prospettiva per incrementare l'innovazione, mettere in discussione le tue ipotesi e ridurre il rischio di conferme parziali. Aumenta l'inclusione, la diversità e l'accessibilità all'interno dei team per ottenere prospettive vantaggiose.

La cultura organizzativa influisce direttamente sulla soddisfazione sul lavoro e sulla conservazione dei membri del team. Sostieni il coinvolgimento e le capacità dei membri del tuo team per ottenere il successo della tua attività.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Cerca opinioni e prospettive diverse: incoraggia la condivisione dei contributi da parte di tutti. Dai voce ai gruppi sottorappresentati. Distribuisci a rotazione i ruoli e responsabilità nelle riunioni.
  - Amplia ruoli e responsabilità: offri ai membri del team l'opportunità di assumere ruoli che altrimenti potrebbero altrimenti non ricoprire mai. Ciò consentirà loro di acquisire esperienza e nuove prospettive grazie anche alle interazioni con i nuovi membri del team, con i quali potrebbero non interagire altrimenti. Un mutuo scambio di esperienze e punti di vista vantaggioso per tutti Con l'aumento della prospettiva, possono emergere ulteriori opportunità di business o possono essere identificate nuove opportunità di miglioramento. Fai in modo che i membri di un team svolgano a turno attività comuni eseguite normalmente da altri affinché comprendano le richieste e l'impatto delle loro prestazioni.
  - Garantisci un ambiente sicuro e ospitale: adotta policy e controlli che consentano di proteggere la sicurezza fisica e mentale dei membri del team all'interno dell'organizzazione. I membri del team devono essere in grado di interagire senza alcun timore. Quando i membri del team si sentono al sicuro e ben accolti, è più probabile che siano coinvolti e produttivi. Più è diversificata la tua organizzazione, migliore sarà la comprensione nei confronti delle persone supportate, inclusi i clienti. Quando i membri del team si sentono a loro agio, sono liberi di parlare e sono sicuri che verranno ascoltati, con maggiori probabilità condivideranno informazioni preziose (ad esempio, opportunità di marketing, esigenze di accessibilità, segmenti di mercato non serviti, rischi non riconosciuti nel tuo ambiente).
  - Consenti la totale partecipazione dei membri del team: fornisci le risorse necessarie ai dipendenti affinché partecipino appieno a tutte le attività correlate al lavoro. I membri del team che affrontano sfide quotidiane hanno sviluppato competenze per superarle. Queste competenze esclusive possono offrire vantaggi significativi alla tua organizzazione. Grazie al supporto di strutture adeguate, i membri del team possono portare in azienda contributi vantaggiosi.

Preparazione

Domande

- OPS 4 In che modo progetti il carico di lavoro al fine di comprenderne lo stato?
- OPS 5 In che modo riduci i difetti, favorisci la correzione e migliori il flusso nella produzione?
- OPS 6 In che modo mitighi i rischi della distribuzione?
- OPS 7 Come fai a sapere che sei pronto a supportare un carico di lavoro?

OPS 4 In che modo progetti il carico di lavoro al fine di comprenderne lo stato?

Progetta il tuo carico di lavoro in modo da ottenere le informazioni necessarie tra i componenti (ad esempio, parametri, log e tracce) per comprenderne lo stato interno. Ciò ti consente di fornire risposte efficaci in base alle esigenze.

Best practice

- OPS04-BP01 Implementazione della telemetria dell'applicazione
- OPS04-BP02 Implementazione e configurazione della telemetria del carico di lavoro
- OPS04-BP03 Implementazione della telemetria dell'attività degli utenti
- OPS04-BP04 Implementazione della telemetria delle dipendenze
- OPS04-BP05 Implementazione della tracciabilità delle transazioni

OPS04-BP01 Implementazione della telemetria dell'applicazione

La telemetria dell'applicazione è la base su cui si fonda l'osservabilità del carico di lavoro. L'applicazione deve trasmettere una telemetria in grado di fornire approfondimenti sullo stato dell'applicazione stessa e sul raggiungimento dei risultati aziendali. Dalla risoluzione dei problemi alla misurazione dell'impatto di una nuova funzionalità, la telemetria dell'applicazione garantisce informazioni su cui basare la creazione, il funzionamento e l'evoluzione del carico di lavoro.

La telemetria dell'applicazione è costituita da parametri e registri. I parametri sono informazioni diagnostiche, ad esempio il polso o la temperatura, e vengono impiegati in maniera collettiva per descrivere lo stato dell'applicazione. La loro raccolta nel tempo può servire per sviluppare indicatori di riferimento e rilevare anomalie. I registri sono messaggi inviati dall'applicazione in merito al suo stato interno o a eventi che si verificano. Sono esempi di eventi registrati i codici di errore, gli identificatori di transazione e le azioni dell'utente.

Risultato desiderato:

- L'applicazione trasmette parametri e registri che forniscono informazioni in merito alla sua integrità e al raggiungimento dei risultati aziendali.
- I parametri e i registri sono archiviati centralmente per tutte le applicazioni del carico di lavoro.

Anti-pattern comuni:

- L'applicazione non emette telemetria. Sei costretto a fare affidamento sui tuoi clienti per informarti quando c'è qualcosa che non va.
- Un cliente ha segnalato che la tua applicazione non risponde. Non disponi di una telemetria e non sei in grado di confermare l'effettiva esistenza del problema o definirlo senza utilizzare l'applicazione per comprendere l'attuale esperienza utente.

Vantaggi dell'adozione di questa best practice:

- Possibilità di conoscere lo stato di integrità dell'applicazione, l'esperienza utente e di sapere se i risultati aziendali sono stati raggiunti.
- Possibilità di reagire rapidamente alle modifiche dello stato di integrità dell'applicazione.
- Possibilità di sviluppare tendenze in merito allo stato di integrità dell'applicazione.
- Possibilità di prendere decisioni più informate sul miglioramento dell'applicazione.
- Possibilità di rilevare e risolvere più rapidamente eventuali problemi con l'applicazione.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

Per implementare l'applicazione della telemetria sono necessari tre passaggi: identificare una posizione in cui archiviare la telemetria, identificare una telemetria che descriva lo stato dell'applicazione e dotare l'applicazione degli strumenti per trasmettere la telemetria.

Prendiamo, a titolo di esempio, un'azienda di e-commerce con un'architettura basata su microservizi. Nell'ambito dell'iter progettuale dell'architettura, l'azienda identifica una telemetria dell'applicazione utile per capire lo stato di ciascun microservizio. Ad esempio, il servizio legato al carrello dell'utente trasmette una telemetria relativa a eventi come l'aggiunta al carrello, l'abbandono del carrello e il tempo che è servito per aggiungere un articolo al carrello. Per tutti i microservizi è prevista la registrazione di errori, avvisi e informazioni sulle transazioni. La telemetria viene inviata a Amazon CloudWatch per l'archiviazione e l'analisi.

Passaggi dell'implementazione

Il primo passaggio consiste nell'identificare una posizione centrale in cui archiviare la telemetria per le applicazioni del carico di lavoro. In assenza di una piattaforma esistente, Amazon CloudWatch può garantire la raccolta della telemetria, dashboard, analisi e funzionalità di generazione di eventi.

Per stabilire il tipo di telemetria necessaria, è utile iniziare ponendosi queste domande:

- La mia applicazione è integra?
- La mia applicazione sta raggiungendo i risultati aziendali attesi?

L'applicazione deve emettere parametri e registri in grado di rispondere in modo collettivo a queste domande. Se non è possibile rispondere con la telemetria dell'applicazione esistente, si potrà lavorare con le figure che si occupano di business e di progettazione per creare un elenco di soluzioni di telemetria in grado di farlo. Nel corso dell'identificazione e dello sviluppo di una nuova telemetria per l'applicazione, è possibile chiedere una consulenza tecnica agli esperti del team Account AWS.

Una volta identificata la soluzione di telemetria da aggiungere, si potrà lavorare con le figure che si occupano della progettazione per dotare l'applicazione degli strumenti necessari. La soluzione AWS Distro for Open Telemetry fornisce API, librerie e agenti per la raccolta della telemetria dell'applicazione. Questo esempio mostra come dotare un'applicazione JavaScript di parametri personalizzati.

Se desiderano comprendere i servizi di osservabilità offerti da AWS, i clienti possono seguire il workshop dal titolo One Observability Workshop in autonomia o richiedere supporto e indicazioni al team Account AWS. Il workshop illustra le soluzioni per l'osservabilità offerte da AWS, con esempi pratici del loro utilizzo.

Per maggiori approfondimenti riguardo alla telemetria dell'applicazione, leggi l'articolo Strumentazione di sistemi distribuiti per visibilità operativa nella Amazon Builder's Library. L'articolo spiega in che modo Amazon dota le applicazioni degli strumenti necessari e può servire come riferimento per sviluppare proprie linee guida sulla strumentazione.

Livello di impegno per il piano di implementazione: Medio

Risorse

Best practice correlate:

the section called “OPS04-BP02 Implementazione e configurazione della telemetria del carico di lavoro” - La telemetria dell'applicazione è un componente della telemetria del carico di lavoro. Per conoscere l'integrità del carico di lavoro in generale è necessario conoscere l'integrità delle singole applicazioni che lo compongono.

the section called “OPS04-BP03 Implementazione della telemetria dell'attività degli utenti” - La telemetria dell'attività utente è spesso un sottoinsieme della telemetria dell'applicazione. L'attività

utente, ad esempio eventi di aggiunta al carrello, flussi di clic o transazioni completate, fornisce informazioni dettagliate sull'esperienza utente.

the section called “OPS04-BP04 Implementazione della telemetria delle dipendenze” - I controlli di dipendenza sono correlati alla telemetria dell'applicazione e possono essere inclusi nella propria applicazione. Se quest'ultima si appoggia a dipendenze esterne, ad esempio DNS o un database, può trasmettere parametri e registri riguardanti la reperibilità, i timeout e altri eventi.

the section called “OPS04-BP05 Implementazione della tracciabilità delle transazioni” - Il tracciamento delle transazioni in un carico di lavoro richiede che ogni applicazione trasmetta informazioni sulle modalità con cui elabora gli eventi condivisi. Il modo in cui le singole applicazioni gestiscono questi eventi viene trasmesso tramite la relativa telemetria dell'applicazione.

the section called “OPS08-BP02 Definizione dei parametri del carico di lavoro” - I parametri del carico di lavoro sono gli indicatori chiave del suo stato di integrità. I parametri chiave dell'applicazione fanno parte dei parametri del carico di lavoro.

Documenti correlati:

- AWS Builders Library: Strumentazione di sistemi distribuiti per visibilità operativa
- AWS Distro for OpenTelemetry
- Whitepaper sull'eccellenza operativa secondo il Canone di architettura AWS - Progettazione della telemetria
- Creazione di parametri da registro eventi mediante filtri
- Progettazione e implementazione di registrazione e monitoraggio con Amazon CloudWatch
- Monitoring application health and performance with AWS Distro for OpenTelemetry (Monitoraggio dell'integrità e delle prestazioni dell'applicazione con AWS Distro for OpenTelemetry)
- Novità - How to better monitor your custom application metrics using Amazon CloudWatch Agent (Come monitorare meglio i parametri personalizzati di un'applicazione con l'agente Amazon CloudWatch)
- Osservabilità in AWS
- Scenario: pubblicazione di parametri su CloudWatch
- Inizia a creare - Come monitorare le applicazioni in modo efficace
- Utilizzo di CloudWatch con AWS SDK

Video correlati:

- AWS re:Invent 2021 - Observability the open-source way (AWS re:Invent 2021 - L'osservabilità con l'open-source)
- Collect Metrics and Logs from Amazon EC2 instances with the CloudWatch Agent (Raccolta di parametri e registri da istanze EC2 con l'agente CloudWatch)
- How to Easily Setup Application Monitoring for Your AWS Workloads - AWS Online Tech Talks (Come configurare facilmente il monitoraggio dell'applicazione per i carichi di lavoro AWS - AWS Online Tech Talks)
- Mastering Observability of Your Serverless Applications - AWS Online Tech Talks (Controllare l'osservabilità delle applicazioni serverless - AWS Online Tech Talks)
- Open Source Observability with AWS - AWS Virtual Workshop (Osservabilità open-source con AWS - Workshop virtuale AWS)

Esempi correlati:

- Risorse di esempio per registrazione e monitoraggio con AWS
- AWS Solution: Amazon CloudWatch Monitoring Framework (Soluzione AWS: framework di monitoraggio di Amazon CloudWatch)
- AWS Solution: Centralized Logging (Soluzione AWS: registrazione centralizzata)
- One Observability Workshop

OPS04-BP02 Implementazione e configurazione della telemetria del carico di lavoro

Progetta e configura il carico di lavoro affinché fornisca informazioni sul suo stato interno e sullo stato corrente, ad esempio volume delle chiamate API, codici di stato HTTP ed eventi di scalabilità. Utilizza queste informazioni per determinare quando è necessaria una risposta.

Puoi avvalerti di un servizio come Amazon CloudWatch per aggregare log e parametri da componenti del carico di lavoro (ad esempio, log API da AWS CloudTrail, parametri AWS Lambda, Registri di flusso Amazon VPCe altri servizi).

Anti-pattern comuni:

- I tuoi clienti lamentano prestazioni scarse. Non sono presenti modifiche recenti all'applicazione, pertanto sospetti un problema con un componente del carico di lavoro. Non disponi della telemetria per analizzare e determinare quali componenti contribuiscono a rendere scarse le prestazioni.

- L'applicazione non è raggiungibile. In mancanza di telemetria, non puoi determinare se si tratta di un problema di rete.

Vantaggi dell'adozione di questa best practice: Comprendere cosa succede all'interno del carico di lavoro ti consente di rispondere, se necessario.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Implementazione della telemetria per registri e parametri: dota il carico di lavoro degli strumenti per fornire informazioni sullo stato interno, sullo stato e sul raggiungimento dei risultati aziendali. Utilizza queste informazioni per stabilire quando è necessaria una risposta.
  - Migliorare la capacità di osservazione delle VM con Amazon CloudWatch - AWS Online Tech Talks
  - Come funziona Amazon CloudWatch
  - Che cos'è Amazon CloudWatch?
  - Utilizzare i parametri Amazon CloudWatch
  - Che cos'è Amazon CloudWatch Logs?
    - Implementazione della telemetria per registri e parametri: progetta e configura il carico di lavoro affinché fornisca informazioni sul suo stato interno e sullo stato corrente, ad esempio volume delle chiamate API, codici di stato HTTP ed eventi di scalabilità.
      - Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch
      - AWS CloudTrail
      - Che cos'è AWS CloudTrail?
      - Log di flusso VPC

Risorse

Documenti correlati:

- AWS CloudTrail
- Documentazione su Amazon CloudWatch
- Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch
- Come funziona Amazon CloudWatch

- Utilizzare i parametri Amazon CloudWatch
- Log di flusso VPC
- Che cos'è AWS CloudTrail?
- Che cos'è Amazon CloudWatch Logs?
- Che cos'è Amazon CloudWatch?

Video correlati:

- Application Performance Management on AWS (Gestione delle prestazioni delle applicazioni su AWS)
- Migliorare la capacità di osservazione delle VM con Amazon CloudWatch
- Migliorare la capacità di osservazione delle VM con Amazon CloudWatch - AWS Online Tech Talks

OPS04-BP03 Implementazione della telemetria dell'attività degli utenti

Implementa il codice dell'applicazione affinché fornisca informazioni sulle attività degli utenti, ad esempio, flussi di clic o transazioni avviate, abbandonate e completate. Utilizza queste informazioni per comprendere come viene utilizzata l'applicazione, i modelli di utilizzo e per stabilire quando è necessaria una risposta.

Anti-pattern comuni:

- Gli sviluppatori hanno distribuito una nuova funzionalità senza telemetria degli utenti e l'utilizzo è aumentato. Non puoi determinare se l'aumento di utilizzo dall'uso della nuova funzionalità o se è un problema introdotto con il nuovo codice.
- Gli sviluppatori hanno distribuito una nuova funzionalità senza telemetria degli utenti. Non è possibile stabilire se i clienti la utilizzano senza contattarli e interrogarli.

Vantaggi dell'adozione di questa best practice: Scopri come i clienti utilizzano la tua applicazione per identificare modelli di utilizzo, comportamenti imprevisti e per consentirti di rispondere, se necessario.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Implementazione della telemetria delle attività degli utenti: progetta il codice dell'applicazione affinché fornisca informazioni sulle attività degli utenti (ad esempio flussi di clic o transazioni

avviate, abbandonate e completate). Utilizza queste informazioni per comprendere come viene utilizzata l'applicazione, i modelli di utilizzo e per stabilire quando è necessaria una risposta.

OPS04-BP04 Implementazione della telemetria delle dipendenze

Progetta e configura il carico di lavoro affinché fornisca informazioni sullo stato (ad esempio, raggiungibilità o tempo di risposta) delle risorse da cui dipende. Esempi di dipendenze esterne possono includere database esterni, DNS e connettività di rete. Utilizza queste informazioni per stabilire quando è necessaria una risposta.

Anti-pattern comuni:

- Non è possibile determinare se il motivo per cui l'applicazione è irraggiungibile è un problema DNS senza eseguire manualmente un controllo per verificare se il provider DNS funziona.
- L'applicazione correlata al carrello non è in grado di completare le transazioni. Non sei in grado di determinare se si tratta di un problema con il provider di elaborazione della carta di credito senza contattarlo per verificarlo.

Vantaggi dell'adozione di questa best practice: Comprendere lo stato delle dipendenze ti consente di rispondere, se necessario.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Implementazione della telemetria delle dipendenze: progetta e configura il carico di lavoro affinché fornisca informazioni sullo stato dei sistemi da cui dipende. Alcuni esempi includono: database esterni, DNS, connettività di rete e servizi esterni di elaborazione delle carte di credito.
  - Agente di Amazon CloudWatch con integrazione AWS Systems Manager - raccolta parametri e registri unificata per Linux e Windows
  - Raccolta di parametri e registri da istanze Amazon EC2 e da server on-premise con l'agente di CloudWatch

Risorse

Documenti correlati:

- Agente di Amazon CloudWatch con integrazione AWS Systems Manager - raccolta parametri e registri unificata per Linux e Windows
- Raccolta di parametri e registri da istanze Amazon EC2 e da server on-premise con l'agente di CloudWatch

Esempi correlati:

- Well-Architected Labs - Dependency Monitoring (Monitoraggio delle dipendenze)

OPS04-BP05 Implementazione della tracciabilità delle transazioni

Implementa il codice dell'applicazione e configura i componenti del carico di lavoro affinché forniscano informazioni sul flusso delle transazioni nel carico di lavoro. Utilizza queste informazioni per stabilire quando è necessaria una risposta e per favorire l'identificazione dei fattori che contribuiscono all'origine di un problema.

In AWS, puoi utilizzare i servizi di tracciamento distribuiti, come AWS X-Ray, per raccogliere e registrare le tracce mentre le transazioni attraversano il tuo carico di lavoro, generare mappe per vedere come le transazioni fluiscono tra il tuo carico di lavoro e i servizi, ottenere informazioni sulle relazioni tra i componenti e identificare e analizzare i problemi in tempo reale.

Anti-pattern comuni:

- Hai implementato un'architettura di microservizi serverless che si estende su più account. I clienti riscontrano problemi di prestazioni intermittenti. Non sei in grado di scoprire quale funzione o componente è responsabile perché mancano le tracce che consentono di individuare il punto in cui si è verificato il problema all'interno dell'applicazione e la relativa causa.
- Stai cercando di determinare dove si trovano i colli di bottiglia delle prestazioni nel carico di lavoro in modo da sviluppare possibili soluzioni. Non sei in grado di vedere la relazione tra i componenti dell'applicazione e i servizi con cui interagiscono, per determinare dove si trovano i colli di bottiglia perché mancano le tracce che ti consentirebbero di approfondire i servizi e i percorsi specifici che influiscono sulle prestazioni dell'applicazione.

Vantaggi dell'adozione di questa best practice: Comprendere il flusso delle transazioni nel il carico di lavoro consente di comprendere il loro comportamento previsto e le variazioni correlate, consentendo di rispondere, se necessario.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Implementazione della tracciabilità delle transazioni: progetta l'applicazione e il carico di lavoro affinché forniscano informazioni sul flusso delle transazioni nei componenti del sistema, come la fase della transazione, il componente attivo e il tempo per completare l'attività. Utilizza queste informazioni per determinare cosa è in corso, cosa è completato e i risultati delle attività completate. Questo è utile per stabilire quando è necessaria una risposta. Ad esempio, i tempi di risposta più lunghi del previsto per le transazioni all'interno di un componente possono essere indicativi di problemi a carico di quel componente.
  - AWS X-Ray
  - Che cos'è AWS X-Ray?

Risorse

Documenti correlati:

- AWS X-Ray
- Che cos'è AWS X-Ray?

OPS 5 In che modo riduci i difetti, favorisci la correzione e migliori il flusso nella produzione?

Adotta prassi che migliorino il flusso delle modifiche nella produzione, che consentano il refactoring e il feedback veloce su qualità e correzione di errori. Tali prassi accelerano l'ingresso in produzione delle modifiche vantaggiose, limitano i problemi distribuiti e consentono una rapida identificazione e risoluzione dei problemi introdotti attraverso le attività di distribuzione.

Best practice

- OPS05-BP01 Utilizzo del controllo delle versioni
- OPS05-BP02 Test e convalida delle modifiche
- OPS05-BP03 Utilizzo di sistemi di gestione delle configurazioni
- OPS05-BP04 Utilizzo di sistemi di gestione della compilazione e implementazione
- OPS05-BP05 Esecuzione della gestione delle patch
- OPS05-BP06 Condivisione degli standard di progettazione
- OPS05-BP07 Implementazione di prassi per migliorare la qualità del codice

- OPS05-BP08 Utilizzo di più ambienti
- OPS05-BP09 Applicazione di modifiche frequenti, minime e reversibili
- OPS05-BP10 Automazione completa dell'integrazione e della distribuzione

OPS05-BP01 Utilizzo del controllo delle versioni

Utilizza il controllo delle versioni per abilitare il monitoraggio di modifiche e rilasci.

Molti servizi AWS offrono funzionalità di controllo delle versioni. Utilizza una revisione o un sistema di controllo del codice sorgente come AWS CodeCommit per gestire il codice e altri artefatti, come i modelli AWS CloudFormation controllati dalla versione della tua infrastruttura.

Anti-pattern comuni:

- Hai sviluppato e archiviato il codice sulla workstation. Si è verificato un errore di storage non recuperabile sulla workstation in cui il codice è andato perso.
- Dopo aver sovrascritto il codice esistente con le modifiche, riavvii l'applicazione e non è più utilizzabile. Non è possibile ripristinare la modifica.
- Hai un blocco di scrittura su un file di report che deve essere modificato da altri utenti. Ti contattano per chiederti di smettere di utilizzarlo in modo che possano completare le loro attività.
- Il team di ricerca ha lavorato a un'analisi dettagliata che definirà il tuo lavoro futuro. Qualcuno ha salvato accidentalmente la lista della spesa nel report finale. Non puoi ripristinare la modifica e dovrai ricreare il report.

Vantaggi dell'adozione di questa best practice: Grazie alle funzionalità di controllo delle versioni, puoi ripristinare facilmente gli stati validi noti, le versioni precedenti e limitare il rischio di perdita degli asset.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Utilizzo del controllo delle versioni: mantieni gli asset in repository con controllo delle versioni. In questo modo si supporta il monitoraggio delle modifiche, la distribuzione di nuove versioni, il rilevamento delle modifiche apportate alle versioni esistenti e il ripristino delle versioni precedenti, ad esempio il rollback a uno stato corretto noto in caso di errore. Integra nelle tue procedure le funzionalità di controllo delle versioni dei sistemi di gestione delle configurazioni.

- Introduzione ad AWS CodeCommit
- What is AWS CodeCommit? (Che cos'è AWS CodeCommit?)

Risorse

Documenti correlati:

- What is AWS CodeCommit? (Che cos'è AWS CodeCommit?)

Video correlati:

- Introduzione ad AWS CodeCommit

OPS05-BP02 Test e convalida delle modifiche

Testa e convalida le modifiche per limitare e rilevare gli errori. Automatizza il testing per ridurre gli errori causati dai processi manuali e il livello di impegno richiesto per il test.

Molti servizi AWS offrono funzionalità di controllo delle versioni. Utilizza una revisione o un sistema di controllo del codice sorgente come AWS CodeCommit per gestire il codice e altri artefatti, come i modelli AWS CloudFormation controllati dalla versione della tua infrastruttura.

Anti-pattern comuni:

- Distribuisci il tuo nuovo codice alla produzione e i clienti iniziano a chiamare perché la tua applicazione non funziona più.
- Applichi nuovi gruppi di sicurezza per migliorare la sicurezza perimetrale. Questo metodo funziona con conseguenze impreviste e gli utenti non sono in grado di accedere alle applicazioni.
- Modifichi un metodo richiamato dalla nuova funzione. Anche un'altra funzione dipendeva da tale metodo e non funziona più. Il problema non viene rilevato ed entra in produzione. L'altra funzione non viene richiamata per un certo periodo di tempo e, infine, viene meno in produzione senza alcuna correlazione con la causa.

Vantaggi dell'adozione di questa best practice: Testando e convalidando le modifiche in anticipo, puoi risolvere i problemi con costi ridotti al minimo e limitare l'impatto sui clienti. Eseguendo il test prima della distribuzione, riduci al minimo la possibilità di errore.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Test e convalida delle modifiche: è necessario testare le modifiche e convalidare i risultati in tutte le fasi del ciclo di vita, ad esempio sviluppo, test e produzione. Utilizza i risultati dei test per confermare le nuove funzionalità e ridurre il rischio e l'impatto delle distribuzioni non riuscite. Automatizza i test e la convalida per assicurare la coerenza della revisione, ridurre gli errori causati dai processi manuali e ridurre il livello di impegno richiesto.
  - What is AWS CodeBuild? (Che cos'è AWS CodeBuild?)
  - Supporto della creazione locale per AWS CodeBuild

Risorse

Documenti correlati:

- Strumenti per gli sviluppatori di AWS
- Supporto della creazione locale per AWS CodeBuild
- What is AWS CodeBuild? (Che cos'è AWS CodeBuild?)

OPS05-BP03 Utilizzo di sistemi di gestione delle configurazioni

L'utilizzo di sistemi di gestione delle configurazioni permette di effettuare modifiche alle stesse e tenerne traccia. Questi sistemi riducono gli errori causati dai processi manuali e il livello di impegno richiesto per la distribuzione delle modifiche.

Durante l'inizializzazione di una risorsa, la gestione delle configurazioni statiche consente di impostare valori che dovrebbero rimanere coerenti per tutta la vita utile della risorsa. Ne sono alcuni esempi l'azione di configurare un server web o applicativo su un'istanza oppure di definire la configurazione di un servizio AWS nella AWS Management Console o tramite la AWS CLI.

Al momento dell'inizializzazione, la gestione delle configurazioni dinamiche consente di impostare valori che possono cambiare nel corso della vita utile di una risorsa. Ad esempio è possibile impostare un interruttore funzionale in grado di abilitare una funzionalità nel codice tramite una modifica della configurazione, oppure modificare il livello di dettaglio del registro durante un incidente per acquisire un maggior numero di dati e cambiarlo in seguito per tornare al livello di dettaglio precedente, risparmiando così in numero di registri e nei relativi costi.

Se sulle applicazioni in esecuzione su istanze, container, funzioni serverless o dispositivi sono attive configurazioni dinamiche, è possibile utilizzare AWS AppConfig per gestirle e distribuirle fra i diversi ambienti.

In AWS, è possibile utilizzare AWS Config per monitorare in modo continuo le configurazioni delle risorse AWS tra i diversi account e regioni. Questa soluzione consente di tenere traccia della cronologia delle configurazioni, di capire che effetto avrebbe la modifica di una configurazione sulle altre risorse e verificarle rispetto alle configurazioni previste o desiderate tramite Regole di AWS Config e i pacchetti di conformità di AWS Config.

In AWS, puoi creare pipeline di integrazione continua/distribuzione continua (CI/CD) utilizzando servizi come gli Strumenti per gli sviluppatori di AWS (ad esempio, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeploye AWS CodeStar).

Mantieni un calendario delle modifiche e verifica quando sono pianificate attività o eventi aziendali significativi che potrebbero essere influenzati dall'implementazione del cambiamento. Regola le attività per gestire i rischi in funzione dei vari eventi. Il calendario delle modifiche di AWS Systems Manager fornisce un meccanismo per documentare intervalli di tempo aperti o chiusi alle modifiche e per quale motivo, e per condividere tali informazioni con altri Account AWS. Gli script di AWS Systems Manager Automation possono essere configurati per rispettare lo stato del calendario delle modifiche.

finestre di manutenzione di AWS Systems Manager possono essere utilizzate per pianificare le prestazioni degli script Run Command o Automation di AWS, delle chiamate AWS Lambda o delle attività di AWS Step Functions in orari specifici. Contrassegna queste attività nel calendario delle modifiche in modo che possano essere incluse nella tua valutazione.

Anti-pattern comuni:

- Aggiorni manualmente la configurazione del server Web all'interno del parco istanze e un certo numero di server non risponde a causa di errori di aggiornamento.
- Aggiorni manualmente il parco istanze del server applicazioni nel corso di molte ore. L'incoerenza nella configurazione durante la modifica causa comportamenti imprevisti.
- Qualcuno ha aggiornato i tuoi gruppi di sicurezza e i server Web non sono più accessibili. Senza sapere cosa è stato modificato, dedichi molto tempo a esaminare il problema prolungando il tempo necessario per il ripristino.

Vantaggi dell'adozione di questa best practice: L'adozione di sistemi di gestione della configurazione riduce il livello di impegno necessario per apportare e tenere traccia delle modifiche e la frequenza degli errori causati dalle procedure manuali.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Utilizzo dei sistemi di gestione delle configurazioni: utilizza i sistemi di gestione delle configurazioni per tenere traccia e implementare le modifiche, ridurre gli errori causati dai processi manuali e ridurre il livello di impegno richiesto.
  - Gestione della configurazione delle infrastrutture
  - AWS Config
  - What is AWS Config? (Che cos'è AWS Config?)
  - Introduzione ad AWS CloudFormation
  - Che cos'è AWS CloudFormation?
  - AWS OpsWorks
  - What is AWS OpsWorks? (Che cos'è AWS OpsWorks?)
  - Introduzione ad AWS Elastic Beanstalk
  - Che cos'è AWS Elastic Beanstalk?

Risorse

Documenti correlati:

- AWS AppConfig
- Strumenti per gli sviluppatori di AWS
- AWS OpsWorks
- Il calendario delle modifiche di AWS Systems Manager
- finestre di manutenzione di AWS Systems Manager
- Gestione della configurazione delle infrastrutture
- Che cos'è AWS CloudFormation?
- What is AWS Config? (Che cos'è AWS Config?)
- Che cos'è AWS Elastic Beanstalk?

- What is AWS OpsWorks? (Che cos'è AWS OpsWorks?)

Video correlati:

- Introduzione ad AWS CloudFormation
- Introduzione ad AWS Elastic Beanstalk

OPS05-BP04 Utilizzo di sistemi di gestione della compilazione e implementazione

Utilizza sistemi di gestione della creazione e distribuzione Questi sistemi riducono gli errori causati dai processi manuali e il livello di impegno richiesto per la distribuzione delle modifiche.

In AWS, puoi compilare pipeline di integrazione continua/implementazione continua (CI/CD) utilizzando servizi come gli Strumenti per sviluppatori in AWS (ad esempio, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeploye AWS CodeStar).

Anti-pattern comuni:

- Dopo aver compilato il codice nel sistema di sviluppo, copi il file eseguibile nei sistemi di produzione e questo non si avvia. I file di log locali indicano che l'operazione è risultata impossibile a causa della mancanza di dipendenze.
- Hai creato l'applicazione con nuove funzionalità nel tuo ambiente di sviluppo e fornisci il codice al controllo qualità. Il controllo qualità non riesce perché mancano asset statici.
- Venerdì, dopo un notevole sforzo, hai creato l'applicazione manualmente nel tuo ambiente di sviluppo, incluse le nuove funzionalità codificate. Lunedì non sei in grado di ripetere le fasi che ti hanno consentito di creare correttamente la tua applicazione.
- Esegui i test creati per la nuova versione. Quindi passi la settimana successiva a configurare un ambiente di test ed eseguire tutti i test di integrazione esistenti seguiti dai test delle prestazioni. Il nuovo codice ha un impatto inaccettabile sulle prestazioni e deve essere risviluppato e quindi ritestato.

Vantaggi dell'adozione di questa best practice: Fornendo meccanismi per gestire le attività di compilazione e distribuzione, riduci il livello di impegno necessario per eseguire attività ripetitive, consenti ai membri del team di concentrarsi liberamente sulle loro attività creative di valore elevato e limiti l'introduzione di errori derivanti da procedure manuali.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Utilizzo di sistemi di gestione della compilazione e implementazione: utilizza sistemi di gestione della compilazione e implementazione per tenere traccia e implementare le modifiche, ridurre gli errori causati dai processi manuali e ridurre il livello di impegno richiesto. Automatizza completamente la pipeline di integrazione e distribuzione dal check-in del codice fino alle fasi di creazione, test, distribuzione e convalida. In questo modo è possibile ridurre il lead time, aumentare la frequenza delle modifiche e ridurre il livello di impegno richiesto.
  - Che cos'è AWS CodeBuild?
  - Best practice di integrazione continua per lo sviluppo del software
  - Slalom: CI/CD per applicazioni serverless su AWS
  - Introduzione ad AWS CodeDeploy - Sviluppo di software automatizzato con Amazon Web Services
  - Che cos'è AWS CodeDeploy?

Risorse

Documenti correlati:

- Strumenti per sviluppatori in AWS
- Che cos'è AWS CodeBuild?
- Che cos'è AWS CodeDeploy?

Video correlati:

- Best practice di integrazione continua per lo sviluppo del software
- Introduzione ad AWS CodeDeploy - Sviluppo di software automatizzato con Amazon Web Services
- Slalom: CI/CD per applicazioni serverless su AWS

OPS05-BP05 Esecuzione della gestione delle patch

La gestione delle patch consente di ottenere funzionalità, risolvere problemi e rispettare i requisiti di governance. Automatizza la gestione delle patch per ridurre gli errori causati dai processi manuali e il livello di impegno richiesto per applicare le patch.

La gestione delle patch e delle vulnerabilità fa parte delle attività di gestione dei rischi e dei vantaggi. È preferibile disporre di infrastrutture immutabili e distribuire carichi di lavoro in stati noti verificati. Se ciò non è realizzabile, l'applicazione di patch sul posto è l'alternativa.

Aggiornare immagini di macchine, immagini di container o, nel caso di Lambda, runtime personalizzati e librerie aggiuntive per rimuovere le vulnerabilità fa parte della gestione delle patch. È consigliabile gestire gli aggiornamenti alle Amazon Machine Image (AMI) per immagini Linux o Windows Server utilizzando EC2 Image Builder. Puoi utilizzare Amazon Elastic Container Registry con la pipeline esistente per gestire le immagini di Amazon ECS e gestire le immagini di Amazon EKS. AWS Lambda include caratteristiche di gestione della versione.

L'applicazione di patch non deve essere eseguita sui sistemi di produzione senza prima eseguire test in un ambiente sicuro. Le patch devono essere applicate solo se supportano risultati operativi o aziendali. In AWS, è possibile utilizzare AWS Systems Manager Patch Manager per automatizzare il processo di applicazione di patch ai sistemi gestiti e pianificare l'attività utilizzando le finestre di manutenzione di AWS Systems Manager.

Anti-pattern comuni:

- Ti viene assegnato il compito di applicare tutte le nuove patch di sicurezza entro 2 ore, il che provoca più interruzioni a causa dell'incompatibilità dell'applicazione con le patch.
- Una libreria senza patch comporta conseguenze indesiderate in quanto parti sconosciute utilizzano vulnerabilità al suo interno per accedere al carico di lavoro.
- L'applicazione di patch agli ambienti per sviluppatori viene eseguita automaticamente senza avvisare gli sviluppatori. Gli sviluppatori ti inviano più reclami perché il loro ambiente non funziona come previsto.
- Non hai applicato patch al software pronto all'uso commerciale su un'istanza persistente. Quando hai problemi con il software e contatti il fornitore, questo ti informerà che la versione non è supportata e dovrai applicare le patch a un livello specifico per ricevere assistenza.
- Una patch rilasciata di recente per il software di crittografia utilizzato offre miglioramenti significativi in termini di prestazioni. Il sistema privo di patch presenta problemi di prestazioni che rimangono in vigore a causa della mancata applicazione di patch.

Vantaggi dell'adozione di questa best practice: Stabilendo un processo di gestione delle patch, inclusi i criteri per l'applicazione di patch e la metodologia per la distribuzione tra gli ambienti, sarai in grado di realizzarne i vantaggi e controllarne l'impatto. In questo modo sarà possibile adottare le caratteristiche e le funzionalità desiderate, eliminare i problemi e mantenere la conformità alla

governance. Implementa sistemi di gestione delle patch e automazione per ridurre il livello di impegno per distribuire le patch e limitare gli errori causati dai processi manuali.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Gestione delle patch: applica patch ai sistemi per correggere gli errori, ottenere le funzionalità o le capacità desiderate e assicurare la conformità alle policy di governance e ai requisiti di supporto del fornitore. Nei sistemi immutabili, distribuisci con il set di patch appropriato per raggiungere il risultato desiderato. Automatizza il meccanismo di gestione delle patch per ridurre il tempo necessario per applicare le patch, ridurre gli errori causati dai processi manuali e il livello di impegno richiesto per applicare le patch.
  - AWS Systems Manager Patch Manager

Risorse

Documenti correlati:

- Strumenti per gli sviluppatori di AWS
- AWS Systems Manager Patch Manager

Video correlati:

- CI/CD per applicazioni serverless su AWS
- Progettare nell'ottica Ops

Esempi correlati:

- Well-Architected Labs - Inventory and Patch Management (Gestione di inventario e patch)

OPS05-BP06 Condivisione degli standard di progettazione

Condividi le best practice con i team per incrementare la consapevolezza e potenziare al massimo i vantaggi delle attività di sviluppo.

In AWS, applicazione, elaborazione, infrastruttura e operazioni possono essere definite e gestite utilizzando metodologie di codice. In questo modo le operazioni di rilascio, condivisione e adozione saranno più semplici.

Molti servizi e risorse AWS sono progettati per essere condivisi tra account, mettendo prodotti e conoscenze a disposizione di tutti i team. Ad esempio, puoi condividere repository CodeCommit , Lambda funzioni, bucket Amazon S3e AMI con account specifici.

Quando pubblichi risorse o aggiornamenti, utilizza Amazon SNS per fornire notifiche tra account diversi. Gli iscritti possono utilizzare Lambda per ottenere le nuove versioni.

Se nella tua organizzazione vengono applicati standard condivisi, è fondamentale che esistano meccanismi per richiedere aggiunte, modifiche ed eccezioni agli standard a supporto delle attività dei team. Senza questa opzione, gli standard diventano un ostacolo per l'innovazione.

Anti-pattern comuni:

- Hai creato il tuo meccanismo di autenticazione utente, così come tutti gli altri team di sviluppo nella tua organizzazione. Gli utenti devono mantenere un set separato di credenziali per ogni parte del sistema cui desiderano accedere.
- Hai creato il tuo meccanismo di autenticazione utente, così come tutti gli altri team di sviluppo nella tua organizzazione. All'organizzazione viene assegnato un nuovo requisito di conformità che deve essere soddisfatto. Ogni singolo team di sviluppo deve ora investire le risorse per implementare il nuovo requisito.
- Hai creato il tuo layout dello schermo, così come tutti gli altri team di sviluppo nella tua organizzazione. Gli utenti si lamentano della difficoltà di navigazione nelle interfacce incoerenti.

Vantaggi dell'adozione di questa best practice: Utilizza standard condivisi per supportare l'adozione delle best practice e per massimizzare i vantaggi degli sforzi delle attività di sviluppo laddove gli standard soddisfano i requisiti di più applicazioni o organizzazioni.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Condivisione degli standard di progettazione: condividi le best practice esistenti, gli standard di progettazione, gli elenchi di controllo, le procedure operative e i requisiti su linee guida e governance tra i team per ridurre la complessità e sfruttare al massimo i vantaggi derivanti dalle attività di sviluppo. Assicurati che siano state predisposte procedure per richiedere modifiche, aggiunte ed eccezioni agli standard di progettazione a supporto del miglioramento e dell'innovazione continua. Verifica che i team siano a conoscenza del contenuto pubblicato affinché possano avvalersene e limitare ripetizioni e sprechi di energie.

- Delega dell'accesso all'ambiente AWS
- Condivisione di un repository AWS CodeCommit
- Autorizzazione semplificata delle funzioni AWS Lambda
- Condivisione di un'AMI con Account AWS specifici
- Condivisione più rapida dei modelli con un URL di AWS CloudFormation Designer
- Utilizzo di AWS Lambda con Amazon SNS

Risorse

Documenti correlati:

- Autorizzazione semplificata delle funzioni AWS Lambda
- Condivisione di un repository AWS CodeCommit
- Condivisione di un'AMI con Account AWS specifici
- Condivisione più rapida dei modelli con un URL di AWS CloudFormation Designer
- Utilizzo di AWS Lambda con Amazon SNS

Video correlati:

- Delega dell'accesso all'ambiente AWS

OPS05-BP07 Implementazione di prassi per migliorare la qualità del codice

Implementa prassi per migliorare la qualità del codice e ridurre al minimo i difetti, ad esempio sviluppo basato su test, revisioni del codice e adozione di standard.

In AWS è possibile integrare servizi come Amazon CodeGuru nella pipeline per identificare automaticamente possibili problemi legati al codice e alla sicurezza tramite analisi del programma e machine learning. CodeGuru fornisce consigli su come implementare le best pratice di AWS per affrontare tali problemi.

Anti-pattern comuni:

- Per poter testare prima la tua funzionalità, hai deciso di non integrare la libreria di purificazione degli input standard. Dopo i test, esegui il commit del codice, ma dimentichi di completare l'incorporazione della libreria.

- Hai poca esperienza con il set di dati che stai elaborando e non sai che questo può presentare dei casi limite. Questi casi limite non sono compatibili con il codice che hai implementato.

Vantaggi dell'adozione di questa best practice: L'adozione di pratiche per migliorare la qualità del codice ti consente di ridurre al minimo i problemi di produzione.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Implementazione di prassi per migliorare la qualità del codice: implementa prassi per migliorare la qualità del codice e ridurre al minimo i difetti e il rischio che vengano distribuiti, ad esempio sviluppo basato su test, programmazione in coppia, revisioni del codice e adozione di standard.
  - Amazon CodeGuru

Risorse

Documenti correlati:

- Amazon CodeGuru

OPS05-BP08 Utilizzo di più ambienti

Utilizza ambienti multipli per sperimentare, sviluppare e testare il carico di lavoro. Utilizza livelli crescenti di controlli man mano che gli ambienti si avvicinano alla fase di produzione per avere la certezza che il carico di lavoro funzionerà come previsto una volta distribuito.

Anti-pattern comuni:

- Stai sviluppando in un ambiente di sviluppo condiviso e un altro sviluppatore sovrascrive le tue modifiche al codice.
- I controlli di sicurezza restrittivi nell'ambiente di sviluppo condiviso impediscono di sperimentare nuovi servizi e funzionalità.
- Esegui test di carico sui tuoi sistemi di produzione e causa un'interruzione per i tuoi utenti.
- Si è verificato un errore critico che ha causato la perdita di dati nella produzione. Nel tuo ambiente di produzione tenti di ricreare le condizioni che portano alla perdita di dati in modo da poter identificare come si è verificata e impedire che si ripeta. Per evitare un'ulteriore perdita di dati durante il test, devi rendere l'applicazione non disponibile per i tuoi utenti.

- Stai operando un servizio multi-tenant e non sei in grado di supportare la richiesta di un cliente per un ambiente dedicato.
- Non esegui sempre test, ma quando lo fai è in produzione.
- Ritieni che la semplicità di un singolo ambiente prevalga sulla portata dell'impatto che possono avere modifiche all'interno dell'ambiente.

Vantaggi dell'adozione di questa best practice: Con la distribuzione di più ambienti, puoi supportare più ambienti di sviluppo, test e produzione simultanei senza creare conflitti tra sviluppatori o community di utenti.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Utilizzo di più ambienti: fornisci agli sviluppatori ambienti sandbox con controlli minimi per abilitare la sperimentazione. Fornisci ambienti di sviluppo individuali per abilitare il lavoro in parallelo, incrementando l'agilità dello sviluppo. Implementa controlli più rigorosi negli ambienti che si avvicinano alla produzione per consentire agli sviluppatori di innovare. Utilizza l'approccio Infrastructure-as-Code e sistemi di gestione delle configurazioni per distribuire ambienti configurati in modo coerente con i controlli presenti in produzione per assicurare che i sistemi funzionino nel modo previsto quando vengono distribuiti. Quando gli ambienti non vengono utilizzati, disattivali per evitare costi associati alle risorse inattive, ad esempio i sistemi di sviluppo nelle ore serali e nei fine settimana. Distribuisci ambienti equivalenti alla produzione quando esegui i test di carico per ottenere risultati validi.
  - Che cos'è AWS CloudFormation?
  - Come arrestare e riavviare le istanze Amazon EC2 a intervalli regolari con AWS Lambda?

Risorse

Documenti correlati:

- Come arrestare e riavviare le istanze Amazon EC2 a intervalli regolari con AWS Lambda?
- Che cos'è AWS CloudFormation?

OPS05-BP09 Applicazione di modifiche frequenti, minime e reversibili

Le modifiche frequenti, minime e reversibili riducono la portata e l'impatto di una modifica. Questo semplifica la risoluzione dei problemi, consente tempi di correzione più rapidi e permette di eseguire il rollback di una modifica.

Anti-pattern comuni:

- Viene distribuita trimestralmente una nuova versione dell'applicazione.
- Le modifiche vengono apportate frequentemente allo schema del database.
- Esegui aggiornamenti manuali sul posto, sovrascrivendo le installazioni e le configurazioni esistenti.

Vantaggi dell'adozione di questa best practice: Riconosci più rapidamente i vantaggi derivanti dalle attività di sviluppo grazie alla distribuzione frequente di piccole modifiche. Quando le modifiche sono piccole, è molto più facile identificare se hanno conseguenze indesiderate. Quando le modifiche sono reversibili, è meno rischioso implementare la modifica poiché il ripristino è più semplice.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Applicazione di modifiche frequenti, minime e reversibili: le modifiche frequenti, minime e reversibili riducono la portata e l'impatto di una modifica. Questo semplifica la risoluzione dei problemi, consente tempi di correzione più rapidi e permette di eseguire il rollback di una modifica. Inoltre, aggiunge più rapidamente valore al business.

OPS05-BP10 Automazione completa dell'integrazione e della distribuzione

Automatizza la creazione, la distribuzione e il test del carico di lavoro. Questo riduce gli errori causati dai processi manuali e l'impegno necessario per distribuire le modifiche.

Applica i metadati utilizzando i tag delle risorse e AWS Resource Groups seguendo una strategia di applicazione dei tag coerente per consentire l'identificazione delle risorse. Applica tag alle risorse per organizzare, monitorare i costi e controllare gli accessi e ottimizza l'esecuzione delle attività operative automatizzate.

Anti-pattern comuni:

- Venerdì termini la creazione del nuovo codice per il ramo delle funzionalità. Lunedì, dopo aver eseguito gli script di test di qualità del codice e ciascuno degli script dei test di unità, verificherai nel codice la prossima versione pianificata.
- Ti verrà assegnato di codificare una correzione per un problema critico che interessa un numero elevato di clienti nella produzione. Dopo aver testato la correzione, esegui il commit del codice e richiedi via e-mail alla gestione delle modifiche l'approvazione per distribuirlo in produzione.

Vantaggi dell'adozione di questa best practice: Implementando sistemi di gestione automatizzati di compilazione e distribuzione, riduci gli errori causati dai processi manuali e l'impegno necessario per distribuire le modifiche, consentendo ai membri del team di concentrarsi su attività aziendali più importanti.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Utilizzo di sistemi di gestione della creazione e distribuzione: utilizza sistemi di gestione della creazione e distribuzione per tenere traccia e implementare le modifiche, ridurre gli errori causati dai processi manuali e ridurre il livello di impegno richiesto. Automatizza completamente la pipeline di integrazione e distribuzione dal check-in del codice fino alle fasi di creazione, test, distribuzione e convalida. In questo modo è possibile ridurre il lead time, aumentare la frequenza delle modifiche e ridurre il livello di impegno richiesto.
  - What is AWS CodeBuild? (Che cos'è AWS CodeBuild?)
  - Best practice di integrazione continua per lo sviluppo del software
  - Slalom: CI/CD for serverless applications on AWS (Slalom: CI/CD per applicazioni serverless su AWS)
  - Introduction to AWS CodeDeploy - automated software deployment with Amazon Web Services (Introduzione ad AWS CodeDeploy – Implementazione automatica del software con Amazon Web Services)
  - What is AWS CodeDeploy? (Che cos'è AWS CodeDeploy?)

Risorse

Documenti correlati:

- What is AWS CodeBuild? (Che cos'è AWS CodeBuild?)

- What is AWS CodeDeploy? (Che cos'è AWS CodeDeploy?)

Video correlati:

- Best practice di integrazione continua per lo sviluppo del software
- Introduction to AWS CodeDeploy - automated software deployment with Amazon Web Services (Introduzione ad AWS CodeDeploy – Implementazione automatica del software con Amazon Web Services)
- Slalom: CI/CD for serverless applications on AWS (Slalom: CI/CD per applicazioni serverless su AWS)

OPS 6 In che modo mitighi i rischi della distribuzione?

Adotta prassi che consentano di fornire un feedback rapido sulla qualità e permettano un ripristino veloce dalle modifiche che non hanno i risultati previsti. L'uso di queste prassi consente di mitigare l'impatto dei problemi introdotti attraverso la distribuzione delle modifiche.

Best practice

- OPS06-BP01 Preparazione di un piano in caso di esito negativo delle modifiche
- OPS06-BP02 Test e convalida delle modifiche
- OPS06-BP03 Utilizzo di sistemi di gestione dell'implementazione
- OPS06-BP04 Test utilizzando implementazioni limitate
- OPS06-BP05 Distribuzione utilizzando ambienti paralleli
- OPS06-BP06 Distribuzione di modifiche frequenti, minime e reversibili
- OPS06-BP07 Automazione completa dell'integrazione e dell'implementazione
- OPS06-BP08 Automazione dei test e del rollback

OPS06-BP01 Preparazione di un piano in caso di esito negativo delle modifiche

Pianifica il ripristino di uno stato corretto noto o la correzione nell'ambiente di produzione nel caso in cui una modifica non produca il risultato desiderato. Questa preparazione riduce i tempi di ripristino grazie a risposte più veloci.

Anti-pattern comuni:

- Hai eseguito una distribuzione e l'applicazione è diventata instabile, ma sembra che ci siano utenti attivi sul sistema. Devi decidere se eseguire il rollback della modifica e influire sugli utenti attivi o attendere il rollback della modifica, sapendo che gli utenti possono essere comunque influenzati.
- Dopo aver apportato una modifica di routine, i nuovi ambienti sono accessibili, ma una delle sottoreti è diventata irraggiungibile. Devi decidere se eseguire il rollback di tutto o provare a correggere la sottorete inaccessibile. Mentre prendi tale decisione, la sottorete rimane irraggiungibile.

Vantaggi dell'adozione di questa best practice: Avere pronto un piano riduce il tempo medio di ripristino (MTTR) dalle modifiche non riuscite, riducendo di conseguenza l'impatto sugli utenti finali.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Preparazione di un piano in caso di esito negativo delle modifiche: pianifica il ripristino di uno stato corretto noto (ovvero, esegui il rollback della modifica) o la correzione nell'ambiente di produzione (ovvero, esegui il roll forward della modifica) nel caso in cui una modifica non produca il risultato desiderato. In presenza di modifiche di cui non è possibile effettuare il rollback in caso di esito negativo, presta la dovuta attenzione prima di eseguire il commit.

OPS06-BP02 Test e convalida delle modifiche

Testa le modifiche e convalida i risultati in tutte le fasi del ciclo di vita per confermare le nuove funzionalità e ridurre al minimo il rischio e l'impatto delle distribuzioni non riuscite.

Su AWS puoi creare ambienti temporanei paralleli che riducono il rischio, lo sforzo e il costo della sperimentazione e dell'esecuzione di test. Automatizza la distribuzione di questi ambienti utilizzando AWS CloudFormation per garantire implementazioni coerenti degli ambienti temporanei.

Anti-pattern comuni:

- Viene distribuita una nuova funzionalità interessante nella tua applicazione. Non funziona. Non lo sai.
- I certificati vengono aggiornati. Installi accidentalmente i certificati nei componenti errati. Non lo sai.

Vantaggi dell'adozione di questa best practice: Testando e convalidando le modifiche dopo la distribuzione, sarai in grado di identificare tempestivamente i problemi offrendo l'opportunità di mitigare l'impatto sui clienti.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Test e convalida delle modifiche: testa le modifiche e convalida i risultati in tutte le fasi del ciclo di vita (ad esempio, distribuzione, test e produzione) per confermare le nuove funzionalità e ridurre al minimo il rischio e l'impatto delle distribuzioni non riuscite.
  - AWS Cloud9
  - Cos'è AWS Cloud9?
  - Come eseguire test e debug con AWS CodeDeploy in locale prima di distribuire il codice

Risorse

Documenti correlati:

- AWS Cloud9
- Strumenti per gli sviluppatori di AWS
- Come eseguire test e debug con AWS CodeDeploy in locale prima di distribuire il codice
- Cos'è AWS Cloud9?

OPS06-BP03 Utilizzo di sistemi di gestione dell'implementazione

Usa sistemi di gestione della distribuzione per monitorare e implementare una modifica. Questo riduce gli errori causati dai processi manuali e l'impegno necessario per distribuire le modifiche.

In AWS, puoi compilare pipeline di integrazione continua/implementazione continua (CI/CD) utilizzando servizi come gli Strumenti per sviluppatori in AWS (ad esempio, AWS CodeCommit, AWS CodeBuild, AWS CodePipeline, AWS CodeDeploye AWS CodeStar).

Anti-pattern comuni:

- Distribuisci manualmente gli aggiornamenti dei server applicazioni all'interno del parco istanze e alcuni server non rispondono a causa di errori di aggiornamento.

- Distribuisci manualmente il parco istanze del server applicazioni nel corso di molte ore. L'incoerenza nelle versioni durante la modifica causa comportamenti imprevisti.

Vantaggi dell'adozione di questa best practice: L'adozione di sistemi di gestione dell'implementazione riduce il livello di impegno necessario per implementare modifiche e la frequenza degli errori causati dalle procedure manuali.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Utilizzo di sistemi di gestione dell'implementazione: utilizza sistemi di gestione della distribuzione per monitorare e implementare le modifiche. Questo ridurrà gli errori causati dai processi manuali e il livello di impegno richiesto per la distribuzione delle modifiche. Automatizza completamente la pipeline di integrazione e distribuzione dal check-in del codice fino alle fasi di test, distribuzione e convalida. In questo modo è possibile ridurre il lead time, aumentare la frequenza delle modifiche e ridurre ulteriormente il livello di impegno richiesto.
  - Introduzione ad AWS CodeDeploy - Sviluppo di software automatizzato con Amazon Web Services
  - Che cos'è AWS CodeDeploy?
  - Che cos'è AWS Elastic Beanstalk?
  - Che cos'è Amazon API Gateway?

Risorse

Documenti correlati:

- Guida per l'utente di AWS CodeDeploy
- Strumenti per sviluppatori in AWS
- Prova un'implementazione blu/verde di esempio in AWS CodeDeploy
- Che cos'è AWS CodeDeploy?
- Che cos'è AWS Elastic Beanstalk?
- Che cos'è Amazon API Gateway?

Video correlati:

- Approfondimento sulle tecniche avanzate di distribuzione continua utilizzando AWS
- Introduzione ad AWS CodeDeploy - Sviluppo di software automatizzato con Amazon Web Services

OPS06-BP04 Test utilizzando implementazioni limitate

Esegui test con distribuzioni limitate accanto ai sistemi esistenti per confermare i risultati desiderati prima della distribuzione su vasta scala. Ad esempio, utilizza test della distribuzione di tipo canary oppure distribuzioni one-box.

Anti-pattern comuni:

- Distribuisci una modifica non riuscita a tutta la produzione contemporaneamente. Non lo sai.

Vantaggi dell'adozione di questa best practice: Testando e convalidando le modifiche dopo la distribuzione limitata, sarai in grado di identificare tempestivamente i problemi con un impatto minimo sui clienti offrendo l'opportunità di mitigare ulteriormente quest'ultimo.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Test con implementazioni limitate: esegui test con implementazioni limitate insieme ai sistemi esistenti per confermare i risultati desiderati prima dell'implementazione su vasta scala. Ad esempio, utilizza test della distribuzione di tipo canary oppure distribuzioni one-box.
  - Guida per l'utente di AWS CodeDeploy
  - Implementazione blu/verde con AWS Elastic Beanstalk
  - Configurazione dell'implementazione di una release canary di API Gateway
  - Prova un'implementazione blu/verde di esempio in AWS CodeDeploy
  - Utilizzo di configurazioni di implementazione in AWS CodeDeploy

Risorse

Documenti correlati:

- Guida per l'utente di AWS CodeDeploy
- Implementazione blu/verde con AWS Elastic Beanstalk
- Configurazione dell'implementazione di una release canary di API Gateway

- Prova un'implementazione blu/verde di esempio in AWS CodeDeploy
- Utilizzo di configurazioni di implementazione in AWS CodeDeploy

OPS06-BP05 Distribuzione utilizzando ambienti paralleli

Implementa le modifiche in ambienti paralleli, quindi esegui la transizione al nuovo ambiente. Mantieni l'ambiente precedente finché non viene confermata la riuscita della distribuzione. In questo modo si riducono i tempi di ripristino grazie alla possibilità di eseguire il rollback all'ambiente precedente.

Anti-pattern comuni:

- Esegui una distribuzione variabile modificando i sistemi esistenti. Dopo aver scoperto che la modifica non è andata a buon fine, devi modificare nuovamente i sistemi per ripristinare la versione precedente estendendo il tempo di ripristino.
- Durante una finestra di manutenzione, disattivi l'ambiente precedente, quindi inizi a creare il nuovo ambiente. Molte ore dopo aver iniziato la procedura, scopri problemi irrecuperabili con la distribuzione. La stanchezza è grande, ma devi trovare le procedure di distribuzione precedenti e iniziare a ricostruire il vecchio ambiente.

Vantaggi dell'adozione di questa best practice: Utilizzando ambienti paralleli puoi distribuire anticipatamente il nuovo ambiente e passare a esso quando lo desideri. Se nel nuovo ambiente ci sono problemi, puoi eseguire rapidamente il ripristino al tuo ambiente originale.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Distribuzione con ambienti paralleli: implementa le modifiche in ambienti paralleli, quindi effettua la transizione o passa al nuovo ambiente. Mantieni l'ambiente precedente finché non viene confermata la riuscita della distribuzione. Questo riduce i tempi di ripristino grazie alla possibilità di eseguire il rollback all'ambiente precedente. Ad esempio, usa infrastrutture immutabili con distribuzioni blue/green.
  - Utilizzo di configurazioni di distribuzione in AWS CodeDeploy
  - Distribuzioni blu/verde con AWS Elastic Beanstalk
  - Configurazione della distribuzione di una release Canary di API Gateway
  - Prova un'implementazione blu/verde di esempio in AWS CodeDeploy

Risorse

Documenti correlati:

- Guida per l'utente di AWS CodeDeploy
- Distribuzioni blu/verde con AWS Elastic Beanstalk
- Configurazione della distribuzione di una release Canary di API Gateway
- Prova un'implementazione blu/verde di esempio in AWS CodeDeploy
- Utilizzo di configurazioni di distribuzione in AWS CodeDeploy

Video correlati:

- Approfondimento sulle tecniche avanzate di distribuzione continua utilizzando AWS

OPS06-BP06 Distribuzione di modifiche frequenti, minime e reversibili

Utilizza modifiche frequenti, minime e reversibili per ridurre la portata e l'impatto di una modifica. Semplificherai così la risoluzione dei problemi, accelerando la correzione e mantenendo la possibilità di rollback delle modifiche.

Anti-pattern comuni:

- Viene distribuita trimestralmente una nuova versione dell'applicazione.
- Le modifiche vengono apportate frequentemente allo schema del database.
- Esegui aggiornamenti manuali sul posto, sovrascrivendo le installazioni e le configurazioni esistenti.

Vantaggi dell'adozione di questa best practice: Riconosci più rapidamente i vantaggi derivanti dalle attività di sviluppo grazie alla distribuzione frequente di piccole modifiche. Quando le modifiche sono piccole, è molto più facile identificare se hanno conseguenze indesiderate. Quando le modifiche sono reversibili, il rischio di implementare la modifica man mano che il ripristino viene semplificato è minore.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Distribuzione di modifiche frequenti, minime e reversibili: utilizza modifiche frequenti, minime e reversibili per ridurre la portata di una modifica. Semplificherai così la risoluzione dei problemi, accelerando la correzione e mantenendo la possibilità di rollback delle modifiche.

OPS06-BP07 Automazione completa dell'integrazione e dell'implementazione

Automatizza la creazione, la distribuzione e il test del carico di lavoro. Questo riduce gli errori causati dai processi manuali e l'impegno necessario per distribuire le modifiche.

Applica i metadati utilizzando i tag delle risorse e AWS Resource Groups seguendo una strategia di applicazione dei tag coerente per consentire l'identificazione delle risorse. Applica tag alle risorse per organizzare, monitorare i costi e controllare gli accessi e ottimizza l'esecuzione delle attività operative automatizzate.

Anti-pattern comuni:

- Venerdì termini la creazione del nuovo codice per il ramo delle funzionalità. Lunedì, dopo aver eseguito gli script di test di qualità del codice e ciascuno degli script dei test di unità, verificherai nel codice la prossima versione pianificata.
- Ti verrà assegnato di codificare una correzione per un problema critico che interessa un numero elevato di clienti nella produzione. Dopo aver testato la correzione, esegui il commit del codice e richiedi via e-mail alla gestione delle modifiche l'approvazione per implementarlo in produzione.

Vantaggi dell'adozione di questa best practice: Implementando sistemi di gestione automatizzati di compilazione e distribuzione, riduci gli errori causati dai processi manuali e lo sforzo di distribuire le modifiche consentendo ai membri del team di concentrarsi sull'offerta di valore aggiunto.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Utilizzo di sistemi di gestione della compilazione e dell'implementazione: utilizza sistemi di gestione della compilazione e dell'implementazione per tenere traccia e realizzare le modifiche, ridurre gli errori causati dai processi manuali e ridurre il livello di impegno richiesto. Automatizza completamente la pipeline di integrazione e distribuzione dal check-in del codice fino alle fasi di creazione, test, distribuzione e convalida. In questo modo è possibile ridurre il lead time, aumentare la frequenza delle modifiche e ridurre il livello di impegno richiesto.

- Che cos'è AWS CodeBuild?
- Best practice di integrazione continua per lo sviluppo del software
- Slalom: CI/CD per applicazioni serverless su AWS
- Introduzione ad AWS CodeDeploy – Implementazione automatica del software con Amazon Web Services)
- What is AWS CodeDeploy? (Che cos'è AWS CodeDeploy?)
- Approfondimento sulle tecniche avanzate di distribuzione continua utilizzando AWS

Risorse

Documenti correlati:

- Prova un'implementazione blu/verde di esempio in AWS CodeDeploy
- Che cos'è AWS CodeBuild?
- What is AWS CodeDeploy? (Che cos'è AWS CodeDeploy?)

Video correlati:

- Best practice di integrazione continua per lo sviluppo del software
- Approfondimento sulle tecniche avanzate di distribuzione continua utilizzando AWS
- Introduzione ad AWS CodeDeploy – Implementazione automatica del software con Amazon Web Services)
- Slalom: CI/CD per applicazioni serverless su AWS

OPS06-BP08 Automazione dei test e del rollback

Automatizza i test degli ambienti distribuiti per confermare i risultati desiderati. Automatizza il rollback a uno stato corretto noto quando non vengono raggiunti i risultati previsti, per ridurre al minimo il tempo di ripristino e gli errori causati dai processi manuali.

Anti-pattern comuni:

- Distribuisci le modifiche al carico di lavoro. Una volta completata la modifica, inizi il test post-implementazione. Al completamento, ti accorgi che il carico di lavoro è inutilizzabile e i clienti sono disconnessi. Inizi quindi a eseguire il rollback alla versione precedente. Dopo un lungo periodo di tempo per rilevare il problema, il tempo di ripristino viene esteso dalla reimplementazione manuale.

Vantaggi dell'adozione di questa best practice: Testando e convalidando le modifiche dopo l'implementazione, puoi identificare immediatamente i problemi. Effettuando automaticamente il rollback alla versione precedente, l'impatto sui clienti viene ridotto al minimo.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Automazione di test e rollback: automatizza i test degli ambienti implementati per confermare i risultati desiderati. Automatizza il rollback a uno stato corretto noto quando non vengono raggiunti i risultati previsti, per ridurre al minimo il tempo di ripristino e gli errori causati dai processi manuali. Ad esempio, esegui transazioni utente sintetiche dettagliate dopo la distribuzione, verifica i risultati ed esegui il rollback se l'esito è negativo.
  - Reimplementazione e rollback di una implementazione con AWS CodeDeploy

Risorse

Documenti correlati:

- Reimplementazione e rollback di una implementazione con AWS CodeDeploy

OPS 7 Come fai a sapere che sei pronto a supportare un carico di lavoro?

Valuta la disponibilità operativa del carico di lavoro, dei processi e delle procedure, nonché del personale per comprendere i rischi operativi correlati al carico di lavoro.

Best practice

- OPS07-BP01 Verifica della capacità del personale
- OPS07-BP02 Revisione costante della prontezza operativa
- OPS07-BP03 Utilizzo di runbook per eseguire le procedure
- OPS07-BP04 Utilizzo dei playbook per analizzare i problemi
- OPS07-BP05 Adozione di decisioni informate per implementare sistemi e modifiche

OPS07-BP01 Verifica della capacità del personale

Predisponi un meccanismo per stabilire se disponi del numero appropriato di risorse qualificate per supportare le esigenze operative. Forma il personale e adegua la dotazione di personale, se necessario, per mantenere un supporto efficace.

Assicurati che il numero di membri del team di cui disponi sia sufficiente per coprire le attività operative, inclusa la reperibilità. Assicurati che i tuoi team dispongano delle competenze necessarie per avere successo con la formazione sul carico di lavoro, gli strumenti operativi e AWS.

AWS fornisce delle risorse, tra cui centro risorse per Nozioni di base su AWS, Blog AWS, AWS OnlineTech Talks, Eventi e webinar AWSe gli AWS Well-Architected Labs, che forniscono indicazioni, esempi e procedure guidate dettagliate per formare i team. Inoltre, AWS Training and Certification offre risorse di formazione gratuite tramite corsi digitali gestiti dall'utente sulle nozioni di base di AWS. Per supportare ulteriormente lo sviluppo delle competenze AWS del tuo team, è anche possibile iscriversi a corsi di formazione con istruttore.

Anti-pattern comuni:

- Distribuire un carico di lavoro senza membri del team qualificati per supportare la piattaforma e i servizi in uso.
- Distribuire un carico di lavoro senza membri del team disponibili durante le ore di supporto previste.
- Distribuire un carico di lavoro senza risorse sufficienti per supportarlo alcuni membri del team sono in congedo o in malattia.
- Distribuire carichi di lavoro aggiuntivi senza rivedere l'impatto aggiuntivo sui membri del team che supportano tale servizio e altri carichi di lavoro.

Vantaggi dell'adozione di questa best practice: Membri del team qualificati costituiscono un supporto efficace al carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Capacità del personale: verifica di disporre di personale sufficientemente qualificato per supportare il carico di lavoro in modo efficace.
  - Dimensioni del team: assicurati che il numero di membri del team di cui disponi sia sufficiente per coprire le attività operative, inclusa la reperibilità.

- Capacità del team: assicurati che i membri del team abbiano una formazione sufficiente su AWS, sul carico di lavoro e sugli strumenti operativi per svolgere il proprio lavoro.
  - Eventi e webinar AWS
  - Ti diamo il benvenuto in AWS Training and Certification
- Esame delle capacità: esamina la dimensione e le competenze del team man mano che cambiano le condizioni operative e i carichi di lavoro, per assicurarti di disporre di capacità sufficienti per mantenere l'eccellenza operativa. Effettua gli aggiustamenti necessari per garantire che la dimensione e le competenze del team siano in linea con i requisiti operativi per i carichi di lavoro supportati dal team.

Risorse

Documenti correlati:

- Blog AWS
- Eventi e webinar AWS
- centro risorse per Nozioni di base su AWS
- AWS OnlineTech Talks
- Ti diamo il benvenuto in AWS Training and Certification

Esempi correlati:

- Well-Architected Labs

OPS07-BP02 Revisione costante della prontezza operativa

Usa le revisioni della prontezza operativa (ORR) per verificare che puoi utilizzare il carico di lavoro. ORR è un meccanismo sviluppato da Amazon per verificare che i team possano utilizzare in sicurezza i propri carichi di lavoro. ORR è un processo di revisione e ispezione che utilizza un elenco di controllo per i requisiti. È un'esperienza self-service che i team utilizzano per certificare i propri carichi di lavoro. Le ORR includono le best practice delle lezioni apprese durante gli anni dedicati alla creazione di software.

Un elenco di controllo ORR è composto da suggerimenti sull'architettura, processo operativo, gestione degli eventi e qualità del rilascio. Il nostro processo di correzione dell'errore (CoE, Correction of Error) è uno dei principali fattori trainanti di questi elementi. L'analisi post-incidente

deve guidare l'evoluzione della ORR. Una ORR non riguarda solo l'adozione delle best practice, ma anche la prevenzione del ripetersi di eventi già visti. Infine, in una ORR possono essere inclusi anche i requisiti di sicurezza, governance e conformità.

Esegui le ORR prima che un carico di lavoro venga lanciato nella disponibilità generale e quindi durante tutto il ciclo di vita dello sviluppo software. L'esecuzione della ORR prima del lancio aumenta la tua capacità di utilizzare il carico di lavoro in sicurezza. Riesegui periodicamente la ORR sul carico di lavoro per cogliere eventuali scostamenti dalle best practice. Puoi usare gli elenchi di controllo ORR per il lancio di nuovi servizi e le ORR per le revisioni periodiche. In tal modo puoi tenerti aggiornato sulle nuove best practice che emergono e incorporare le lezioni apprese dall'analisi post-incidente. Man mano che l'utilizzo del cloud cresce, puoi creare i requisiti di ORR nella tua architettura come valori predefiniti.

Risultato desiderato: hai un elenco di controllo ORR con le best practice per la tua organizzazione. Le ORR vengono eseguite prima dell'avvio dei carichi di lavoro. Le ORR vengono eseguite periodicamente nel corso del ciclo di vita del carico di lavoro.

Anti-pattern comuni:

- Avvii un carico di lavoro senza sapere se puoi utilizzarlo.
- I requisiti di governance e sicurezza non sono inclusi nella certificazione di un carico di lavoro per l'avvio.
- I carichi di lavoro non vengono rivalutati periodicamente.
- I carichi di lavoro vengono avviati senza le procedure richieste.
- Si osserva la ripetizione di errori con la stessa causa principale in più carichi di lavoro.

Vantaggi dell'adozione di questa best practice:

- I tuoi carichi di lavoro includono le best practice di architettura, processo e gestione.
- Le lezioni apprese sono incorporate nel processo ORR.
- Le procedure richieste sono in atto all'avvio dei carichi di lavoro.
- Le ORR vengono eseguite durante l'intero ciclo di vita del software dei carichi di lavoro.

Livello di rischio se questa best practice non fosse adottata: alto

Guida all'implementazione

Una ORR è composta da un processo e un elenco di controllo. Il processo ORR deve essere adottato dall'organizzazione e supportato da uno sponsor esecutivo. Come minimo, le ORR devono essere eseguite prima che il carico di lavoro venga lanciato nella disponibilità generale. Esegui la ORR durante tutto il ciclo di vita dello sviluppo software per mantenerlo aggiornato con le best practice o i nuovi requisiti. L'elenco di controllo ORR deve includere elementi di configurazione, requisiti di sicurezza e governance e best practice dell'organizzazione. Nel tempo, puoi utilizzare i servizi, come AWS Config, AWS Security Hube AWS Control Tower Guardrailsper creare le best practice dalla ORR nei guardrail per il rilevamento automatico delle best practice.

Esempio del cliente

Dopo diversi incidenti di produzione, AnyCompany Retail ha deciso di implementare un processo ORR. Ha creato un elenco di controllo composto da best practice, requisiti di governance e conformità e lezioni apprese dalle interruzioni. I nuovi carichi di lavoro conducono le ORR prima dell'avvio. Ogni carico di lavoro esegue una ORR annuale con un sottoinsieme di best practice per incorporare nuove best practice e requisiti che vengono aggiunti all'elenco di controllo ORR. Nel tempo, AnyCompany Retail ha utilizzato AWS Config per individuare le best practices, accelerando il processo ORR.

Passaggi dell'implementazione

Per ulteriori informazioni sulle ORR, consulta il whitepaper Operational Readiness Reviews (ORR) (Revisioni della prontezza operativa (ORR)). Fornisce informazioni dettagliate sulla cronologia del processo ORR, su come creare la procedura ORR e su come sviluppare il proprio elenco di controllo ORR. I passaggi seguenti costituiscono una versione abbreviata di quel documento. Per una comprensione approfondita di cosa sono le ORR e di come crearne una, ti consigliamo di leggere il whitepaper.

1. Riunisci gli stakeholder importanti, inclusi i rappresentanti della sicurezza, delle operazioni e dello sviluppo.
2. Chiedi a ogni stakeholder di indicare almeno un requisito. Per la prima iterazione, prova a limitare il numero di elementi a trenta al massimo.
   - Appendix B: Example ORR questions (Appendice B: Domande ORR di esempio) del whitepaper Operational Readiness Reviews (ORR) (Revisioni della prontezza operativa (ORR)) contiene domande di esempio che puoi utilizzare per iniziare.
3. Raccogli i tuoi requisiti in un foglio di calcolo.

- Puoi utilizzare gli obiettivi personalizzati nella funzione AWS Well-Architected Tool per sviluppare la ORR e condividerla tra i tuoi account e l'organizzazione AWS.

4. Identifica un carico di lavoro su cui condurre la ORR. L'ideale è un carico di lavoro pre-lancio o un carico di lavoro interno.
5. Scorri l'elenco di controllo ORR e prendi nota di tutti i rilevamenti fatti. I rilevamenti potrebbero non essere validi se è in atto una mitigazione. Aggiungi qualsiasi rilevamento privo di mitigazione al tuo backlog di elementi e implementalo prima del lancio.
6. Continua ad aggiungere le best practice e i requisiti all'elenco di controllo ORR nel corso del tempo.

I clienti di AWS Support con supporto Enterprise possono richiedere il workshop Operational Readiness Review (Revisione sulla prontezza operativa) al proprio Technical Account Manager (TAM). Il workshop è una sessione interattiva di lavoro a ritroso per sviluppare il tuo elenco di controllo ORR.

Livello di impegno per il piano di implementazione: alto. L'adozione di una procedura ORR nella tua organizzazione richiede la sponsorizzazione dell'esecutivo e l'adesione degli stakeholder. Crea e aggiorna l'elenco di controllo con input provenienti da tutta l'organizzazione.

Risorse

Best practice correlate:

- OPS01-BP03 Valutazione dei requisiti di governance - I requisiti di governance sono una scelta naturale per un elenco di controllo ORR.
- OPS01-BP04 Valutazione dei requisiti di conformità - I requisiti di conformità sono talvolta inclusi in un elenco di controllo ORR. Altre volte costituiscono un processo separato.
- OPS03-BP07 Fornitura di risorse appropriate ai team - La capacità del team è un buon requisito ORR.
- OPS06-BP01 Preparazione di un piano in caso di esito negativo delle modifiche - Prima di avviare il carico di lavoro, è necessario stabilire un piano di rollback o rollforward.
- OPS07-BP01 Verifica della capacità del personale - Per supportare un carico di lavoro è necessario disporre del personale necessario.
- SEC01-BP03 Identificazione e convalida degli obiettivi di controllo - Gli obiettivi di controllo della sicurezza costituiscono eccellenti requisiti ORR.

- REL13-BP01 Definizione degli obiettivi di ripristino in caso di downtime e perdita di dati - I piani di ripristino di emergenza sono un buon requisito ORR.
- COST02-BP01 Sviluppo di politiche basate sui requisiti dell'organizzazione - Le policy di gestione dei costi sono utili da includere nell'elenco di controllo ORR.

Documenti correlati:

- AWS Control Tower - Guardrails in AWS Control Tower (Guardrail in AWS Control Tower)
- AWS Well-Architected Tool - Custom Lenses (Obiettivi personalizzati)
- Operational Readiness Review Template by Adrian Hornsby (Modello di revisione della prontezza operativa di Adrian Hornsby)
- Whitepaper Operational Readiness Reviews (ORR) (Revisioni della prontezza operativa (ORR))

Video correlati:

- AWS Supports You | Building an Effective Operational Readiness Review (ORR) (AWS ti supporta | Creazione di un'efficace revisione della prontezza operativa (ORR))

Esempi correlati:

- Sample Operational Readiness Review (ORR) Lens (Esempio di obiettivi per la revisione della prontezza operativa (ORR))

Servizi correlati:

- AWS Config
- AWS Control Tower
- AWS Security Hub
- AWS Well-Architected Tool

OPS07-BP03 Utilizzo di runbook per eseguire le procedure

Un runbook è un processo documentato finalizzato al raggiungimento di un determinato risultato. I runbook sono composti da una serie di passaggi che è necessario eseguire per conseguire un obiettivo. L'uso dei runbook può essere fatto risalire agli albori dell'aviazione. Nelle operazioni cloud,

è possibile utilizzare i runbook per ridurre i rischi e ottenere i risultati desiderati. In estrema sintesi, un runbook è un elenco di controllo da seguire per completare un'attività.

I runbook costituiscono una parte essenziale del funzionamento dei carichi di lavoro. Dall'inserimento di un nuovo membro in un team all'implementazione di una versione principale, i runbook sono processi codificati che garantiscono risultati coerenti indipendentemente da chi li utilizza. I runbook devono essere pubblicati a livello centralizzato e aggiornati in base all'evoluzione del processo. L'aggiornamento dei runbook rappresenta infatti un elemento chiave dell'intero processo di gestione delle modifiche. Devono inoltre includere le linee guida relative a gestione degli errori, strumenti, autorizzazioni, eccezioni ed escalation in caso di problemi.

A mano a mano che l'organizzazione cresce, è consigliabile automatizzare i runbook. Inizia con runbook concisi e di frequente utilizzo. Utilizza un linguaggio di scripting per automatizzare le procedure o semplificarne l'esecuzione. Dopo aver automatizzato i primi runbook, potrai dedicare altro tempo all'automazione dei runbook più complessi. Gradualmente dovrai automatizzare la maggior parte dei runbook.

Risultato desiderato: il team dispone di una raccolta di linee guida dettagliate per l'esecuzione delle attività relative ai carichi di lavoro. I runbook contengono il risultato desiderato, gli strumenti e le autorizzazioni necessari e le istruzioni per la gestione degli errori. Vengono archiviati in una posizione centralizzata e aggiornati di frequente.

Anti-pattern comuni:

- Ricorso alla memoria per completare i singoli passaggi di un processo.
- Implementazione manuale delle modifiche senza utilizzare un elenco di controllo.
- Vari membri dei team eseguono lo stesso processo con procedure o risultati diversi.
- Mancato aggiornamento dei runbook in base alle modifiche o ai processi di automazione del sistema.

Vantaggi dell'adozione di questa best practice:

- Riduzione della percentuale degli errori per le attività manuali.
- Le operazioni vengono eseguite in modo coerente.
- I nuovi membri dei team possono essere operativi da subito.
- I runbook possono essere automatizzati per semplificare le operazioni più impegnative.

Livello di rischio associato se questa best practice non fosse adottata: medio

Guida all'implementazione

I runbook possono avere vari formati, a seconda del livello di "maturità" dell'organizzazione. Nella loro formulazione minima, devono essere un documento di testo in cui sono dettagliate le procedure. Il risultato desiderato deve essere indicato in modo chiaro e preciso. Devono inoltre documentare in modo chiaro le autorizzazioni e gli strumenti speciali necessari. Devono includere linee guida dettagliate relative alle gestione degli errori e ai livelli di escalation nel caso in cui si verifichino problemi o errori. I runbook devono riportare il nome del proprietario ed essere pubblicati in una posizione centralizzata. Dopo averlo compilato, un runbook deve essere convalidato. A tale scopo, devi far eseguire il runbook da un membro diverso del tuo team. A mano a mano che la procedura si evolve, aggiorna i runbook in base al processo di gestione delle modifiche.

I runbook in formato testuale devono essere automatizzati a seconda dell'evoluzione dell'organizzazione. Utilizzando servizi come Automazioni AWS Systems Manager, puoi trasformare un testo non formattato in automazioni che possono essere eseguite nell'ambito di un carico di lavoro. Queste automazioni possono essere eseguite in risposta a eventi, per ridurre il carico operativo a salvaguardia del carico di lavoro.

Esempio del cliente

AnyCompany Retail deve eseguire aggiornamenti dello schema del database durante le implementazioni del software. Il team responsabile delle operazioni cloud ha lavorato assieme al team addetto all'amministrazione del database per redigere un runbook per l'implementazione manuale di queste modifiche. Nel runbook sono incluse le procedure dettagliate sotto forma di elenco di controllo. È presente anche una sezione sulla gestione degli errori in caso di problemi. Il runbook è stato pubblicato assieme ad altri runbook sul wiki interno. Il team responsabile delle operazioni cloud pensa di pianificare l'automazione del runbook in futuro.

Passaggi dell'implementazione

Se non è presente un repository di documenti, è consigliabile creare una libreria di runbook utilizzando un repository per il controllo delle versioni. Puoi creare i runbook utilizzando Markdown. Di seguito è riportato un modello di runbook di esempio che è possibile utilizzare come riferimento per la creazione dei runbook.

```
# Titolo runbook ## Informazioni runbook | ID runbook | Descrizione | Strumenti
 utilizzati | Autorizzazioni speciali | Autore runbook | Data ultimo aggiornamento |
 POC escalation | |-------|-------|-------|-------|-------|-------|-------| | RUN001 |
```

```
Argomento runbook Risultato desiderato | Strumenti | Autorizzazioni | Nome e cognome |
21-09-2022 | Nome escalation | ## Passaggi 1. Passaggio 1 2. Passaggio 2
```

1. Se non disponi di un repository o di un wiki per la documentazione, crea un repository per il controllo delle versioni nel sistema di controllo delle versioni in uso.
2. Individua un processo che non ha un runbook. Un processo ideale è un processo eseguito a cadenza più o meno regolare, con un numero limitato di passaggi e con errori a basso impatto.
3. Nel repository di documenti, crea una nuova bozza di documento Markdown utilizzando il modello. Compila il campo `Titolo runbook` e i campi obbligatori nell'area `Informazioni runbook`.
4. Partendo dal primo passaggio, compila l'area `Passaggi` del runbook.
5. Associa il runbook a un membro del team. Chiedi a tale membro di utilizzare il runbook per convalidare i passaggi. In caso di informazioni mancanti o poca chiarezza, aggiorna il runbook.
6. Pubblica il runbook nell'archivio della documentazione interna. Comunica l'avvenuta pubblicazione al team e alle altre parti interessate.
7. In questo modo, nel corso del tempo creerai una libreria di runbook. A mano a mano che la libreria cresce, comincia a pensare di automatizzare i runbook.

Livello di impegno per il piano di implementazione: basso Lo standard minimo previsto per i runbook è una guida dettagliata in formato testo. L'automazione dei runbook può aumentare l'impegno a livello di implementazione.

Risorse

Best practice correlate:

- OPS02-BP02 Assegnazione di proprietari identificati a processi e procedure: i runbook devono avere un proprietario responsabile della loro manutenzione.
- OPS07-BP04 Utilizzo dei playbook per analizzare i problemi: i runbook e i playbook sono pressoché simili, con un'unica differenza, ovvero in un runbook è previsto un risultato desiderato. In molti casi, i runbook vengono attivati dopo che un playbook ha individuato una causa principale.
- OPS10-BP01 Utilizzo di un processo per la gestione di eventi, incidenti e problemi: i runbook costituiscono una best practice per la gestione di eventi, incidenti e problemi.
- OPS10-BP02 Definizione di un processo per ogni avviso: i runbook e i playbook devono essere utilizzati in risposta agli avvisi. Nel corso del tempo queste reazioni devono essere automatizzate.
- OPS11-BP04 Gestione delle conoscenze: la gestione dei runbook è un elemento fondamentale della gestione delle conoscenze.

Documenti correlati:

- Achieving Operational Excellence using automated playbook and runbook (Eccellenza operativa mediante playbook e runbook automatizzati)
- AWS Systems Manager: Working with runbooks (AWS Systems Manager: Utilizzo dei runbook)
- Migration playbook for AWS large migrations - Task 4: Improving your migration runbooks (Playbook per la migrazione per migrazioni AWS di grandi dimensioni - Attività 4: Ottimizzazione dei runbook per la migrazione)
- Use AWS Systems Manager Automation runbooks to resolve operational tasks (Utilizzo dei runbook di Automazione AWS Systems Manager per la risoluzione delle attività operative)

Video correlati:

- AWS re:Invent 2019: DIY guide to runbooks, incident reports, and incident response (SEC318-R1) (Guida fai da te per runbook, report e risposte relativi agli incidenti [SEC318-R1])
- How to automate IT Operations on AWS | Amazon Web Services (Procedure di automazione delle operazioni IT in AWS | Amazon Web Services)
- Integrate Scripts into AWS Systems Manager (Integrazione di script in AWS Systems Manager)

Esempi correlati:

- AWS Systems Manager: Automation walkthroughs (Procedure di automazione dettagliate)
- AWS Systems Manager: Restore a root volume from the latest snapshot runbook (Runbook per il ripristino di un volume root volume dallo snapshot più recente)
- Building an AWS incident response runbook using Jupyter notebooks and CloudTrail Lake (Creazione di un runbook per le risposte agli incidenti AWS mediante notebook Jupyter e data lake CloudTrail)
- Gitlab - Runbooks (Runbook)
- Rubix - A Python library for building runbooks in Jupyter Notebooks (Rubix - Libreria Python per la creazione di runbook in notebook Jupyter)
- Using Document Builder to create a custom runbook (Utilizzo di Document Builder per creare un runbook personalizzato)
- Well-Architected Labs: Automating operations with Playbooks and Runbooks (Automazione delle operazioni con playbook e runbook)

Servizi correlati:

- AWS Systems Manager Automation (Automazione AWS Systems Manager)

OPS07-BP04 Utilizzo dei playbook per analizzare i problemi

I playbook sono guide dettagliate che vengono utilizzate quando si verificano incidenti per analizzare, valutare l'impatto e identificare la causa principale del problema. I playbook sono utili in molti scenari diversi, dalle implementazioni non riuscite agli incidenti di sicurezza. In molti casi, i playbook identificano la causa principale che viene poi mitigata tramite un runbook. I playbook costituiscono un componente essenziale dei piani di risposta agli incidenti di ogni organizzazione.

Un buon playbook include diverse funzionalità chiave che guidano l'utente, passo dopo passo, nel processo di rilevamento. Ma quali passaggi deve eseguire l'utente per diagnosticare un incidente? Illustra chiaramente nel playbook se sono necessari strumenti speciali o autorizzazioni elevate. È essenziale predisporre un piano di comunicazione per aggiornare gli stakeholder sullo stato dell'analisi. Nelle situazioni in cui non è possibile identificare la causa principale, il playbook deve prevedere un piano di escalation. Se viene identificata la causa principale, il playbook deve includere il riferimento di un runbook che descrive come risolvere il problema. I playbook devono essere archiviati centralmente e aggiornati regolarmente. Se i playbook vengono utilizzati per avvisi specifici, fornisci al team i riferimenti dei playbook all'interno degli avvisi.

Man mano che l'organizzazione acquisisce maturità, puoi automatizzare i playbook. Inizia con i playbook che trattano incidenti a basso rischio. Utilizza gli script per automatizzare i passaggi di rilevamento. Assicurati di avere i relativi runbook per mitigare le cause principali più comuni.

Risultato desiderato: l'organizzazione dispone dei playbook per gli incidenti comuni. I playbook sono archiviati in una posizione centrale e disponibili per i membri del team. I playbook vengono aggiornati frequentemente. Per qualsiasi causa principale nota, vengono creati i relativi runbook.

Anti-pattern comuni:

- Non esiste un modo standard per analizzare un incidente.
- I membri del team confidano nella "memoria muscolare" o nelle conoscenze istituzionali per risolvere i problemi di un'implementazione non riuscita.
- I nuovi membri del team apprendono come analizzare i problemi attraverso tentativi ed errori.
- Le best practice per l'analisi dei problemi non sono condivise tra i team.

Vantaggi dell'adozione di questa best practice:

- I playbook rendono più efficaci le tue attività per mitigare gli incidenti.
- Uno stesso playbook può essere utilizzato da diversi membri del team in modo da identificare la causa principale in modo coerente.
- Le cause principali note possono già disporre di runbook appositamente sviluppati, accelerando i tempi di ripristino.
- I playbook accelerano la collaborazione tra i membri del team.
- I team possono applicare i processi su vasta scala tramite i playbook ripetibili.

Livello di rischio associato se questa best practice non fosse adottata: medio

Guida all'implementazione

Il modo in cui crei e utilizzi i playbook dipende dalla maturità della tua organizzazione. Se non hai familiarità con il cloud, crea i playbook in formato testo in un repository per i documenti centrale. Man mano che l'organizzazione acquisisce maturità, i playbook possono diventare semi automatizzati tramite script scritti in linguaggi come Python. Questi script possono essere eseguiti all'interno di un notebook Jupyter per accelerare il rilevamento. Le organizzazioni avanzate dispongono di playbook completamente automatizzati per i problemi comuni che vengono risolti automaticamente con i runbook.

Inizia a creare i playbook elencando gli incidenti comuni che si verificano nel tuo carico di lavoro. Scegli i playbook per gli incidenti a basso rischio e in cui la causa principale è riconducibile a pochi problemi. Una volta creati i playbook per gli scenari più semplici, passa agli scenari a rischio più elevato o in cui la causa principale non è ancora nota.

I playbook in formato testo vengono automatizzati man mano che l'organizzazione acquisisce maturità. Utilizzando servizi come Automazione AWS Systems Manager, il testo normale può essere trasformato in automazioni che possono essere eseguite sul carico di lavoro per accelerare le analisi. Queste automazioni possono essere attivate in risposta agli eventi, riducendo il tempo medio per rilevare e risolvere gli incidenti.

I clienti possono utilizzare AWS Systems Manager Incident Manager per rispondere agli incidenti. Questo servizio fornisce un'unica interfaccia per valutare gli incidenti, informare gli stakeholder circa il rilevamento e la mitigazione e collaborare per tutta la durata dell'incidente. Utilizza Automazione AWS Systems Manager per accelerare il rilevamento e il ripristino.

Esempio del cliente

Si è verificato un incidente che ha avuto un impatto sulla produzione della società AnyCompany Retail. L'ingegnere di turno utilizza un playbook per analizzare il problema e man mano che esegue i passaggi, mantiene aggiornati gli stakeholder indicati nel playbook. L'ingegnere identifica la causa principale come una race condition di un servizio di back-end. Utilizzando un runbook, l'ingegnere riavvia il servizio e riporta quindi AnyCompany Retail online.

Passaggi dell'implementazione

Se non è già presente, è consigliabile creare un repository per i documenti con il controllo delle versioni per la libreria di playbook. Puoi creare i tuoi playbook utilizzando Markdown, che è compatibile con la maggior parte dei sistemi di automazione dei playbook. Se parti da zero, utilizza il seguente modello di playbook come esempio.

```
# Titolo del playbook ## Informazioni sul playbook | ID playbook | Descrizione
 | Strumenti utilizzati | Autorizzazioni speciali | Autore del playbook | Ultimo
 aggiornamento | POC di escalation | Stakeholder | Piano di comunicazione |
 |-------|-------|-------|-------|-------|-------|-------|-------|-------| | RUN001
 | A cosa serve questo playbook? Per quale incidente viene utilizzato? | Strumenti
 | Autorizzazioni | Il tuo nome | 21-09-2022 | Nome dell'escalation | Nome dello
 stakeholder | Come vengono comunicati gli aggiornamenti durante l'analisi? | ##
 Passaggi 1. Passaggio 1 2. Passaggio 2
```

1. Se non disponi di un repository o di un wiki per i documenti, crea nel sistema di controllo delle versioni in uso un nuovo repository con il controllo delle versioni per i tuoi playbook.
2. Identifica un problema comune che richieda un'analisi, vale a dire uno scenario in cui la causa principale è riconducibile a pochi problemi e la risoluzione è a basso rischio.
3. Utilizzando il modello Markdown, compila la sezione `Titolo del playbook` e i campi in `Informazioni sul playbook`.
4. Includi i passaggi per la risoluzione dei problemi. Illustra nel modo più chiaro possibile le azioni da eseguire o le aree da analizzare.
5. Chiedi a un membro del team di esaminare e convalidare il tuo playbook. Se manca un'informazione o è necessario un chiarimento, aggiorna il playbook.
6. Pubblica il tuo playbook nel repository per i documenti e informa il tuo team e tutti gli stakeholder.
7. Questa libreria diventerà sempre più ricca man mano che aggiungi altri playbook. Una volta che sono disponibili diversi playbook, inizia ad automatizzarli con strumenti come Automazione AWS Systems Manager per mantenere sincronizzati l'automazione e i playbook.

Livello di impegno per il piano di implementazione: basso. I playbook sono documenti di testo archiviati in una posizione centrale. Le organizzazioni che hanno acquisito maturità applicano l'automazione dei playbook.

Risorse

Best practice correlate:

- OPS02-BP02 Assegnazione di proprietari identificati a processi e procedure: i playbook devono avere un proprietario responsabile della manutenzione.
- OPS07-BP03 Utilizzo di runbook per eseguire le procedure: i runbook e i playbook sono praticamente simili con un'unica differenza, ovvero in un runbook è previsto un risultato desiderato. In molti casi, i runbook vengono utilizzati dopo che un playbook ha individuato la causa principale.
- OPS10-BP01 Utilizzo di un processo per la gestione di eventi, incidenti e problemi: i playbook costituiscono una best practice per la gestione di eventi, incidenti e problemi.
- OPS10-BP02 Definizione di un processo per ogni avviso: i runbook e i playbook devono essere utilizzati in risposta agli avvisi. Nel corso del tempo queste reazioni devono essere automatizzate.
- OPS11-BP04 Gestione delle conoscenze: la manutenzione dei playbook è un elemento chiave della gestione delle conoscenze.

Documenti correlati:

- Achieving Operational Excellence using automated playbook and runbook (Eccellenza operativa mediante playbook e runbook automatizzati)
- AWS Systems Manager: Utilizzo di runbook
- Use AWS Systems Manager Automation runbooks to resolve operational tasks (Utilizzo dei runbook di Automazione AWS Systems Manager per la risoluzione delle attività operative)

Video correlati:

- AWS re:Invent 2019: DIY guide to runbooks, incident reports, and incident response (SEC318-R1) (Guida fai da te per runbook, report e risposte relativi agli incidenti (SEC318-R1))
- AWS Systems Manager Incident Manager - AWS Virtual Workshops (Workshop virtuali AWS)
- Integrate Scripts into AWS Systems Manager (Integrazione di script in AWS Systems Manager)

Esempi correlati:

- AWS Customer Playbook Framework (Framework di playbook del cliente AWS)
- AWS Systems Manager: Automation walkthroughs (Procedure di automazione dettagliate)
- Building an AWS incident response runbook using Jupyter notebooks and CloudTrail Lake (Creazione di un runbook per le risposte agli incidenti AWS mediante notebook Jupyter e data lake CloudTrail)
- Rubix - A Python library for building runbooks in Jupyter Notebooks (Rubix - Libreria Python per la creazione di runbook in notebook Jupyter)
- Using Document Builder to create a custom runbook (Utilizzo di Document Builder per creare un runbook personalizzato)
- Well-Architected Labs: Automating operations with Playbooks and Runbooks (Automazione delle operazioni con playbook e runbook)
- Well-Architected Labs: Incident response playbook with Jupyter (Well-Architected Labs: playbook di risposta agli incidenti con Jupyter)

Servizi correlati:

- Automazione AWS Systems Manager
- AWS Systems Manager Incident Manager

OPS07-BP05 Adozione di decisioni informate per implementare sistemi e modifiche

Valuta la capacità del team di supportare il carico di lavoro e la conformità del carico di lavoro alla governance. Confronta questi aspetti con i vantaggi della distribuzione quando decidi se eseguire il passaggio di un sistema o di una modifica in produzione. Per prendere decisioni informate, tieni conto dei rischi e dei benefici.

Si definisce "pre-mortem" un esercizio in cui il team simula un errore per sviluppare strategie di mitigazione. Utilizza questo esercizio per prevedere errori e creare procedure ove opportuno. Quando apporti modifiche agli elenchi di controllo che utilizzi per valutare i tuoi carichi di lavoro, pianifica quello che farai con i sistemi live che non risultano più conformi.

Anti-pattern comuni:

- Decidere di distribuire un carico di lavoro senza comprendere i rischi di sicurezza presenti al suo interno.

- Decidere di distribuire un carico di lavoro senza capire se è conforme alla governance e agli standard.
- Decidere di distribuire un carico di lavoro senza capire se il team è in grado di supportarlo.
- Decidere di distribuire un carico di lavoro senza comprendere i vantaggi dell'organizzazione.

Vantaggi dell'adozione di questa best practice: Membri del team qualificati costituiscono un supporto efficace al carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Decisioni informate per implementare sistemi e modifiche: valuta le capacità del team di supportare il carico di lavoro e la relativa conformità alla governance. Confronta questi aspetti con i vantaggi della distribuzione quando decidi se eseguire il passaggio di un sistema o di una modifica in produzione. Per prendere decisioni informate, tieni conto dei rischi e dei vantaggi.

Operatività

Domande

- OPS 8 Come fai a comprendere lo stato del tuo carico di lavoro?
- OPS 9 Come fai a comprendere lo stato delle operazioni?
- OPS 10 In che modo gestisci gli eventi del carico di lavoro e delle operazioni?

OPS 8 Come fai a comprendere lo stato del tuo carico di lavoro?

Definisci, acquisisci e analizza i parametri del carico di lavoro per ottenere visibilità sugli eventi del carico di lavoro, in modo da intraprendere le azioni appropriate.

Best practice

- OPS08-BP01 Identificazione degli indicatori chiave di prestazione
- OPS08-BP02 Definizione dei parametri del carico di lavoro
- OPS08-BP03 Raccolta e analisi dei parametri del carico di lavoro
- OPS08-BP04 Definizione di baseline per i parametri del carico di lavoro
- OPS08-BP05 Studio dei modelli di attività previsti per il carico di lavoro

- OPS08-BP06 Attivazione di un avviso quando i risultati del carico di lavoro sono a rischio
- OPS08-BP07 Attivazione di un avviso quando vengono rilevate delle anomalie nel carico di lavoro
- OPS08-BP08 Convalida del raggiungimento dei risultati e dell'efficacia dei KPI e dei parametri

OPS08-BP01 Identificazione degli indicatori chiave di prestazione

Identifica gli indicatori chiave delle prestazioni (KPI) in base ai risultati aziendali desiderati (ad esempio, tasso di ordini, tasso di conservazione dei clienti e profitti rispetto alle spese operative) e ai risultati dei clienti (ad esempio, soddisfazione dei clienti). Valuta i KPI per determinare il successo del carico di lavoro.

Anti-pattern comuni:

- I dirigenti aziendali ti chiedono in che modo un carico di lavoro sia stato in grado di soddisfare le esigenze aziendali, ma non disponi di un quadro di riferimento per determinare il successo.
- Non sei in grado di stabilire se l'applicazione pronta all'uso commerciale utilizzata per la tua organizzazione è conveniente.

Vantaggi dell'adozione di questa best practice: Identificando indicatori chiave delle prestazioni, puoi ottenere risultati aziendali da utilizzare come test dello stato e del successo del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Identificazione degli indicatori chiave di prestazione: identifica gli indicatori chiave di prestazione (KPI) basati su risultati attesi in termini di business e clienti. Valuta i KPI per determinare il successo del carico di lavoro.

OPS08-BP02 Definizione dei parametri del carico di lavoro

Definisci i parametri del carico di lavoro per misurare il raggiungimento dei KPI (ad esempio, carrelli degli acquisti abbandonati, ordini effettuati, costo, prezzo e spesa allocata per il carico di lavoro). Definisci i parametri del carico di lavoro per misurarne lo stato (ad esempio, tempo di risposta dell'interfaccia, percentuale di errori, richieste effettuate, richieste completate e utilizzo). Valuta i parametri per stabilire se il carico di lavoro raggiunge i risultati previsti e per comprendere lo stato del carico di lavoro.

È consigliabile inviare i dati di log a un servizio come CloudWatch Logs e generare parametri a partire dalle osservazioni dei contenuti di log necessari.

CloudWatch dispone di funzionalità specializzate quali Amazon CloudWatch Insights per .NET e SQL Server e Container Insights che possono aiutarti a identificare e configurare parametri chiave, log e allarmi per le risorse applicative e lo stack tecnologico specificamente supportati.

Anti-pattern comuni:

- Hai definito parametri standard, non associati ad alcun KPI o personalizzati per qualsiasi carico di lavoro.
- Si verificano errori nei calcoli dei parametri che produrranno risultati non validi.
- Non sono stati definiti parametri per il carico di lavoro.
- Le misurazioni riguardano solo la disponibilità.

Vantaggi dell'adozione di questa best practice: Definendo e valutando i parametri del carico di lavoro, puoi determinarne lo stato e misurare i risultati aziendali ottenuti.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Definizione dei parametri del carico di lavoro: definisci i parametri del carico di lavoro per misurare il raggiungimento dei KPI. Definisci i parametri del carico di lavoro per misurare lo stato del carico di lavoro e dei singoli componenti. Valuta i parametri per stabilire se il carico di lavoro raggiunge i risultati previsti e per comprendere lo stato del carico di lavoro.
  - Pubblicazione di parametri personalizzati
  - Ricerca e filtraggio dei dati di log
  - Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch

Risorse

Documenti correlati:

- Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch
- Pubblicazione di parametri personalizzati
- Ricerca e filtraggio dei dati di log

OPS08-BP03 Raccolta e analisi dei parametri del carico di lavoro

Esegui revisioni proattive regolari dei parametri per identificare le tendenze e stabilire dove sono necessarie risposte adeguate.

Devi aggregare i dati di log della tua applicazione, dei componenti del carico di lavoro, dei servizi e delle chiamate API in un servizio come CloudWatch Logs. Genera parametri dalle osservazioni dei contenuti di log necessari per consentire analisi approfondite delle prestazioni delle attività operative.

In AWS puoi analizzare i parametri del carico di lavoro e identificare i problemi operativi grazie alla funzionalità di machine learning di Amazon DevOps Guru. AWS DevOps Guru fornisce notifiche relative ai problemi operativi, con consigli mirati e proattivi per risolvere tali problemi e mantenere integra l'applicazione.

Nel modello di responsabilità condivisa AWS, le parti relative al monitoraggio vengono passate a te attraverso AWS Health Dashboard. Questo pannello di controllo fornisce avvisi e indicazioni per la correzione dei problemi quando in AWS si verificano eventi che possono avere ripercussioni su di te. I clienti iscritti al supporto Business ed Enterprise hanno a loro volta accesso all' API AWS Health, il che consente loro di integrare i propri sistemi di gestione degli eventi.

In AWS è possibile esporta i dati di log in Amazon S3 oppure inviare log direttamente a Amazon S3 per lo storage a lungo termine. Utilizzando AWS Glue, è possibile individuare e preparare i dati di log in Amazon S3 per l'analisi, archiviando i metadati associati nel AWSAWS Glue Data Catalog. Amazon Athena, grazie all'integrazione nativa con AWS Glue, quindi, può essere utilizzato per analizzare i dati di log, eseguendo query tramite SQL standard. Utilizzando uno strumento di business intelligence come Amazon QuickSight puoi visualizzare, esplorare e analizzare i tuoi dati.

Una soluzione alternativa sarebbe quella di utilizzare Amazon OpenSearch Service e i pannelli di controllo di OpenSearch per raccogliere, analizzare e visualizzare i log su AWS per più account e Regioni AWS.

Anti-pattern comuni:

- Il team di progettazione della rete ti chiede le tariffe correnti di utilizzo della larghezza di banda della rete. Secondo i parametri correnti, l'utilizzo della rete è al 35%. Come misura di risparmio sui costi, viene ridotta la capacità del circuito causando problemi di connettività diffusi, poiché la misurazione effettuata in un determinato momento non rifletteva l'andamento dei tassi di utilizzo.
- Il router ha generato errori. Ha registrato errori di memoria non critici con frequenza maggiore fino al completamento dell'errore. Non hai rilevato questo andamento e di conseguenza non hai sostituito la memoria difettosa prima che il router causasse un'interruzione del servizio.

Vantaggi dell'adozione di questa best practice: Raccogliendo e analizzando i parametri del carico di lavoro, puoi comprenderne lo stato e ottenere informazioni sulle tendenze che possono avere un impatto di esso o sul raggiungimento dei risultati aziendali.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Raccolta e analisi dei parametri del carico di lavoro: esegui regolarmente revisioni proattive dei parametri per identificare le tendenze e stabilire dove occorrono risposte adeguate.
  - Utilizzare i parametri Amazon CloudWatch
  - Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch
  - Raccolta di parametri e registri da istanze Amazon EC2 e da server on-premise con l'agente di CloudWatch

Risorse

Documenti correlati:

- Amazon Athena
- Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch
- Amazon DevOps Guru
- AWS Glue
- AWSAWS Glue Data Catalog
- Amazon OpenSearch Service
- AWS Health Dashboard
- Amazon QuickSight
- Raccolta di parametri e registri da istanze Amazon EC2 e da server on-premise con l'agente di CloudWatch
- Utilizzare i parametri Amazon CloudWatch

OPS08-BP04 Definizione di baseline per i parametri del carico di lavoro

Definisci le baseline per i parametri in modo da fornire i valori previsti di base per il confronto e l'identificazione dei componenti con prestazioni basse o alte. Identifica le soglie di miglioramento, verifica e intervento.

Anti-pattern comuni:

- Un server è in esecuzione al 95% di utilizzo della CPU e ti viene chiesto se ciò è positivo o negativo. L'utilizzo della CPU su tale server non segue la baseline, quindi non hai idea se ciò sia positivo o negativo.

Vantaggi dell'adozione di questa best practice: Definendo i valori dei parametri di base, è possibile valutare i valori dei parametri correnti e le tendenze dei parametri per determinare se è necessaria un'azione.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Definizione di baseline per i parametri del carico di lavoro: definisci le baseline per i parametri del carico di lavoro in modo da fornire i valori previsti di base per il confronto.
  - Creazione di allarmi Amazon CloudWatch

Risorse

Documenti correlati:

- Creazione di allarmi Amazon CloudWatch

OPS08-BP05 Studio dei modelli di attività previsti per il carico di lavoro

Definisci modelli di attività del carico di lavoro per identificare comportamenti anomali in modo da rispondere in modo appropriato, se necessario.

CloudWatch, tramite la funzionalità CloudWatch Anomaly Detection , applica algoritmi statistici e di machine learning per generare una gamma di valori previsti che rappresentano il normale comportamento dei parametri.

Amazon DevOps Guru può servire per identificare comportamenti anomali tramite la correlazione di eventi, l'analisi dei registri e l'applicazione del machine learning per analizzare la telemetria del carico di lavoro. Se vengono rilevati comportamenti inattesi, fornisce i parametri e gli eventi correlati con consigli per affrontare il comportamento.

Anti-pattern comuni:

- Stai esaminando i log di utilizzo della rete e vedi che questo è aumentato tra le 11:30 e le 13:30 e poi di nuovo alle 16:30 alle 18:00. Non sai se questo deve essere considerato normale o meno.
- I server Web si riavviano ogni sera alle 03:00. Non sai se questo è un comportamento previsto.

Vantaggi dell'adozione di questa best practice: Acquisendo modelli di comportamento, puoi riconoscere comportamenti imprevisti e intervenire, se necessario.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Studio di modelli di attività previsti per il carico di lavoro: definisci modelli di attività del carico di lavoro per stabilire quando un comportamento non rientra nei valori previsti, in modo da poter rispondere adeguatamente se necessario.

Risorse

Documenti correlati:

- Amazon DevOps Guru
- CloudWatch Anomaly Detection

OPS08-BP06 Attivazione di un avviso quando i risultati del carico di lavoro sono a rischio

Attiva un avviso quando i risultati del carico di lavoro sono a rischio, in modo da poter rispondere adeguatamente, se necessario.

Idealmente, in precedenza hai identificato una soglia di parametro raggiunta la quale si deve attivare l'allarme, oppure un evento che puoi utilizzare per attivare una risposta automatizzata.

In AWS, è possibile utilizzare Amazon CloudWatch Synthetics per creare script canary e monitorare endpoint e API, eseguendo le stesse azioni intraprese dai clienti. Grazie alla telemetria generata e alle informazioni ottenute identifichi i problemi prima che interessino i clienti.

Puoi anche utilizzare CloudWatch Logs Insights per cercare e analizzare in modo interattivo i dati di log utilizzando un linguaggio di query appositamente creato. Gli approfondimenti CloudWatch Logs rilevano automaticamente i campi nei log dei servizi AWS e i registri eventi personalizzati in JSON. Si dimensiona in base al volume di log e alla complessità delle query e fornisce risposte in pochi secondi, aiutandoti a individuare i fattori associati all'incidente.

Anti-pattern comuni:

- Non disponi della connettività di rete. Nessuno lo sa. Nessuno sta cercando di identificare il motivo o intraprendere un'azione per ripristinare la connettività.
- Dopo una patch, le istanze persistenti non sono disponibili, creando disagi agli utenti. I tuoi utenti hanno aperto casi di supporto. Nessuno ha ricevuto notifiche. Nessuno sta intervenendo.

Vantaggi dell'adozione di questa best practice: Riconoscendo lo stato di rischio dei risultati aziendali e attivando avvisi sulla necessità di intervento, hai l'opportunità di prevenire o mitigare l'impatto di un incidente.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Attivazione di un avviso quando i risultati del carico di lavoro sono a rischio: invia un avviso quando i risultati del carico di lavoro sono a rischio, in modo da rispondere in modo appropriato, se necessario.
  - Che cos'è Amazon CloudWatch Events?
  - Creazione di allarmi Amazon CloudWatch
  - Richiamo di funzioni Lambda utilizzando le notifiche di Amazon SNS

Risorse

Documenti correlati:

- Amazon CloudWatch Synthetics
- CloudWatch Logs Insights
- Creazione di allarmi Amazon CloudWatch
- Richiamo di funzioni Lambda utilizzando le notifiche di Amazon SNS
- Che cos'è Amazon CloudWatch Events?

OPS08-BP07 Attivazione di un avviso quando vengono rilevate delle anomalie nel carico di lavoro

Attiva un avviso quando vengono rilevate delle anomalie nel carico di lavoro, in modo da poter rispondere adeguatamente, se necessario.

L'analisi dei parametri del carico di lavoro nel corso del tempo potrebbe stabilire modelli di comportamento che puoi quantificare in modo sufficiente per definire un evento o attivare un allarme in risposta.

Una volta addestrata, la funzionalità CloudWatch Anomaly Detection può essere utilizzata per inviare allarmi in caso di anomalie rilevate o può fornire i valori previsti sovrapposti su un grafico di dati dei parametri per un confronto continuo.

Anti-pattern comuni:

- Le vendite del sito Web di vendita al dettaglio sono aumentate improvvisamente e in modo significativo. Nessuno lo sa. Nessuno sta cercando di identificare ciò che ha portato a questo picco. Nessuno interviene per garantire ai clienti un'esperienza di qualità sotto il carico aggiuntivo.
- Dopo l'applicazione di una patch, i tuoi server persistenti si riavviano spesso creando disagi gli utenti. In genere i server si riavviano al massimo fino a tre volte. Nessuno lo sa. Nessuno sta cercando di identificare il motivo per cui ciò si verifica.

Vantaggi dell'adozione di questa best practice: Comprendendo i modelli di comportamento del carico di lavoro, puoi identificare comportamenti imprevisti e intervenire, se necessario.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Attivazione di un avviso in caso di rilevamento di anomalie: invia un avviso quando vengono rilevate anomalie del carico di lavoro, in modo da rispondere in modo appropriato, se necessario.
  - Che cos'è Amazon CloudWatch Events?
  - Creazione di allarmi Amazon CloudWatch
  - Richiamo di funzioni Lambda utilizzando le notifiche di Amazon SNS

Risorse

Documenti correlati:

- Creazione di allarmi Amazon CloudWatch
- CloudWatch Anomaly Detection
- Richiamo di funzioni Lambda utilizzando le notifiche di Amazon SNS
- Che cos'è Amazon CloudWatch Events?

OPS08-BP08 Convalida del raggiungimento dei risultati e dell'efficacia dei KPI e dei parametri

Crea una vista a livello di business delle tue operazioni del carico di lavoro, per stabilire se le esigenze sono soddisfatte e per identificare gli aspetti da migliorare per raggiungere gli obiettivi di business. Convalida l'efficacia dei KPI e dei parametri e rivedili, se necessario.

AWS, inoltre, fornisce supporto per strumenti di business intelligence e sistemi per l'analisi di registri di terze parti attraverso API e SDK del servizio AWS (ad esempio Grafana, Kibana e Logstash).

Anti-pattern comuni:

- Il tempo di risposta della pagina non è mai stato considerato determinante per la soddisfazione dei clienti. Non hai mai stabilito un parametro o una soglia per il tempo di risposta della pagina. I tuoi clienti si lamentano della lentezza.
- Non hai raggiunto i tuoi obiettivi minimi in termini di tempo di risposta. Nel tentativo di migliorare il tempo di risposta, hai ampliato i server delle applicazioni. Ora stai superando gli obiettivi di tempo di risposta con un margine significativo e disponi anche di una notevole capacità inutilizzata che stai pagando.

Vantaggi dell'adozione di questa best practice: Esaminando e rivedendo i KPI e i parametri, puoi capire in che modo il carico di lavoro supporta il raggiungimento dei risultati aziendali e identificare i punti di miglioramento per ottenerli.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Convalida del raggiungimento dei risultati e dell'efficacia dei KPI e dei parametri: crea una panoramica a livello aziendale delle operazioni dei carichi di lavoro per stabilire se le esigenze sono soddisfatte e identificare le aree migliorabili per raggiungere gli obiettivi aziendali. Convalida l'efficacia dei KPI e dei parametri e rivedili, se necessario.
  - Utilizzo dei pannelli di controllo Amazon CloudWatch
  - Che cos'è l'analisi dei log?

Risorse

Documenti correlati:

- Utilizzo dei pannelli di controllo Amazon CloudWatch

- Che cos'è l'analisi dei log?

OPS 9 Come fai a comprendere lo stato delle operazioni?

Definisci, acquisisci e analizza i parametri delle operazioni per ottenere visibilità sugli eventi delle operazioni, in modo da intraprendere le azioni appropriate.

Best practice

- OPS09-BP01 Identificazione degli indicatori chiave di prestazione
- OPS09-BP02 Definizione dei parametri delle operazioni
- OPS09-BP03 Raccolta e analisi dei parametri delle operazioni
- OPS09-BP04 Definizione delle baseline per i parametri delle operazioni
- OPS09-BP05 Acquisizione dei modelli di attività previsti per le operazioni
- OPS09-BP06 Attivazione di un avviso quando i risultati delle operazioni sono a rischio
- OPS09-BP07 Attivazione di un avviso quando vengono rilevate delle anomalie nelle operazioni
- OPS09-BP08 Convalida del raggiungimento dei risultati e dell'efficacia dei KPI e dei parametri

OPS09-BP01 Identificazione degli indicatori chiave di prestazione

Identifica gli indicatori chiave di prestazione (KPI) in base all'obiettivo desiderato (ad esempio, fornitura di nuove caratteristiche) e ai risultati dei clienti (ad esempio, casi del servizio clienti). Valuta i KPI per determinare il successo delle operazioni.

Anti-pattern comuni:

- I dirigenti aziendali ti chiedono in che si raggiungono gli obiettivi aziendali con operazioni di successo, ma non disponi di un quadro di riferimento per determinare il successo.
- Non sei in grado di determinare se le finestre di manutenzione hanno un impatto sui risultati aziendali.

Vantaggi dell'adozione di questa best practice: Identificando indicatori chiave delle prestazioni, puoi ottenere risultati aziendali da utilizzare come test dello stato e del successo delle tue operazioni.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Identificazione degli indicatori chiave di prestazione: identifica gli indicatori chiave di prestazione (KPI) basati su risultati attesi in termini di business e clienti. Valuta i KPI per determinare il successo delle operazioni.

OPS09-BP02 Definizione dei parametri delle operazioni

Definisci i parametri delle operazioni per misurare il raggiungimento dei KPI (ad esempio, distribuzioni riuscite e distribuzioni non riuscite). Definisci i parametri delle operazioni per misurare lo stato delle attività operative (ad esempio, tempo medio per rilevare un incidente (MTTD) e tempo medio per il ripristino (MTTR) in seguito a un incidente). Valuta i parametri per stabilire se le operazioni raggiungono i risultati previsti e per comprendere lo stato delle loro attività.

Anti-pattern comuni:

- I parametri delle operazioni sono basati su ciò che il team ritiene ragionevole.
- Si verificano errori nei calcoli dei parametri che produrranno risultati non validi.
- Non disponi di parametri definiti per le attività operative.

Vantaggi dell'adozione di questa best practice: Definendo e valutando i parametri delle operazioni, puoi determinare lo stato delle tue attività operative e misurare i risultati aziendali ottenuti.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Definizione dei parametri delle operazioni: definisci i parametri delle operazioni per misurare il raggiungimento dei KPI. Definisci i parametri delle operazioni per misurare lo stato delle operazioni e delle relative attività. Valuta i parametri per stabilire se le operazioni raggiungono i risultati previsti e comprendere lo stato delle operazioni.
  - Pubblicazione di parametri personalizzati
  - Ricerca e filtraggio dei dati di log
  - Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch

Risorse

Documenti correlati:

- AWS Answers: Centralized Logging (AWS Answers: registrazione centralizzata)
- Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch
- Rilevare e reagire a modifiche dello stato della pipeline con Amazon CloudWatch Events
- Pubblicazione di parametri personalizzati
- Ricerca e filtraggio dei dati di log

Video correlati:

- Creazione di un piano di monitoraggio

OPS09-BP03 Raccolta e analisi dei parametri delle operazioni

Esegui regolarmente revisioni proattive dei parametri per identificare le tendenze e stabilire dove sono necessarie risposte adeguate.

È consigliabile aggregare i dati di registro relativi all'esecuzione delle attività e delle chiamate API operative in un servizio come CloudWatch Logs. Genera parametri dalle osservazioni dei contenuti di log necessari per ottenere informazioni dettagliate sulle prestazioni delle attività operative.

In AWS è possibile esporta i dati di log in Amazon S3 oppure inviare log direttamente a Amazon S3 per lo storage a lungo termine. Utilizzando AWS Glue, è possibile individuare e preparare i dati di log in Amazon S3 per l'analisi, archiviando i metadati associati nel AWSAWS Glue Data Catalog. Amazon Athena, grazie all'integrazione nativa con AWS Glue, quindi, può essere utilizzato per analizzare i dati di log, eseguendo query tramite SQL standard. Utilizzando uno strumento di business intelligence come Amazon QuickSight puoi visualizzare, esplorare e analizzare i tuoi dati.

Anti-pattern comuni:

- La distribuzione coerente di nuove funzionalità è considerata un indicatore chiave delle prestazioni. Non disponi di un metodo per misurare la frequenza delle distribuzioni.
- Registri le distribuzioni, le distribuzioni sottoposte a rollback, le patch e le patch di rollback per tenere traccia delle attività operative, ma nessuno esamina i parametri.
- L'obiettivo RTO (Recovery Time Objective) per ripristinare il database perso è di al massimo 15 minuti, definiti quando il sistema è stato distribuito ed era privo di utenti. Ora hai 10.000 utenti e sei in attività da due anni. Un ripristino recente ha richiesto più di due ore. Questo non è stato registrato e nessuno lo sa.

Vantaggi dell'adozione di questa best practice: Raccogliendo e analizzando i parametri delle operazioni, puoi comprenderne lo stato e ottenere informazioni sulle tendenze che possono avere un impatto di esse o sul raggiungimento dei risultati aziendali.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Raccolta e analisi dei parametri delle operazioni: esegui regolarmente revisioni proattive dei parametri per identificare le tendenze e determinare dove occorrono risposte adeguate.
  - Utilizzare i parametri Amazon CloudWatch
  - Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch
  - Raccolta di parametri e registri da istanze Amazon EC2 e da server on-premise con l'agente di CloudWatch

Risorse

Documenti correlati:

- Amazon Athena
- Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch
- Amazon QuickSight
- AWS Glue
- AWSAWS Glue Data Catalog
- Raccolta di parametri e registri da istanze Amazon EC2 e da server on-premise con l'agente di CloudWatch
- Utilizzare i parametri Amazon CloudWatch

OPS09-BP04 Definizione delle baseline per i parametri delle operazioni

Definisci le baseline per i parametri in modo da fornire i valori previsti di base per il confronto e l'identificazione delle attività operative con prestazioni basse e alte.

Anti-pattern comuni:

- Ti è stato chiesto di indicare il tempo previsto per la distribuzione. Non hai misurato il tempo necessario per la distribuzione e non sei in grado di determinare i tempi previsti.

- Ti è stato chiesto di indicare il tempo necessario per risolvere un problema con i server applicazioni. Non disponi di informazioni sul tempo necessario per il ripristino dal primo contatto del cliente. Non disponi di informazioni sul tempo necessario per il ripristino dalla prima identificazione di un problema tramite il monitoraggio.
- Ti è stato chiesto il numero del personale di supporto necessario durante il fine settimana. Non hai idea del numero di casi di supporto tipici durante un fine settimana e non puoi fornire una stima.
- L'obiettivo del tempo di ripristino per recuperare i database persi è di al massimo 15 minuti, definiti quando il sistema è stato implementato ed era privo di utenti. Ora hai 10.000 utenti e sei in attività da due anni. Non disponi di alcuna informazione su come il tempo di ripristino è cambiato per il database.

Vantaggi dell'adozione di questa best practice: Definendo i valori dei parametri di base, è possibile valutare i valori dei parametri correnti e le tendenze dei parametri per determinare se è necessaria un'azione.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Studio di modelli di attività previsti per le operazioni: definisci modelli di attività operative per stabilire quando un comportamento non rientra nei valori previsti, in modo da poter rispondere adeguatamente se necessario.

OPS09-BP05 Acquisizione dei modelli di attività previsti per le operazioni

Definisci modelli di attività operative per identificare comportamenti anomali in modo da rispondere in modo appropriato, se necessario.

Anti-pattern comuni:

- La percentuale di errori di distribuzione è aumentata sostanzialmente di recente. Gli errori vengono risolti in modo indipendente. Non ti accorgi che gli errori corrispondono alle distribuzioni di un nuovo dipendente che non ha familiarità con il sistema di gestione della distribuzione.

Vantaggi dell'adozione di questa best practice: Studiando i modelli di comportamento, puoi riconoscere comportamenti imprevisti e intervenire, se necessario.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Studio di modelli di attività previsti per le operazioni: definisci modelli di attività operative per stabilire quando un comportamento non rientra nei valori previsti, in modo da poter rispondere adeguatamente se necessario.

OPS09-BP06 Attivazione di un avviso quando i risultati delle operazioni sono a rischio

Ogni volta che i risultati delle operazioni sono a rischio, è necessario attivare un avviso e determinare le azioni da intraprendere. I risultati delle operazioni sono costituiti da qualsiasi attività che supporta un carico di lavoro in produzione. Sono incluse tutte le operazioni, dall'implementazione di nuove versioni delle applicazioni al ripristino da interruzione. I risultati delle operazioni devono essere trattati con la stessa importanza dei risultati aziendali.

I team del software devono identificare i parametri e le attività delle operazioni chiave e creare i relativi avvisi. Gli avvisi devono essere tempestivi e fruibili. Se viene generato un avviso, è necessario includere un riferimento a un runbook o un playbook corrispondente. Gli avvisi senza un'azione corrispondente possono portare al cosiddetto affaticamento dagli avvisi ("alert fatigue").

Risultato desiderato: quando le attività operative sono a rischio, vengono inviati avvisi per individuare l'azione da intraprendere. Gli avvisi spiegano il motivo per cui sono stati generati e includono il riferimento a un playbook per analizzare o a un runbook per mitigare. Ove possibile, i runbook vengono automatizzati e vengono inviate le notifiche.

Anti-pattern comuni:

- Si analizza un incidente e vengono compilati i casi di supporto. I casi di supporto stanno violando l'Accordo sul livello di servizio (SLA) ma non vengono generati avvisi.
- Un'implementazione in produzione pianificata per mezzanotte è stata ritardata a causa di modifiche del codice dell'ultimo minuto. Non viene generato alcun avviso e l'implementazione si blocca.
- Si verifica un'interruzione della produzione ma non vengono inviati avvisi.
- Il tempo di implementazione è costantemente al di sotto delle stime. Non viene intrapresa alcuna azione per analizzare.

Vantaggi dell'adozione di questa best practice:

- Gli avvisi per i risultati delle operazioni a rischio aumentano la tua capacità di supportare il carico di lavoro anticipando i problemi.

- I risultati aziendali sono migliorati grazie all'integrità delle operazioni.
- Il rilevamento e la risoluzione dei problemi operativi sono migliorati.
- L'integrità operativa complessiva è aumentata.

Livello di rischio associato se questa best practice non fosse adottata: medio

Guida all'implementazione

I risultati delle operazioni devono essere definiti prima di poter inviare gli avvisi. Inizia stabilendo quali attività operative sono più importanti per l'organizzazione: eseguire l'implementazione in produzione in meno di due ore o rispondere a una richiesta di supporto entro un determinato periodo di tempo? L'organizzazione deve definire le attività operative chiave e come vengono misurate in modo che possano essere monitorate, migliorate e segnalate. È necessaria una posizione centrale in cui archiviare e analizzare la telemetria del carico di lavoro e delle operazioni. Lo stesso meccanismo deve essere in grado di attivare un avviso quando l'esito di un'operazione è a rischio.

Esempio del cliente

È stato attivato un allarme CloudWatch durante un'implementazione di routine presso AnyCompany Retail. Il lead time per l'implementazione è stato violato. Amazon EventBridge ha creato un OpsItem in AWS Systems Manager OpsCenter. Il team delle operazioni cloud utilizza un playbook per analizzare il problema e nota che una modifica dello schema richiede più tempo del previsto. Avvisa lo sviluppatore di turno e continua a monitorare l'implementazione. Una volta completata l'implementazione, il team delle operazioni cloud risolve OpsItem. Il team esamina l'incidente per l'analisi dopo il completamento.

Passaggi dell'implementazione

1. Se non hai identificato KPI, parametri e attività delle operazioni, lavora sull'implementazione delle best practice precedenti per questa domanda (da OPS09-BP01 a OPS09-BP05).
   - I clienti AWS Support con Supporto Enterprise possono richiedere il workshop sui KPI operativi al proprio Technical Account Manager (TAM). Questo workshop collaborativo ti aiuta a definire i KPI e i parametri delle operazioni allineati agli obiettivi di business, senza costi aggiuntivi. Contatta il Technical Account Manager per ulteriori informazioni.
2. Dopo aver stabilito le attività operative, i KPI e i parametri, configura gli avvisi nella piattaforma di osservabilità. Gli avvisi devono avere un'azione associata, come un playbook o un runbook. Gli avvisi senza un'azione devono essere evitati.

3. Occorre valutare nel tempo i parametri, i KPI e le attività delle operazioni per identificare le aree di miglioramento. Acquisisci i feedback in runbook e playbook dagli operatori per identificare le aree di miglioramento nella risposta agli avvisi.
4. Gli avvisi devono includere un meccanismo per contrassegnarli come falsi positivi che porta alla revisione delle soglie dei parametri.

Livello di impegno per il piano di implementazione: medio. Prima di implementare questa best practice, ne esistono diverse altre che devono essere applicate. Una volta identificate le attività operative e stabiliti i KPI operativi, è necessario definire gli avvisi.

Risorse

Best practice correlate:

- OPS02-BP03 Assegnazione di proprietari identificati alle operazioni che siano responsabili delle relative prestazioni: ogni attività e risultato dell'operazione deve avere un proprietario identificato che è il responsabile e deve essere avvisato quando i risultati sono a rischio.
- OPS03-BP02 Potere di intervento dei membri del team quando i risultati sono a rischio: quando vengono generati avvisi, il tuo team deve disporre dell'autorità per agire e risolvere il problema.
- OPS09-BP01 Identificazione degli indicatori chiave di prestazione: gli avvisi sui risultati delle operazioni iniziano con l'identificazione dei KPI delle operazioni.
- OPS09-BP02 Definizione dei parametri delle operazioni: adotta questa best practice prima di iniziare a generare avvisi.
- OPS09-BP03 Raccolta e analisi dei parametri delle operazioni: la raccolta centralizzata dei parametri delle operazioni è necessaria per creare gli avvisi.
- OPS09-BP04 Definizione delle baseline per i parametri delle operazioni: le linee di base dei parametri operativi offrono la possibilità di ottimizzare gli avvisi ed evitare il cosiddetto affaticamento dagli avvisi ("alert fatigue").
- OPS09-BP05 Acquisizione dei modelli di attività previsti per le operazioni: puoi migliorare la precisione dei tuoi avvisi comprendendo gli schemi di attività per gli eventi operativi.
- OPS09-BP08 Convalida del raggiungimento dei risultati e dell'efficacia dei KPI e dei parametri: valuta il raggiungimento dei risultati delle operazioni per assicurarti che i tuoi KPI e i tuoi parametri siano validi.
- OPS10-BP02 Definizione di un processo per ogni avviso: ogni avviso deve avere un runbook o un playbook associato e fornire il contesto per la persona che viene avvisata.

- OPS11-BP02 Esecuzione di analisi post-incidente: conduci un'analisi post-incidente dopo l'avviso per identificare le aree di miglioramento.

Documenti correlati:

- AWS Deployment Pipelines Reference Architecture: Application Pipeline Architecture (Architettura di riferimento per l'implementazione di pipeline AWS: architettura della pipeline dell'applicazione)
- GitLab: Getting Started with Agile / DevOps Metrics (GitLab: Introduzione ai parametri Agile/DevOps)

Video correlati:

- Aggregate and Resolve Operational Issues Using AWS Systems Manager OpsCenter (Aggregazione e risoluzione dei problemi operativi utilizzando AWS Systems Manager OpsCenter)
- Integrate AWS Systems Manager OpsCenter with Amazon CloudWatch Alarms (Integrazione degli allarmi Amazon CloudWatch in AWS Systems Manager OpsCenter)
- Integrate Your Data Sources into AWS Systems Manager OpsCenter Using Amazon EventBridge (Integrazione delle origini dei dati in AWS Systems Manager OpsCenter utilizzando Amazon EventBridge)

Esempi correlati:

- Automate remediation actions for Amazon EC2 notifications and beyond using Amazon EC2 Systems Manager Automation and AWS Health (Automatizzazione delle azioni di correzioni per le notifiche Amazon Amazon EC2 e oltre utilizzando Automazione Amazon Amazon EC2 Systems Manager e AWS Health)
- AWS Management and Governance Tools Workshop - Operations 2022 (Workshop sugli strumenti di gestione e governanceAWS - Operazioni)
- Ingesting, analyzing, and visualizing metrics with DevOps Monitoring Dashboard on AWS (Inserimento, analisi e visualizzazione di parametri con il dashboard di monitoraggio DevOps in AWS)

Servizi correlati:

- Amazon EventBridge

- AWS Support Proactive Services - Operations KPI Workshop (Servizi proattivi di Supporto AWS - Workshop sui KPI operativi)
- AWS Systems Manager OpsCenter
- CloudWatch Eventi

OPS09-BP07 Attivazione di un avviso quando vengono rilevate delle anomalie nelle operazioni

Attiva un avviso quando vengono rilevate delle anomalie nelle operazioni in modo da poter rispondere adeguatamente, se necessario.

L'analisi dei parametri delle operazioni nel corso del tempo potrebbe stabilire modelli di comportamento che puoi quantificare in modo sufficiente per definire un evento o attivare un allarme in risposta.

Una volta addestrata, la funzionalità CloudWatch Anomaly Detection può essere utilizzata per inviare allarmi in caso di anomalie rilevate o può fornire i valori previsti sovrapposti su un grafico di dati dei parametri per un confronto continuo.

Amazon DevOps Guru può servire per identificare comportamenti anomali tramite la correlazione di eventi, l'analisi dei registri e l'applicazione del machine learning per analizzare la telemetria del carico di lavoro. Al informazioni ottenute vengono presentate con i dati e i consigli pertinenti.

Anti-pattern comuni:

- Stai applicando una patch al parco istanze. La patch è stata testata correttamente nell'ambiente di test. La patch ha esito negativo per una grande percentuale di istanze nel parco istanze. Non fai nulla.
- Noti che le distribuzioni sono disponibili a partire da venerdì a fine giornata. L'organizzazione ha finestre di manutenzione predefinite il martedì e il giovedì. Non fai nulla.

Vantaggi dell'adozione di questa best practice: Comprendendo i modelli di comportamento delle operazioni puoi identificare comportamenti imprevisti e intervenire, se necessario.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Attivazione di un avviso quando vengono rilevate anomalie nelle operazioni: invia un avviso quando vengono rilevate anomalie nelle operazioni, in modo da rispondere in modo appropriato, se necessario.
  - Che cos'è Amazon CloudWatch Events?
  - Creazione di allarmi Amazon CloudWatch
  - Richiamo di funzioni Lambda utilizzando le notifiche di Amazon SNS

Risorse

Documenti correlati:

- Amazon DevOps Guru
- CloudWatch Anomaly Detection
- Creazione di allarmi Amazon CloudWatch
- Rilevare e reagire a modifiche dello stato della pipeline con Amazon CloudWatch Events
- Richiamo di funzioni Lambda utilizzando le notifiche di Amazon SNS
- Che cos'è Amazon CloudWatch Events?

OPS09-BP08 Convalida del raggiungimento dei risultati e dell'efficacia dei KPI e dei parametri

Crea una vista a livello di business delle attività operative, per stabilire se le esigenze sono soddisfatte e per identificare gli aspetti da migliorare per raggiungere gli obiettivi di business. Convalida l'efficacia dei KPI e dei parametri e rivedili, se necessario.

AWS inoltre fornisce supporto per strumenti di business intelligence e sistemi per l'analisi di registri di terze parti attraverso API e SDK del servizio AWS (ad esempio Grafana, Kibana e Logstash).

Anti-pattern comuni:

- La frequenza delle distribuzioni è aumentata con l'aumento del numero di team di sviluppo. Il numero previsto di distribuzioni definito è una volta alla settimana. La distribuzione viene effettuata regolarmente ogni giorno. Quando è presente un problema con il sistema di distribuzione e le distribuzioni non sono possibili, tale problema non viene rilevato per giorni.
- Quando precedentemente la tua azienda ha fornito supporto solo durante le ore lavorative principali dal lunedì al venerdì. Per gli incidenti hai stabilito un obiettivo relativo al tempo di risposta

che corrisponde al giorno lavorativo successivo. Di recente hai iniziato a offrire una copertura di supporto 24x7 con un obiettivo di tempo di risposta di 2 ore. Il personale notturno è sovraccarico e i clienti non sono contenti. Non vi sono indicazioni in merito all'esistenza di problemi legati ai tempi di risposta agli incidenti perché i report si riferiscono a un obiettivo specificato come "giorno lavorativo successivo".

Vantaggi dell'adozione di questa best practice: Esaminando e rivedendo i KPI e i parametri, puoi capire in che modo il carico di lavoro supporta il raggiungimento dei risultati aziendali e puoi identificare i punti di miglioramento per ottenerli.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Convalida del raggiungimento dei risultati e dell'efficacia dei KPI e dei parametri: crea una panoramica a livello aziendale delle attività operative per determinare se le esigenze sono soddisfatte e per identificare le aree che devono essere migliorate per raggiungere gli obiettivi aziendali. Convalida l'efficacia dei KPI e dei parametri e rivedili, se necessario.
  - Utilizzo dei pannelli di controllo Amazon CloudWatch
  - Che cos'è l'analisi dei log?

Risorse

Documenti correlati:

- Utilizzo dei pannelli di controllo Amazon CloudWatch
- Che cos'è l'analisi dei log?

OPS 10 In che modo gestisci gli eventi del carico di lavoro e delle operazioni?

Prepara e convalida le procedure in risposta agli eventi per ridurre al minimo il loro impatto sul tuo carico di lavoro.

Best practice

- OPS10-BP01 Utilizzo di un processo per la gestione di eventi, incidenti e problemi
- OPS10-BP02 Definizione di un processo per ogni avviso
- OPS10-BP03 Definizione della priorità degli eventi operativi in base agli effetti sul business

- OPS10-BP04 Definizione dei percorsi di escalation
- OPS10-BP05 Abilitazione delle notifiche push
- OPS10-BP06 Comunicazione dello stato tramite pannelli di controllo
- OPS10-BP07 Automazione delle risposte agli eventi

OPS10-BP01 Utilizzo di un processo per la gestione di eventi, incidenti e problemi

L'organizzazione dispone di processi per gestire eventi, incidenti e problemi. Gli eventi sono costituiti da quanto accade nel carico di lavoro che non necessita di un intervento umano. Gli incidenti sono invece eventi che richiedono un intervento. I problemi sono eventi ricorrenti che richiedono un intervento o che non possono essere risolti. È necessario disporre di processi per ridurre l'impatto degli eventi sull'azienda e accertarsi di reagire in modo tempestivo e appropriato.

Quando nel carico di lavoro si verificano problemi o incidenti, è necessario utilizzare i processi per gestirli. In che modo puoi comunicare lo stato dell'evento alle parti coinvolte? Chi supervisiona la gestione delle risposte? Quali sono gli strumenti da utilizzare per ridurre l'impatto dell'evento? Questi sono solo alcuni esempi delle domande a cui devi rispondere per creare un processo di risposta affidabile.

I processi devono essere documentati in una posizione centralizzata, nonché essere disponibili a chiunque sia coinvolto nel carico di lavoro. Se non è presente un wiki o un archivio di documenti centralizzato, è possibile utilizzare un repository per il controllo delle versioni. In questo modo sarà possibile mantenere aggiornati i piani in modo conforme all'evoluzione dei processi.

I problemi possono essere automatizzati. Il tempo richiesto per la gestione di questo tipo di eventi potrebbe essere altrimenti destinato all'innovazione. Comincia a creare un processo ripetibile per ridurre il più possibile l'impatto del problema. Gradualmente cerca di concentrarti sull'automazione della riduzione o risoluzione del problema sottostante. In questo modo il tempo risparmiato potrà essere dedicato a migliorare il carico di lavoro.

Risultato desiderato: l'organizzazione dispone di un processo per gestire eventi, incidenti e problemi. Questi processi sono documentati e archiviati in una posizione centralizzata e vengono aggiornati in base alle modifiche apportate.

Anti-pattern comuni:

- Un incidente si verifica durante il fine settimana e il tecnico di turno non sa cosa fare.

- Un cliente invia un messaggio e-mail indicando che l'applicazione non è disponibile. Riavvii il server per correggere il problema. Questo incidente si verifica di frequente.
- Si verifica un incidente e più team si mettono a lavorare in modo indipendente per risolvere il problema.
- Le implementazioni vengono eseguite nel carico di lavoro senza essere documentate.

Vantaggi dell'adozione di questa best practice:

- Nel carico di lavoro è presente un itinerario di audit degli eventi.
- Viene ridotto il tempo necessario per il ripristino in seguito a un incidente.
- I membri dei team riescono a risolvere incidenti e problemi in modo coerente.
- Durante l'analisi di un incidente, l'approccio è condiviso e più consolidato.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

L'implementazione di questa best practice prevede la registrazione degli eventi dei carichi di lavoro. Per la gestione di incidenti e problemi, è necessario ricorrere ai processi. I processi sono documentati, condivisi e aggiornati con frequenza. I problemi vengono identificati, classificati in base alla priorità e corretti.

Esempio del cliente

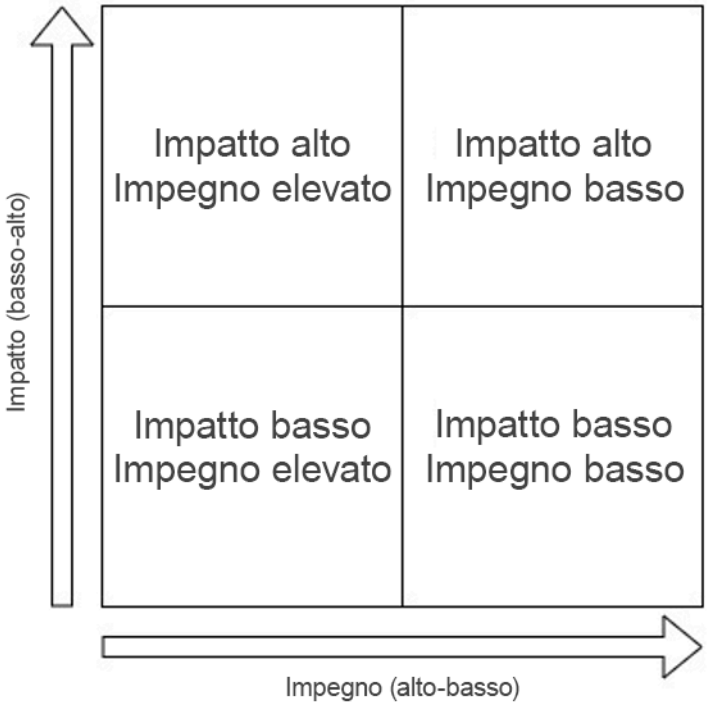
AnyCompany Retail ha dedicato una parte del proprio wiki interno ai processi destinati alla gestione di eventi, incidenti e problemi. Tutti gli eventi vengono inviati ad Amazon EventBridge. I problemi vengono classificati come OpsItems (elementi di lavoro operativi) in AWS Systems Manager OpsCenter e classificati in base alla loro priorità al fine della loro risoluzione, in modo da ridurre eventuali attività indifferenziate. Quando i processi subiscono variazioni, vengono aggiornati nel wiki interno. Viene utilizzato AWS Systems Manager Incident Manager per gestire gli incidenti e coordinare le attività di riduzione dell'impatto.

Passaggi dell'implementazione

1. Eventi
   - Tieni traccia degli eventi che si verificano nel carico di lavoro, anche se non è richiesto alcun intervento umano.

- Collabora con le parti coinvolte a livello di piano di lavoro per redigere un elenco di eventi di cui tenere traccia, ad esempio implementazioni completate o applicazioni di patch riuscite.
- Puoi utilizzare servizi come Amazon EventBridge oppure Amazon Simple Notification Service per generare eventi personalizzati per il monitoraggio.

2. Incidenti
   - Per prima cosa, definisci il piano di comunicazione per gli incidenti. Quali parti coinvolte devono essere informate? In che modo le tieni costantemente aggiornate? Chi supervisiona il coordinamento di tutte queste attività? È consigliabile creare un canale di chat per le comunicazioni e il coordinamento.
   - Definisci un percorso di escalation per i team di supporto del carico di lavoro, soprattutto se il team non dispone di turni di rotazione della disponibilità. A seconda del livello di supporto, è possibile segnalare un caso anche mediante il AWS Support.
   - Crea un playbook per l'analisi dell'incidente. È necessario includere il piano di comunicazione e, in dettaglio, i passaggi del processo di indagine. Includi il controllo del AWS Health Dashboard nel processo di indagine.
   - Documenta il piano di risposta agli incidenti. Comunica il piano di gestione degli incidenti in modo che i clienti esterni siano consapevoli delle regole da seguire e dei comportamenti richiesti previsti. Fornisci formazione ai membri dei team su come utilizzare tale piano di gestione.
   - I clienti possono utilizzare Incident Manager per configurare e gestire il piano di risposta agli incidenti.
   - I clienti del supporto Enterprise possono richiedere di seguire il workshop relativo alla gestione degli incidenti al proprio Technical Account Manager (TAM). Questo workshop guidato consente di verificare il piano di risposta agli incidenti esistente e ti aiuta a individuare eventuali aree da migliorare.
3. Problemi
   - I problemi devono essere identificati e registrati nel sistema ITSM in uso.
   - Identifica tutti i problemi noti ed eseguine una catalogazione in base all'impegno necessario per correggerli e al relativo impatto sul carico di lavoro.



- Per prima cosa risolvi i problemi caratterizzati dall'impatto più alto e dal minore impegno. Dopodiché, passa alla risoluzione dei problemi che rientrano nel quadrante basso impatto/basso impegno.
- Puoi utilizzare Systems Manager OpsCenter per identificare i problemi, associarvi runbook e tenerne traccia.

Livello di impegno per il piano di implementazione: medio. Devi disporre sia di un processo che degli strumenti per implementare questa best practice. Documenta i processi e rendili disponibili a chiunque sia coinvolto nel carico di lavoro. Aggiornali con frequenza. È disponibile un processo per la gestione e la migrazione o la risoluzione dei problemi.

Risorse

Best practice correlate:

- OPS07-BP03 Utilizzo di runbook per eseguire le procedure: i problemi noti necessitano di un runbook associato in modo tale che le attività di attenuazione dell'impatto siano coerenti.
- OPS07-BP04 Utilizzo dei playbook per analizzare i problemi: gli incidenti devono essere analizzati con il supporto di playbook.
- OPS11-BP02 Esecuzione di analisi post-incidente: esegui sempre un post-mortem dopo aver eseguito un ripristino in seguito a un incidente.

Documenti correlati:

- Atlassian - Incident management in the age of DevOps (Atlassian - Gestione degli incidenti nell'era di DevOps)
- AWS Security Incident Response Guide (Guida alle risposte agli incidenti di sicurezza di AWS)
- Incident Management in the Age of DevOps and SRE (Gestione degli incidenti nell'era di DevOps e SRE)
- PagerDuty - What is Incident Management? (PagerDuty - Che cos'è la gestione degli incidenti?)

Video correlati:

- AWS re:Invent 2020: Incident management in a distributed organization (Gestione degli incidenti in un'organizzazione distribuita)
- AWS re:Invent 2021 - Building next-gen applications with event-driven architectures (Sviluppo di applicazioni di nuova generazione con architetture basate su eventi)
- AWS Supports You | Exploring the Incident Management Tabletop Exercise (Esplorazione degli esercizi di simulazione relativi alla gestione degli incidenti)
- AWS Systems Manager Incident Manager - AWS Virtual Workshops (Workshop virtuali AWS)
- AWS What's Next ft. Incident Manager | AWS Events (Novità di AWS e Incident Manager | Eventi AWS)

Esempi correlati:

- AWS Management and Governance Tools Workshop - OpsCenter (Workshop sugli strumenti di gestione e governance AWS - OpsCenter)
- AWS Proactive Services – Incident Management Workshop (Servizi AWS proattivi – Workshop relativo alla gestione degli incidenti)
- Building an event-driven application with Amazon EventBridge (Sviluppo di un'applicazione basata su eventi con Amazon EventBridge)
- Building event-driven architectures on AWS (Sviluppo di architetture basate su eventi in AWS)

Servizi correlati:

- Amazon EventBridge
- Amazon SNS

- AWS Health Dashboard
- AWS Systems Manager Incident Manager
- AWS Systems Manager OpsCenter

OPS10-BP02 Definizione di un processo per ogni avviso

Predisponi una risposta specifica (runbook o playbook), con un proprietario espressamente identificato, per ogni evento per cui viene generato un avviso. Questo consente di rispondere agli eventi operativi in modo rapido ed efficace, evitando che gli eventi che richiedono un'azione vengano oscurati da notifiche meno importanti.

Anti-pattern comuni:

- Il sistema di monitoraggio presenta un flusso di connessioni approvate insieme ad altri messaggi. Il volume di messaggi è così grande che vengono ignorati dei messaggi di errore periodici che richiedono il tuo intervento.
- Ricevi un avviso che informa che il sito Web è inattivo. Non esiste un processo definito per quando ciò si verifica. Sei costretto ad adottare un approccio ad hoc per diagnosticare e risolvere il problema. Lo sviluppo di questo processo durante l'esecuzione prolunga il tempo di ripristino.

Vantaggi dell'adozione di questa best practice: Generando avvisi solo quando è necessaria un'operazione, eviti che gli avvisi di basso valore nascondano quelli più importanti. Creando un processo per ogni avviso che richiede un'azione, puoi attivare una risposta coerente e immediata agli eventi nel tuo ambiente.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Un processo per ogni avviso: a ogni evento per cui viene generato un avviso deve corrispondere una risposta specifica (runbook o playbook) con un responsabile specificatamente identificato (ad esempio, una persona, un team o un ruolo) a cui spetta il compito di completare correttamente l'azione. L'esecuzione della risposta può essere automatizzata o condotta da un altro team, ma il proprietario è tenuto ad assicurarsi che il processo produca i risultati previsti. Questi processi consentono di rispondere agli eventi operativi in modo rapido ed efficace, evitando che gli eventi che richiedono un'azione vengano oscurati da notifiche meno importanti. Ad esempio, è possibile applicare l'auto scaling per ridimensionare un front-end Web, ma il team operativo può essere

tenuto a garantire che le regole e i limiti di auto scaling siano appropriati per le esigenze del carico di lavoro.

Risorse

Documenti correlati:

- Funzionalità di Amazon CloudWatch
- Che cos'è Amazon CloudWatch Events?

Video correlati:

- Creazione di un piano di monitoraggio

OPS10-BP03 Definizione della priorità degli eventi operativi in base agli effetti sul business

Quando più eventi richiedono un intervento, assicurati che quelli più significativi per il business vengano affrontati per primi. Sono esempi di effetti il decesso o l'infortunio, le perdite finanziarie o i danni alla reputazione o alla fiducia.

Anti-pattern comuni:

- Ricevi una richiesta di supporto per aggiungere una configurazione della stampante per un utente. Durante la risoluzione del problema, ricevi una richiesta di supporto per sito di vendita al dettaglio non disponibile. Dopo aver completato la configurazione della stampante per l'utente, inizi a lavorare sul problema del sito Web.
- Ti viene segnalato che il sito Web di vendita al dettaglio e il sistema delle buste paga non sono disponibili. Non sai quale deve avere la priorità.

Vantaggi dell'adozione di questa best practice: Dare priorità alle risposte agli incidenti che determinano il maggiore impatto sull'azienda consente di gestire tale impatto.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Definizione della priorità degli eventi operativi in base agli effetti sul business: assicurati che quando più eventi richiedono un intervento, quelli più significativi per l'azienda vengano affrontati

per primi. Sono esempi di effetti il decesso o l'infortunio, le perdite finanziarie, le violazioni alle normative o i danni alla reputazione o alla fiducia.

OPS10-BP04 Definizione dei percorsi di escalation

Definisci percorsi di escalation nei tuoi runbook e playbook, compresi gli eventi che attivano l'escalation e le procedure di escalation. In particolare, identifica i proprietari per ogni azione per assicurare risposte rapide ed efficaci agli eventi operativi.

Stabilisci in quali circostanze serve una decisione umana prima che venga intrapresa un'azione. Collabora con i responsabili delle decisioni affinché questa decisione venga presa in anticipo e l'operazione sia preapprovata, in modo che la MTTR non si prolunghi in attesa di una risposta.

Anti-pattern comuni:

- Il sito di vendita al dettaglio non è disponibile. Il runbook per il ripristino del sito non è chiaramente comprensibile. Inizi a chiamare i colleghi sperando che qualcuno possa aiutarti.
- Ricevi un caso di supporto per un'applicazione irraggiungibile. Non disponi delle autorizzazioni per amministrare il sistema. Non sai a chi compete questo compito. Tenti di contattare il proprietario del sistema che ha aperto il caso ma non ricevi risposta. Né tu né i tuoi colleghi sapete chi bisogna contattare per il sistema.

Vantaggi dell'adozione di questa best practice: Definendo le escalation e i trigger e le procedure per l'escalation, abiliti l'aggiunta sistematica di risorse a un incidente con una rapidità adeguata all'impatto.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Definizione di percorsi di escalation: definisci percorsi di escalation nei tuoi runbook e playbook, compresi gli eventi che attivano l'escalation e le relative procedure. Ad esempio, l'escalation di un problema dai tecnici del supporto ai tecnici del supporto senior quando i runbook non riescono a risolvere il problema o quando è trascorso un determinato periodo di tempo. Un altro esempio di percorso di escalation appropriato è l'inoltro dai tecnici del supporto senior al team di sviluppo per un carico di lavoro quando i playbook non sono in grado di identificare un percorso di correzione o quando è trascorso un determinato periodo di tempo. In particolare, identifica i proprietari per ogni azione per assicurare risposte rapide ed efficaci agli eventi operativi. Le escalation possono

includere terze parti, ad esempio un provider di connettività di rete o un produttore di software. Possono anche includere i responsabili decisionali autorizzati identificati per i sistemi interessati.

OPS10-BP05 Abilitazione delle notifiche push

Informa direttamente gli utenti (ad esempio tramite e-mail o SMS) quando i servizi che utilizzano sono interessati e quando vengono ripristinate le normali condizioni operative, per consentire loro di adottare le misure appropriate.

Anti-pattern comuni:

- La tua applicazione sta sperimentando un Denial of Service distribuito e non risponde da giorni. Non è presente alcun messaggio di errore. Non hai inviato un'e-mail di notifica. Non hai inviato notifiche testuali. Non hai condiviso le informazioni sui social media. I clienti sono frustrati e sono alla ricerca di altri fornitori in grado di supportarli.
- Lunedì la tua applicazione ha avuto problemi dopo una patch ed è rimasta fuori uso per un paio d'ore. Martedì la tua applicazione ha avuto problemi dopo la distribuzione del codice e per un paio d'ore ha mostrato segni di instabilità. Mercoledì l'applicazione ha avuto problemi a seguito di una distribuzione del codice per mitigare una vulnerabilità di sicurezza associata alla patch non riuscita e non è stata disponibile per un paio d'ore. Giovedì i tuoi clienti, infastiditi, hanno iniziato a cercare un altro fornitore che potesse supportarli.
- Per il fine settimana avevi già previsto interventi di manutenzione che avrebbero reso irraggiungibile l'applicazione. Non ne dai comunicazione ai tuoi clienti. Alcuni dei tuoi clienti avevano pianificato attività che comportavano l'uso della tua applicazione. Sono molto frustrati dal fatto che la tua applicazione non è disponibile.

Vantaggi dell'adozione di questa best practice: Definendo le notifiche e i trigger e le procedure per le notifiche, i clienti vengono informati e possono rispondere nel caso siano interessati dai problemi riguardanti il carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Abilitazione delle notifiche push: informa direttamente gli utenti (ad esempio tramite e-mail o SMS) quando i servizi che utilizzano sono interessati e quando vengono ripristinate le normali condizioni operative, per consentire loro di adottare le misure appropriate.
  - Caratteristiche di Amazon SES

- Che cos'è Amazon SES?
- Impostazione delle notifiche Amazon SNS

Risorse

Documenti correlati:

- Caratteristiche di Amazon SES
- Impostazione delle notifiche Amazon SNS
- Che cos'è Amazon SES?

OPS10-BP06 Comunicazione dello stato tramite pannelli di controllo

Fornisci pannelli di controllo personalizzati in base ai destinatari, ad esempio i team tecnici interni, la dirigenza e i clienti, per comunicare lo stato operativo corrente del business e fornire i parametri desiderati.

Puoi creare pannelli di controllo utilizzando Amazon CloudWatch Dashboards sulle home page personalizzabili nella console di CloudWatch. Utilizzando servizi di business intelligence come Amazon QuickSight è possibile creare e pubblicare pannelli di controllo interattivi sullo stato del carico di lavoro e delle operazioni (ad esempio tassi di ordinazione, utenti connessi e tempi di transazione). Crea pannelli di controllo che mostrino visualizzazioni dei parametri a livello di sistema e a livello di azienda.

Anti-pattern comuni:

- Su richiesta, esegui un report sull'utilizzo corrente dell'applicazione per la gestione.
- Durante un incidente, vieni contattato ogni 20 minuti da un responsabile di sistema preoccupato, che desidera sapere se il problema è stato risolto.

Vantaggi dell'adozione di questa best practice: Creando pannelli di controllo, abiliti l'accesso self-service alle informazioni consentendo ai clienti di informarsi autonomamente e decidere se devono intervenire.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Comunicazione dello stato tramite pannelli di controllo: fornisci pannelli di controllo personalizzati in base ai destinatari, ad esempio i team tecnici, la leadership e i clienti, per comunicare l'attuale stato operativo del business e fornire i parametri rilevanti. Offrire un'opzione self-service per le informazioni di stato riduce le interruzioni derivanti dalla gestione delle richieste di stato da parte dei team operativi. Ne sono esempi i pannelli di controllo di Amazon CloudWatch e AWS Health Dashboard.
  - I pannelli di controllo di CloudWatch creano e utilizzano visualizzazioni dei parametri personalizzate

Risorse

Documenti correlati:

- Amazon QuickSight
- I pannelli di controllo di CloudWatch creano e utilizzano visualizzazioni dei parametri personalizzate

OPS10-BP07 Automazione delle risposte agli eventi

Automatizza le risposte agli eventi per ridurre gli errori causati dai processi manuali e assicurare risposte rapide e coerenti.

I modi per automatizzare le azioni di runbook o playbook su AWS sono molteplici. Per rispondere a un evento dovuto a una modifica dello stato nelle risorse AWS o a eventi personalizzati, è necessario creare regole CloudWatch Events per attivare risposte tramite destinazioni CloudWatch (ad esempio funzioni Lambda, argomenti Amazon Simple Notification Service (Amazon SNS), attività Amazon ECS e AWS Systems Manager Automation).

Per rispondere a un determinato parametro che supera una soglia per una certa risorsa (ad es. il tempo di attesa), è consigliabile creare avvisi CloudWatch per eseguire una o più azioni utilizzando azioni Amazon EC2 e azioni Auto Scaling o per inviare una notifica a un argomento Amazon SNS. Se è necessario eseguire azioni personalizzate in risposta a un avviso, richiama Lambda con una notifica Amazon SNS. Utilizza Amazon SNS per pubblicare notifiche di eventi e messaggi di escalation, in modo tale che le persone ne siano informate.

AWS supporta, inoltre, sistemi di terze parti attraverso API e SDK del servizio AWS. Esistono numerosi strumenti forniti da partner AWS e da terze parti che consentono di monitorare e inviare

notifiche e risposte. Alcuni di questi strumenti sono New Relic, Splunk, Loggly, SumoLogic e Datadog.

Rendi disponibili le procedure manuali cruciali in modo tale che possano essere utilizzate quando le procedure automatiche non riescono.

Anti-pattern comuni:

- Uno sviluppatore controlla il proprio codice. Questo evento avrebbe potuto essere utilizzato per avviare una compilazione e quindi eseguire il test, ma non accade nulla.
- L'applicazione registra un errore specifico prima di smettere di funzionare. La procedura per riavviare l'applicazione è ben nota e può essere creata con script. Puoi utilizzare l'evento di log per richiamare uno script e riavviare l'applicazione. Ricevi, invece, una chiamata alle 3 di domenica mattina, quando si verifica l'errore, perché sei reperibile come risorsa responsabile della correzione del sistema.

Vantaggi dell'adozione di questa best practice: Utilizzando le risposte automatizzate agli eventi, riduci il tempo necessario per rispondere e limiti l'introduzione di errori da attività manuali.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Automazione delle risposte agli eventi: automatizza le risposte agli eventi per ridurre gli errori causati dai processi manuali e per assicurare risposte rapide e coerenti.
  - Che cos'è Amazon CloudWatch Events?
  - Creazione di una regola di CloudWatch Events che si attiva al verificarsi di un evento
  - Creazione di una regola di CloudWatch Events che si attiva con una chiamata API AWS tramite AWS CloudTrail
  - Esempi di eventi CloudWatch Events dai servizi supportati

Risorse

Documenti correlati:

- Funzionalità di Amazon CloudWatch
- Esempi di eventi CloudWatch Events dai servizi supportati

- Creazione di una regola di CloudWatch Events che si attiva con una chiamata API AWS tramite AWS CloudTrail
- Creazione di una regola di CloudWatch Events che si attiva al verificarsi di un evento
- Che cos'è Amazon CloudWatch Events?

Video correlati:

- Creazione di un piano di monitoraggio

Esempi correlati:

Evoluzione

Domanda

- OPS 11 In che modo fai evolvere le operazioni?

OPS 11 In che modo fai evolvere le operazioni?

Dedica tempo e risorse al miglioramento incrementale continuo, per far evolvere l'efficacia e l'efficienza delle tue operazioni.

Best practice

- OPS11-BP01 Definizione di un processo per il miglioramento continuo
- OPS11-BP02 Esecuzione di analisi post-incidente
- OPS11-BP03 Implementazione di cicli di feedback
- OPS11-BP04 Gestione delle conoscenze
- OPS11-BP05 Definizione dei fattori che promuovono il miglioramento
- OPS11-BP06 Convalida delle informazioni
- OPS11-BP07 Revisione dei parametri delle operazioni
- OPS11-BP08 Documentazione e condivisione delle conoscenze acquisite
- OPS11-BP09 Allocazione di tempo per i miglioramenti

OPS11-BP01 Definizione di un processo per il miglioramento continuo

Valuta regolarmente e dai priorità alle opportunità di miglioramento per concentrare gli sforzi laddove è possibile ottenere i maggiori vantaggi.

Anti-pattern comuni:

- Hai documentato le procedure necessarie per creare un ambiente di sviluppo o di test. Puoi utilizzare CloudFormation per automatizzare il processo, ma decidi di farlo manualmente dalla console.
- I test mostrano che gran parte dell'utilizzo della CPU all'interno dell'applicazione riguarda un piccolo set di funzioni inefficienti. Potresti concentrarti sul miglioramento e sulla riduzione dei costi, ma sei stato incaricato di creare una nuova funzione di fruibilità.

Vantaggi dell'adozione di questa best practice: Il miglioramento continuo fornisce un meccanismo per valutare regolarmente le opportunità di miglioramento, assegnare priorità alle opportunità e concentrare gli sforzi laddove è possibile ottenere i maggiori vantaggi.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Definizione di un processo per il miglioramento continuo: valuta regolarmente e dai priorità alle opportunità di miglioramento per concentrare gli sforzi dove producono i vantaggi maggiori. Apporta modifiche per migliorare e valutare i risultati per favorire il successo. Se i risultati non sono in linea con gli obiettivi e il miglioramento resta prioritario, riprova utilizzando linee d'azione alternative. I processi operativi devono prevedere l'allocazione di tempo e risorse per rendere possibile il miglioramento incrementale continuo.

OPS11-BP02 Esecuzione di analisi post-incidente

Esamina gli eventi che influiscono sui clienti e identifica i fattori che contribuiscono e le azioni preventive. Utilizza queste informazioni per sviluppare modi per limitare o prevenire il ripetersi degli incidenti. Sviluppa procedure per attivare risposte rapide ed efficaci. Comunica i fattori che hanno contribuito al presentarsi dell'imprevisto e le azioni correttive secondo necessità, specificamente mirate per il pubblico di destinazione.

Anti-pattern comuni:

- Sei amministratore di un server di applicazioni. Circa ogni 23 ore e 55 minuti tutte le sessioni attive vengono terminate. Hai tentato di identificare ciò che non va a buon fine sul server di applicazioni. Sospetti che potrebbe trattarsi di un problema di rete, ma non riesci a ottenere la collaborazione dal team di rete perché i suoi membri sono troppo occupati per supportarti. Ti manca un processo predefinito da seguire per ottenere supporto e raccogliere le informazioni necessarie per stabilire che cosa sta accadendo.
- Si è verificata una perdita di dati all'interno del carico di lavoro. Questa è la prima volta che si è verificata e la causa non è immediatamente identificabile. Decidi che non è importante perché puoi ricreare i dati. La perdita di dati inizia a verificarsi con maggiore frequenza e influisce sui clienti. Questo comporta inoltre un ulteriore onere operativo quando ripristini i dati mancanti.

Vantaggi dell'adozione di questa best practice: Avere un processo predefinito per determinare i componenti, le condizioni, le azioni e gli eventi che hanno contribuito a un incidente consente di identificare le opportunità di miglioramento.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Utilizzo di un processo per determinare i fattori che contribuiscono al verificarsi di un incidente: esamina tutti gli incidenti che influiscono sul cliente. Predisponi un processo per identificare e documentare i fattori che contribuiscono a un incidente, in modo da sviluppare azioni di mitigazione in grado di limitare o impedire il suo ripetersi e per sviluppare procedure che consentano risposte rapide ed efficaci. Condividi la causa principale nel modo appropriato, personalizzando la comunicazione in base ai destinatari.

OPS11-BP03 Implementazione di cicli di feedback

I cicli di feedback forniscono informazioni fruibili che guidano il processo decisionale. Vanno creati nelle procedure e nei carichi di lavoro per identificare i problemi e le aree che necessitano di miglioramenti. Inoltre, convalidano gli investimenti effettuati nei miglioramenti. Questi cicli di feedback sono la base per migliorare continuamente il carico di lavoro.

I cicli di feedback si dividono in due categorie: feedback immediato e analisi retrospettiva. Il feedback immediato viene raccolto con la revisione delle prestazioni e dei risultati delle attività operative. Questo feedback proviene dai membri del team, dai clienti o dall'output automatizzato dell'attività. Il feedback immediato viene ricevuto ad esempio dal test A/B e dall'offerta di nuove funzionalità, ed è essenziale per anticipare l'errore (fail fast).

L'analisi retrospettiva viene eseguita regolarmente per acquisire il feedback della revisione dei risultati operativi e dei parametri nel tempo. Queste retrospettive si svolgono alla fine di uno sprint, in base a una cadenza o dopo importanti rilasci o eventi. Questo tipo di ciclo di feedback convalida gli investimenti nelle operazioni o nel carico di lavoro, consente di misurare il successo e comprova la tua strategia.

Risultato desiderato: l'uso del feedback immediato e dell'analisi retrospettiva per guidare i miglioramenti. L'applicazione di un meccanismo per acquisire il feedback di utenti e membri del team. L'uso dell'analisi retrospettiva per identificare le tendenze che guidano i miglioramenti.

Anti-pattern comuni:

- Lanci una nuova funzionalità ma non hai modo di ricevere il feedback dei clienti.
- Dopo aver investito in miglioramenti delle operazioni, non conduci una retrospettiva per convalidare gli investimenti.
- Raccogli il feedback dei clienti ma non lo esamini regolarmente.
- I cicli di feedback portano alla proposta di elementi di azione non sono inclusi nel processo di sviluppo software.
- I clienti non ricevono un feedback sui miglioramenti che hanno proposto.

Vantaggi dell'adozione di questa best practice:

- Puoi lavorare a ritroso con il cliente per promuovere nuove funzionalità.
- La cultura della tua organizzazione può reagire più rapidamente ai cambiamenti.
- Le tendenze vengono utilizzate per identificare le opportunità di miglioramento.
- Le retrospettive convalidano gli investimenti effettuati per il carico di lavoro e le operazioni.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

L'implementazione di questa best practice comporta l'utilizzo del feedback immediato e dell'analisi retrospettiva. Questi cicli di feedback guidano i miglioramenti. Esistono molti meccanismi per il feedback immediato, inclusi questionari, sondaggi dei clienti o moduli di feedback. La tua organizzazione utilizza anche le retrospettive per identificare le opportunità di miglioramento e convalidare le iniziative.

Esempio del cliente

AnyCompany Retail crea un modulo Web in cui i clienti possono fornire il feedback o segnalare problemi. Durante lo Scrum settimanale, il feedback degli utenti viene valutato dal team di sviluppo software. Il feedback viene regolarmente utilizzato per guidare l'evoluzione della piattaforma. Viene eseguita una retrospettiva alla fine di ogni sprint per identificare gli elementi che devono essere migliorati.

Passaggi dell'implementazione

1. Feedback immediato
   - Hai bisogno di un meccanismo per ricevere il feedback dai clienti e dai membri del team. Le attività operative possono anche essere configurate per fornire un feedback automatizzato.
   - L'organizzazione ha bisogno di un processo per rivedere il feedback, determinare cosa migliorare e pianificare il miglioramento.
   - Il feedback deve essere aggiunto al processo di sviluppo software.
   - Quando apporti miglioramenti, contatta l'autore del feedback.
     - Puoi utilizzare AWS Systems Manager OpsCenter per creare e monitorare questi miglioramenti come OpsItems.
2. Analisi retrospettiva
   - Conduci le retrospettive alla fine di un ciclo di sviluppo, a una cadenza prestabilita o dopo un rilascio importante.
   - Riunisci gli stakeholder coinvolti nel carico di lavoro per la riunione retrospettiva.
   - Crea tre colonne sulla lavagna o in un foglio di lavoro: Fine, Inizio e Mantenimento.
     - Fine è per tutto ciò che vuoi che il team smetta di fare.
     - Inizio è per le idee che vuoi iniziare ad applicare.
     - Mantenimento è per ciò che vuoi continuare a fare.
   - Raccogli il feedback dagli stakeholder.
   - Dai priorità al feedback. Assegna le azioni e gli stakeholder a qualsiasi elemento nelle colonne Inizio e Mantenimento.
   - Aggiungi le azioni al processo di sviluppo software e comunica gli aggiornamenti sullo stato agli stakeholder mentre apporti i miglioramenti.

Livello di impegno per il piano di implementazione: medio. Per implementare questa best practice è necessario un modo per ricevere il feedback immediato e analizzarlo. Inoltre, è necessario stabilire un processo di analisi retrospettiva.

Risorse

Best practice correlate:

- OPS01-BP01 Valutazione delle esigenze dei clienti esterni: i cicli di feedback sono un meccanismo per raccogliere le esigenze dei clienti esterni.
- OPS01-BP02 Valutazione delle esigenze dei clienti interni: gli stakeholder interni possono utilizzare i cicli di feedback per comunicare necessità e requisiti.
- OPS11-BP02 Esecuzione di analisi post-incidente: le analisi successive agli incidenti sono una forma importante di analisi retrospettiva da condurre dopo gli incidenti.
- OPS11-BP07 Revisione dei parametri delle operazioni: le revisioni dei parametri operativi identificano tendenze e aree di miglioramento.

Documenti correlati:

- 7 Pitfalls to Avoid When Building CCOE (7 errori da evitare durante la creazione di un Centro di eccellenza del Cloud (CCoE))
- Atlassian Team Playbook - Retrospectives (Playbook Atlassian Team - Retrospettive)
- Email Definitions: Feedback Loops (Definizioni di e-mail: cicli di feedback)
- Establishing Feedback Loops Based on the AWS Well-Architected Framework Review (Applicazione dei cicli di feedback in base alla revisione di Framework AWS Well-Architected)
- IBM Garage Methodology - Hold a retrospective (Metodologia IBM Garage - Condurre una retrospettiva)
- Investopedia - The PDCS Cycle (Investopedia - Il ciclo PDCA)
- Maximizing Developer Effectiveness by Tim Cochran (Massimizzazione dell'efficacia degli sviluppatori di Tim Cochran)
- Operations Readiness Reviews (ORR) Whitepaper - Iteration (Whitepaper per le revisioni della preparazione delle operazioni - Iterazione)
- TIL CSI - Continual Service Improvement (TIL CSI - Miglioramento continuo del servizio)
- When Toyota met e-commerce: Lean at Amazon (Toyota incontra l'e-commerce: semplificazione con Amazon)

Video correlati:

- Building Effective Customer Feedback Loops (Creazione di efficaci cicli di feedback dei clienti)

Esempi correlati:

- Astuto - Open source customer feedback tool (Astuto - Strumento di feedback dei clienti open source)
- AWS Solutions - QnABot on AWS (Soluzioni AWS - QnABot in AWS)
- Fider - A platform to organize customer feedback (Fider - Una piattaforma per organizzare il feedback dei clienti)

Servizi correlati:

- AWS Systems Manager OpsCenter

OPS11-BP04 Gestione delle conoscenze

I membri del tuo team dispongono di meccanismi per trovare le informazioni che cercano in modo tempestivo, per accedervi e per verificare che siano attuali e complete. Sono disponibili meccanismi per identificare i contenuti necessari, i contenuti che necessitano di aggiornamento e i contenuti che devono essere archiviati per non essere più richiamati.

Anti-pattern comuni:

- Un singolo cliente frustrato apre un caso di supporto in relazione a una nuova richiesta di funzionalità del prodotto per risolvere un problema percepito. La richiesta viene aggiunta all'elenco dei miglioramenti prioritari.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Gestione delle conoscenze: assicurati che i membri del team dispongano di meccanismi per trovare le informazioni che cercano in modo tempestivo, per accedervi e per verificare che siano attuali e complete. Mantieni meccanismi per identificare i contenuti necessari, i contenuti che necessitano di aggiornamento e i contenuti che devono essere archiviati per non essere più richiamati.

OPS11-BP05 Definizione dei fattori che promuovono il miglioramento

Identifica i fattori che promuovono il miglioramento, in modo da valutare e dare priorità alle opportunità.

Su AWS, è possibile aggregare i registri di tutte le tue attività operative, i tuoi carichi di lavoro e le tue infrastrutture per creare una cronologia dettagliata dell'attività. Puoi quindi utilizzare gli strumenti AWS per analizzare lo stato delle tue operazioni e del carico di lavoro nel tempo (ad esempio identificare le tendenze, correlare eventi e attività ai risultati, nonché confrontare ed evidenziare le differenze tra ambienti e all'interno di sistemi) per rilevare le opportunità di miglioramento in base ai fattori che hai definito.

Potresti utilizzare CloudTrail per tracciare l'attività API (attraverso la AWS Management Console, CLI, SDK e API) per scoprire cosa sta succedendo nei tuoi account. Traccia le tue attività di distribuzione degli Strumenti per sviluppatori AWS con CloudTrail e CloudWatch. In questo modo sarà aggiunta ai dati di log di CloudWatch Logs una cronologia dettagliata delle attività delle distribuzioni e dei loro risultati.

Esporta i dati di log in Amazon S3 per lo storage a lungo termine. Utilizzando AWS Glue, puoi individuare e preparare i dati di log in Amazon S3 per l'analisi. Utilizzo Amazon Athena, attraverso l'integrazione nativa con AWS Glue, per analizzare i dati di registro. Utilizza uno strumento di business intelligence come Amazon QuickSight per visualizzare, esplorare e analizzare i dati.

Anti-pattern comuni:

- Hai uno script che funziona ma non è scritto nel modo migliore. Dedichi del tempo alla sua riscrittura. Ora è un'opera d'arte.
- La tua start-up sta cercando di ottenere un'altra serie di finanziamenti da un investitore in capitali di rischio. Ti viene richiesto di dimostrare la conformità allo standard PCI DSS. Per documentare la conformità non riesci a rispettare una data di consegna e, di conseguenza, perdi un cliente. Non era una scelta sbagliata ma ora ti chiedi se fosse la cosa giusta da fare.

Vantaggi dell'adozione di questa best practice: Stabilendo quali criteri desideri utilizzare per migliorare, puoi ridurre al minimo l'impatto delle motivazioni basate sugli eventi o degli investimenti influenzati da fattori emotivi.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Comprensione dei fattori che promuovono il miglioramento: è consigliabile apportare modifiche a un sistema solo quando un risultato desiderato è supportato.
  - Funzionalità desiderate: prendi in considerazione le funzionalità e le capacità desiderate quando valuti le opportunità di miglioramento.
    - Novità di AWS
  - Problemi inaccettabili: tieni in considerazione i problemi, i bug e le vulnerabilità inaccettabili quando valuti le opportunità di miglioramento.
    - Ultimi bollettini di sicurezza AWS
    - AWS Trusted Advisor
  - Requisiti di conformità: quando esamini le opportunità di miglioramento, prendi in considerazione gli aggiornamenti e le modifiche necessarie per mantenere la conformità a normative e policy o per avere diritto al supporto di terze parti.
    - Conformità di AWS
    - Programmi per la conformità di AWS
    - Ultime novità sulla conformità di AWS

Risorse

Documenti correlati:

- Amazon Athena
- Amazon QuickSight
- Conformità di AWS
- Ultime novità sulla conformità di AWS
- Programmi per la conformità di AWS
- AWS Glue
- Ultimi bollettini di sicurezza AWS
- AWS Trusted Advisor
- Esporta i dati di log in Amazon S3
- Novità di AWS

OPS11-BP06 Convalida delle informazioni

Rivedi i risultati dell'analisi e le risposte con i team trasversali e i proprietari dell'azienda. Utilizza queste revisioni per definire una visione comune, identificare ulteriori impatti e stabilire le linee d'azione. Adatta le risposte, se necessario.

Anti-pattern comuni:

- Noti che su un sistema l'utilizzo della CPU è al 95% e decidi che è prioritario trovare un modo per ridurre il carico sul sistema. Stabilisci che la soluzione migliore è dimensionare verso l'alto. Il sistema, un transcodificatore, è stato calibrato per funzionare con un utilizzo costante della CPU al 95%. Se l'avessi contattato, il responsabile del sistema avrebbe potuto spiegarti la situazione. Hai sprecato il tuo tempo.
- Il responsabile di un sistema ritiene che il sistema sia mission critical. Il sistema non è stato inserito in un ambiente a sicurezza elevata. Per migliorare la sicurezza, adotti i controlli di rilevazione e prevenzione aggiuntivi necessari per i sistemi mission critical. Comunichi al proprietario del sistema che il lavoro è completo e che gli verranno addebitati i costi per le risorse aggiuntive. Nella discussione che segue a questa notifica, il responsabile del sistema apprende che esiste una definizione formale di mission critical che il suo sistema non soddisfa.

Vantaggi dell'adozione di questa best practice: Convalidando le informazioni con i responsabili aziendali e con gli esperti in materia, è possibile stabilire una comprensione comune e gestire il miglioramento in modo più efficace.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Convalida delle informazioni: interagisci con i responsabili aziendali e gli esperti in materia per garantire la comprensione e l'accordo comuni sul significato dei dati raccolti. Individua ulteriori problemi e impatti potenziali e stabilisci le azioni da intraprendere.

OPS11-BP07 Revisione dei parametri delle operazioni

Esegui regolarmente un'analisi retrospettiva dei parametri operativi con i partecipanti di vari team da diverse aree del business. Utilizza queste revisioni per identificare opportunità di miglioramento e potenziali linee d'azione e per condividere le conoscenze acquisite.

Cerca opportunità di miglioramento in tutti i tuoi ambienti (per esempio sviluppo, test e produzione).

Anti-pattern comuni:

- Un'importante promozione al dettaglio è stata interrotta da uno dei tuoi interventi di manutenzione. L'azienda non è al corrente del fatto che i normali interventi di manutenzione possono essere rimandati nel caso vi siano altri eventi di particolare rilievo per l'azienda.
- Per l'uso di una libreria contenente degli errori comunemente utilizzata nella tua organizzazione si è verificata una prolungata interruzione del servizio. Successivamente hai eseguito la migrazione a una libreria affidabile. Gli altri team della tua organizzazione non sanno di essere a rischio. Se si svolgessero incontri periodici durante i quali esaminare questo incidente, anch'essi sarebbero al corrente del rischio.
- Le prestazioni del transcodificatore hanno avuto un peggioramento, con conseguenti problemi prolungati per il team multimediale. Il problema non è ancora grave e non lo diventerà finché non sarà tanto avanzato da provocare un incidente. Se esaminassi i parametri operativi con il team multimediale, ci sarebbe l'occasione di modificare i parametri, riconoscendo l'esperienza dei tuoi interlocutori e affrontando il problema.
- Non stai tenendo sotto controllo i contratti sul livello di servizio (SLA) che misurano la soddisfazione dei clienti. Le tendenze indicano un andamento negativo per quanto riguarda il rispetto degli SLA. In caso di mancato rispetto degli SLA, sono previste sanzioni economiche. Se si tenessero incontri periodici per esaminare i parametri di questi Accordi sul livello di servizio (SLA), ci sarebbe l'opportunità riconoscere e risolvere il problema.

Vantaggi dell'adozione di questa best practice: riunendosi regolarmente per esaminare i parametri operativi, gli eventi e gli incidenti, si mantiene una comprensione comune tra i team e si condividono i risultati ottenuti, assegnando priorità e indirizzando i miglioramenti in modo più preciso.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Revisioni dei parametri operativi: esegui regolarmente un'analisi retrospettiva dei parametri operativi con i partecipanti di vari team da diverse aree del business. Coinvolgi i soggetti interessati, compresi i team che si occupano di business, sviluppo e operazioni, per convalidare ciò che è emerso dal feedback immediato e dall'analisi retrospettiva e per condividere le conoscenze acquisite. Utilizza le informazioni di cui dispongono per identificare opportunità di miglioramento e possibili linee d'azione.
  - Amazon CloudWatch
  - Utilizzare i parametri Amazon CloudWatch

- Pubblicazione di parametri personalizzati
- Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch

Risorse

Documenti correlati:

- Amazon CloudWatch
- Documentazione di riferimento su parametri e dimensioni di Amazon CloudWatch
- Pubblicazione di parametri personalizzati
- Utilizzare i parametri Amazon CloudWatch

OPS11-BP08 Documentazione e condivisione delle conoscenze acquisite

Documenta e condividi le conoscenze acquisite durante le attività operative per metterle a frutto internamente e nei vari team.

La condivisione di quanto appreso dai team comporta maggiori vantaggi all'interno dell'organizzazione. Dovrai condividere informazioni e risorse per impedire che si verifichino errori evitabili, nonché semplificare le attività di sviluppo. In questo modo potrai concentrarti sulla distribuzione delle funzionalità desiderate.

Utilizza AWS Identity and Access Management (IAM) per definire i permessi che consentono un accesso controllato alle risorse che desideri condividere all'interno e tra i vari account. Dovrai utilizzare repository AWS CodeCommit dotati di controllo versione per condividere librerie dell'applicazione, procedure di scripting, documentazione di procedure e altra documentazione di sistema. Metti a disposizione i tuoi standard di elaborazione condividendo l'accesso ai tuoi AMI e fornendo l'autorizzazione a utilizzare le tue funzioni Lambda nei vari account. È consigliabile condividere i tuoi standard infrastrutturali come modelli AWS CloudFormation.

Grazie ad API e SDK di AWS, hai modo di integrare strumenti e repository esterni e di parti terze (ad es. GitHub, BitBucket e SourceForge). Quando condividi ciò che hai appreso e sviluppato, fai attenzione a strutturare i permessi in modo tale da garantire l'integrità dei repository condivisi.

Anti-pattern comuni:

- Per l'uso di una libreria contenente degli errori comunemente utilizzata nella tua organizzazione si è verificata una prolungata interruzione del servizio. Successivamente hai eseguito la migrazione

a una libreria affidabile. Gli altri team della tua organizzazione non sanno di essere a rischio. Se tu documentassi e condividessi la tua esperienza con questa libreria, sarebbero al corrente del rischio.

- Hai identificato un caso limite in un microservizio condiviso internamente che causa l'interruzione delle sessioni. Hai aggiornato le chiamate al servizio per evitare questo caso limite. Gli altri team della tua organizzazione non sanno di essere a rischio. Se tu documentassi e condividessi la tua esperienza con questa libreria, sarebbero al corrente del rischio.
- Hai trovato un modo per ridurre in modo significativo i requisiti di utilizzo della CPU per uno dei tuoi microservizi. Non sai se altri team potrebbero sfruttare questa tecnica. Se tu documentassi e condividessi la tua esperienza con questa libreria, avrebbero l'opportunità di farlo.

Vantaggi dell'adozione di questa best practice: condividere le lezioni apprese a supporto del miglioramento e per trarre il massimo vantaggio dall'esperienza.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Documentazione e condivisione delle conoscenze acquisite: predisponi procedure per documentare le conoscenze acquisite dall'esecuzione delle attività operative e dalle analisi retrospettive affinché tali informazioni possano essere utilizzate dal altri team.
  - Condividi le conoscenze acquisite: predisponi procedure per condividere con tutti i team le conoscenze acquisite e gli artefatti associati. Ad esempio condividi le procedure, le istruzioni, la governance e le best practice aggiornate tramite un wiki accessibile. Condividi script, codice e librerie tramite un repository comune.
    - Delega dell'accesso all'ambiente AWS
    - Condivisione di un repository AWS CodeCommit
    - Autorizzazione semplificata delle funzioni AWS Lambda
    - Condivisione di un'AMI con account AWS specifici
    - Condivisione più rapida dei modelli con un URL di AWS CloudFormation Designer
    - Utilizzo di AWS Lambda con Amazon SNS

Risorse

Documenti correlati:

- Autorizzazione semplificata delle funzioni AWS Lambda
- Condivisione di un repository AWS CodeCommit
- Condivisione di un'AMI con account AWS specifici
- Condivisione più rapida dei modelli con un URL di AWS CloudFormation Designer
- Utilizzo di AWS Lambda con Amazon SNS

Video correlati:

- Delega dell'accesso all'ambiente AWS

OPS11-BP09 Allocazione di tempo per i miglioramenti

Dedica tempo e risorse all'interno dei processi per rendere possibile il miglioramento incrementale continuo.

Su AWS puoi creare duplicati temporanei paralleli di ambienti per ridurre il rischio, lo sforzo e il costo della sperimentazione e dell'esecuzione di test. Questi ambienti duplicati possono essere utilizzati per testare le conclusioni di analisi ed esperimenti, ma anche per sviluppare e testare i miglioramenti pianificati.

Anti-pattern comuni:

- Si è verificato un problema di prestazioni noto nel server di applicazioni. Il problema viene aggiunto al backlog, dopo l'implementazione prevista delle varie funzionalità. Se la velocità con cui vengono aggiunte le funzionalità pianificate rimane costante, il problema di prestazioni non verrà mai risolto.
- Per supportare il miglioramento continuo, autorizzi amministratori e sviluppatori a utilizzare tutto il loro tempo aggiuntivo per selezionare e implementare miglioramenti. I miglioramenti non vengono mai completati.

Vantaggi dell'adozione di questa best practice: Dedicando tempo e risorse all'interno dei processi, renderai possibile il miglioramento incrementale continuo.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Allocazione di tempo per apportare miglioramenti: dedica tempo e risorse all'interno dei processi per rendere possibili miglioramenti graduali e continui. Implementa modifiche per migliorare

e valutare i risultati per favorire il successo. Se i risultati non sono in linea con gli obiettivi e il miglioramento resta prioritario, valuta linee d'azione alternative.

Sicurezza

Argomenti

Nozioni di base sulla sicurezza

Domanda

SEC 1 Come gestisci in modo sicuro un carico di lavoro?

Per gestire il carico di lavoro in modo sicuro, è necessario applicare le best practice globali a ogni area di sicurezza. Segui i requisiti e i processi definiti in termini di eccellenza operativa a livello organizzativo e di carico di lavoro e applicali a tutte le aree. Rimanere aggiornati con le raccomandazioni di AWS e del settore nonché con l'intelligence sulle minacce aiuta a sviluppare il modello di rischio e gli obiettivi di controllo. L'automazione dei processi di sicurezza, i test e la convalida consentono di ricalibrare le operazioni di sicurezza.

Best practice

- SEC01-BP06 Automatizzazione dei test e della convalida dei controlli di sicurezza nelle pipeline
- SEC01-BP07 Identificazione e assegnazione di priorità ai rischi utilizzando un modello di minaccia
- SEC01-BP08 Valutazione e implementazione periodiche di nuovi servizi e funzionalità di sicurezza

SEC01-BP01 Separazione dei carichi di lavoro tramite account

Inizia tenendo conto della sicurezza e dell'infrastruttura per consentire alla tua organizzazione di impostare guardrail comuni al crescere dei carichi di lavoro. Questo approccio fornisce limiti e controlli tra i carichi di lavoro. La separazione a livello di account è fortemente consigliata per isolare gli ambienti di produzione dagli ambienti di sviluppo e test, oppure per fornire un forte limite logico tra i carichi di lavoro che elaborano dati con diversi livelli di sensibilità, secondo quanto definito da requisiti di conformità esterni (ad esempio PCI-DSS o HIPAA), e i carichi di lavoro che non lo fanno.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Utilizzo di AWS Organizations: utilizza AWS Organizations per applicare a livello centrale una gestione basata su policy per più account Account AWS.
  - Nozioni di base su AWS Organizations
  - Come usare le policy di controllo dei servizi per impostare guardrail di permessi negli account della tua AWS Organization
- Considerazioni su AWS Control Tower: AWS Control Tower rappresenta un modo semplice per configurare e gestire un nuovo ambiente AWS sicuro e multi-account in base alle best practice.
  - AWS Control Tower

Risorse

Documenti correlati:

- Best practice IAM
- Bollettini sulla sicurezza
- AWS Security Audit Guidelines

Video correlati:

- Gestire ambienti AWS multi-account tramite AWS Organizations

- Security Best Practices the Well-Architected Way
- Uso di AWS Control Tower per gestire ambienti AWS multi-account

SEC01-BP02 Protezione Account AWS

La protezione degli account Account AWS prevede diversi aspetti, tra cui la protezione di e il non utilizzo dell' utente roote l'aggiornamento costante delle informazioni di contatto. Puoi utilizzare AWS Organizations per gestire e amministrare centralmente i tuoi account man mano che i tuoi carichi di lavoro in AWS crescono e li ridimensioni. AWS Organizations ti aiuta a gestire gli account, impostare controlli e configurare i servizi tra gli account.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Utilizzo di AWS Organizations: utilizza AWS Organizations per applicare a livello centrale una gestione basata su policy per più account Account AWS.
  - Nozioni di base su AWS Organizations
  - Come usare le policy di controllo dei servizi per impostare guardrail di permessi negli account della tua AWS Organization
- Limita l'uso dell'utente root AWS. Utilizza l'utente root solo per eseguire attività che lo richiedono specificamente.
  - Attività AWS che richiedono le credenziali di un utente root dell'account AWS
- Abilita l'autenticazione a più fattori (MFA) per l'utente root: abilita MFA sull'utente root Account AWS, se AWS Organizations non gestisce gli utenti root per te.
  - Utente root
- Modifica periodica della password dell'utente root: modificare periodicamente la password dell'utente root riduce il rischio di utilizzo di una password salvata. Si tratta di un aspetto particolarmente importante se non utilizzi AWS Organizations e chiunque dispone di un accesso fisico.
  - Modifica della password dell'utente root Account AWS
- Abilita le notifiche quando viene usato l'utente root Account AWS: ricevere una notifica in automatico riduce i rischi.
  - Come ricevere notifiche quando si utilizzano chiavi di accesso root di Account AWS

- Accesso limitato a regioni aggiunte di recente: per le nuove Regioni AWS, le risorse IAM, ad esempio utenti e ruoli, verranno propagate solo alle regioni abilitate.
  - Definizione dei permessi per abilitare gli account per Regioni AWS imminenti
- Considera AWS CloudFormation StackSets: CloudFormation StackSets consente di distribuire risorse, tra cui policy, ruoli e gruppi IAM, in una serie di regioni e Account AWS a partire da un modello approvato.
  - Uso di CloudFormation StackSets

Risorse

Documenti correlati:

- AWS Control Tower
- AWS Security Audit Guidelines
- Best practice IAM
- Bollettini sulla sicurezza

Video correlati:

- Abilita l'adozione AWS su scala con automazione e governance
- Security Best Practices the Well-Architected Way

Esempi correlati:

- Laboratorio: Account AWS e utente root

SEC01-BP03 Identificazione e convalida degli obiettivi di controllo

In base ai requisiti di conformità e ai rischi identificati dal modello di rischio, deriva e convalida gli obiettivi di controllo e i controlli da applicare al carico di lavoro. La convalida continua degli obiettivi di controllo e dei controlli aiuta a misurare l'efficacia della mitigazione dei rischi.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Identifica i requisiti di conformità. Scopri i requisiti organizzativi, legali e di conformità perché il tuo carico di lavoro risulti conforme.
- Identifica le risorse di conformità AWS: identifica le risorse che AWS mette a disposizione per aiutarti nei processi di conformità.
  - https://aws.amazon.com/compliance/
  - https://aws.amazon.com/artifact/

Risorse

Documenti correlati:

- AWS Security Audit Guidelines
- Bollettini sulla sicurezza

Video correlati:

- AWS Security Hub: gestire gli avvisi di sicurezza e automatizzare la conformità
- Security Best Practices the Well-Architected Way

SEC01-BP04 Aggiornamento costante sulle minacce alla sicurezza

Per definire e implementare controlli appropriati, riconosci i vettori di attacco rimanendo aggiornato sulle minacce alla sicurezza più recenti. Usa AWS Managed Services per semplificare la ricezione di notifiche in seguito a comportamenti inaspettati o inusuali nei tuoi account AWS. Esegui delle indagini avvalendoti degli strumenti AWS Partner o di feed di informazioni sulle minacce di terze parti come parte del tuo flusso di informazioni di sicurezza. Al CVE (Common Vulnerabilities and Exposures) contiene vulnerabilità di sicurezza informatica pubbliche che puoi utilizzare come aggiornamento.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Iscrizione alle fonti di informazione sulle minacce: consulta regolarmente le informazioni sulle minacce da varie fonti attinenti alle tecnologie che utilizzi per il tuo carico di lavoro.
  - Elenco CVE (Common Vulnerabilities and Exposures)

- Considera il servizio AWS Shield Advanced : fornisce visibilità quasi in tempo reale sulle fonti di intelligence, se il tuo carico di lavoro è accessibile da Internet.

Risorse

Documenti correlati:

- AWS Security Audit Guidelines
- AWS Shield
- Bollettini sulla sicurezza

Video correlati:

- Security Best Practices the Well-Architected Way

SEC01-BP05 Aggiornamento costante sulle raccomandazioni di sicurezza

Tieniti aggiornato sulle raccomandazioni di sicurezza di AWS e del settore, così da revisionare l'assetto di sicurezza del tuo carico di lavoro. Bollettini sulla sicurezza AWS contengono informazioni importanti sulla sicurezza e notifiche relative alla privacy.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Segui gli aggiornamenti di AWS: segui o verifica regolarmente la presenza di nuovi consigli, suggerimenti e trucchi.
  - AWS Well-Architected Labs
  - Blog sulla sicurezza AWS
  - Documentazione del servizio AWS
- Sottoscrivi gli aggiornamenti di settore: consulta regolarmente le notizie da varie fonti attinenti alle tecnologie impiegate nel tuo carico di lavoro.
  - Esempio: Elenco CVE (Common Vulnerabilities and Exposures)

Risorse

Documenti correlati:

- Bollettini sulla sicurezza

Video correlati:

- Security Best Practices the Well-Architected Way

SEC01-BP06 Automatizzazione dei test e della convalida dei controlli di sicurezza nelle pipeline

Stabilisci previsioni e modelli sicuri per i meccanismi di sicurezza testati e convalidati come parte della compilazione, delle pipeline e dei processi. Utilizza strumenti e l'automazione per testare e convalidare tutti i controlli di sicurezza in modo continuo. Ad esempio, scansiona elementi quali immagini di macchine e modelli di infrastrutture come codice per individuare vulnerabilità di sicurezza, irregolarità e deviazioni da una previsione stabilita in ogni fase. AWS CloudFormation Guard può aiutarti a verificare la sicurezza dei modelli CloudFormation, a risparmiare tempo e a ridurre il rischio che si verifichino errori di configurazione.

È fondamentale ridurre il numero di errori di sicurezza introdotti in un ambiente di produzione, quindi più operazioni di controllo di qualità e riduzione dei difetti è possibile eseguire nel processo di compilazione, più efficace sarà il risultato. Progetta pipeline di integrazione e distribuzione continue (CI/CD) per testare eventuali problemi di sicurezza quando possibile. Le pipeline CI/CD offrono l'opportunità di migliorare la sicurezza in ogni fase della compilazione e della distribuzione. Anche gli strumenti di sicurezza CI/CD devono essere mantenuti aggiornati per mitigare le minacce in continua evoluzione.

Monitora le modifiche alla configurazione del tuo carico di lavoro per facilitare gli audit di conformità, la gestione delle modifiche e le indagini che possono essere applicate al tuo caso. Puoi usare AWS Config per registrare e valutare le tue risorse AWS e di terze parti. Consente di eseguire audit costanti e di valutare la conformità generale a regole e pacchetti di conformità, ossia raccolte di regole con azioni di correzione.

Nel monitoraggio delle modifiche sono incluse modifiche pianificate, parte del processo di controllo delle modifiche della tua organizzazione (a cui a volte si fa riferimento con l'acronimo MACD: Move, Add, Change, Delete), le modifiche non pianificate e le modifiche inaspettate, come gli incidenti. Le modifiche possono avvenire a livello di infrastruttura, ma essere relative anche ad altre categorie, come le modifiche nei repository di codice, le modifiche delle immagini di macchine e degli inventari di applicazioni, le modifiche di processi e policy o le modifiche alla documentazione.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Automazione della gestione della configurazione: applica e convalida in automatico configurazioni sicure sfruttando un apposito servizio o strumento di gestione.
  - AWS Systems Manager
  - AWS CloudFormation
  - Configurazione di una pipeline CI/CD in AWS

Risorse

Documenti correlati:

- Come usare le policy di controllo dei servizi per impostare guardrail di permessi negli account della tua AWS Organization

Video correlati:

- Gestire ambienti AWS multi-account tramite AWS Organizations
- Security Best Practices the Well-Architected Way

SEC01-BP07 Identificazione e assegnazione di priorità ai rischi utilizzando un modello di minaccia

Utilizza un modello di rischio per identificare e mantenere un registro aggiornato delle potenziali minacce. Classifica le minacce in ordine di priorità e adatta i controlli di sicurezza in modo da prevenirle, rilevarle e affrontarle. Rivedi e mantieni questo approccio nel contesto dell'evoluzione del panorama della sicurezza.

La modellazione delle minacce offre un approccio sistematico di supporto nell'individuazione e nella risoluzione di problematiche di sicurezza nella fase iniziale del processo di progettazione. Se le mitigazioni avvengono prima è meglio, perché i costi sono inferiori a una fase più avanzata del ciclo di vita.

I passaggi di base tipici del processo di modellazione delle minacce sono:

1. Identificare gli asset, gli attori, i punti d'ingresso, i componenti, i casi d'uso e i livelli di fiducia e includerli in un diagramma di progettazione.
2. Identificare un elenco di minacce.

3. Per ogni minaccia identificare le mitigazioni, che possono includere implementazioni di controlli di sicurezza.
4. Creare e rivedere una matrice dei rischi per stabilire se la minaccia è stata correttamente mitigata.

La modellazione delle minacce è più efficace se viene eseguita a livello di carico di lavoro (o di funzionalità del carico di lavoro), garantendo la disponibilità dell'intero contesto per la valutazione. Mantenere e aggiornare questa matrice in linea con l'evoluzione dello scenario di sicurezza.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Creazione di un modello di minaccia: un modello di minaccia aiuta a identificare e affrontare potenziali minacce alla sicurezza.
  - NIST: Guida alla modellazione delle minacce del sistema incentrato sui dati

Risorse

Documenti correlati:

- AWS Security Audit Guidelines
- Bollettini sulla sicurezza

Video correlati:

- Security Best Practices the Well-Architected Way

SEC01-BP08 Valutazione e implementazione periodiche di nuovi servizi e funzionalità di sicurezza

Valuta e implementa servizi e funzionalità di sicurezza di AWS e partner AWS che consentano di sviluppare l'assetto di sicurezza del carico di lavoro. Il blog sulla sicurezza AWS evidenzia nuovi servizi e funzionalità AWS, guide all'implementazione e linee guida generali sulla sicurezza. Novità di AWS è un'ottima scelta per essere aggiornati su tutte le nuove funzionalità, i servizi e gli annunci AWS.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Pianificazione di revisioni regolari: crea un calendario di attività di revisione che preveda requisiti di conformità, valutazione delle nuove funzionalità e dei nuovi servizi di sicurezza AWS e l'aggiornamento costante rispetto alle novità del settore.
- Funzionalità e servizi AWS: scopri le funzionalità di sicurezza disponibili per i servizi che utilizzi e approfondisci le nuove caratteristiche al momento del rilascio.
  - Blog sulla sicurezza AWS
  - Bollettini sulla sicurezza AWS
  - Documentazione del servizio AWS
- Definizione del processo di onboarding del servizio AWS: definisci i processi per l'onboarding di nuovi servizi AWS. Includi il modo in cui valuti la funzionalità dei nuovi servizi AWS e i requisiti di conformità per il tuo carico di lavoro.
- Test di nuovi servizi e funzionalità: testa nuovi servizi e funzionalità al momento del rilascio in un ambiente non di produzione che replica in maniera fedele quello di produzione.
- Implementazione di altri meccanismi di difesa: implementa meccanismi automatizzati per difendere il carico di lavoro, esplora le opzioni disponibili.
  - Correzione di risorse AWS non conformi in base alle regole di Regole di AWS Config

Risorse

Video correlati:

- Security Best Practices the Well-Architected Way

Gestione di identità e accessi

Domande

- SEC 2 Come gestisci l'autenticazione per persone e macchine?
- SEC 3 Come gestisci le autorizzazioni per persone e macchine?

SEC 2 Come gestisci l'autenticazione per persone e macchine?

Ci sono due tipi di identità da gestire quando ci si avvicina all'utilizzo di carichi di lavoro AWS sicuri. Comprendere il tipo di identità necessaria per gestire e concedere l'accesso ti aiuta a garantire che le identità corrette abbiano accesso alle risorse giuste nelle condizioni adeguate.

Identità umane: amministratori, sviluppatori, operatori e utenti finali necessitano di un'identità per accedere agli ambienti e alle applicazioni AWS. Si tratta di membri dell'organizzazione o utenti esterni con cui collabori e che interagiscono con le tue risorse AWS tramite browser Web, applicazioni client o strumenti a riga di comando interattivi.

Identità di macchine: le applicazioni di servizio, gli strumenti operativi e i carichi di lavoro necessitano di un'identità per effettuare richieste ai servizi AWS, ad esempio per leggere i dati. Queste identità includono macchine in esecuzione nell'ambiente AWS, ad esempio istanze Amazon EC2 o funzioni AWS Lambda. Puoi gestire le identità di macchine anche per soggetti esterni che necessitano dell'accesso. Inoltre, potresti disporre di macchine al di fuori di AWS che devono accedere al tuo ambiente AWS.

Best practice

- SEC02-BP01 Utilizzo meccanismi di accesso efficaci
- SEC02-BP02 Utilizzo di credenziali temporanee
- SEC02-BP03 Archiviazione e utilizzo dei segreti in modo sicuro
- SEC02-BP04 Fai affidamento su un provider di identità centralizzato
- SEC02-BP05 Verifica e rotazione periodica delle credenziali
- SEC02-BP06 Utilizzo dei gruppi di utenti e degli attributi

SEC02-BP01 Utilizzo meccanismi di accesso efficaci

Imposta la lunghezza minima della password e spiega agli utenti la necessità di evitare password comuni o utilizzate in precedenza. Applica la Multi-Factor Authentication (MFA) con meccanismi software o hardware per garantire un ulteriore livello di verifica. Ad esempio, quando utilizzi Centro identità IAM come origine di identità, configura l'impostazione "Compatibile con il contesto" o "Sempre attivo" per MFA e consenti agli utenti di registrare i propri dispositivi MFA per accelerare l'adozione. Quando utilizzi un provider di identità (IdP) esterno, configura il provider di identità per MFA.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Creazione di una policy Identify and Access Management (IAM) per applicare l'accesso MFA: crea una policy IAM gestita del cliente che vieta tutte le azioni IAM eccetto quelle che consentono a un utente di assumere ruoli, cambiare le proprie credenziali e gestire i dispositivi MFA sulla pagina Le mie credenziali di sicurezza.
- Abilitazione di MFA nel provider di identità: abilita MFA nel provider di identità o sul servizio single sign-on, come AWS IAM Identity Center, che usi.
- Configurazione di una policy di password sicura: configura una policy di password sicura in IAM e nei sistemi di identità federate per contribuire alla protezione da attacchi di forza bruta.

- Rotazione regolare delle credenziali: assicurati che gli amministratori del carico di lavoro modifichino le password e le chiavi di accesso (se utilizzate) con cadenza regolare.

Risorse

Documenti correlati:

- Nozioni di base su AWS Secrets Manager
- Best Practice IAM
- Provider di identità e federazione
- L'utente root dell'account AWS
- Nozioni di base su AWS Secrets Manager
- Credenziali di sicurezza temporanee
- Soluzioni dei partner per la sicurezza: accesso e controllo degli accessi
- Credenziali di sicurezza temporanee
- L'utente root dell'account AWS

Video correlati:

- Best practice per la gestione, il recupero e la rotazione di segreti su scala
- Gestione delle autorizzazioni utente su scala con IAM Identity Center
- Mastering identity at every layer of the cake

SEC02-BP02 Utilizzo di credenziali temporanee

richiedi alle identità di acquisire dinamicamente credenziali temporanee. Per le identità della forza lavoro, utilizza AWS IAM Identity Center o la federazione con ruoli AWS Identity and Access Management (IAM) per accedere a Account AWS. Per le identità di macchine, ad esempio le istanze Amazon Elastic Compute Cloud(Amazon EC2) o le funzioni AWS Lambda, è necessario utilizzare ruoli IAM anziché utenti IAM con chiavi di accesso a lungo termine.

Per le identità umane che utilizzano AWS Management Console, è necessario che gli utenti acquisiscano credenziali temporanee ed eseguano la federazione in AWS. Puoi farlo con il portale utenti AWS IAM Identity Center. Per gli utenti che richiedono l'accesso alla CLI, assicurati di utilizzare AWS CLI v2, che supporta l'integrazione diretta con IAM Identity Center. Gli utenti possono creare profili CLI collegati ad account e ruoli IAM Identity Center. La CLI recupera automaticamente le credenziali AWS da IAM Identity Center e le aggiorna per tuo conto. In questo modo non è più necessario copiare e incollare credenziali AWS temporanee dalla console IAM Identity Center. Per l'SDK, gli utenti devono fare affidamento su AWS Security Token Service (AWS STS) per acquisire ruoli per ricevere credenziali temporanee. In alcuni casi, le credenziali temporanee potrebbero non essere pratiche. È necessario conoscere i rischi che comporta l'archiviazione delle chiavi di accesso, ruotarle spesso e richiedere l'autenticazione a più fattori (MFA) come condizione quando possibile. Uso delle informazioni a cui è stato eseguito l'ultimo accesso per stabilire quando ruotare o rimuovere le chiavi di accesso.

Per i casi in cui è necessario concedere ai consumatori l'accesso alle risorse AWS, utilizza i pool di identità di Amazon Cognito e assegna loro un set di credenziali temporanee con credenziali limitate per accedere alle risorse AWS. Le autorizzazioni per ciascun utente sono controllate tramite i ruoli IAM creati. Puoi definire regole per scegliere il ruolo per ogni utente in base alle registrazioni nel token ID dell'utente. Puoi definire un ruolo predefinito per gli utenti autenticati. Puoi anche definire un ruolo IAM separato con autorizzazioni limitate per gli utenti guest non autenticati.

Per le identità di macchine, è necessario fare affidamento sui ruoli IAM per concedere l'accesso ad AWS. Per le istanze Amazon Elastic Compute Cloud(Amazon EC2), puoi utilizzare i ruoli per Amazon EC2. Puoi collegare un ruolo IAM all'istanza Amazon EC2 per consentire alle applicazioni in esecuzione su Amazon EC2 di utilizzare credenziali di sicurezza temporanee create, distribuite e fatte ruotare automaticamente da AWS tramite Instance Metadata Service (IMDS). La versione più recente di IMDS aiuta a proteggere da vulnerabilità che espongono le credenziali temporanee e devono essere implementate. Per accedere alle istanze Amazon EC2 con chiavi o password, AWS Systems Manager è un modo più sicuro per accedere e gestire le istanze utilizzando un agente preinstallato senza il segreto archiviato. Inoltre, altri servizi AWS, ad esempio AWS Lambda,

consentono di configurare un ruolo del servizio IAM per concedere le autorizzazioni al servizio per eseguire operazioni AWS utilizzando credenziali temporanee. In situazioni in cui non è possibile usare credenziali temporanee, usa strumenti programmatici come AWS Secrets Manager, per automatizzare la rotazione e la gestione delle credenziali.

Verifica e ruota periodicamente le credenziali: La convalida periodica, preferibilmente tramite uno strumento automatizzato, è necessaria per verificare che vengano applicati i controlli corretti. Per le identità umane, è necessario richiedere agli utenti di modificare periodicamente le password e ritirare le chiavi di accesso a favore delle credenziali temporanee. Nella fase di passaggio da utenti IAM a identità centralizzate, puoi generare un report sulle credenziali per un audit degli utenti IAM. Ti consigliamo inoltre di monitorare le impostazioni MFA nel tuo provider di identità. È possibile configurare Regole di AWS Config per monitorare queste impostazioni. Per le identità di macchine, devi fare affidamento sulle credenziali temporanee utilizzando i ruoli IAM. Per situazioni in cui ciò non è possibile, è necessario eseguire audit frequenti e ruotare le chiavi di accesso.

Archivia e utilizza i segreti in modo sicuro: Per le credenziali non correlate a IAM e che non possono sfruttare le credenziali temporanee, ad esempio gli accessi al database, utilizza un servizio progettato per gestire i segreti, ad esempio Secrets Manager. Secrets Manager semplifica la gestione, la rotazione e l'archiviazione sicura delle chiavi segrete crittografate utilizzando i servizi supportati. Le chiamate per accedere ai segreti vengono registrate in AWS CloudTrail ai fini dell'audit e le autorizzazioni IAM possono concedere loro un accesso con privilegi minimi.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Implementazione di policy con privilegi minimi: assegna policy di accesso con privilegi minimi a gruppi e ruoli IAM in modo da rispecchiare il ruolo o la funzione dell'utente che hai definito.
  - Assegnare il privilegio minimo
- Rimozione delle autorizzazioni non necessarie: implementa il privilegio minimo rimuovendo le autorizzazioni superflue.
  - Riduzione dell'ambito di applicazione della policy mediante visualizzazione dell'attività dell'utente
  - Visualizzazione dell'accesso al ruolo
- Considerazioni sui limiti delle autorizzazioni: un limite delle autorizzazioni è una caratteristica avanzata per utilizzare una policy gestita che imposta il numero massimo di autorizzazioni che una policy basata su identità può concedere a un'entità IAM. Il limite delle autorizzazioni di un'entità le

permette di eseguire solo le operazioni consentite dalle policy basate su identità e dai limiti delle autorizzazioni.

  - Laboratorio: limiti delle autorizzazioni IAM per delegare la creazione di ruoli
- Tag di risorse per le autorizzazioni: puoi utilizzare i tag per controllare l'accesso alle risorse AWS che supportano il tagging. Puoi anche applicare tag a utenti e ruoli IAM per controllare a cosa possono accedere.
  - Laboratorio: controllo degli accessi basato su tag IAM per EC2
  - Controllo dell'accesso basato su attributi (Attribute-Based Access Control, ABAC)

Risorse

Documenti correlati:

- Nozioni di base su AWS Secrets Manager
- Best practice IAM
- Provider di identità e federazione
- Soluzioni dei partner per la sicurezza: accesso e controllo degli accessi
- Credenziali di sicurezza temporanee
- L'utente root dell'account AWS

Video correlati:

- Best practice per la gestione, il recupero e la rotazione di segreti su scala
- Gestione delle autorizzazioni utente su scala con AWS IAM Identity Center
- Mastering identity at every layer of the cake

SEC02-BP03 Archiviazione e utilizzo dei segreti in modo sicuro

Per la forza lavoro e le identità di macchine che richiedono segreti, come password per applicazioni di terze parti, archiviale con rotazione automatica utilizzando gli standard di settore più recenti in un servizio specializzato, mentre per le credenziali che non sono legate a IAM e che non possono sfruttare credenziali temporanee, come gli accessi al database, usa un servizio ideato per la gestione dei segreti, come AWS Secrets Manager. Secrets Manager semplifica la gestione, la rotazione e l'archiviazione sicura di segreti crittografati tramite servizi supportati. Le chiamate per accedere

ai segreti vengono registrate in AWS CloudTrail ai fini dell'audit e le autorizzazioni IAM possono concedere loro un accesso con privilegi minimi.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Uso di AWS Secrets Manager: AWS Secrets Manager è un servizio AWS che ti facilita il compito di gestire i segreti. I segreti possono essere credenziali di database, password, chiavi API di terze parti e persino testo arbitrario.

Risorse

Documenti correlati:

- Nozioni di base su AWS Secrets Manager
- Provider di identità e federazione

Video correlati:

- Best practice per la gestione, il recupero e la rotazione di segreti su scala

SEC02-BP04 Fai affidamento su un provider di identità centralizzato

Per le identità della forza lavoro, affidati a un provider di identità che ti consenta di gestire le identità in un luogo centralizzato. In questo modo è più semplice gestire l'accesso tra più applicazioni e servizi, perché crei, gestisci e revochi l'accesso da una singola posizione. Ad esempio, se qualcuno lascia la tua organizzazione, puoi revocare l'accesso per tutte le applicazioni e i servizi (incluso AWS) da un'unica posizione. Ciò riduce la necessità di molteplici credenziali e offre l'opportunità di integrarsi con i processi delle risorse umane esistenti.

Per la federazione con singoli account AWS, puoi utilizzare identità centralizzate per AWS con un provider basato su SAML 2.0 con AWS Identity and Access Management. Puoi utilizzare qualsiasi provider: in hosting su AWS, esterno ad AWS o fornito dalla AWS Partner, compatibile con il protocollo SAML 2.0 . Puoi utilizzare la federazione tra l'account AWS e il provider scelto per concedere a un utente o a un'applicazione l'accesso per chiamare le operazioni API AWS utilizzando un'asserzione SAML per ottenere le credenziali di sicurezza temporanee. È, inoltre, supportato il Single Sign-On basato sul Web, che consente agli utenti di accedere alla AWS Management Console dal portale di accesso.

Per la federazione a più account in AWS Organizations, puoi configurare l'origine di identità in AWS IAM Identity Center (IAM Identity Center)e specificare dove sono archiviati gli utenti e i gruppi. Una volta configurato, il provider di identità è la tua fonte di attendibilità e puoi sincronizzare le informazioni utilizzando il protocollo System for Cross-domain Identity Management (SCIM) v2.0. Puoi, quindi, cercare utenti o gruppi e concedere loro l'accesso IAM Identity Center ad account AWS, applicazioni cloud o entrambi.

IAM Identity Center si integra con AWS Organizations consentendoti di configurare il provider di identità una volta e quindi concedere l'accesso agli account nuovi e esistenti gestiti nella tua organizzazione. IAM Identity Center fornisce uno store predefinito che puoi utilizzare per gestire utenti e gruppi. Se scegli di utilizzare lo store IAM Identity Center, crea utenti e gruppi e assegna il loro livello di accesso agli account e alle applicazioni AWS, tenendo presente la best practice del privilegio minimo. In alternativa, puoi scegliere di connetterti al provider di identità esterno utilizzando SAML 2.0 o connetterti a Microsoft AD Directory utilizzando AWS Directory Service. Una volta configurate, puoi accedere alla AWS Management Console o all'app mobile AWS, eseguendo l'autenticazione tramite il tuo provider di identità centrale.

Per gestire gli utenti finali o i consumatori dei tuoi carichi di lavoro, ad esempio un'app per dispositivi mobili, puoi utilizzare Amazon Cognito. Ti consente di autenticare, autorizzare e gestire utenti per applicazioni Web e per dispositivi mobili. Gli utenti possono accedere direttamente con un nome utente e una password oppure tramite terze parti, ad esempio Amazon, Apple, Facebook o Google.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Centralizzazione dell'accesso amministrativo: crea un'entità di gestore dell'identità digitale Identity and Access Management (IAM) per stabilire una relazione consolidata tra l'Account AWS e il gestore dell'identità digitale (IdP). IAM supporta gli IdP compatibili con OpenID Connect (OIDC) o SAML 2.0 (Security Assertion Markup Language 2.0).
  - Provider di identità e federazione
- Centralizzazione dell'accesso alle applicazioni: considera Amazon Cognito per centralizzare l'accesso alle applicazioni. Si tratta di un servizio che consente di aggiungere il controllo di registrazione e accessi alle tue app Web e per dispositivi mobili in modo rapido e semplice. Amazon Cognito ricalibra le risorse per milioni di utenti e supporta l'accesso con gestori di identità social, come Facebook, Google e Amazon, e gestori di identità aziendali attraverso SAML 2.0.

- Rimozione di gruppi e utenti IAM obsoleti: dopo avere cominciato a utilizzare un gestore dell'identità digitale (IdP), rimuovi gli utenti e i gruppi IAM non più necessari.
  - Trovare credenziali non utilizzate
  - Eliminare un gruppo IAM

Risorse

Documenti correlati:

- Best Practice IAM
- Soluzioni dei partner per la sicurezza: accesso e controllo degli accessi
- Credenziali di sicurezza temporanee
- L'utente root dell'account AWS

Video correlati:

- Best practice per la gestione, il recupero e la rotazione di segreti su scala
- Gestione delle autorizzazioni utente su scala con AWS IAM Identity Center
- Mastering identity at every layer of the cake

SEC02-BP05 Verifica e rotazione periodica delle credenziali

Quando non puoi fare affidamento sulle credenziali temporanee e devi richiedere credenziali a lungo termine, verificale per assicurarti che siano applicati i controlli prestabiliti (ad esempio, la MFA), siano soggette regolarmente a rotazione e dispongano di un livello di accesso appropriato. La convalida periodica, preferibilmente tramite uno strumento automatizzato, è necessaria per verificare che vengano applicati i controlli corretti. Per le identità umane, è necessario richiedere agli utenti di modificare periodicamente le password e ritirare le chiavi di accesso a favore delle credenziali temporanee. Nella fase di passaggio da utenti AWS Identity and Access Management (IAM) a identità centralizzate, puoi generare un report sulle credenziali per un audit degli utenti IAM. Ti consigliamo inoltre di monitorare le impostazioni MFA nel tuo provider di identità. È possibile configurare Regole di AWS Config per monitorare queste impostazioni. Per le identità di macchine, devi fare affidamento sulle credenziali temporanee utilizzando i ruoli IAM. Per situazioni in cui ciò non è possibile, è necessario eseguire audit frequenti e ruotare le chiavi di accesso.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Esegui una verifica regolare delle credenziali: usa report di credenziali e Identify and Access Management (IAM) Access Analyzer per verificare le credenziali e le autorizzazioni IAM.
  - IAM Access Analyzer
  - Come ottenere un report sulle credenziali
  - Laboratorio: Pulizia automatica degli utenti IAM
- Usa i livelli di accesso per verificare le autorizzazioni IAM: per migliorare la sicurezza dell'account Account AWS, rivedi e monitora regolarmente ciascuna delle policy IAM, assicurandoti che concedano il privilegio minimo necessario per eseguire solo le operazioni indispensabili.
  - Utilizzo di livelli di accesso per rivedere le autorizzazioni IAM
- Prendi in considerazione l'automazione della creazione e degli aggiornamenti delle risorse IAM: AWS CloudFormation può essere utilizzato per automatizzare la distribuzione di risorse IAM, inclusi ruoli e policy, per ridurre gli errori umani, perché i modelli possono essere verificati e controllati a livello di versione.
  - Laboratorio: Distribuzione automatizzata di gruppi e ruoli IAM

Risorse

Documenti correlati:

- Nozioni di base su AWS Secrets Manager
- Best practice IAM
- Provider di identità e federazione
- Soluzioni dei partner per la sicurezza: accesso e controllo degli accessi
- Credenziali di sicurezza temporanee

Video correlati:

- Best practice per la gestione, il recupero e la rotazione di segreti su scala
- Gestione delle autorizzazioni utente su scala con AWS IAM Identity Center
- Mastering identity at every layer of the cake

SEC02-BP06 Utilizzo dei gruppi di utenti e degli attributi

Man mano che il numero di utenti gestiti cresce, sarà necessario determinare i modi per organizzarli in modo da poterli gestire su vasta scala. Inserisci gli utenti con requisiti di sicurezza comuni in gruppi definiti dal provider di identità e metti in atto meccanismi per garantire che gli attributi utente che potrebbero essere utilizzati per il controllo degli accessi (ad esempio, reparto o posizione) siano corretti e aggiornati. Utilizza questi gruppi e attributi, anziché i singoli utenti, per controllare l'accesso. In questo modo puoi gestire l'accesso centralmente, modificando una volta sola l'appartenenza o gli attributi di un gruppo utente con un set di autorizzazioni, anziché aggiornare numerose policy individuali quando le esigenze di accesso di un utente cambiano. Puoi utilizzare AWS IAM Identity Center (IAM Identity Center) per gestire gruppi di utenti e attributi. IAM Identity Center supporta la maggior parte degli attributi utilizzati, indipendentemente dal fatto che vengano inseriti manualmente durante la creazione dell'utente o assegnati automaticamente utilizzando un motore di sincronizzazione, come definito nella specifica System for Cross-Domain Identity Management (SCIM).

Inserisci gli utenti con requisiti di sicurezza comuni in gruppi definiti dal provider di identità e metti in atto meccanismi per garantire che gli attributi utente che potrebbero essere utilizzati per il controllo degli accessi (ad esempio, reparto o posizione) siano corretti e aggiornati. Utilizza questi gruppi e attributi, anziché i singoli utenti, per controllare l'accesso. In questo modo è possibile gestire l'accesso centralmente, modificando una volta sola l'appartenenza o gli attributi di un gruppo utente, anziché aggiornare numerose policy individuali quando le esigenze di accesso di un utente cambiano.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Se stai utilizzando AWS IAM Identity Center (IAM Identity Center), configura i gruppi: IAM Identity Center offre la possibilità di configurare gruppi di utenti e di assegnare ai gruppi il livello di autorizzazione desiderato.
  - AWS Single Sign-On - Gestione delle identità
- Scopri il controllo degli accessi basato su attributi (ABAC): ABAC è una strategia di autorizzazione che definisce i permessi in base agli attributi.
  - Che cos'è ABAC per AWS?
  - Laboratorio: controllo degli accessi basato su tag IAM per EC2

Risorse

Documenti correlati:

- Nozioni di base su AWS Secrets Manager
- Best practice IAM
- Provider di identità e federazione
- L'utente root dell'account AWS

Video correlati:

- Best practice per la gestione, il recupero e la rotazione di segreti su scala
- Gestione delle autorizzazioni utente su scala con AWS IAM Identity Center
- Mastering identity at every layer of the cake

Esempi correlati:

- Laboratorio: controllo degli accessi basato su tag IAM per EC2

SEC 3 Come gestisci le autorizzazioni per persone e macchine?

Gestisci le autorizzazioni per controllare l'accesso alle identità di persone e macchine che richiedono l'accesso ad AWS e al tuo carico di lavoro. Le autorizzazioni controllano chi può accedere a cosa e a quali condizioni.

Best practice

- SEC03-BP01 Definizione dei requisiti di accesso
- SEC03-BP02 Concessione dell'accesso con privilegio minimo
- SEC03-BP03 Determinazione di un processo per l'accesso di emergenza
- SEC03-BP04 Riduzione delle autorizzazioni in modo continuo
- SEC03-BP05 Definizione dei guardrail per le autorizzazioni dell'organizzazione
- SEC03-BP06 Gestione degli accessi in base al ciclo di vita
- SEC03-BP07 Analisi dell'accesso pubblico e multi-account
- SEC03-BP08 Condivisione delle risorse in modo sicuro

SEC03-BP01 Definizione dei requisiti di accesso

Ogni componente o risorsa del carico di lavoro deve essere accessibile da amministratori, utenti finali o altri componenti. Individua una definizione chiara di chi o cosa deve avere accesso a ciascun componente, quindi scegli il tipo di identità e il metodo di autenticazione e autorizzazione appropriati.

Anti-pattern comuni:

- Codifica fissa o archiviazione dei segreti nell'applicazione.
- Concessione di autorizzazioni personalizzate per ogni utente.
- Utilizzo di credenziali di lunga durata.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

Ogni componente o risorsa del carico di lavoro deve essere accessibile da amministratori, utenti finali o altri componenti. Individua una definizione chiara di chi o cosa deve avere accesso a ciascun componente, quindi scegli il tipo di identità e il metodo di autenticazione e autorizzazione appropriati.

L'accesso regolare agli Account AWS dell'organizzazione viene fornito utilizzando l'accesso federato o un gestore dell'identità centralizzato. Occorre anche centralizzare la gestione delle identità e garantire la presenza di una procedura consolidata per integrare l'accesso ad AWS nel ciclo di vita dell'accesso dei dipendenti. Ad esempio, quando un dipendente passa a un ruolo lavorativo con un livello di accesso diverso, anche la sua appartenenza al gruppo deve cambiare per riflettere i nuovi requisiti di accesso.

Quando si definiscono i requisiti di accesso per le identità non umane, determina quali applicazioni e componenti devono accedere e come vengono concesse le autorizzazioni. L'utilizzo di ruoli IAM creati con il modello di accesso con privilegi minimi è un approccio consigliato. Le policy gestite da AWS forniscono le policy IAM predefinite che coprono la maggior parte dei casi d'uso comuni.

I servizi AWS, come AWS Secrets Manager e Archivio dei parametri AWS Systems Managerti consentono di scollegare i segreti dall'applicazione o dal carico di lavoro in modo sicuro nei casi in cui non è possibile utilizzare i ruoli IAM. In Secrets Manager puoi adottare la rotazione automatica delle credenziali. Puoi usare Systems Manager per fare riferimento a parametri negli script, comandi, documenti SSM, configurazione e flussi di lavoro di automazione utilizzando il nome univoco specificato al momento della creazione del parametro.

Puoi usare AWS Identity and Access Management Roles Anywhere per ottenere credenziali di sicurezza temporanee in IAM per i carichi di lavoro eseguiti all'esterno di AWS. I tuoi carichi di lavoro possono utilizzare le stesse policy IAM e ruoli IAM che usi con le applicazioni AWS per accedere alle risorse AWS.

Ove possibile, prediligi le credenziali temporanee a breve termine rispetto a quelle statiche a lungo termine. Per gli scenari in cui gli utenti IAM devono avere l'accesso programmatico e credenziali a lungo termine, utilizza le ultime informazioni usate per la chiave di accesso per ruotare e rimuovere le chiavi di accesso.

Risorse

Documenti correlati:

- Il controllo dell'accesso basato su attributi (Attribute-Based Access Control, ABAC)
- AWS IAM Identity Center
- IAM Roles Anywhere
- AWS Managed policies for IAM Identity Center (Policy gestite da AWS per IAM Identity Center)
- AWS IAM policy conditions (Condizioni delle policy AWS IAM)
- Casi d'uso IAM
- Rimozione di credenziali non necessarie
- Lavorare con le policy
- How to control access to AWS resources based on Account AWS, OU, or organization (Come controllare l'accesso alle risorse AWS in base all'account, all'unità organizzativa o all'organizzazione AWS)
- Identify, arrange, and manage secrets easily using enhanced search in AWS Secrets Manager (Identificazione, organizzazione e gestione semplificate dei segreti con la ricerca avanzata di AWS Secrets Manager)

Video correlati:

- Become an IAM Policy Master in 60 Minutes or Less (Diventa un IAM Policy Master in 60 minuti)
- Separation of Duties, Least Privilege, Delegation, and CI/CD (Separazione dei compiti, privilegio minimo, delega e integrazione e distribuzione continua (CI/CD)
- Streamlining identity and access management for innovation (Semplificazione della gestione delle identità e degli accessi per l'innovazione)

SEC03-BP02 Concessione dell'accesso con privilegio minimo

Concedi alle identità soltanto il livello di accesso di cui hanno bisogno, specificando le operazioni che possono effettuare, le risorse AWS su cui possono operare e a quali condizioni. Affidati ai gruppi e agli attributi di identità per impostare dinamicamente le autorizzazioni su vasta scala, anziché definire le autorizzazioni per i singoli utenti. Ad esempio, puoi concedere a un gruppo di sviluppatori le autorizzazioni per gestire solo le risorse del loro progetto. In questo modo, quando uno sviluppatore lascia il gruppo, perderà l'accesso a tutte le risorse gestite tramite il gruppo e non sarà necessario modificare le policy di accesso.

Anti-pattern comuni:

- L'impostazione predefinita è la concessione delle autorizzazioni di amministratore agli utenti.
- L'utilizzo dell'account root per le attività quotidiane.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

Stabilire un principio di privilegio minimo assicura che alle identità venga concesso di eseguire il minimo set di funzioni necessarie alla realizzazione di un'attività specifica, bilanciando al tempo stesso usabilità ed efficienza. Il funzionamento di questo principio limita l'accesso involontario e ti consente di verificare chi ha accesso a quali risorse. In AWS per impostazione predefinita le identità non dispongono di autorizzazioni ad eccezione dell'utente root. Le credenziali per l'utente root devono essere rigorosamente controllate e utilizzate solo per poche attività specifiche.

Puoi utilizzare le policy per concedere esplicitamente autorizzazioni collegate a IAM o a entità di risorse, ad esempio un ruolo IAM utilizzato da identità federate, macchine o risorse, come un bucket S3. Quando crei e colleghi una policy, puoi specificare le azioni del servizio, le risorse e le condizioni che devono essere vere affinché AWS consenta l'accesso. AWS supporta una varietà di condizioni che contribuiscono a ridurre l'accesso. Ad esempio, utilizzando `PrincipalOrgID, una` chiave di condizione, l'identificatore di AWS Organizations viene verificato in modo che l'accesso possa essere concesso all'interno della tua organizzazione AWS.

Puoi anche controllare le richieste effettuate dai servizi AWS per tuo conto, ad esempio AWS CloudFormation per la creazione di una funzione AWS Lambda, utilizzando la chiave di condizione `CalledVia` . Occorre suddividere in livelli i diversi tipi di policy per limitare efficacemente le autorizzazioni complessive di un account. Ad esempio, puoi consentire ai team dell'applicazione di

creare le proprie policy IAM usando un limite delle autorizzazioni per contenere il numero massimo di autorizzazioni che possono concedere.

Sono disponibili diverse funzionalità AWS per consentirti di applicare su vasta scala la gestione delle autorizzazioni e aderire al principio del privilegio minimo. Il controllo degli accessi basato su attributi ti consente di limitare le autorizzazioni in base al tag di una risorsa, per prendere le decisioni di autorizzazione in base ai tag applicati alla risorsa e al principale IAM chiamante. Ti consente di combinare nella tua policy tag e autorizzazioni per ottenere l'accesso granulare alle risorse senza la necessità di creare molte policy personalizzate.

Un altro modo per accelerare la creazione di una policy con privilegi minimi consiste nel basare la policy sulle autorizzazioni CloudTrail dopo l'esecuzione di un'attività. Il Sistema di analisi degli accessi IAM può generare automaticamente una policy IAM basata su attività. È inoltre possibile utilizzare IAM Access Advisor a livello di organizzazione o singolo account per monitorare le ultime informazioni a cui si accede per una determinata policy.

Stabilisci una cadenza di revisione di questi dettagli e rimozione delle autorizzazioni non necessarie. Occorre impostare dei guardrail di autorizzazione all'interno della tua organizzazione AWS per controllare il numero massimo di autorizzazioni di qualsiasi account membro. I servizi come AWS Control Tower dispongono di controlli preventivi gestiti prescrittivi che ti permettono di definire i tuoi controlli.

Risorse

Documenti correlati:

- Limiti delle autorizzazioni per le entità IAM
- Techniques for writing least privilege IAM policies (Tecniche per la scrittura di policy IAM con privilegio minimo)
- IAM Access Analyzer makes it easier to implement least privilege permissions by generating IAM policies based on access activity (IAM Access Analyzer semplifica l'implementazione delle autorizzazioni con privilegio minimo generando policy IAM basate sull'attività di accesso)
- Perfezionamento delle autorizzazioni in AWS utilizzando le informazioni sull'ultimo accesso
- IAM policy types and when to use them (Tipi di policy IAM e quando utilizzarle)
- Test delle policy IAM con il simulatore di policy IAM
- Guardrails in AWS Control Tower (Guardrail in AWS Control Tower)
- Zero Trust architectures: An AWS perspective (Architetture Zero Trust: una prospettiva AWS)

- How to implement the principle of least privilege with CloudFormation StackSets (Come implementare il principio del privilegio minimo con CloudFormation StackSets)

Video correlati:

- Next-generation permissions management (Gestione delle autorizzazioni di ultima generazione)
- Zero Trust: An AWS perspective (Zero Trust: una prospettiva AWS)
- How can I use permissions boundaries to limit IAM users and roles to prevent privilege escalation? (Come utilizzare i limiti delle autorizzazioni per limitare utenti e ruoli IAM e impedire l'escalation dei privilegi?)

Esempi correlati:

- Laboratorio: limiti delle autorizzazioni IAM per delegare la creazione di ruoli

SEC03-BP03 Determinazione di un processo per l'accesso di emergenza

Un processo che consente l'accesso di emergenza al carico di lavoro nell'improbabile caso di un problema a un processo automatizzato o a una pipeline. Questo consente di utilizzare criteri di accesso con privilegi minimi, ma garantisce che gli utenti possano ottenere il corretto livello di accesso quando ne hanno bisogno. Ad esempio, puoi definire un processo per gli amministratori per verificare e approvare la loro richiesta, ad esempio un ruolo AWS tra account di emergenza per l'accesso o un processo specifico che gli amministratori devono seguire per convalidare e approvare una richiesta di emergenza.

Anti-pattern comuni:

- La mancanza di un processo di emergenza per il ripristino da interruzione con la configurazione dell'identità esistente.
- La concessione di autorizzazioni con privilegi elevati a lungo termine per la risoluzione dei problemi o per scopi di ripristino.

Livello di rischio associato se questa best practice non fosse adottata: medio

Guida all'implementazione

L'applicazione dell'accesso di emergenza può assumere diverse forme per le quali occorre essere preparati. Il primo è un errore del gestore dell'identità principale. In questo caso, è necessario impiegare un secondo metodo di accesso con le autorizzazioni necessarie per il ripristino che potrebbe essere eseguito da un gestore dell'identità di backup o un utente IAM. Questo secondo metodo deve essere rigorosamente controllato, monitorato e notificato nel caso venga utilizzato. L'identità di accesso di emergenza deve provenire da un account specifico per questo scopo e disporre solo delle autorizzazioni per assumere un ruolo appositamente progettato per il ripristino.

Occorre anche essere preparati per l'accesso di emergenza in cui è necessario un accesso amministrativo temporaneo con privilegi elevati. Uno scenario comune consiste nel limitare le autorizzazioni mutevoli a un processo automatizzato utilizzato per l'implementazione delle modifiche. Nel caso in cui questo processo riscontri un problema, gli utenti potrebbero avere la necessità di autorizzazioni con privilegi elevati per ripristinare la funzionalità. In questa situazione, stabilisci un processo in cui gli utenti possono richiedere l'accesso con privilegi elevati e gli amministratori possono convalidarlo e approvarlo. I piani di implementazione che includono i dettagli delle best practice per la preassegnazione dell'accesso e la creazione di ruoli di emergenza di tipo break-glass sono forniti come parte di SEC10-BP05 Preassegnazione dell'accesso.

Risorse

Documenti correlati:

- Monitor and Notify on AWS (Monitoraggio e invio di notifiche in AWS)
- Managing temporary elevated access (Gestione dell'accesso temporaneo con privilegi elevati)

Video correlato:

- Become an IAM Policy Master in 60 Minutes or Less (Diventa un IAM Policy Master in 60 minuti)

SEC03-BP04 Riduzione delle autorizzazioni in modo continuo

Man mano che i team e i carichi di lavoro determinano l'accesso di cui hanno bisogno, rimuovi le autorizzazioni che non utilizzano più e stabilisci processi di revisione per applicare le autorizzazioni con privilegi minimi. Monitora e riduci continuamente le identità e le autorizzazioni non utilizzate.

A volte, quando i team e i progetti stanno per iniziare, puoi scegliere di concedere un accesso estensivo (in un ambente di sviluppo o di test) per promuovere innovazione e agilità. Ti suggeriamo

di valutare gli accessi con regolarità e, soprattutto in un ambiente di produzione, di limitare l'accesso solo alle autorizzazioni richieste e di ottenere il privilegio minimo. AWS offre funzionalità di analisi degli accessi per identificare accessi inutilizzati. Per identificare gli utenti, i ruoli, le autorizzazioni e le credenziali inutilizzati, AWS analizza le attività di accesso e fornisce informazioni sull'ultimo ruolo e chiave di accesso utilizzati. Puoi utilizzare il timestamp dell'ultimo accesso a identificare utenti e ruoli inutilizzatie rimuoverli. Inoltre, puoi rivedere le informazioni sull'ultimo accesso al servizio e sull'ultima azione per identificare e restringere le autorizzazioni per specifici utenti e ruoli. Ad esempio, puoi utilizzare le informazioni sull'ultimo accesso per identificare le operazioni Amazon Simple Storage Service(Amazon S3) specifiche richieste dal ruolo dell'applicazione e limitare l'accesso solo a quelle. Queste funzionalità sono disponibili nella AWS Management Console e a livello di programmazione per consentirti di incorporarle nei flussi di lavoro dell'infrastruttura e negli strumenti automatizzati.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Configurazione di AWS Identify and Access Management (IAM) Access Analyzer: AWS IAM Access Analyzer identifica le risorse dell'organizzazione e gli account, ad esempio i bucket Amazon Simple Storage Service (Amazon S3) o i ruoli IAM, che sono condivisi con un'entità esterna.
  - AWS IAM Access Analyzer

Risorse

Documenti correlati:

- Controllo dell'accesso basato su attributi (Attribute-Based Access Control, ABAC)
- Assegnare il privilegio minimo
- Rimozione di credenziali non necessarie
- Lavorare con le policy

Video correlati:

- Become an IAM Policy Master in 60 Minutes or Less (Diventa un IAM Policy Master in 60 minuti)
- Separation of Duties, Least Privilege, Delegation, and CI/CD (Separazione dei compiti, privilegio minimo, delega e integrazione e implementazione continua CI/CD)

SEC03-BP05 Definizione dei guardrail per le autorizzazioni dell'organizzazione

Stabilisci controlli comuni che limitano l'accesso a tutte le identità nella tua organizzazione. Ad esempio, puoi limitare l'accesso a Regioni AWS specifiche o impedire agli operatori di eliminare risorse comuni, come ad esempio un ruolo IAM utilizzato per il team di sicurezza centrale.

Anti-pattern comuni:

- Esecuzione di carichi di lavoro nell'account di amministratore dell'organizzazione.
- Esecuzione di carichi di lavoro di produzione e non di produzione nello stesso account.

Livello di rischio associato se questa best practice non fosse adottata: medio

Guida all'implementazione

Man mano che aumenti e gestisci carichi di lavoro aggiuntivi in AWS, devi separarli utilizzando gli account e gestire questi ultimi utilizzando AWS Organizations. Ti consigliamo di stabilire limiti di autorizzazione comuni che limitano l'accesso a tutte le identità nella tua organizzazione. Ad esempio, puoi limitare l'accesso a Regioni AWS specifiche o impedire al tuo team di eliminare risorse comuni, come ad esempio un ruolo IAM utilizzato dal team di sicurezza centrale.

Puoi iniziare implementando delle policy di controllo dei servizi di esempio, come una policy che impedisce agli utenti di disabilitare i servizi chiave. Le policy di controllo dei servizi utilizzano il linguaggio di policy IAM e consentono di stabilire i controlli a cui aderiscono tutti i principali IAM (utenti e ruoli). Puoi limitare l'accesso a specifiche azioni del servizio, risorse e in base a condizioni specifiche per soddisfare le esigenze di controllo degli accessi della tua organizzazione. Se necessario, puoi definire eccezioni ai limiti definiti. Ad esempio, puoi limitare le azioni del servizio per tutte le entità IAM nell'account tranne per un ruolo amministratore specifico.

Ti consigliamo di evitare di eseguire carichi di lavoro nell'account di gestione. L'account di gestione deve essere utilizzato per governare e distribuire i guardrail di sicurezza che influiscono sugli account membri. Alcuni servizi AWS supportano l'uso di un account amministratore delegato. Se è disponibile, devi utilizzare questo account delegato anziché l'account di gestione. È necessario limitare scrupolosamente l'accesso all'account dell'amministratore dell'organizzazione.

L'utilizzo di una strategia multi-account ti consente di avere una maggiore flessibilità nell'applicazione di guardrail ai tuoi carichi di lavoro. L'architettura di riferimento per la sicurezza AWS fornisce le indicazioni prescrittive su come progettare la struttura del tuo account. I servizi AWS come AWS Control Tower forniscono le funzionalità per gestire centralmente i controlli preventivi e investigativi

all'interno dell'organizzazione. Definisci uno scopo chiaro per ogni account o unità organizzativa all'interno della tua organizzazione e limita i controlli in linea con tale scopo.

Risorse

Documenti correlati:

- AWS Organizations
- Policy di controllo dei servizi (Service Control Policies, SCP)
- Get more out of service control policies in a multi-account environment (Ottieni di più dalle policy di controllo dei servizi in un ambiente multi-account)
- AWS Security Reference Architecture (AWS SRA) (Architettura di riferimento per la sicurezza AWS (AWS SRA))

Video correlati:

- Enforce Preventive Guardrails using Service Control Policies (Applicazione di guardrail preventivi utilizzando le policy di controllo dei servizi)
- Building governance at scale with AWS Control Tower (Creazione di una governance su vasta scala con AWS Control Tower)
- AWS Identity and Access Management deep dive (Approfondimenti su AWS Identity and Access Management)

SEC03-BP06 Gestione degli accessi in base al ciclo di vita

Integra i controlli degli accessi con il ciclo di vita degli operatori e delle applicazioni e con il tuo provider di federazione centralizzata. Ad esempio, rimuovi l'accesso di un utente quando lascia l'organizzazione o cambia ruolo.

Quando gestisci i carichi di lavoro utilizzando account separati, in alcuni casi sarà necessario condividere le risorse tra tali account. Ti consigliamo di condividere le risorse utilizzando AWS Resource Access Manager (AWS RAM). Questo servizio ti consente di condividere in modo semplice e sicuro le risorse AWS all'interno della tua organizzazione AWS Organizations e delle unità organizzative. Con AWS RAM, l'accesso alle risorse condivise viene automaticamente concesso o revocato quando gli account vengono spostati da e verso l'organizzazione o l'unità organizzativa con cui sono condivisi. In questo modo puoi garantire che le risorse vengano condivise solo con gli account desiderati.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Ciclo di vita degli accessi utente: implementa una policy per il ciclo di vita degli accessi utente per i nuovi entranti, le modifiche alle funzioni lavorative e gli uscenti per garantire l'accesso solo agli utenti attuali.

Risorse

Documenti correlati:

- Controllo dell'accesso basato su attributi (Attribute-Based Access Control, ABAC)
- Assegnare il privilegio minimo
- IAM Access Analyzer
- Rimozione di credenziali non necessarie
- Lavorare con le policy

Video correlati:

- Become an IAM Policy Master in 60 Minutes or Less (Diventa un IAM Policy Master in 60 minuti)
- Separation of Duties, Least Privilege, Delegation, and CI/CD (Separazione dei compiti, privilegio minimo, delega e integrazione e distribuzione continua (CI/CD)

SEC03-BP07 Analisi dell'accesso pubblico e multi-account

Monitora continuamente i risultati che evidenziano l'accesso pubblico e multi-account. Limita l'accesso pubblico e multi-account alle risorse che ne hanno bisogno.

Anti-pattern comuni:

- La mancanza di un processo per governare l'accesso multi-account e l'accesso pubblico alle risorse.

Livello di rischio associato se questa best practice non fosse adottata: basso

Guida all'implementazione

In AWS puoi concedere l'accesso a risorse in un altro account. Concedi l'accesso diretto multi-account utilizzando le policy collegate alle risorse, ad esempio le policy di bucket Amazon Simple Storage Service (Amazon S3), o consentendo a un'identità di assumere un ruolo IAM in un altro account. Quando si utilizzano le policy di risorse, verifica che l'accesso sia concesso alle identità dell'organizzazione e rendi pubbliche le risorse. Definisci un processo per approvare tutte le risorse che devono essere pubblicamente disponibili.

IAM Access Analyzer usa la sicurezza comprovabile per identificare tutti i percorsi di accesso a una risorsa dall'esterno del proprio account. Esamina continuamente le policy delle risorse e segnala i risultati dell'accesso pubblico e tra account per semplificare l'analisi di accessi potenzialmente estensivi. Prendi in considerazione la configurazione di IAM Access Analyzer con AWS Organizations per verificare di avere visibilità su tutti i tuoi account. IAM Access Analyzer ti permette anche di visualizzare in anteprima i risultati del sistema di analisi degli accessiprima di implementare le autorizzazioni delle risorse. Questo consente di convalidare che le modifiche alla policy concedono solo l'accesso multi-account e pubblico autorizzati alle risorse. Quando si progetta per l'accesso a più account, le policy di attendibilità consentono di controllare in quali casi è possibile assumere un ruolo. Ad esempio, puoi limitare l'assunzione del ruolo a un particolare intervallo di IP di origine.

Puoi anche utilizzare AWS Config per segnalare e correggere le risorse per qualsiasi configurazione di accesso pubblico accidentale, tramite i controlli delle policy AWS Config. Servizi come AWS Control Tower e AWS Security Hub semplificano l'implementazione di controlli e guardrail in AWS Organizations per identificare e correggere le risorse pubblicamente esposte. Ad esempio, AWS Control Tower ha un guardrail gestito in grado di rilevare se gli snapshot Amazon EBS sono ripristinabili da tutti gli account AWS.

Risorse

Documenti correlati:

- Utilizzo di AWS Identity and Access Management Access Analyzer
- Guardrails in AWS Control Tower (Guardrail in AWS Control Tower)
- AWS Foundational Security Best Practices standard (Standard delle best practice per la sicurezza di base di AWS)
- AWS Config Managed Rules (Regole gestite di AWS Config)
- Riferimento dei controlli AWS Trusted Advisor

Video correlati:

- Best Practices for securing your multi-account environment (Best practice per la protezione dell'ambiente multi-account)
- Dive Deep into IAM Access Analyzer (Approfondimenti sul Sistema di analisi degli accessi IAM)

SEC03-BP08 Condivisione delle risorse in modo sicuro

Regola il consumo di risorse condivise tra account diversi o all'interno della tua AWS Organizations. Monitora le risorse condivise e rivedi l'accesso alle stesse.

Anti-pattern comuni:

- Utilizzo della policy di attendibilità IAM predefinita quando si concede l'accesso multi-account di terze parti.

Livello di rischio associato se questa best practice non fosse adottata: basso

Guida all'implementazione

Quando gestisci i tuoi carichi di lavoro utilizzando più account AWS, potrebbe essere necessario condividere le risorse tra gli account. Molto spesso si tratta della condivisione tra account all'interno di AWS Organizations. Diversi servizi AWS come AWS Security Hub, Amazon GuardDutye AWS Backup dispongono di funzionalità per più account integrate in Organizations. Puoi utilizzare AWS Resource Access Manager per condividere altre risorse comuni, come i collegamenti del gateway di transito alla VPN o le sottoreti VPC, AWS Network Firewallo le pipeline di Amazon SageMaker Runtime. Se vuoi assicurarti che il tuo account condivida solo risorse all'interno di Organizations, ti consigliamo di utilizzare le policy di controllo dei servizi per impedire l'accesso a principali esterni.

Quando condividi le risorse, devi mettere in atto le misure per proteggerti da accessi non intenzionali. Ti consigliamo di combinare i controlli basati sull'identità e i controlli di rete per creare un perimetro di dati per la tua organizzazione. Questi controlli devono porre limiti rigorosi a quali risorse possono essere condivise e impedire la condivisione o l'esposizione non consentite delle risorse. Ad esempio, come parte del tuo perimetro di dati puoi utilizzare le policy dell'endpoint VPC e la condizione `aws:PrincipalOrgId` per garantire che le identità che accedono ai bucket Amazon S3 appartengano all'organizzazione.

In alcuni casi, potresti voler consentire la condivisione di risorse al di fuori di Organizations o concedere a terze parti l'accesso al tuo account. Ad esempio, un partner può fornire una soluzione

di monitoraggio che deve accedere alle risorse del tuo account. In questi casi, devi creare un ruolo tra più account IAM con i soli privilegi necessari alla terza parte. Dovresti anche creare una policy di attendibilità usando la condizione ID esterno. Quando si utilizza un ID esterno, è necessario generare un ID univoco per ciascuna terza parte. L'ID univoco non deve essere fornito o controllato da terzi. Se la terza parte non ha più bisogno di accedere al tuo ambiente, occorre rimuovere il ruolo. In ogni caso, devi evitare di fornire a terze parti credenziali IAM a lungo termine. Tieni presente gli altri servizi AWS che supportano in modo nativo la condivisione. Ad esempio AWS Well-Architected Tool consente la condivisione di un carico di lavoro con altri account AWS.

Quando si usa un servizio come Amazon S3, si consiglia di disabilitare le liste di controllo degli accessi (ACL) per il bucket Amazon S3 e usare le policy IAM per definire il controllo degli accessi. Per limitare l'accesso a un'origine Amazon S3 da Amazon CloudFrontmigra dall'identità di accesso origine (OAI) al controllo degli accessi di origine (OAC) che supporta funzionalità aggiuntive tra cui la crittografia lato server con AWS KMS.

Risorse

Documenti correlati:

- Il proprietario del bucket concede autorizzazioni multi-account per gli oggetti che non sono di sua proprietà
- How to use Trust Policies with IAM (Come utilizzare le policy di attendibilità con IAM)
- Building Data Perimeter on AWS (Creazione del perimetro dei dati in AWS)
- How to use an external ID when granting a third party access to your AWS resources (Come utilizzare un ID esterno quando si concede a una terza parte l'accesso alle risorse AWS)

Video correlati:

- Granular Access with AWS Resource Access Manager (Accesso granulare con Gestione degli accessi alle risorse AWS)
- Securing your data perimeter with VPC endpoints (Protezione del perimetro dei dati con gli endpoint VPC)
- Establishing a data perimeter on AWS (Applicazione di un perimetro dei dati in AWS)

Rilevamento

Domanda

SEC 4 In che modo individui ed esamini gli eventi di sicurezza?

Acquisisci ed analizza gli eventi a partire da log e parametri per acquistare visibilità. Agisci su eventi di sicurezza e potenziali minacce per contribuire a rendere sicuro il carico di lavoro.

Best practice

- SEC04-BP01 Configurazione dei registri di servizi e applicazioni
- SEC04-BP02 Analisi di log, risultati e parametri a livello centrale
- SEC04-BP03 Automazione delle risposte agli eventi
- SEC04-BP04 Implementazione di eventi di sicurezza fruibili

SEC04-BP01 Configurazione dei registri di servizi e applicazioni

Configura i registri per tutto il carico di lavoro, inclusi registri di applicazioni, di risorse e di servizi AWS. Ad esempio, assicurati che AWS CloudTrail, Amazon CloudWatch Logs, Amazon GuardDuty e AWS Security Hub siano abilitati per tutti gli account all'interno della tua organizzazione.

Una pratica di base è quella di stabilire un set di meccanismi di rilevamento a livello di account. Questo set di meccanismi di base ha lo scopo di registrare e rilevare un'ampia gamma di operazioni su tutte le risorse nel tuo account. Tali meccanismi consentono di creare una funzionalità di rilevamento completa con opzioni che includono la correzione automatizzata e integrazioni dei partner per renderla ancora più funzionale.

In AWS, i servizi che possono implementare questo set di base includono:

- AWS CloudTrail fornisce uno storico degli eventi delle attività del tuo account AWS, incluse le operazioni eseguite dalla AWS Management Console, gli SDK AWS, gli strumenti a riga di comando e altri servizi AWS.
- AWS Config monitora e registra le configurazioni delle risorse AWS e consente di automatizzare la valutazione e la correzione rispetto alle configurazioni desiderate.
- Amazon GuardDuty è un servizio di rilevamento delle minacce che esegue un monitoraggio costante per individuare attività dannose e comportamenti non autorizzati al fine di proteggere i tuoi carichi di lavoro e account AWS.

- AWS Security Hub offre un unico punto di aggregazione, organizzazione e assegnazione di priorità per gli avvisi di sicurezza o i risultati provenienti da diversi servizi AWS e da prodotti opzionali di terze parti per fornire una panoramica completa degli avvisi di sicurezza e dello stato di conformità.

Partendo dalla base esistente a livello di account, molti servizi AWS principali, ad esempio Amazon Virtual Private Cloud Console (Amazon VPC), forniscono funzionalità di registrazione a livello di servizio. Registri di flusso Amazon VPC consentono di acquisire informazioni sul traffico IP da e verso le interfacce di rete che possono fornire approfondimenti preziosi sulla cronologia della connettività e attivare azioni automatizzate in base a comportamenti anomali.

Per le istanze Amazon Elastic Compute Cloud (Amazon EC2) e la registrazione basata su applicazioni che non proviene dai servizi AWS, i registri possono essere archiviati e analizzati utilizzando Amazon CloudWatch Logs. Un' agente raccoglie i log dal sistema operativo e dalle applicazioni in esecuzione e li archivia automaticamente. Quando i log sono disponibili in CloudWatch Logs, puoi elaborarli in tempo realeo analizzarli utilizzando CloudWatch Logs Insights.

Oltre alla raccolta e all'aggregazione dei log, è altrettanto importante la capacità di estrarre informazioni significative dai grandi volumi di dati di log ed eventi generati da architetture complesse. Consulta la scheda Monitoraggio del Whitepaper sul principio dell'affidabilità per maggiori dettagli. I log stessi possono contenere dati considerati sensibili, sia quando i dati dell'applicazione sono stati erroneamente inseriti nei file di log acquisiti dall'agente di CloudWatch Logs, sia quando la registrazione tra regioni è configurata per l'aggregazione dei log e vi sono considerazioni legislative sulla spedizione di determinati tipi di informazioni oltre confine.

Un approccio consiste nell'utilizzare le funzioni AWS Lambda, attivate su eventi quando vengono distribuiti i registri, per filtrare e redigere i dati di registro prima di inoltrarli a una posizione di registrazione centrale, ad esempio un bucket Amazon Simple Storage Service (Amazon S3). I registri non redatti possono essere conservati in un bucket locale fino a quando non è trascorso un "periodo di tempo ragionevole" (secondo quanto stabilito dalla legislazione e dal team legale) e a quel punto una regola del ciclo di vita di Amazon S3 può eliminarli automaticamente. Si possono proteggere ulteriormente i log in Amazon S3 utilizzando Amazon S3 Object Lock, dove è possibile archiviare oggetti utilizzando un modello WORM (Write Once Read Many).

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Abilitazione della registrazione dei servizi AWS: abilita la registrazione dei servizi AWS per soddisfare i tuoi requisiti. Le funzionalità di registrazione includono quanto segue: registri di flussi

Amazon VPC, registri Elastic Load Balancing (ELB), registri di bucket Amazon S3, registri di accesso CloudFront, registri di query Amazon Route 53 e registri Amazon Relational Database Service (Amazon RDS).

  - AWS Answers: capacità native di registrazioni di sicurezza AWS
- Valuta e abilita la registrazione di sistemi operativi e log specifici per l'applicazione, così da rilevare eventuali comportamenti sospetti.
  - Nozioni di base su CloudWatch Logs
  - Strumenti per gli sviluppatori e analisi dei registri
- Applicazione di controlli adeguati ai registri: i registri contengono informazioni sensibili e solo gli utenti autorizzati devono avere accesso a tali dati. Considera la possibilità di limitare le autorizzazioni per i bucket Amazon S3 e i gruppi di logCloudWatch Logs.
  - Autenticazione e controllo degli accessi per Amazon CloudWatch
  - Gestione di identità e accessi in Amazon S3
- Configurazione Amazon GuardDuty: GuardDuty è un servizio di rilevamento delle minacce che esegue un monitoraggio costante per individuare attività dannose e comportamenti non autorizzati al fine di proteggere i tuoi carichi di lavoro e Account AWS. Abilita GuardDuty e configura gli avvisi automatici per e-mail utilizzando il laboratorio.
- Configurazione di un percorso personalizzato in CloudTrail: la configurazione di un percorso permette di memorizzare registri per un tempo maggiore del periodo predefinito e analizzarli in un secondo momento.
- Abilita AWS Config: AWS Config fornisce una visualizzazione dettagliata della configurazione delle risorse AWS nel tuo Account AWS. In essa sono inclusi il modo in cui le risorse sono correlate tra loro e il modo in cui erano configurate in precedenza, in modo da poter vedere il cambiamento di configurazioni e relazioni nel corso del tempo.
- Abilita AWS Security Hub: Security Hub offre una panoramica generale del tuo assetto di sicurezza in AWS e ti aiuta a verificare la conformità rispetto a standard di sicurezza di settore e best practice. Security Hub raccoglie dati sulla sicurezza da Account AWS, servizi e prodotti di partner di terze parti e aiuta ad analizzare i trend di sicurezza e a identificare le problematiche di sicurezza con priorità maggiore.

Risorse

Documenti correlati:

- Amazon CloudWatch

- Amazon EventBridge
- Nozioni di base: Amazon CloudWatch Logs
- Soluzione dei partner per la sicurezza: registrazione e monitoraggio

Video correlati:

- Centrally Monitoring Resource Configuration & Compliance (Monitoraggio centrale della configurazione e della conformità delle risorse)
- Remediating Amazon GuardDuty and AWS Security Hub Findings (Correzione Amazon GuardDuty e risultati AWS Security Hub)
- Threat management in the cloud: Amazon GuardDuty and AWS Security Hub (Gestione delle minacce nel cloud: Amazon GuardDuty e AWS Security Hub)

Esempi correlati:

- Laboratorio: Implementazione automatizzata di controlli di rilevamento

SEC04-BP02 Analisi di log, risultati e parametri a livello centrale

i team delle operazioni di sicurezza confidano nella raccolta di log e nell'utilizzo di strumenti di ricerca per scoprire potenziali eventi di interesse, che potrebbero indicare attività non autorizzate o modifiche involontarie. Tuttavia, la semplice analisi dei dati raccolti e l'elaborazione manuale delle informazioni non sono sufficienti per tenere il passo con il volume di informazioni provenienti da architetture complesse. Le sole analisi e i soli resoconti non facilitano l'assegnazione delle risorse giuste per lavorare a un evento in modo adeguato e nei tempi giusti.

Una best practice per creare un team per le operazioni di sicurezza preparato è integrare profondamente il flusso degli eventi di sicurezza e le scoperte in un sistema di notifica e flusso di lavoro, come un sistema di ticketing, un sistema di bug o altri sistemi riguardanti le informazioni di sicurezza o la gestione degli eventi (SIEM). Ciò elimina il flusso di lavoro da e-mail e report statici e consente di instradare, inoltrare e gestire eventi o risultati. Molte organizzazioni integrano anche gli avvisi di sicurezza nelle loro piattaforme di chat, collaborazione e di produttività per sviluppatori. Per le aziende che intraprendono la strada dell'automazione, un sistema di ticketing basato su API a bassa latenza offre una notevole flessibilità quando si pianifica "cosa automatizzare prima".

Questa best practice si applica non solo agli eventi di sicurezza generati dai messaggi di log che illustrano l'attività degli utenti o gli eventi di rete, ma anche a quelli generati dalle modifiche rilevate

nell'infrastruttura stessa. La possibilità di rilevare le modifiche, determinare se una modifica è appropriata e quindi instradare tali informazioni al flusso di lavoro di correzione adatto è essenziale per mantenere e convalidare un'architettura sicura, in un contesto di modifiche difficili da individuare come indesiderabili per impedirne l'esecuzione tramite una combinazione di configurazioni AWS Identity and Access Management(IAM) e AWS Organizations.

Amazon GuardDuty e AWS Security Hub forniscono meccanismi di aggregazione, deduplicazione e analisi per i record di log che vengono resi disponibili anche tramite altri servizi AWS. GuardDuty acquisisce, aggrega e analizza le informazioni da origini come AWS CloudTrail management and data events, log di VPC DNS e log di flusso VPC. Security Hub può acquisire, aggregare e analizzare output di GuardDuty, AWS Config, Amazon Inspector, Amazon Macie, AWS Firewall Manager e un numero significativo di prodotti di sicurezza di terze parti disponibili in Marketplace AWS, nonché il codice proprietario, se è stato compilato in modo adeguato. Sia GuardDuty sia Security Hub hanno un modello membro-amministratore che può aggregare risultati e informazioni su più account. Inoltre, Security Hub viene spesso utilizzato dai clienti che dispongono di un sistema SIEM on-premise, come un preprocessore e aggregatore di avvisi e log lato AWS, da cui possono quindi acquisire Amazon EventBridge tramite un processore e un server di inoltro basati su AWS Lambda.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Valuta le opzioni per l'elaborazione dei log: valuta le opzioni disponibili per l'elaborazione dei log.
  - Utilizza Amazon OpenSearch Service per registrare e monitorare (quasi) tutto
  - Individuazione di un partner specializzato in soluzioni di registrazione e monitoraggio
- Come inizio per analizzare i log CloudTrail, testa Amazon Athena.
  - Configurazione di Athena per analizzare i log CloudTrail.
- Implementa la registrazione centralizzata in AWS: guarda la soluzione di esempio AWS seguente per centralizzare le registrazioni da più origini.
  - Centralize logging solution
- Implementa la registrazione centralizzata con il partner: i partner APN hanno soluzioni per aiutarti ad analizzare i log centralmente.
  - Registrazione e Monitoraggio

Risorse

Documenti correlati:

- AWS Answers: Centralized Logging (AWS Answers: registrazione centralizzata)
- AWS Security Hub
- Amazon CloudWatch
- Amazon EventBridge
- Nozioni di base: Amazon CloudWatch Logs
- Soluzione dei partner per la sicurezza: registrazione e monitoraggio

Video correlati:

- Centrally Monitoring Resource Configuration & Compliance
- Remediating Amazon GuardDuty and AWS Security Hub Findings
- Threat management in the cloud: Amazon GuardDuty and AWS Security Hub

SEC04-BP03 Automazione delle risposte agli eventi

L'utilizzo dell'automazione per analizzare e correggere gli eventi riduce l'impegno e il rischio di errori umani e consente di dimensionare le capacità di analisi. Le revisioni periodiche ti aiuteranno a ottimizzare gli strumenti di automazione e a effettuare un'iterazione costante.

In AWS, è possibile analizzare gli eventi di interesse e le informazioni relative alle modifiche potenzialmente impreviste in un flusso di lavoro automatizzato utilizzando Amazon EventBridge. Questo servizio fornisce un motore di regole scalabile progettato per gestire sia i formati di eventi AWS nativi (ad esempio eventi AWS CloudTrail), sia gli eventi personalizzati che puoi generare dalla tua applicazione. Amazon GuardDuty consente inoltre di instradare gli eventi a un sistema di flusso di lavoro per i sistemi di risposta agli incidenti (AWS Step Functions), a un account di sicurezza centrale o a un bucket per ulteriori analisi.

È inoltre possibile rilevare le modifiche e instradare queste informazioni al flusso di lavoro corretto utilizzando Regole di AWS Config e Pacchetti di conformità. AWS Config individua le modifiche ai servizi coperti (con una latenza maggiore rispetto a EventBridge) e genera eventi che possono essere analizzati tramite le regole di Regole di AWS Config per il rollback, per rafforzare le policy di conformità e per inviare le informazioni ai sistemi, ad esempio le piattaforme di gestione delle modifiche e i sistemi di ticketing operativi. Oltre a scrivere funzioni Lambda personalizzate per rispondere agli eventi di AWS Config, puoi utilizzare il kit per lo sviluppo di regole di Regole di AWS Confige una libreria di Regole di AWS Config open source. I pacchetti di conformità sono una raccolta

di Regole di AWS Config e di azioni di correzione che distribuisci come entità singola creata come modello YAML. Un modello di pacchetto di conformità di esempio è disponibile per il Well-Architected Security Pillar.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Implementa un avviso automatizzato con GuardDuty: GuardDuty è un servizio di rilevamento delle minacce che esegue un monitoraggio continuo per individuare attività dannose e comportamenti non autorizzati al fine di proteggere carichi di lavoro e Account AWS. Abilita GuardDuty e configura gli avvisi automatici.
- Automatizza i processi di indagine: sviluppa processi automatizzati per indagare su un evento e riferire informazioni a un amministratore per risparmiare tempo.
  - Laboratorio: Amazon GuardDuty hands on

Risorse

Documenti correlati:

- AWS Answers: Centralized Logging (AWS Answers: registrazione centralizzata)
- AWS Security Hub
- Amazon CloudWatch
- Amazon EventBridge
- Nozioni di base: Amazon CloudWatch Logs
- Soluzione dei partner per la sicurezza: registrazione e monitoraggio
- Configurazione di Amazon GuardDuty

Video correlati:

- Centrally Monitoring Resource Configuration & Compliance
- Remediating Amazon GuardDuty and AWS Security Hub Findings
- Threat management in the cloud: Amazon GuardDuty and AWS Security Hub

Esempi correlati:

- Laboratorio: Distribuzione automatizzata di controlli di rilevamento

SEC04-BP04 Implementazione di eventi di sicurezza fruibili

Crea e invia al tuo team avvisi fruibili. Assicurati che includano informazioni pertinenti affinché il team possa intervenire. per ogni meccanismo di rilevamento di cui disponi, devi disporre anche di un processo, sotto forma di runbook oppure playbook, da analizzare. Ad esempio, quando abiliti Amazon GuardDuty, vengono generati risultati diversi. È necessario disporre di una voce runbook per ogni tipo di risultato; ad esempio, se viene rilevato un trojan , il runbook contiene istruzioni semplici che indicano come eseguire l'analisi e correggere il problema.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Identificazione delle metriche disponibili per i servizi AWS: scopri le metriche a disposizione attraverso Amazon CloudWatch per i servizi in uso.
  - Documentazione del servizio AWS
  - Utilizzare i parametri Amazon CloudWatch
- Configurazione degli avvisi Amazon CloudWatch.
  - Utilizzo degli allarmi di Amazon CloudWatch

Risorse

Documenti correlati:

- Amazon CloudWatch
- Amazon EventBridge
- Soluzione dei partner per la sicurezza: registrazione e monitoraggio

Video correlati:

- Centrally Monitoring Resource Configuration & Compliance (Monitoraggio centrale della configurazione e della conformità delle risorse)
- Remediating Amazon GuardDuty and AWS Security Hub Findings (Correzione Amazon GuardDuty e risultati AWS Security Hub)

- Threat management in the cloud: Amazon GuardDuty and AWS Security Hub (Gestione delle minacce nel cloud: Amazon GuardDuty e AWS Security Hub)

Protezione dell'infrastruttura

Domande

- SEC 5 In che modo proteggi le risorse di rete?
- SEC 6 In che modo proteggi le risorse di calcolo?

SEC 5 In che modo proteggi le risorse di rete?

Qualsiasi carico di lavoro che abbia una qualche forma di connettività di rete, che si tratti di Internet o di una rete privata, richiede più livelli di difesa per proteggere da minacce esterne e interne basate sulla rete.

Best practice

- SEC05-BP01 Creazione di livelli di rete
- SEC05-BP02 Controllo del traffico a tutti i livelli
- SEC05-BP03 Automatizzazione della protezione di rete
- SEC05-BP04 Implementazione di funzioni di ispezione e protezione

SEC05-BP01 Creazione di livelli di rete

Raggruppa i componenti che condividono requisiti di raggiungibilità in vari livelli. Ad esempio, un cluster di database in un VPC senza necessità di accesso a Internet deve essere posizionato in sottoreti senza routing da o verso Internet. In un carico di lavoro serverless che opera senza un VPC, livelli e segmentazione simili con microservizi possono raggiungere lo stesso obiettivo.

Componenti come istanze Amazon Elastic Compute Cloud (Amazon EC2), cluster di database Amazon Relational Database Service (Amazon RDS) e funzioni AWS Lambda che condividono i requisiti di raggiungibilità possono essere segmentati in livelli formati da sottoreti. Ad esempio, un cluster di database Amazon RDS in un VPC senza necessità di accesso a Internet deve essere posizionato in sottoreti senza routing da o verso Internet. Questo approccio a più livelli per i controlli mitiga l'impatto di una configurazione errata di un livello singolo, che potrebbe consentire l'accesso involontario. Per Lambda, è possibile eseguire le funzioni nel VPC per sfruttare i controlli basati su VPC.

Per la connettività di rete che può includere migliaia di VPC, account AWS e reti on-premise, è consigliabile utilizzare AWS Transit Gateway. Funge da hub che controlla il modo in cui il traffico viene instradato tra tutte le reti connesse, che agiscono come raggi. Il traffico tra Amazon Virtual Private Cloud e AWS Transit Gateway rimane sulla rete privata AWS, riducendo i vettori di minacce esterni, ad esempio attacchi DDoS (Distributed Denial of Service) ed exploit comuni, come iniezione SQL, script tra siti, false richieste tra siti o uso illecito del codice di autenticazione. Il peering AWS Transit Gateway tra Regioni crittografa il relativo traffico senza un singolo punto di errore o un collo di bottiglia a livello di banda.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Creazione di sottoreti in VPC: crea sottoreti per ogni livello (in gruppi che includono più zone di disponibilità) e associa tabelle di instradamento per controllare il routing.
  - VPC e sottoreti
  - Tabelle di instradamento

Risorse

Documenti correlati:

- AWS Firewall Manager
- Amazon Inspector
- Sicurezza Amazon VPC
- Nozioni di base su AWS WAF

Video correlati:

- AWS Transit Gateway reference architectures for many VPCs (Architetture di referenza AWS Transit Gateway per molti VPC)
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield (Accelerazione e protezione delle applicazioni con Amazon CloudFront, AWS Web Application Firewall e AWS Shield)

Esempi correlati:

- Laboratorio: Implementazione automatizzata di VPC

SEC05-BP02 Controllo del traffico a tutti i livelli

durante la progettazione della topologia di rete, è necessario esaminare i requisiti di connettività di ciascun componente. Ad esempio, va esaminato se un componente richiede accessibilità a Internet (in entrata e in uscita), connettività a VPC, servizi edge e data center esterni.

Un VPC consente di definire la topologia di rete che si estende su una regione Regione AWS con un intervallo di indirizzi IPv4 privati impostato dall'utente o un intervallo di indirizzi IPv6 selezionato da AWS. È necessario applicare più controlli con un approccio di difesa avanzata sia per il traffico in entrata che per quello in uscita, tra cui l'uso di gruppi di sicurezza (firewall di ispezione stateful), liste di controllo degli accessi di rete, sottoreti e tabelle di routing. All'interno di un VPC, puoi creare sottoreti in una zona di disponibilità. Ogni sottorete può avere una tabella di routing associata che definisce le regole di instradamento per la gestione dei percorsi del traffico all'interno della sottorete. Puoi definire una sottorete Internet instradabile tramite un percorso che va a un gateway Internet o NAT collegato al VPC o attraverso un altro VPC.

Un'istanza, un database Amazon Relational Database Service(Amazon RDS) o un altro servizio che viene avviato all'interno di un VPC ha un proprio gruppo di sicurezza per interfaccia di rete. Questo firewall è esterno al livello del sistema operativo e può essere utilizzato per definire le regole per il traffico consentito in entrata e in uscita. Puoi anche definire le relazioni tra i gruppi di sicurezza. Ad esempio, le istanze all'interno di un gruppo di sicurezza a livello di database accettano solo il traffico dalle istanze all'interno del livello dell'applicazione, in riferimento ai gruppi di sicurezza applicati alle istanze coinvolte. A meno che non utilizzi protocolli non TCP, non dovrebbe essere necessario disporre di un'istanza Amazon Elastic Compute Cloud(Amazon EC2) accessibile direttamente da internet (anche con porte limitate da gruppi di sicurezza) senza un sistema di bilanciamento del carico o CloudFront. Questo aiuta a proteggerla da accessi non intenzionali dovuti a un problema del sistema operativo o dell'applicazione. Una sottorete può anche avere una lista di controllo degli accessi di rete collegata, che funge da firewall stateless. È necessario configurare la lista di controllo degli accessi di rete per limitare l'ambito del traffico consentito tra i livelli; tieni presente che è necessario definire le regole sia in entrata che in uscita.

Alcuni servizi AWS richiedono dei componenti per accedere a internet per le chiamate API, in cui si trovano gli endpoint API AWS . Altri servizi AWS usano Endpoint VPC all'interno dei Amazon VPC. Molti servizi AWS, tra cui Amazon S3 e Amazon DynamoDB, supportano gli endpoint VPC e questa tecnologia è stata generalizzata in AWS PrivateLink. Ti consigliamo di usare questo approccio per accedere ai servizi AWS, ai servizi di terze parti e ai servizi proprietari ospitati in sicurezza in altri

VPC. Tutto il traffico di rete su AWS PrivateLink rimane sul backbone AWS globale e non attraversa mai internet. La connettività può solo essere avviata dal consumatore del servizio e non dal provider del servizio. Usando l'accesso AWS PrivateLink per i servizi esterni è possibile creare VPC isolati senza accesso a internet e proteggere i VPC da vettori di minacce esterni. I servizi di terze parti possono usare AWS PrivateLink per consentire ai propri clienti di connettersi ai servizi dai propri VPC su indirizzi IP privati. Per gli asset del VPC che devono effettuare connessioni in uscita a Internet, queste possono essere effettuate solo in uscita (unidirezionale) tramite un gateway NAT gestito da AWS, un gateway Internet per connessioni solo in uscita o proxy Web creati e gestiti dall'utente.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Controlla il traffico di rete in un VPC: implementa le best practice di VPC per controllare il traffico.
  - Sicurezza Amazon VPC
  - Endpoint VPC
  - Gruppo di sicurezza Amazon VPC
  - ACL di rete
- Controlla il traffico a livello di edge: implementa servizi edge, come Amazon CloudFront, per fornire un ulteriore livello di protezione e altre funzionalità.
  - Casi d'uso Amazon CloudFront
  - AWS Global Accelerator
  - AWS Web Application Firewall (AWS WAF)
  - Amazon Route 53
  - Amazon VPC Ingress Routing
- Controlla il traffico di rete privato: implementa servizi in grado di proteggere il traffico privato per il carico di lavoro.
  - Amazon VPC Peering
  - Amazon VPC Endpoint Services (AWS PrivateLink)
  - Amazon VPC Transit Gateway
  - AWS Direct Connect
  - AWS Site-to-Site VPN
  - AWS Client VPN
  - Amazon S3 Access Points

Risorse

Documenti correlati:

- AWS Firewall Manager
- Amazon Inspector
- Nozioni di base su AWS WAF

Video correlati:

- AWS Transit Gateway reference architectures for many VPCs
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield

Esempi correlati:

- Laboratorio: Distribuzione automatizzata di VPC

SEC05-BP03 Automatizzazione della protezione di rete

Automatizza i meccanismi di protezione per creare una rete in grado di difendersi da sola grazie alle informazioni sulle minacce e al rilevamento delle anomalie. Ad esempio, strumenti di rilevamento e prevenzione delle intrusioni in grado di adattarsi alle minacce attuali e di ridurre il loro impatto. Un firewall per applicazioni Web è un esempio di dove è possibile automatizzare la protezione della rete, ad esempio utilizzando la soluzione Automatismi di sicurezza di AWS WAF (https://github.com/awslabs/aws-waf-security-automations) per bloccare automaticamente le richieste provenienti da indirizzi IP associati a noti attori di minacce.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Automatizza la protezione per il traffico basato sul Web: AWS offre una soluzione che usa AWS CloudFormation per distribuire automaticamente una serie di regole AWS WAF progettate per filtrare gli attacchi comuni basati sul Web. Gli utenti hanno la possibilità di scegliere tra caratteristiche di protezione preconfigurate che definiscono le regole incluse in una lista di controllo accessi Web (ACL Web) di AWS WAF.
  - Automazioni di sicurezza AWS WAF

- Considera le soluzioni AWS Partner: i partner AWS offrono centinaia di prodotti leader nel settore che sono equivalenti, identici o si integrano ai controlli esistenti negli ambienti on-premise. Questi prodotti integrano i servizi AWS esistenti per permettere di distribuire un'architettura di sicurezza completa e un'esperienza più fluida nel cloud e negli ambienti on-premise.
  - Sicurezza dell'infrastruttura

Risorse

Documenti correlati:

- AWS Firewall Manager
- Amazon Inspector
- Sicurezza di Amazon VPC
- Nozioni di base su AWS WAF

Video correlati:

- AWS Transit Gateway reference architectures for many VPCs
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield

Esempi correlati:

- Laboratorio: Distribuzione automatizzata di VPC

SEC05-BP04 Implementazione di funzioni di ispezione e protezione

Ispeziona e filtra il traffico a ogni livello. Puoi ispezionare le configurazioni VPC per rilevare potenziali accessi indesiderati con VPC Network Access Analyzer. Puoi specificare i requisiti di accesso alla rete e individuare percorsi di rete potenziali che non li soddisfano. Per i componenti che eseguono transazioni tramite protocolli basati su HTTP, un firewall per applicazioni Web può aiutare a proteggere dagli attacchi comuni. AWS WAF è un firewall per applicazioni Web che consente di monitorare e bloccare le richieste HTTP che corrispondono alle regole configurabili inoltrate a un'API di Amazon API Gateway, ad Amazon CloudFront o a un Application Load Balancer. Per iniziare a usare AWS WAF, puoi utilizzare Regole gestite da AWS in combinazione con le tue oppure puoi utilizzare integrazioni dei partner esistenti.

Per gestire le protezioni di AWS WAF, AWS Shield Advanced e i gruppi di sicurezza di Amazon VPC in AWS Organizations, puoi utilizzare AWS Firewall Manager. Questo consente di configurare e gestire centralmente le regole del firewall tra gli account e le applicazioni, rendendo più semplice il dimensionamento dell'applicazione delle regole comuni. Consente inoltre di rispondere rapidamente agli attacchi utilizzando AWS Shield Advancedo soluzioni che bloccano automaticamente le richieste indesiderate alle applicazioni Web. Firewall Manager funziona anche con AWS Network Firewall. AWS Network Firewall è un servizio gestito che usa un motore di regole per garantire un controllo granulare sul traffico di rete stateful e stateless. Supporta le specifiche dell'intrusion prevention system (IPS) open source compatibile con Suricata per le regole che contribuiscono alla protezione del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Configura Amazon GuardDuty: GuardDuty è un servizio di rilevamento delle minacce che esegue un monitoraggio continuo per individuare attività dannose e comportamenti non autorizzati al fine di proteggere carichi di lavoro e account Account AWS. Abilita GuardDuty e configura gli avvisi automatici.
  - Amazon GuardDuty
  - Laboratorio: Distribuzione automatizzata di controlli di rilevamento
- Configura i log di flusso del cloud privato virtuale (VPC): Log di flusso VPC è una funzione che ti permette di acquisire informazioni sul traffico IP in entrata e in uscita dalle interfacce di rete nel tuo VPC. I dati del log di flusso possono essere pubblicati su Amazon CloudWatch Logs e Amazon Simple Storage Service (Amazon S3). Dopo aver creato un log di flusso, puoi recuperarne e visualizzarne i dati nella destinazione scelta.
- Considera il mirroring del traffico VPC: il mirroring del traffico è una caratteristica di Amazon VPC che puoi utilizzare per copiare il traffico di rete da un'interfaccia di rete elastica di istanze Amazon Elastic Compute Cloud (Amazon EC2) e quindi inviarlo ad appliance di sicurezza e monitoraggio fuori banda per l'ispezione dei contenuti, il monitoraggio delle minacce e la risoluzione dei problemi.
  - Mirroring del traffico del VPC

Risorse

Documenti correlati:

- AWS Firewall Manager

- Amazon Inspector
- Sicurezza Amazon VPC
- Nozioni di base su AWS WAF

Video correlati:

- AWS Transit Gateway reference architectures for many VPCs
- Application Acceleration and Protection with Amazon CloudFront, AWS WAF, and AWS Shield

Esempi correlati:

- Laboratorio: Distribuzione automatizzata di VPC

SEC 6 In che modo proteggi le risorse di calcolo?

Le risorse di calcolo nel carico di lavoro richiedono più livelli di difesa per contribuire alla protezione da minacce esterne ed interne. Le risorse di calcolo includono istanze EC2, container, funzioni di AWS Lambda, servizi di database, dispositivi IoT e altro.

Best practice

- SEC06-BP01 Gestione delle vulnerabilità
- SEC06-BP02 Riduzione della superficie d'attacco
- SEC06-BP03 Implementazione di servizi gestiti
- SEC06-BP04 Automatizzazione della protezione delle risorse di calcolo
- SEC06-BP05 Concessione del permesso di eseguire azioni a distanza
- SEC06-BP06 Convalida dell'integrità del software

SEC06-BP01 Gestione delle vulnerabilità

Scansiona e correggi frequentemente le vulnerabilità del codice, delle dipendenze e dell'infrastruttura per proteggere da nuove minacce.

Partendo dalla configurazione della tua infrastruttura di calcolo, puoi automatizzare la creazione e l'aggiornamento delle risorse tramite AWS CloudFormation. CloudFormation consente di creare modelli scritti in YAML o JSON, tramite esempi AWS o scrivendone di propri. In questo modo è

possibile creare modelli di infrastruttura sicuri per impostazione predefinita che puoi verificare con CloudFormation Guard, per risparmiare tempo e diminuire il rischio di errori di configurazione. Puoi creare la tua infrastruttura e distribuire le tue applicazioni tramite la distribuzione continua, ad esempio con AWS CodePipeline, per automatizzare le fasi di creazione, test e rilascio.

Sei responsabile della gestione delle patch per le tue risorse AWS, incluse le istanze Amazon Elastic Compute Cloud (Amazon EC2), Amazon Machine Images (AMI) e molte altre risorse di calcolo. Per le istanze Amazon EC2, AWSSystems Manager Patch Manager automatizza il processo di applicazione di patch alle istanze gestite con aggiornamenti correlati alla sicurezza e di altro tipo. Puoi utilizzare il gestore patch per applicare patch sia per i sistemi operativi sia per le applicazioni. (Sul server di Windows, il supporto per le applicazioni è limitato agli aggiornamenti per le applicazioni Microsoft). Puoi usare Patch Manager per installare i Service Pack sulle istanze Windows ed eseguire aggiornamenti minori di versione sulle istanze Linux. Puoi applicare le patch ai parchi delle istanze Amazon EC2 o ai server on-premise e alle macchine virtuali (VM) secondo il tipo di sistema operativo. Questo comprende le versioni supportate dei server Windows, Amazon Linux, Amazon Linux 2, CentOS, Debian Server, Oracle Linux, Red Hat Enterprise Linux (RHEL), SUSE Linux Enterprise Server (SLES) e Ubuntu Server. Puoi eseguire la scansione delle istanze per visualizzare solo un report delle patch mancanti oppure puoi eseguire la scansione e installare automaticamente tutte le patch mancanti.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Configura Amazon Inspector: Amazon Inspector testa l'accessibilità della rete delle tue istanze Amazon Elastic Compute Cloud (Amazon EC2) e lo stato di sicurezza delle applicazioni eseguite su tali istanze. Amazon Inspector valuta le applicazioni relativamente a esposizione, vulnerabilità e deviazioni dalle best practice.
  - What is Amazon Inspector? (What is Amazon Inspector?)
- Esegui la scansione del codice sorgente: esegui la scansione di librerie e dipendenze per rilevare eventuali vulnerabilità.
  - Amazon CodeGuru
  - OWASP: strumenti di analisi del codice sorgente

Risorse

Documenti correlati:

- AWS Systems Manager
- Sostituzione di un host bastione con Amazon EC2 Systems Manager
- Panoramica sulla sicurezza di AWS Lambda

Video correlati:

- Esecuzione di carichi di lavoro con livello di sicurezza elevato su Amazon EKS
- Securing Serverless and Container Services
- Best practice di sicurezza per il servizio di metadati dell'istanza Amazon EC2

Esempi correlati:

- Laboratorio: Distribuzione automatizzata di firewall per applicazioni Web

SEC06-BP02 Riduzione della superficie d'attacco

Riduci la superficie di attacco ad accessi non intenzionali attraverso la protezione avanzata dei sistemi operativi e riducendo al minimo i componenti, le librerie e i servizi di consumo esterni in uso. Inizia riducendo i componenti inutilizzati, siano essi pacchetti del sistema operativo o applicazioni per carichi di lavoro basati su Amazon Elastic Compute Cloud (Amazon EC2) o moduli software esterni nel codice (per tutti i carichi di lavoro). Esistono molte guide per la configurazione della protezione avanzata e della sicurezza dei sistemi operativi e dei software dei server comuni. Ad esempio, puoi iniziare dal Center for Internet Security e iterare.

In Amazon EC2 puoi creare Amazon Machine Image (AMI), con patch e rafforzamento, per soddisfare i requisiti di sicurezza specifici della tua organizzazione. Le patch e altri controlli di sicurezza che applichi sulle AMI diventano effettivi nel momento in cui vengono creati: non sono dinamici, a meno che tu non decida di modificarli subito dopo l'avvio, ad es. con AWS Systems Manager.

Puoi semplificare il processo di creazione di AMI sicure con EC2 Image Builder. EC2 Image Builder riduce in modo significativo l'impegno richiesto per creare e mantenere immagini "golden" senza scrivere e aggiornare la manutenzione. Quando sono disponibili gli aggiornamenti software, Image Builder produce automaticamente una nuova immagine senza richiedere agli utenti di iniziare una creazione manuale. EC2 Image Builder consente di convalidare con facilità la funzionalità e la sicurezza delle immagini prima di usarle in produzione con test tuoi e forniti da AWS. Puoi anche

applicare impostazioni di sicurezza fornite da AWS per proteggere ulteriormente le immagine e rispettare i criteri di sicurezza interni, Ad esempio, puoi produrre immagini conformi allo standard Security Technical Implementation Guide (STIG) con modelli forniti da AWS.

Con l'utilizzo di strumenti di analisi del codice statico di terze parti puoi identificare problemi di sicurezza comuni, ad esempio limiti di input delle funzioni non controllati e CVE applicabili. Puoi utilizzare Amazon CodeGuru per le lingue supportate. Possono anche essere utilizzati strumenti di controllo delle dipendenze per stabilire se le librerie a cui si collega il codice sono le versioni più recenti, se le stesse sono prive di CVE e se le condizioni di licenza soddisfano i requisiti delle policy del software.

Con Amazon Inspector puoi eseguire valutazioni della configurazione a fronte delle istanze per CVE note, confrontare i valori rispetto a benchmark di sicurezza e automatizzare la notifica dei difetti. Amazon Inspector viene eseguito sulle istanze di produzione o in una pipeline di compilazione e invia una notifica agli sviluppatori e agli ingegneri quando sono disponibili nuovi risultati. Puoi accedere in modo programmatico ai risultati e indirizzare i tuoi team ai sistemi di backlog e rilevamento dei bug. EC2 Image Builder può essere utilizzato per mantenere le immagini del server (AMI) tramite l'applicazione di patch automatizzata, l'applicazione di policy di sicurezza fornite da AWS e altre personalizzazioni. Quando utilizzi i container, implementa la scansione delle immagini ECR nella pipeline di compilazione regolarmente confrontandola con il repository di immagini per cercare le CVE nei container.

Anche se Amazon Inspector e altri strumenti sono efficaci per identificare configurazioni ed eventuali CVE presenti, sono necessari altri metodi per testare il carico di lavoro a livello di applicazione. Il fuzzing è un metodo noto di individuazione dei bug mediante l'automazione per inserire dati malformati nei campi di input e in altre aree dell'applicazione.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Rafforzamento del sistema operativo: configura i sistemi operativi per adeguarli alle best practice.
  - Protezione di Amazon Linux
  - Protezione di Microsoft Windows Server
- Rafforzamento delle risorse containerizzate: configura le risorse containerizzate per la conformità alle best practice in materia di sicurezza.
- Implementa le best practice AWS Lambda.
  - Best practice di AWS Lambda

Risorse

Documenti correlati:

- AWS Systems Manager
- Sostituzione di un host bastione con Amazon EC2 Systems Manager
- Panoramica sulla sicurezza di AWS Lambda

Video correlati:

- Esecuzione di carichi di lavoro con livello di sicurezza elevato su Amazon EKS
- Securing Serverless and Container Services
- Best practice di sicurezza per il servizio di metadati dell'istanza Amazon EC2

Esempi correlati:

- Laboratorio: Distribuzione automatizzata di firewall per applicazioni Web

SEC06-BP03 Implementazione di servizi gestiti

Implementa servizi che gestiscono le risorse, ad esempio Amazon Relational Database Service (Amazon RDS), AWS Lambda e Amazon Elastic Container Service (Amazon ECS), per ridurre le attività di manutenzione della sicurezza nell'ambito del modello di responsabilità condivisa. Ad esempio, Amazon RDS aiuta a configurare, gestire e dimensionare un database relazionale e automatizza le attività di amministrazione quali provisioning di hardware, configurazione di database, applicazione di patch e backup. Ciò significa che hai più tempo libero per concentrarti sulla protezione dell'applicazione in altri modi descritti nel Framework AWS Well-Architected. Lambda consente di eseguire il codice senza dover effettuare il provisioning o gestire server, perciò è sufficiente focalizzarsi su connettività, invocazione e sicurezza a livello di codice, anziché sull'infrastruttura o sul sistema operativo.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Identificazione dei servizi disponibili: esplora, testa e implementa servizi che gestiscono le risorse, come Amazon RDS, AWS Lambda e Amazon ECS.

Risorse

Documenti correlati:

- Sito Web AWS
- AWS Systems Manager
- Sostituzione di un host bastione con Amazon EC2 Systems Manager
- Panoramica sulla sicurezza di AWS Lambda

Video correlati:

- Esecuzione di carichi di lavoro con livello di sicurezza elevato su Amazon EKS
- Securing Serverless and Container Services
- Best practice di sicurezza per il servizio di metadati dell'istanza Amazon EC2

Esempi correlati:

- Laboratorio: Richiesta di certificati pubblichi da parte di Gestione certificati AWS

SEC06-BP04 Automatizzazione della protezione delle risorse di calcolo

Automatizza i meccanismi di protezione delle risorse di calcolo, tra cui la gestione delle vulnerabilità, la riduzione della superficie di attacco e la gestione delle risorse. L'automazione ti consentirà di investire tempo nella protezione di altri aspetti del carico di lavoro e di ridurre il rischio di errori umani.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Automazione della gestione della configurazione: applica e convalida in automatico configurazioni sicure sfruttando un apposito servizio o strumento di gestione.
  - AWS Systems Manager
  - AWS CloudFormation
  - Laboratorio: Implementazione automatizzata di VPC
  - Laboratorio: Implementazione automatizzata di applicazioni Web EC2

- Automazione dell'applicazione delle patch alle istanze Amazon Elastic Compute Cloud (Amazon EC2): AWS Systems Manager Patch Manager automatizza il processo di applicazione di patch alle istanze gestite con aggiornamenti correlati alla sicurezza e di altro tipo. Puoi utilizzare il gestore patch per applicare patch sia per i sistemi operativi sia per le applicazioni.
  - AWS Systems Manager Patch Manager
  - Applicazione di patch centralizzata multi-regione e muli-account con AWS Systems Manager Automation.

- Implementazione della prevenzione e del rilevamento delle intrusioni: implementa uno strumento di rilevamento e prevenzione delle intrusioni per monitorare e bloccare le attività sospette sulle istanze.
- Considerazione delle soluzioni AWS Partner: i partner AWS offrono centinaia di prodotti leader nel settore che sono equivalenti, identici o si integrano ai controlli esistenti negli ambienti on-premise. Questi prodotti integrano i servizi AWS esistenti per permettere di implementare un'architettura di sicurezza completa e un'esperienza più fluida nel cloud e negli ambienti on-premise.
  - Sicurezza dell'infrastruttura

Risorse

Documenti correlati:

- AWS CloudFormation
- AWS Systems Manager
- AWS Systems Manager Patch Manager
- Applicazione di patch centralizzata multi-regione e muli-account con AWS Systems Manager Automation.
- Sicurezza dell'infrastruttura
- Sostituzione di un host bastione con Amazon EC2 Systems Manager
- Panoramica sulla sicurezza di AWS Lambda

Video correlati:

- Esecuzione di carichi di lavoro con livello di sicurezza elevato su Amazon EKS
- Securing Serverless and Container Services

- Best practice di sicurezza per il servizio di metadati dell'istanza Amazon EC2

Esempi correlati:

- Laboratorio: Implementazione automatizzata di firewall per applicazioni Web
- Laboratorio: Implementazione automatizzata di applicazioni Web EC2

SEC06-BP05 Concessione del permesso di eseguire azioni a distanza

Eliminare la possibilità di accesso interattivo riduce il rischio di errore umano e la potenziale configurazione o gestione manuale. Ad esempio, utilizza un flusso di lavoro per la gestione delle modifiche per distribuire le istanze Amazon Elastic Compute Cloud (Amazon EC2) tramite infrastructure-as-code, quindi gestire le istanze Amazon EC2 utilizzando strumenti come AWS Systems Manager invece di consentire l'accesso diretto o tramite un host bastione. AWS Systems Manager può automatizzare un'ampia gamma di attività di manutenzione e distribuzione utilizzando funzionalità quali automazione di automazione, documenti (playbook) e il Run Command. Gli stack di AWS CloudFormation si basano su pipeline e possono automatizzare le attività di distribuzione e gestione dell'infrastruttura senza utilizzare direttamente la AWS Management Console o le API.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Sostituisci l'accesso della console: sostituisci l'accesso via console (SSH o RDP) alle istanze con AWS Systems Manager Run Command per automatizzare le attività di gestione.

- AWS Systems Manager Run Command

Risorse

Documenti correlati:

- AWS Systems Manager
- AWS Systems Manager Run Command
- Sostituzione di un host bastione con Amazon EC2 Systems Manager
- Panoramica sulla sicurezza di AWS Lambda

Video correlati:

- Esecuzione di carichi di lavoro con livello di sicurezza elevato su Amazon EKS
- Securing Serverless and Container Services
- Best practice di sicurezza per il servizio di metadati dell'istanza Amazon EC2

Esempi correlati:

- Laboratorio: Distribuzione automatizzata di firewall per applicazioni Web

SEC06-BP06 Convalida dell'integrità del software

Implementa meccanismi (ad esempio la firma del codice) per verificare che il software, il codice e le librerie utilizzati nel carico di lavoro provengano da origini attendibili e non siano stati manomessi. Ad esempio, devi verificare il certificato di firma del codice dei file binari e degli script per confermare l'autore e accertarti che non sia stato manomesso da quando è stato creato dall'autore. AWS Signer può aiutare a garantire l'affidabilità e l'integrità del tuo codice tramite una sua gestione centralizzata, registrando il ciclo di vita, incluso la registrazione delle certificazioni e delle chiavi pubbliche e private. Puoi imparare come usare modelli avanzati e best practice per la registrazione del codice con AWS Lambda. Inoltre, un confronto tra il checksum del software scaricato e quello del provider può garantire che non sia stato manomesso.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Analizza i meccanismi: la firma del codice è uno dei meccanismi utili per convalidare l'integrità del software.
  - NIST: considerazioni sulla sicurezza per la registrazione del codice

Risorse

Documenti correlati:

- AWS Signer
- Nuovo – Code Signing, a Trust and Integrity Control for AWS Lambda

Protezione dei dati

Domande

- SEC 7 In che modo classifichi i dati?
- SEC 8 In che modo proteggi i dati inattivi?
- SEC 9 In che modo proteggi i dati in transito?

SEC 7 In che modo classifichi i dati?

La classificazione fornisce un modo per categorizzare i dati in base ai livelli di criticità e sensibilità, in modo da aiutarti a determinare i controlli di protezione e conservazione appropriati.

Best practice

- SEC07-BP01 Identificazione dei dati all'interno del carico di lavoro
- SEC07-BP02 Definizione dei controlli di protezione dei dati
- SEC07-BP03 Automazione dell'identificazione e della classificazione
- SEC07-BP04 Definizione della gestione del ciclo di vita dei dati

SEC07-BP01 Identificazione dei dati all'interno del carico di lavoro

è necessario comprendere il tipo e la classificazione dei dati elaborati dal carico di lavoro, i processi aziendali associati, il proprietario dei dati, i requisiti legali e di conformità applicabili, il luogo di archiviazione e i controlli risultanti da applicare. Ciò può includere classificazioni per indicare se i dati sono destinati a essere disponibili al pubblico, se i dati sono solo di uso interno, ad esempio informazioni che consentono l'identificazione personale del cliente (PII, Personally Identifiable Information), oppure se i dati riguardano un accesso più limitato, ad esempio relativi alla proprietà intellettuale, dati confidenziali o sensibili e altro ancora. L'attenta gestione di un sistema appropriato di classificazione dei dati e dei requisiti di protezione di ciascun livello del carico di lavoro consente di mappare i controlli e il livello di accesso/protezione dei dati adeguato. Ad esempio, i contenuti destinati al pubblico sono accessibili a tutti, ma i contenuti importanti sono crittografati e archiviati in modo protetto e richiedono l'accesso autorizzato a una chiave per essere decrittati.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Prendi in considerazione la possibilità di scoprire i dati con Amazon Macie: Macie riconosce i dati sensibili, come le Informazioni personali di identificazione (PII) o la proprietà intellettuale.
  - Amazon Macie

Risorse

Documenti correlati:

- Amazon Macie
- Whitepaper sulla classificazione dei dati
- Nozioni di base su Amazon Macie

Video correlati:

- Introducing the New Amazon Macie

SEC07-BP02 Definizione dei controlli di protezione dei dati

Proteggi i dati in base al livello di classificazione. Ad esempio, puoi mettere in sicurezza le informazioni classificate come pubbliche utilizzando raccomandazioni pertinenti e allo stesso tempo proteggere i dati sensibili con controlli aggiuntivi.

Utilizzando tag di risorse, account AWS separati per livelli di sensibilità (e potenzialmente anche per avvertimento/enclave/community di interesse), policy IAM, SCP di AWS Organizations, AWS Key Management Service (AWS KMS) e AWS CloudHSM, puoi definire e implementare le policy per la classificazione e la protezione dei dati tramite la crittografia. Ad esempio, se in un progetto sono presenti bucket S3 che contengono dati estremamente critici o istanze Amazon Elastic Compute Cloud (Amazon EC2) che elaborano dati riservati, essi possono essere contrassegnati con un tag `Project=ABC` . Solo il team ristretto conosce il significato del codice del progetto e rappresenta un modo per utilizzare il controllo degli accessi basato su attributi. Puoi definire i livelli di accesso alle chiavi di crittografia AWS KMS tramite policy e concessioni delle chiavi per garantire che solo i servizi appropriati abbiano accesso ai contenuti sensibili tramite un meccanismo sicuro. Se prendi decisioni in merito alle autorizzazioni in base ai tag, devi assicurarti che le autorizzazioni sui tag siano definite in modo appropriato utilizzando le policy dei tag in AWS Organizations.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Definizione dello schema di identificazione e classificazione dei dati: l'identificazione e la classificazione dei dati è utile a valutare l'impatto potenziale e il tipo di dati archiviati e a stabilire chi può accedervi.
  - Documentazione di AWS
- Identificazione dei controlli AWS disponibili: scopri i controlli di sicurezza per i servizi AWS che stai utilizzando o che intendi utilizzare. Molti servizi dispongono di una sezione sulla sicurezza nella documentazione.
  - Documentazione di AWS
- Identificazione delle risorse di conformità AWS: identifica le risorse che AWS mette a disposizione per facilitare i processi di conformità.
  - https://aws.amazon.com/compliance/

Risorse

Documenti correlati:

- Documentazione di AWS
- Whitepaper sulla classificazione dei dati
- Nozioni di base su Amazon Macie
- Testo mancante

Video correlati:

- Introducing the New Amazon Macie (Presentazione del nuovo Amazon Macie)

SEC07-BP03 Automazione dell'identificazione e della classificazione

automatizzare l'identificazione e la classificazione dei dati può aiutarti a implementare i controlli corretti. L'utilizzo dell'automazione per queste operazioni invece dell'accesso diretto da parte di una persona riduce il rischio di errori umani e di esposizione delle persone. È consigliabile valutare l'utilizzo di uno strumento, ad esempio Amazon Macie, che utilizza il machine learning per rilevare, classificare e proteggere automaticamente i dati sensibili in AWS. Amazon Macie riconosce i dati sensibili, quali informazioni personali di identificazione (PII) o di proprietà intellettuale e fornisce

pannelli di controllo e allarmi che offrono visibilità su come viene effettuato l'accesso a tali dati o come vengono spostati.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Utilizzo di Amazon Simple Storage Service (Amazon S3) Inventory: Amazon S3 Inventory è uno degli strumenti che utilizzabili per eseguire audit e segnalare lo stato di replica e crittografia degli oggetti.
  - Amazon S3 Inventory
- Considerazione di Amazon Macie: Amazon Macie sfrutta il machine learning per scoprire e classificare automaticamente i dati archiviati in Amazon S3.
  - Amazon Macie

Risorse

Documenti correlati:

- Amazon Macie
- Amazon S3 Inventory
- Whitepaper sulla classificazione dei dati
- Nozioni di base su Amazon Macie

Video correlati:

- Introducing the New Amazon Macie (Presentazione del nuovo Amazon Macie)

SEC07-BP04 Definizione della gestione del ciclo di vita dei dati

la strategia del ciclo di vita definita deve basarsi sul livello di sensibilità e sui requisiti legali e aziendali. Gli aspetti da considerare includono la durata di conservazione dei dati, i processi di distruzione dei dati, la gestione degli accessi ai dati, la trasformazione dei dati e la condivisione dei dati. Nella scelta di una metodologia di classificazione dei dati, è necessario valutare l'usabilità rispetto all'accesso. Devi inoltre gestire vari livelli di accesso e particolarità per implementare un approccio sicuro e utilizzabile per ogni livello. Utilizza sempre un approccio di difesa avanzata e riduci l'accesso umano ai dati e ai meccanismi per trasformare, eliminare o copiare i dati. Ad esempio,

richiedi agli utenti di effettuare l'autenticazione in un'applicazione e fornisci all'applicazione, anziché agli utenti, l'autorizzazione di accesso necessaria per eseguire "operazioni a distanza". Inoltre, assicurati che gli utenti provengano da un percorso di rete sicuro e richiedi l'accesso alle chiavi di decrittografia. Utilizza strumenti, pannelli di controllo e generazione di report automatizzata, per fornire agli utenti informazioni ricavate dai dati piuttosto che concedere loro l'accesso diretto ai dati.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Identifica i tipi di dati: identifica i tipi di dati che stai archiviando o elaborando nel carico di lavoro. Questi potrebbero consistere in testo, immagini, database binari e così via.

Risorse

Documenti correlati:

- Whitepaper sulla classificazione dei dati
- Nozioni di base su Amazon Macie

Video correlati:

- Introducing the New Amazon Macie

SEC 8 In che modo proteggi i dati inattivi?

Proteggi i dati inattivi implementando più controlli, per ridurre il rischio di accessi non autorizzati o altri comportamenti impropri.

Best practice

- SEC08-BP01 Implementazione della gestione sicura delle chiavi
- SEC08-BP02 Applicazione della crittografia dei dati inattivi
- SEC08-BP03 Automatizzazione della protezione dei dati a riposo
- SEC08-BP04 Applicazione del controllo degli accessi
- SEC08-BP05 Utilizzo di meccanismi per tenere le persone a distanza dai dati

SEC08-BP01 Implementazione della gestione sicura delle chiavi

Definendo un approccio alla crittografia che include l'archiviazione, la rotazione e il controllo degli accessi delle chiavi, proteggi i tuoi contenuti da utenti non autorizzati e dall'esposizione superflua a utenti autorizzati. AWS Key Management Service (AWS KMS) ti aiuta a gestire le chiavi di crittografia e si integra con molti servizi AWS. Si tratta di un servizio che fornisce un'archiviazione durevole, sicura e ridondante per le tue chiavi AWS KMS. Puoi definire i tuoi alias delle chiavi e le policy a livello di chiave. Le policy ti aiutano a definire gli amministratori della chiave e i suoi utenti. Inoltre, AWS CloudHSM è un modulo di sicurezza hardware (HSM, Hardware Security Module) basato sul cloud che consente di generare e utilizzare chiavi di crittografia personalizzate nel Cloud AWS. Ti aiuta a rispettare i requisiti di conformità aziendali, contrattuali e normativi per la sicurezza dei dati utilizzando HSM conformi allo standard FIPS 140-2 Level 3.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Implementazione di AWS KMS: AWS KMS semplifica la creazione e la gestione di chiavi e controlla l'uso della crittografia in un'ampia gamma di servizi AWS e nelle tue applicazioni. AWS KMS è un servizio sicuro e resiliente che usa moduli di sicurezza hardware convalidati FIPS 140-2 per proteggere le tue chiavi.
  - Nozioni di base: AWS Key Management Service (AWS KMS)
- Considerazione dell'SDK di crittografia AWS: utilizza l'SDK di crittografia AWS con integrazione di AWS KMS quando la tua applicazione richiede la crittografia dei dati lato client.
  - SDK di crittografia AWS

Risorse

Documenti correlati:

- AWS Key Management Service
- Servizi e strumenti di crittografia di AWS
- Nozioni di base: AWS Key Management Service (AWS KMS)
- Protezione dei dati Amazon S3 tramite la crittografia

Video correlati:

- Come funziona la crittografia in AWS
- Securing Your Block Storage on AWS (Protezione dello storage a blocchi in AWS).

SEC08-BP02 Applicazione della crittografia dei dati inattivi

Devi accertarti che l'unico modo per archiviare i dati sia l'utilizzo della crittografia. AWS Key Management Service (AWS KMS) si integra perfettamente con molti servizi AWS per semplificare la crittografia di tutti i dati a riposo. Ad esempio, in Amazon Simple Storage Service (Amazon S3) puoi impostare la crittografia predefinita su un bucket in modo che tutti i nuovi oggetti vengano crittografati automaticamente. Inoltre, Le istanze dei server virtuali Amazon Elastic Compute Cloud (Amazon EC2) e Amazon S3 supportano l'applicazione della crittografia impostando la crittografia predefinita. Puoi utilizzare Regole di AWS Config per verificare automaticamente che stai utilizzando la crittografia, ad esempio, per i volumi Amazon Elastic Block Store (Amazon EBS), le istanze Amazon Relational Database Service (Amazon RDS)e bucket Amazon S3.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Applicazione della crittografia dei dati inattivi per Amazon Simple Storage Service (Amazon S3): implementa la crittografia predefinita del bucket Amazon S3.
  - Come abilitare la crittografia predefinita per un bucket S3?
- Utilizzo di AWS Secrets Manager: Secrets Manager è il servizio AWS che facilita il compito di gestire i segreti. I segreti possono essere credenziali di database, password, chiavi API di terze parti e persino testo arbitrario.
  - AWS Secrets Manager
- Configurazione della crittografia predefinita per nuovi volumi EBS: specifica che desideri che tutti i volumi EBS appena creati vengano creati in forma crittografata, con la possibilità di utilizzare la chiave predefinita fornita da AWS oppure una chiave creata da te.
  - Crittografia predefinita per i volumi EBS
- Configurazione di Amazon Machine Images (AMI) crittografate: copiando un'AMI esistente con crittografia abilitata verrà eseguita la crittografia automatica di volumi root e snapshot.
  - AMI con snapshot crittografati
- Configurazione della crittografia Amazon Relational Database Service (Amazon RDS): configura la crittografia per cluster e snapshot del database Amazon RDS inattivi abilitando l'opzione di crittografia.

    - Crittografia delle risorse Amazon RDS
- Configurazione della crittografia in servizi AWS aggiuntivi: per i servizi AWS che usi, stabilisci le funzionalità di crittografia.
    - Documentazione di AWS

Risorse

Documenti correlati:

- AMI con snapshot crittografati
- AWS Crypto Tools
- Documentazione di AWS
- SDK di crittografia AWS
- Whitepaper per i dettagli della crittografia di AWS KMS
- AWS Key Management Service
- AWS Secrets Manager
- Servizi e strumenti di crittografia di AWS
- Crittografia Amazon EBS
- Crittografia predefinita per i volumi EBS
- Crittografia delle risorse Amazon RDS
- Come abilitare la crittografia predefinita per un bucket S3?
- Protezione dei dati Amazon S3 tramite la crittografia

Video correlati:

- Come funziona la crittografia in AWS
- Securing Your Block Storage on AWS (Protezione dello storage a blocchi in AWS).

SEC08-BP03 Automatizzazione della protezione dei dati a riposo

utilizza strumenti automatizzati per convalidare e applicare la protezione dei dati a riposo in modo continuo; ad esempio verifica che siano presenti solo risorse di storage crittografate. Puoi automatizzare la convalida della crittografia di tutti i volumi EBS utilizzando Regole di AWS Config. AWS Security Hub può anche verificare una serie di controlli diversi tramite verifiche

automatiche a fronte di standard di sicurezza. Inoltre, le Regole di AWS Config possono correggere automaticamente le risorse non conformi.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

I dati a riposo rappresentano tutti i dati conservati nello storage non volatile per qualsiasi durata del carico di lavoro. Sono inclusi storage a blocchi, storage di oggetti, database, archivi, dispositivi IoT e qualsiasi altro supporto di storage su cui sono conservati i dati. La protezione dei dati a riposo riduce il rischio di accesso non autorizzato quando vengono implementati crittografia e controlli degli accessi adeguati.

Applica la crittografia dei dati a riposo: devi accertarti che l'unico modo per archiviare i dati sia l'utilizzo della crittografia. AWS KMS si integra perfettamente con molti servizi AWS per semplificare la crittografia di tutti i dati inattivi. Ad esempio, in Amazon Simple Storage Service (Amazon S3) puoi impostare la crittografia predefinita su un bucket in modo che tutti i nuovi oggetti vengano crittografati automaticamente. Inoltre, Amazon EC2 e Amazon S3 supportano l'applicazione della crittografia impostando la crittografia predefinita. Puoi utilizzare AWS Managed Config Rules per verificare automaticamente che stai utilizzando la crittografia, ad esempio, per i volumi EBS, le istanze Amazon Relational Database Service (Amazon RDS)e bucket Amazon S3.

Risorse

Documenti correlati:

- AWS Crypto Tools
- SDK di crittografia AWS

Video correlati:

- How Encryption Works in AWS (Come funziona la crittografia in AWS)
- Securing Your Block Storage on AWS

SEC08-BP04 Applicazione del controllo degli accessi

Applica il controllo degli accessi con privilegi minimi e meccanismi come backup, isolamento e controllo delle versioni, per favorire la protezione dei dati a riposo. Impedisci agli operatori di concedere l'accesso pubblico ai tuoi dati.

Controlli diversi inclusi l'accesso (tramite il privilegio minimo), i backup (vedi il whitepaper sull'affidabilità), l'isolamento e il controllo delle versioni possono tutti aiutare a proteggere i dati a riposo. L'accesso ai dati deve essere controllato utilizzando i meccanismi di rilevamento trattati in precedenza in questo documento, tra cui CloudTrail e il registro del livello di servizio, ad esempio i registri di accesso Amazon Simple Storage Service (Amazon S3). Devi eseguire un inventario dei dati accessibili al pubblico e pianificare come ridurre la quantità di dati disponibili nel tempo. Amazon S3 Glacier Vault Lock e Amazon S3 Object Lock sono funzionalità che forniscono un controllo degli accessi obbligatorio. Una volta bloccata una policy Vault con l'opzione di conformità, nemmeno l'utente root può modificarla fino alla scadenza del blocco. Il meccanismo soddisfa i requisiti di gestione di libri e record di SEC, CFTC e FINRA. Per ulteriori dettagli, consulta questo whitepaper.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Applicazione del controllo degli accessi: applica il controllo degli accessi con privilegio minimo, incluso l'accesso alle chiavi di crittografia.
  - Introduzione alla gestione delle autorizzazioni di accesso per le risorse Amazon S3
- Separazione dei dati in base a diversi livelli di classificazione: usa diversi Account AWS per i livelli di classificazione dei dati gestiti da AWS Organizations.
  - AWS Organizations
- Revisione delle policy AWS KMS: rivedi il livello di accesso consentito nelle policy di AWS KMS.
  - Panoramica sulla gestione dell'accesso alle risorse di AWS KMS
- Esame dei bucket Amazon S3 e delle autorizzazioni degli oggetti: rivedi con regolarità il livello di accesso concesso nelle policy dei bucket Amazon S3. Le best practice prevedono l'assenza di bucket pubblicamente leggibili o scrivibili. Valuta l'utilizzo di AWS Config per rilevare i bucket disponibili pubblicamente e di Amazon CloudFront per fornire contenuti provenienti da Amazon S3.
  - Regole di AWS Config
  - Amazon S3 + Amazon CloudFront: un abbinamento nel cloud
- Abilitazione del controllo delle versioni e del blocco degli oggetti di Amazon S3.
  - Utilizzo del controllo delle versioni
  - Blocco degli oggetti con Amazon S3 Object Lock
- Utilizzo di Amazon S3 Inventory: Amazon S3 Inventory è uno degli strumenti utilizzabili per eseguire audit e segnalare lo stato di replica e crittografia degli oggetti.
  - Amazon S3 Inventory

- Esame delle autorizzazioni di condivisione Amazon EBS e AMI: le autorizzazioni di condivisione consentono la condivisione di immagini e volumi con Account AWS esterni al tuo carico di lavoro.
  - Condivisione di uno snapshot Amazon EBS
  - AMI condivise

Risorse

Documenti correlati:

- Whitepaper per i dettagli della crittografia di AWS KMS

Video correlati:

- Securing Your Block Storage on AWS (Protezione dello storage a blocchi in AWS).

SEC08-BP05 Utilizzo di meccanismi per tenere le persone a distanza dai dati

Evita a tutti gli utenti di accedere direttamente a dati e sistemi sensibili in circostanze operative normali. Ad esempio, usa un flusso di lavoro per la gestione delle modifiche per gestire le istanze Amazon Elastic Compute Cloud (Amazon EC2) tramite strumenti, invece di consentire l'accesso diretto o tramite un host bastione. A tal fine puoi utilizzare AWS Systems Manager Automation, che utilizza documenti di automazione che contengono le fasi utilizzate per eseguire le attività. Questi documenti possono essere archiviati nel controllo sorgente, revisionati in peering prima dell'esecuzione e testati accuratamente per ridurre al minimo i rischi rispetto all'accesso alla shell. Gli utenti aziendali possono utilizzare un pannello di controllo anziché accedere direttamente a un datastore per eseguire query. Se non vengono utilizzate le pipeline CI/CD, determina quali controlli e processi sono necessari per fornire in modo adeguato un meccanismo di accesso di tipo break-glass normalmente disabilitato.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Implementazione di meccanismi per tenere le persone lontane dai dati: i meccanismi includono l'utilizzo di pannelli di controllo, ad esempio Amazon QuickSight, per mostrare i dati agli utenti anziché eseguire query direttamente.
  - Amazon QuickSight

- Automazione della gestione della configurazione: esegui azioni a distanza, applica e convalida in automatico configurazioni sicure sfruttando un apposito servizio o strumento di gestione. Evita l'uso di bastion host o l'accesso diretto alle istanze EC2.
  - AWS Systems Manager
  - AWS CloudFormation
  - Pipeline CI/CD per modelli AWS CloudFormation su AWS

Risorse

Documenti correlati:

- Whitepaper per i dettagli della crittografia di AWS KMS

Video correlati:

- Come funziona la crittografia in AWS
- Securing Your Block Storage on AWS (Protezione dello storage a blocchi in AWS).

SEC 9 In che modo proteggi i dati in transito?

Proteggi i dati in transito implementando più controlli, per ridurre il rischio di accessi non autorizzati o perdita.

Best practice

- SEC09-BP01 Implementazione della gestione sicura delle chiavi e dei certificati
- SEC09-BP02 Applicazione della crittografia dei dati in transito
- SEC09-BP03 Automatizzazione del rilevamento degli accessi indesiderati ai dati
- SEC09-BP04 Autenticazione delle comunicazioni di rete

SEC09-BP01 Implementazione della gestione sicura delle chiavi e dei certificati

archivia le chiavi di crittografia e i certificati in modo sicuro e ruotali a intervalli di tempo appropriati tramite un controllo rigoroso degli accessi. Il modo migliore per farlo è utilizzare un servizio gestito, ad esempio AWS Certificate Manager (ACM). Questo servizio consente di effettuare il provisioning, gestire e distribuire facilmente certificati TLS (Transport Layer Security) pubblici e privati da utilizzare

con i servizi AWS e le risorse interne connesse. I certificati TLS vengono utilizzati per proteggere le comunicazioni di rete e stabilire l'identità dei siti Web su Internet e delle risorse su reti private. ACM si integra con le risorse AWS, come Elastic Load Balancer (ELB), distribuzioni AWS e API su API Gateway, gestendo anche i rinnovi automatici dei certificati. Se utilizzi ACM per implementare un'autorità di certificazione (CA, Certificate Authority) root privata, esso può fornire sia certificati sia chiavi private da utilizzare in istanze Amazon Elastic Compute Cloud (Amazon EC2), container e così via.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Implementazione della gestione sicura delle chiavi e dei certificati: implementa una soluzione di gestione dei certificati e delle chiavi sicura e definita.
  - Gestion certificati AWS
  - Come ospitare e gestire un'intera infrastruttura di certificati privata in AWS
- Implementazione dei protocolli di sicurezza: utilizza protocolli sicuri che offrono autenticazione e riservatezza, come TLS (Transport Layer Security) o IPsec, per ridurre il rischio di manomissione o perdita dei dati. Consulta la documentazione di AWS per i protocolli e la sicurezza attinenti ai servizi in uso.

Risorse

Documenti correlati:

- Documentazione di AWS

SEC09-BP02 Applicazione della crittografia dei dati in transito

Applica i requisiti di crittografia definiti in base ad appropriati standard e raccomandazioni in modo da soddisfare i requisiti aziendali, legali e di conformità. I servizi AWS forniscono endpoint HTTPS utilizzando TLS per le comunicazioni e pertanto forniscono crittografia in transito quando comunicano con le API AWS. I protocolli non sicuri, come HTTP, possono essere controllati e bloccati in un VPC tramite l'uso di gruppi di sicurezza. Le richieste HTTP possono anche essere reindirizzate automaticamente a HTTPS in Amazon CloudFront o in un Application Load Balancer. Hai il controllo completo sulle tue risorse informatiche per implementare la crittografia in transito nei tuoi servizi. Inoltre, puoi utilizzare la connettività VPN nel VPC da una rete esterna per facilitare la crittografia del traffico. Per requisiti particolari, in Marketplace AWS sono disponibili soluzioni di terze parti.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Applica la crittografia in transito: i requisiti di crittografia definiti dovrebbero essere basati sugli standard e sulle best practice più recenti e consentire solo protocolli sicuri. Configura ad esempio un solo gruppo di sicurezza per consentire l'utilizzo del protocollo HTTPS a un Application Load Balancer o a un'istanza Amazon Elastic Compute Cloud (Amazon EC2).
- Configura protocolli di sicurezza nei servizi edge: configura HTTPS con Amazon CloudFront e la crittografia richiesta.
  - Utilizzo di HTTPS con CloudFront
- Usa un VPN per la connettività esterna: valuta l'impiego di una VPN IPsec per la protezione delle connessioni punto a punto o rete a rete al fine di garantire la riservatezza e l'integrità dei dati.
  - Connessioni VPN
- Configura protocolli sicuri nei sistemi di bilanciamento del carico: abilita listener HTTPS per la protezione delle connessioni verso i sistemi di bilanciamento del carico.
  - Listener HTTPS per Application Load Balancer
- Configura protocolli sicuri per le istanze: valuta la possibilità di configurare la crittografia HTTPS sulle istanze.
  - Tutorial: configurazione del server Web Apache su Amazon Linux 2 per l'utilizzo di SSL/TLS
- Configura protocolli di sicurezza in Amazon Relational Database Service (Amazon RDS): usa Secure Socket Layer (SSL) o Transport Layer Security (TLS) per crittografare la connessione a istanze di database.
  - Utilizzo di SSL per crittografare una connessione a un'istanza DB
- Configura protocolli di sicurezza in Amazon Redshift: configura il cluster per richiedere una connessione Secure Socket Layer (SSL) o Transport Layer Security (TLS).
  - Configurazione delle opzioni di sicurezza per le connessioni
- Configura protocolli di sicurezza in servizi AWS aggiuntivi Per i servizi AWS che usi, stabilisci le funzionalità di crittografia in transito.

Risorse

Documenti correlati:

- Documentazione di AWS

SEC09-BP03 Automatizzazione del rilevamento degli accessi indesiderati ai dati

Usa strumenti come Amazon GuardDuty per rilevare in automatico attività o tentativi sospetti di trasferire i dati al di fuori di limiti predefiniti. Ad esempio, GuardDuty può rilevare attività di lettura di Amazon Simple Storage Service (Amazon S3) inusuale con Exfiltration:S3/AnomalousBehavior finding. Oltre a GuardDuty, si possono utilizzare i Registri di flusso Amazon VPC, che acquisiscono informazioni sul traffico di rete, con Amazon EventBridge per attivare il rilevamento di connessioni anomale, riuscite e negate. Amazon S3 Access Analyzer aiuta a valutare quali dati sono accessibili a chi nei bucket Amazon S3.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Automazione del rilevamento di accessi ai dati non intenzionali: utilizza uno strumento o un meccanismo di rilevamento per rilevare automaticamente i tentativi di spostamento dei dati all'esterno dei confini definiti, ad esempio, per individuare un sistema di database che copia i dati su un host sconosciuto.
  - Log di flusso VPC
- Valutazione di Amazon Macie: Amazon Macie è un servizio di sicurezza e privacy dei dati completamente gestito che utilizza il machine learning e il pattern matching per rilevare e proteggere i dati sensibili all'interno di AWS.
  - Amazon Macie

Risorse

Documenti correlati:

- Log di flusso VPC
- Amazon Macie

SEC09-BP04 Autenticazione delle comunicazioni di rete

Verifica l'identità delle comunicazioni utilizzando protocolli che supportano l'autenticazione, ad esempio Transport Layer Security (TLS) o IPsec.

L'utilizzo di protocolli di rete che supportano l'autenticazione consente di stabilire l'attendibilità tra le parti. Questo si aggiunge alla crittografia utilizzata nel protocollo per ridurre il rischio che le

comunicazioni vengano alterate o intercettate. I protocolli comuni che implementano l'autenticazione includono il protocollo TLS (Transport Layer Security), che viene utilizzato in molti servizi AWS, e IPsec, utilizzato in AWS Virtual Private Network (AWS VPN).

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Implementazione di protocolli sicuri: utilizza protocolli sicuri che offrono autenticazione e riservatezza, come TLS (Transport Layer Security) o IPsec, per ridurre il rischio di manomissione o perdita dei dati. Verifica la Documentazione di AWS per i protocolli e la sicurezza attinenti ai servizi in uso.

Risorse

Documenti correlati:

- Documentazione di AWS

Risposta agli imprevisti

Domanda

- SEC 10 In che modo prevedi, reagisci a e risolvi gli incidenti?

SEC 10 In che modo prevedi, reagisci a e risolvi gli incidenti?

La preparazione è cruciale per un esame tempestivo ed efficace degli incidenti di sicurezza, nonché per la risposta e il ripristino, così da ridurre al minimo potenziali interruzioni dell'organizzazione.

Best practice

- SEC10-BP01 Identificazione del personale chiave e delle risorse esterne
- SEC10-BP02 Sviluppo di piani di gestione degli incidenti
- SEC10-BP03 Preparazione di funzionalità forensi
- SEC10-BP04 Automatizzazione della capacità di contenimento
- SEC10-BP05 Preassegnazione dell'accesso
- SEC10-BP06 Distribuzione anticipata degli strumenti

- SEC10-BP07 Esecuzione di giornate di gioco

SEC10-BP01 Identificazione del personale chiave e delle risorse esterne

Identifica personale, risorse e requisiti legali interni ed esterni che potrebbero aiutare l'organizzazione a rispondere a un incidente.

Quando definisci come affrontare la risposta agli incidenti nel cloud, insieme ad altri team (ad esempio il consulente legale, la leadership dell'organizzazione, le parti interessate, i servizi AWS Support e altri), devi identificare il personale chiave, le parti interessate e i contatti pertinenti. Per ridurre le dipendenze e i tempi di risposta, assicurati che il personale, i team di sicurezza specializzati e i team che rispondono agli incidenti ricevano informazioni sui servizi che utilizzi e abbiano l'opportunità di esercitarsi direttamente.

Ti invitiamo a identificare i partner di sicurezza AWS esterni in grado di fornirti competenze e una prospettiva diversa per potenziare le tue capacità di risposta. I partner di sicurezza affidabili possono aiutarti a identificare potenziali rischi o minacce che potresti non conoscere.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Identifica il personale chiave all'interno dell'organizzazione: conserva un elenco di contatti del personale interno alla tua organizzazione che potrebbe essere necessario coinvolgere per rispondere a un incidente ed effettuare il ripristino.
- Identifica i partner esterni: se necessario, coinvolgi partner esterni che possano aiutarti a rispondere a un incidente e a effettuare il ripristino.

Risorse

Documenti correlati:

- AWS Incident Response Guide

Video correlati:

- Prepare for and respond to security incidents in your AWS environment

Esempi correlati:

SEC10-BP02 Sviluppo di piani di gestione degli incidenti

Crea piani che ti aiutino a rispondere a un incidente, comunicare durante lo stesso e ripristinare in seguito le risorse. Ad esempio, puoi avviare un piano di risposta agli incidenti con gli scenari più probabili per il carico di lavoro e l'organizzazione. Includi il modo in cui gestiresti la comunicazione e l'escalation internamente ed esternamente.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

Un piano di gestione degli incidenti è fondamentale per rispondere, mitigare e ripristinare lo stato a seguito del potenziale impatto degli incidenti di sicurezza. Un piano di gestione degli incidenti è un processo strutturato per identificare, correggere e rispondere tempestivamente agli incidenti di sicurezza.

Il cloud ha molti degli stessi ruoli e requisiti operativi che si trovano in un ambiente on-premise. Quando si crea un piano di gestione degli incidenti è importante tenere conto delle strategie di risposta e ripristino che meglio si allineano ai risultati aziendali e ai requisiti di conformità. Ad esempio, se gestisci carichi di lavoro in AWS conformi a FedRAMP negli Stati Uniti, è utile attenersi a NIST SP 800-61 Computer Security Handling Guide (NIST SP 800-61 Guida alla gestione della sicurezza informatica). Analogamente, quando gestisci carichi di lavoro con dati PII (informazioni personali di identificazione) europei, considera ad esempio come potresti proteggere e rispondere a problemi relativi alla residenza dei dati come richiesto dalle normative del Regolamento generale sulla protezione dei dati (GDPR) dell'Unione europea.

Quando crei un piano di gestione degli incidenti per i carichi di lavoro eseguiti in AWS, inizia con il Modello di responsabilità condivisa AWSper creare un approccio di difesa in profondità in risposta agli incidenti. In questo modello, AWS gestisce la sicurezza del cloud e tu sei responsabile della sicurezza nel cloud. Ciò significa che mantieni il controllo e sei responsabile dei controlli di sicurezza che scegli di implementare. La AWS Security Incident Response Guide (Guida alle risposte agli incidenti di sicurezza di AWS) illustra i concetti chiave e le linee guida di base per la creazione di un piano di gestione degli incidenti incentrato sul cloud.

Un piano di gestione degli incidenti efficace deve essere continuamente iterato per rimanere in linea con l'obiettivo delle operazioni cloud. Prendi in considerazione l'utilizzo dei piani di implementazione descritti di seguito durante la creazione e l'evoluzione del tuo piano di gestione degli incidenti.

- Istruzione e formazione per la risposta agli incidenti: quando si verifica una deviazione dalla linea di base definita (ad esempio, un'implementazione o una configurazione errata), potrebbe essere

necessario rispondere e analizzare. Per farlo correttamente, è necessario comprendere quali controlli e capacità è possibile utilizzare per la risposta agli incidenti di sicurezza all'interno del proprio ambiente AWS, nonché i processi che è necessario implementare per preparare, istruire e formare i team cloud che partecipano alla risposta agli incidenti.

- Playbook e runbook sono meccanismi efficaci per creare coerenza nella formazione su come rispondere agli incidenti. Inizia con la creazione di un elenco di procedure eseguite di frequente per rispondere agli incidenti e continua a ripetere le operazioni mentre apprendi o utilizzi nuove procedure.
- Acquisisci familiarità con i playbook e i runbook con i previsti game day. Durante i game day, simula la risposta agli incidenti in un ambiente controllato in modo che i team possano apprendere come rispondere e per verificare che i team coinvolti nella risposta agli incidenti conoscano bene i flussi di lavoro. Esamina i risultati dell'evento simulato per identificare i miglioramenti e determinare le necessità di ulteriore formazione o strumenti aggiuntivi.
- La sicurezza deve essere considerata un impegno per tutti. Sviluppa una conoscenza collettiva del processo di gestione degli incidenti coinvolgendo tutto il personale che normalmente gestisce i carichi di lavoro. Includi tutti gli aspetti dell'azienda, come le operazioni, i test, lo sviluppo, la sicurezza, la direzione e l'esecutivo.

- Documentazione del piano di gestione degli incidenti: documenta gli strumenti e il processo per registrare, agire, comunicare lo stato di avanzamento e fornire notifiche sugli incidenti attivi. L'obiettivo del piano di gestione degli incidenti è verificare che il normale funzionamento venga ripristinato il più rapidamente possibile, l'impatto sul business sia ridotto al minimo e tutte le parti interessate siano informate. Esempi di incidenti includono, tra gli altri, la perdita o il deterioramento della connettività di rete, un processo o un'API che non risponde, un'attività pianificata che non viene eseguita (ad esempio le patch non riuscite), l'indisponibilità dei dati o del servizio dell'applicazione, l'interruzione del servizio non pianificata a causa di eventi di sicurezza, la perdita di credenziali o gli errori di configurazione.
  - Identifica il proprietario principale responsabile della risoluzione degli incidenti, ad esempio il proprietario del carico di lavoro. Predisponi una guida chiara su chi guiderà la risposta all'incidente e come verrà gestita la comunicazione. Quando più di una parte partecipa al processo di risoluzione degli incidenti, ad esempio un fornitore esterno, prendi in considerazione la creazione di una matrice di responsabilità (RACI)dettagliando i ruoli e le responsabilità di vari team o persone necessari per la risoluzione degli incidenti.

    La matrice RACI descrive quanto segue:

    - R: Responsible, la parte responsabile che svolge il lavoro per completare l'attività.

    - A: Accountable, parte o stakeholder predisposta con l'autorità finale sul completamento corretto dell'attività specifica.
    - C: Consulted, parte consultata le cui opinioni sono richieste, tipicamente come esperti in materia.
    - I: Informed, parte a cui viene notificato lo stato di avanzamento, spesso solo al completamento dell'attività o del risultato finale.
- Classificazione degli incidenti: la definizione e la classificazione degli incidenti in base alla gravità e al punteggio di impatto consente un approccio strutturato al triage e alla risoluzione degli incidenti. Le seguenti raccomandazioni illustrano una matrice di urgenza dall'impatto alla risoluzione per quantificare un incidente. Ad esempio, un incidente a basso impatto e a bassa urgenza è considerato un incidente di bassa gravità.
    - Alto: l'impatto sulla tua attività è significativo. Le funzioni critiche dell'applicazione relative alle risorse AWS non sono disponibili. Questa categoria è riservata agli eventi più critici che interessano i sistemi produttivi. L'impatto dell'incidente aumenta rapidamente poiché la correzione è soggetta a requisiti di tempo.
    - Medio: un servizio aziendale o un'applicazione correlata alle risorse AWS ha subito un impatto moderato e continua a funzionare in uno stato degradato. Le applicazioni che contribuiscono agli obiettivi del livello di servizio (SLO) sono interessate entro i limiti dell'Accordo sul livello di servizio (SLA). I sistemi possono funzionare con capacità ridotte senza grande impatto finanziario e reputazionale.
    - Basso: sono interessate le funzioni non critiche del servizio aziendale o dell'applicazione relative alle risorse AWS. I sistemi possono funzionare con capacità ridotta con minimo impatto finanziario e reputazionale.
- Standardizzazione dei controlli di sicurezza: l'obiettivo della standardizzazione dei controlli di sicurezza è ottenere coerenza, tracciabilità e ripetibilità per quanto riguarda i risultati operativi. Promuovi la standardizzazione tra le attività chiave che sono critiche per la risposta agli incidenti, ad esempio:
    - Gestione di identità e accessi: stabilisci i meccanismi per controllare l'accesso ai tuoi dati e gestire i privilegi per le identità di persone fisiche e macchine. Estendi la tua gestione di identità e accessi al cloud, utilizzando la sicurezza federata con autenticazione unica e privilegi basati sui ruoli per ottimizzare la gestione degli accessi. Per i suggerimenti sulle best practice e i piani di miglioramento per standardizzare la gestione degli accessi, consulta la sezione della gestione di identità e accessi del whitepaper Security Pillar (Principio della sicurezza).

  - Gestione delle vulnerabilità: stabilisci i meccanismi per identificare le vulnerabilità del tuo ambiente AWS che potrebbero essere utilizzate dagli aggressori per compromettere e abusare del tuo sistema. Implementa i controlli preventivi e investigativi come meccanismi di sicurezza per rispondere e mitigare il potenziale impatto degli incidenti di sicurezza. Standardizza i processi come la modellazione delle minacce come parte della creazione dell'infrastruttura e del ciclo di vita della distribuzione delle applicazioni.
  - Gestione delle configurazioni: definisci le configurazioni standard e automatizza le procedure per l'implementazione delle risorse nel Cloud AWS. La standardizzazione dell'infrastruttura e del provisioning delle risorse aiuta a mitigare il rischio di configurazioni errate dovute a implementazioni o configurazioni errate per incidente umano. Consulta la sezione Design Principles (principi di progettazione) del whitepaper Operational Excellence Pillar (Principio dell'eccellenza operativa) per linee guida e piani di miglioramento per l'applicazione di questo controllo.
  - Registrazione e monitoraggio per il controllo di audit: implementa i meccanismi per monitorare le tue risorse per errori, degrado delle prestazioni e problemi di sicurezza. La standardizzazione di questi controlli fornisce anche gli audit trail delle attività che si verificano nel sistema, aiutando il triage tempestivo e la risoluzione dei problemi. Le best practice incluse in SEC 4 ("In che modo individui ed esamini gli eventi di sicurezza?") forniscono le indicazioni per l'applicazione di questo controllo.
- Utilizzo dell'automazione: l'automazione consente una risoluzione tempestiva degli incidenti su vasta scala. AWS fornisce diversi servizi per automatizzare nel contesto della strategia di risposta agli incidenti. Concentrati sulla ricerca di un equilibrio appropriato tra automazione e intervento manuale. Quando crei la risposta agli incidenti nei playbook e nei runbook, automatizza i passaggi ripetibili. Usa i servizi AWS come AWS Systems Manager Incident Manager per risolvere gli incidenti IT più velocemente. Utilizza gli strumenti per sviluppatori per fornire il controllo delle versioni e automatizzare le implementazioni di Amazon Machine Images (AMI) e Infrastruttura come codice (IaC) senza l'intervento umano. Ove applicabile, automatizza il rilevamento e la valutazione della conformità utilizzando servizi gestiti come Amazon GuardDuty, Amazon Inspector, AWS Security Hub, AWS Config e Amazon Macie. Ottimizza le capacità di rilevamento con soluzioni di machine learning come Amazon DevOps Guru per rilevare problemi di schemi operativi anomali prima che si verifichino.
- Esecuzione dell'analisi della causa principale e acquisizione delle lezioni apprese: implementa i meccanismi per acquisire le lezioni apprese come parte di una revisione della risposta successiva all'incidente. Quando la causa principale di un incidente rivela un difetto più grande, un difetto di progettazione, una configurazione errata o una possibilità di ricorrenza, essa viene classificata

come problema. In questi casi, analizza e risolvi il problema per ridurre al minimo l'interruzione delle normali operazioni.

Risorse

Documenti correlati:

- AWS Security Incident Response Guide (Guida alle risposte agli incidenti di sicurezza di AWS)
- NIST: Guida alla gestione degli incidenti di sicurezza informatica

Video correlati:

- Automating Incident Response and ForensicsAWS
- DIY guide to runbooks, incident reports, and incident response
- Prepare for and respond to security incidents in your AWS environment

Esempi correlati:

- Laboratorio: Incident Response Playbook with Jupyter - AWS IAM
- Lab: Incident Response with AWS Console and CLI (Laboratorio: risposta agli incidenti con la Console AWS e l'interfaccia della riga di comando)

SEC10-BP03 Preparazione di funzionalità forensi

È importante che le persone che intervengono dopo un incidente siano in grado di capire quando e come un'indagine forense è richiesta nel piano di risposta. L'organizzazione deve stabilire le prove da raccogliere e gli strumenti da utilizzare nel processo. Identifica e prepara capacità di indagine forensi idonee, tra cui specialisti esterni, strumenti e automazione. Una decisione importante da prendere in anticipo è se raccogliere i dati da un sistema live. Alcuni dati, come i contenuti della memoria volatile o le connessioni di rete attive, andranno perse se il sistema viene spento o riavviato.

Il team di risposta agli incidenti può abbinare strumenti, come AWS Systems Manager, Amazon EventBridge e AWS Lambda, per eseguire in automatico strumenti forensi all'interno di un sistema operativo e il mirroring del traffico VPC e ottenere un pacchetto di rete, per raccogliere prove non persistenti. Conduci altre attività, come l'analisi dei log o l'analisi delle immagini del disco, in un account di sicurezza dedicato con workstation forensi personalizzate e strumenti accessibili per i soccorritori.

Invia con regolarità i log rilevanti a un data store che garantisce durabilità e integrità elevate. I soccorritori devono avere accesso a tali log. AWS offre diversi strumenti che possono semplificare l'analisi dei log, come Amazon Athena, Amazon OpenSearch Service (OpenSearch Service) e Amazon CloudWatch Logs Insights. Inoltre, conserva le prove in modo sicuro con Amazon Simple Storage Service (Amazon S3) Object Lock. Questo servizio è in linea con il modello WORM (scrivi uno - leggi molti) e impedisce l'eliminazione o la sovrascrittura di oggetti per un periodo definito. Poiché le tecniche di indagine forensi richiedono una formazione specializzata, potrebbe essere necessario coinvolgere specialisti esterni.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Identifica le capacità forensi: ricerca le capacità di indagine forense della tua organizzazione, gli strumenti disponibili e gli specialisti esterni.
- Automatizzazione delle indagini e della risposta agli incidenti

Risorse

Documenti correlati:

- How to automate forensic disk collection in AWS

SEC10-BP04 Automatizzazione della capacità di contenimento

Automatizza il contenimento di un incidente e il successivo ripristino per ridurre i tempi di risposta e l'impatto sull'organizzazione.

Dopo aver creato e utilizzato i processi e gli strumenti dai playbook, puoi decostruire la logica in una soluzione basata su codice, che può essere utilizzata come strumento dal team di risposta per automatizzare la risposta e rimuovere la varianza o le supposizioni. Questo può accelerare il ciclo di vita di una risposta. L'obiettivo successivo è abilitare questo codice in modo che sia completamente automatizzato e che possa essere richiamato dagli avvisi o dagli eventi stessi, piuttosto che da un addetto alle risposte, per creare una risposta basata sugli eventi. Questi processi devono anche aggiungere automaticamente dati pertinenti ai sistemi di sicurezza. Ad esempio, un incidente che comporta traffico da un indirizzo IP non desiderato può automaticamente popolare un elenco i blocco AWS WAF o un gruppo di regole del firewall di rete per prevenire ulteriori attività.

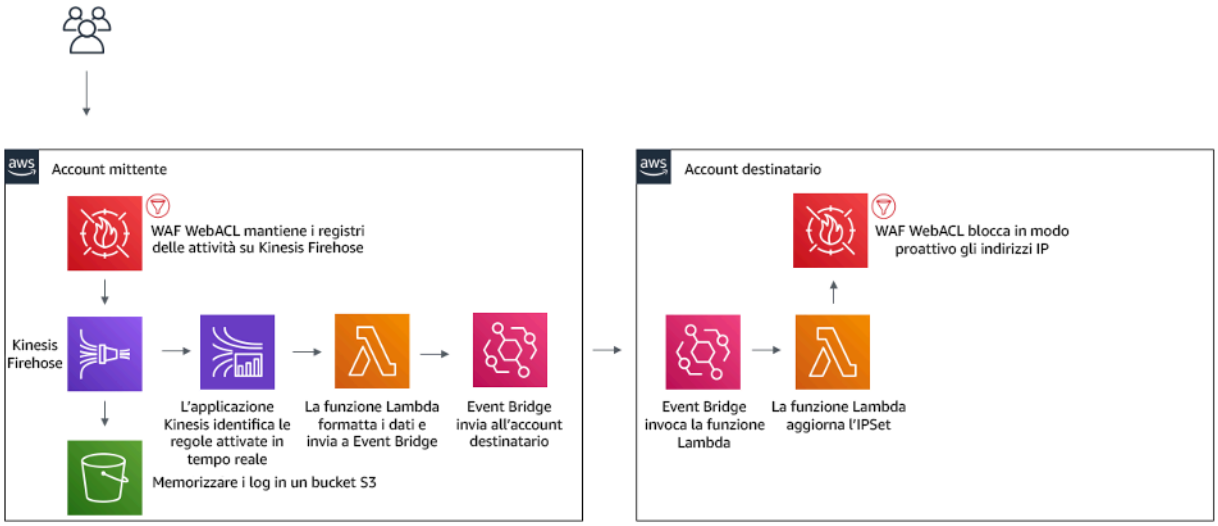


Figura 3: AWS WAF automatizza il blocco di indirizzi IP dannosi noti.

Tramite un sistema di risposta basata sugli eventi, un meccanismo di rilevamento attiva un meccanismo di risposta per correggere automaticamente l'evento. Puoi utilizzare le funzionalità di risposta basata sugli eventi per ridurre il time-to-value tra meccanismi di rilevamento e di risposta. Per creare questa architettura basata sugli eventi, puoi utilizzare AWS Lambda, un servizio di elaborazione serverless che esegue il codice in risposta a eventi e gestisce automaticamente le risorse di calcolo sottostanti per tuo conto. Ad esempio, supponiamo che tu disponga di un account AWS con il servizio AWS CloudTrail abilitato. Se AWS CloudTrail è disabilitato (tramite la chiamata API `cloudtrail:StopLogging`), puoi utilizzare Amazon EventBridge per monitorare l'evento specifico `cloudtrail:StopLogging` e richiamare una funzione AWS Lambda al fine di chiamare `cloudtrail:StartLogging` per riavviare la registrazione.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

Automatizzazione della capacità di contenimento.

Risorse

Documenti correlati:

- AWS Incident Response Guide

Video correlati:

- Prepare for and respond to security incidents in your AWS environment

SEC10-BP05 Preassegnazione dell'accesso

Verifica che il team di risposta agli incidenti disponga degli opportuni diritti di accesso allocati in AWS per ridurre i tempi necessari per l'analisi e il ripristino.

Anti-pattern comuni:

- L'utilizzo dell'account root per la risposta agli incidenti.
- La modifica degli account utente esistenti.
- La manipolazione diretta delle autorizzazioni IAM quando si fornisce l'elevazione dei privilegi just-in-time.

Livello di rischio associato se questa best practice non fosse adottata: medio

Guida all'implementazione

AWS raccomanda di ridurre o eliminare, ove possibile, la dipendenza da credenziali di lunga durata, a favore delle credenziali temporanee e dei meccanismi di escalation dei privilegi just-in-time . Le credenziali di lunga durata sono soggette a rischi per la sicurezza e aumentano il sovraccarico operativo. Per la maggior parte delle attività di gestione, nonché per le attività di risposta agli incidenti, consigliamo di implementare la federazione delle identità insieme all'escalation temporanea per l'accesso amministrativo. In questo modello, un utente richiede l'elevazione a un livello di privilegio superiore (come un ruolo di risposta agli incidenti) e, se è idoneo all'elevazione, la richiesta viene inviata al responsabile dell'approvazione. Se la richiesta viene approvata, l'utente riceve un set di credenziali AWS temporanee che può utilizzare per eseguire le sue attività. Alla scadenza di queste credenziali, l'utente deve inviare una nuova richiesta di elevazione.

Si consiglia l'uso dell'escalation temporanea dei privilegi nella maggior parte degli scenari di risposta agli incidenti. Il modo corretto per farlo è utilizzare AWS Security Token Service e le policy di sessione per definire l'ambito di accesso.

Esistono scenari in cui le identità federate non sono disponibili, come nei casi di:

- Interruzione correlata a un gestore dell'identità digitale (IdP) compromesso.
- Configurazione errata o errore umano che causa l'interruzione del sistema di gestione dell'accesso federato.

- Attività dannose come un evento DDoS (Distributed Denial of Service) o indisponibilità del sistema.

Nei casi precedenti, si deve configurare un accesso di emergenza di tipo break-glass per consentire l'analisi e la tempestiva risoluzione degli incidenti. Ti consigliamo di utilizzare un utente IAM con le autorizzazioni appropriate per eseguire le attività e accedere alle risorse AWS. Utilizza le credenziali root solo per le attività che richiedono l'accesso come utente root. Per verificare che i team di risposta agli incidenti dispongano del corretto livello di accesso ad AWS e ad altri sistemi pertinenti, ti consigliamo di eseguire la pre-assegnazione di account utente dedicati. Gli account utente richiedono l'accesso con privilegi e devono essere rigorosamente controllati e monitorati. Gli account devono essere creati con il minor numero di privilegi richiesti per eseguire le attività e il livello di accesso deve essere basato sui playbook inclusi nel piano di gestione degli incidenti.

Utilizza utenti e ruoli specifici e dedicati come best practice. L'escalation temporanea dell'accesso di utenti o ruoli tramite l'aggiunta di policy IAM rende poco chiaro quale fosse l'accesso degli utenti durante l'incidente e rischia di non revocare i privilegi oggetto di escalation.

È importante rimuovere il maggior numero possibile di dipendenze per verificare che sia possibile ottenere l'accesso nel maggior numero possibile di scenari di errore. Per supportare questa esigenza, crea un playbook per verificare che gli utenti dei team di risposta agli incidenti vengano creati come utenti AWS Identity and Access Management in un account di sicurezza dedicato e non gestiti tramite una federazione esistente o una soluzione di autenticazione unica (SSO). Ogni singolo utente dei team di risposta deve avere il proprio account denominato. La configurazione dell'account deve applicare una policy di password complesse e l'autenticazione a più fattori (MFA). Se i playbook di risposta agli incidenti richiedono solo l'accesso alla AWS Management Console, non è necessario che l'utente disponga di chiavi di accesso configurate né che sia esplicitamente autorizzato a creare chiavi di accesso. A tale scopo è possibile configurare le policy IAM o le policy di controllo dei servizi come menzionato in AWS Security Best Practices (Best practice di sicurezza AWS) per le policy di controllo dei servizi AWS Organizations. Gli utenti non devono avere privilegi oltre la capacità di assumere i ruoli di risposta agli incidenti in altri account.

Durante un incidente potrebbe essere necessario concedere l'accesso ad altre persone interne o esterne per supportare le attività di analisi, correzione o ripristino. In questo caso, utilizza il meccanismo del playbook menzionato in precedenza e un processo per verificare che qualsiasi accesso aggiuntivo venga revocato immediatamente dopo il completamento dell'incidente.

Per verificare che l'uso dei ruoli di risposta agli incidenti possa essere adeguatamente monitorato e controllato, è essenziale che gli account utente IAM creati a tale scopo non siano condivisi tra le persone e che l'utente root Account AWS non venga utilizzato se non per un'attività specifica. Se è

richiesto l'utente root (ad esempio, l'accesso IAM a un account specifico non è disponibile), utilizza un processo separato con un playbook disponibile per verificare la disponibilità della password dell'utente root e del token MFA.

Per configurare le policy IAM per i ruoli di risposta agli incidenti, prendi in considerazione di usare IAM Access Analyzer per generare le policy sulla base dei log AWS CloudTrail. In questo caso, concedi l'accesso come amministratore al ruolo di risposta agli incidenti per un account non di produzione ed esegui i playbook. Al termine, potrà essere creata una policy che consenta solo le azioni da intraprendere. Questa policy può quindi essere applicata a tutti i ruoli di risposta agli incidenti in tutti gli account. Puoi anche creare una policy IAM separata per ogni playbook per avere una gestione e un controllo più semplici. Esempi di playbook possono essere piani di risposta per ransomware, violazioni dei dati, perdita dell'accesso alla produzione e altri scenari.

Utilizza gli account utente di risposta agli incidenti per assumere i ruoli di risposta IAM dedicati in altri Account AWS. Questi ruoli devono essere configurati in modo che possano essere assunti solo dagli utenti nell'account di sicurezza e la relazione di trust deve richiedere che il principale chiamante sia autenticato tramite MFA. I ruoli devono utilizzare policy IAM con ambito limitato per controllare l'accesso. Assicurati che tutte le richieste `AssumeRole` per questi ruoli vengano registrate in CloudTrail e notificate e che tutte le azioni intraprese utilizzando questi ruoli vengano registrate.

Ti consigliamo vivamente di nominare chiaramente gli account utente IAM e i ruoli IAM per trovarli facilmente nei log CloudTrail. Un esempio potrebbe essere quello di nominare gli account IAM *<ID_UTENTE>*`-BREAK-GLASS` e i ruoli IAM `RUOLO-BREAK-GLASS`.

CloudTrail viene utilizzato per registrare l'attività API negli account AWS e deve essere utilizzato per configurare gli avvisi sull'utilizzo dei ruoli di risposta agli incidenti. Fai riferimento al post del blog sulla configurazione degli avvisi quando vengono utilizzate le chiavi root. Le istruzioni possono essere modificate per configurare il parametro Amazon CloudWatch da filtro a filtro negli eventi `AssumeRole` correlati al ruolo IAM di risposta agli incidenti:

```
{ $.eventName = "AssumeRole" && $.requestParameters.roleArn =
 "<ARN_RUOLO_DI_RISPOSTA_AGLI_INCIDENTI>" && $.userIdentity.invokedBy NOT EXISTS &&
 $.eventType != "AwsServiceEvent" }
```

Poiché è probabile che i ruoli di risposta agli incidenti abbiano un livello di accesso elevato, è importante che questi avvisi vengano inviati a un gruppo ampio e vengano gestiti tempestivamente.

Durante un incidente, è possibile che un membro del team di risposta richieda l'accesso a sistemi che non sono direttamente protetti da IAM, ad esempio istanze Amazon Elastic Compute Cloud,

database Amazon Relational Database Service o piattaforme Software-as-a-service (SaaS). Anziché i protocolli nativi come SSH o RDP, ti consigliamo vivamente di utilizzare AWS Systems Manager Session Manager per l'accesso amministrativo completo alle istanze Amazon EC2. Questo accesso può essere monitorato utilizzando IAM, che è sicuro e controllato. Puoi anche automatizzare parti dei tuoi playbook utilizzando i documenti di AWS Systems Manager Run Commandche possono ridurre gli errori dell'utente e migliorare i tempi di ripristino. Per l'accesso a database e strumenti di terze parti, ti consigliamo di archiviare le credenziali di accesso in AWS Secrets Manager e di concedere l'accesso ai ruoli degli utenti dei team di risposta agli incidenti.

Infine, la gestione degli account utente IAM di risposta agli incidenti deve essere aggiunta ai processi degli utenti che si uniscono, si spostano o lasciano l'organizzazione e deve rivista e testata periodicamente per verificare che sia consentito solo l'accesso previsto.

Risorse

Documenti correlati:

- Managing temporary elevated access to your AWS environment (Gestione dell'accesso temporaneo con privilegi elevati all'ambiente AWS)
- AWS Security Incident Response Guide (Guida alle risposte agli incidenti di sicurezza di AWS)
- AWS Elastic Disaster Recovery
- AWS Systems Manager Incident Manager
- Impostazione di una policy delle password dell'account per utenti IAM
- Utilizzo dell'autenticazione a più fattori (MFA) in AWS
- Configuring Cross-Account Access with MFA (Configurazione dell'accesso multi-account con MFA)
- Using IAM Access Analyzer to generate IAM policies (Utilizzo di IAM Access Analyzer per generare policy IAM)
- Best Practices for AWS Organizations Service Control Policies in a Multi-Account Environment (Best practice per le policy di controllo dei servizi di AWS Organizations in un ambiente multi-account)
- How to Receive Notifications When Your AWS Account’s Root Access Keys Are Used (Come ricevere le notifiche quando vengono utilizzate le chiavi di accesso root dell'account AWS)
- Create fine-grained session permissions using IAM managed policies (Creazione di autorizzazioni di sessione dettagliate utilizzando le policy gestite da IAM)

Video correlati:

- Automating Incident Response and ForensicsAWS
- DIY guide to runbooks, incident reports, and incident response
- Prepare for and respond to security incidents in your AWS environment

Esempi correlati:

- Lab: AWS Account Setup and Root User (Laboratorio: configurazione dell'account AWS e dell'utente root)
- Lab: Incident Response with AWS Console and CLI (Laboratorio: risposta agli incidenti con la Console AWS e l'interfaccia della riga di comando)

SEC10-BP06 Distribuzione anticipata degli strumenti

assicurati che il personale addetto alla sicurezza disponga degli strumenti giusti pre-distribuiti in AWS per ridurre i tempi di verifica fino al ripristino.

Per automatizzare le funzioni delle operazioni e la progettazione della sicurezza, puoi utilizzare un set completo di API e strumenti di AWS. Puoi automatizzare completamente le funzionalità di gestione delle identità, sicurezza della rete, protezione dei dati e monitoraggio e distribuirle utilizzando metodi di sviluppo software comuni già esistenti. Quando crei l'automazione della sicurezza, il sistema può monitorare, rivedere e avviare una risposta, invece di far monitorare alle persone il comportamento di sicurezza e reagire manualmente agli eventi. Un modo efficace per fornire in automatico dati di log pertinenti e su cui è possibile effettuare ricerche nei servizi AWS a chi si attiva in seguito a un incidente è abilitare Amazon Detective.

Se i team di risposta agli incidenti continuano a rispondere agli avvisi nello stesso modo, rischiano il cosiddetto affaticamento dagli avvisi ("alert fatigue"). Ciò significa che, nel corso del tempo, il team può diventare desensibilizzato agli avvisi e può commettere errori nella gestione di situazioni ordinarie o farsi sfuggire avvisi insoliti. L'automazione aiuta a evitare l'affaticamento dagli avvisi utilizzando funzioni che elaborano gli avvisi ripetitivi e ordinari, lasciando alle persone la gestione degli incidenti sensibili e univoci. Se si integrano sistemi di rilevamento delle anomalie, come Amazon GuardDuty, AWS CloudTrail Insights e Amazon CloudWatch Anomaly Detection, è possibile ridurre l'impatto di avvisi frequenti basati su soglie.

Puoi migliorare i processi manuali automatizzando le fasi del processo a livello di programmazione. Dopo aver definito il modello di correzione di un evento, puoi decomporre tale modello in una logica fruibile e scrivere il codice per eseguire tale logica. Il team di risposta può quindi eseguire il

codice per risolvere il problema. Nel corso del tempo, puoi automatizzare più fasi e, infine, gestire automaticamente intere classi di incidenti comuni.

Per gli strumenti eseguiti all'interno del sistema operativo dell'istanza Amazon Elastic Compute Cloud (Amazon EC2), considera l'utilizzo di AWS Systems Manager Run Command, che consente di amministrare le istanze in remoto e in modo sicuro utilizzando un agente installato nel sistema operativo delle istanze Amazon EC2. È richiesto l'agente Systems Manager (Agente SSM), installato per impostazione predefinita su molte Amazon Machine Image (AMI). Tieni presente, tuttavia, che una volta che un'istanza è stata compromessa, nessuna risposta da parte di strumenti o agenti in esecuzione su di essa va considerata affidabile.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Distribuisci gli strumenti in anticipo: verifica che il personale addetto alla sicurezza disponga dei corretti strumenti pre-distribuiti in AWS affinché si possa implementare una risposta adeguata a un incidente.
  - Laboratorio: Incident response with AWS Management Console and CLI
  - Incident Response Playbook with Jupyter - AWS IAM
  - Automazione della sicurezza in AWS
- Implementa l'applicazione di tag alle risorse: applica tag alle risorse con informazioni, ad esempio un codice per la risorsa sottoposta a verifica, in modo da poter identificare le risorse durante un incidente.
  - Strategie di applicazione di tag AWS

Risorse

Documenti correlati:

- AWS Incident Response Guide (Guida alle risposte agli incidenti)

Video correlati:

- DIY guide to runbooks, incident reports, and incident response

SEC10-BP07 Esecuzione di giornate di gioco

i game day, noti anche come simulazioni o esercizi, sono eventi interni che offrono un'opportunità strutturata per mettere in pratica i piani e le procedure di gestione degli incidenti in uno scenario realistico. Tali attività sono importanti per esercitare le capacità dei partecipanti, con gli stessi strumenti e le stesse tecniche del mondo reale e persino gli stessi ambienti. I game day riguardano fondamentalmente la preparazione e il miglioramento iterativo delle capacità di risposta. Alcuni dei motivi per cui potresti trovare utile l'organizzazione di game day includono:

- Convalida della preparazione
- Sviluppo delle competenze: apprendimento da simulazioni e dal personale preposto alla formazione
- Rispetto degli obblighi contrattuali o di conformità
- Generazione di artefatti per l'accreditamento
- Agilità: miglioramento incrementale
- Maggiore rapidità e miglioramento degli strumenti
- Perfezionamento della comunicazione e dell'escalation
- Gestione più sicura delle situazioni rare e inaspettate

Per questi motivi, il valore derivato dalla partecipazione a un'attività di simulazione aumenta l'efficacia di un'organizzazione durante gli eventi stressanti. Sviluppare un'attività di simulazione realistica e utile può essere un esercizio difficile. Anche se testare le procedure o l'automazione che gestisce eventi noti presenta alcuni vantaggi, è altrettanto utile partecipare alle attività Security Incident Response Simulations (SIRS) creative per mettersi alla prova in situazioni impreviste e migliorare continuamente.

Crea simulazioni personalizzate in base al tuo ambiente, al tuo team e ai tuoi strumenti. Trova un problema e progetta una simulazione. Potrebbe trattarsi di una credenziale compromessa, di un server che comunica con sistemi indesiderati o di una configurazione errata che comporta un'esposizione non autorizzata. Identifica gli ingegneri che conoscono l'organizzazione per creare lo scenario e per la partecipazione di un altro gruppo. Lo scenario deve essere realistico e abbastanza impegnativo per essere rilevante. Deve offrire la possibilità di fare pratica con registri, notifiche, escalation ed esecuzioni di runbook o automazioni. Durante la simulazione, i soccorritori devono esercitare le proprie capacità tecniche e organizzative e i leader devono essere coinvolti per sviluppare le competenze necessarie per la gestione degli incidenti. Alla fine della simulazione, riconosci l'impegno del team e trova il modo di iterare, ripetere e ampliare nuove simulazioni.

AWS ha creato modelli di Runbook di risposta agli incidenti che puoi usare non solo per preparare la tua risposta, ma anche come base per una simulazione. Nella fase di pianificazione, una simulazione può essere suddivisa in cinque fasi.

Raccolta delle prove: In questa fase, un team riceverà avvisi tramite diversi mezzi, come, ad esempio, un sistema di ticket interno, avvisi da strumenti di monitoraggio, suggerimenti anonimi o persino notizie pubbliche. I team cominciano quindi a esaminare i log di infrastrutture e applicazioni per stabilire l'origine della compromissione. Questa fase deve coinvolgere anche escalation interne e leadership degli incidenti. Una volta identificate queste informazioni, i team passano al contenimento dell'incidente

Contenimento dell'incidente: I team avranno stabilito che si è verificato un incidente e avranno identificato l'origine del compromissione. I team devono ora agire per contenerlo disabilitando ad esempio le credenziali compromesse, isolando una risorsa di calcolo o revocando l'autorizzazione di un ruolo.

Eliminazione dell'incidente: Ora che l'incidente è stato contenuto, i team lavoreranno per mitigare le vulnerabilità nelle applicazioni o nelle configurazioni dell'infrastruttura che sono state coinvolte nella compromissione. Potrebbe essere necessario ruotare tutte le credenziali utilizzate per un carico di lavoro, modificare le liste di controllo degli accessi (ACL) o cambiare le configurazioni di rete.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Esegui game day: esegui eventi di risposta a incidenti simulati (game day) per minacce diverse che coinvolgono personale e dirigenza chiave.
- Integrazione dei concetti appresi: le lezioni apprese dall'esecuzione di game day devono essere parte del loop di feedback per migliorare i processi.

Risorse

Documenti correlati:

- AWS Incident Response Guide
- AWS Elastic Disaster Recovery (Ripristino di emergenza elastico AWS)

Video correlati:

- DIY guide to runbooks, incident reports, and incident response

Affidabilità

Argomenti

- Fondamenti
- Architettura del carico di lavoro
- Gestione delle modifiche
- Gestione degli errori

Fondamenti

Domande

- REL 1 In che modo gestisci quote e vincoli di servizio?
- REL 2 In che modo pianifichi la topologia di rete?

REL 1 In che modo gestisci quote e vincoli di servizio?

Per le architetture di carichi di lavoro basate sul cloud, esistono quote di servizio (definite anche come restrizioni dei servizi). Queste quote sono presenti per evitare di effettuare accidentalmente il provisioning di più risorse di quelle necessarie e limitare i tassi di richiesta sulle operazioni API in modo da proteggere i servizi da un uso illecito. Esistono anche vincoli di risorse, ad esempio la velocità con cui è possibile trasferire i bit su un cavo in fibra ottica o la quantità di storage su un disco fisico.

Best practice

- REL01-BP01 Consapevolezza su quote e vincoli di servizio
- REL01-BP02 Gestione delle quote di servizio in più account e regioni
- REL01-BP03 Adattamento di quote e vincoli di servizio fissi mediante l'architettura
- REL01-BP04 Monitoraggio e gestione delle quote
- REL01-BP05 Automazione della gestione delle quote
- REL01-BP06 Creazione di un divario sufficiente tra le quote attuali e l'utilizzo massimo per consentire eventuali failover

REL01-BP01 Consapevolezza su quote e vincoli di servizio

Conosci le quote predefinite e le richieste di aumento delle quote per l'architettura del carico di lavoro. Inoltre, sai quali vincoli delle risorse, ad esempio disco o rete, sono potenzialmente influenti.

Service Quotas è un servizio AWS che ti aiuta a gestire le quote per oltre 100 servizi AWS da un'unica posizione. Oltre a cercare i valori delle quote, puoi anche richiedere e monitorare gli aumenti delle quote stesse tramite la console Service Quotas o tramite l'SDK AWS. AWS Trusted Advisor offre un controllo delle quote di servizio che mostra l'utilizzo e le quote per alcuni aspetti di determinati servizi. Le quote predefinite per ciascun servizio sono riportate anche nella rispettiva documentazione di AWS. Consulta ad esempio le quote di Amazon VPC. I limiti di velocità sulle API con throttling vengono impostati all'interno del API Gateway stesso configurando un piano di utilizzo. Altri limiti impostati come configurazione per i rispettivi servizi includono Provisioned IOPS, storage RDS allocato e allocazioni di volumi EBS. Amazon Elastic Compute Cloud (Amazon EC2) dispone di un proprio pannello di controllo sui limiti del servizio che consente di gestire l'istanza, Amazon Elastic Block Store (Amazon EBS) e i limiti degli indirizzi IP elastici. Se hai un caso d'uso in cui le quote di servizio influiscono sulle prestazioni della tua applicazione e non sono adattabili alle tue esigenze, contatta AWS Support per vedere se sono possibili riduzioni.

Anti-pattern comuni:

- Implementazione di un carico di lavoro senza tenere conto delle quote di servizio sui servizi AWS utilizzati.
- Progettazione di un carico di lavoro senza esaminare e soddisfare i vincoli di progettazione dei servizi AWS.
- Implementazione di un carico di lavoro con un utilizzo significativo che sostituisce un carico di lavoro noto esistente senza contattare AWS Support in anticipo.
- Pianificazione di un evento per indirizzare il traffico verso il carico di lavoro, ma senza configurare le quote necessarie o contattare AWS Support in anticipo.

Vantaggi dell'adozione di questa best practice: Essere a conoscenza delle quote di servizio, dei limiti di throttling delle API e dei vincoli di progettazione ti consentirà di tenerne conto nella progettazione, nell'implementazione e nel funzionamento del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Esamina le quote dei servizi AWS nella documentazione pubblicata e in Service Quotas

    - AWS Service Quotas (precedentemente note come restrizioni dei servizi)
- Stabilisci tutti i servizi necessari per il tuo carico di lavoro analizzando il codice di implementazione.
- Utilizza AWS Config per trovare tutte le risorse AWS utilizzate in Account AWS.
    - Tipi di risorse e relazioni tra risorse AWS Config supportate da AWS
- Puoi anche utilizzare AWS CloudFormation per individuare le risorse AWS utilizzate. Esamina le risorse create nella AWS Management Console o tramite il comando list-stack-resources dell'interfaccia a riga di comando. Puoi anche visualizzare le risorse configurate per essere distribuite nel modello stesso.
    - Visualizzazione delle risorse e dei dati dello stack AWS CloudFormation sulla AWS Management Console
    - AWS CLI per CloudFormation: list-stack-resources
- Determina le quote di servizio applicabili. Utilizza le informazioni accessibili in modo programmatico tramite Trusted Advisor e Service Quotas.

Risorse

Documenti correlati:

- Marketplace AWS: prodotti CMDB per il monitoraggio delle restrizioni
- AWS Service Quotas (precedentemente note come restrizioni dei servizi)
- Elenco di controllo delle best practice di AWS Trusted Advisor (consulta la sezione Restrizioni dei servizi)
- AWS limit monitor on AWS answers (Monitoraggio quota AWS su risposte AWS)
- Quote di servizio di Amazon EC2
- What is Service Quotas? (Che cos'è Service Quotas?)

Video correlati:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP02 Gestione delle quote di servizio in più account e regioni

Se utilizzi più Account AWS o Regioni AWS, assicurati di richiedere le quote appropriate in tutti gli ambienti in cui vengono eseguiti i carichi di lavoro di produzione.

Le quote di servizio vengono monitorate per account. Salvo diversa indicazione, ogni quota è specifica della Regione AWS. Oltre agli ambienti di produzione, gestisci anche le quote in tutti gli ambienti non di produzione applicabili, in modo che i test e lo sviluppo non siano ostacolati.

Anti-pattern comuni:

- Consentire l'aumento dell'utilizzo delle risorse in una zona di isolamento senza alcun meccanismo per mantenere la capacità nelle altre.
- Impostazione manuale di tutte le quote in modo indipendente nelle zone di isolamento.
- Non avere la garanzia che le implementazioni isolate a livello regionale siano dimensionate per accogliere l'aumento del traffico da un'altra regione in caso di perdita di un'implementazione.

Vantaggi dell'adozione di questa best practice: Avere la garanzia di poter gestire il carico corrente se una zona di isolamento non è disponibile può aiutare a ridurre il numero di errori che si verificano durante il failover, invece di causare un denial of service ai clienti.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Seleziona gli account e le regioni pertinenti in base ai tuoi requisiti di servizio, di latenza, normativi e di ripristino di emergenza.
- Identifica le quote dei servizi per tutti gli account, le regioni e le zone di disponibilità pertinenti. Le restrizioni si riferiscono ad account e regione.
- What is Service Quotas? (Che cos'è Service Quotas?)

Risorse

Documenti correlati:

- Marketplace AWS: prodotti CMDB per il monitoraggio delle restrizioni
- AWS Service Quotas (precedentemente note come restrizioni dei servizi)
- Elenco di controllo delle best practice di AWS Trusted Advisor (consulta la sezione Restrizioni dei servizi)
- AWS limit monitor on AWS answers (Monitoraggio quota AWS su risposte AWS)
- Quote di servizio di Amazon EC2

- What is Service Quotas? (Che cos'è Service Quotas?)

Video correlati:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP03 Adattamento di quote e vincoli di servizio fissi mediante l'architettura

Considera le quote di servizio immutabili e le risorse fisiche e progetta per evitare che queste compromettano l'affidabilità.

Alcuni esempi includono larghezza di banda di rete, dimensioni di payload di AWS Lambda, velocità di ottimizzazione del throttling per API Gateway e connessioni utente simultanee a un cluster Amazon Redshift.

Anti-pattern comuni:

- Eseguire il benchmarking per un periodo di tempo troppo breve, utilizzando il limite di picco, ma aspettandosi poi che il servizio mantenga tale capacità per periodi prolungati.
- Scegliere un progetto che utilizza una risorsa di un servizio per utente o cliente, ignorando che ci sono vincoli di progettazione che causeranno un errore durante il dimensionamento.

Vantaggi dell'adozione di questa best practice: monitorare le quote fisse nei servizi AWS e i vincoli in altre parti del carico di lavoro, ad esempio vincoli di connettività, vincoli di indirizzo IP e vincoli nei servizi di terze parti, ti consente di capire quando ti stai avvicinando a una quota e di gestirla prima che venga superata.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Essere consapevoli delle quote di servizio fisse Essere consapevoli delle quote di servizio fisse e dei vincoli e progettare in base a questi.
  - AWS Service Quotas

Risorse

Documenti correlati:

- Marketplace AWS: prodotti CMDB per il monitoraggio delle restrizioni
- AWS Service Quotas (precedentemente note come restrizioni dei servizi)
- Elenco di controllo delle best practice di AWS Trusted Advisor (consulta la sezione Restrizioni dei servizi)
- AWS limit monitor on AWS answers (Monitoraggio quota AWS su risposte AWS)
- Quote di servizio di Amazon EC2
- Che cos'è Service Quotas?

Video correlati:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP04 Monitoraggio e gestione delle quote

Valuta il tuo utilizzo potenziale e aumenta le quote in modo appropriato per una crescita pianificata dell'utilizzo.

Per i servizi supportati, puoi gestire le quote configurando gli allarmi CloudWatch affinché monitorino l'utilizzo e ti inviino una notifica in caso di raggiungimento delle quote. Questi allarmi possono essere attivati da Service Quotas o da Trusted Advisor. Puoi anche utilizzare i filtri dei parametri su CloudWatch Logs per cercare ed estrarre modelli nei log al fine di determinare se l'utilizzo è vicino alle soglie delle quote.

Anti-pattern comuni:

- Configurare avvisi che si attivano quando le Service Quotas stanno per essere raggiunte, ma senza avere alcun processo sulle modalità di risposta a un avviso.
- Configurare allarmi solo per i servizi supportati da Service Quotas, escludendo il monitoraggio di altri servizi.

Vantaggi dell'adozione di questa best practice: il monitoraggio automatico delle quote di servizio AWS e il monitoraggio dell'utilizzo rispetto a tali quote ti consentiranno di sapere quando stai per raggiungere il limite di una quota. Puoi anche utilizzare questi dati di monitoraggio per valutare quando puoi ridurre le quote per ridurre i costi.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Monitora e gestisci le quote Valuta l'utilizzo potenziale su AWS, aumenta le quote dei servizi regionali in modo appropriato e consenti una crescita pianificata dell'utilizzo.
  - Acquisisci l'attuale consumo di risorse, ad esempio bucket e istanze. Utilizza le operazioni delle API di servizi come l'API DescribeInstances di Amazon EC2 per raccogliere informazioni sul consumo attuale delle risorse.
  - Acquisisci le quote correnti Utilizza la documentazione di AWS Service Quotas, AWS Trusted Advisor e AWS.
    - AWS Service Quotas è un servizio AWS che ti aiuta a gestire le quote per oltre 100 servizi AWS da un'unica posizione.
    - Utilizza le restrizioni dei servizi di Trusted Advisor per determinare le restrizioni dei servizi attuali.
    - Utilizza le operazioni delle API di servizi per determinare le attuali quote di servizio, quando supportate.
    - Tieni un registro degli aumenti di quota richiesti e del loro stato Dopo l'approvazione di un aumento di quota, assicurati di aggiornare i registri per riflettere la modifica della quota.

Risorse

Documenti correlati:

- Marketplace AWS: prodotti CMDB per il monitoraggio delle restrizioni
- AWS Service Quotas (precedentemente note come restrizioni dei servizi)
- Controlli delle best practice AWS Trusted Advisor per i limiti del servizio
- AWS limit monitor on AWS answers (Monitoraggio quota AWS su risposte AWS)
- Quote di servizio di Amazon EC2
- Che cos'è Service Quotas?
- Monitora Service Quotas utilizzando allarmi Amazon CloudWatch

Video correlati:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP05 Automazione della gestione delle quote

Implementa strumenti per ricevere avvisi quando le soglie stanno per essere raggiunte. Puoi automatizzare le richieste di aumento delle quote utilizzando le API AWS Service Quotas.

Se integri il tuo database di gestione della configurazione (CMDB) o il sistema di ticketing con le Service Quotas, puoi automatizzare il monitoraggio delle richieste di aumento delle quote e delle quote correnti. Oltre all'SDK AWS, Service Quotas offre automazione utilizzando AWS Command Line Interface (AWS CLI).

Anti-pattern comuni:

- Monitoraggio delle quote e dell'utilizzo nei fogli di calcolo.
- Esecuzione di report sull'utilizzo giornaliero, settimanale o mensile e successivo confronto dell'utilizzo con le quote.

Vantaggi dell'adozione di questa best practice: Il monitoraggio automatico delle quote di servizio AWS e il monitoraggio dell'utilizzo rispetto a tale quota ti consentiranno di sapere quando stai per raggiungere una quota. Puoi configurare l'automazione affinché ti aiuti a richiedere un aumento della quota quando necessario. Puoi decidere di ridurre alcune quote quando il tuo utilizzo tende alla direzione opposta per ottenere i vantaggi di riduzione del rischio (in caso di credenziali compromesse) e dei costi.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Impostazione del monitoraggio automatico: implementa strumenti utilizzando gli SDK per ricevere avvisi quando le soglie stanno per essere raggiunte.
  - Utilizza Service Quotas e potenzia il servizio con una soluzione di monitoraggio automatico delle quote come AWS Limit Monitor o un'offerta di Marketplace AWS.
    - What is Service Quotas? (Che cos'è Service Quotas?)
    - Monitoraggio delle quota su AWS – Soluzione AWS
  - Impostazione di risposte attivate in base alle soglie delle quote tramite l'utilizzo delle API di Amazon SNS e AWS Service Quotas.
  - Automazione dei test.
    - Configura le soglie delle restrizioni.

- Integrazione con eventi di modifica di AWS Config, pipeline di implementazione, Amazon EventBridge o terze parti.
- Imposta artificialmente soglie basse per le quote in modo da testare le risposte.
- Configura i trigger per eseguire azioni adeguate in seguito alle notifiche e contatta AWS Support se necessario.
- Attiva manualmente gli eventi di modifica.
- Esegui una giornata di gioco per testare il processo di modifica dell'aumento delle quote.

Risorse

Documenti correlati:

- Partner APN: partner per la gestione della configurazione
- Marketplace AWS: prodotti CMDB per il monitoraggio delle restrizioni
- AWS Service Quotas (precedentemente note come restrizioni dei servizi)
- Elenco di controllo delle best practice di AWS Trusted Advisor (consulta la sezione Restrizioni dei servizi)
- Monitoraggio delle quota su AWS – Soluzione AWS
- Quote di servizio di Amazon EC2
- What is Service Quotas? (Che cos'è Service Quotas?)

Video correlati:

- AWS Live re:Inforce 2019 - Service Quotas

REL01-BP06 Creazione di un divario sufficiente tra le quote attuali e l'utilizzo massimo per consentire eventuali failover

Quando una risorsa presenta un errore, può continuare a essere conteggiata ai fini del raggiungimento delle quote fino a quando non viene terminata correttamente. Assicurati che le quote coprano la sovrapposizione di tutte le risorse non riuscite con sostituzioni prima che le risorse non riuscite vengano terminate. Nel calcolo di questo intervallo dovresti considerare un errore nella zona di disponibilità.

Anti-pattern comuni:

- Impostazione delle quote di servizio in base alle esigenze attuali senza tenere conto degli scenari di failover.

Vantaggi dell'adozione di questa best practice: Quando gli eventi hanno un impatto potenziale sulla disponibilità, il cloud consente di implementare strategie per mitigare o recuperare tali eventi. Queste strategie spesso includono la creazione di risorse aggiuntive per sostituire quelle in errore. La tua strategia di quote deve tenere conto di queste risorse aggiuntive.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Assicurati che ci sia un intervallo sufficiente tra la quota di servizio e l'utilizzo massimo per consentire un eventuale failover.
  - Determina le quote di servizio, specificando i pattern di implementazione, i requisiti di disponibilità e la crescita dei consumi.
  - Richiedi aumenti delle quote, se necessario. Pianifica tenendo conto del tempo necessario affinché le richieste di aumento delle quote siano soddisfatte.
    - Determina i requisiti di affidabilità, chiamati anche "numero di 9".
    - Determina gli scenari di errore (ad esempio, perdita di un componente, una zona di disponibilità o una regione).
    - Stabilisci la metodologia di implementazione (ad esempio, canary, blu/verde, rosso/nero o rolling).
    - Includi un buffer appropriato (ad esempio, 15%) rispetto alla restrizione attuale.
    - Pianifica la crescita dei consumi (ad esempio, monitora le tendenze dei consumi).

Risorse

Documenti correlati:

- Marketplace AWS: prodotti CMDB per il monitoraggio delle restrizioni
- AWS Service Quotas (precedentemente note come restrizioni dei servizi)
- Elenco di controllo delle best practice di AWS Trusted Advisor (consulta la sezione Restrizioni dei servizi)
- Quote di servizio di Amazon EC2
- Che cos'è Service Quotas?

Video correlati:

- AWS Live re:Inforce 2019 - Service Quotas

REL 2 In che modo pianifichi la topologia di rete?

I carichi di lavoro sono spesso presenti in più ambienti. Questi includono più ambienti cloud (sia pubblicamente accessibili sia privati) e, possibilmente, l'infrastruttura del data center esistente. I piani devono includere considerazioni di rete, ad esempio connettività intrasistema e intersistema, gestione di indirizzi IP pubblici, gestione di indirizzi IP privati e risoluzione dei nomi di dominio.

Best practice

- REL02-BP01 Utilizzo di una connettività di rete a disponibilità elevata per gli endpoint pubblici del carico di lavoro
- REL02-BP02 Esecuzione del provisioning di connettività ridondante tra reti private nel cloud e negli ambienti on-premise.
- REL02-BP03 Verifica che l'allocazione delle sottoreti IP consenta l'espansione e la disponibilità:
- REL02-BP04 Preferire topologie hub-and-spoke rispetto a mesh da-molti-a-molti
- REL02-BP05 Applicazione di intervalli di indirizzi IP privati non sovrapposti in tutti gli spazi con indirizzi privati a cui sono connessi

REL02-BP01 Utilizzo di una connettività di rete a disponibilità elevata per gli endpoint pubblici del carico di lavoro

Questi endpoint e il routing verso di essi devono essere altamente disponibili. Per ottenere questo risultato, utilizza DNS ad alta disponibilità, reti di distribuzione di contenuti (CDN), API Gateway, bilanciamento del carico o proxy inversi.

Amazon Route 53, AWS Global Accelerator, Amazon CloudFront, Amazon API Gateway e Elastic Load Balancing (ELB) offrono tutti endpoint pubblici altamente disponibili. Puoi anche scegliere di valutare le appliance software di Marketplace AWS per il bilanciamento del carico e il proxy.

I consumatori del servizio fornito dal carico di lavoro, che siano utenti finali o altri servizi, effettuano richieste su questi endpoint del servizio. Sono disponibili diverse risorse AWS che ti consentono di fornire endpoint a disponibilità elevata.

Elastic Load Balancing fornisce bilanciamento del carico tra le zone di disponibilità, esegue l'instradamento di livello 4 (TCP) o 7 (http/https) e si integra con AWS WAF e con AWS Auto Scaling

per contribuire a creare un'infrastruttura con riparazione automatica e assorbire gli aumenti di traffico, mentre rilascia risorse quando questo diminuisce.

Amazon Route 53 è un servizio del sistema di nomi di dominio (DNS) scalabile e altamente disponibile che collega le richieste degli utenti all'infrastruttura in esecuzione in AWS, come istanze Amazon EC2, load balancer Elastic Load Balancing o bucket Amazon S3 e può essere utilizzato anche per instradare gli utenti a un'infrastruttura esterna ad AWS.

AWS Global Accelerator è un servizio a livello di rete che puoi utilizzare per indirizzare il traffico verso endpoint ottimali sulla rete globale AWS.

Gli attacchi DDoS (Distributed Denial of Service) rischiano di chiudere il traffico legittimo e di ridurre la disponibilità per gli utenti. AWS Shield fornisce protezione automatica da questi attacchi senza costi aggiuntivi per gli endpoint del servizio AWS sul carico di lavoro. Puoi potenziare queste caratteristiche con appliance virtuali dei partner APN e di Marketplace AWS per soddisfare le tue esigenze.

Anti-pattern comuni:

- Utilizzo di indirizzi Internet pubblici su istanze o container e gestione della connettività tramite DNS.
- Utilizzo degli indirizzi del protocollo Internet anziché dei nomi di dominio per l'individuazione dei servizi.
- Fornitura di contenuti (pagine Web, asset statici, file multimediali) a un'area geografica di grandi dimensioni senza l'utilizzo di una rete di distribuzione di contenuti.

Vantaggi dell'adozione di questa best practice: Implementando servizi ad alta disponibilità nel carico di lavoro, ti assicuri che il carico di lavoro sarà disponibile per i tuoi utenti.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Assicurati di avere una connettività altamente disponibile per gli utenti del carico di lavoro. Amazon Route 53, AWS Global Accelerator, Amazon CloudFront, Amazon API Gateway e Elastic Load Balancing (ELB) forniscono tutti endpoint rivolti al pubblico altamente disponibili. Puoi anche scegliere di valutare le appliance software di Marketplace AWS per il bilanciamento del carico e il proxy.

- Assicurati di avere una connessione altamente disponibile per i tuoi utenti.
- Accertati di utilizzare un DNS altamente disponibile per gestire i nomi di dominio degli endpoint delle applicazioni.

    - Se gli utenti accedono alla tua applicazione tramite Internet, utilizza le operazioni delle API di servizio per confermare il corretto utilizzo degli Internet gateway. Assicurati inoltre che le voci delle tabelle di routing per le sottoreti che ospitano gli endpoint dell'applicazione siano corrette.
        - DescribeInternetGateways
        - DescribeRouteTables
- Assicurati di utilizzare un proxy inverso o un load balancer altamente disponibile prima dell'applicazione.
    - Se gli utenti accedono all'applicazione tramite l'ambiente on-premise, verifica che la connettività tra quest'ultimo e AWS sia altamente disponibile.
    - Utilizza Route 53 per gestire i nomi di dominio.
        - Che cos'è Amazon Route 53?
    - Utilizza un provider DNS di terze parti che soddisfi i tuoi requisiti.
    - Utilizza Elastic Load Balancing.
        - What is Elastic Load Balancing? (Che cos'è Elastic Load Balancer)
    - Utilizza un'appliance di Marketplace AWS che soddisfi i tuoi requisiti.

Risorse

Documenti correlati:

- Partner APN: partner per la pianificazione della rete
- Suggerimenti sulla resilienza di AWS Direct Connect
- Marketplace AWS per l'infrastruttura di rete
- Whitepaper: Opzioni di connettività di Amazon Virtual Private Cloud
- Connettività di rete di elevata disponibilità in più data center
- Utilizzo del kit di strumenti di resilienza di Direct Connect per iniziare
- Endpoint VPC e servizi di endpoint VPC (AWS PrivateLink)
- Che cos'è AWSGlobal Accelerator?
- Che cos'è Amazon VPC?
- Che cos'è un Transit Gateway?
- Che cos'è Amazon CloudFront?
- Che cos'è Amazon Route 53?

- What is Elastic Load Balancing? (Che cos'è Elastic Load Balancer)
- Lavorare con gateway Direct Connect

Video correlati:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (Progettazione avanzata di VPC e nuove funzionalità per Amazon VPC) (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (Architetture di riferimento del Gateway di transito AWS per molte VPC) (NET406-R1)

REL02-BP02 Esecuzione del provisioning di connettività ridondante tra reti private nel cloud e negli ambienti on-premise.

Utilizza più connessioni AWS Direct Connect o tunnel VPN tra reti private implementate separatamente. Utilizza più ubicazioni Direct Connect per un'elevata disponibilità. Se utilizzi più Regioni AWS, garantisci la ridondanza in almeno due di esse. È possibile valutare le appliance Marketplace AWS che terminano le VPN. Se utilizzi appliance di Marketplace AWS, distribuisci le istanze ridondanti per la disponibilità elevata in diverse zone di disponibilità.

AWS Direct Connect è un servizio cloud che semplifica la creazione di una connessione di rete dedicata dall'ambiente on-premise ad AWS. Utilizzando il gateway Direct Connect, il data center on-premise può essere collegato a più VPC AWS distribuiti in più Regioni AWS.

Questa ridondanza risolve possibili errori che condizionano la resilienza della connettività:

- Come pensi di essere resiliente ai fallimenti nella topologia?
- Cosa succede se configuri qualcosa in modo errato e rimuovi la connettività?
- Sarai in grado di gestire un inaspettato aumento del traffico o dell'utilizzo dei tuoi servizi?
- Sarai in grado di assorbire un tentativo di attacco DDoS (Distributed Denial of Service)?

Quando si connette il VPC al data center in locale tramite VPN, si devono considerare i requisiti di resilienza e larghezza di banda necessari quando si seleziona la dimensione del fornitore e dell'istanza su cui è necessario eseguire l'appliance. Se si utilizza un'appliance VPN non resiliente nella sua implementazione, è necessario disporre di una connessione ridondante tramite una seconda appliance. Per tutti questi scenari, è necessario definire un orario accettabile per il ripristino e il test per garantire che sia possibile soddisfare tali requisiti.

Se scegli di connettere il VPC al data center utilizzando una connessione Direct Connect e hai bisogno che questa connessione sia altamente disponibile, predisponi connessioni Direct Connect ridondanti da ogni data center. La connessione ridondante dovrebbe utilizzare una seconda connessione Direct Connect da una posizione diversa rispetto alla prima. Se disponi di più data center, assicurati che le connessioni terminino in posizioni diverse. Utilizza il Kit di strumenti di resilienza Direct Connect come ausilio per la configurazione.

Se scegli di eseguire il failover sul VPN su Internet utilizzando AWS VPN, è importante capire che supporta fino a 1,25 Gbps di velocità di trasmissione effettiva per tunnel VPN, ma non supporta Equal Cost Multi Path (ECMP) per il traffico in uscita nel caso di più tunnel VPN gestiti da AWS che terminano sullo stesso gateway privato virtuale (VGW). Non è consigliabile utilizzare VPN gestite da AWS come backup per le connessioni Direct Connect, a meno che non sia possibile tollerare velocità inferiori a 1 Gbps durante il failover.

Puoi anche utilizzare gli endpoint VPC per connettere privatamente il tuo VPC ai servizi AWS supportati e ai servizi endpoint VPC basati su AWS PrivateLink senza dover attraversare la rete Internet pubblica. Gli endpoint sono dispositivi virtuali. Sono componenti VPC a scalabilità orizzontale, ridondanti e ad alta disponibilità. Consentono la comunicazione tra le istanze nel VPC e i servizi senza imporre rischi di disponibilità o vincoli di larghezza di banda sul traffico di rete.

Anti-pattern comuni:

- Avere un solo provider di connettività tra la rete in locale e AWS.
- Utilizzare le funzionalità di connettività della connessione AWS Direct Connect, ma con una sola connessione.
- Disporre di un solo percorso per la connettività VPN.

Vantaggi dell'adozione di questa best practice: implementando una connettività ridondante tra il tuo ambiente cloud e l'ambiente aziendale/on-premise, puoi garantire che i servizi dipendenti tra i due ambienti possano comunicare in maniera affidabile.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Garantire una connettività altamente disponibile tra AWS e l'ambiente on-premise. Utilizza più connessioni AWS Direct Connect o tunnel VPN tra reti private implementate separatamente. Utilizza più ubicazioni Direct Connect per un'elevata disponibilità. Se utilizzi più Regioni AWS, garantisci la ridondanza in almeno due di esse. È possibile valutare le appliance Marketplace AWS

che terminano le VPN. Se utilizzi appliance di Marketplace AWS, distribuisci le istanze ridondanti per la disponibilità elevata in diverse zone di disponibilità.

- Assicurati di avere una connessione ridondante con l'ambiente on-premise Potresti aver bisogno di connessioni ridondanti a più Regioni AWS per soddisfare le tue esigenze di disponibilità.
  - Suggerimenti sulla resilienza di AWS Direct Connect
  - Utilizzo di connessioni VPN da sito a sito ridondanti per fornire il failover
    - Utilizza le operazioni delle API di servizi per identificare l'utilizzo corretto dei circuiti Direct Connect.
      - DescribeConnections
      - DescribeConnectionsOnInterconnect
      - DescribeDirectConnectGatewayAssociations
      - DescribeDirectConnectGatewayAttachments
      - DescribeDirectConnectGateways
      - DescribeHostedConnections
      - DescribeInterconnects
    - Se esiste una sola connessione Direct Connect o se non ne hai nessuna, crea dei tunnel VPN ridondanti verso i tuoi gateway privati virtuali (VGW).
      - Cos'è VPN sito-sito AWS?
- Acquisisci la tua attuale connettività (ad esempio, Direct Connect, gateway privati virtuali, appliance Marketplace AWS).
  - Utilizza le operazioni delle API di servizi per eseguire la query della configurazione delle connessioni Direct Connect.
    - DescribeConnections
    - DescribeConnectionsOnInterconnect
    - DescribeDirectConnectGatewayAssociations
    - DescribeDirectConnectGatewayAttachments
    - DescribeDirectConnectGateways
    - DescribeHostedConnections
    - DescribeInterconnects
  - Utilizza le operazioni delle API di servizi per raccogliere i gateway privati virtuali (VGW) dove vengono utilizzati dalle tabelle di instradamento.
    - DescribeVpnGateways

    - DescribeRouteTables
- Utilizza le operazioni delle API di servizi per raccogliere le applicazioni di Marketplace AWS dove vengono utilizzate dalle tabelle di instradamento.
    - DescribeRouteTables

Risorse

Documenti correlati:

- Partner APN: partner per la pianificazione della rete
- Suggerimenti sulla resilienza di AWS Direct Connect
- Marketplace AWS per l'infrastruttura di rete
- Whitepaper: Opzioni di connettività di Amazon Virtual Private Cloud
- Connettività di rete di elevata disponibilità in più data center
- Utilizzo di connessioni VPN da sito a sito ridondanti per fornire il failover
- Utilizzo del kit di strumenti di resilienza di Direct Connect per iniziare
- Endpoint VPC e servizi di endpoint VPC (AWS PrivateLink)
- Che cos'è Amazon VPC?
- Che cos'è un Transit Gateway?
- Cos'è VPN sito-sito AWS?
- Lavorare con gateway Direct Connect

Video correlati:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (Progettazione avanzata di VPC e nuove funzionalità per Amazon VPC) (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (Architetture di riferimento del Gateway di transito AWS per molte VPC) (NET406-R1)

REL02-BP03 Verifica che l'allocazione delle sottoreti IP consenta l'espansione e la disponibilità:

Gli intervalli di indirizzi IP dei Amazon VPC devono essere sufficientemente ampi per soddisfare i requisiti del carico di lavoro, tenendo conto anche dell'espansione futura e dell'allocazione degli

indirizzi IP alle sottoreti nelle zone di disponibilità. Sono inclusi sistemi di bilanciamento del carico, istanze EC2 e applicazioni basate su container.

Quando si pianifica la topologia di rete, il primo passo è definire lo spazio stesso degli indirizzi IP. Gli intervalli di indirizzi IP privati (secondo le linee guida RFC 1918) dovrebbero essere allocati per ogni VPC. Nell'ambito di questo processo, soddisfa i seguenti requisiti:

- Lascia spazi per indirizzi IP per più di un VPC per Regione.
- All'interno di un VPC, lascia spazio per più sottoreti che coprono più zone di disponibilità.
- Lascia sempre spazio per un blocco CIDR inutilizzato all'interno di un VPC per un'espansione futura.
- Assicurati che sia disponibile spazio per gli indirizzi IP, al fine di soddisfare le esigenze di qualsiasi parco istanze EC2 transitorio che puoi utilizzare, ad esempio parchi istanze Spot per il machine learning, cluster Amazon EMR o cluster Amazon Redshift.
- Tieni presente che i primi quattro indirizzi IP e l'ultimo indirizzo IP in ogni blocco CIDR della sottorete sono riservati e non disponibili per l'uso.
- È consigliabile pianificare la distribuzione di blocchi CIDR VPC di grandi dimensioni. Tieni presente che il blocco CIDR VPC iniziale allocato al VPC non può essere modificato o eliminato, ma puoi aggiungere ulteriori blocchi CIDR non sovrapposti al VPC. I CIDR IPv4 della sottorete non possono essere modificati, mentre ciò è possibile con i CIDR IPv6. Tieni presente che la distribuzione del VPC più grande possibile (/16) genera oltre 65.000 indirizzi IP. Solo nello spazio degli indirizzi IP di base 10.x.x.x potresti effettuare il provisioning di 255 VPC di questo tipo. Pertanto, dovresti peccare per eccesso piuttosto che per difetto per semplificare la gestione dei VPC.

Anti-pattern comuni:

- Creazione di VPC di piccole dimensioni.
- Creare sottoreti di piccole dimensioni e dover quindi aggiungere sottoreti alle configurazioni man mano che cresci.
- Stima erronea del numero di indirizzi IP che un elastic load balancer può utilizzare.
- Distribuzione di numerosi sistemi di bilanciamento del carico a traffico elevato nelle stesse sottoreti.

Vantaggi dell'adozione di questa best practice: In questo modo puoi consentire la crescita dei carichi di lavoro e continuare a fornire disponibilità man mano che incrementi le dimensioni.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Pianificazione della rete in base a crescita, compliance normativa e integrazione con altre reti. Senza una pianificazione adeguata, la crescita può essere sottovalutata, la compliance normativa può cambiare e l'implementazione di acquisizioni o di connessioni a reti private può rivelarsi difficile.
  - Seleziona gli Account AWS e le Regioni pertinenti in base ai tuoi requisiti di servizio, di latenza, normativi e di ripristino di emergenza.
  - Identifica le esigenze delle implementazioni di VPC regionali.
  - Identifica le dimensioni dei VPC.
    - Stabilisci se intendi implementare connettività multi-VPC.
      - Che cos'è un Transit Gateway?
      - Connettività multi-VPC a singola Regione
    - Stabilisci se hai bisogno di reti separate a causa di requisiti normativi.
    - Fai in modo che i VPC abbiano le dimensioni maggiori possibili. Il blocco CIDR VPC iniziale allocato al VPC non può essere modificato o eliminato, ma puoi aggiungere ulteriori blocchi CIDR non sovrapposti al VPC. Tuttavia, questo potrebbe frammentare gli intervalli degli indirizzi.
    - Fai in modo che i VPC abbiano le dimensioni maggiori possibili. Il blocco CIDR VPC iniziale allocato al VPC non può essere modificato o eliminato, ma puoi aggiungere ulteriori blocchi CIDR non sovrapposti al VPC. Tuttavia, questo potrebbe frammentare gli intervalli degli indirizzi.

Risorse

Documenti correlati:

- Partner APN: partner per la pianificazione della rete
- Marketplace AWS per l'infrastruttura di rete
- Whitepaper: Opzioni di connettività di Amazon Virtual Private Cloud
- Connettività di rete di elevata disponibilità in più data center
- Connettività multi-VPC a singola Regione
- Che cos'è Amazon VPC?

Video correlati:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (Progettazione avanzata di VPC e nuove funzionalità per Amazon VPC) (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (Architetture di riferimento del Gateway di transito AWS per molte VPC) (NET406-R1)

REL02-BP04 Preferire topologie hub-and-spoke rispetto a mesh da-molti-a-molti

Se più di due spazi di indirizzi di rete (ad esempio, VPC e reti on-premise) sono connessi tramite peering VPC, AWS Direct Connect o VPN, utilizza un modello hub-and-spoke, come quello fornito da AWS Transit Gateway.

Se disponi solo di due reti di questo tipo, puoi semplicemente connetterle tra loro, tuttavia, man mano che il numero di reti cresce, la complessità di tali connessioni mesh diventa insostenibile. AWS Transit Gateway offre un modello hub-and-spoke di facile manutenzione, consentendo l'instradamento del traffico su più reti.

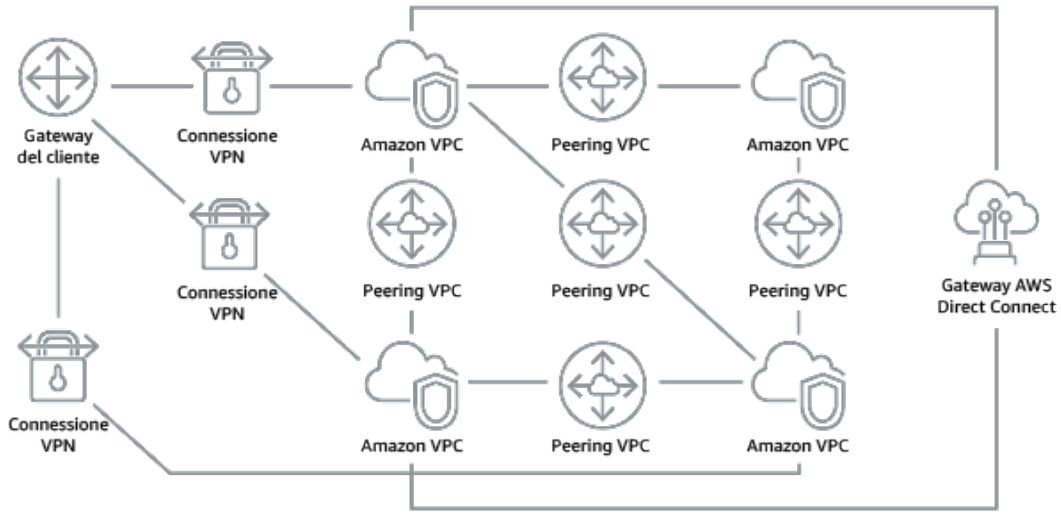


Figura 1. Senza AWS Transit Gateway, è necessario eseguire il peering tra Amazon VPC e tra ciascuna ubicazione locale utilizzando una connessione VPN, che può diventare complessa man mano che si dimensiona.

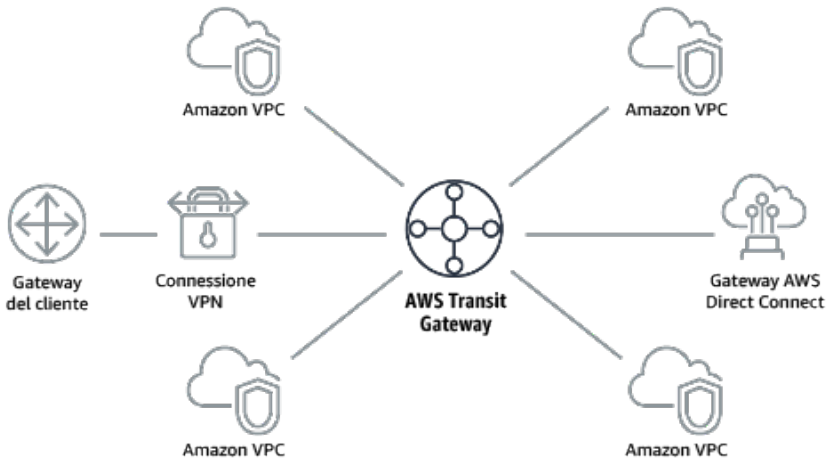


Figura 2. Con AWS Transit Gateway: è sufficiente connettere ogni Amazon VPC o VPN ad AWS Transit Gateway e instradare il traffico da e verso ogni VPC o VPN.

Anti-pattern comuni:

- Utilizzo del peering VPC per connettere più di due VPC.
- Creazione di più sessioni BGP per ogni VPC per stabilire una connettività che si estende su cloud privati virtuali (VPC, Virtual Private Cloud) distribuiti in più Regioni AWS.

Vantaggi dell'adozione di questa best practice: Man mano che il numero di reti cresce, la complessità di tali connessioni mesh diventa insostenibile. AWS Transit Gateway offre un modello hub-and-spoke di facile manutenzione, consentendo l'instradamento del traffico su più reti.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Preferire topologie hub-and-spoke rispetto a mesh da-molti-a-molti. Se più di due spazi di indirizzi di rete (VPC, reti on-premise) sono connessi tramite peering VPC, AWS Direct Connect o VPN, utilizza un modello hub-and-spoke, come quello fornito da AWS Transit Gateway.
  - Se disponi solo di due reti di questo tipo, puoi semplicemente connetterle tra loro, tuttavia, man mano che il numero di reti cresce, la complessità di tali connessioni mesh diventa insostenibile. AWS Transit Gateway offre un modello hub-and-spoke di facile manutenzione, consentendo l'instradamento del traffico su più reti.
    - Che cos'è un Transit Gateway?

Risorse

Documenti correlati:

- Partner APN: partner per la pianificazione della rete
- Marketplace AWS per l'infrastruttura di rete
- Connettività di rete di elevata disponibilità in più data center
- Endpoint VPC e servizi di endpoint VPC (AWS PrivateLink)
- Che cos'è Amazon VPC?
- Che cos'è un Transit Gateway?

Video correlati:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (Progettazione avanzata di VPC e nuove funzionalità per Amazon VPC) (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (Architetture di riferimento del Gateway di transito AWS per molte VPC) (NET406-R1)

REL02-BP05 Applicazione di intervalli di indirizzi IP privati non sovrapposti in tutti gli spazi con indirizzi privati a cui sono connessi

Gli intervalli di indirizzi IP di ogni VPC non devono sovrapporsi quando collegati in peering o connessi tramite VPN. Analogamente, è necessario evitare conflitti di indirizzi IP tra un VPC e ambienti in locale o con altri provider di servizi cloud utilizzati. Bisogna inoltre disporre di un modo per allocare gli intervalli di indirizzi IP privati quando necessario.

Un sistema di gestione degli indirizzi IP (IPAM) può aiutarti in questo. Su Marketplace AWS sono disponibili diversi IPAM.

Anti-pattern comuni:

- Utilizzo nel VPC dello stesso intervallo IP utilizzato in locale o nella rete aziendale.
- Non tenere traccia degli intervalli IP dei VPC utilizzati per distribuire i carichi di lavoro.

Vantaggi dell'adozione di questa best practice: La pianificazione attiva della rete garantisce di non avere più occorrenze dello stesso indirizzo IP nelle reti interconnesse. In questo modo si evitano problemi di instradamento in parti del carico di lavoro che utilizzano le diverse applicazioni.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Monitora e gestisci l'uso di CIDR. Valuta il tuo utilizzo potenziale su AWS, aggiungi intervalli CIDR ai VPC esistenti e crea i VPC per consentire la crescita pianificata dell'utilizzo.
  - Acquisisci il consumo attuale di CIDR (ad esempio, VPC e sottoreti)
    - Utilizza le operazioni delle API di servizi per raccogliere il consumo attuale di CIDR.
  - Acquisisci l'utilizzo attuale delle sottoreti.
    - Utilizza le operazioni delle API di servizio per raccogliere le sottoreti per VPC in ogni Regione.
      - DescribeSubnets
    - Registra l'uso attuale.
    - Verifica se hai creato intervalli di indirizzi IP sovrapposti.
    - Calcola la capacità inutilizzata.
    - Individua gli intervalli di indirizzi IP sovrapposti. Puoi eseguire la migrazione a un nuovo intervallo di indirizzi o utilizzare le appliance NAT (Network and Port Translation) di Marketplace AWS se hai l'esigenza di connettere gli intervalli sovrapposti.

Risorse

Documenti correlati:

- Partner APN: partner per la pianificazione della rete
- Marketplace AWS per l'infrastruttura di rete
- Whitepaper: Opzioni di connettività di Amazon Virtual Private Cloud
- Connettività di rete di elevata disponibilità in più data center
- Che cos'è Amazon VPC?
- Che cos'è IPAM?

Video correlati:

- AWS re:Invent 2018: Advanced VPC Design and New Capabilities for Amazon VPC (Progettazione avanzata di VPC e nuove funzionalità per Amazon VPC) (NET303)
- AWS re:Invent 2019: AWS Transit Gateway reference architectures for many VPCs (Architetture di riferimento del Gateway di transito AWS per molte VPC) (NET406-R1)

Architettura del carico di lavoro

Domande

- REL 3 In che modo progetti l'architettura del servizio di carico di lavoro?
- REL 4 In che modo progetti le interazioni in un sistema distribuito per evitare errori?
- REL 5 In che modo progetti le interazioni in un sistema distribuito per mitigare o affrontare gli errori?

REL 3 In che modo progetti l'architettura del servizio di carico di lavoro?

Creazione di carichi di lavoro altamente scalabili e affidabili utilizzando un'architettura orientata ai servizi (SOA) o un'architettura di microservizi. L'architettura orientata ai servizi (SOA) è la pratica di rendere i componenti software riutilizzabili tramite interfacce di servizio. L'architettura dei microservizi va oltre, per rendere i componenti più piccoli e semplici.

Best practice

- REL03-BP01 Scelta del tipo di segmentazione del carico di lavoro
- REL03-BP02 Creazione di servizi focalizzati su domini e funzionalità aziendali specifici
- REL03-BP03 Fornitura di contratti di servizio per API

REL03-BP01 Scelta del tipo di segmentazione del carico di lavoro

La segmentazione del carico di lavoro è importante quando vengono determinati i requisiti di resilienza dell'applicazione. L'architettura monolitica deve essere evitata se possibile. Valuta invece con particolare attenzione quali componenti dell'applicazione possono essere suddivisi in microservizi. A seconda dei requisiti dell'applicazione, ciò potrebbe risultare in una combinazione di architettura orientata ai servizi (SOA) e microservizi, laddove possibile. I carichi di lavoro stateless sono maggiormente idonei a essere implementati come microservizi.

Risultato desiderato: i carichi di lavoro devono essere supportabili, scalabili e devono essere caratterizzati dalla minore interdipendenza possibile.

Quando scegli come segmentare il carico di lavoro, trova il giusto compromesso tra i vantaggi e le complessità. Ciò che è giusto per un nuovo prodotto al primo lancio è diverso dai requisiti di un carico di lavoro creato per ridimensionare le risorse. Durante la rifattorizzazione (riprogettazione) di un monolito, dovrai considerare la capacità dell'applicazione di supportare la suddivisione in

servizi stateless. La suddivisione dei servizi in elementi più piccoli consente a team ristretti e ben definiti di svilupparli e gestirli. Tuttavia, servizi di piccole dimensioni possono introdurre complessità, che includono un eventuale aumento della latenza, un debug più complesso e un maggiore carico operativo.

Anti-pattern comuni:

- Il microservizio Death Star rappresenta una situazione in cui i componenti atomici diventano così interdipendenti che un errore verificatosi in un componente genera un errore molto più grande, rendendo i componenti rigidi e fragili se considerati come monolito.

Vantaggi dell'adozione di questa best practice:

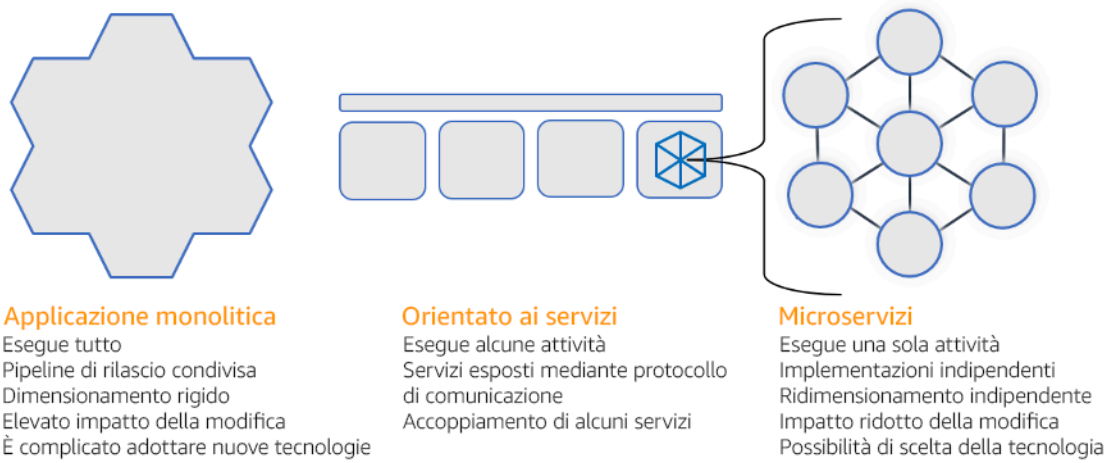
- Segmenti più specifici comportano maggiore agilità, flessibilità organizzativa e scalabilità.
- Riduzione dell'impatto derivante dall'interruzione dei servizi.
- I componenti dell'applicazione possono avere requisiti di disponibilità diversi, che a loro volta possono essere supportati da una segmentazione più atomica.
- Responsabilità ben definite per i team che supportano il carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

Scegli il tipo di architettura in base al tipo di segmentazione del carico di lavoro. Scegli una SOA o un'architettura di microservizi (o, in alcuni rari casi, un'architettura monolitica). Anche se scegli di iniziare con un'architettura monolitica, devi assicurarti che sia modulare e possa evolvere in SOA o microservizi man mano che il prodotto si dimensiona con l'adozione da parte degli utenti. La SOA e i microservizi offrono rispettivamente una segmentazione più piccola, preferita come architettura moderna scalabile e affidabile, ma ci sono compromessi da considerare soprattutto quando si distribuisce un'architettura di microservizi.

Uno dei principali compromessi è che ora disponi di un'architettura di calcolo distribuita che può rendere più difficile il raggiungimento dei requisiti di latenza degli utenti ed è presente un'ulteriore complessità nel debug e nel tracciamento delle interazioni degli utenti. Puoi utilizzare AWS X-Ray per risolvere questo problema. Un altro effetto da considerare è l'aumento della complessità operativa man mano che aumenta il numero di applicazioni che gestisci, che richiede la distribuzione di più componenti di indipendenza.



Architettura monolitica, orientata ai servizi e di microservizi

Passaggi dell'implementazione

- Determina l'architettura più appropriata per rifattorizzare (riprogettare) o creare l'applicazione. SOA e microservizi offrono segmentazione rispettivamente di dimensioni minori, preferita in quanto architettura moderna, scalabile e affidabile. SOA può essere un buon compromesso per ottenere una segmentazione di dimensioni minori, evitando al contempo alcune delle complessità dei microservizi. Per ulteriori dettagli, consulta I compromessi dei microservizi.
- Se il carico di lavoro è adatto e la tua organizzazione può supportarla, è consigliabile utilizzare un'architettura di microservizi per ottenere la massima agilità e affidabilità. Per ulteriori dettagli, consulta Implementing Microservices on AWS (Implementazione di microservizi in AWS).
- Considera l'ipotesi di attenerti al modello Strangler Fig per eseguire la rifattorizzazione (riprogettazione) di un monolito in componenti più piccoli. Ciò comporta la graduale sostituzione di componenti specifici dell'applicazione con nuove applicazioni e nuovi servizi. AWS Migration Hub Refactor Spaces funge da punto di partenza per la rifattorizzazione incrementale. Per ulteriori dettagli, consulta Seamlessly migrate on-premises legacy workloads using a strangler pattern (Migrazione senza problemi di carichi di lavoro legacy on-premise mediante un modello Strangler).
- L'implementazione di microservizi può richiedere un meccanismo di individuazione dei servizi per consentire ai servizi distribuiti di comunicare tra loro. AWS App Mesh può essere utilizzato con architetture orientate ai servizi per offrire rilevamento e accesso affidabili ai servizi. AWS Cloud Map può inoltre essere utilizzato per il rilevamento dinamico dei servizi basato su DNS.
- In caso di migrazione da un monolito a una SOA, Amazon MQ può aiutare a colmare il divario come bus del servizio durante la riprogettazione delle applicazioni legacy nel cloud.

- Per i monoliti esistenti con un unico database condiviso, scegli come riorganizzare i dati in segmenti più piccoli. Questa riorganizzazione può avvenire per unità aziendale, schema di accesso o struttura dei dati. A questo punto del processo di rifattorizzazione (riprogettazione), deve orientare la scelta verso un database di tipo relazionale o non relazionale (NoSQL). Per ulteriori dettagli, consulta From SQL to NoSQL (Da SQL a NoSQL).

Livello di impegno per il piano di implementazione: alto

Risorse

Best practice correlate:

- REL03-BP02 Creazione di servizi focalizzati su domini e funzionalità aziendali specifici

Documenti correlati:

- Amazon API Gateway: configurazione di una REST API mediante OpenAPI
- Cosa si intende per architettura orientata ai servizi?
- Bounded Context (un modello centrale in Domain-Driven Design)
- Implementazione di microservizi in AWS
- I compromessi dei microservizi
- Microservizi: una definizione di questo nuovo termine di architettura
- Implementazione di microservizi in AWS
- What is AWS App Mesh? (Che cos'è AWS App Mesh?)

Esempi correlati:

- Iterative App Modernization Workshop (Workshop sulla modernizzazione delle applicazioni interattive)

Video correlati:

- Delivering Excellence with Microservices on AWS (Implementazione dell'eccellenza con i microservizi in AWS)

REL03-BP02 Creazione di servizi focalizzati su domini e funzionalità aziendali specifici

L'architettura orientata ai servizi (SOA) crea servizi con funzioni ben delineate definite dalle esigenze aziendali. I microservizi utilizzano modelli di dominio e contesto delimitato per restringere ulteriormente questa operazione, in modo che ogni servizio esegua una sola operazione. Focalizzarsi su funzionalità specifiche consente di differenziare i requisiti di affidabilità dei diversi servizi e mirare agli investimenti in modo più specifico. Un problema aziendale conciso e l'associazione di un piccolo team a ciascun servizio facilitano il dimensionamento dell'organizzazione.

Nella progettazione di un'architettura di microservizi, è utile impiegare Domain-Driven Design (DDD) per modellare il problema aziendale utilizzando le entità. Ad esempio, per il sito Web Amazon.com, le entità possono includere pacchetti, consegna, pianificazione, prezzo, sconto e valuta. Quindi il modello viene ulteriormente suddiviso in modelli più piccoli utilizzando il contesto delimitato, dove le entità che condividono caratteristiche e attributi simili vengono raggruppate insieme. Pertanto, utilizzando il pacchetto di esempio di Amazon.com, la consegna e la pianificazione sarebbero parte del contesto di spedizione, mentre il prezzo, lo sconto e la valuta fanno parte del contesto dei prezzi. Con il modello diviso in contesti, emerge un modello su come delimitare i microservizi.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Progetta il carico di lavoro in base ai domini aziendali e alle loro rispettive funzionalità. Focalizzarsi su funzionalità specifiche consente di differenziare i requisiti di affidabilità dei diversi servizi e mirare agli investimenti in modo più specifico. Un problema aziendale conciso e l'associazione di un piccolo team a ciascun servizio facilitano il dimensionamento dell'organizzazione.
  - Esegui l'analisi di dominio per mappare una progettazione basata sul dominio (DDD, domain-driven design) per il carico di lavoro. In seguito, puoi scegliere un tipo di architettura per soddisfare le esigenze del carico di lavoro.
    - How to break a Monolith into Microservices (Come trasformare un monolite in microservizi)
    - Getting Started with DDD when Surrounded by Legacy Systems (Iniziare con il DDD quando si è circondati da sistemi legacy)
    - Eric Evans "Domain-Driven Design: Tackling Complexity in the Heart of Software"

    - Implementazione di microservizi in AWS
- Suddividi i tuoi servizi in componenti più piccoli possibile. Con l'architettura di microservizi, puoi dividere il tuo carico di lavoro in componenti dotati della funzionalità minima per consentire agilità e ridimensionamento dell'organizzazione.
  - Definisci l'API per il carico di lavoro e i suoi obiettivi di progettazione, limiti e qualsiasi altra considerazione per l'uso.
    - Definizione dell'API.
      - La definizione dell'API deve consentire la crescita e parametri aggiuntivi.
    - Definizione delle disponibilità progettate.
      - La tua API può avere più obiettivi di progettazione per funzioni differenti.
    - Definizione di limiti
      - Esegui test per definire i limiti delle tue capacità di carico di lavoro.

Risorse

Documenti correlati:

- Amazon API Gateway: configurazione di una REST API mediante OpenAPI
- Bounded Context (un modello centrale in Domain-Driven Design)
- Eric Evans "Domain-Driven Design: Tackling Complexity in the Heart of Software"
- Getting Started with DDD when Surrounded by Legacy Systems (Iniziare con il DDD quando si è circondati da sistemi legacy)
- How to break a Monolith into Microservices (Come trasformare un monolite in microservizi)
- Implementazione di microservizi in AWS
- I compromessi dei microservizi
- Microservizi: una definizione di questo nuovo termine di architettura
- Implementazione di microservizi in AWS

REL03-BP03 Fornitura di contratti di servizio per API

I contratti di servizio sono accordi documentati tra i team sull'integrazione dei servizi e includono una definizione API leggibile dal computer, limiti di velocità e aspettative di prestazioni. Una strategia di controllo delle versioni consente ai clienti di continuare a utilizzare l'API esistente e migrare le

applicazioni all'API più recente quando sono pronte. La distribuzione può avvenire in qualsiasi momento, purché il contratto non venga violato. Il team del fornitore di servizi può utilizzare lo stack tecnologico scelto per soddisfare il contratto API. Analogamente, l'utente del servizio può utilizzare la propria tecnologia.

I microservizi portano il concetto dell'architettura orientata ai servizi (SOA) al punto della creazione di servizi che hanno una serie minima di funzionalità. Ogni servizio pubblica un'API e obiettivi di progettazione, limiti e altre considerazioni per l'utilizzo del servizio. Questo stabilisce un contratto con le applicazioni di chiamata. Questo comporta tre vantaggi principali:

- Il servizio ha un problema aziendale circoscritto da risolvere e un piccolo team proprietario del problema aziendale. Questo consente un miglior ridimensionamento organizzativo.
- Ciascun team può effettuare un'implementazione in qualsiasi momento purché questa soddisfi i rispettivi requisiti "contrattuali" e dell'API.
- Il team può utilizzare qualsiasi stack tecnologico a condizione che soddisfi le proprie API e altri requisiti di "contratto".

Amazon API Gateway è un servizio completamente gestito che semplifica agli sviluppatori la creazione, la pubblicazione, la manutenzione, il monitoraggio e la protezione delle API su qualsiasi scala. Gestisce tutte le attività coinvolte nell'accettazione e nell'elaborazione di fino a centinaia di migliaia di chiamate API simultanee, tra cui la gestione del traffico, il controllo delle autorizzazioni e degli accessi, il monitoraggio e la gestione delle versioni delle API. Utilizzando OpenAPI Specification (OAS), precedentemente noto come Swagger Specification, è possibile definire il contratto API e importarlo in API Gateway. Con API Gateway, puoi eseguire la versione e la distribuzione delle API.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Fornisci contratti di servizio per API: i contratti di servizio sono accordi documentati tra i team sull'integrazione dei servizi e includono una definizione di API leggibile meccanicamente, limiti di velocità e aspettative di prestazioni.
  - Amazon API Gateway: configurazione di una REST API mediante OpenAPI
    - Una strategia di controllo delle versioni consente ai client di continuare a utilizzare l'API esistente e migrare le applicazioni all'API più recente quando sono pronte.
    - Amazon API Gateway è un servizio completamente gestito che semplifica agli sviluppatori la creazione delle API su qualsiasi scala. Utilizzando OpenAPI Specification (OAS),

precedentemente noto come Swagger Specification, puoi definire il contratto API e importarlo in API Gateway. Con API Gateway, puoi eseguire la versione e la distribuzione delle API.

Risorse

Documenti correlati:

- Amazon API Gateway: configurazione di una REST API mediante OpenAPI
- Bounded Context (un modello centrale in Domain-Driven Design)
- Implementazione di microservizi in AWS
- I compromessi dei microservizi
- Microservizi: una definizione di questo nuovo termine di architettura
- Implementazione di microservizi in AWS

REL 4 In che modo progetti le interazioni in un sistema distribuito per evitare errori?

I sistemi distribuiti si basano sulle reti di comunicazione per interconnettere i componenti (ad esempio server o servizi). Il carico di lavoro deve funzionare in modo affidabile nonostante la perdita o la latenza dei dati in queste reti. I componenti del sistema distribuito devono funzionare in modo da non influire negativamente su altri componenti o sul carico di lavoro. Queste best practice prevengono gli errori e migliorano il tempo medio tra errori (MTBF).

Best practice

- REL04-BP01 Identificazione del tipo di sistema distribuito necessario
- REL04-BP02 Implementazione di dipendenze "loosely coupled"
- REL04-BP03 Esecuzione di un lavoro costante
- REL04-BP04 Rendere tutte le risposte idempotenti

REL04-BP01 Identificazione del tipo di sistema distribuito necessario

I sistemi distribuiti hard real-time richiedono risposte che devono essere fornite in modo sincrono e rapido, mentre i sistemi soft real-time hanno una finestra temporale più generosa di minuti o più per la risposta. I sistemi offline gestiscono le risposte tramite elaborazione in batch o asincrona. I sistemi distribuiti hard real-time hanno i requisiti di affidabilità più severi.

Le difficoltà maggiori con i sistemi distribuiti riguardano i sistemi distribuiti hard real-time, noti anche come servizi di richiesta/risposta. La difficoltà sta nel fatto che le richieste arrivino in modo imprevedibile e le risposte debbano essere fornite rapidamente (ad esempio, il cliente è attivamente in attesa della risposta). Alcuni esempi includono server Web front-end, pipeline degli ordini, transazioni con carte di credito, ogni API AWS e telefonia.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Identifica il tipo di sistema distribuito necessario. Le sfide nell'ambito dei sistemi distribuiti includevano la latenza, il dimensionamento, la comprensione delle API di rete, i dati di marshalling e non-marshalling e la complessità di algoritmi come Paxos. Man mano che i sistemi diventano più grandi e più distribuiti, quelli che erano casi teorici limite diventano eventi regolari.
  - The Amazon Builders' Library: Difficoltà dei sistemi distribuiti
    - I sistemi distribuiti hard real-time richiedono risposte da fornire in modo sincrono e rapido.
    - I sistemi soft real-time hanno una finestra temporale più generosa di minuti o più per la risposta.
    - I sistemi offline gestiscono le risposte tramite elaborazione in batch o asincrona.
    - I sistemi distribuiti hard real-time hanno i requisiti di affidabilità più severi.

Risorse

Documenti correlati:

- Amazon EC2: Ensuring Idempotency (EC2: garantire l'idempotenza)
- The Amazon Builders' Library: Difficoltà dei sistemi distribuiti
- The Amazon Builders' Library: Reliability, constant work, and a good cup of coffee
- Che cos'è Amazon EventBridge?
- Che cos'è Amazon Simple Queue Service?

Video correlati:

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (AWS New York Summit 2019: Introduzione alle architetture guidate dagli eventi e ad Amazon EventBridge) (MAD205)

- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (Chiudere i cicli e aprire le menti: come prendere il controllo dei sistemi, grandi e piccoli) (sono inclusi accoppiamento debole, lavoro costante e stabilità statica) (ARC337)
- AWS re:Invent 2019: Moving to event-driven architectures (Passare alle architetture basate sugli eventi) (SVS308)

REL04-BP02 Implementazione di dipendenze "loosely coupled"

Le dipendenze come sistemi di accodamento, sistemi di streaming, flussi di lavoro e sistemi di bilanciamento del carico sono "loosely coupled" (con accoppiamento debole). L'accoppiamento debole aiuta a isolare il comportamento di un componente dagli altri componenti che dipendono da esso, aumentando la resilienza e l'agilità.

Se i cambiamenti apportati a un componente forzano la modifica anche di altri componenti basati sullo stesso, allora si parla di "tightly coupling" (accoppiamento stretto). Il "loose coupling" (accoppiamento debole) interrompe questa dipendenza, in modo che i componenti dipendenti debbano conoscere solo l'interfaccia con versione e pubblicata. L'implementazione di un accoppiamento debole tra dipendenze isola un errore all'interno di una dipendenza affinché non influenzi l'altra.

L'accoppiamento debole consente di aggiungere liberamente ulteriore codice o caratteristiche a un componente, riducendo al minimo i rischi per i componenti che dipendono da esso. Inoltre, la scalabilità è migliorata in quanto è possibile aumentare orizzontalmente o persino modificare l'implementazione sottostante della dipendenza.

Per migliorare ulteriormente la resilienza tramite accoppiamento debole, rendi le interazioni dei componenti asincrone laddove possibile. Questo modello è idoneo a qualsiasi interazione che non richieda una risposta immediata e laddove la conferma della registrazione di una richiesta sia sufficiente. Include un componente che genera eventi e un altro che li utilizza. I due componenti non si integrano tramite un'interazione diretta point-to-point, ma in genere attraverso un livello di archiviazione intermedio durevole, come una coda SQS o una piattaforma di dati in streaming come Amazon Kinesis o AWS Step Functions.

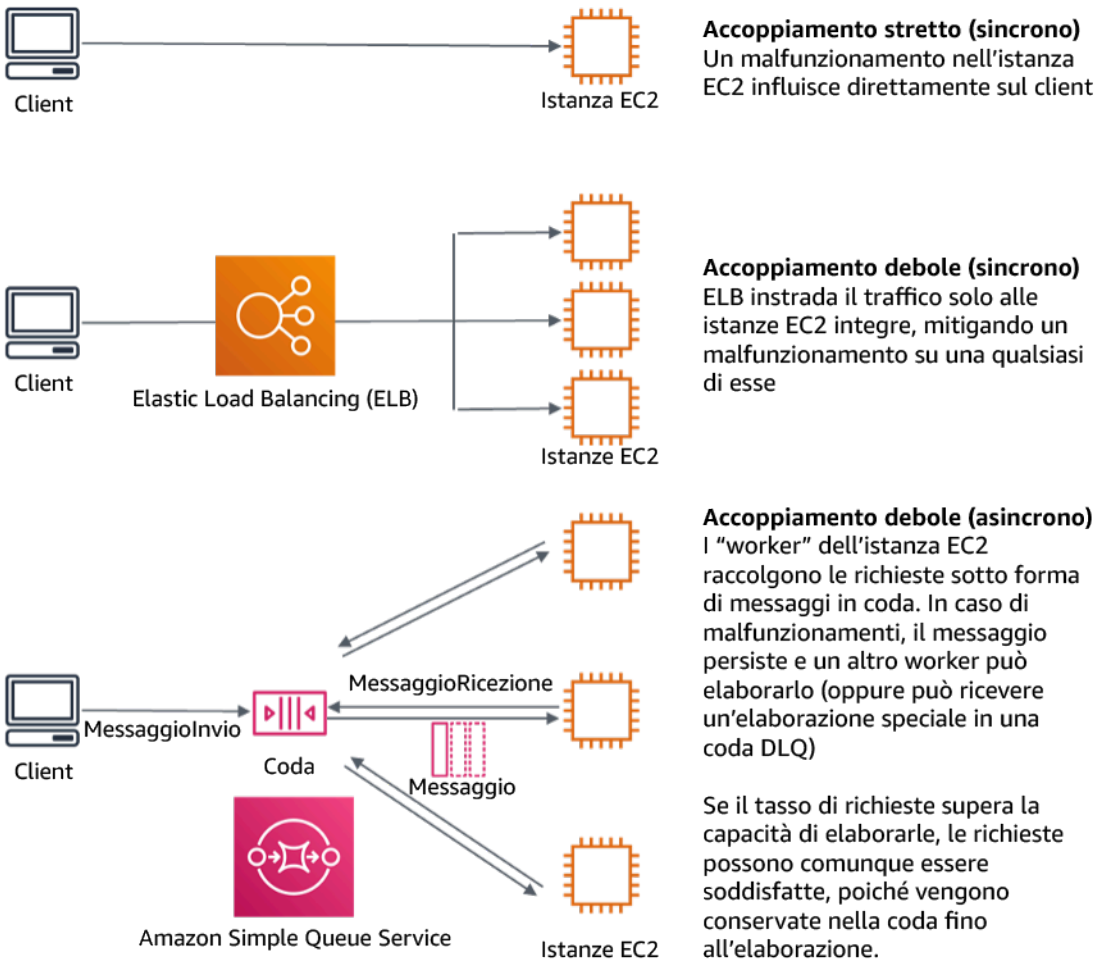


Figura 4. Le dipendenze come i sistemi di accodamento e i sistemi di bilanciamento del carico sono "loosely coupled"

Le code Amazon SQS ed Elastic Load Balancer sono solo due modi per aggiungere un livello intermedio per l'accoppiamento debole. Le architetture basate su eventi possono anche essere create in Cloud AWS utilizzando Amazon EventBridge, che può astrarre i client (produttori di eventi) dai servizi a cui fanno affidamento (consumatori di eventi). Amazon Simple Notification Service (Amazon SNS) è una soluzione efficace quando hai bisogno di messaggistica da-molti-a-molti, dalla velocità di trasmissione effettiva elevata e basata su push. Utilizzando gli argomenti di Amazon SNS, i sistemi di pubblicazione possono inviare messaggi a un numero elevato di endpoint sottoscrittori per l'elaborazione parallela.

Mentre le code offrono diversi vantaggi, nella maggior parte dei sistemi hard real-time, le richieste più vecchie di una soglia temporale (spesso secondi) dovrebbero essere considerate obsolete (il client ha abbandonato e non è più in attesa di una risposta) e non elaborate. In questo modo, è possibile elaborare invece le richieste più recenti (e probabilmente ancora valide).

Anti-pattern comuni:

- Distribuzione di un singleton come parte di un carico di lavoro.
- Invocazione diretta di API tra livelli di carico di lavoro senza funzionalità di failover o elaborazione asincrona della richiesta.

Vantaggi dell'adozione di questa best practice: L'accoppiamento debole aiuta a isolare il comportamento di un componente dagli altri componenti che dipendono da esso, aumentando la resilienza e l'agilità. L'errore in un componente è isolato dagli altri.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Implementazione di dipendenze "loosely coupled". Le dipendenze come sistemi di accodamento, sistemi di streaming, flussi di lavoro e sistemi di bilanciamento del carico sono "loosely coupled" (con accoppiamento debole). L'accoppiamento debole aiuta a isolare il comportamento di un componente dagli altri componenti che dipendono da esso, aumentando la resilienza e l'agilità.
  - AWS re:Invent 2019: Moving to event-driven architectures (Passare alle architetture basate sugli eventi) (SVS308)
  - Che cos'è Amazon EventBridge?
  - Che cos'è Amazon Simple Queue Service?
    - Amazon EventBridge consente di creare architetture basate su eventi caratterizzate da accoppiamento e distribuzione deboli.
      - AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (AWS New York Summit 2019: Introduzione alle architetture guidate dagli eventi e ad Amazon EventBridge) (MAD205)
    - Se i cambiamenti apportati a un componente forzano la modifica anche di altri componenti che si basano su esso, allora sono strettamente accoppiati. L'accoppiamento debole interrompe questa dipendenza, in modo che i componenti dipendenti debbano conoscere solo l'interfaccia con versione e pubblicata.
    - Rendere le interazioni dei componenti asincrone, laddove possibile. Questo modello è idoneo a qualsiasi interazione che non richieda una risposta immediata e laddove la conferma della registrazione di una richiesta sia sufficiente.

- AWS re:Invent 2019: Scalable serverless event-driven applications using Amazon SQS and Lambda (Applicazioni scalabili serverless basate sugli eventi con l'utilizzo di Amazon SQS e Lambda) (API304)

Risorse

Documenti correlati:

- AWS re:Invent 2019: Moving to event-driven architectures (Passare alle architetture basate sugli eventi) (SVS308)
- Amazon EC2: Ensuring Idempotency (EC2: garantire l'idempotenza)
- The Amazon Builders' Library: Difficoltà dei sistemi distribuiti
- The Amazon Builders' Library: Reliability, constant work, and a good cup of coffee
- Che cos'è Amazon EventBridge?
- Che cos'è Amazon Simple Queue Service?

Video correlati:

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (AWS New York Summit 2019: Introduzione alle architetture guidate dagli eventi e ad Amazon EventBridge) (MAD205)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (Chiudere i cicli e aprire le menti: come prendere il controllo dei sistemi, grandi e piccoli) (sono inclusi accoppiamento debole, lavoro costante e stabilità statica) (ARC337)
- AWS re:Invent 2019: Moving to event-driven architectures (Passare alle architetture basate sugli eventi) (SVS308)
- AWS re:Invent 2019: Scalable serverless event-driven applications using Amazon SQS and Lambda (Applicazioni scalabili serverless basate sugli eventi con l'utilizzo di Amazon SQS e Lambda) (API304)

REL04-BP03 Esecuzione di un lavoro costante

I sistemi possono fallire quando si verificano modifiche rapide e di grandi dimensioni nel carico. Ad esempio, se il carico di lavoro effettua un controllo dell'integrità di migliaia di server deve inviare ogni volta lo stesso payload delle dimensioni (uno snapshot completo dello stato corrente).

Indipendentemente dal fatto che non ci siano server guasti, o che lo siano tutti, il sistema di controllo dello stato esegue un lavoro costante con modifiche rapide e di piccole dimensioni.

Ad esempio, se il sistema di controllo dello stato monitora 100.000 server, il carico su di esso è nominale al di sotto del tasso di errore normalmente basso del server. Tuttavia, se un evento importante rendesse la metà di questi server non integra, il sistema di controllo dello stato sarebbe sovraccarico nel tentativo di aggiornare i sistemi di notifica e comunicare lo stato con i client. Pertanto, il sistema di controllo dello stato dovrebbe ogni volta inviare lo snapshot completo dello stato corrente. 100.000 stati di integrità del server, ciascuno rappresentato da un bit, sarebbero solo un payload di 12,5 KB. Indipendentemente dal fatto che non ci siano server guasti, o che lo siano tutti, il sistema di controllo dello stato esegue un lavoro costante e le modifiche rapide e di grandi dimensioni non rappresentano una minaccia per la stabilità del sistema. Questo è in realtà il modo in cui Amazon Route 53 gestisce i controlli dell'integrità degli endpoint (come gli indirizzi IP) per stabilire come gli utenti finali vengono instradati verso di loro.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Esegui un lavoro costante in modo che i sistemi non falliscano quando si verificano cambiamenti rapidi e significativi nel carico.
- Implementazione di dipendenze "loosely coupled". Le dipendenze come sistemi di accodamento, sistemi di streaming, flussi di lavoro e sistemi di bilanciamento del carico sono "loosely coupled" (con accoppiamento debole). L'accoppiamento debole aiuta a isolare il comportamento di un componente dagli altri componenti che dipendono da esso, aumentando la resilienza e l'agilità.
  - The Amazon Builders' Library: Reliability, constant work, and a good cup of coffee
  - AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (Chiudere i cicli e aprire le menti: come prendere il controllo dei sistemi, grandi e piccoli), include lavoro costante (ARC337)
    - Per l'esempio di un sistema di controllo dell'integrità che monitora 100.000 server, progetta i carichi di lavoro in modo che le dimensioni dei payload rimangano costanti indipendentemente dal numero di successi o di fallimenti.

Risorse

Documenti correlati:

- Amazon EC2: garantire l'idempotenza

- The Amazon Builders' Library: Difficoltà dei sistemi distribuiti
- The Amazon Builders' Library: Reliability, constant work, and a good cup of coffee

Video correlati:

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (Introduzione alle architetture guidate dagli eventi e ad Amazon EventBridge) (MAD205)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (Chiudere i cicli e aprire le menti: come prendere il controllo dei sistemi, grandi e piccoli), include lavoro costante (ARC337)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (Chiudere i cicli e aprire le menti: come prendere il controllo dei sistemi, grandi e piccoli), sono inclusi accoppiamento debole, lavoro costante e stabilità statica (ARC337)
- AWS re:Invent 2019: passare alle architetture basate sugli eventi (SVS308)

REL04-BP04 Rendere tutte le risposte idempotenti

Un servizio idempotente promette il completamento di ogni richiesta esattamente una volta, in modo tale che effettuare più richieste identiche abbia lo stesso effetto di effettuare una singola richiesta. Un servizio idempotente semplifica ad un client l'implementazione di nuovi tentativi senza temere che una richiesta venga elaborata erroneamente più volte. Per eseguire questa operazione, i client possono inviare richieste API con un token di idempotenza: viene utilizzato lo stesso token ogni volta che si ripete la richiesta. Un'API del servizio idempotente utilizza il token per restituire una risposta identica a quella restituita la prima volta che la richiesta è stata completata.

In un sistema distribuito, è facile eseguire un'operazione al massimo una volta (il client effettua una sola richiesta) o almeno una volta (la richiesta continua finché il client non ottiene la conferma dell'esito positivo). Tuttavia, è difficile garantire che un'operazione sia idempotente, il che significa che viene eseguita esattamente una volta, in modo tale che effettuare più richieste identiche abbia lo stesso effetto di effettuare una singola richiesta. Utilizzando i token di idempotenza nelle API, i servizi possono ricevere una richiesta di mutazione una o più volte senza creare record duplicati o effetti collaterali.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Rendi tutte le risposte idempotenti. Un servizio idempotente promette il completamento di ogni richiesta esattamente una volta, in modo tale che effettuare più richieste identiche abbia lo stesso effetto di effettuare una singola richiesta.
  - I client possono inviare richieste API con un token di idempotenza: viene utilizzato lo stesso token ogni volta che si ripete la richiesta. Un'API del servizio idempotente utilizza il token per restituire una risposta identica a quella restituita la prima volta che la richiesta è stata completata.
    - Amazon EC2: Ensuring Idempotency (EC2: garantire l'idempotenza)

Risorse

Documenti correlati:

- Amazon EC2: Ensuring Idempotency (EC2: garantire l'idempotenza)
- The Amazon Builders' Library: Difficoltà dei sistemi distribuiti
- The Amazon Builders' Library: Reliability, constant work, and a good cup of coffee

Video correlati:

- AWS New York Summit 2019: Intro to Event-driven Architectures and Amazon EventBridge (AWS New York Summit 2019: Introduzione alle architetture guidate dagli eventi e ad Amazon EventBridge) (MAD205)
- AWS re:Invent 2018: Close Loops and Opening Minds: How to Take Control of Systems, Big and Small (Chiudere i cicli e aprire le menti: come prendere il controllo dei sistemi, grandi e piccoli) (sono inclusi accoppiamento debole, lavoro costante e stabilità statica) (ARC337)
- AWS re:Invent 2019: Moving to event-driven architectures (Passare alle architetture basate sugli eventi) (SVS308)

REL 5 In che modo progetti le interazioni in un sistema distribuito per mitigare o affrontare gli errori?

I sistemi distribuiti si basano sulle reti di comunicazione per interconnettere i componenti (ad esempio server o servizi). Il carico di lavoro deve funzionare in modo affidabile nonostante la perdita o la latenza dei dati su queste reti. I componenti del sistema distribuito devono funzionare in modo da non influire negativamente su altri componenti o sul carico di lavoro. Queste best practice consentono ai

carichi di lavoro di affrontare stress o guasti, recuperare più rapidamente e mitigare l'impatto di tali problemi. Il risultato è un miglioramento del tempo medio di ripristino (MTTR).

Best practice

- REL05-BP01 Implementazione del degrado elegante per trasformare le dipendenze forti applicabili in dipendenze deboli
- REL05-BP02 Richieste di limitazione (della larghezza di banda della rete)
- REL05-BP03 Controllo e limitazione delle chiamate di ripetizione
- REL05-BP04 Errore rapido e limitazione delle code
- REL05-BP05 Impostazione dei timeout dei client
- REL05-BP06 Rendere i servizi stateless laddove possibile
- REL05-BP07 Implementazione di leve di emergenza

REL05-BP01 Implementazione del degrado elegante per trasformare le dipendenze forti applicabili in dipendenze deboli

Quando le dipendenze di un componente non sono integre, il componente stesso può comunque funzionare, anche se in modo degradato. Ad esempio, quando una chiamata di dipendenza non riesce, utilizza invece una risposta statica predeterminata.

Considera un servizio B chiamato dal servizio A che a sua volta chiama il servizio C.

Figura 5. Il servizio C ha esito negativo quando viene chiamato dal servizio B. Il servizio B restituisce una risposta degradata al servizio A.

Quando il servizio B chiama il servizio C, ha ricevuto da quest'ultimo un errore o un timeout. Il servizio B, senza una risposta dal servizio C (e dai dati che contiene) restituisce invece ciò che può. Questo può essere l'ultimo valore buono memorizzato nella cache oppure il servizio B può sostituire una risposta statica predeterminata a ciò che avrebbe ricevuto dal servizio C. Può quindi restituire una risposta degradata all'intermediario, il servizio A. Senza questa risposta statica, l'errore nel servizio C si propagherebbe attraverso il servizio B fino al servizio A, causando una perdita di disponibilità.

Secondo il fattore moltiplicativo nell'equazione di disponibilità per le dipendenze forti (consulta Calcolo della disponibilità con dipendenze forti), qualsiasi calo della disponibilità di C influisce notevolmente sulla disponibilità effettiva di B. Restituendo il servizio di risposta statica B mitiga l'errore in C e, sebbene degradato, rende la disponibilità del servizio C simile alla disponibilità del 100% (presupponendo che restituisca in modo affidabile la risposta statica in condizioni di errore). La risposta statica è una semplice alternativa alla restituzione di un errore e non è un tentativo di ricalcolare la risposta utilizzando metodi diversi. Tali tentativi a livello di un meccanismo completamente diverso che cercano di ottenere lo stesso risultato sono chiamati comportamento di fallback e sono un anti-modello da evitare.

Un altro esempio di degrado elegante è il modello dell'interruttore. Le strategie di ripetizione devono essere utilizzate quando l'errore è transitorio. Quando non è il caso e l'operazione potrebbe non riuscire, il modello dell'interruttore impedisce al client di eseguire una richiesta che potrebbe non riuscire. Quando le richieste vengono elaborate normalmente, l'interruttore viene chiuso e le richieste scorrono. Quando il sistema remoto inizia a restituire errori o presenta una latenza elevata, l'interruttore si apre e la dipendenza viene ignorata o i risultati vengono sostituiti con risposte ottenute più semplicemente, ma meno complete (che potrebbero essere semplicemente una cache di risposta). Periodicamente, il sistema tenta di chiamare la dipendenza per determinare se è stata ripristinata. In questo caso, l'interruttore viene chiuso.

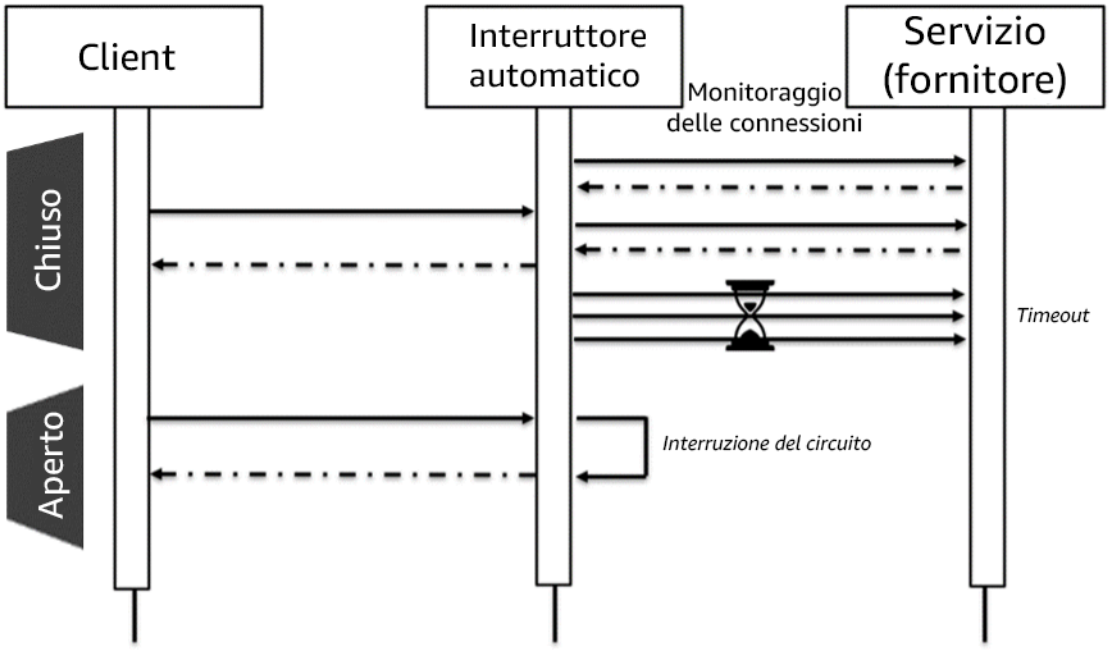


Figura 6. L'interruttore che mostra gli stati chiusi e aperti.

Oltre agli stati chiusi e aperti mostrati nel diagramma, dopo un periodo di tempo configurabile nello stato aperto, l'interruttore può passare allo stato semiaperto. In questo stato, tenta periodicamente di chiamare il servizio a una velocità molto inferiore rispetto al normale. Questa indagine viene utilizzata per controllare lo stato del servizio. Dopo un certo numero di successi nello stato semiaperto, l'interruttore passa allo stato chiuso e le normali richieste riprendono.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Implementa il degrado elegante per trasformare le dipendenze forti applicabili in dipendenze deboli. Quando le dipendenze di un componente non sono integre, il componente stesso può comunque funzionare, anche se in modo degradato. Ad esempio, quando una chiamata di dipendenza non riesce, utilizza invece una risposta statica predeterminata.
  - Restituendo una risposta statica, il carico di lavoro mitiga gli errori che si verificano nelle sue dipendenze.
    - Corso Well-Architected: Level 300: Implementing Health Checks and Managing Dependencies to Improve Reliability
  - Rileva quando è probabile che l'operazione di ripetizione non vada a buon fine e impedisci al client di effettuare chiamate non riuscite con il modello dell'interruttore.
    - CircuitBreaker

Risorse

Documenti correlati:

- Amazon API Gateway: throttling delle richieste API per migliorare le prestazioni
- CircuitBreaker (riepilogo dal libro Circuit Breaker da "Release It!")
- Ripetizione dei tentativi in caso di errore e backoff esponenziale in AWS
- Michael Nygard "Release It! Design and Deploy Production-Ready Software"
- The Amazon Builders' Library: Evitare il fallback nei sistemi distribuiti
- The Amazon Builders' Library: Evitare insormontabili backlog di code
- The Amazon Builders' Library: Sfide e strategie del caching
- The Amazon Builders' Library: Timeout, nuovi tentativi e backoff con jitter

Video correlati:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (Presentazione della libreria dei costruttori di Amazon) (DOP328)

Esempi correlati:

- Corso Well-Architected: Level 300: Implementing Health Checks and Managing Dependencies to Improve Reliability

REL05-BP02 Richieste di limitazione (della larghezza di banda della rete)

La richiesta di limitazione (della larghezza di banda della rete) è un modello di mitigazione per rispondere a un aumento imprevisto della domanda. Alcune richieste vengono soddisfatte, ma quelle che superano un limite definito vengono rifiutate e restituiscono un messaggio che indica che sono state sottoposte a throttling. L'aspettativa per i client è che si ritirino e abbandonino la richiesta o riprovino a una velocità più lenta.

I servizi devono essere progettati per gestire una capacità nota di richieste che ogni nodo o cella può elaborare. Questa capacità può essere stabilita mediante test di carico. È quindi necessario tenere traccia del tasso di arrivo delle richieste e se il tasso di arrivo temporaneo supera questo limite, la risposta appropriata è segnalare che la richiesta è stata limitata. Ciò consente all'utente di riprovare, potenzialmente su un nodo o una cella differente che potrebbe avere capacità disponibile. Amazon API Gateway fornisce metodi per la limitazione (della larghezza di banda della rete) delle richieste. Amazon SQS e Amazon Kinesis possono eseguire il buffer delle richieste, livellare il tasso di richiesta e alleggerire la necessità di limitazione (della larghezza di banda della rete) per le richieste che possono essere gestite in modo asincrono.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Richieste di limitazione (della larghezza di banda della rete). Si tratta di un modello di mitigazione per rispondere a un aumento imprevisto della domanda. Alcune richieste vengono soddisfatte, ma quelle che superano un limite definito vengono rifiutate e restituiscono un messaggio che indica che sono state sottoposte a throttling. L'aspettativa per i client è che si ritirino e abbandonino la richiesta o riprovino a una velocità più lenta.
  - Utilizzo di Amazon API Gateway
    - throttling delle richieste API per migliorare il throughput

Risorse

Documenti correlati:

- Amazon API Gateway: throttling delle richieste API per migliorare le prestazioni
- Ripetizione dei tentativi in caso di errore e backoff esponenziale in AWS
- The Amazon Builders' Library: Evitare il fallback nei sistemi distribuiti
- The Amazon Builders' Library: Evitare insormontabili backlog di code
- The Amazon Builders' Library: Timeout, nuovi tentativi e backoff con jitter
- throttling delle richieste API per migliorare il throughput

Video correlati:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (Presentazione della libreria dei costruttori di Amazon) (DOP328)

REL05-BP03 Controllo e limitazione delle chiamate di ripetizione

Utilizza il backoff esponenziale per eseguire nuovi tentativi dopo intervalli progressivamente più lunghi. Introduci il jitter per randomizzare gli intervalli di ripetizione e limitare il numero massimo di tentativi.

I componenti tipici di un sistema software distribuito includono server, sistemi di bilanciamento del carico, database e server DNS. Durante il funzionamento, e sempre soggetti ad anomalie, uno qualsiasi tra questi componenti può iniziare a generare errori. La tecnica predefinita per gestire gli errori consiste nell'implementare nuovi tentativi lato client. Questa tecnica aumenta l'affidabilità e la disponibilità dell'applicazione. Tuttavia, su vasta scala, e se i client tentano di riprovare l'operazione fallita non appena si verifica un errore, la rete può diventare rapidamente satura di richieste nuove e riproposte, ognuna delle quali compete per la larghezza di banda della rete. Ciò può causare una tempesta di ripetizione dei tentativi, che ridurrà la disponibilità del servizio. Questo modello potrebbe continuare finché non si verifica un errore completo del sistema.

Per evitare tali scenari, è necessario utilizzare gli algoritmi di backoff come il backoff esponenziale comune. Gli algoritmi di backoff esponenziale riducono gradualmente la velocità con cui vengono eseguiti i nuovi tentativi, evitando così la congestione della rete.

Molti SDK e librerie software, inclusi quelli di AWS, implementano una versione di questi algoritmi. Tuttavia, non dare mai per scontato che esista un algoritmo di backoff: esegui sempre test e verificane la presenza.

Il backoff semplice da solo non è sufficiente perché nei sistemi distribuiti tutti i client possono eseguire simultaneamente il backoff, creando cluster di chiamate ripetute. Nel suo post del blog Exponential Backoff and Jitter (Jitter e backoff esponenziale) , spiega come modificare la funzione wait() nel backoff esponenziale per evitare cluster di chiamate riproposte. La soluzione consiste nell'aggiungere jitter nella funzione wait(). Per evitare di eseguire nuovi tentativi per troppo tempo, le implementazioni dovrebbero limitare il backoff a un valore massimo.

Infine, è importante configurare un numero massimo di tentativi o di tempo trascorso, dopo il quale i nuovi tentativi semplicemente falliranno. Gli SDK AWS lo implementano per impostazione predefinita e può essere configurato. Per i servizi di livello inferiore, un limite massimo di tentativi di risposta pari a zero o a uno può limitare il rischio ed essere comunque efficace in quanto i tentativi di risposta sono delegati ai servizi di livello superiore.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Controlla e limita le chiamate riproposte. Utilizza il backoff esponenziale per eseguire nuovi tentativi dopo intervalli progressivamente più lunghi. Introduci il jitter per randomizzare gli intervalli di ripetizione e limitare il numero massimo di tentativi.
  - Ripetizione dei tentativi in caso di errore e backoff esponenziale in AWS
    - Gli SDK di Amazon implementano i nuovi tentativi e il backoff esponenziale per impostazione predefinita. Potrai implementare una logica similare nel tuo livello di dipendenze quando effettui chiamate ai tuoi servizi dipendenti. Potrai decidere quali sono i timeout e quando cessare i tentativi in base al tuo caso d'uso.

Risorse

Documenti correlati:

- Amazon API Gateway: throttling delle richieste API per migliorare le prestazioni
- Ripetizione dei tentativi in caso di errore e backoff esponenziale in AWS
- The Amazon Builders' Library: Evitare il fallback nei sistemi distribuiti
- The Amazon Builders' Library: Evitare insormontabili backlog di code

- The Amazon Builders' Library: Sfide e strategie del caching
- The Amazon Builders' Library: Timeout, nuovi tentativi e backoff con jitter

Video correlati:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (Presentazione della libreria dei costruttori di Amazon) (DOP328)

REL05-BP04 Errore rapido e limitazione delle code

Se il carico di lavoro non è in grado di rispondere correttamente a una richiesta, restituisce rapidamente un errore. Ciò consente il rilascio delle risorse associate a una richiesta e permette al servizio di recuperare se le risorse sono in esaurimento. Se il carico di lavoro è in grado di rispondere correttamente, ma la frequenza delle richieste è troppo elevata, utilizza una coda per eseguire il buffer delle richieste. Tuttavia, non consentire code lunghe che possono comportare l'elaborazione di richieste obsolete a cui il client ha già rinunciato.

Questa best practice si applica al lato server, o ricevitore, della richiesta.

Tieni presente che le code possono essere create a più livelli di un sistema e possono compromettere notevolmente la possibilità di recuperare rapidamente quando le richieste obsolete (che non necessitano più di una risposta) vengono elaborate prima di richieste più recenti. Fai attenzione ai luoghi in cui sono presenti code. Spesso si nascondono nei flussi di lavoro o nel lavoro registrato in un database.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Errore rapido e limitazione delle code. Se il carico di lavoro non è in grado di rispondere correttamente a una richiesta, restituisce rapidamente un errore. Ciò consente il rilascio delle risorse associate a una richiesta e permette al servizio di recuperare se le risorse sono in esaurimento. Se il carico di lavoro è in grado di rispondere correttamente, ma la frequenza delle richieste è troppo elevata, utilizza una coda per eseguire il buffer delle richieste. Tuttavia, non consentire code lunghe che possono comportare l'elaborazione di richieste obsolete a cui il client ha già rinunciato.
  - Implementazione d'errore rapido quando il servizio è eccessivamente sollecitato.it
    - Errore rapido

- Limita le code in un sistema basato su code, quando l'elaborazione si interrompe ma i messaggi continuano ad arrivare, il debito di messaggi può accumularsi in un backlog di grandi dimensioni, determinando un aumento del tempo di elaborazione. Il lavoro potrebbe essere completato troppo tardi perché i risultati siano utili, provocando essenzialmente il danneggiamento della disponibilità che l'accodamento doveva evitare.
  - The Amazon Builders' Library: Evitare insormontabili backlog di code

Risorse

Documenti correlati:

- Ripetizione dei tentativi in caso di errore e backoff esponenziale in AWS
- Errore rapido
- The Amazon Builders' Library: Evitare il fallback nei sistemi distribuiti
- The Amazon Builders' Library: Evitare insormontabili backlog di code
- The Amazon Builders' Library: Sfide e strategie del caching
- The Amazon Builders' Library: Timeout, nuovi tentativi e backoff con jitter

Video correlati:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (Presentazione della libreria dei costruttori di Amazon) (DOP328)

REL05-BP05 Impostazione dei timeout dei client

Imposta i timeout in modo appropriato, verificali sistematicamente e non fare affidamento sui valori predefiniti poiché sono generalmente troppo alti.

Questa best practice si applica al lato client, o al mittente, della richiesta.

Imposta sia un timeout di connessione che un timeout di richiesta su qualsiasi chiamata remota e, generalmente, su qualsiasi chiamata tra i processi. Molti framework offrono funzionalità di timeout integrate, ma fai attenzione perché molti hanno valori predefiniti infiniti o troppo alti. Un valore troppo elevato riduce l'utilità del timeout perché le risorse continuano a essere consumate mentre il client attende che si verifichi il timeout. Un valore troppo basso può generare un aumento del traffico sul back-end e una maggiore latenza perché vengono ritentate troppe richieste. In alcuni casi, questo può portare a interruzioni complete perché tutte le richieste vengono ritentate.

Per ulteriori informazioni su come Amazon utilizza timeout, nuovi tentativi e backoff con jitter, consulta la Builders' Library: timeout, nuovi tentativi e backoff con jitter.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Imposta sia un timeout di connessione che un timeout di richiesta su qualsiasi chiamata remota e, generalmente, su qualsiasi chiamata tra i processi. Molti framework offrono funzionalità di timeout integrate, ma fai attenzione perché molti hanno valori predefiniti infiniti o troppo alti. Un valore troppo elevato riduce l'utilità del timeout perché le risorse continuano a essere consumate mentre il client attende che si verifichi il timeout. Un valore troppo basso può generare un aumento del traffico sul back-end e una maggiore latenza perché vengono ritentate troppe richieste. In alcuni casi, questo può portare a interruzioni complete perché tutte le richieste vengono ritentate.
  - AWS SDK: Retries and Timeouts (SDK AWS: nuovi tentativi e timeout)

Risorse

Documenti correlati:

- AWS SDK: Retries and Timeouts (SDK AWS: nuovi tentativi e timeout)
- Amazon API Gateway: throttling delle richieste API per migliorare le prestazioni
- Ripetizione dei tentativi in caso di errore e backoff esponenziale in AWS
- The Amazon Builders' Library: Timeout, nuovi tentativi e backoff con jitter

Video correlati:

- Retry, backoff, and jitter: AWS re:Invent 2019: Introducing The Amazon Builders' Library (Presentazione della libreria dei costruttori di Amazon) (DOP328)

REL05-BP06 Rendere i servizi stateless laddove possibile

I servizi non devono richiedere lo stato oppure devono eseguire l'offload dello stato in modo tale che, tra diverse richieste client, non vi sia alcuna dipendenza dai dati archiviati localmente su disco o in memoria. In questo modo i server possono essere sostituiti a piacimento senza compromettere la disponibilità. Amazon ElastiCache o Amazon DynamoDB sono ottime destinazioni per lo stato di offload.

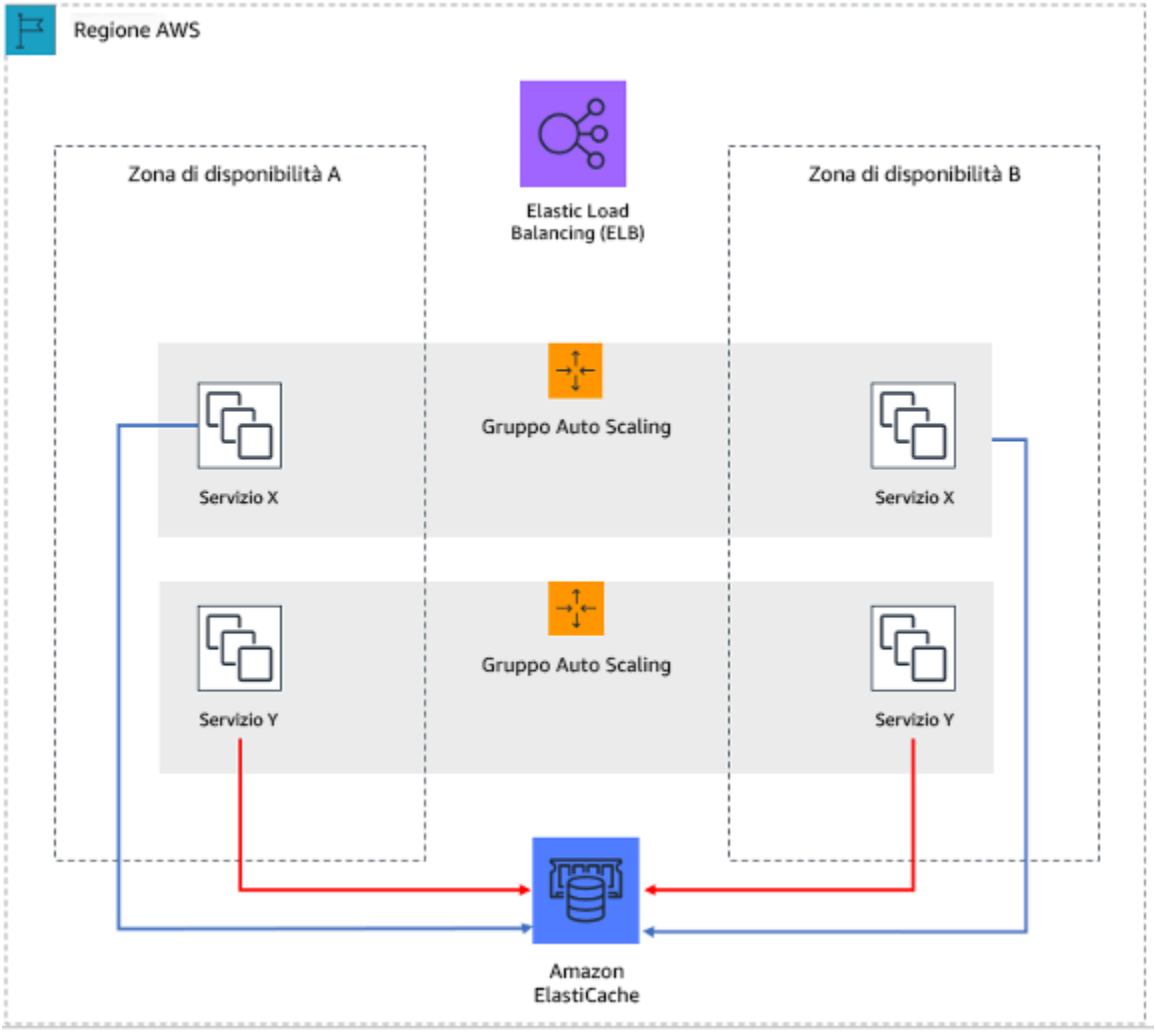


Figura 7. In questa applicazione Web stateless, viene eseguito l'offload dello stato della sessione in Amazon ElastiCache.

Quando gli utenti o i servizi interagiscono con un'applicazione, spesso eseguono una serie di interazioni che formano una sessione. Una sessione è un dato univoco per gli utenti che persistono tra le richieste mentre utilizzano l'applicazione. Un'applicazione stateless è un'applicazione che non richiede la conoscenza delle interazioni precedenti e non memorizza le informazioni sulla sessione.

Una volta progettata per essere stateless, puoi utilizzare servizi di elaborazione serverless, come AWS Lambda o AWS Fargate.

Oltre alla sostituzione del server, un altro vantaggio delle applicazioni stateless è che possono ricalibrare orizzontalmente perché qualsiasi risorsa di calcolo disponibile (ad esempio istanze EC2 e funzioni AWS Lambda) può soddisfare ogni richiesta.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Trasforma le applicazioni in stateless. Applicazioni stateless consentono un dimensionamento orizzontale e sono tolleranti al guasto di un singolo nodo.
  - Eliminazione dello stato che potrebbe effettivamente essere memorizzato nei parametri di richiesta.
  - Dopo aver esaminato se lo stato è necessario, sposta qualsiasi tracciamento dello stato in una cache multizona resiliente o in un archivio di dati come Amazon ElastiCache, Amazon RDS, Amazon DynamoDB o una soluzione di dati distribuiti di terze parti. Memorizza uno stato impossibile da spostare in datastore resilienti.
    - Alcuni dati (come i cookie) possono passare nei titoli o nei parametri di query.
    - Effettua il refactoring per rimuovere uno stato che può essere passato velocemente nelle richieste.
    - È possibile che alcuni dati non siano effettivamente necessari per richiesta e possano essere recuperati on demand.
    - Rimuovi i dati recuperabili in modo asincrono.
    - Scegli un datastore che soddisfi i requisiti per uno stato necessario.
    - Valuta l'utilizzo di un database NoSQL per dati non relazionali.

Risorse

Documenti correlati:

- The Amazon Builders' Library: Evitare il fallback nei sistemi distribuiti
- The Amazon Builders' Library: Evitare insormontabili backlog di code
- The Amazon Builders' Library: Sfide e strategie del caching

REL05-BP07 Implementazione di leve di emergenza

Le leve di emergenza sono processi rapidi che possono mitigare l'impatto sulla disponibilità sul carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Implementa leve di emergenza. Si tratta di processi rapidi che possono mitigare l'impatto della disponibilità sul carico di lavoro. Possono essere utilizzati in assenza di una causa principale. Una leva di emergenza ideale riduce a zero il carico cognitivo dei resolver fornendo criteri di attivazione e disattivazione completamente deterministici. Le leve sono spesso manuali, ma possono anche essere automatizzate
  - Esempi di leve includono
    - Bloccare tutto il traffico dei robot
    - Servire pagine statiche anziché dinamiche
    - Ridurre la frequenza delle chiamate a una dipendenza
    - Limitare le chiamate dalle dipendenze
  - Suggerimenti per l'implementazione e l'utilizzo di leve di emergenza
    - Quando le leve sono attivate, fai di meno, non di più
    - Rendi le cose semplici, evita comportamenti bimodali
    - Testare periodicamente le leve
  - Di seguito sono elencati alcuni esempi di operazioni che NON rappresentano leve di emergenza
    - Aggiunta di capacità
    - Chiamare i proprietari dei servizi dei client che dipendono dal tuo servizio e chiedere loro di ridurre le chiamate
    - Apportare una modifica al codice e rilasciarlo

Gestione delle modifiche

Domande

- REL 6 In che modo monitori le risorse del carico di lavoro?
- REL 7 In che modo progetti il carico di lavoro per adattarti ai cambiamenti della domanda?
- REL 8 In che modo implementi le modifiche?

REL 6 In che modo monitori le risorse del carico di lavoro?

I log e i parametri sono strumenti molto efficaci per ottenere informazioni sullo stato del tuo carico di lavoro. È possibile configurare il carico di lavoro in modo da monitorare i log e i parametri e

inviare notifiche quando vengono superate le soglie o si verificano eventi significativi. Il monitoraggio consente al carico di lavoro di riconoscere quando vengono superate le soglie di prestazioni basse o si verificano errori, in modo che possa essere ripristinato automaticamente di rimando.

Best practice

- REL06-BP01 Monitoraggio di tutti i componenti per il carico di lavoro (generazione)
- REL06-BP02 Definizione e calcolo dei parametri (aggregazione)
- REL06-BP03 Invio di notifiche (elaborazione e avvisi in tempo reale)
- REL06-BP04 Automatizzazione delle risposte (elaborazione e avvisi in tempo reale)
- REL06-BP05 Analisi
- REL06-BP06 Esecuzione di revisioni periodiche
- REL06-BP07 Monitoraggio del tracciamento end-to-end delle richieste attraverso il sistema

REL06-BP01 Monitoraggio di tutti i componenti per il carico di lavoro (generazione)

monitora i componenti del carico di lavoro con Amazon CloudWatch o con strumenti di terze parti. Monitora i servizi AWS con il pannello di controllo AWS Health.

Occorre monitorare tutti i componenti del carico di lavoro, inclusi front-end, logica aziendale e livelli di storage. Definisci i parametri chiave e come estrarli dai registri, se necessario, e imposta soglie per l'attivazione degli eventi di allarme corrispondenti. Assicurati che i parametri siano pertinenti agli indicatori chiave di prestazione (KPI) del tuo carico di lavoro e utilizza i parametri e i registri per identificare i primi segnali di degrado del servizio. Ad esempio, un parametro legato ai risultati aziendali, come il numero di ordini elaborati con successo al minuto, può indicare problemi di carico di lavoro più rapidamente di un parametro tecnico, come l'utilizzo della CPU. Utilizza il pannello di controllo AWS Health per una visualizzazione personalizzata delle prestazioni e della disponibilità dei servizi AWS sottostanti alle risorse AWS.

Il monitoraggio nel cloud offre nuove opportunità. La maggior parte dei provider cloud ha sviluppato hook personalizzabili e può fornire approfondimenti per aiutarti a monitorare più livelli del carico di lavoro. I servizi AWS come Amazon CloudWatch applicano algoritmi statistici e di apprendimento automatico per analizzare continuamente i parametri di sistemi e applicazioni, determinare le normali linee di base e far emergere le anomalie con un intervento minimo da parte dell'utente. Gli algoritmi di rilevamento delle anomalie tengono conto della stagionalità e delle variazioni di tendenza dei parametri.

AWS mette a disposizione una grande quantità di informazioni di monitoraggio e di registro che possono essere utilizzate per definire parametri specifici per i carichi di lavoro, processi di variazione della domanda e per l'adozione di tecniche di apprendimento automatico indipendentemente dalle competenze di ML.

Inoltre, monitora tutti gli endpoint esterni per avere la certezza che siano indipendenti dall'implementazione di base. Questo monitoraggio attivo può essere effettuato con transazioni sintetiche (talvolta indicate come canary utente,ma da non confondere con le implementazioni canary) che eseguono periodicamente una serie di attività comuni che corrispondono alle azioni eseguite dai client del carico di lavoro. Mantieni queste attività di breve durata e assicurati di non sovraccaricare il carico di lavoro durante il test. Amazon CloudWatch Synthetics ti consente di creare canary sintetici per monitorare gli endpoint e le API. Puoi anche combinare i nodi client sintetici Canary con la console AWS X-Ray per individuare quali Canary sintetiche stanno riscontrando problemi con errori, guasti o velocità di throttling per l'intervallo di tempo selezionato.

Risultato desiderato:

raccogliere e utilizzare i parametri critici di tutti i componenti del carico di lavoro per garantire l'affidabilità del carico di lavoro e un'esperienza utente ottimale. Rilevare che un carico di lavoro non sta raggiungendo i risultati aziendali consente di dichiarare rapidamente un disastro e di riprendersi da un incidente.

Anti-pattern comuni:

- Solo monitoraggio delle interfacce esterne per il carico di lavoro.
- Non generare parametri specifici per il carico di lavoro e affidati solo ai parametri forniti dai servizi AWS utilizzati dal carico di lavoro.
- Utilizzare solo parametri tecnici nel carico di lavoro e non monitorare i parametri relativi agli indicatori chiave di prestazione (KPI) non tecnici a cui il carico di lavoro contribuisce.
- Affidarsi al traffico di produzione e a semplici controlli di integrità per monitorare e valutare lo stato del carico di lavoro.

Vantaggi dell'adozione di questa best practice: il monitoraggio a tutti i livelli del carico di lavoro consente di prevedere e risolvere più rapidamente i problemi dei componenti che costituiscono il carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

1. Abilitazione della registrazione ove disponibile. I dati di monitoraggio devono essere ottenuti da tutti i componenti dei carichi di lavoro. Attiva ulteriori registri, come i registri di accesso S3, e abilita il carico di lavoro per registrare i dati specifici del carico di lavoro. Raccogli i parametri per le medie di CPU, I/O di rete e I/O su disco da servizi come Amazon ECS, Amazon EKS, Amazon EC2, Elastic Load Balancing, AWS Auto Scaling ed Amazon EMR. Consulta Servizi AWS che pubblicano parametri CloudWatch Servizi AWS che pubblicano parametri su CloudWatch.
2. Esamina tutti i parametri predefiniti ed esplora eventuali lacune nella raccolta dei dati. Tutti i servizi generano parametri predefiniti. La raccolta di parametri predefiniti consente di comprendere meglio le dipendenze tra i componenti del carico di lavoro e il modo in cui l'affidabilità e le prestazioni dei componenti influiscono sul carico di lavoro. Puoi anche creare e pubblicare parametri propri affinché CloudWatch utilizzi la AWS CLI o un'API. Questo
3. valuta tutti i parametri per decidere quelli a cui inviare avvisi per ogni servizio AWS nel carico di lavoro. Puoi scegliere di selezionare un sottoinsieme di parametri che hanno un impatto importante sull'affidabilità del carico di lavoro. La focalizzazione su soglie e parametri critici consente di affinare il numero di avvisi informativi e può contribuire a ridurre al minimo i falsi positivi.
4. Definisci gli avvisi e il processo di recupero del carico di lavoro dopo l'attivazione dell'avviso. La definizione degli avvisi consente di notificare, intensificare e seguire rapidamente le fasi necessarie per il ripristino da un incidente e il rispetto dell'obiettivo di tempo di ripristino (RTO) prescritto. Puoi utilizzare avvisi Amazon CloudWatch per invocare flussi di lavoro automatici e avviare procedure di ripristino in base a soglie definite.
5. Esplora l'uso di transazioni sintetiche per raccogliere dati rilevanti sullo stato dei carichi di lavoro. Il monitoraggio sintetico segue gli stessi percorsi ed esegue le stesse azioni di un cliente, il che consente di verificare continuamente l'esperienza del cliente anche quando non c'è traffico di clienti sui carichi di lavoro. Utilizzando le transazioni sintetiche,puoi individuare i problemi prima dei clienti.

Risorse

Best practice correlate:

- REL11-BP03 Automatizzazione della riparazione a tutti i livelli

Documenti correlati:

- Getting started with your AWS Health Dashboard – Your account health (Nozioni di base su AWS HealthDashboard: stato del tuo account)
- Servizi AWS che pubblicano parametri CloudWatch
- Log di accesso per Network Load Balancer
- Log di accesso per Application Load Balancer
- Accesso a Amazon CloudWatch Logs per AWS Lambda
- Registrazione delle richieste con registrazione dell'accesso al server Amazon S3
- Abilita i log di accesso per Classic Load Balancer
- Esportazione di dati di registro in Amazon S3
- Installazione dell'agente CloudWatch su un'istanza Amazon EC2
- Pubblicazione di parametri personalizzati
- Utilizzo dei pannelli di controllo Amazon CloudWatch
- Utilizzare i parametri Amazon CloudWatch
- Utilizzo di Canary (Amazon CloudWatch Synthetics)
- Cosa sono i Amazon CloudWatch Logs?

Guide per l'utente:

- Creazione di un trail
- Monitoraggio dei parametri di memoria e del disco per le istanze Amazon EC2 Linux
- Utilizzo di CloudWatch Logs con istanze di container
- Log di flusso VPC
- Che cos'è Amazon DevOps Guru?
- Che cos'è AWS X-Ray?

Blog correlati:

- Effettuare il debug con Amazon CloudWatch Synthetics e AWS X-Ray

Esempi e workshop correlati:

- AWS Well-Architected Labs: Operational Excellence - Dependency Monitoring (Laboratori ben strutturati AWS: Eccellenza operativa - Monitoraggio delle dipendenze)
- The Amazon Builders' Library: Dotazione dei sistemi distribuiti per la visibilità operativa

- Workshop sull'osservabilità

REL06-BP02 Definizione e calcolo dei parametri (aggregazione)

Archivia i dati di registro e applica i filtri, laddove necessari, per calcolare i parametri, ad esempio i conteggi di un evento di registro specifico o la latenza calcolata dai timestamp del registro eventi.

Amazon CloudWatch e Amazon S3 fungono da principali livelli di aggregazione e storage. Per alcuni servizi, come AWS Auto Scaling e Elastic Load Balancing, i parametri predefiniti vengono forniti per impostazione predefinita per il carico della CPU o la latenza media delle richieste in un cluster o in un'istanza. Per i servizi di streaming, come i registri di flusso VPC e AWS CloudTrail, i dati degli eventi vengono inoltrati a CloudWatch Logs ed è necessario definire e applicare filtri di parametri per estrarre i parametri dai dati dell'evento. In questo modo vengono forniti dati di serie temporali, che possono fungere da input per gli allarmi CloudWatch definiti dall'utente per attivare gli avvisi.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Aggregazione: definisci e calcola i parametri. Archivia i dati di log e applica filtri, se necessario, per calcolare i parametri, ad esempio i conteggi di un evento di log specifico o la latenza calcolata dai timestamp degli eventi di log
  - I filtri dei parametri definiscono i termini e i modelli da ricercare nei dati di registro inviati a CloudWatch Logs. CloudWatch Logs utilizza questi filtri di parametri per trasformare i dati di registro in parametri CloudWatch numerici che è possibile rappresentare su un grafico o un avviso.
    - Ricerca e filtraggio dei dati di log
  - Utilizza una terza parte affidabile per aggregare i registri.
    - Segui le istruzioni che ti vengono fornite dalle terze parti. La maggior parte dei prodotti di terze parti si integra con CloudWatch e Amazon S3.
  - Alcuni servizi AWS possono pubblicare registri direttamente in Amazon S3. Se il requisito principale per i registri è l'archiviazione in Amazon S3, si può facilmente fare in modo che il servizio che produce i registri li invii direttamente a Amazon S3, senza dover creare un'infrastruttura aggiuntiva.
    - Invio di registri direttamente a Amazon S3

Risorse

Documenti correlati:

- Query di esempio di Amazon CloudWatch Logs Insights
- Effettuare il debug con Amazon CloudWatch Synthetics e AWS X-Ray
- One Observability Workshop
- Ricerca e filtraggio dei dati di log
- Invio di registri direttamente a Amazon S3
- The Amazon Builders' Library: Dotazione dei sistemi distribuiti per la visibilità operativa

REL06-BP03 Invio di notifiche (elaborazione e avvisi in tempo reale)

Le organizzazioni interessate ricevono le notifiche quando si verificano eventi significativi.

Gli avvisi possono essere inviati ad argomenti Amazon Simple Notification Service (Amazon SNS) e poi inoltrati a un numero qualsiasi di iscritti. Ad esempio, Amazon SNS può inoltrare avvisi a un alias e-mail in modo che il personale tecnico possa rispondere.

Anti-pattern comuni:

- La configurazione di avvisi a una soglia troppo bassa causa l'invio di troppe notifiche.
- Non archiviare avvisi per l'esplorazione futura.

Vantaggi dell'adozione di questa best practice: le notifiche sugli eventi (anche quelle che è possibile gestire e risolvere in automatico) consentono di avere un record di eventi e di affrontarli potenzialmente in modo diverso in futuro.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Elaborazione e avvisi in tempo reale. Le organizzazioni che devono essere messe al corrente ricevono le notifiche nel caso si verifichino eventi significativi
  - I pannelli di controllo di Amazon CloudWatch sono home page personalizzabili nella console CloudWatch che puoi utilizzare per monitorare le tue risorse in un'unica visualizzazione, anche quelle distribuite tra regioni diverse.
    - Utilizzo dei pannelli di controllo Amazon CloudWatch

- Crea un avviso quando un parametro supera un limite.
  - Utilizzo degli allarmi di Amazon CloudWatch

Risorse

Documenti correlati:

- One Observability Workshop
- The Amazon Builders' Library: Dotazione dei sistemi distribuiti per la visibilità operativa
- Utilizzo degli allarmi di Amazon CloudWatch
- Utilizzo dei pannelli di controllo Amazon CloudWatch
- Utilizzare i parametri Amazon CloudWatch

REL06-BP04 Automatizzazione delle risposte (elaborazione e avvisi in tempo reale)

utilizza l'automazione per agire quando viene rilevato un evento; ad esempio, per sostituire i componenti guasti.

Gli avvisi possono attivare eventi di AWS Auto Scaling, in modo che i cluster reagiscano ai cambiamenti della domanda. Gli avvisi possono essere inviati a Amazon Simple Queue Service (Amazon SQS), che può fungere da punto di integrazione per sistemi di ticket di terze parti. AWS Lambda può anche effettuare l'iscrizione ad avvisi, fornendo agli utenti un modello serverless asincrono che reagisce alle modifiche in modo dinamico. AWS Config monitora e registra continuamente le configurazioni delle risorse AWS e può attivare AWS Systems Manager Automation per risolvere i problemi.

Amazon DevOps Guru monitora automaticamente le risorse dell'applicazione per rilevare comportamenti anomali e fornisce raccomandazioni mirate per accelerare i tempi di identificazione e riparazione dei problemi.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Utilizza Amazon DevOps Guru per eseguire azioni automatizzate. Amazon DevOps Guru monitora automaticamente le risorse dell'applicazione per rilevare comportamenti anomali e fornisce raccomandazioni mirate per accelerare i tempi di identificazione e riparazione dei problemi.
  - What is Amazon DevOps Guru? (Che cos'è Amazon DevOps Guru?)

- Utilizza AWS Systems Manager per eseguire azioni automatizzate. AWS Config monitora e registra in modo continuo le configurazioni delle risorse AWS e può attivare AWS Systems Manager per risolvere i problemi.
  - AWS Systems Manager Automation
    - Crea e utilizza documenti Systems Manager Automation. Questi definiscono le operazioni che Systems Manager esegue sulle istanze gestite e su altre risorse AWS quando si avvia un processo di automazione.
    - Gestione dei documenti di automazione (playbook)
- Amazon CloudWatch invia eventi di modifica dello stato di avviso a Amazon EventBridge. Crea regole di EventBridge per automatizzare le risposte.
  - Creazione di una regola EventBridge che si attivi su un evento da una risorsa AWS
- Crea ed esegui un piano per automatizzare le risposte.
  - Effettua l'inventario di tutte le procedure di risposta agli avvisi. Devi pianificare le risposte agli avvisi prima di classificare le attività.
  - Effettua l'inventario di tutte le attività con azioni specifiche da intraprendere. La maggior parte di queste azioni è documentata nei runbook. È inoltre necessario disporre di playbook per gli avvisi relativi a eventi imprevisti.
  - Esamina i runbook e i playbook per tutte le azioni automatizzabili. In generale, se è possibile definire un'azione, è molto probabile che si possa anche automatizzare.
  - Classifica innanzitutto le attività soggette a errori o dispendiose in termini di tempo. È molto utile eliminare le fonti di errore e ridurre i tempi di risoluzione.
  - Definisci un piano per completare l'automazione. Mantieni un piano attivo per automatizzare e aggiornare l'automazione.
  - Esamina i requisiti manuali per le opportunità di automazione. Metti alla prova il processo manuale per scoprire opportunità di automazione.

Risorse

Documenti correlati:

- AWS Systems Manager Automation
- Creazione di una regola EventBridge che si attivi su un evento da una risorsa AWS
- One Observability Workshop
- The Amazon Builders' Library: Dotazione dei sistemi distribuiti per la visibilità operativa

- What is Amazon DevOps Guru? (Che cos'è Amazon DevOps Guru?)
- Gestione dei documenti di automazione (playbook)

REL06-BP05 Analisi

raccogli i file di log e le cronologie dei parametri e analizzali per ottenere informazioni più ampie sulle tendenze e sui carichi di lavoro.

Amazon CloudWatch Logs Insights supporta un linguaggio di query semplice ma potente che puoi utilizzare per analizzare i dati di log. Amazon CloudWatch Logs supporta anche le sottoscrizioni che consentono ai dati di fluire in modo ottimale verso Amazon S3, dove puoi utilizzare o Amazon Athena per eseguire query sui dati. Supporta, inoltre, le query su un'ampia gamma di formati. Consulta SerDe e formati di dati supportati nella Guida per l'utente Amazon Athena per ulteriori informazioni. Per l'analisi di enormi set di file di log, puoi eseguire un cluster Amazon EMR per effettuare analisi con capacità nell'ordine dei petabyte.

Esistono numerosi strumenti forniti da Partner AWS e terze parti che consentono aggregazione, elaborazione, archiviazione e analisi. Questi strumenti includono New Relic, Splunk, Loggly, Logstash, CloudHealth e Nagios. Tuttavia, la generazione esterna di log di sistema e applicazioni è univoca per ciascun provider di servizi cloud e spesso per ciascun servizio.

Una parte spesso trascurata del processo di monitoraggio è la gestione dei dati. È necessario determinare i requisiti di conservazione per il monitoraggio dei dati, quindi applicare le policy del ciclo di vita di conseguenza. Amazon S3 supporta la gestione del ciclo di vita a livello di bucket S3. Questa gestione del ciclo di vita può essere applicata in modo diverso ai diversi percorsi nel bucket. Verso la fine del ciclo di vita è possibile trasferire i dati su Amazon S3 Glacier per l'archiviazione a lungo termine fino alla scadenza, al termine del periodo di conservazione. La classe di storage S3 Intelligent-Tiering è progettata per ottimizzare i costi trasferendo automaticamente i dati nel livello di accesso più conveniente, senza impatto sulle prestazioni o sovraccarico operativo.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Gli approfondimenti CloudWatch Logs consentono di cercare e analizzare in modo interattivo i dati di registro in Amazon CloudWatch Logs.
  - Analisi dei dati di registro con gli approfondimenti CloudWatch Logs
  - Query di esempio di Amazon CloudWatch Logs Insights

- Utilizza Amazon CloudWatch Logs per inviare registri a Amazon S3 dove puoi utilizzare Amazon Athena per le query dei dati.
  - Come faccio ad analizzare i miei registri di accesso al server Amazon S3 utilizzando Athena?
    - Crea una policy del ciclo di vita di S3 per il bucket dei log di accesso al server. Configura la policy del ciclo di vita per rimuovere periodicamente i file di log. In questo modo si riduce la quantità di dati che Athena deve analizzare per ogni query.
      - Come faccio a creare una policy del ciclo di vita per un bucket S3?

Risorse

Documenti correlati:

- Query di esempio di Amazon CloudWatch Logs Insights
- Analisi dei dati di registro con gli approfondimenti CloudWatch Logs
- Effettuare il debug con Amazon CloudWatch Synthetics e AWS X-Ray
- Come faccio a creare una policy del ciclo di vita per un bucket S3?
- Come faccio ad analizzare i miei registri di accesso al server Amazon S3 utilizzando Athena?
- One Observability Workshop
- The Amazon Builders' Library: Dotazione dei sistemi distribuiti per la visibilità operativa

REL06-BP06 Esecuzione di revisioni periodiche

Esegui verifiche frequenti delle modalità di implementazione del monitoraggio del carico di lavoro e aggiornalo in base a eventi e modifiche significativi.

Il monitoraggio efficace è basato su parametri aziendali chiave. Assicurati che questi parametri siano presenti nel carico di lavoro man mano che le priorità aziendali cambiano.

L'audit del monitoraggio consente di sapere quando un'applicazione sta raggiungendo gli obiettivi di disponibilità. L'analisi delle cause principali richiede la capacità di scoprire cosa è successo in caso di errori. AWS consente di monitorare lo stato dei tuoi servizi durante un incidente:

- Amazon CloudWatch Logs: è possibile archiviare i log in questo servizio e controllarne i contenuti.
- Amazon CloudWatch Logs Insights: è un servizio completamente gestito che consente di eseguire analisi di registri di grandi dimensioni in pochi secondi. Offre query e visualizzazioni rapide e interattive.

- AWS Config: è possibile vedere quale infrastruttura AWS era in uso in momenti differenti.
- AWS CloudTrail: è possibile vedere quali API AWS sono state richiamate, a che ora e da quale principale.

In AWS, conduciamo meeting settimanali per esaminare le prestazioni operative e condividere quanto appreso tra i team. Dato l'elevato numero di team presenti in AWS, abbiamo creato La ruota per scegliere casualmente un carico di lavoro da esaminare. Stabilire una cadenza regolare per le revisioni delle prestazioni operative e la condivisione delle conoscenze migliora la capacità di ottenere prestazioni più elevate dai team operativi.

Anti-pattern comuni:

- Raccolta dei soli parametri predefiniti.
- Impostazione di una strategia di monitoraggio senza alcuna revisione.
- Nessuna discussione sul monitoraggio quando vengono distribuite modifiche importanti.

Vantaggi dell'adozione di questa best practice: la verifica periodica del monitoraggio consente di prevedere potenziali problemi, invece di rispondere alle notifiche quando un problema previsto si verifica effettivamente.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Crea più pannelli di controllo per il carico di lavoro. È necessario disporre di un pannello di controllo di primo livello contenente i parametri aziendali chiave, nonché i parametri tecnici che hai identificato come i più rilevanti per lo stato previsto del carico di lavoro al variare dell'utilizzo. È inoltre importante disporre di pannelli di controllo per vari livelli di applicazione e dipendenze che è possibile ispezionare.
  - Utilizzo dei pannelli di controllo Amazon CloudWatch
- Pianifica ed effettua revisioni periodiche dei pannelli di controllo del carico di lavoro. Effettua un'ispezione regolare dei pannelli di controllo. La frequenza può essere diversa a seconda di quanto l'ispezione sia approfondita.
  - Ispeziona l'andamento nei parametri. Confronta i valori dei parametri con i valori storici per vedere se ci sono tendenze che potrebbero suggerire l'esame di un particolare aspetto. Riportiamo alcuni esempi: aumento della latenza, riduzione della funzione aziendale primaria e aumento delle risposte all'errore.

- Identificazione di outlier/anomalie nei parametri. Le medie o mediane possono nascondere outlier e anomalie. Osserva i valori più alti e più bassi nell'intervallo di tempo e analizza le cause dei risultati estremi. Man mano che continui a eliminare tali cause, la riduzione del numero di valori estremi ti consente di continuare a migliorare la coerenza delle prestazioni del carico di lavoro.
- Ricerca di bruschi cambiamenti nel comportamento. Un cambiamento repentino della quantità o della direzione di un parametro può indicare un cambiamento nell'applicazione o fattori esterni che potrebbero richiedere l'aggiunta di ulteriori parametri da monitorare.

Risorse

Documenti correlati:

- Query di esempio di Amazon CloudWatch Logs Insights
- Effettuare il debug con Amazon CloudWatch Synthetics e AWS X-Ray
- One Observability Workshop
- The Amazon Builders' Library: Dotazione dei sistemi distribuiti per la visibilità operativa
- Utilizzo dei pannelli di controllo Amazon CloudWatch

REL06-BP07 Monitoraggio del tracciamento end-to-end delle richieste attraverso il sistema

Utilizza AWS X-Ray o strumenti di terze parti per consentire agli sviluppatori di eseguire più facilmente l'analisi e il debug di sistemi distribuiti, per comprendere l'andamento delle prestazioni delle loro applicazioni e dei relativi servizi sottostanti.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Monitora il tracciamento end-to-end delle richieste attraverso il sistema. AWS X-Ray è un servizio che raccoglie dati sulle richieste elaborate dalla tua applicazione e fornisce strumenti che puoi utilizzare per visualizzare, filtrare e ottenere informazioni approfondite su tali dati per identificare problemi e opportunità di ottimizzazione. Per qualsiasi richiesta tracciata alla tua applicazione, puoi visualizzare informazioni dettagliate non solo sulla richiesta e sulla risposta, ma anche sulle chiamate effettuate dall'applicazione verso microservizi, database, API Web e risorse AWS a valle.
  - Che cos'è AWS X-Ray?
  - Effettuare il debug con Amazon CloudWatch Synthetics e AWS X-Ray

Risorse

Documenti correlati:

- Effettuare il debug con Amazon CloudWatch Synthetics e AWS X-Ray
- One Observability Workshop
- The Amazon Builders' Library: Dotazione dei sistemi distribuiti per la visibilità operativa
- Utilizzo di Canary (Amazon CloudWatch Synthetics)
- Che cos'è AWS X-Ray?

REL 7 In che modo progetti il carico di lavoro per adattarti ai cambiamenti della domanda?

Un carico di lavoro scalabile fornisce elasticità per aggiungere o rimuovere risorse automaticamente, in modo che vi sia una stretta corrispondenza con la domanda attuale in un dato momento.

Best practice

- REL07-BP01 Utilizzo dell'automazione per l'acquisizione o il dimensionamento delle risorse
- REL07-BP02 Ottenimento di risorse quando viene rilevata la compromissione di un carico di lavoro
- REL07-BP03 Ottenimento di risorse dopo aver rilevato che sono necessarie più risorse per un carico di lavoro
- REL07-BP04 Esecuzione di un test di carico sul carico di lavoro

REL07-BP01 Utilizzo dell'automazione per l'acquisizione o il dimensionamento delle risorse

Quando sostituisci risorse danneggiate o esegui il dimensionamento del carico di lavoro, puoi automatizzare il processo utilizzando servizi AWS gestiti, come Amazon S3 e AWS Auto Scaling. Puoi anche utilizzare strumenti di terze parti e SDK AWS per automatizzare il dimensionamento.

I servizi gestiti AWS includono Amazon S3, Amazon CloudFront, AWS Auto Scaling, AWS Lambda, Amazon DynamoDB, AWS Fargate e Amazon Route 53.

AWS Auto Scaling consente di rilevare e sostituire le istanze danneggiate. Inoltre, permette di creare piani di dimensionamento per le risorse, tra cui istanze e parchi istanze Amazon EC2 , attività Amazon ECS , tabelle e indici Amazon DynamoDB e repliche di Amazon Aurora .

Durante il dimensionamento di istanze EC2, assicurati di utilizzare più zone di disponibilità (preferibilmente almeno tre) e di aggiungere o rimuovere capacità per mantenere il bilanciamento tra queste zone. Anche le attività ECS o i pod Kubernetes (quando si utilizza Amazon Elastic Kubernetes Service) devono essere distribuiti su più zone di disponibilità.

Quando utilizzi AWS Lambda, le istanze subiscono un dimensionamento automatico. Ogni volta che viene ricevuta una notifica di evento per la funzione, AWS Lambda individua rapidamente la capacità libera all'interno del parco istanze di calcolo ed esegue il codice fino alla simultaneità allocata. Devi assicurarti che la simultaneità necessaria sia configurata sulla Lambda specifica e nelle tue Service Quotas.

Amazon S3 ricalibra automaticamente le risorse per gestire elevati tassi di richiesta. Ad esempio, l'applicazione può ottenere almeno 3.500 richieste PUT/COPY/POST/DELETE o 5.500 richieste GET /HEAD al secondo per prefisso in un bucket. Non ci sono limiti al numero di prefissi in un bucket. Puoi aumentare le prestazioni di lettura o scrittura parallelizzando le letture. Ad esempio, se crei 10 prefissi in un bucket Amazon S3 per parallelizzare le letture, potresti dimensionare le prestazioni di lettura a 55.000 richieste al secondo.

Configura e utilizza Amazon CloudFront o una rete di distribuzione di contenuti (CDN) attendibile. Una CDN può fornire tempi di risposta più rapidi agli utenti finali e può servire le richieste di contenuti dalla cache, riducendo così la necessità di dimensionare il carico di lavoro.

Anti-pattern comuni:

- Implementare gruppi Auto Scaling per la correzione automatica, ma senza elasticità.
- Utilizzare l'auto scaling per rispondere a grandi aumenti di traffico.
- Distribuire applicazioni altamente stateful, eliminando l'opzione di elasticità.

Vantaggi dell'adozione di questa best practice: L'automazione elimina il potenziale di errori manuali nella distribuzione e nella disattivazione delle risorse. L'automazione elimina il rischio di superamento dei costi e di rifiuto del servizio a causa della risposta lenta alle esigenze di distribuzione o disattivazione.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Configura e utilizza AWS Auto Scaling. In questo modo è possibile monitorare le applicazioni e regolare automaticamente la capacità per mantenere prestazioni stabili e prevedibili al minor costo

possibile. Grazie ad AWS Auto Scaling, puoi configurare il dimensionamento delle applicazioni per più risorse in vari servizi.

- Che cos'è AWS Auto Scaling?
  - Configura il dimensionamento automatico su serie di istanze Spot e istanze Amazon EC2, attività Amazon ECS, indici e tabelle Amazon DynamoDB, repliche Amazon Aurora e applicazioni Marketplace AWS, come applicabile.
    - Gestione automatica della capacità di throughput con DynamoDB Auto Scaling
      - Utilizza le operazioni delle API di servizi per specificare gli avvisi, le policy di ridimensionamento e i tempi di riscaldamento e raffreddamento.

- Utilizza Elastic Load Balancing. I sistemi di bilanciamento del carico possono distribuire il carico in base al percorso o alla connettività di rete.
  - Che cos'è Elastic Load Balancing?
    - Application Load Balancers può distribuire il carico per percorso.
      - What is an Application Load Balancer? (Che cos'è un Application Load Balancer?)
        - Configura un Application Load Balancer per distribuire il traffico su diversi carichi di lavoro in base a un percorso nello stesso nome di dominio.
        - Gli Application Load Balancers possono essere utilizzati per distribuire i carichi in modo da gestire la domanda attraverso l'integrazione con AWS Auto Scaling.
          - Uso di un sistema di bilanciamento del carico con un gruppo Auto Scaling
    - I Network Load Balancer possono distribuire il carico in base alla connessione.
      - Che cos'è un Network Load Balancer?
        - Configura un Network Load Balancer per distribuire il traffico su diversi carichi di lavoro tramite TCP o per disporre di un set costante di indirizzi IP per il carico di lavoro.
        - I Network Load Balancer possono essere utilizzati per distribuire i carichi in modo da gestire la domanda attraverso l'integrazione con AWS Auto Scaling.

- Uso di un provider DNS altamente disponibile I nomi DNS consentono agli utenti di accedere ai carichi di lavoro utilizzando nomi anziché indirizzi IP e distribuire queste informazioni in un ambito definito, solitamente a livello globale per gli utenti del carico di lavoro.
  - Utilizza Amazon Route 53 o un provider DNS affidabile.
    - Che cos'è Amazon Route 53?
  - Utilizza Route 53 per gestire le distribuzioni CloudFront e i load balancer.
    - Individua i domini e i sottodomini da gestire.

    - Crea set di record appropriati utilizzando record ALIAS o CNAME.
        - Uso dei record
- Utilizza la rete globale AWS per ottimizzare il percorso dagli utenti alle applicazioni. AWS Global Accelerator monitora costantemente l'integrità degli endpoint delle applicazioni e reindirizza il traffico verso endpoint integri in meno di 30 secondi.
    - AWS Global Accelerator è un servizio che migliora la disponibilità e le prestazioni delle applicazioni con utenti locali o globali, fornendo indirizzi IP statici che fungono da punto di ingresso fisso agli endpoint delle applicazioni in una o più regioni Regioni AWS, ad esempio Application Load Balancers, Network Load Balancer o istanze Amazon EC2.
        - Che cos'è AWS Global Accelerator?
- Configura e utilizza Amazon CloudFront o una rete di distribuzione di contenuti (CDN) attendibile. Una rete di distribuzione di contenuti (CDN) può fornire tempi di risposta più rapidi agli utenti finali e soddisfare richieste di contenuti che possono causare un dimensionamento non necessario dei carichi di lavoro.
    - Che cos'è Amazon CloudFront?
        - Configura le distribuzioni di Amazon CloudFront per i carichi di lavoro oppure utilizza una CDN di terze parti.
            - Puoi limitare l'accesso ai tuoi carichi di lavoro in modo che siano accessibili solo da CloudFront utilizzando gli intervalli di indirizzi IP per CloudFront nelle policy di accesso o nei gruppi di sicurezza degli endpoint.

Risorse

Documenti correlati:

- Partner APN: partner per la creazione di soluzioni di elaborazione automatizzate
- AWS Auto Scaling: come funzionano i piani di dimensionamento
- Marketplace AWS: prodotti che possono essere utilizzati con Auto Scaling
- Gestione automatica della capacità di throughput con DynamoDB Auto Scaling
- Uso di un sistema di bilanciamento del carico con un gruppo Auto Scaling
- Che cos'è AWS Global Accelerator?
- Che cos'è Amazon EC2 Auto Scaling?
- Che cos'è AWS Auto Scaling?

- Che cos'è Amazon CloudFront?
- Che cos'è Amazon Route 53?
- Che cos'è Elastic Load Balancing?
- Che cos'è un Network Load Balancer?
- What is an Application Load Balancer? (Che cos'è un Application Load Balancer?)
- Uso dei record

REL07-BP02 Ottenimento di risorse quando viene rilevata la compromissione di un carico di lavoro

All'occorrenza, ridimensiona le risorse in modo reattivo se la disponibilità è influenzata per ripristinare la disponibilità del carico di lavoro.

Devi prima configurare i controlli dello stato e i criteri su questi controlli per indicare quando la disponibilità è influenzata dalla mancanza di risorse. Quindi notificare al personale appropriato di dimensionare manualmente la risorsa o attivare l'automazione per dimensionarla automaticamente.

Il dimensionamento può essere regolato manualmente in base al carico di lavoro, ad esempio modificando il numero di istanze EC2 in un gruppo con scalabilità automatica o modificando la velocità di trasmissione effettiva di una tabella DynamoDB tramite la AWS Management Console o la AWS CLI. Tuttavia, l'automazione deve essere utilizzata ogni qualvolta sia possibile (consulta Utilizzo dell'automazione per l'acquisizione o il dimensionamento delle risorse).

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Ottieni le risorse quando viene rilevata la compromissione di un carico di lavoro All'occorrenza, ridimensiona le risorse in modo reattivo se la disponibilità è influenzata per ripristinare la disponibilità del carico di lavoro.
  - Utilizza i piani di dimensionamento, che sono il componente principale di AWS Auto Scaling, per configurare una serie di istruzioni per dimensionare le risorse. Se lavori con AWS CloudFormation o aggiungi tag alle risorse AWS, puoi impostare piani di dimensionamento per diversi set di risorse, per ogni applicazione. AWS Auto Scaling fornisce raccomandazioni per strategie di dimensionamento personalizzate per ogni risorsa. Dopo aver creato il piano, AWS Auto Scaling combina i metodi di dimensionamento dinamico e predittivo per supportare la tua strategia di dimensionamento.
    - AWS Auto Scaling: come funzionano i piani di dimensionamento

- Amazon EC2 Auto Scaling aiuta a garantire che sia disponibile il numero corretto di istanze Amazon EC2 per gestire il carico dell'applicazione. È possibile creare raccolte di istanze EC2, denominate gruppi Auto Scaling. Puoi specificare il numero minimo di istanze in ciascun gruppo con scalabilità automatica, mentre Amazon EC2 Auto Scaling garantisce che il gruppo non scenda mai al di sotto di tale quantità. Puoi specificare il numero massimo di istanze in ciascun gruppo con scalabilità automatica, mentre Amazon EC2 Auto Scaling garantisce che il gruppo non scenda mai al di sotto di tale quantità.
  - Che cos'è Amazon EC2 Auto Scaling?
- Il dimensionamento automatico Amazon DynamoDB utilizza il servizio di dimensionamento automatico dell'applicazione AWS per regolare dinamicamente la capacità effettiva di trasmissione assegnata per tuo conto, in risposta ai modelli di traffico effettivi. Ciò consente a una tabella o a un indice secondario globale di aumentare la capacità di lettura e scrittura assegnata per gestire aumenti di traffico improvvisi, senza throttling.
  - Gestione automatica della capacità di throughput con DynamoDB Auto Scaling

Risorse

Documenti correlati:

- Partner APN: partner per la creazione di soluzioni di elaborazione automatizzate
- AWS Auto Scaling: come funzionano i piani di dimensionamento
- Marketplace AWS: prodotti che possono essere utilizzati con Auto Scaling
- Gestione automatica della capacità di throughput con DynamoDB Auto Scaling
- Che cos'è Amazon EC2 Auto Scaling?

REL07-BP03 Ottenimento di risorse dopo aver rilevato che sono necessarie più risorse per un carico di lavoro

Dimensiona le risorse in modo proattivo per soddisfare la domanda ed evitare l'impatto sulla disponibilità.

Molti servizi AWS dimensionano automaticamente le risorse per soddisfare la domanda. Se si utilizzano istanze Amazon EC2 o cluster Amazon ECS, puoi configurare la scalabilità automatica di tali istanze in base ai parametri di utilizzo corrispondenti alla richiesta del carico di lavoro. Per Amazon EC2, è possibile impiegare l'utilizzo medio della CPU, il conteggio delle richieste del sistema di bilanciamento del carico o la larghezza di banda di rete per aumentare (o ridurre) le istanze EC2.

Per Amazon ECS, è possibile impiegare l'utilizzo medio della CPU, il conteggio delle richieste del load balancer e l'utilizzo della memoria per aumentare orizzontalmente (o ridurre orizzontalmente) le attività ECS. Utilizzando il dimensionamento automatico di destinazione su AWS, l'autoscaler si comporta come un termostato domestico, aggiungendo o rimuovendo risorse per mantenere il valore di destinazione (ad esempio, il 70% di utilizzo della CPU) specificato.

AWS Auto Scaling può anche eseguire l' Auto Scaling predittivo, che utilizza il machine learning per analizzare il carico di lavoro cronologico di ciascuna risorsa e prevede regolarmente il carico futuro per i due giorni successivi.

La legge di Little aiuta a calcolare il numero di istanze di calcolo (istanze EC2, funzioni Lambda simultanee, ecc.) necessarie.

$L = \lambda W$

L = numero di istanze (o simultaneità media nel sistema)

λ = velocità media alla quale arrivano le richieste (richieste/sec)

W = tempo medio trascorso da ogni richiesta nel sistema (sec)

Ad esempio, a 100 rps, se ogni richiesta impiega 0,5 secondi per l'elaborazione, avrai bisogno di 50 istanze per tenere il passo con la domanda.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Ottieni risorse dopo aver rilevato che sono necessarie più risorse per un carico di lavoro Dimensiona le risorse in modo proattivo per soddisfare la domanda ed evitare l'impatto sulla disponibilità.
  - Valuta quante risorse di calcolo sono necessarie (simultaneità di calcolo) per gestire un determinato tasso di richiesta
    - Telling Stories About Little's Law
  - Quando disponi di un modello cronologico per l'utilizzo, imposta il dimensionamento programmato per il dimensionamento automatico Amazon EC2.
    - Dimensionamento programmato per Amazon EC2 Auto Scaling
  - Utilizza il dimensionamento predittivo di AWS.
    - Dimensionamento predittivo per EC2, alimentato dal machine learning

Risorse

Documenti correlati:

- AWS Auto Scaling: come funzionano i piani di dimensionamento
- Marketplace AWS: prodotti che possono essere utilizzati con Auto Scaling
- Gestione automatica della capacità di throughput con DynamoDB Auto Scaling
- Dimensionamento predittivo per EC2, alimentato dal machine learning
- Dimensionamento programmato per Amazon EC2 Auto Scaling
- Telling Stories About Little's Law
- Che cos'è Amazon EC2 Auto Scaling?

REL07-BP04 Esecuzione di un test di carico sul carico di lavoro

Adotta un metodo di test del carico per misurare se l'attività di dimensionamento soddisfa i requisiti del carico di lavoro.

È importante eseguire test di carico prolungati. I test di carico devono rilevare il punto di rottura e testare le prestazioni del carico di lavoro. AWS consente di creare facilmente ambienti di test temporanei che riproducono la scala del carico di lavoro di produzione. Nel cloud, puoi creare un ambiente di test su scala produttiva on demand, completare i test e disattivare le risorse. Poiché paghi per l'ambiente di test solo quando è in esecuzione, puoi simulare un ambiente live a un costo notevolmente inferiore rispetti ai test in locale.

I test di carico in produzione dovrebbero anche essere considerati come parte dei game day in cui il sistema di produzione viene messo alla prova, durante le ore di utilizzo inferiore del cliente, con tutto il personale a disposizione per interpretare i risultati e risolvere eventuali problemi che si presentano.

Anti-pattern comuni:

- Eseguire test di carico su distribuzioni che non presentano la stessa configurazione della tua produzione.
- Eseguire test di carico solo su singole parti del carico di lavoro e non sulla sua interezza.
- Eseguire test di carico con un sottoinsieme di richieste e non con un set rappresentativo delle richieste effettive.
- Eseguire test di carico su un fattore di sicurezza di poco superiore al carico previsto.

Vantaggi dell'adozione di questa best practice: Saprai quali sono i componenti dell'architettura che non funzionano sotto carico e potrai identificare per tempo i parametri che indicano l'avvicinamento al carico in questione, così da affrontare il problema e prevenire l'impatto dell'esito negativo.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Esegui test di carico per identificare quali aspetti del carico di lavoro indicano la necessità di aggiungere o rimuovere capacità. Il test di carico deve avere un traffico rappresentativo simile a quello che ricevi nella produzione. Aumenta il carico mentre osservi i parametri implementati per stabilire quale di questi indica quando è necessario aggiungere o rimuovere risorse.
  - Distributed Load Testing on AWS (Test di carico distribuito su AWS): simula migliaia di utenti connessi
    - Identifica la combinazione di richieste. Potresti avere diverse combinazioni di richieste, quindi dovresti esaminare vari intervalli di tempo per identificare la combinazione di traffico.
    - Implementa un driver di caricamento. Puoi utilizzare codice personalizzato, software open source o software commerciale per implementare un driver di carico.
    - Esegui un test di carico iniziale con una capacità ridotta. Puoi vedere alcuni effetti immediati applicando il carico su una capacità inferiore, possibilmente pari a un'istanza o a un container.
    - Esegui un test di carico con una capacità maggiore. Gli effetti saranno diversi su un carico distribuito, quindi è necessario eseguire il test in condizioni quanto più simili possibili all'ambiente del prodotto.

Risorse

Documenti correlati:

- Distributed Load Testing on AWS (Test di carico distribuito su AWS): simula migliaia di utenti connessi

REL 8 In che modo implementi le modifiche?

Per distribuire nuove funzionalità e garantire che i carichi di lavoro e l'ambiente operativo eseguano software noti e che sia possibile applicare patch o sostituirli in modo prevedibile, sono necessarie modifiche controllate. Se invece non sono controllate, risulta difficile prevederne l'effetto o risolvere eventuali problemi che causano.

Best practice

- REL08-BP01 Utilizzo di runbook per attività standard come l'implementazione
- REL08-BP02 Esecuzione di test funzionali come parte integrante dell'implementazione
- REL08-BP03 Esecuzione di test di resilienza come parte integrante dell'implementazione
- REL08-BP04 Esecuzione dell'implementazione utilizzando un'infrastruttura immutabile
- REL08-BP05 Implementazione delle modifiche tramite automazione

REL08-BP01 Utilizzo di runbook per attività standard come l'implementazione

I runbook sono le procedure predefinite per ottenere risultati specifici. Utilizza i runbook per eseguire attività standard, o manualmente o automaticamente. Alcuni esempi includono l'implementazione di un carico di lavoro, l'applicazione di patch a un carico di lavoro o la realizzazione di modifiche DNS.

Ad esempio, metti in atto processi per garantire la sicurezza del rollback durante le distribuzioni. Garantire la possibilità di eseguire il rollback di una distribuzione senza interruzioni per i clienti è fondamentale per rendere un servizio affidabile.

Per le procedure di runbook, inizia da un processo manuale valido ed efficace, implementalo nel codice e attivalo per l'esecuzione automatica, se necessario.

Anche per carichi di lavoro sofisticati e altamente automatizzati, i runbook rimangono utili per eseguire game day o soddisfare rigorosi requisiti di reportistica e audit.

Tieni presente che i playbook vengono utilizzati in risposta a incidenti specifici e i runbook vengono utilizzati per ottenere risultati specifici. Spesso, i runbook sono per attività di routine, mentre i playbook vengono utilizzati per rispondere a eventi non di routine.

Anti-pattern comuni:

- Eseguire modifiche impreviste alla configurazione nella produzione.
- Ignorare le fasi del piano per velocizzare l'implementazione, compromettendone la riuscita.
- Apportare modifiche senza testarne l'annullamento.

Vantaggi dell'adozione di questa best practice: Una pianificazione efficace aumenta la capacità di eseguire correttamente la modifica, perché sei a conoscenza di tutti i sistemi interessati. Convalidare la modifica negli ambienti di test aumenta la sicurezza.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Abilita risposte coerenti e tempestive agli eventi noti documentando le procedure nei runbook.
  - Framework AWS Well-Architected – Concetti – Runbook
- Uso del principio di infrastruttura come codice per definire l'infrastruttura Utilizzando AWS CloudFormation o una terza parte affidabile per definire la tua infrastruttura, puoi utilizzare un software per il controllo delle versioni per gestire le versioni e tenere traccia delle modifiche.
  - Utilizza AWS CloudFormation o un provider di terze parti affidabile per definire l'infrastruttura.
    - Che cos'è AWS CloudFormation?
  - Crea modelli unici e disaccoppiati, utilizzando solidi principi di progettazione del software.
    - Stabilisci le autorizzazioni, i modelli e le parti responsabili dell'implementazione
      - Controllo degli accessi con AWS Identity and Access Management
    - Utilizza un controllo sorgente come AWS CodeCommit o uno strumento di terze parti affidabili per il controllo delle versioni.
      - Che cos'è AWS CodeCommit?

Risorse

Documenti correlati:

- Partner APN: partner per la creazione di soluzioni di distribuzione automatizzate
- Marketplace AWS: prodotti per l'automazione delle distribuzioni
- Framework AWS Well-Architected – Concetti – Runbook
- Che cos'è AWS CloudFormation?
- Che cos'è AWS CodeCommit?

Esempi correlati:

- Automating operations with Playbooks and Runbooks (Automazione delle operazioni con Playbook e Runbook)

REL08-BP02 Esecuzione di test funzionali come parte integrante dell'implementazione

I test funzionali vengono eseguiti come parte integrante della distribuzione automatizzata. Se non vengono soddisfatti i criteri di esito positivo, la pipeline viene arrestata o ripresa dall'inizio.

Questi test vengono eseguiti in un ambiente di pre-produzione, gestito per fasi prima della produzione nella pipeline. Idealmente, questa operazione viene eseguita come parte di una pipeline di distribuzione.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Esegui test funzionali come parte integrante dell'implementazione. I test funzionali vengono eseguiti come parte integrante della distribuzione automatizzata. Se non vengono soddisfatti i criteri di esito positivo, la pipeline viene arrestata o ripresa dall'inizio.
  - Richiama AWS CodeBuild durante l'azione di test delle pipeline di rilascio di software modellate in AWS CodePipeline. Questa funzionalità consente di eseguire facilmente un'ampia gamma di test sul codice, tra cui test delle unità, analisi del codice statico e test di integrazione.
    - AWS CodePipeline Adds Support for Unit and Custom Integration Testing with AWS CodeBuild (AWS CodePipeline aggiunge il supporto per i test di unità e integrazione personalizzati con AWS CodeBuild)
  - Utilizza le soluzioni Marketplace AWS per eseguire test automatizzati come parte integrante della tua pipeline di distribuzione di software.
    - Automazione e test del software

Risorse

Documenti correlati:

- AWS CodePipeline Adds Support for Unit and Custom Integration Testing with AWS CodeBuild (AWS CodePipeline aggiunge il supporto per i test di unità e integrazione personalizzati con AWS CodeBuild)
- Automazione e test del software
- Che cos'è AWS CodePipeline?

REL08-BP03 Esecuzione di test di resilienza come parte integrante dell'implementazione

I test di resilienza (eseguiti utilizzando i Principles of Chaos Engineering) vengono svolti nell'ambito della pipeline di implementazione automatizzata in un ambiente di pre-produzione.

Questi test vengono gestiti per fasi ed eseguiti nella pipeline di pre-produzione. Devono anche essere eseguiti in produzione, ma come parte di game day.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Esegui test di resilienza come parte integrante della distribuzione Utilizza l'ingegneria del caos, la disciplina che consiste nello sperimentare su un carico di lavoro per aumentare la fiducia nella capacità del carico di lavoro di resistere a condizioni turbolente in produzione.
  - I test di resilienza inseriscono errori o causano un degrado delle risorse per valutare se il carico di lavoro risponde con la resilienza progettata
    - Corso Well-Architected: Level 300: Testing for Resiliency of EC2 RDS and S3
  - Questi test possono essere eseguiti regolarmente in ambienti di pre-produzione nelle pipeline di distribuzione automatizzate.
  - È opportuno eseguirli anche in produzione, nell'ambito delle giornate di gioco pianificate.
  - A partire dai principi di ingegneristica del caos, avanza ipotesi sulle prestazioni del carico di lavoro in caso di vari problemi, quindi mettile alla prova utilizzando i test di resilienza.
    - Principles of Chaos Engineering

Risorse

Documenti correlati:

- Principles of Chaos Engineering
- Che cos'è AWS Fault Injection Simulator?

Esempi correlati:

- Corso Well-Architected: Level 300: Testing for Resiliency of EC2 RDS and S3

REL08-BP04 Esecuzione dell'implementazione utilizzando un'infrastruttura immutabile

L'infrastruttura immutabile è un modello che richiede che non vengano applicati aggiornamenti, patch di sicurezza o modifiche di configurazione sui carichi di lavoro di produzione. Quando è necessaria una modifica, l'architettura viene costruita su una nuova infrastruttura e distribuita alla produzione.

L'implementazione più comune del paradigma dell'infrastruttura immutabile è il server immutabile. Ciò significa che se un server necessita di un aggiornamento o di una correzione, vengono distribuiti nuovi server invece di aggiornare quelli già in uso. Pertanto, invece di accedere al server tramite SSH

e aggiornare la versione del software, ogni modifica nell'applicazione inizia con un push del software al repository di codice, ad esempio git push. Poiché non sono consentite modifiche nell'infrastruttura immutabile, puoi essere sicuro dello stato del sistema distribuito. Le infrastrutture immutabili sono intrinsecamente più coerenti, affidabili e prevedibili e semplificano molti aspetti dello sviluppo e delle operazioni di software.

Utilizza una distribuzione Canary o blue/green durante la distribuzione di applicazioni in infrastrutture immutabili.

Distribuzione Canary : è la pratica di indirizzare un piccolo numero di clienti alla nuova versione, in genere in esecuzione su una singola istanza di servizio (la release Canary). Quindi analizzerai in modo approfondito le modifiche di comportamento o gli errori generati. Puoi rimuovere il traffico dalla release Canary in caso di problemi critici e reindirizzare gli utenti alla versione precedente. Se la distribuzione viene completata correttamente, puoi continuare a distribuire alla velocità desiderata, monitorando le modifiche alla ricerca di errori, fino a quando non sarai completamente distribuito. AWS CodeDeploy può essere configurato con una configurazione di distribuzione che abilita una distribuzione Canary.

Distribuzione blue/green : è simile alla distribuzione Canary, tranne per il fatto che un intero parco dell'applicazione è distribuito in parallelo. Puoi alternare le distribuzioni tra i due stack (blue e green). Ancora una volta, puoi inviare il traffico alla nuova versione e tornare alla versione precedente in caso di problemi con la distribuzione. Generalmente, tutto il traffico viene trasferito contemporaneamente, tuttavia puoi anche utilizzare frazioni del traffico verso ciascuna versione per accelerare l'adozione della nuova versione mediante le funzionalità di instradamento DNS ponderato di Amazon Route 53. AWS CodeDeploy e AWS Elastic Beanstalk possono essere impostati con una configurazione di implementazione che abilita un'implementazione blu/verde.

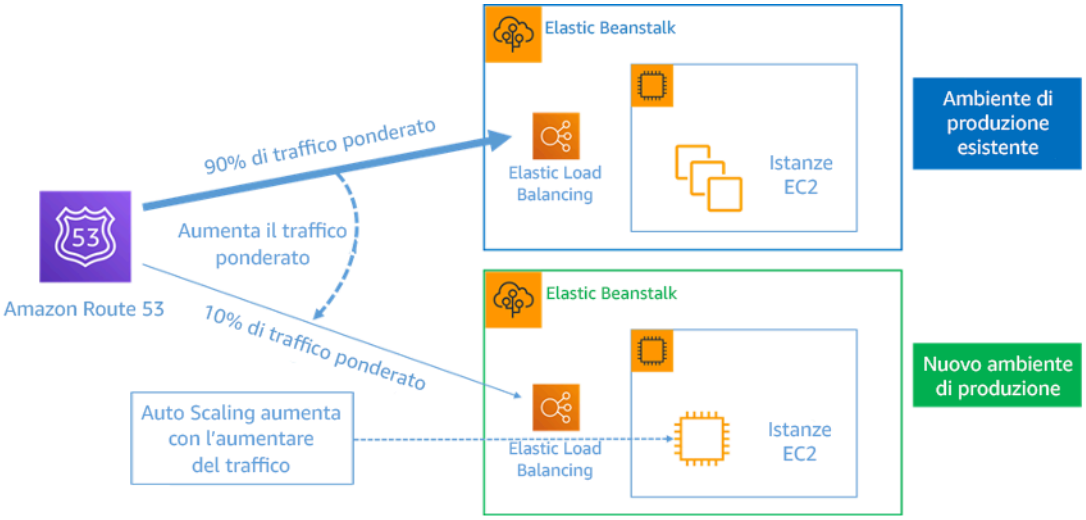


Figura 8: Implementazione blu/verde con AWS Elastic Beanstalk e Amazon Route 53

Vantaggi dell'infrastruttura immutabile:

- Riduzione delle deviazioni di configurazione: sostituendo frequentemente i server da una configurazione di base, nota e controllata dalla versione, l'infrastruttura viene reimpostata a uno stato noto, evitando deviazioni di configurazione.
- Distribuzioni semplificate: le distribuzioni sono semplificate perché non devono supportare gli aggiornamenti. Gli aggiornamenti sono solo nuove distribuzioni.
- Distribuzioni atomiche affidabili: le distribuzioni vengono completate correttamente o non cambia nulla. Offre maggiore fiducia nel processo di distribuzione.
- Distribuzioni più sicure con processi di rollback e ripristino rapidi: Le distribuzioni sono più sicure perché la versione funzionante precedente non viene modificata. Puoi eseguire il rollback se vengono rilevati errori.
- Ambienti di test e debug ottimizzati: poiché tutti i server utilizzano la stessa immagine, non ci sono differenze tra gli ambienti. Una build viene distribuita in più ambienti. Inoltre, evita ambienti incoerenti e semplifica test e debug.
- Maggiore scalabilità: poiché i server utilizzano un'immagine di base, sono coerenti e ripetibili, la scalabilità automatica è intrinseca.
- Toolchain semplificata: la toolchain è semplificata poiché è possibile eliminare gli strumenti di gestione della configurazione che gestiscono gli aggiornamenti del software di produzione. Non vengono installati altri strumenti o agenti sui server. Le modifiche vengono apportate all'immagine di base, testate e implementate.
- Maggiore sicurezza: negando tutte le modifiche ai server, puoi disabilitare SSH sulle istanze e rimuovere le chiavi. Questo riduce il vettore di attacco, migliorando l'assetto di sicurezza dell'organizzazione.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Distribuisci utilizzando un'infrastruttura immutabile. Un'infrastruttura immutabile è un modello che impone che non vengano applicati aggiornamenti, patch di sicurezza o modifiche sui carichi di lavoro di produzione. Quando è necessaria una modifica, viene creata una nuova versione dell'architettura e distribuita alla produzione.
  - Panoramica di una distribuzione Blue/Green

- Distribuzione graduale di applicazioni serverless
- Infrastruttura immutabile: affidabilità, coerenza e fiducia attraverso l'immutabilità
- Release Canary

Risorse

Documenti correlati:

- Release Canary
- Distribuzione graduale di applicazioni serverless
- Infrastruttura immutabile: affidabilità, coerenza e fiducia attraverso l'immutabilità
- Panoramica di una distribuzione Blue/Green
- The Amazon Builders' Library: Garantire la sicurezza del rollback durante le distribuzioni

REL08-BP05 Implementazione delle modifiche tramite automazione

Le distribuzioni e l'applicazione di patch sono automatizzate per eliminare l'impatto negativo.

Apportare modifiche ai sistemi produttivi è una delle maggiori aree di rischio per molte organizzazioni. Riteniamo che le distribuzioni siano un problema prioritario da risolvere insieme ai problemi aziendali affrontati dal software. Oggi, ciò significa l'uso dell'automazione ovunque sia pratica nelle operazioni, inclusi test e distribuzione di modifiche, aggiunta o rimozione di capacità e migrazione dei dati. AWS CodePipeline consente di gestire le fasi necessarie per rilasciare il carico di lavoro. Questo include uno stato di distribuzione che utilizza AWS CodeDeploy per automatizzare la distribuzione del codice dell'applicazione su istanze Amazon EC2, istanze in locale, funzioni Lambda serverless o servizi Amazon ECS.

> **Consiglio**
>
> Anche se la prassi comune suggerisce di includere le persone nelle procedure operative più difficili, suggeriamo di automatizzare le procedure più difficili proprio per questo motivo.

Anti-pattern comuni:

- Eseguire le modifiche manualmente.

- Ignorare le fasi dell'automazione attraverso i flussi di lavoro di emergenza.
- Non seguire i piani.

Vantaggi dell'adozione di questa best practice: L'utilizzo dell'automazione per distribuire tutte le modifiche scongiura il rischio di introdurre errori umani e consente di effettuare test prima di modificare la produzione, così da garantire che i piani siano completi.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Automatizzazione della pipeline di distribuzione Le pipeline di distribuzione permettono di richiamare test automatici, rilevare le anomalie e interrompere la pipeline a una determinata fase prima della distribuzione in produzione o eseguire automaticamente il ripristino di una modifica.
  - The Amazon Builders' Library: Garantire la sicurezza del rollback durante le distribuzioni
  - The Amazon Builders' Library: Più velocità con una consegna continua
    - Utilizza AWS CodePipeline (o un prodotto di terze parti affidabile) per definire ed eseguire le tue pipeline.
      - Configura la pipeline in modo che inizi quando si effettua il commit di una modifica al repository del codice.
        - Che cos'è AWS CodePipeline?
      - Utilizza Amazon Simple Notification Service (Amazon SNS) e Amazon Simple Email Service (Amazon SES) per inviare notifiche sui problemi nella pipeline o integrarti utilizzando uno strumento di chat per team, ad esempio Amazon Chime.
        - Che cos'è Amazon Simple Notification Service?
        - Che cos'è Amazon SES?
        - Che cos'è Amazon Chime?
        - Automatizza i messaggi delle chat con webhook.

Risorse

Documenti correlati:

- Partner APN: partner per la creazione di soluzioni di distribuzione automatizzate
- Marketplace AWS: prodotti per l'automazione delle distribuzioni

- Automatizza i messaggi delle chat con webhook.
- The Amazon Builders' Library: Garantire la sicurezza del rollback durante le distribuzioni
- The Amazon Builders' Library: Più velocità con una consegna continua
- Che cos'è AWS CodePipeline?
- Che cos'è CodeDeploy?
- AWS Systems Manager Patch Manager
- Che cos'è Amazon SES?
- Che cos'è Amazon Simple Notification Service?

Video correlati:

- AWS Summit 2019: CI/CD su AWS (AWS Summit: CI/CD su AWS)

Gestione degli errori

Domande

- REL 9 In che modo esegui il backup dei dati?
- REL 10 In che modo utilizzi l'isolamento dei guasti per proteggere il carico di lavoro?
- REL 11 In che modo progetti il carico di lavoro affinché resista ai guasti dei componenti?
- REL 12 In che modo testi l'affidabilità?
- REL 13 Come pianifichi il disaster recovery (DR)?

REL 9 In che modo esegui il backup dei dati?

Esegui il backup dei dati, delle applicazioni e della configurazione per soddisfare i tuoi requisiti relativi agli obiettivi di tempo di ripristino (recovery time objective, RTO) e agli obiettivi di punto di ripristino (recovery point objective, RPO).

Best practice

- REL09-BP01 Identificazione e backup di tutti i dati che richiedono un backup o riproduzione dei dati dalle origini
- REL09-BP02 Protezione e codifica dei backup
- REL09-BP03 Esecuzione del backup dei dati in automatico

- REL09-BP04 Ripristino periodico dei dati per verificare l'integrità e i processi di backup:

REL09-BP01 Identificazione e backup di tutti i dati che richiedono un backup o riproduzione dei dati dalle origini

Tutti i data store AWS offrono funzionalità di backup. Servizi come Amazon RDS e Amazon DynamoDB supportano inoltre il backup automatico che consente il ripristino point-in-time (PITR), grazie al quale è possibile ripristinare un backup in qualsiasi momento fino a cinque minuti o meno rispetto all'ora corrente. Molti servizi AWS offrono la possibilità di copiare i backup su un'altra Regione AWS. AWS Backup è uno strumento che consente di centralizzare e automatizzare la protezione dei dati tra i vari servizi AWS.

Amazon S3 può essere utilizzato come destinazione di backup per le origini dei dati gestite dal cliente e gestite da AWS. I servizi AWS come Amazon EBS, Amazon RDS e Amazon DynamoDB hanno funzionalità incorporate per creare i backup. È anche possibile utilizzare software di backup di terze parti.

È possibile eseguire il backup dei dati on-premise in Cloud AWS utilizzando AWS Storage Gateway oppure AWS DataSync. I bucket Amazon S3 possono essere utilizzati per archiviare questi dati su AWS. Amazon S3 offre più livelli di archiviazione, quali Amazon S3 Glacier oppure S3 Glacier Deep Archive per ridurre i costi di archiviazione dei dati.

Potresti essere in grado di soddisfare le esigenze di recupero dei dati riproducendo i dati da altre origini. Ad esempio, I nodi di replica Amazon Elasticache oppure Repliche di lettura RDS possono essere utilizzati per riprodurre i dati in caso di perdita dei dati primari. Nei casi in cui origini di questo tipo possono essere utilizzate per raggiungere l'Obiettivo del punto di ripristino (RPO) e l'Obiettivo del tempo di ripristino (RTO),potrebbe non essere necessario un backup. Un altro esempio: se con Amazon EMR, potrebbe non essere necessario eseguire il backup del data store HDFS, purché sia possibile riprodurre i dati in EMR da S3.

Quando scegli una strategia di backup, devi considerare il tempo necessario per il ripristino dei dati. Il tempo necessario per il ripristino dei dati dipende dal tipo di backup (nel caso di una strategia di backup) o dalla complessità del meccanismo di riproduzione dei dati. Questo tempo deve rientrare nell'RTO per il carico di lavoro.

Risultato desiderato:

le origini dei dati sono state identificate e classificate in base alla criticità. Quindi, stabilisci una strategia per il recupero dei dati in base all'RPO. Questa strategia prevede il backup di queste

origini dei dati o la possibilità di riprodurre i dati da altre origini. In caso di perdita di dati, la strategia implementata consente il recupero o la riproduzione dei dati entro i termini RPO e RTO definiti.

Fase di maturità del cloud: Foundational

Anti-pattern comuni:

- Mancata conoscenza di tutte le origini dei dati per il carico di lavoro e della loro criticità.
- Non si eseguono backup delle origini dei dati critiche.
- Esecuzione di backup solo di alcune origini dei dati senza utilizzare la criticità come criterio.
- Non esiste un RPO definito o la frequenza di backup non può soddisfare l'RPO.
- Nessuna valutazione della necessità di un backup o della possibilità di riprodurre i dati da altre origini.

Vantaggi dell'adozione di questa best practice: L'identificazione dei punti in cui sono necessari i backup e l'implementazione di un meccanismo per la creazione di backup, o la possibilità di riprodurre i dati da una fonte esterna, migliorano la capacità di ripristinare e recuperare i dati durante un'interruzione.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Scopri e utilizza le funzionalità di backup dei servizi e delle risorse AWS utilizzati dal carico di lavoro. La maggior parte dei servizi AWS offre funzionalità per eseguire il backup dei dati del carico di lavoro.

Passaggi dell'implementazione

1. Identificazione di tutte le origini dei dati per il carico di lavoro. I dati possono essere memorizzati su diverse risorse, come ad esempio database, volumi, filesystem, sistemi di registrazionee archiviazione di oggetti. Consulta la sezione Risorse per trovare Documenti correlati ai diversi servizi AWS in cui vengono archiviati i dati e la capacità di backup che questi servizi offrono.
2. Classificazione delle origini dei dati in base alla criticità. I diversi set di dati avranno diversi livelli di criticità per un carico di lavoro e quindi diversi requisiti di resilienza. Ad esempio, alcuni dati possono essere critici e richiedere un RPO prossimo allo zero, mentre altri dati possono essere meno critici e tollerare un RPO più elevato e una certa perdita di dati. Allo stesso modo, anche i diversi set di dati possono avere requisiti RTO diversi.

3. Utilizza i servizi AWS o di terze parti per creare i backup dei dati. AWS Backup è un servizio gestito che permette di creare backup di varie origini dei dati su AWS. La maggior parte di questi servizi dispone anche di funzionalità native per la creazione di backup. Marketplace AWS ha molte soluzioni che offrono anche queste funzionalità. Consulta lo Risorse elencate di seguito per informazioni su come creare backup dei dati da vari servizi AWS.
4. Per i dati non sottoposti a backup, stabilire un meccanismo di riproduzione dei dati. Puoi decidere di non eseguire il backup di dati riproducibili da altre origini per vari motivi. Potrebbe essere più conveniente riprodurre i dati dalle origini, quando necessario, piuttosto che creare un backup, dato che l'archiviazione dei backup può comportare dei costi. Un altro esempio è quello in cui il ripristino da un backup richiede più tempo rispetto alla riproduzione dei dati dalle origini, con conseguente violazione dell'RTO. In queste situazioni, è necessario considerare i compromessi e stabilire un processo ben definito per la riproduzione dei dati da queste origini quando è necessario il ripristino dei dati. Ad esempio, se hai caricato dati da Amazon S3 su un data warehouse (come Amazon Redshift) o su un cluster MapReduce (come Amazon EMR) per compiere analisi, ottieni un esempio pratico di riproduzione dati da oltre origini. Finché i risultati di queste analisi vengono archiviati o sono riproducibili, non subirai una perdita di dati a causa di un guasto nel data warehouse o nel cluster MapReduce. Altri esempi che possono essere riprodotti dalle origini includono le cache (ad esempio Amazon ElastiCache) o le repliche di lettura RDS.
5. Stabilisci una cadenza per il backup dei dati. La creazione di backup delle origini dei dati è un processo periodico e la frequenza deve dipendere dall'RPO.

Livello di impegno per il piano di implementazione: Moderato

Risorse

Best practice correlate:

REL13-BP01 Definizione degli obiettivi di ripristino in caso di downtime e perdita di dati

REL13-BP02 Utilizzo di strategie di ripristino definite per conseguire gli obiettivi di ripristino

Documenti correlati:

- Che cos'è AWS Backup?
- What is AWS DataSync? (Che cos'è AWS DataSync?)
- What is Volume Gateway? (Che cos'è il Gateway di volumi?)
- Partner APN: partner per il backup

- Marketplace AWS: prodotti che possono essere utilizzati per il backup
- Snapshot Amazon EBS
- Backing Up Amazon EFS (Backup di Elastic File System)
- Backup di Amazon FSx per Windows File Server
- Backup e ripristino di ElastiCache for Redis
- Creating a DB Cluster Snapshot in Neptune (Creazione di uno snapshot cluster DB in Neptune)
- Creazione di uno snapshot DB
- Creating an EventBridge Rule That Triggers on a Schedule (Creazione di una regola EventBridge che viene eseguita in base a una pianificazione)
- Replica tra Regioni con Amazon S3
- EFS-to-EFS AWS Backup (Backup da EFS a EFS)
- Esportazione di dati di registro in Amazon S3
- Gestione del ciclo di vita dell'applicazione
- Backup e ripristino on demand per DynamoDB
- Point-in-time recovery for DynamoDB (Ripristino point-in-time per DynamoDB)
- Gestione di snapshot degli indici Amazon OpenSearch Service

Video correlati:

- AWS re:Invent 2021 - Backup, disaster recovery, and ransomware protection with AWS (Backup, ripristino di emergenza e protezione ransomware con AWS)
- AWS Backup Demo: Cross-Account and Cross-Region Backup (Backup trasversale tra account e tra regioni)
- AWS re:Invent 2019: Deep dive on AWS Backup(Approfondimento si AWS Backup), ft. Rackspace (STG341)

Esempi correlati:

- Well-Architected lab: Implementing Bi-Directional Cross-Region Replication (CRR) for Amazon S3 (Laboratorio Well-Architected: Implementazione della replica bi-direzionale tra regioni (CRR) per Amazon S3)
- Corso Well-Architected: Testing Backup and Restore of Data

- Well-Architected lab: Backup and Restore with Failback for Analytics Workload (Laboratorio Well-Architected: Backup e ripristino con failback per il carico di lavoro analitico)
- Well-Architected lab: Disaster Recovery - Backup and Restore (Laboratorio Well-Architected: Ripristino di emergenza – Backup e ripristino)

REL09-BP02 Protezione e codifica dei backup

Controlla e rileva l'accesso ai backup utilizzando l'autenticazione e l'autorizzazione, come ad esempio AWS IAM. Previeni e rileva se l'integrità dei dati dei backup è compromessa utilizzando la crittografia.

Amazon S3 supporta diversi metodi di crittografia dei dati archiviati. Utilizzando la crittografia lato server, Amazon S3 accetta anche dati non crittografati e li crittografa man mano che vengono memorizzati. Utilizzando la crittografia lato client, l'applicazione del carico di lavoro è responsabile della crittografia dei dati prima che vengano inviati a Amazon S3. Entrambi i metodi ti consentono di utilizzare AWS Key Management Service (AWS KMS) per creare ed archiviare la chiave di crittografia dei dati, oppure di utilizzarne una personalizzata (della quale sarai responsabile). Tramite AWS KMS puoi impostare delle policy utilizzando IAM per regolare l'accesso alle chiavi dei dati, oltre che ai dati privi di crittografia.

Per Amazon RDS, se hai scelto di crittografare i database, anche i backup verranno crittografati. I backup di DynamoDB sono sempre crittografati.

Anti-pattern comuni:

- Disporre di un accesso identico sia per i backup e l'automazione del ripristino sia per i dati.
- Non codificare i backup.

Vantaggi dell'adozione di questa best practice: La protezione dei backup previene la manomissione dei dati, mentre la crittografia dei dati impedisce l'accesso in caso di esposizione accidentale.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Utilizzo della crittografia su ciascuno dei datastore. Se i dati di origine sono crittografati, lo sarà anche il backup.
  - Abilitazione della crittografia in RDS. Puoi configurare la crittografia dei dati inattivi utilizzando AWS Key Management Service al momento della creazione di un'istanza RDS.

    - Crittografia delle risorse Amazon RDS
- Abilitazione della crittografia sui volumi EBS. Puoi configurare la crittografia predefinita o specificare una chiave univoca al momento della creazione del volume.
    - Crittografia Amazon EBS
- Utilizza la crittografia Amazon DynamoDB richiesta. DynamoDB crittografa tutti i dati a riposo. Puoi utilizzare una chiave AWS KMS di proprietà di AWS o una chiave KMS gestita da AWS specificando una chiave archiviata nel tuo account.
    - Crittografia a riposo per DynamoDB
    - Gestione di tabelle crittografate
- Codifica dei dati archiviati in Amazon EFS. Configura la crittografia al momento della creazione del file system.
    - Crittografia dei dati e dei metadati in EFS
- Configura la crittografia nelle regioni di origine e di destinazione. Puoi configurare la crittografia dei dati inattivi in Amazon S3 utilizzando le chiavi archiviate in KMS, ma le chiavi sono specifiche per regione. Puoi specificare le chiavi di destinazione quando configuri la replica.
    - Configurazione aggiuntiva CRR: replica di oggetti creati con crittografia lato server (SSE) utilizzando le chiavi di crittografia archiviate in AWS KMS

- Implementazione delle autorizzazioni con privilegi minimi per accedere ai backup. Segui le best practice per limitare l'accesso a backup, snapshot e repliche in conformità con le best practice di sicurezza.
    - Pilastro della sicurezza – AWS Well-Architected

Risorse

Documenti correlati:

- Marketplace AWS: prodotti che possono essere utilizzati per il backup
- Crittografia Amazon EBS
- Amazon S3: protezione dei dati tramite la crittografia
- Configurazione aggiuntiva CRR: replica di oggetti creati con crittografia lato server (SSE) utilizzando le chiavi di crittografia archiviate in AWS KMS
- Crittografia a riposo per DynamoDB
- Crittografia delle risorse Amazon RDS

- Crittografia dei dati e dei metadati in EFS
- Encryption for Backups in AWS (Crittografia per i backup in AWS Backup)
- Gestione di tabelle crittografate
- Pilastro della sicurezza – AWS Well-Architected

Esempi correlati:

- Well-Architected lab: Implementing Bi-Directional Cross-Region Replication (CRR) for Amazon S3 (Laboratorio Well-Architected: Implementazione della replica bi-direzionale tra regioni (CRR) per Amazon S3)

REL09-BP03 Esecuzione del backup dei dati in automatico

Configura i backup in modo che vengano eseguiti automaticamente in base a una pianificazione periodica informata dall'Obiettivo del punto di ripristino (RPO) o dalle modifiche apportate al set di dati. I set di dati critici con bassi requisiti di perdita di dati devono essere sottoposti a backup automatico su base frequente, mentre i dati meno critici, per i quali è accettabile una certa perdita, possono essere sottoposti a backup meno frequenti.

AWS Backup può essere utilizzato per creare backup automatici di varie origini dei dati AWS. Il backup delle istanze Amazon RDS può essere eseguito quasi ininterrottamente ogni cinque minuti e quello degli oggetti Amazon S3 quasi ininterrottamente ogni quindici minuti, consentendo il ripristino point-in-time (PITR) a un punto specifico della cronologia di backup. Per altre origini dei dati AWS, come volumi Amazon EBS, tabelle Amazon DynamoDB o file system Amazon FSx, AWS Backup può eseguire il backup automatico con una frequenza di un'ora. Questi servizi offrono anche funzionalità di backup nativo. I servizi AWS che offrono un backup automatizzato con ripristino point-in-time includono Amazon DynamoDB, Amazon RDSe Amazon Keyspaces (per Apache Cassandra) ; questi possono essere ripristinati a un punto specifico della cronologia di backup. La maggior parte degli altri servizi di archiviazione di dati AWS offre la possibilità di programmare backup periodici, anche ogni ora.

Amazon RDS e Amazon DynamoDB offrono un backup continuo con ripristino point-in-time. Una volta abilitato, il controllo delle versioni Amazon S3 è automatico. Amazon Data Lifecycle Manager può essere utilizzato per automatizzare la creazione, la copia e l'eliminazione degli snapshot Amazon EBS. Può anche automatizzare la creazione, la copia, la rimozione e la cancellazione di Amazon Machine Images (AMI) con backup Amazon EBS e dei relativi snapshot Amazon EBS sottostanti.

Per una visualizzazione centralizzata dell'automazione e della cronologia dei backup, AWS Backup fornisce una soluzione di backup completamente gestita basata su policy. Centralizza e automatizza il backup dei dati su più servizi AWS nel cloud e on-premise utilizzando AWS Storage Gateway.

Oltre a quella di controllo delle versioni, Amazon S3 offre tutte le funzioni di replica. L'intero bucket S3 può essere replicato automaticamente in un altro bucket in una Regione AWS diversa.

Risultato desiderato:

un processo automatizzato che crea backup delle origini dei dati con una cadenza stabilita.

Anti-pattern comuni:

- Eseguire i backup manualmente.
- Utilizzare risorse che dispongono di funzionalità di backup, ma non includere il backup nell'automazione.

Vantaggi dell'adozione di questa best practice: L'automazione dei backup garantisce che vengano eseguiti regolarmente in base all'RPO e avvisa se non vengono eseguiti.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

1. Identifica le origini dei dati al momento sottoposte a backup manuale. Consulta REL09-BP01 Identificazione e backup di tutti i dati che richiedono un backup o riproduzione dei dati dalle origini per avere una guida.
2. Determina l'RPO per il carico di lavoro. Consulta REL13-BP01 Definizione degli obiettivi di ripristino in caso di downtime e perdita di dati per avere una guida.
3. Utilizza una soluzione di backup automatico o un servizio gestito. AWS Backup è un servizio completamente gestito che semplifica la centralizzazione e l'automazione della protezione dei dati tra i servizi AWS, nel cloud e on-premise. I piani di backup sono una funzionalità di AWS Backup che consente di creare regole che definiscono le risorse da sottoporre a backup e la frequenza con cui questi backup devono essere creati. Questa frequenza deve essere informata dall'RPO stabilito al punto 2. Questo laboratorio WA fornisce una guida pratica su come creare backup automatizzati utilizzando AWS Backup. La maggior parte dei servizi AWS di archiviazione dei dati offre funzionalità di backup native. Ad esempio, RDS può essere sfruttato per backup automatici con ripristino point-in-time (PITR).

4. Per le origini dei dati non supportate da una soluzione di backup automatico o da un servizio gestito, come le origini dei dati on-premise o le code di messaggi, è consigliabile utilizzare una soluzione di terze parti affidabile per creare backup automatici. In alternativa, puoi creare un'automazione utilizzando la AWS CLI o gli SDK. Puoi utilizzare le funzioni AWS Lambda o AWS Step Functions per definire la logica di creazione di un backup dei dati e utilizzare Amazon EventBridge per eseguirlo con una frequenza basata sull'RPO (come stabilito nel passaggio 2).

Livello di impegno per il piano di implementazione: Bassa

Risorse

Documenti correlati:

- Partner APN: partner per il backup
- Marketplace AWS: prodotti che possono essere utilizzati per il backup
- Creating an EventBridge Rule That Triggers on a Schedule (Creazione di una regola EventBridge che viene eseguita in base a una pianificazione)
- Che cos'è AWS Backup?
- Che cos'è AWS Step Functions?

Video correlati:

- AWS re:Invent 2019: Deep dive on AWS Backup(Approfondimento si AWS Backup), ft. Rackspace (STG341)

Esempi correlati:

- Corso Well-Architected: Testing Backup and Restore of Data

REL09-BP04 Ripristino periodico dei dati per verificare l'integrità e i processi di backup:

Esegui un test di ripristino per verificare che l'implementazione del processo di backup soddisfi gli obiettivi di tempo di ripristino (recovery time objective, RTO) e gli obiettivi di punto di ripristino (recovery point objective, RPO).

Con AWS, puoi creare un ambiente di test e ripristinare i backup per valutare le funzionalità RTO e RPO ed eseguire test sul contenuto e l'integrità dei dati.

Inoltre, Amazon RDS e Amazon DynamoDB consentono il ripristino point-in-time (PITR). Utilizzando il backup continuo, puoi ripristinare il set di dati allo stato in cui si trovava in una data e un'ora specificate.

Risultato desiderato: I dati dei backup vengono ripristinati periodicamente utilizzando meccanismi ben definiti per garantire che il ripristino sia possibile entro l'Obiettivo del tempo di ripristino (RTO) stabilito per il carico di lavoro. Verifica che il ripristino da un backup porti a una risorsa che contiene i dati originali senza che questi siano danneggiati o inaccessibili e con una perdita di dati entro l'Obiettivo del punto di ripristino (RPO).

Anti-pattern comuni:

- Ripristinare un backup, senza però eseguire query o recuperare dati per garantire che il ripristino sia utilizzabile.
- Presupporre l'esistenza di un backup.
- Presupporre che il backup di un sistema sia pienamente operativo e che i dati possano essere recuperati da esso.
- Presupporre che il tempo di ripristino o di recupero dei dati da un backup rientri nell'RTO del carico di lavoro.
- Presupporre che i dati contenuti nel backup rientrino nell'RPO del carico di lavoro.
- Ripristino ad hoc, senza l'utilizzo di un runbook o al di fuori di una procedura automatizzata consolidata.

Vantaggi dell'adozione di questa best practice: la verifica del ripristino dei backup assicura che i dati possano essere ripristinati quando necessario senza preoccuparsi che possano essere mancanti o danneggiati, che il ripristino e il recupero siano possibili entro l'RTO per il carico di lavoro e che qualsiasi perdita di dati rientri nell'RPO per il carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

La verifica delle capacità di backup e ripristino aumenta la fiducia nella capacità di eseguire queste azioni durante un'interruzione. Ripristina periodicamente i backup in una nuova posizione ed esegui test per verificare l'integrità dei dati. Alcuni test comuni da eseguire sono la verifica che

tutti i dati siano disponibili, non siano danneggiati, siano accessibili e che qualsiasi perdita di dati rientri nell'RPO del carico di lavoro. Questi test possono anche aiutare a verificare se i meccanismi di ripristino sono sufficientemente veloci per soddisfare l'RTO del carico di lavoro.

1. Identifica le origini dei dati di cui si sta eseguendo il backup e dove sono archiviati i backup. Consulta REL09-BP01 Identificazione e backup di tutti i dati che richiedono un backup o riproduzione dei dati dalle origini per una guida all'implementazione.
2. Stabilisci i criteri per la convalida dei dati per ogni origine dei dati. Tipi di dati differenti avranno proprietà diverse che potrebbero richiedere meccanismi di convalida diversi. Considera il modo in cui potrebbero essere convalidati questi dati prima di poterli utilizzare in produzione. Alcuni modi comuni per convalidare i dati sono l'uso delle loro proprietà dei dati e del backup, come il tipo di dati, il formato, la somma di controllo, la dimensione o la combinazione di questi elementi con una logica di convalida personalizzata. Ad esempio, può trattarsi di un confronto dei valori di checksum tra la risorsa ripristinata e l'origine dei dati al momento della creazione del backup.
3. Stabilisci l'RTO e l'RPO per il ripristino dei dati in base alla loro criticità. Consulta REL13-BP01 Definizione degli obiettivi di ripristino in caso di downtime e perdita di dati per una guida all'implementazione.
4. Valuta la capacità di ripristino dei dati. Rivedi la strategia di backup e ripristino per capire se è in grado di soddisfare RTO e RPO e modifica la strategia se necessario. Utilizzando Hub di resilienza AWS, puoi eseguire una valutazione del carico di lavoro. La valutazione esamina la configurazione dell'applicazione rispetto alle policy sulla resilienza e indica se gli obiettivi RTO e RPO possono essere raggiunti.
5. Esegui un ripristino di prova utilizzando i processi attualmente in uso in produzione per il ripristino dei dati. Questi processi dipendono dal modo in cui è stato eseguito il backup dell'origine dei dati iniziale, dal formato e dalla posizione di archiviazione del backup stesso o dalla riproduzione dei dati da altre fonti. Ad esempio, utilizzi un servizio gestito come AWS Backup, questo potrebbe essere semplice come il ripristino del backup in una nuova risorsa. Se hai utilizzato il Ripristino di emergenza elastico AWS, puoi avviare un'analisi di ripristino.
6. Convalida il ripristino dei dati dalla risorsa ripristinata (dal passo precedente) in base ai criteri stabiliti in precedenza per la convalida dei dati al passo 2. I dati ripristinati e recuperati contengono il record/la voce più recente al momento del backup? Questi dati rientrano nell'RPO per il carico di lavoro?
7. Misura il tempo richiesto per il ripristino e il recupero e confrontalo con l'RTO stabilito in precedenza nel passaggio 3. Questo tempo deve rientrare nell'RTO per il carico di lavoro? Ad esempio, confronta i timestamp dell'inizio del processo di ripristino e del completamento della convalida del ripristino per calcolare la durata del processo. Tutte le chiamate API AWS hanno

una datazione temporale e queste informazioni sono disponibili in AWS CloudTrail. Sebbene queste informazioni possano fornire dettagli sull'inizio del processo di ripristino, la logica di convalida dovrebbe registrare il timestamp finale del completamento della convalida. Se utilizzi un processo automatizzato, puoi utilizzare servizi come Amazon DynamoDB per l'archiviazione di queste informazioni. Inoltre, molti servizi AWS offrono una cronologia degli eventi che fornisce informazioni con data e ora in cui si sono verificate determinate azioni. All'interno di AWS Backup, le azioni di backup e di ripristino sono denominate processie questi processi contengono informazioni sulla data e l'ora come parte dei metadati che possono essere utilizzati per misurare il tempo necessario per il ripristino e il recupero.

8. Invia notifica alle parti interessate (stakeholder) se la convalida dei dati non riesce o se il tempo necessario per il ripristino e il recupero supera l'RTO stabilito per il carico di lavoro. Quando si implementa l'automazione per farlo, come in questo laboratorio,servizi come Amazon Simple Notification Service (Amazon SNS) possono essere utilizzati per inviare notifiche push, come e-mail o SMS, alle parti interessate. Questi messaggi possono anche essere pubblicati su applicazioni di messaggistica come Amazon Chime, Slack o Microsoft Teams o utilizzati per creare attività come OpsItem utilizzando OpsCenter di AWS Systems Manager.
9. Automatizzare questo processo per eseguirlo periodicamente. Ad esempio, per automatizzare i processi di ripristino e recupero si possono utilizzare servizi come AWS Lambda o una State Machine in AWS Step Functions, mentre Amazon EventBridge può essere utilizzato per attivare periodicamente questo flusso di lavoro di automazione, come mostrato nel diagramma di architettura sottostante. Scopri come automatizzare la convalida del ripristino dati con AWS Backup. Inoltre, questo laboratorio Well-Architected fornisce un'esperienza pratica su come realizzare l'automazione di alcuni dei passaggi qui descritti.

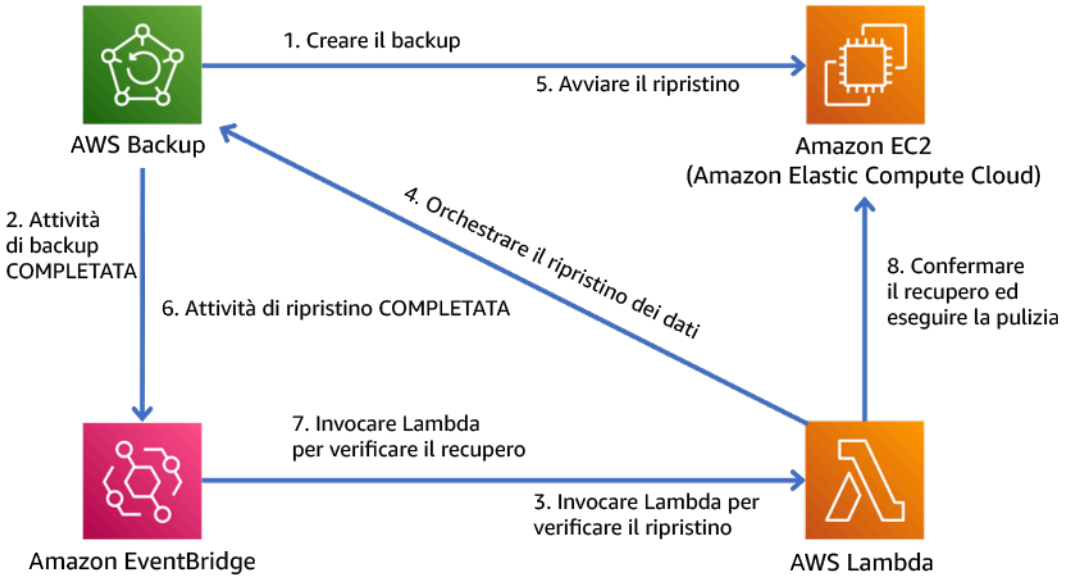


Figura 9. Un processo di backup e ripristino automatizzato

Livello di impegno per il piano di implementazione: da moderato a elevato, a seconda della complessità dei criteri di convalida.

Risorse

Documenti correlati:

- automatizzare la convalida del ripristino dati con AWS Backup
- Partner APN: partner per il backup
- Marketplace AWS: prodotti che possono essere utilizzati per il backup
- Creating an EventBridge Rule That Triggers on a Schedule (Creazione di una regola EventBridge che viene eseguita in base a una pianificazione)
- Backup e ripristino on demand per DynamoDB
- Che cos'è AWS Backup?
- Che cos'è AWS Step Functions?
- What is AWS Elastic Disaster Recovery (Che cos'è il ripristino di emergenza elastico AWS?)
- AWS Elastic Disaster Recovery (Ripristino di emergenza elastico AWS)

Esempi correlati:

- Corso Well-Architected: Testing Backup and Restore of Data

REL 10 In che modo utilizzi l'isolamento dei guasti per proteggere il carico di lavoro?

Le barriere per l'isolamento dei guasti limitano l'effetto di un errore all'interno di un carico di lavoro a un numero limitato di componenti. I componenti al di fuori della barriera non subiscono gli effetti del guasto. Utilizzando più barriere per l'isolamento dei guasti, puoi limitare l'impatto sul carico di lavoro.

Best practice

- REL10-BP01 Implementazione del carico di lavoro in diversi luoghi
- REL10-BP02 Selezione delle posizioni appropriate per la tua implementazione multiposizione
- REL10-BP03 Ripristino automatico dei componenti vincolati a una singola posizione
- REL10-BP04 Utilizzo di architetture a paratie per limitare la portata dell'impatto

REL10-BP01 Implementazione del carico di lavoro in diversi luoghi

Distribuisci i dati e le risorse del carico di lavoro su più zone di disponibilità o, se necessario, su diverse Regioni AWS. Questi luoghi possono essere diversi a seconda delle necessità.

Uno dei principi fondamentali per la progettazione di servizi su AWS è l'eliminazione di singoli punti di errore nell'infrastruttura fisica sottostante. Questo ci spinge a creare software e sistemi che utilizzano più zone di disponibilità e sono resistenti al fallimento di una singola zona. Allo stesso modo, i sistemi sono costruiti per resistere ai guasti di un singolo nodo di calcolo, singolo volume di archiviazione o singola istanza di un database. Quando si costruisce un sistema che si basa su componenti ridondanti, è importante garantire che i componenti funzionino in modo indipendente e, nel caso delle Regioni AWS, in modo autonomo. I vantaggi ottenuti dai calcoli di disponibilità teorica con componenti ridondanti sono validi solo se questo continua a essere vero.

Zone di disponibilità (AZ)

Le Regioni AWS sono composte da almeno due zone di disponibilità progettate per essere indipendenti. Ogni zona di disponibilità è separata da una distanza fisica significativa da altre zone per evitare scenari di guasto correlati, dovuti a rischi ambientali come incendi, inondazioni e tornado. Ogni zona di disponibilità ha anche un'infrastruttura fisica indipendente: connessioni dedicate di alimentazione di rete, fonti di alimentazione di backup autonome, servizi meccanici indipendenti e connettività di rete indipendente all'interno e all'esterno della zona di disponibilità. Questa struttura limita gli errori di uno qualsiasi di questi sistemi alla sola AZ interessata. Nonostante siano geograficamente separate, le zone di disponibilità sono situate nella stessa area regionale, il che consente una rete a velocità di trasmissione effettiva elevata e bassa latenza. L'intera Regione AWS (in tutte le zone di disponibilità, costituite da più data center fisicamente indipendenti) può

essere trattata come un unico obiettivo logico di implementazione per il carico di lavoro, compresa la possibilità di replicare i dati in modo sincrono (ad esempio, tra i database). Ciò ti consente di utilizzare le zone di disponibilità in una configurazione attiva/attiva o attiva/standby.

Le zone di disponibilità sono indipendenti e pertanto la disponibilità del carico di lavoro aumenta quando il carico di lavoro è progettato per utilizzare più zone di disponibilità. Alcuni servizi AWS (tra cui il piano dati dell'istanza Amazon EC2) sono implementati come servizi strettamente zonali nei quali hanno un destino condiviso con la zona di disponibilità in cui si trovano. Le istanze Amazon EC2 nelle altre AZ non saranno, tuttavia, interessate e continueranno a funzionare. Allo stesso modo, se un errore in una zona di disponibilità causa l'errore di un database Amazon Aurora, un'istanza Aurora di lettura-replica in una AZ non interessata può essere automaticamente promossa a primaria. I servizi regionali AWS, ad esempio Amazon DynamoDB, utilizzano internamente più zone di disponibilità in una configurazione attiva/attiva per raggiungere gli obiettivi di progettazione della disponibilità per quel servizio, senza che sia necessario configurare il posizionamento delle AZ.

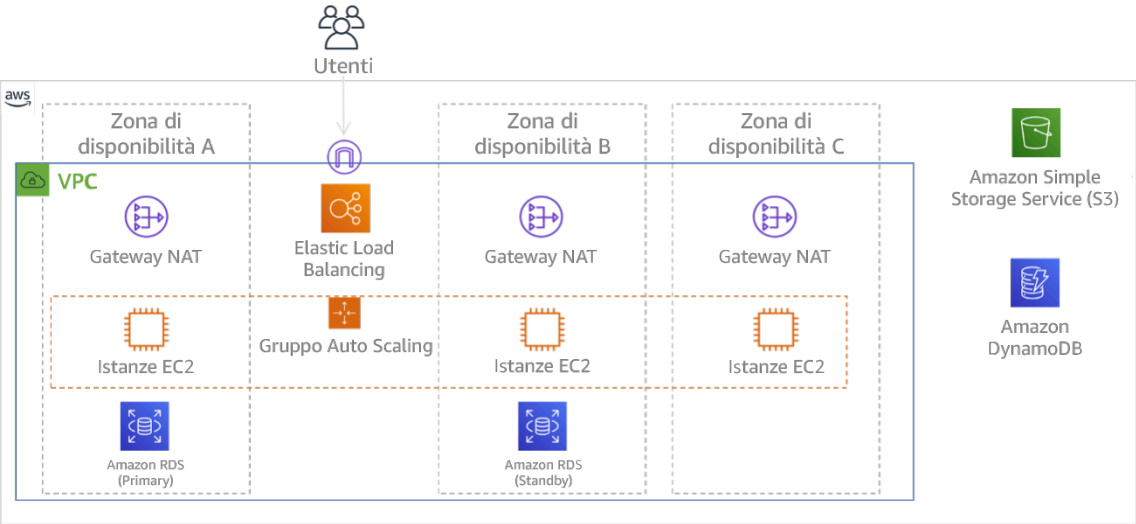


Figura 9. Architettura multi-livello distribuita su tre zone di disponibilità. Tieni presente che Amazon S3 e Amazon DynamoDB sono sempre Multi-AZ automaticamente. L'ELB viene inoltre distribuito in tutte e tre le zone.

Mentre i piani di controllo AWS in genere offrono la possibilità di gestire le risorse all'interno dell'intera Regione (più zone di disponibilità), alcuni piani di controllo (inclusi Amazon EC2 ed Amazon EBS) hanno la capacità di filtrare i risultati per una singola zona di disponibilità. Con questo approccio, la richiesta viene elaborata solo nella zona di disponibilità specificata, riducendo l'esposizione all'interruzione in altre zone di disponibilità. Questo esempio di AWS CLI illustra come ottenere informazioni su un'istanza Amazon EC2 dalla sola zona di disponibilità us-east-2c:

```
AWS ec2 describe-instances --filters Name=availability-zone,Values=us-east-2c
```

Zone locali AWS

Le Zone locali AWS agiscono in modo simile alle zone di disponibilità nella rispettiva Regione AWS, in quanto possono essere selezionate come ubicazione di posizionamento per le risorse AWS zonali come le sottoreti e le istanze EC2. Ciò che le rende speciali è che non si trovano nella Regione AWS associata, ma vicino a grandi popolazioni, settori e centri IT in cui al momento non esiste alcuna Regione AWS. Tuttavia, mantengono una connessione sicura e a larghezza di banda elevata tra i carichi di lavoro locali nella zona locale e quelli in esecuzione nella Regione AWS. È consigliabile utilizzare le Zone locali AWS per implementare i carichi di lavoro più vicini agli utenti per requisiti a bassa latenza.

Amazon Global Edge Network

Amazon Global Edge Network è costituito da posizioni edge in città di tutto il mondo. Amazon CloudFront utilizza questa rete per fornire contenuti agli utenti finali con una latenza inferiore. AWS Global Accelerator consente di creare gli endpoint del carico di lavoro in queste posizioni edge per fornire l'onboarding alla rete globale AWS vicino agli utenti. Amazon API Gateway permette agli endpoint API ottimizzati per l'edge che utilizzano una distribuzione CloudFront di facilitare l'accesso dei clienti attraverso la posizione edge più vicina.

Regioni AWS

Le Regioni AWS sono progettate per essere autonome; pertanto, per utilizzare un approccio multi-regione, puoi implementare copie dedicate dei servizi in ciascuna Regione.

Un approccio multi-regione è comune per le strategie di ripristino di emergenza per raggiungere gli obiettivi di ripristino quando si verificano eventi unici su larga scala. Consulta Pianificazione per il disaster recovery (DR) per ulteriori informazioni su queste strategie. Qui, tuttavia, si focalizza l'attenzione sulla disponibilità, che cerca di fornire un obiettivo medio di operatività nel tempo. Per gli obiettivi di alta disponibilità, un'architettura multi-regione sarà generalmente progettata per essere attiva/attiva, dove ogni copia del servizio (nelle rispettive Regioni) è attiva (serve le richieste).

> **Consiglio**
>
> Gli obiettivi di disponibilità per la maggior parte dei carichi di lavoro possono essere soddisfatti utilizzando una strategia multi-AZ all'interno di una singola Regione AWS.

Considera le architetture multi-regione solo quando i carichi di lavoro hanno requisiti di disponibilità estremi o altri obiettivi aziendali che richiedono un'architettura multi-regione.

AWS offre ai clienti la possibilità di gestire servizi in più Regioni. Ad esempio, AWS fornisce una replica continua e asincrona dei dati utilizzando la replica Amazon Simple Storage Service (Amazon S3), le repliche di lettura Amazon RDS (incluse le repliche di lettura Aurora) e le tabelle globali Amazon DynamoDB. Con la replica continua, le versioni dei dati sono disponibili per un uso quasi immediato in ogni Regione attiva.

Utilizzando AWS CloudFormation, puoi definire l'infrastruttura e implementarla in modo coerente sugli Account AWS e sulle Regioni AWS. Invece, AWS CloudFormation StackSets estende questa funzionalità consentendo di creare, aggiornare o eliminare stack AWS CloudFormation su più account e regioni con un'unica operazione. Per le implementazioni di istanza Amazon EC2, si utilizza un'immagine AMI (Amazon Machine Image) per fornire informazioni quali la configurazione hardware e il software installato. È possibile implementare una pipeline di Amazon EC2 Image Builder che crea le AMI necessarie e le copia nelle regioni attive. Ciò garantisce che le Golden AMI abbiano tutto ciò che serve per implementare e dimensionare il carico di lavoro in ogni nuova regione.

Per instradare il traffico, sia Amazon Route 53 sia AWS Global Accelerator abilitano la definizione di criteri che determinano quali utenti indirizzare a ogni endpoint regionale attivo. Con Global Accelerator imposti un valore di traffico per controllare la percentuale di traffico diretta a ciascun endpoint dell'applicazione. Route 53 supporta questo approccio percentuale e anche diverse altre policy disponibili, tra cui quelle basate sulla geoprossimità e sulla latenza. Global Accelerator sfrutta automaticamente la vasta rete di server edge AWS per convogliare il traffico verso la dorsale di rete AWS il prima possibile, con conseguente riduzione delle latenze delle richieste.

Tutte queste capacità operano in modo da preservare l'autonomia di ogni Regione. Ci sono pochissime eccezioni a questo approccio, inclusi i nostri servizi che forniscono distribuzione edge globale (ad esempio Amazon CloudFront e Amazon Route 53), insieme al piano di controllo per il servizio AWS Identity and Access Management (IAM). La maggior parte dei servizi opera interamente all'interno di una singola Regione.

Data center in locale

Per i carichi di lavoro eseguiti in un data center on-premise, puoi progettare un'esperienza ibrida quando possibile. AWS Direct Connect fornisce una connessione di rete dedicata dalla tua sede ad AWS che consente l'esecuzione in entrambi.

Un'altra opzione è quella di eseguire l'infrastruttura AWS e i servizi on-premise utilizzando AWS Outposts. AWS Outposts è un servizio completamente gestito che estende l'infrastruttura AWS, i servizi AWS, le API e gli strumenti al tuo data center. La stessa infrastruttura hardware utilizzata nel Cloud AWS viene installata nel data center. AWS Outposts è, quindi, connesso alla Regione AWS più vicina. Puoi quindi utilizzare AWS Outposts per supportare i carichi di lavoro che hanno requisiti di bassa latenza o di elaborazione dei dati locali.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Utilizza zone di disponibilità multiple e Regioni AWS. Distribuisci i dati e le risorse del carico di lavoro su più zone di disponibilità o, se necessario, su diverse Regioni AWS. Questi luoghi possono essere diversi a seconda delle necessità.
  - I servizi regionali sono distribuiti intrinsecamente in zone di disponibilità.
    - Sono inclusi Amazon S3, Amazon DynamoDB e AWS Lambda (se non collegati a un VPC)
  - Distribuisci il tuo container, istanza e carichi di lavoro basati su funzioni in più zone di disponibilità. Utilizza datastore multi-zona, inclusi sistemi di cache. Utilizza le funzionalità di dimensionamento automatico EC2, posizionamento di attività ECS, configurazione della funzione AWS Lambda in esecuzione nel tuo VPC e i cluster ElastiCache.
    - Utilizza sottoreti che sono in zone di disponibilità separate nella distribuzione di gruppi Auto Scaling.
      - Esempio: distribuzione di istanze in più zone di disponibilità
      - Strategie di posizionamento dei processi di Amazon ECS
      - Configurazione di una funzione AWS Lambda per accedere alle risorse in un Amazon VPC
      - Scelta di regioni e zone di disponibilità
    - Utilizza sottoreti in zone di disponibilità separate quando distribuisci gruppi Auto Scaling.
      - Esempio: distribuzione di istanze in più zone di disponibilità
    - Utilizza parametri di posizionamento attività ECS, specificando i gruppi di sottorete DB.
      - Strategie di posizionamento dei processi di Amazon ECS
    - Utilizza sottoreti in più zone di disponibilità quando configuri una funzione da eseguire nel tuo VPC.
      - Configurazione di una funzione AWS Lambda per accedere alle risorse in un Amazon VPC
    - Utilizza più zone di disponibilità con cluster ElastiCache.
      - Scelta di regioni e zone di disponibilità

- Se il carico di lavoro deve essere implementato in più Regioni, scegli una strategia multi-regione. La maggior parte delle esigenze di affidabilità può essere soddisfatta all'interno di una singola Regione AWS utilizzando una strategia a più zone di disponibilità. Quando necessario, utilizza una strategia multi-Regione per soddisfare le tue esigenze aziendali.
  - AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Modelli architetturali per applicazioni attive-attive su più Regioni) (ARC209-R2)
    - Il backup in un'altra Regione AWS può garantire ulteriormente che i dati saranno disponibili quando necessario.
    - Alcuni carichi di lavoro hanno requisiti normativi che prevedono l'utilizzo di una strategia multi-regione.
- Valuta AWS Outposts per il tuo carico di lavoro. Se il carico di lavoro richiede bassa latenza nel data center locale o ha requisiti di elaborazione dei dati locali. In tal caso esegui l'infrastruttura e i servizi AWS on-premise utilizzando AWS Outposts.
  - Che cos'è AWS Outposts?
- Stabilisci se le Zone locali AWS ti aiutano a fornire il servizio ai tuoi utenti. Se hai requisiti di bassa latenza, verifica se le Zone locali AWS si trovano vicino ai tuoi utenti. Se sì, utilizzale per implementare carichi di lavoro più vicini a tali utenti.
  - Domande frequenti sulle Zone locali AWS

Risorse

Documenti correlati:

- Infrastruttura globale di AWS
- Domande frequenti sulle Zone locali AWS
- Strategie di posizionamento dei processi di Amazon ECS
- Scelta di regioni e zone di disponibilità
- Esempio: distribuzione di istanze in più zone di disponibilità
- Tabelle globali: replica multi-regione con DynamoDB
- Using Amazon Aurora global databases (Utilizzo di database Amazon Aurora globali)
- Creating a Multi-Region Application with AWS Services blog series (Creazione di un'applicazione multi-regione con la serie di blog sui servizi AWS)
- Che cos'è AWS Outposts?

Video correlati:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Modelli architetturali per applicazioni attive-attive su più Regioni) (ARC209-R2)
- AWS re:Invent 2019: Innovation and operation of the AWS global network infrastructure (Innovazione e gestione dell'infrastruttura di rete globale AWS) (NET339)

REL10-BP02 Selezione delle posizioni appropriate per la tua implementazione multiposizione

Risultato desiderato

Per ottenere un'elevata disponibilità, distribuisci sempre (quando possibile) i componenti del carico di lavoro in più zone di disponibilità (AZ), come illustrato nella Figura 10. Per i carichi di lavoro con requisiti di resilienza estremi, valuta attentamente le opzioni per un'architettura multiregione.

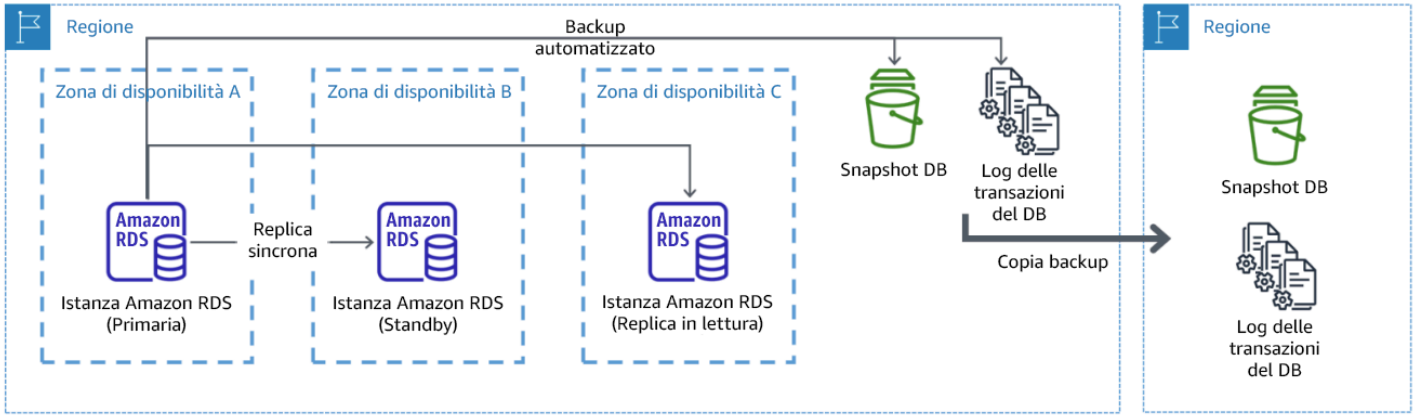


Figura 10: Distribuzione resiliente di un database multi-AZ con backup in un'altra regione AWS

Anti-pattern comuni

- Scelta di progettare un'architettura multi-regione quando un'architettura multi-AZ soddisferebbe i requisiti.
- Non si tiene conto delle dipendenze tra i componenti dell'applicazione se i requisiti di resilienza e multi-sede differiscono tra questi componenti.

Vantaggi dell'adozione di questa best practice

Per la resilienza, devi utilizzare un approccio che costruisca livelli di difesa. Un livello protegge dalle interruzioni più piccole e più comuni costruendo un'architettura ad alta disponibilità utilizzando più

AZ. Un altro livello di difesa è destinato a proteggere da eventi rari come disastri naturali diffusi e interruzioni a livello regionale. Questo secondo livello implica l'architettura dell'applicazione in modo che si estenda su più Regioni AWS.

- La differenza tra una disponibilità del 99,5% e una del 99,99% è di oltre 3,5 ore al mese. La disponibilità prevista di un carico di lavoro può raggiungere i "quattro nove" solo se si trova in più AZ.
- Eseguendo il carico di lavoro in più AZ, puoi isolare gli errori di alimentazione, raffreddamento e rete e la maggior parte dei disastri naturali come incendi e inondazioni.
- L'implementazione di una strategia multi-regione per il tuo carico di lavoro aiuta a proteggerlo da disastri naturali diffusi che colpiscono un'ampia regione geografica di un paese o da guasti tecnici di portata regionale. Tieni presente che l'implementazione di un'architettura multi-regione può essere molto complessa e di solito non è necessaria per la maggior parte dei carichi di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Per un evento disastroso basato sull'interruzione o la perdita parziale di una zona di disponibilità, l'implementazione di un carico di lavoro a disponibilità elevata in più zone di disponibilità all'interno di una singola Regione AWS aiuta a mitigare i disastri naturali e tecnici. Ogni Regione AWS è composta da più zone di disponibilità, ciascuna isolata dagli errori nelle altre zone e separate da una distanza significativa. Tuttavia, per un evento di disastro che include il rischio di perdere più componenti della zona di disponibilità, che si trovano a una distanza significativa l'uno dall'altro, è necessario implementare opzioni di ripristino di emergenza per mitigare gli errori di portata regionale. Per i carichi di lavoro che richiedono un'estrema resilienza (infrastrutture critiche, applicazioni sanitarie, infrastrutture di sistemi finanziari e così via), può essere necessaria una strategia multi-regione.

Passaggi dell'implementazione

1. Valutare il carico di lavoro e determinare se le esigenze di resilienza possono essere soddisfatte da un approccio multi-AZ (Regione AWS singola) o se richiedono un approccio multi-regione. L'implementazione di un'architettura multi-regione per soddisfare questi requisiti introdurrà un'ulteriore complessità, quindi considera attentamente il tuo caso d'uso e i suoi requisiti. I requisiti di resilienza possono quasi sempre essere soddisfatti utilizzando un singolo Regione AWS. Per stabilire se è necessario utilizzare più Regioni, considera i seguenti possibili requisiti:
   a. Ripristino di emergenza: per un evento disastroso basato sull'interruzione o la perdita parziale di una zona di disponibilità, l'implementazione di un carico di lavoro a disponibilità elevata in più

zone di disponibilità all'interno di una singola Regione AWS aiuta a mitigare i disastri naturali e tecnici. In caso di eventi disastrosi che comportano il rischio di perdere più componenti della zone di disponibilità, che si trovano a una distanza significativa l'uno dall'altro, è necessario implementare il ripristino di emergenza in più regioni per mitigare i disastri naturali o gli errori tecnici di portata regionale.

b. Alta disponibilità: è possibile utilizzare un'architettura multi-regione (utilizzando più AZ in ogni regione) per ottenere una disponibilità superiore a quattro 9 (> 99,99%).

c. Localizzazione delle risorse: quando si distribuisce un carico di lavoro a un pubblico globale, è possibile distribuire stack localizzati in diverse Regioni AWS per servire il pubblico di quelle regioni. La localizzazione può includere la lingua, la valuta e i tipi di dati memorizzati.

d. Prossimità agli utenti: quando si distribuisce un carico di lavoro a un pubblico globale, è possibile ridurre la latenza distribuendo gli stack alle regioni Regioni AWS in prossimità degli utenti finali.

e. Posizione fisica dei dati: alcuni carichi di lavoro sono soggetti a requisiti di residenza dei dati, in base ai quali i dati di determinati utenti devono rimanere all'interno dei confini di un determinato Paese. In base alla normativa in questione, è possibile scegliere di distribuire un intero stack o solo i dati nella Regione AWS all'interno di tali confini.

2. Ecco alcuni esempi di funzionalità multi-AZ fornite dai servizi AWS:

a. Per proteggere i carichi di lavoro che utilizzano EC2 o ECS, è necessario distribuire un Elastic Load Balancer davanti alle risorse di calcolo. Elastic Load Balancing quindi fornisce la soluzione per rilevare le istanze nelle zone non integre e instradare il traffico verso quelle integre.

   i. Nozioni di base su Application Load Balancers

   ii. Nozioni di base su Network Load Balancer

b. Nel caso di istanze EC2 che eseguono software commerciale pronto all'uso e che non supportano il bilanciamento del carico, puoi ottenere una forma di tolleranza ai guasti implementando una metodologia di ripristino di emergenza multi-AZ.

   i. the section called “REL13-BP02 Utilizzo di strategie di ripristino definite per conseguire gli obiettivi di ripristino”

c. Per le attività Amazon ECS, distribuire il servizio in modo uniforme su tre AZ per ottenere un equilibrio tra disponibilità e costi.

   i. Amazon ECS availability best practices | Containers (Best practice di disponibilità ECS | Container)

d. Per non Aurora Amazon RDS, puoi scegliere multi-AZ come opzione di configurazione. In caso di errore dell'istanza del database primario, Amazon RDS promuove automaticamente

un database standby per ricevere il traffico in un'altra zona di disponibilità. Puoi inoltre creare repliche di lettura multi-regione per migliorare la resilienza.

   i. Implementazioni Multi-AZ Amazon RDS

   ii. Creazione di una replica di lettura in un'altra Regione AWS

3. Ecco alcuni esempi di funzionalità multi-AZ fornite dai servizi AWS:

   a. Per i carichi di lavoro Amazon S3 in cui la disponibilità multi-AZ è fornita automaticamente dal servizio, considera i punti di accesso multi-regione se è necessaria un'implementazione multi-regione.

      i. Punti di accesso multi-regione in Amazon S3

   b. Per le tabelle DynamoDB in cui la disponibilità multi-AZ è fornita automaticamente dal servizio, è possibile convertire facilmente le tabelle esistenti in tabelle globali per sfruttare più regioni.

      i. Convert Your Single-Region Amazon DynamoDB Tables to Global Tables (Convertire le tabelle Amazon DynamoDB di una singola regione in tabelle globali)

   c. Se il carico di lavoro è gestito da Application Load Balancers o da Network Load Balancer, utilizza AWS Global Accelerator per migliorare la disponibilità dell'applicazione indirizzando il traffico verso più regioni che contengono endpoint integri.

      i. Endpoints for standard accelerators in AWS Global Accelerator - AWS Global Accelerator (Endpoint per acceleratori standard in AWS Global Accelerator) (amazon.com)

   d. Per le applicazioni che sfruttano AWS EventBridge, considera i bus tre regioni per inoltrare gli eventi ad altre regioni selezionate.

      i. Sending and receiving Amazon EventBridge events between Regioni AWS (Invio e ricezione di eventi Amazon EventBridge tra regioni AWS)

   e. Per i database Amazon Aurora, considera i database globali Aurora, che si estendono su più regioni AWS. I cluster esistenti possono essere modificati per aggiungere anche nuove Regioni.

      i. Nozioni di base sui database globali Amazon Aurora

   f. Se il carico di lavoro include chiavi di crittografia AWS Key Management Service (AWS KMS), valuta se le chiavi multi-regione sono adatte all'applicazione.

      i. Chiavi multi-regione in AWS KMS

   g. Per altre funzionalità del servizio AWS, vedi questa serie di blog su Creating a Multi-Region Application with AWS Services series (Creazione di un'applicazione multi-regione con la serie di servizi AWS)

Livello di impegno per il piano di implementazione: da moderato ad alto

Risorse

Documenti correlati:

- Creating a Multi-Region Application with AWS Services series (Creazione di un'applicazione multi-regione con la serie di servizi AWS)
- Disaster Recovery (DR) Architecture on AWS, Part IV: Multi-site Active/Active (Architettura di ripristino di emergenza su AWS, parte IV: attiva/attiva multi-sito)
- Infrastruttura globale di AWS
- Domande frequenti su AWS Local Zones
- Disaster Recovery (DR) Architecture on AWS, Part I: Strategies for Recovery in the Cloud (Architettura di ripristino di emergenza su AWS parte I: strategie per il ripristino nel cloud)
- Il ripristino di emergenza è differente nel cloud
- Tabelle globali: replica multi-regione con DynamoDB

Video correlati:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Modelli di architettura per applicazioni attive-attive multiregione) (ARC209-R2)
- Auth0: architettura ad alta disponibilità multi-Regione che raggiunge più di 1,5 miliardi di accessi al mese con failover automatico

Esempi correlati:

- Disaster Recovery (DR) Architecture on AWS, Part I: Strategies for Recovery in the Cloud (Architettura di ripristino di emergenza su AWS parte I: strategie per il ripristino nel cloud)
- DTCC raggiunge livelli di resilienza superiori a quelli che raggiunge on-premise
- Expedia Group utilizza un'architettura multi-regione, a più zone di disponibilità con un servizio DNS proprietario per aggiungere resilienza alle applicazioni
- Uber: ripristino di emergenza per Kafka multi-Regione
- Netflix: attivo-attivo per la resilienza multi-regione
- Come costruiamo la posizione fisica dei dati per Atlassian Cloud
- Intuit TurboTax funziona in due regioni

REL10-BP03 Ripristino automatico dei componenti vincolati a una singola posizione

Se i componenti del carico di lavoro possono essere eseguiti solo in una singola zona di disponibilità o in un data center on-premise, è necessario implementare la capacità di eseguire una ricostruzione completa del carico di lavoro entro gli obiettivi di ripristino definiti.

Se, a causa di vincoli tecnologici, non è possibile seguire le linee guida per distribuire il carico di lavoro in più posizioni, è necessario implementare un percorso alternativo mirato alla resilienza. È necessario automatizzare la possibilità di ricreare l'infrastruttura necessaria, ridistribuire le applicazioni e ricreare i dati necessari per questi casi.

Ad esempio, Amazon EMR lancia tutti i nodi per un determinato cluster nella stessa zona di disponibilità: eseguire un cluster nella stessa zona migliora le prestazioni dei flussi di lavoro poiché fornisce una velocità di accesso ai dati più elevata. Se questo componente è necessario per la resilienza del carico di lavoro, è necessario disporre di un modo per implementare nuovamente il cluster e i relativi dati. Inoltre, per Amazon EMR è necessario effettuare il provisioning della ridondanza in modi diversi dall'utilizzo di Multi-AZ. È possibile effettuare il provisioning di nodi multipli. Utilizzando EMR File System (EMRFS), i dati in EMR possono essere memorizzati in Amazon S3, che a sua volta può essere replicato su più zone di disponibilità o Regioni AWS.

Analogamente, Amazon Redshift per impostazione predefinita effettua il provisioning del cluster in una zona di disponibilità casuale all'interno della Regione AWS selezionata. Tutti i nodi del cluster vengono assegnati nella stessa zona.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Implementa l'autoriparazione. Distribuisci le tue istanze o container utilizzando, quando possibile, il ridimensionamento automatico. Se non è possibile utilizzare il ridimensionamento automatico, utilizza il ripristino automatico per istanze EC2 o implementa l'automazione di autoriparazione in base agli eventi del ciclo di vita di container Amazon EC2 o ECS.
  - Utilizza gruppi Auto Scaling per carichi di lavoro di container e istanze che non richiedono un indirizzo IP di una singola istanza, un indirizzo IP privato, un indirizzo IP elastico o metadati di istanza.
    - Che cos'è EC2 Auto Scaling?
    - Scalabilità automatica del servizio
      - È possibile impiegare i dati utente di configurazione del lancio per implementare l'automazione capace di autoriparare la maggior parte dei carichi di lavoro.

- Utilizza il ripristino automatico delle istanze EC2 per carichi di lavoro che richiedono un indirizzo ID di una singola istanza, indirizzo IP privato, indirizzo IP elastico e metadati di istanza.
  - Recover your instance.
    - Il ripristino automatico invierà avvisi sullo stato del ripristino a un argomento SNS quando viene rilevato l'errore dell'istanza.
- Utilizza eventi del ciclo di vita di istanze EC2 o eventi ECS per automatizzare l'autoriparazione dove non è possibile utilizzare l'Auto Scaling o il ripristino EC2.
  - EC2 Auto Scaling lifecycle hooks
  - Eventi Amazon ECS
    - Utilizza gli eventi per invocare l'automazione che riparerà il tuo componente secondo la logica di processo richiesta.

Risorse

Documenti correlati:

- Eventi Amazon ECS
- EC2 Auto Scaling lifecycle hooks
- Recover your instance.
- Scalabilità automatica del servizio
- Che cos'è EC2 Auto Scaling?

REL10-BP04 Utilizzo di architetture a paratie per limitare la portata dell'impatto

Come per le paratie su una nave, questo modello garantisce il contenimento di un guasto in un piccolo sottoinsieme di richieste o clienti, in modo che il numero di richieste danneggiate sia limitato e si possa comunque continuare senza errori. Le paratie per i dati sono spesso chiamate partizioni, mentre le paratie per i servizi sono note come celle.

In una architettura basata su celle, ogni cella è un'istanza completa e indipendente del servizio e ha una dimensione massima fissa. Con l'aumentare del carico, i carichi di lavoro aumentano aggiungendo più celle. Una chiave di partizione viene utilizzata sul traffico in entrata per determinare quale cella elaborerà la richiesta. Qualsiasi guasto è contenuto nella singola cella in cui si verifica, in modo che il numero di richieste danneggiate sia limitato man mano che le altre celle continuano senza errori. È importante identificare la chiave di partizione corretta per ridurre al minimo le

interazioni tra celle ed evitare la necessità di coinvolgere servizi di mappatura complessi in ogni richiesta. I servizi che richiedono una mappatura complessa finiscono semplicemente per spostare il problema ai servizi di mappatura, là dove i servizi che richiedono interazioni cross-cell creano dipendenze tra celle (e questo riduce i miglioramenti della disponibilità che ne deriverebbero).

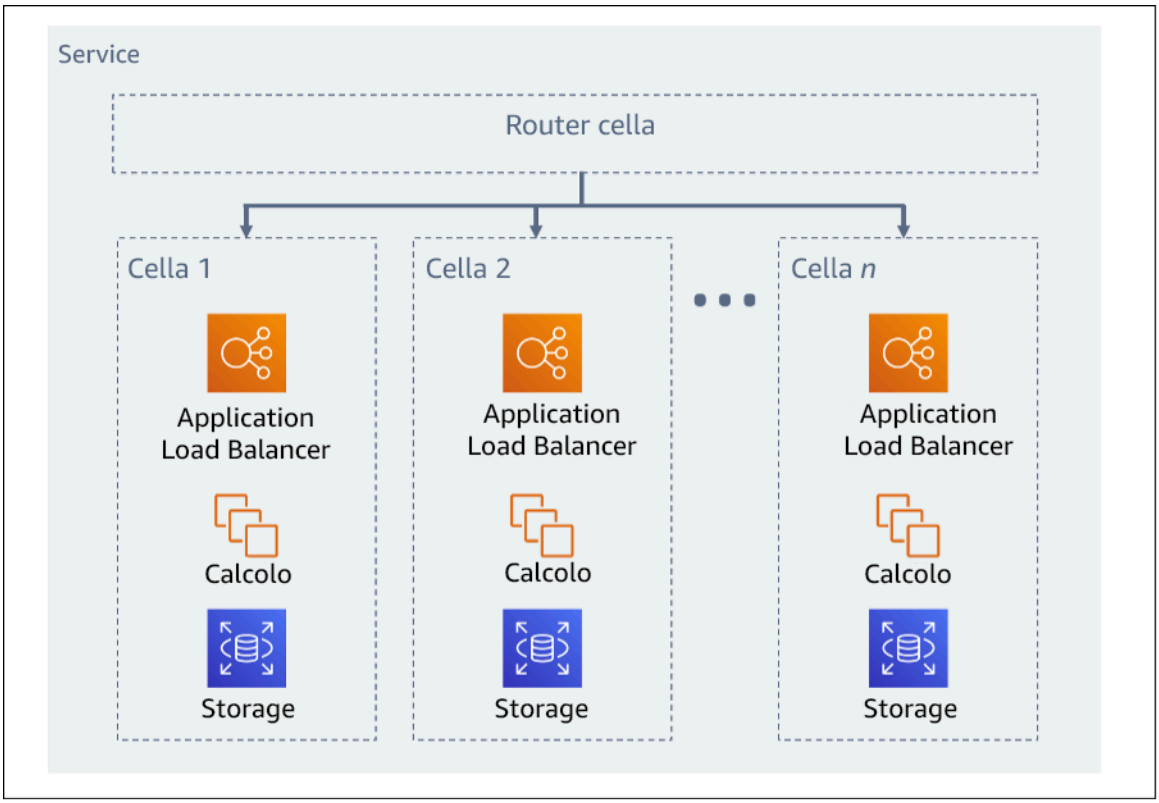


Figura 11. Architettura basata su celle

In un post del suo blog AWS, Colm MacCarthaigh spiega in che modo Amazon Route 53 utilizza il concetto di sharding casuale per isolare le richieste dei clienti negli shard. Uno shard in questo caso è costituito da due o più celle. In base alla chiave di partizione, il traffico da un cliente (o risorse o qualsiasi altra cosa desideri isolare) viene instradato allo shard assegnato. Nel caso di otto celle con due celle per shard e clienti divisi tra i quattro shard, il 25% dei clienti riscontrerebbe un impatto in caso di problema.

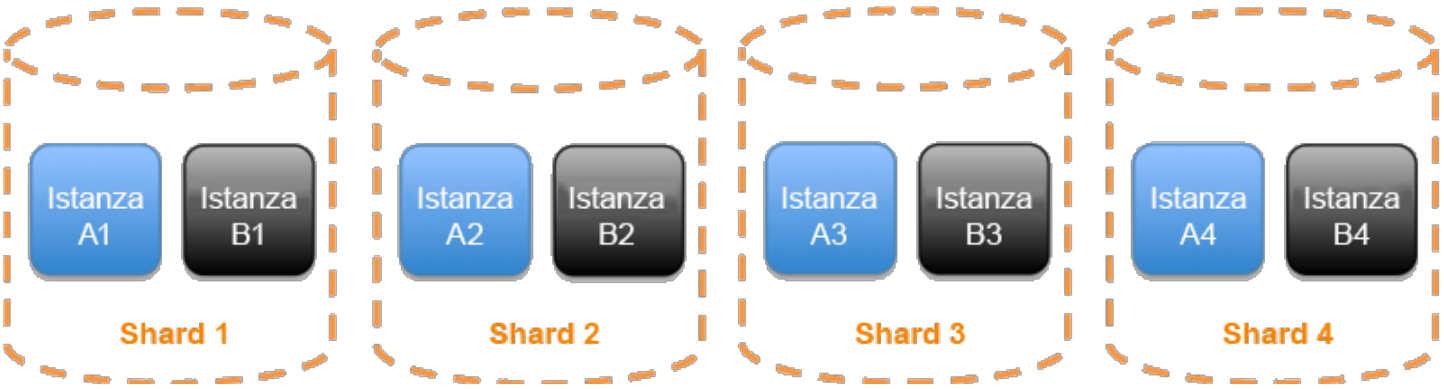


Figura 12. Servizio suddiviso in quattro shard tradizionali di due celle ciascuno

Con lo sharding casuale, puoi creare shard virtuali di due celle ciascuno e assegnare i clienti a uno di questi shard virtuali. Quando si verifica un problema, puoi comunque perdere un quarto dell'intero servizio, ma il modo in cui vengono assegnati i clienti o le risorse significa che l'ambito dell'impatto con lo sharding casuale è notevolmente inferiore al 25%. Con otto celle, ci sono 28 combinazioni univoche di due celle, il che significa che ci sono 28 possibili shard casuali (shard virtuali). Se disponi di centinaia o migliaia di clienti e assegni ogni cliente a uno shard casuale, l'impatto causato da un problema è di solo 1/28. Questo è sette volte superiore rispetto allo sharding normale.

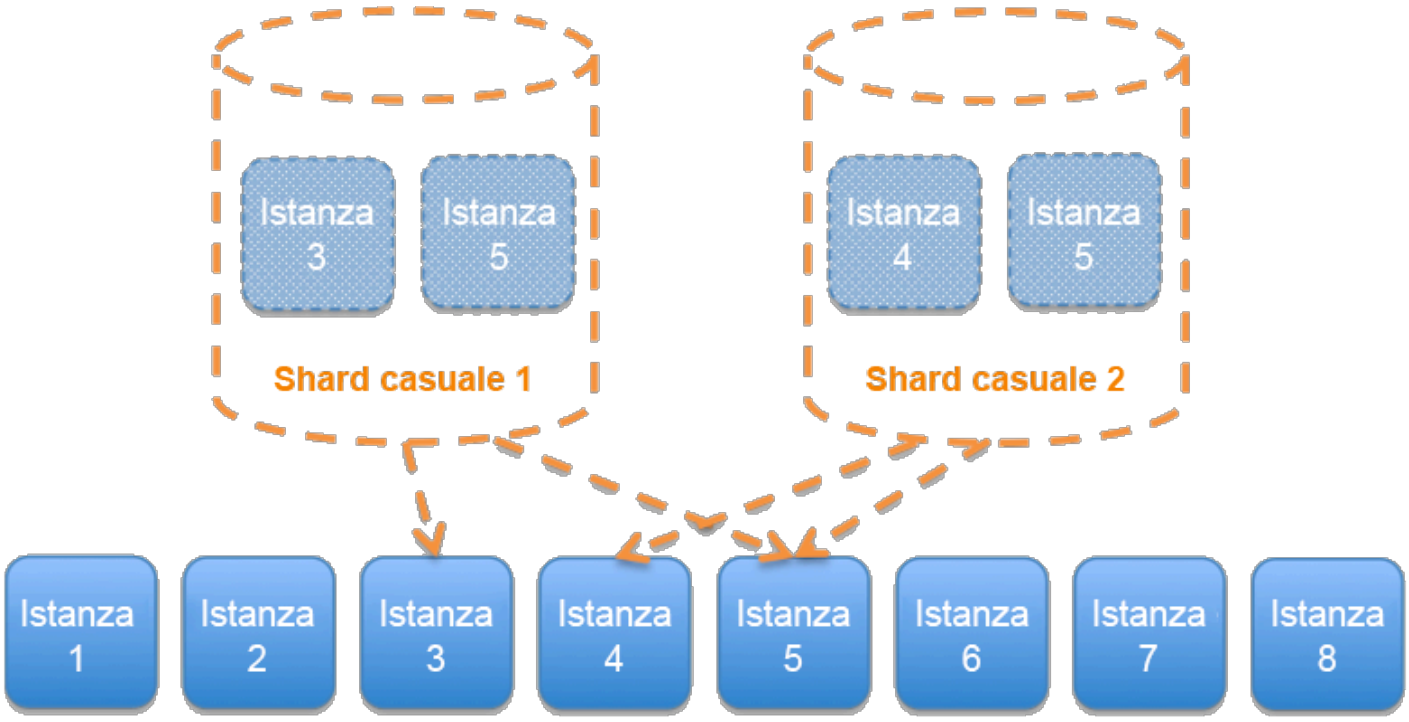


Figura 13. Servizio suddiviso in 28 shard casuali (shard virtuali) di due celle ciascuna (vengono mostrati solo due shard casuali su 28 possibili)

Uno shard può essere utilizzato per server, code o altre risorse in aggiunta alle celle.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Utilizzo di architetture paratie Come per le paratie su una nave, questo modello garantisce il contenimento di un guasto in un piccolo sottoinsieme di richieste/utenti, in modo che il numero di richieste danneggiate sia limitato e si possa comunque continuare senza errori. Le paratie per i dati sono spesso chiamate partizioni, mentre le paratie per i servizi sono note come celle.
  - Well-Architected lab: Fault isolation with shuffle sharding
  - Shuffle-sharding: AWS re:Invent 2019: Introducing The Amazon Builders' Library (Introduzione alla libreria dei costruttori di Amazon) (DOP328)
  - AWS re:Invent 2018: How AWS Minimizes the Blast Radius of Failures (Come AWS riduce al minimo il raggio di esplosione dei guasti) (ARC338)
- Valutazione dell'architettura basata su celle per il carico di lavoro In un'architettura basata su celle, ogni cella è un'istanza completa e indipendente del servizio e ha una dimensione massima fissa. Con l'aumentare del carico, i carichi di lavoro aumentano aggiungendo più celle. Una chiave di partizione viene utilizzata sul traffico in entrata per determinare quale cella elaborerà la richiesta. Qualsiasi guasto è contenuto nella singola cella in cui si verifica, in modo che il numero di richieste danneggiate sia limitato man mano che le altre celle continuano senza errori. È importante identificare la chiave di partizione corretta per ridurre al minimo le interazioni tra celle ed evitare la necessità di coinvolgere servizi di mappatura complessi in ogni richiesta. I servizi che richiedono una mappatura complessa finiscono semplicemente per spostare il problema ai servizi di mappatura, mentre i servizi che richiedono interazioni tra celle riducono l'autonomia delle celle (e quindi i presunti miglioramenti della disponibilità che ne deriverebbero).
  - Nel suo post del blog AWS, Colm MacCarthaigh spiega in che modo Amazon Route 53 utilizza il concetto di partizione casuale per isolare le richieste dei clienti nelle partizioni
    - Shuffle Sharding: Massive and Magical Fault Isolation

Risorse

Documenti correlati:

- Shuffle Sharding: Massive and Magical Fault Isolation
- The Amazon Builders' Library: Isolamento del carico di lavoro utilizzando lo sharding casuale

Video correlati:

- AWS re:Invent 2018: How AWS Minimizes the Blast Radius of Failures (Come AWS riduce al minimo il raggio di esplosione dei guasti) (ARC338)
- Shuffle-sharding: AWS re:Invent 2019: Introducing The Amazon Builders' Library (Introduzione alla libreria dei costruttori di Amazon) (DOP328)

Esempi correlati:

- Well-Architected lab: Fault isolation with shuffle sharding

REL 11 In che modo progetti il carico di lavoro affinché resista ai guasti dei componenti?

I carichi di lavoro con requisiti di disponibilità elevata e MTTR (Mean Time To Recovery) basso devono essere progettati per garantire la resilienza.

Best practice

- REL11-BP01 Monitoraggio di tutti i componenti del carico di lavoro per la rilevazione dei guasti
- REL11-BP02 Failover e passaggio a risorse integre
- REL11-BP03 Automatizzazione della riparazione a tutti i livelli
- REL11-BP04 Fare affidamento al piano dati invece che al piano di controllo durante il ripristino
- REL11-BP05 Utilizzo della stabilità statica per evitare un comportamento bimodale
- REL11-BP06 Invio di notifiche quando gli eventi influiscono sulla disponibilità

REL11-BP01 Monitoraggio di tutti i componenti del carico di lavoro per la rilevazione dei guasti

Monitora continuamente lo stato del carico di lavoro, in modo che tu e i tuoi sistemi automatizzati siate consapevoli del deterioramento o del guasto non appena questo si verifica. Monitora gli indicatori chiave di prestazioni (KPI) in base al valore aziendale.

Tutti i meccanismi di ripristino e correzione devono essere in grado di rilevare rapidamente i problemi. I guasti tecnici devono essere rilevati prima in modo che possano essere risolti. Tuttavia, la disponibilità si basa sulla capacità del carico di lavoro di fornire valore aziendale, quindi gli indicatori chiave di prestazione (KPI) che misurano questo aspetto devono far parte della strategia di rilevamento e correzione.

Anti-pattern comuni:

- Non sono stati configurati allarmi, pertanto le interruzioni si verificano senza notifica.
- Gli allarmi esistono, ma a soglie che non forniscono tempo adeguato per reagire.
- I parametri non vengono raccolti abbastanza spesso da soddisfare l'obiettivo di tempo di ripristino (RTO, recovery time objective).
- Solo il livello del carico di lavoro rivolto al cliente viene monitorato attivamente.
- Viene effettuata solo la raccolta di parametri tecnici, senza includere quelli delle funzioni aziendali.
- Non è presente alcun parametro che misuri l'esperienza utente del carico di lavoro.

Vantaggi dell'adozione di questa best practice: Eseguire un monitoraggio appropriato a tutti i livelli consente di ridurre i tempi di ripristino riducendo i tempi di rilevamento.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Determina l'intervallo di raccolta per i componenti in base agli obiettivi di ripristino.
  - L'intervallo di monitoraggio dipende dalla velocità con cui è necessario ripristinare Il tempo di ripristino dipende dal tempo necessario a ripristinare, perciò è necessario determinare la frequenza della raccolta considerando tale tempo e l'obiettivo di tempo di ripristino (RTO, recovery time objective).
- Configura il monitoraggio dettagliato per i componenti.
  - Determinare se è necessario un monitoraggio dettagliato per le istanze EC2 e l'Auto Scaling Il monitoraggio dettagliato fornisce parametri con un intervallo di 1 minuto, mentre il monitoraggio predefinito fornisce parametri con un intervallo di 5 minuti.
    - Abilitare o disabilitare il monitoraggio dettagliato della propria istanza
    - Monitoraggio di gruppi con scalabilità automatica e istanze con Amazon CloudWatch
  - Determinare se è necessario un monitoraggio avanzato per RDS Il monitoraggio avanzato utilizza un agente sulle istanze RDS per ottenere informazioni utili su diversi processi o thread in un'istanza RDS.
    - Monitoraggio avanzato
- Creazione di parametri personalizzati per misurare indicatori chiave di prestazione (KPI) aziendali I carichi di lavoro implementano funzioni aziendali chiave. Queste funzioni devono essere utilizzate come KPI che aiutano a identificare quando si verifica un problema indiretto.
  - Pubblicazione di parametri personalizzati

- Monitoraggio della presenza di errori nell'esperienza utente tramite le canary degli utenti Il test sintetico delle transazioni (noto anche come "test canary", ma da non confondere con le distribuzioni canary) in grado di eseguire e simulare il comportamento dei clienti è uno dei processi di test più importanti. Esegui questi test costantemente sugli endpoint del carico di lavoro da diverse posizioni remote.
  - Amazon CloudWatch Synthetics consente di creare i Canary dell'utente
- Creazione di parametri personalizzati che monitorino l'esperienza dell'utente Dotare l'esperienza del cliente di strumenti consente di determinare quando essa peggiora.
  - Pubblicazione di parametri personalizzati
- Imposta gli allarmi per rilevare quando una qualsiasi parte del carico di lavoro non funziona correttamente e per indicare quando effettuare l'Auto Scaling delle risorse. Gli allarmi possono essere visualizzati sui pannelli di controllo, possono essere inviati avvisi tramite Amazon SNS o e-mail e il dimensionamento automatico può essere utilizzato per aumentare o ridurre le risorse per un carico di lavoro.
  - Utilizzo degli allarmi di Amazon CloudWatch
- Crea pannelli di controllo per visualizzare i parametri. I pannelli di controllo possono essere utilizzati per visualizzare tendenze, valori anomali e altri indicatori di potenziali problemi, oppure per fornire un'indicazione dei problemi che potresti voler esaminare.
  - Utilizzo dei pannelli di controllo CloudWatch

Risorse

Documenti correlati:

- Amazon CloudWatch Synthetics consente di creare i Canary dell'utente
- Abilitare o disabilitare il monitoraggio dettagliato della propria istanza
- Monitoraggio avanzato
- Monitoraggio di gruppi con scalabilità automatica e istanze con Amazon CloudWatch
- Pubblicazione di parametri personalizzati
- Utilizzo degli allarmi di Amazon CloudWatch
- Utilizzo dei pannelli di controllo CloudWatch

Esempi correlati:

- Corso Well-Architected: Level 300: Implementing Health Checks and Managing Dependencies to Improve Reliability

REL11-BP02 Failover e passaggio a risorse integre

Garantisce che laddove si verifichi un errore con una risorsa, le risorse integre possano continuare a soddisfare le richieste. Per gli errori legati alle posizioni (ad esempio una zona di disponibilità o una Regione AWS), assicurati di disporre di sistemi che possano eseguire il failover e passare a risorse integre in posizioni non danneggiate.

I servizi AWS, come Elastic Load Balancing e AWS Auto Scaling, aiutano a distribuire il carico tra le risorse e le zone di disponibilità. Pertanto, il guasto di una singola risorsa (come un'istanza EC2) o la compromissione di una zona di disponibilità possono essere mitigati spostando il traffico sulle risorse integre rimanenti. Per i carichi di lavoro multi-regione, questa operazione è più complicata. Ad esempio, le repliche di lettura tra Regioni consentono di implementare i dati in più Regioni AWS, ma è comunque necessario promuovere la replica di lettura a primaria e indirizzare il traffico verso di essa in caso di failover. Amazon Route 53 e AWS Global Accelerator possono aiutare a instradare il traffico tra Regioni AWS.

Se il carico di lavoro utilizza servizi AWS, ad esempio Amazon S3 o Amazon DynamoDB, questi vengono automaticamente implementati in più zone di disponibilità. In caso di errore, il piano di controllo AWS instrada automaticamente il traffico verso le posizioni integre per te. I dati sono archiviati in modo ridondante in più zone di disponibilità e rimangono disponibili. Per Amazon RDS, è necessario scegliere l'opzione di configurazione Multi-AZ; quindi, in caso di errore, AWS indirizzerà automaticamente il traffico verso l'istanza integra. Per le istanze Amazon EC2, le attività Amazon ECS o i pod Amazon EKS, puoi scegliere le zone di disponibilità in cui implementarli. Elastic Load Balancing, quindi, fornisce la soluzione per rilevare le istanze nelle zone non integre e instradare il traffico verso quelle integre. Elastic Load Balancing può anche instradare il traffico verso i componenti del data center on-premise.

Per gli approcci multi-regione (che potrebbero includere anche data center on-premise), Amazon Route 53 offre un modo per definire domini Internet e assegnare policy di instradamento che possono includere controlli dell'integrità per garantire che il traffico venga instradato verso regioni integre. In alternativa, AWS Global Accelerator fornisce indirizzi IP statici che fungono da punto di ingresso fisso alla tua applicazione, quindi, instrada verso endpoint nelle Regioni AWS a tua scelta, utilizzando la rete globale AWS, anziché Internet, per migliorare le prestazioni e l'affidabilità.

AWS si avvicina alla progettazione dei servizi pensando al ripristino degli errori. Progettiamo servizi per ridurre al minimo i tempi di recupero da guasti e l'impatto sui dati. I nostri servizi utilizzano principalmente archivi di dati che riconoscono le richieste solo dopo che queste sono state archiviate in modo duraturo su più repliche in una Regione. Questi servizi e risorse includono Amazon Aurora, istanze database Multi-AZ Amazon Relational Database Service (Amazon RDS), Amazon S3, Amazon DynamoDB, Amazon Simple Queue Service (Amazon SQS) e Amazon Elastic File System (Amazon EFS). Sono costruiti con il criterio dell'isolamento basato sulle celle ed utilizzano l'isolamento dei guasti fornito dalle zone di disponibilità. Facciamo ampio uso dell'automazione nelle nostre procedure operative. Ottimizziamo anche la nostra funzionalità di sostituzione e riavvio per un ripristino rapidamente dalle interruzioni.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Failover su risorse integre. Garantisce che laddove si verifichi un errore con una risorsa, le risorse integre possano continuare a soddisfare le richieste. Per gli errori legati alle posizioni (ad esempio una zona di disponibilità o una Regione AWS), assicurati di disporre di sistemi che possano eseguire il failover e passare a risorse integre in posizioni non danneggiate.
  - Se il carico di lavoro utilizza servizi AWS, ad esempio Amazon S3 o Amazon DynamoDB, questi vengono automaticamente implementati in più zone di disponibilità. In caso di errore, il piano di controllo AWS instrada automaticamente il traffico verso le posizioni integre per te.
  - Per Amazon RDS, è necessario scegliere l'opzione di configurazione Multi-AZ; quindi, in caso di errore, AWS indirizzerà automaticamente il traffico verso l'istanza integra.
    - Alta disponibilità (Multi-AZ) per Amazon RDS
  - Per le istanze Amazon EC2 o le attività Amazon ECS, puoi scegliere le zone di disponibilità su cui effettuare la distribuzione.Elastic Load Balancing quindi rileverà le istanze in zone non integre e instraderà il traffico verso quelle integre. Elastic Load Balancing può persino instradare il traffico ai componenti nel tuo data center locale.
  - Per approcci multi-regione (che potrebbero includere anche data center in locale), assicurati che i dati e le risorse provenienti da posizioni integre possano continuare a servire le richieste
    - Ad esempio, le repliche di lettura tra Regioni consentono di implementare i dati in più Regioni AWS, ma è comunque necessario promuovere la replica di lettura per dominare e indirizzare il traffico verso di essa in caso di guasto di una posizione primaria.
      - Panoramica delle repliche di lettura Amazon RDS

- Amazon Route 53 offre un modo per definire domini Internet e assegnare policy di instradamento, che potrebbero includere controlli dell'integrità, per garantire che il traffico venga instradato verso Regioni integre. In alternativa, AWS Global Accelerator fornisce indirizzi IP statici che fungono da punto di ingresso fisso alla tua applicazione, quindi, instrada verso endpoint nelle Regioni AWS a tua scelta, utilizzando la rete globale AWS, anziché Internet, per migliorare le prestazioni e l'affidabilità.
  - Amazon Route 53: scelta di una policy di instradamento
  - Che cos'è AWSGlobal Accelerator?

Risorse

Documenti correlati:

- Partner APN: partner che possono essere d'aiuto con l'automazione della tua tolleranza ai guasti
- Marketplace AWS: prodotti utilizzabili per la tolleranza ai guasti
- AWS OpsWorks: Using Auto Healing to Replace Failed Instances (Utilizzo della riparazione automatica per sostituire le istanze in errore)
- Amazon Route 53: scelta di una policy di instradamento
- Alta disponibilità (Multi-AZ) per Amazon RDS
- Panoramica delle repliche di lettura Amazon RDS
- Strategie di posizionamento dei processi di Amazon ECS
- Creating Kubernetes Auto Scaling Groups for Multiple Availability Zones (Creazione di gruppi con scalabilità automatica Kubernetes per più zone di disponibilità)
- Che cos'è AWS Global Accelerator?

Esempi correlati:

- Corso Well-Architected: Level 300: Implementing Health Checks and Managing Dependencies to Improve Reliability

REL11-BP03 Automatizzazione della riparazione a tutti i livelli

Al rilevamento di un guasto, utilizza funzionalità automatizzate per eseguire azioni da correggere.

La capacità di riavvio è uno strumento importante per risolvere gli errori. Come illustrato in precedenza per i sistemi distribuiti, una best practice consiste nel rendere i servizi stateless laddove possibile. In questo modo si evita la perdita di dati o la disponibilità al riavvio. Nel cloud, puoi (e generalmente dovresti) sostituire l'intera risorsa (ad esempio, l'istanza EC2 o la funzione Lambda) come parte del riavvio. Il riavvio stesso è un modo semplice e affidabile per eseguire il ripristino in caso di guasto. Molti tipi diversi di guasto si verificano nei carichi di lavoro. Possono verificarsi guasti a livello di hardware, software, comunicazione e operazioni. Anziché creare nuovi meccanismi per intrappolare, identificare e correggere ciascuno dei diversi tipi di guasto, mappa diverse categorie di guasto alla stessa strategia di ripristino. Un'istanza può restituire un guasto causato da un guasto hardware, da un bug del sistema operativo, da una memory leak o da altre cause. Anziché creare una correzione personalizzata per ogni situazione, considera una di esse come un guasto dell'istanza. Termina l'istanza e consenti ad AWS Auto Scaling di sostituirla. In un secondo momento, esegui l'analisi sulla risorsa guasta fuori banda.

Un altro esempio è la possibilità di riavviare una richiesta di rete. Adotta lo stesso approccio di ripristino sia a un timeout di rete sia a un guasto di dipendenza in cui la dipendenza restituisce un guasto. Entrambi gli eventi hanno un effetto simile sul sistema, quindi piuttosto che tentare di trasformare entrambi gli eventi in un "caso speciale", adotta una strategia analoga di nuovi tentativi limitati con un back-off e un jitter esponenziali.

La capacità di riavvio è un meccanismo di ripristino presente nelle architetture di cluster ROC (Recovery Oriented Computing) e ad alta disponibilità.

Amazon EventBridge può essere utilizzato per monitorare e filtrare eventi come allarmi CloudWatch o cambiamenti di stato in altri servizi AWS. In base alle informazioni sugli eventi, può quindi attivare AWS Lambda, AWS Systems Manager Automation o altri target per eseguire una logica di riparazione sul carico di lavoro.

Amazon EC2 Auto Scaling può essere configurato per verificare lo stato dell'istanza EC2. Se l'istanza è in uno stato diverso da quello in esecuzione o se lo stato del sistema è danneggiato, Amazon EC2 Auto Scaling considera l'istanza come non integra e ne avvia una sostitutiva. Se utilizzi AWS OpsWorks, puoi configurare la riparazione automatica delle istanze EC2 a livello del layer OpsWorks.

Per le sostituzioni su larga scala (ad esempio la perdita di un'intera zona di disponibilità), anziché cercare di ottenere nuove risorse contemporaneamente è preferibile adottare la stabilità statica per trarre vantaggio dall'elevata disponibilità.

Anti-pattern comuni:

- Implementazione individuale di applicazioni in istanze/container.

- Distribuzione di applicazioni che non possono essere distribuite in più posizioni senza utilizzare il ripristino automatico.
- Riparazione manuale delle applicazioni che il dimensionamento e il ripristino automatici non sono stati in grado di riparare.

Vantaggi dell'adozione di questa best practice: Il risanamento automatico, anche se il carico di lavoro può essere distribuito in una sola posizione alla volta, ridurrà il tempo medio di ripristino e garantirà la disponibilità del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Utilizzo dei gruppi con scalabilità automatica per implementare livelli in un carico di lavoro. Auto Scaling è in grado di eseguire il risanamento automatico sulle applicazioni stateless e aggiungere e rimuovere capacità.
  - Come funziona AWS Auto Scaling
- Implementa il ripristino automatico sulle istanze EC2 che includono applicazioni distribuite non distribuibili in più posizioni e possono tollerare il riavvio in caso di guasti. Il ripristino automatico può essere utilizzato per sostituire l'hardware guasto e riavviare l'istanza quando l'applicazione non è in grado di essere distribuita in più posizioni. Vengono conservati i metadati dell'istanza e gli indirizzi IP associati, nonché i volumi Amazon EBS e i punti di montaggio su Elastic File System o file system per Lustre e Windows.
  - Ripristino automatico Amazon EC2
  - Amazon Elastic Block Store (Amazon EBS)
  - Amazon Elastic File System (Amazon EFS)
  - What is Amazon FSx for Lustre? Che cos'è Amazon FSx for Lustre?)
  - What is Amazon FSx for Windows File Server? (Che cos'è What is FSx for Windows File Server?)
    - Se utilizzi AWS OpsWorks, puoi configurare il la riparazione automatica delle istanze EC2 a livello del layer.
      - AWS OpsWorks: Using Auto Healing to Replace Failed Instances (Utilizzo della riparazione automatica per sostituire le istanze in errore)
- Implementa il ripristino automatico utilizzando AWS Step Functions e AWS Lambda quando non è possibile utilizzare il dimensionamento automatico o il ripristino automatico oppure quando

il ripristino automatico non riesce. Quando non puoi utilizzare il dimensionamento automatico né il ripristino automatico o il ripristino automatico non riesce, puoi automatizzare la riparazione utilizzando AWS Step Functions e AWS Lambda.

- What is AWS Step Functions? (Che cos'è AWS Step Functions?)
- Cos'è AWS Lambda?
  - Amazon EventBridge può essere utilizzato per monitorare e filtrare eventi come allarmi CloudWatch o cambiamenti di stato in altri servizi AWS. In base alle informazioni sugli eventi, può quindi attivare AWS Lambda (o altri target) per eseguire una logica di riparazione personalizzata sul tuo carico di lavoro.
    - Che cos'è Amazon EventBridge?
    - Utilizzo degli allarmi di Amazon CloudWatch

Risorse

Documenti correlati:

- Partner APN: partner che possono essere d'aiuto con l'automazione della tua tolleranza ai guasti
- Marketplace AWS: prodotti utilizzabili per la tolleranza ai guasti
- AWS OpsWorks: Using Auto Healing to Replace Failed Instances (Utilizzo della riparazione automatica per sostituire le istanze in errore)
- Ripristino automatico Amazon EC2
- Amazon Elastic Block Store (Amazon EBS)
- Amazon Elastic File System (Amazon EFS)
- Come funziona AWS Auto Scaling
- Utilizzo degli allarmi di Amazon CloudWatch
- Che cos'è Amazon EventBridge?
- Cos'è AWS Lambda?
- AWS Systems Manager Automation
- What is AWS Step Functions? (Che cos'è AWS Step Functions?)
- What is Amazon FSx for Lustre? Che cos'è Amazon FSx for Lustre?)
- What is Amazon FSx for Windows File Server? (Che cos'è What is FSx for Windows File Server?)

Video correlati:

- Static stability in AWS: AWS re:Invent 2019: Introducing The Amazon Builders' Library (Introduzione alla libreria dei costruttori di Amazon) (DOP328)

Esempi correlati:

- Corso Well-Architected: Level 300: Implementing Health Checks and Managing Dependencies to Improve Reliability

REL11-BP04 Fare affidamento al piano dati invece che al piano di controllo durante il ripristino

Il piano di controllo è utilizzato per configurare le risorse, mentre il piano dati fornisce servizi. I piani dati hanno tipicamente obiettivi di progettazione della disponibilità più elevati rispetto ai piani di controllo e sono solitamente meno complessi. Quando si implementano risposte di ripristino o mitigazione a eventi potenzialmente dannosi per la resilienza, l'utilizzo di operazioni sul piano di controllo può ridurre la resilienza complessiva della tua architettura. Per esempio, puoi fare affidamento sul piano dati di Amazon Route 53 per instradare in modo affidabile le query DNS basate sui controlli dell'integrità, ma l'aggiornamento delle policy di instradamento Route 53 utilizza il piano di controllo, quindi non fare affidamento su di esso per il ripristino.

I piani dati di Route 53 rispondono alle query DNS ed eseguono e valutano i controlli di integrità. Sono distribuiti a livello globale e progettati per un accordo sul livello di servizio (SLA) con disponibilità al 100%. Le API e le console di gestione di Route 53, dove si creano, aggiornano ed eliminano le risorse di Route 53, funzionano su piani di controllo progettati per privilegiare la forte coerenza e la durata necessarie per la gestione del DNS. A tal fine, i piani di controllo sono situati in un'unica regione, US East (N. Virginia). Sebbene entrambi i sistemi siano costruiti per essere molto affidabili, i piani di controllo non sono inclusi nello SLA. Possono verificarsi eventi rari in cui la progettazione resiliente del piano dati consente di mantenere la disponibilità mentre i piani di controllo non lo fanno. Per i meccanismi di ripristino di emergenza e failover, utilizzare le funzioni del piano dati per garantire la migliore affidabilità possibile.

Per ulteriori informazioni sui piani dati, sui piani di controllo e come AWS costruisce i servizi per soddisfare gli obiettivi di alta disponibilità, consulta il documento stabilità statica utilizzando le zone di disponibilità e la Libreria degli sviluppatori di Amazon.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Affidati al piano dati e non al piano di controllo quando utilizzi Amazon Route 53 per il ripristino di emergenza. Route 53 Application Recovery Controller aiuta a gestire e coordinare il failover utilizzando i controlli di disponibilità e i controlli di instradamento. Queste funzionalità monitorano continuamente la capacità dell'applicazione di riprendersi dai guasti e permettono di controllarne il ripristino su più Regioni AWS, zone di disponibilità e on-premise.
  - What is Route 53 Application Recovery Controller (What is Amazon Route 53 Application Recovery Controller?)
  - Creating Disaster Recovery Mechanisms Using Amazon Route 53 (Creazione di meccanismi di ripristino di emergenza con Amazon Route 53)
  - Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 1: Single-Region stack (Creazione di applicazioni altamente resilienti con Amazon Route 53 Application Recovery Controller, parte 1: stack a singola regione)
  - Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 2: Multi-Region stack (Creazione di applicazioni altamente resilienti con Amazon Route 53 Application Recovery Controller, parte 2: stack multi-regione)
- Capire quali operazioni sono sul piano dati e quali sul piano di controllo.
  - The Amazon Builders' Library: Avoiding overload in distributed systems by putting the smaller service in control
  - API Amazon DynamoDB (piano di controllo e piano dati)
  - AWS Lambda Executions (Esecuzioni Lambda) (suddivise in piano di controllo e piano dati)
  - AWS Lambda Executions (Esecuzioni Lambda) (suddivise in piano di controllo e piano dati)

Risorse

Documenti correlati:

- Partner APN: partner che possono essere d'aiuto con l'automazione della tua tolleranza ai guasti
- Marketplace AWS: prodotti utilizzabili per la tolleranza ai guasti
- The Amazon Builders' Library: Avoiding overload in distributed systems by putting the smaller service in control
- API Amazon DynamoDB (piano di controllo e piano dati)
- AWS Lambda Executions (Esecuzioni Lambda) (suddivise in piano di controllo e piano dati)

- Piano dati AWS Elemental MediaStore
- Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 1: Single-Region stack (Creazione di applicazioni altamente resilienti con Amazon Route 53 Application Recovery Controller, parte 1: stack a singola regione)
- Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 2: Multi-Region stack (Creazione di applicazioni altamente resilienti con Amazon Route 53 Application Recovery Controller, parte 2: stack multi-regione)
- Creating Disaster Recovery Mechanisms Using Amazon Route 53 (Creazione di meccanismi di ripristino di emergenza con Amazon Route 53)
- What is Route 53 Application Recovery Controller (What is Amazon Route 53 Application Recovery Controller?)

Esempi correlati:

- Introduzione a Amazon Route 53 Application Recovery Controller (Introduzione ad Amazon Route 53 Application Recovery Controller)

REL11-BP05 Utilizzo della stabilità statica per evitare un comportamento bimodale

Si ha un comportamento bimodale quando il carico di lavoro mostra un comportamento diverso in modalità normale e di guasto, ad esempio facendo affidamento sull'avvio di nuove istanze se una zona di disponibilità ha esito negativo. Devi invece creare carichi di lavoro che siano staticamente stabili e operino in una sola modalità. In questo caso, effettua il provisioning di istanze sufficienti in ciascuna zona di disponibilità per gestire il carico di lavoro se una zona di disponibilità è stata rimossa, quindi utilizza i controlli dello stato di Elastic Load Balancing o Amazon Route 53 per spostare il carico dalle istanze danneggiate.

La stabilità statica per la distribuzione di calcolo (ad esempio istanze EC2 o container) determinerà la massima affidabilità. Questa operazione deve essere valutata in base ai problemi relativi ai costi. Eseguire il provisioning di minore capacità di elaborazione e affidarsi all'avvio di nuove istanze in caso di guasto è meno costoso. Tuttavia, per i guasti su larga scala (ad esempio un errore nella zona di disponibilità), questo approccio è meno efficace perché si basa sulla reazione ai guasti nel momento in cui si verificano, piuttosto che prepararsi a tali problemi prima che accadano. La soluzione deve valutare l'affidabilità rispetto alle esigenze di costo per il carico di lavoro. Utilizzando più zone di disponibilità, la quantità di elaborazione aggiuntiva necessaria per la stabilità statica diminuisce.

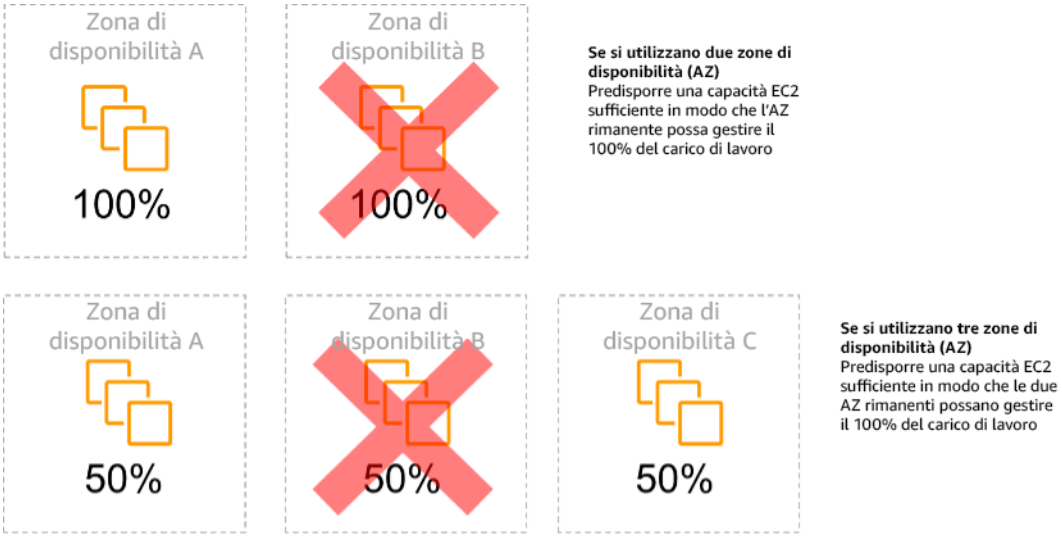


Figura 14: Stabilità statica delle istanze EC2 nelle zone di disponibilità

Dopo il trasferimento del traffico, utilizza AWS Auto Scaling per sostituire in modo asincrono le istanze dalla zona interessata dal guasto e avviarle nelle zone integre.

Un altro esempio di comportamento bimodale potrebbe essere un timeout di rete che potrebbe causare un tentativo di aggiornamento dello stato di configurazione dell'intero sistema. Ciò aggiungerebbe un carico imprevisto a un altro componente, che potrebbe quindi causare un errore, innescando altre conseguenze impreviste. Questo loop di feedback negativo influisce sulla disponibilità del tuo carico di lavoro. Al contrario, è necessario creare sistemi che siano staticamente stabili e funzionino in una sola modalità. Un progetto staticamente stabile sarebbe quello di eseguire un lavoro costante e aggiornare sempre, con cadenze fisse, lo stato di configurazione. Quando una chiamata non riesce, il carico di lavoro utilizza il valore precedentemente memorizzato nella cache e attiva un allarme.

Un altro esempio di comportamento bimodale è consentire ai client di bypassare la cache del carico di lavoro quando si verificano guasti. Potrebbe sembrare una soluzione che soddisfi le esigenze del client, ma non dovrebbe essere consentita perché modifica in modo significativo le richieste sul carico di lavoro e potrebbe causare guasti.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Utilizzo della stabilità statica per evitare un comportamento bimodale. Si ha un comportamento bimodale quando il carico di lavoro mostra un comportamento diverso in modalità normale e di

guasto, ad esempio facendo affidamento sull'avvio di nuove istanze se una zona di disponibilità ha esito negativo.

- Minimizing Dependencies in a Disaster Recovery Plan
- The Amazon Builders' Library: Stabilità statica con le zone di disponibilità
- Static stability in AWS: AWS re:Invent 2019: Introducing The Amazon Builders' Library (Introduzione alla libreria dei costruttori di Amazon) (DOP328)
    - Devi invece creare sistemi che siano staticamente stabili e operino in una sola modalità. In questo caso, effettua il provisioning di istanze sufficienti in ciascuna zona di disponibilità per gestire il carico di lavoro se una zona di disponibilità è stata rimossa, quindi utilizza i controlli dell'integrità di Elastic Load Balancing o Amazon Route 53 per spostare il carico dalle istanze danneggiate.
    - Un altro esempio di comportamento bimodale è consentire ai client di bypassare la cache del carico di lavoro quando si verificano guasti. Potrebbe sembrare una soluzione per soddisfare le esigenze del client, ma non dovrebbe essere consentita perché modifica in modo significativo le richieste sul carico di lavoro e potrebbe causare guasti.

Risorse

Documenti correlati:

- Minimizing Dependencies in a Disaster Recovery Plan
- The Amazon Builders' Library: Stabilità statica con le zone di disponibilità

Video correlati:

- Static stability in AWS: AWS re:Invent 2019: Introducing The Amazon Builders' Library (Introduzione alla libreria dei costruttori di Amazon) (DOP328)

REL11-BP06 Invio di notifiche quando gli eventi influiscono sulla disponibilità

Le notifiche vengono inviate al rilevamento di eventi significativi, anche se il problema causato dall'evento è stato risolto automaticamente.

Il ripristino automatizzato consente al tuo carico di lavoro di essere affidabile. Tuttavia, potrebbe anche oscurare problemi sottostanti che hanno bisogno di essere risolti. Implementa il monitoraggio e gli eventi appropriati in modo da poter rilevare i modelli di problemi, inclusi quelli risolti dalla

diagnostica automatica e risolvere così i problemi della causa principale. Gli allarmi di Amazon CloudWatch possono essere attivati in base ai guasti che si verificano. Possono anche attivarsi in base alle operazioni di ripristino automatizzato eseguite. Gli allarmi CloudWatch possono essere configurati per l'invio di e-mail o per la registrazione di file di log nei sistemi di monitoraggio di terze parti tramite l'integrazione con Amazon SNS.

Anti-pattern comuni:

- Invio di allarmi su cui nessuno agisce.
- Esecuzione dell'automazione del risanamento automatico, ma senza la notifica della necessità di una correzione.

Vantaggi dell'adozione di questa best practice: Le notifiche degli eventi di ripristino ti consentiranno di non ignorare i problemi che si verificano di rado.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Allarmi su indicatori chiave di prestazione aziendali al superamento di una soglia minima Un allarme su indicatori chiave di prestazione aziendali consente di sapere quando il carico di lavoro non è disponibile o non funziona.
  - Creare un allarme CloudWatch basato su una soglia statica
- Allarme su eventi che invocano l'automazione della riparazione Puoi invocare direttamente un'API SNS per inviare notifiche con qualsiasi automazione creata.
  - Che cos'è Amazon Simple Notification Service?

Risorse

Documenti correlati:

- Creare un allarme CloudWatch basato su una soglia statica
- Che cos'è Amazon EventBridge?
- Che cos'è Amazon Simple Notification Service?

REL 12 In che modo testi l'affidabilità?

Dopo aver progettato il carico di lavoro in modo da essere resiliente alle sollecitazioni della produzione, i test sono l'unico modo per garantire il funzionamento corretto e offrire la resilienza prevista.

Best practice

- REL12-BP01 Utilizzo dei playbook per analizzare gli errori
- REL12-BP02 Esecuzione di analisi post-incidente
- REL12-BP03 Test dei requisiti funzionali
- REL12-BP04 Test dei requisiti di dimensionamento e prestazioni
- REL12-BP05 Test della resilienza tramite l'utilizzo dell'ingegneria del caos
- REL12-BP06 Esecuzione regolare di giornate di gioco

REL12-BP01 Utilizzo dei playbook per analizzare gli errori

Abilita risposte coerenti e tempestive a scenari di guasto che non sono ben compresi, documentando il processo di analisi nei playbook. I playbook sono le fasi predefinite eseguite per identificare i fattori che contribuiscono a uno scenario di guasto. I risultati provenienti da un passaggio del processo vengono utilizzati per stabilire i passaggi successivi da intraprendere fino all'identificazione o alla risoluzione del problema.

Il playbook è una pianificazione proattiva che è necessario eseguire, in modo da potere intraprendere azioni reattive in modo efficace. Quando durante la produzione si verificano scenari di guasto non coperti dal playbook, risolvi innanzitutto il problema (spegni l'incendio). Quindi torna indietro e osserva le fasi intraprese per risolvere il problema e utilizzale per aggiungere una nuova voce al playbook.

Tieni presente che i playbook vengono utilizzati in risposta a specifici incidenti, mentre i runbook vengono utilizzati per ottenere esiti specifici. Spesso, i runbook vengono utilizzati per le attività di routine e i playbook vengono utilizzati per rispondere a eventi non di routine.

Anti-pattern comuni:

- Pianificare la distribuzione di un carico di lavoro senza conoscere i processi per diagnosticare i problemi o rispondere agli incidenti.
- Decisioni non pianificate sui sistemi da cui raccogliere log e parametri durante l'analisi di un evento.

- Non conservare parametri e eventi abbastanza a lungo da poter recuperare i dati.

Vantaggi dell'adozione di questa best practice: L'acquisizione di playbook garantisce l'esecuzione coerente dei processi. La codifica dei playbook limita l'introduzione di errori derivanti dall'attività manuale. L'automazione dei playbook riduce il tempo necessario per rispondere a un evento eliminando il requisito per l'intervento dei membri del team o fornendo loro informazioni aggiuntive quando inizia l'intervento.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Utilizza playbook per identificare i problemi. I playbook sono processi documentati per eseguire indagini sui problemi. Abilita risposte coerenti e tempestive agli scenari di errore documentando i processi nei playbook. I playbook devono contenere le informazioni e le istruzioni necessarie affinché una persona adeguatamente qualificata possa raccogliere le informazioni applicabili, identificare potenziali fonti di errore, isolare i guasti e stabilire i fattori che contribuiscono all'origine di un problema (eseguire l'analisi post-incidente).
  - Implementazione dei playbook come codice. Esegui le operazioni come codice mediante lo scripting dei playbook per assicurare coerenza e ridurre gli errori causati dai processi manuali. I playbook possono essere composti da più script che rappresentano le diverse fasi che potrebbero essere necessarie per identificare i fattori che contribuiscono all'origine di un problema. Le attività dei runbook possono essere attivate o eseguite nell'ambito delle attività dei playbook oppure possono richiedere l'esecuzione di un playbook in risposta agli eventi identificati.
    - Automazione dei playbook operativi con AWS Systems Manager
    - AWS Systems Manager Run Command
    - AWS Systems Manager Automation
    - Cos'è AWS Lambda?
    - Che cos'è Amazon EventBridge?
    - Utilizzo degli allarmi di Amazon CloudWatch

Risorse

Documenti correlati:

- AWS Systems Manager Automation
- AWS Systems Manager Run Command
- Automazione dei playbook operativi con AWS Systems Manager
- Utilizzo degli allarmi di Amazon CloudWatch
- Utilizzo di Canary (Amazon CloudWatch Synthetics)
- Che cos'è Amazon EventBridge?
- Cos'è AWS Lambda?

Esempi correlati:

- Automating operations with Playbooks and Runbooks (Automazione delle operazioni con Playbook e Runbook)

REL12-BP02 Esecuzione di analisi post-incidente

Esamina gli eventi che influiscono sui clienti e identifica i fattori che vi hanno contribuito e gli elementi di azione preventivi. Utilizza queste informazioni per sviluppare modi per limitare o prevenire il ripetersi degli imprevisti. Sviluppa procedure per attivare risposte rapide ed efficaci. Comunica i fattori che hanno contribuito al presentarsi dell'imprevisto e le azioni correttive secondo necessità, specificamente mirate per il pubblico di destinazione. All'occorrenza, adotta un metodo per comunicare queste cause ad altri.

Valuta perché i test esistenti non hanno individuato il problema. Aggiungi i test per questo caso se i test non esistono già.

Anti-pattern comuni:

- Individuare i fattori che hanno contribuito al verificarsi dell'incidente, ma non continuare a cercare in maniera più approfondita altri potenziali problemi e approcci da mitigare.
- Identificare le cause degli errori umani senza fornire alcuna formazione o automazione che potrebbe prevenirli.

Vantaggi dell'adozione di questa best practice: L'esecuzione di analisi post-incidente e la condivisione dei risultati consente ad altri carichi di lavoro di mitigare il rischio se hanno implementato gli stessi fattori che hanno contribuito al verificarsi dell'incidente e consente loro di implementare la mitigazione o il ripristino automatico prima che si verifichi un incidente.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Definizione di uno standard per l'analisi post-incidente. Una buona analisi post-incidente fornisce opportunità per proporre soluzioni comuni a problemi con modelli di architettura utilizzati in altri punti nei tuoi sistemi.
  - Assicurati che i fattori che hanno contribuito al verificarsi dell'incidente siano onesti e non presentino colpe.
  - Se non documenti i tuoi problemi, non puoi correggerli.
    - Assicurati che l'analisi post-incidente sia esente da colpe, in modo da poter essere obiettivo riguardo alle azioni correttive proposte e promuovere autovalutazione e collaborazione oneste nei team applicativi.
- Utilizza un processo per determinare i fattori che concorrenti. Predisponi un processo per identificare e documentare i fattori che contribuiscono al verificarsi di un evento, in modo da sviluppare azioni di mitigazione in grado di limitare o impedire il suo ripetersi e per sviluppare procedure che consentano risposte rapide ed efficaci. Comunica i fattori che hanno contribuito al verificarsi dell'incidente in maniera appropriata, specificamente mirati al pubblico di destinazione.
  - Che cos'è l'analisi dei log?

Risorse

Documenti correlati:

- Che cos'è l'analisi dei log?
- Why you should develop a correction of error (COE) (Perché sviluppare una correzione dell'errore)

REL12-BP03 Test dei requisiti funzionali

Utilizza tecniche come i test unitari e i test di integrazione per convalidare le funzionalità richieste.

Puoi ottenere i migliori risultati quando questi test vengono eseguiti automaticamente come parte delle operazioni di sviluppo e distribuzione. Ad esempio, utilizzando AWS CodePipeline, gli sviluppatori affidano le modifiche a un repository di origine in cui CodePipeline rileva automaticamente le modifiche. Queste modifiche vengono create e vengono eseguiti test. Una volta completati i test, il codice creato viene distribuito ai server temporaneo per il test. Dal server temporaneo, CodePipeline esegue più test, come quelli di integrazione o caricamento. Una volta

completati con successo i test, CodePipeline distribuisce il codice testato e approvato alle istanze di produzione.

Inoltre, l'esperienza dimostra che i test sintetici delle transazioni (noti anche come test canary, ma da non confondere con le implementazioni canary) in grado di eseguire e simulare il comportamento dei clienti sono uno dei processi di test più importanti. Esegui questi test costantemente sugli endpoint del carico di lavoro da diverse posizioni remote. Amazon CloudWatch Synthetics ti consente di creare "canary" per monitorare gli endpoint e le API.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Test dei requisiti funzionali. Includono test delle unità e test di integrazione che convalidano la funzionalità richiesta.
  - Utilizzo di CodePipeline con AWS CodeBuild per testare il codice ed eseguire compilazioni
  - AWS CodePipeline Adds Support for Unit and Custom Integration Testing with AWS CodeBuild (AWS CodePipeline aggiunge il supporto per i test di unità e integrazione personalizzati con AWS CodeBuild)
  - Distribuzione continua e integrazione continua
  - Utilizzo di Canary (Amazon CloudWatch Synthetics)
  - Automazione e test del software

Risorse

Documenti correlati:

- Partner APN: partner che possono essere d'aiuto nell'implementazione di una pipeline di integrazione continua
- AWS CodePipeline Adds Support for Unit and Custom Integration Testing with AWS CodeBuild (AWS CodePipeline aggiunge il supporto per i test di unità e integrazione personalizzati con AWS CodeBuild)
- Marketplace AWS: prodotti utilizzabili per l'integrazione continua
- Distribuzione continua e integrazione continua
- Automazione e test del software
- Utilizzo di CodePipeline con AWS CodeBuild per testare il codice ed eseguire compilazioni

- Utilizzo di Canary (Amazon CloudWatch Synthetics)

REL12-BP04 Test dei requisiti di dimensionamento e prestazioni

Utilizza tecniche come i test di carico per convalidare che il carico di lavoro soddisfi i requisiti di dimensionamento e prestazioni.

Nel cloud, puoi creare un ambiente di test su scala di produzione on demand per il tuo carico di lavoro. Se esegui questi test su un'infrastruttura ridotta, devi dimensionare i risultati osservati in base a ciò che pensi accadrà in produzione. I test di carico e prestazioni possono essere eseguiti anche in produzione se si fa attenzione a non influire sugli utenti effettivi e si contrassegna con tag i dati di test in modo da non utilizzare dati utente reali e non danneggiare le statistiche di utilizzo o i report di produzione.

Con i test, assicurati che le risorse di base, le impostazioni di dimensionamento, le quote di servizio e la progettazione di resilienza funzionino come previsto sotto carico.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

- Test dei requisiti di dimensionamento e prestazioni. Esegui test del carico per verificare che il carico di lavoro soddisfi i requisiti di dimensionamento e prestazioni.
  - Distributed Load Testing on AWS (Test di carico distribuito su AWS): simula migliaia di utenti connessi
  - Apache JMeter
    - Distribuisci la tua applicazione in un ambiente identico al tuo ambiente di produzione ed esegui un test di carico.
      - Utilizza un'infrastruttura come code concept per creare un ambiente il più simile possibile al tuo ambiente di produzione.

Risorse

Documenti correlati:

- Distributed Load Testing on AWS (Test di carico distribuito su AWS): simula migliaia di utenti connessi
- Apache JMeter

REL12-BP05 Test della resilienza tramite l'utilizzo dell'ingegneria del caos

Esegui regolarmente esperimenti di ingegneria del caos in ambienti di produzione o per quanto possibile ambienti analoghi per capire in che modo il sistema risponde a condizioni avverse.

Risultato desiderato:

La resilienza del carico di lavoro viene regolarmente verificata mediante l'applicazione dell'ingegneria del caos sotto forma di esperimenti di iniezione di errori o di inserimento di carichi imprevisti, nonché mediante il test della resilienza che convalida i comportamenti previsti noti del carico di lavoro durante un evento. Combina l'ingegneria del caos e i test della resilienza per verificare se il carico di lavoro è in grado di superare i guasti dei componenti ed eseguire il ripristino da interruzioni del servizio impreviste con un impatto minimo o nullo.

Anti-pattern comuni:

- Progettazione della resilienza, ma mancata verifica del funzionamento del carico di lavoro nel suo complesso in caso di errori.
- Mancata sperimentazione in scenari reali e con carichi previsti.
- Mancato trattamento degli esperimenti come codice o loro conservazione durante il ciclo di sviluppo.
- Mancata esecuzione degli esperimenti di ingegneria del caos sia nella pipeline CI/CD che esternamente alle implementazioni.
- Mancato utilizzo delle precedenti analisi post-incidente durante la determinazione degli errori su cui eseguire i test.

Vantaggi dell'adozione di questa best practice: l'introduzione di errori per verificare la resilienza del carico di lavoro consente di verificare che le procedure di ripristino della progettazione resiliente funzionerà se viene generato un vero e proprio errore.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

L'ingegneria del caos offre ai team la possibilità di continuare a inserire scenari di errore reali (simulazioni) in modo controllato a livello di fornitore di servizi, infrastruttura, carico di lavoro e componente con un impatto minimo o nullo per i clienti. Consente inoltre ai team di imparare dagli errori e osservare, misurare e migliorare la resilienza dei carichi di lavoro, nonché verificare l'attivazione degli avvisi e se tali avvisi vengono recapitati ai team se si verifica un evento definito.

Se applicata in modo continuativo, l'ingegneria del caos può mettere in evidenza i difetti del carico di lavoro che, se non risolti, possono avere ripercussioni negative sulla disponibilità e sulle operazioni.

> **Note**
>
> L'ingegneria del caos è la disciplina che sperimenta un sistema per creare fiducia nella capacità del sistema di affrontare condizioni turbolenti nella produzione. – Principi di ingegneria del caos

Se un sistema è in grado di sopportare queste interruzioni, l'esperimento di ingegneria del caos deve essere convertito in test automatico di regressione. In questo modo, gli esperimenti di ingegneria del caos devono essere eseguiti nell'ambito del ciclo di vita dello sviluppo dei sistemi (SDLC) e della pipeline CI/CD.

Per garantire che il carico di lavoro sia in grado di gestire un guasto del componente, esegui l'iniezione di eventi di errore reali durante l'esecuzione degli esperimenti. Ad esempio, esegui esperimenti relativi alla perdita di istanze Amazon EC2 o a eventi di failover delle istanze database Amazon RDS primario e quindi verifica che il carico di lavoro non sia stato compromesso oppure o che si stato interessato solo in minima parte. Utilizza una combinazione di errori dei componenti per simulare gli eventi che possono essere causati da un'interruzione del servizio in una zona di disponibilità.

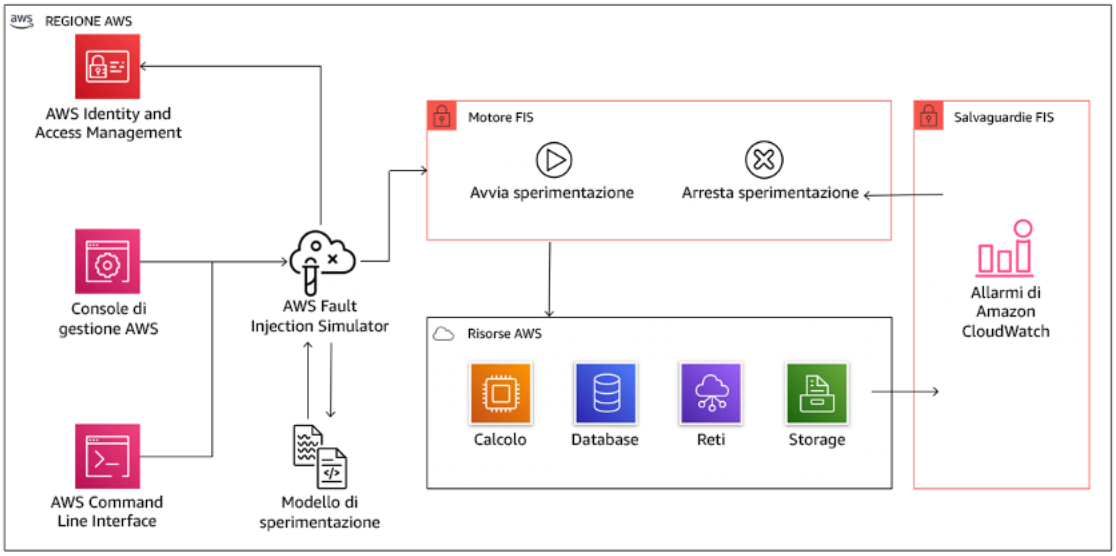
Per gli errori a livello di applicazione, ad esempio gli arresti anomali, puoi iniziare utilizzando fattori di stress, ad esempio l'esaurimento della memoria o della CPU.

Per convalidare i meccanismi di fallback o failover per le dipendenze esterne causate da interruzioni intermittenti dei servizi di rete, i componenti devono simulare tale evento bloccando l'accesso ai fornitori di terze parti per una durata specificata, che può durare da pochi secondi ad alcune ore.

Altre modalità di degrado possono causare funzionalità ridotte e risposte lente, spesso con conseguente interruzione dei servizi. Le fonti comuni di questo degrado sono una maggiore latenza nei servizi critici e una comunicazione di rete inaffidabile (pacchetti persi). Gli esperimenti basati su questi errori, inclusi gli effetti a livello di rete come latenza, messaggi eliminati ed errori DNS, possono prevedere l'incapacità di risolvere un nome, raggiungere il servizio DNS o stabilire connessioni a servizi dipendenti.

Strumenti dell'ingegneria del caos

AWS Fault Injection Service (AWS FIS) è un servizio completamente gestito per l'esecuzione di esperimenti di iniezione di errori che possono essere utilizzati come parte della pipeline di CD o al suo esterno. AWS FIS è una soluzione estremamente valida da utilizzare durante i giorni di gioco dell'ingegneria del caos. Supporta l'introduzione simultanea di errori in diversi tipi di risorse, ad esempio Amazon EC2, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS) e Amazon RDS. Questi errori includono la cessazione delle risorse, la forzatura dei failover, l'applicazione di fattori di stress a CPU o memoria, la limitazione della lunghezza di banda della rete, la latenza e la perdita di pacchetti. Poiché è integrato con gli allarmi Amazon CloudWatch, è possibile impostare condizioni di arresto come guardrail per eseguire il rollback di un esperimento se causa un impatto inatteso.



AWS Fault Injection Service è integrato con le risorse AWS per consentire l'esecuzione di esperimenti di iniezione di errori per i carichi di lavoro.

Esistono anche diverse opzioni di terze parti per gli esperimenti di iniezione di errori. Queste includono strumenti open source, ad esempio Chaos Toolkit, Chaos Meshe Litmus Chaos, nonché opzioni commerciali come Gremlin. Per ampliare l'ambito degli errori che possono essere inseriti in AWS, AWS FIS si integra con Chaos Mesh e Litmus Chaose ciò consente di coordinare i flussi di lavoro relativi all'iniezione di errori tra più strumenti. Ad esempio, puoi eseguire un test di stress sulla CPU di un pod utilizzando gli errori di Chaos Mesh o Litmus Chaos durante la cessazione di una percentuale casualmente selezionata di nodi di cluster mediante le operazioni di errore di AWS FIS.

Passaggi dell'implementazione

- Determinazione degli errori da utilizzare per gli esperimenti.

Valutazione della progettazione del carico di lavoro a livello di resilienza. Tali progettazioni, create mediante le best practice del Canone di architettura AWS) giustificano i rischi in base alle dipendenze critiche, agli eventi pregressi, alle problematiche note e ai requisiti di conformità. Elenca i singoli elementi della progettazione che devono conservare la resilienza e gli errori per mitigare i quali è stata sviluppata. Per ulteriori informazioni su questi elenchi, consulta il whitepaper relativo alla prontezza operativa , contenente linee guida su come creare un processo per impedire che si verifichino di nuovo incidenti già noti. Il processo FMEA (Failure Modes and Effects Analysis) fornisce un framework per l'esecuzione di un'analisi degli errori a livello di componente e del relativo impatto sul carico di lavoro. Il processo FMEA è descritto più in dettaglio nell'articolo di Adrian Cockcroft su modalità di errore e resilienza continua.

- Assegna una priorità a ogni errore.

  Comincia con una categorizzazione approssimativa, ad esempio alta, media o bassa. Per assegnare la priorità, considera la frequenza dell'errore e l'impatto dell'errore sul carico di lavoro nel suo complesso.

  Durante la valutazione della frequenza di un errore specifico, analizza i precedenti dati per lo stesso carico di lavoro, se disponibili. Se non sono disponibili, utilizza i dati di altri carichi di lavoro eseguiti in un ambiente simile.

  Durante la valutazione dell'impatto di un errore specifico, in genere maggiore è l'ambito dell'errore, maggiore sarà l'impatto. Considera la progettazione e lo scopo del carico di lavoro. Ad esempio, la capacità di accedere ai datastore di origine è di cruciale importanza per un carico di lavoro responsabile della trasformazione e dell'analisi dei dati. In questo caso, darai la precedenza agli esperimenti relativi agli errori di accesso, nonché a quelli con accesso limitato a livello di larghezza di banda e inserimento di latenza.

  Le analisi post-incidente rappresentano un'ottima fonte di dati per la comprensione della frequenza e dell'impatto delle modalità di errore.

  Utilizza la priorità assegnata per determinare il primo errore su cui eseguire l'esperimento e l'ordine in cui sviluppare i nuovi esperimenti di iniezione di errori.
- Per ogni esperimento eseguito, attieniti ai principi del volano dell'ingegneria del caos e della resilienza continua.

Volano dell'ingegneria del caos e della resilienza continua, che utilizza il metodo scientifico di Adrian Hornsby.

- Definisci lo stato stazionario come output misurabile di un carico di lavoro che indica un comportamento normale.

  Il carico di lavoro è associato allo stato stazionario se il suo funzionamento è affidabile e conforme a quanto previsto. Verifica pertanto che il carico di lavoro sia integro prima di definire lo stato stazionario. Lo stato stazionario non necessariamente indica l'assenza di impatto sul carico di lavoro se si verifica un errore in quanto una data percentuale di errori può rientrare nei limiti di valori accettabili. Lo stato stazionario rappresenta il punto di riferimento che verrà osservato durante l'esperimento e che metterà in evidenza le anomalie se le ipotesi definite nel passaggio successivo non sono conformi alle previsioni.

Ad esempio, lo stato stazionario di un sistema di pagamento può essere definito come elaborazione di 300 TPS con una percentuale di successo pari al 99% e un tempo di round trip pari a 500 ms.

- Definisci un'ipotesi in merito alle reazioni del carico di lavoro all'errore.

  Un'ipotesi ottimale fa riferimento al modo in cui il carico di lavoro presumibilmente è in grado di ridurre l'impatto dell'errore e salvaguardare lo stato stazionario. Nell'ipotesi è definito che, dato un errore di un tipo specifico, il sistema o il carico di lavoro rimarrà nello stato stazionario perché la progettazione del carico di lavoro ha previsto sistemi specifici di attenuazione degli errori. Il tipo di errore specifico e i sistemi di attenuazione devono essere specificati nell'ipotesi.

  Per l'ipotesi è possibile utilizzare il seguente modello, anche se è accettabile una formulazione diversa:

  > **Note**
  >
  > Se si verifica un *`errore specifico`* , il carico di lavoro *`nome del carico di lavoro`* descriverà *`i controlli di attenuazione`* per controbilanciare *`l'impatto sulle metriche aziendali o tecniche`*.

  Ad esempio:

  - In caso di arresto del 20% dei nodi nel gruppo di nodi Amazon EKS, l'API di creazione delle transazioni continua a servire il 99° percentile delle richieste in meno di 100 ms (stato stazionario). Verrà eseguito il ripristino dei nodi Amazon EKS entro cinque minuti; i pod verranno riprogrammati ed elaboreranno il traffico entro otto minuti dall'inizio dell'esperimento. Gli avvisi verranno attivati entro tre minuti.
  - Se si verifica un errore in un'istanza Amazon EC2, il controllo dell'integrità Elastic Load Balancing del sistema degli ordini farà sì che Elastic Load Balancing si limiti a inviare richieste alle rimanenti istanze integre, mentre la funzionalità Amazon EC2 Auto Scaling sostituirà l'istanza in errore, garantendo un incremento inferiore allo 0,01% degli errori (5xx) lato server (stato stazionario).
  - Se l'istanza database primario Amazon RDS restituisce un errore, il carico di lavoro della raccolta di dati della catena di approvvigionamento eseguirà il failover e si connetterà all'istanza database in standby Amazon RDS per mantenere meno di un minuto di errori di lettura o scrittura del database (stato stazionario).

- Esegui l'esperimento inserendo l'errore.

  Per impostazione predefinita, un esperimento deve essere a prova di errore e tollerato dal carico di lavoro. Se sei consapevole del fatto che il carico di lavoro avrà esito negativo, non eseguire l'esperimento. L'ingegneria del caos deve essere utilizzata per individuare scenari noti sconosciuti o scenari completamente sconosciuti. "Scenari noti sconosciuti" fanno riferimento a quegli scenari di cui sei consapevole, ma non ne comprendi completamente la natura, mentre con "scenari completamente sconosciuti" si intendono quegli scenari a te non noti e di cui non ne comprendi la natura o i motivi. L'esecuzione di esperimenti su un carico di lavoro non funzionante non può fornire nuovi approfondimenti chiarificatori. L'esperimento deve infatti essere pianificato con attenzione, essere caratterizzato da un ambito ben definito relativamente al suo impatto, nonché fornire un meccanismo di rollback applicabile in caso di esiti negativi imprevisti. Se il criterio di due diligence indica che il carico di lavoro è in grado di sostenere l'esperimento, procedi ed esegui l'esperimento. Sono disponibili varie opzioni per l'inserimento degli errori. Per i carichi di lavoro in AWS, AWS FIS fornisce numerose simulazioni di errore predefinite denominate operazioni. Puoi anche definire operazioni personalizzate eseguibili in AWS FIS utilizzando i documenti AWS Systems Manager.

  È sconsigliato l'uso di script personalizzati per gli esperimenti di ingegneria del caos, a meno che gli script non siano in grado di rilevare lo stato corrente del carico di lavoro, generare log e fornire meccanismi di rollback e condizioni di arresto, laddove possibile.

  Un framework o set di strumenti efficace che supporta l'ingegneria del caos deve tenere traccia dello stato corrente di un esperimento, generare log e fornire meccanismi di rollback a supporto dell'esecuzione controllata di un esperimento. Inizia utilizzando un servizio noto, ad esempio AWS FIS, che consente di eseguire esperimenti con ambiti e meccanismi di sicurezza ben definiti in grado di eseguire il rollback dell'esperimento in caso di esiti negativi imprevisti. Per ulteriori informazioni sull'intera gamma di esperimenti che utilizzano AWS FIS, consulta anche la sezione relativa al laboratorio relativo alle app Well-Architected resilienti con ingegneria del caos. Inoltre, AWS Resilience Hub analizzerà il carico di lavoro e creerà gli esperimenti che potrai scegliere di implementare ed eseguire in AWS FIS.

  > **Note**
  >
  > Per ogni esperimento, devi essere consapevole del suo ambito e del relativo impatto. È consigliabile eseguire la simulazione dell'errore in un ambiente non di produzione prima di eseguirla in un ambiente di produzione vero e proprio.

Gli esperimenti devono essere eseguiti in ambienti di produzione con un carico reale mediante implementazioni canary , che attivano sistemi sperimentali e di controllo, laddove possibile. L'esecuzione degli esperimenti durante gli orari non di punta è altamente consigliata al fine di ridurre al massimo potenziali eventi negativi durante la prima esecuzione dell'esperimento negli ambienti di produzione. Inoltre, se l'utilizzo dell'effettivo traffico clienti costituisce un rischio eccessivo, puoi eseguire gli esperimenti utilizzando una sintesi del traffico nell'infrastruttura di produzione utilizzando implementazioni sperimentali e di controllo. Se l'utilizzo di un ambiente di produzione non è possibile, esegui gli esperimenti in ambienti di pre-produzione il più simili possibile agli effettivi ambienti di produzione.

Devi definire e monitorare i guardrail per essere sicuro che l'esperimento non abbia un impatto sul traffico di produzione o sugli altri sistemi che superi i limiti accettabili. Definisci condizioni di arresto per interrompere l'esperimento se viene raggiunta la soglia definita nella metrica del guardrail. In tali condizioni devono essere incluse le metriche relative allo stato stazionario del carico di lavoro e le metriche riferite ai componenti in cui inserisci l'errore. Un monitor sintetico (definito anche canary utente) è una metrica che in genere deve essere inclusa come proxy utente. Le condizioni di arresto per AWS FIS sono supportate nel modello di esperimento, nella misura di un massimo di cinque condizioni di arresto per modello.

Uno dei principi dell'ingegneria del caos prevede la riduzione dell'ambito dell'esperimento e del relativo impatto.

Se da un lato deve essere prevista la possibilità di un determinato impatto negativo a breve termine, dall'altro il contenimento e la riduzione delle conseguenze negative degli esperimenti sono una responsabilità esclusiva dell'addetto all'ingegneria del caos.

Un metodo per verificare l'ambito e il potenziale impatto prevede l'esecuzione dell'esperimento dapprima in un ambiente non di produzione, la verifica che le soglie delle condizioni di arresto vengano attivate come previsto durante lo svolgimento di un esperimento e l'utilizzo effettivo delle misure di osservabilità finalizzate all'acquisizione di un'eccezione, anziché eseguire l'esperimento direttamente in produzione.

Durante l'esecuzione di esperimenti di iniezione di errori, verifica che tutte le parti responsabili ne siano a conoscenza. Comunica ai team appropriati, ad esempio i team responsabili delle operazioni, dell'affidabilità dei servizi e del supporto clienti, quando verranno eseguiti gli esperimenti e l'impatto previsto. Metti a disposizione di questi team strumenti di comunicazione che consentano loro di informare i responsabili dell'esperimento di eventuali effetti avversi.

È necessario ripristinare lo stato originario del carico di lavoro e dei relativi sistemi sottostanti. La progettazione resiliente del carico di lavoro è spesso caratterizzata da funzionalità di riparazione automatica. Tuttavia, alcune progettazioni difettose o alcuni esperimenti non riusciti possono compromettere in modo imprevisto lo stato del carico di lavoro. Entro la fine dell'esperimento dovrai essere consapevole di questa situazione e ripristinare il carico di lavoro e i sistemi.
Con AWS FIS puoi impostare una configurazione di rollback, definita anche post-operazione, all'interno dei parametri operativi. Una post-operazione ripristina una destinazione allo stato in cui si trovava prima dell'esecuzione dell'operazione stessa. Indipendentemente dal fatto che vengano eseguite in modalità automatica, ad esempio utilizzando AWS FIS, o manuale, queste post-operazioni devono essere incluse in un playbook in cui vengono descritte le procedure di rilevamento e gestione degli errori.

- Verifica l'ipotesi.

  Principi di ingegneria del caos è un documento contenente le linee guida su come verificare lo stato stazionario del carico di lavoro.

  È necessario concentrarsi sull'output misurabile di un sistema e non sugli attributi interni del sistema. Le misurazioni di tale output in un breve periodo di tempo costituiscono un'attestazione dello stato stazionario del sistema. La velocità di trasmissione effettiva del sistema nel suo complesso, le percentuali di errori e i percentili della latenza possono essere considerati metriche di interesse che rappresentano il comportamento di uno stato stazionario. Sulla base dei rilevamenti dei modelli di comportamento sistematico durante gli esperimenti, l'ingegneria del caos verifica che il sistema funzioni correttamente anziché tentare di convalidare il modo in cui funziona.

  Nei due esempi precedenti sono state incluse le metriche dello stato stazionario relative a un incremento inferiore allo 0,01% di errori (5xx) lato server e inferiore a un minuto di errori di lettura e scrittura del database.

  Gli errori 5xx rappresentano una buona metrica perché sono la conseguenza della modalità di errore che un client del carico di lavoro sperimenterà direttamente. La misurazione degli errori del database risulta valida come conseguenza diretta dell'errore, ma deve essere supportata da una misurazione diretta dell'impatto, ad esempio le richieste cliente non riuscite o gli errori restituiti a livello di client. Includi anche un monitor sintetico, definito canary utente, in qualsiasi API o URI a cui il client del carico di lavoro ha accesso diretto.
- Migliora la progettazione del carico di lavoro con un occhio di riguardo per la resilienza.

Se lo stato stazionario non è stato preservato, analizza in che modo puoi migliorare la progettazione del flusso di lavoro per azzerare l'impatto dell'errore applicando le best practice descritte nel Pilastro AWS Well-Architected relativo all'affidabilità. Ulteriori linee guida e risorse sono disponibili nella libreria di AWS Builder, dove sono contenuti articoli su come migliorare i controlli dell'integrità oppure impiegare nuovi tentativi con backoff nel codice dell'applicazione.

Dopo aver implementato queste modifiche, esegui di nuovo l'esperimento (rappresentato dalla linea punteggiata nel volano relativo all'ingegneria del caos) per determinare la relativa efficacia. Se nella fase di verifica risulta che l'ipotesi è vera, il carico di lavoro sarà in stato stazionario e il ciclo continuerà.

- Esegui gli esperimenti con regolarità.

  Un esperimento di ingegneria del caos è un ciclo e gli esperimenti devono essere eseguiti regolarmente nell'ambito dell'ingegneria del caos. Se un carico di lavoro è conforme all'ipotesi dell'esperimento, l'esperimento deve essere automatizzato affinché venga eseguito continuamente come fase di regressione della pipeline CI/CD. Per ulteriori informazioni in merito, consulta questo blog relativamente alle procedure di esecuzione degli esperimenti AWS FIS utilizzando AWS CodePipeline. Questo laboratorio relativo a esperimenti AWS FIS ricorrenti in una pipeline CI/CD ti consente di fare esperienza pratica.

  Gli esperimenti di iniezione di errori fanno inoltre parte delle giornate di gioco (consulta REL12-BP06 Esecuzione regolare di giornate di gioco). Le giornate di gioco simulano un errore o un evento per verificare sistemi, processi e risposte dei team. Lo scopo è di eseguire effettivamente le azioni che compirebbe il team come se si verificasse un evento eccezionale.

- Acquisisci e archivia i risultati degli esperimenti.

  I risultati degli esperimenti di iniezione di errori devono essere acquisiti e resi persistenti. Includi tutti i dati necessari, ad esempio orari, carico di lavoro e condizioni, in modo da essere in grado di analizzare i risultati e i trend in un secondo momento. I risultati potrebbero includere, ad esempio, screenshot dei pannelli di controllo, dump in formato CSV del database delle metriche oppure appunti scritti a mano relativi a eventi e osservazioni associati all'esperimento. La registrazione degli esperimenti mediante AWS FIS può rientrare nel processo di acquisizione dei dati.

Risorse

Best practice correlate:

- REL08-BP03 Esecuzione di test di resilienza come parte integrante dell'implementazione
- REL13-BP03 Esecuzione di test sull'implementazione del ripristino di emergenza per convalidare l'implementazione

Documenti correlati:

- What is AWS Fault Injection Service? (Che cos'è AWS Fault Injection Service?)
- What is AWS Resilience Hub? (Che cos'è AWS Resilience Hub?)
- Principi di ingegneria del caos
- Chaos Engineering: Planning your first experiment (Ingegneria del caos: pianificazione del primo esperimento)
- Resilience Engineering: Learning to Embrace Failure
- Chaos Engineering stories (Storie relative all'ingegneria del caso)
- Evitare fallback nei sistemi distribuiti
- Canary Deployment for Chaos Experiments (Implementazione canary per gli esperimenti di ingegneria del caos)

Video correlati:

- AWS re:Invent 2020: Testing resiliency using chaos engineering (ARC316) (Esecuzione di test di resilienza mediante l'ingegneria del caos [ARC316])
- AWS re:Invent 2019: migliorare la resilienza con l'ingegneria del caos (DOP309-R1)
- AWS re:Invent 2019: Performing chaos engineering in a serverless world (CMY301) (Esecuzione dell'ingegneria del caos in uno scenario serverless [CMY301])

Esempi correlati:

- Well-Architected lab: Level 300: Testing for Resiliency of Amazon EC2, Amazon RDS, and Amazon S3 (Test della resilienza di Amazon EC2, Amazon RDS e Amazon S3)
- Chaos Engineering on AWS lab (Laboratorio relativo all'ingegneria del caos in AWS)
- Resilient and Well-Architected Apps with Chaos Engineering lab (Laboratorio relativo alle app Well-Architected resilienti con ingegneria del caos)
- Serverless Chaos lab (Laboratorio relativi a esperimenti di ingegneria del caos per architetture serverless)

- Measure and Improve Your Application Resilience with AWS Resilience Hub lab (Laboratorio di misurazione e ottimizzazione della resilienza dell'applicazione con AWS Resilience Hub)

Strumenti correlati:

- AWS Fault Injection Service
- Marketplace AWS: Gremlin Chaos Engineering Platform (Piattaforma di ingegneria del caos di Gremlin)
- Chaos Toolkit
- Chaos Mesh
- Litmus

REL12-BP06 Esecuzione regolare di giornate di gioco

Utilizza le giornate di gioco per provare regolarmente le procedure per rispondere a eventi ed errori nel modo più vicino possibile alla produzione (anche negli ambienti di produzione) con le persone che si occuperanno di eventuali scenari di errore reali. Le giornate di gioco applicano misure per garantire che gli eventi di produzione non influiscano sugli utenti.

Le giornate di gioco simulano un errore o un evento per testare sistemi, processi e risposte dei team. Lo scopo è di eseguire effettivamente le azioni che compirebbe il team come se si verificasse un evento eccezionale. Questi ti aiuta a capire dove puoi apportare dei miglioramenti e ti può aiutare a sviluppare un'esperienza organizzativa nella gestione degli eventi. Tali azioni devono essere svolte regolarmente in modo che il team costruisca una memoria muscolare su come rispondere.

Quando la progettazione per la resilienza è in loco ed è stata testata in ambienti non di produzione, un game day è il modo per garantire che tutto funzioni come pianificato in produzione. Una giornata di gioco, soprattutto la prima, è un'attività di duro lavoro per tutti, in cui tutti gli ingegneri e i team operativi vengono informati in merito a quando accadrà e cosa accadrà. I runbook sono in loco. Gli eventi simulati, compresi i possibili eventi di guasto, vengono eseguiti nei sistemi di produzione nel modo prescritto e ne viene valutato l'impatto. Se tutti i sistemi funzionano come progettato, il rilevamento e la correzione automatica avvengono con un impatto minimo o nullo. Tuttavia, se si osserva un impatto negativo, viene eseguito il rollback del test e i problemi relativi al carico di lavoro vengono risolti, se necessario manualmente (utilizzando il runbook). Poiché le giornate di gioco hanno spesso luogo in produzione, è necessario prendere tutte le precauzioni per garantire che non vi sia alcun impatto sulla disponibilità per i clienti.

Anti-pattern comuni:

- Documentare le procedure senza mai esercitarle.
- Non includere i responsabili delle decisioni aziendali negli esercizi di test.

Vantaggi dell'adozione di questa best practice: Eseguire giornate di gioco garantisce che tutto il personale segua le policy e le procedure quando si verifica un incidente reale e convalida che tali policy e procedure siano appropriate.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

- Programma giornate di gioco per provare regolarmente i tuoi runbook e playbook. Le giornate di gioco devono coinvolgere tutte le persone implicate in un evento di produzione: proprietari di aziende, personale addetto allo sviluppo, personale operativo e team di risposta agli incidenti.
  - Esegui i test di carico o delle prestazioni e successivamente esegui l'iniezione degli errori.
  - Ricerca anomalie nei tuoi runbook e opportunità di provare i tuoi playbook.
    - In caso di deviazione dai tuoi runbook, perfeziona il runbook o correggi il comportamento. Se ti eserciti sul tuo playbook, identifica il runbook che avrebbe dovuto essere usato, oppure creane uno nuovo.

Risorse

Documenti correlati:

- Che cos'è AWS GameDay?

Video correlati:

- AWS re:Invent 2019: migliorare la resilienza con l'ingegneria del caos (DOP309-R1)

Esempi correlati:

- AWS Well-Architected Labs – Test di resilienza

REL 13 Come pianifichi il disaster recovery (DR)?

Avere backup e componenti del carico di lavoro ridondanti in loco è l'inizio della strategia di disaster recovery. RTO e RPO sono i tuoi obiettivi per il ripristino del carico di lavoro. Imposta questi valori in base alle esigenze aziendali. Implementa una strategia per raggiungere questi obiettivi, prendendo in considerazione le posizioni e la funzione delle risorse e dei dati del carico di lavoro. La probabilità di interruzione e il costo del ripristino sono fattori chiave che aiutano a comunicare il valore aziendale che può avere il ripristino di emergenza per un carico di lavoro.

Best practice

- REL13-BP01 Definizione degli obiettivi di ripristino in caso di downtime e perdita di dati
- REL13-BP02 Utilizzo di strategie di ripristino definite per conseguire gli obiettivi di ripristino
- REL13-BP03 Esecuzione di test sull'implementazione del ripristino di emergenza per convalidare l'implementazione
- REL13-BP04 Gestione della deviazione di configurazione nel sito o nella Regione del ripristino di emergenza
- REL13-BP05 Automatizzazione del ripristino

REL13-BP01 Definizione degli obiettivi di ripristino in caso di downtime e perdita di dati

Il carico di lavoro ha un Recovery Time Objective (RTO) e Recovery Point Objective (RPO).

Il Recovery Time Objective (RTO) è il ritardo massimo accettabile tra l'interruzione del servizio e il suo ripristino. Questo determina ciò che viene considerato un intervallo di tempo accettabile quando il servizio non è disponibile.

Recovery Point Objective (RPO) è il periodo di tempo massimo accettabile dall'ultimo punto di ripristino dei dati. Questo determina ciò che viene considerato una perdita di dati accettabile tra l'ultimo punto di ripristino e l'interruzione del servizio.

RTO e RPO sono valori importanti quando si seleziona una strategia adeguata di ripristino di emergenza per il proprio carico di lavoro. Tali obiettivi sono stabiliti dall'azienda e poi vengono utilizzati dai team tecnici per selezionare e implementare una strategia di ripristino di emergenza.

Risultato desiderato:

Ogni carico di lavoro ha un RTO e un RPO assegnati, definiti in base all'impatto aziendale. Il carico di lavoro viene assegnato a un livello predefinito, che stabilisce la disponibilità del servizio e la perdita

accettabile di dati, con un RTO e un RPO associati. Se tale livello non è raggiungibile, è possibile assegnare un livello personalizzato per carico di lavoro, con l'obiettivo di creare i livelli in un secondo momento. RTO e RPO sono valori fondamentali per la selezione di una strategia di ripristino di emergenza da implementare per il carico di lavoro. Altre riflessioni nel momento della scelta di una strategia di ripristino di emergenza sono i vincoli economici, le dipendenze del carico di lavoro e i requisiti operativi.

Per l'RTO è necessario comprendere l'impatto in base alla durata di un'interruzione. È lineare o ci sono implicazioni non lineari? (Ad esempio, dopo 4 ore, chiudi una linea di produzione fino l'inizio del turno successivo).

Una matrice di ripristino di emergenza, come quella seguente, può aiutarti a capire come la criticità del carico di lavoro sia collegata agli obiettivi di ripristino. (Da notare che i valori reali per gli assi X e Y devono essere personalizzati in base alle esigenze della tua organizzazione).

Matrice di ripristino di emergenza						
		Obiettivo del punto di ripristino				
		meno di 1 minuto	meno di 1 ora	meno di 6 ore	meno di 1 giorno	Più di 1 giorno
Obiettivo del tempo di ripristino	meno di 10 minuti	Critica	Critica	Alta	Medio	Medio
	meno di 2 ore	Critica	Alta	Medio	Medio	Bassa
	meno di 8 ore	Alta	Medio	Medio	Bassa	Bassa
	meno di 24 ore	Medio	Medio	Bassa	Bassa	Bassa
	Più di 24 ore	Medio	Bassa	Bassa	Bassa	Bassa

Figura 16: Matrice di ripristino di emergenza

Anti-pattern comuni:

- Nessun obiettivo di ripristino definito.
- Selezione di obiettivi di ripristino arbitrari.
- Selezione di obiettivi di ripristino troppo tolleranti e che non soddisfano gli obiettivi di business.
- Mancanza di comprensione dell'impatto dei tempi di inattività e perdita dei dati.
- Selezione di obiettivi di ripristino non realistici, come tempo zero di ripristino e nessuna perdita di dati, che potrebbero non essere raggiungibili per la configurazione del tuo carico di lavoro.

- Selezione di obiettivi di ripristino più severi rispetto agli obiettivi aziendali effettivi. Questo costringe a effettuare implementazioni di ripristino di emergenza più costose e complicate rispetto alle esigenze del carico di lavoro.
- Selezione di obiettivi di ripristino non compatibili con quelli di un carico di lavoro dipendente.
- I tuoi obiettivi di ripristino non considerano i requisiti di conformità normativa.
- RTO e RPO definiti per un carico di lavoro, ma mai testati.

Vantaggi dell'adozione di questa best practice: Gli obiettivi di ripristino in termini di tempo e perdita di dati sono necessari per guidare l'implementazione del disaster recovery.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Per un dato carico di lavoro devi considerare l'impatto dei tempi di inattività e della perdita dei dati per la tua azienda. L'impatto generalmente aumenta all'aumentare dei tempi di inattività o della perdita dei dati, ma il ritmo di tale crescita cambia in base al tipo di carico di lavoro. Ad esempio, potresti tollerare l'inattività per massimo un'ora con conseguenze minime, ma successivamente l'impatto diventerebbe rapidamente più serio. L'impatto sull'azienda si manifesta in forme diverse, tra cui costi economici (come perdita di fatturato), fiducia del cliente (e impatto sulla reputazione), problematiche operative (come stipendi in ritardo o diminuzione della produttività) e rischi normativi. Usa i passaggi seguenti per comprendere questi aspetti e impostare i valori RTO e RPO per il tuo carico di lavoro.

Passaggi dell'implementazione

1. Individua gli stakeholder aziendali per questo carico di lavoro e collabora con loro per implementare questi passaggi. Gli obiettivi di ripristino di un carico di lavoro sono il frutto di una decisione aziendale. I team tecnici, quindi, lavorano con gli stakeholder aziendali e usano questi obiettivi per selezionare una strategia di ripristino di emergenza.

   > **Note**
   >
   > Per i passaggi 2 e 3 puoi usare the section called “Foglio di lavoro di implementazione”.

2. Raccogli le informazioni necessarie per prendere una decisione rispondendo alle domande qui di seguito.
3. Hai categorie o livelli di criticità in termini di impatto del tuo carico di lavoro nella tua organizzazione?

a. Se sì, assegna questo carico di lavoro a una categoria

b. Se no, definisci queste categorie. Crea al massimo cinque categorie e perfeziona l'intervallo del tuo Obiettivo del tempo di ripristino (RTO) per ognuna. Ecco alcune categorie di esempio: critico, alto, medio, basso. Per capire come mappare i carichi di lavoro rispetto alle categorie devi considerare se il carico di lavoro è mission-critical, importante per l'azienda o non trainante.

c. Imposta i valori RTO e RPO del carico di lavoro in base alla categoria. Scegli sempre una categoria più severa (RTO e RPO inferiori) rispetto ai valori grezzi calcolati in questa fase. Se ciò comporta una variazione significativa di valore non rispondente alle esigenze, prendi in considerazione la possibilità di creare una nuova categoria.

4. In base alle risposte assegna i valori RTO e RPO al carico di lavoro. Puoi farlo direttamente o assegnando il carico di lavoro a un livello predefinito di servizio.
5. Crea un documento con il piano di ripristino di emergenza (DRP) per questo carico di lavoro, che sarà parte del piano di continuità aziendale della tua organizzazione (BCP), in un punto accessibile al team del carico di lavoro e agli stakeholder.

   a. Registra i valori RTO e RPO e le informazioni usate per definire questi valori. Includi la strategia utilizzata per valutare l'impatto del carico di lavoro sull'azienda.

   b. Registra altre metriche, oltre ai valori RTO e RPO che stai monitorando o che pensi di monitorare per gli obiettivi di ripristino di emergenza.

   c. Dopo aver creato questi valori, potrai aggiungere i dettagli della tua strategia di ripristino di emergenza e il runbook.

6. Osservando le criticità del carico di lavoro in una matrice come quella della Figura 15, puoi iniziare a stabilire livelli predefiniti di servizio per la tua organizzazione.
7. Dopo aver implementato una strategia di ripristino di emergenza (o un proof of concept per una strategia di ripristino di emergenza) secondo quanto stabilito da the section called “REL13-BP02 Utilizzo di strategie di ripristino definite per conseguire gli obiettivi di ripristino”, testa questa strategia per stabilire i valori reali di RTC (Recovery Time Capability) e di RPC (Recovery Point Capability) del carico di lavoro. Se questi valori non sono in linea con gli obiettivi target di ripristino, puoi collaborare con gli stakeholder della tua azienda per modificarli o cambiare la strategia di ripristino di emergenza in modo che possa soddisfare tali obiettivi.

Domande principali

1. Qual è il tempo massimo durante il quale il carico di lavoro può essere inattivo prima che questo abbia un impatto grave sull'attività?

a. Definisci il costo monetario (impatto finanziario diretto) sull'attività al minuto se il carico di lavoro è inattivo.

b. Considera che l'impatto non è sempre lineare. L'impatto può essere limitato all'inizio e poi aumentare rapidamente oltre un punto critico specifico.

2. Qual è la quantità massima di dati che possiamo perdere prima che questo abbia un impatto grave sull'attività?

a. Considera questo valore per gli archivi di dati più strategici. identifica le criticità relative ad altri archivi di dati.

b. I dati del carico di lavoro possono essere ricreati se persi? Se questo è operativamente più facile rispetto al backup e al ripristino, scegli il valore RPO in base alla criticità dei dati di origine utilizzati per ricreare i dati del carico di lavoro.

3. Quali sono gli obiettivi di ripristino e le aspettative di disponibilità dei carichi di lavoro da cui questo valore dipende (downstream) o i carichi di lavoro che dipendono da questo valore (upstream)?

a. Scegli obiettivi di ripristino che consentono a questo carico di lavoro di soddisfare i requisiti delle dipendenze upstream.

b. Scegli obiettivi di ripristino che sono raggiungibili considerate le funzionalità di ripristino delle dipendenze downstream. Possono essere escluse le dipendenze downstream non critiche (quelle che puoi "aggirare"). In alternativa, lavora con dipendenze downstream critiche per migliorare le funzionalità di ripristino, laddove necessario.

Domande aggiuntive

Considera queste domande e come possono essere applicate a questo carico di lavoro:

4. Hai RTO e RPO diversi a seconda del tipo di interruzione (Regione rispetto ad AZ e così via)?
5. Esiste un periodo specifico (stagionalità, eventi commerciali, lanci di prodotto) in cui RTO/RPO possono cambiare? Se sì, qual è la misurazione diversa e il vincolo temporale?
6. Se il carico di lavoro viene perturbato, quanti clienti ne subiranno l'impatto?
7. Qual è l'impatto sulla reputazione se il carico di lavoro è perturbato?
8. Quali altri impatti operativi possono verificarsi se il carico di lavoro subisce perturbazioni? Ad esempio, l'impatto sulla produttività dei dipendenti se i sistemi e-mail non sono disponibili o sei i sistemi di buste paga non sono in grado di inviare le transazioni.
9. In che modo il carico di lavoro e i valori RTO e RPO si allineano alla linea di business e alla strategia di ripristino di emergenza dell'organizzazione?

10. Esistono obblighi contrattuali interni per fornire un servizio? Esistono delle penali nel caso in cui non siano soddisfatti?
11. Quali sono i limiti normativi o di conformità dei dati?

Foglio di lavoro di implementazione

Puoi usare questo foglio di lavoro per le fasi 2 e 3 dell'implementazione. Adegua questo foglio di lavoro in base alle tue esigenze specifiche, aggiungendo, ad esempio, altre domande.

Passo 2: domande principali	Si applica al carico di lavoro?	RTO del carico di lavoro	RPO del carico di lavoro	RTO rettif.	RPO rettif.	Istruzioni
[1] tempo massimo di inattività del carico di lavoro						misurato in tempo dall'inizio del malfunzionamento al ripristino
[2] quantità massima di dati che possono essere persi						misurato in tempo trascorso dall'ultimo set di dati integro ripristinabile
[3a] dipendenze a monte						inserire gli obiettivi di recupero a monte più rigorosi
[3b] riconciliazione delle dipendenze a valle						inserire gli obiettivi di recupero a valle meno rigorosi
[3a] riconciliazione delle dipendenze a monte						Se il valore a monte è inferiore ai valori attuali e il valore a valle è superiore,
[3b] riconciliazione delle dipendenze a valle						operare sulle dipendenze per riconciliare i valori e inserirli qui.
[3] dipendenze						ridurre i valori per soddisfare le dipendenze a monte o alzarli in base alle capacità delle dipendenza a valle
Passo 2: domande aggiuntive						Indicare se la domanda è pertinente. Saltarla in caso affermativo
RTO/RPO di base						Riportare qui i valori di RTO e RPO sopra indicati
[4] tipo di malfunzionamento	[ ]S / [ ]N					Inserire gli obiettivi di recupero per i tipi di evento con i requisiti più rigorosi
[5] obiettivi specifici basati sul tempo	[ ]S / [ ]N					Inserire gli obiettivi di recupero per i tempi con i requisiti più rigorosi
[6] clienti che sperimentano il disservizio	[ ]S / [ ]N					Tracciare un grafico dei clienti che sperimentano il disservizio in funzione del tempo di inattività o dei dati persi. Utilizzare tale grafico per inserire i valori massimi di RTO e RPO ammissibili in base all'impatto sui clienti
[7] impatto reputazionale	[ ]S / [ ]N					Lavorare in modo congiunto con l'azienda per determinare i massimi valori di RTO e RPO in base all'impatto sulla reputazione
[8] impatto operativo	[ ]S / [ ]N					Inserire i valori massimi di RTO e RPO sulla base dell'impatto operativo
[9] allineamento aziendale	[ ]S / [ ]N					Inserire i valori massimi di RTO e RPO per i carichi di lavoro di questo tipo in base ai requisiti LOB e organizzativi
[10] obblighi contrattuali	[ ]S / [ ]N					Inserire i valori massimi di RTO e RPO sulla base degli obblighi contrattuali
[11] conformità normativa	[ ]S / [ ]N					Inserire i valori massimi di RTO e RPO sulla base delle norme di conformità applicabili
obiettivo sulla base delle domande aggiuntive						Selezionare il valore minimo (valore più rigoroso) dalle domande 4-11 e inserirlo qui
obiettivo rettificato						Se non è possibile raggiungere gli obiettivi indicati nella riga precedente, collaborare con le parti interessate per allentare i vincoli e inserire un nuovo minimo qui.
RTO/RPO rettificato						Inserire il valore inferiore tra RPO/RTO di base e valore obiettivo rettificato
Passo 3						
Mappatura su categorie o livelli predefiniti						Regolare entrambi i valori verso il basso (requisito più rigoroso) per allinearsi al livello più vicino definito

Foglio di lavoro

Livello di impegno per il piano di implementazione: Bassa

Risorse

Best practice correlate:

- the section called “REL09-BP04 Ripristino periodico dei dati per verificare l'integrità e i processi di backup:”
- the section called “REL13-BP02 Utilizzo di strategie di ripristino definite per conseguire gli obiettivi di ripristino”

- the section called “REL13-BP03 Esecuzione di test sull'implementazione del ripristino di emergenza per convalidare l'implementazione”

Documenti correlati:

- AWS Architecture Blog: Disaster Recovery Series
- Ripristino di emergenza dei carichi di lavoro su AWS: ripristino nel cloud (whitepaper di AWS)
- Gestire le policy di resilienza con AWS Resilience Hub
- Partner APN: partner che possono assistere con disaster recovery
- Marketplace AWS: prodotti utilizzabili per il disaster recovery

Video correlati:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Modelli architetturali per applicazioni attive-attive su più Regioni) (ARC209-R2)
- Ripristino di emergenza di carichi di lavoro su AWS

REL13-BP02 Utilizzo di strategie di ripristino definite per conseguire gli obiettivi di ripristino

Definisci una strategia di ripristino di emergenza (DR) che soddisfi gli obiettivi di ripristino del carico di lavoro. Scegli una strategia come: backup e ripristino, standby (attivo/passivo) o attivo/attivo.

Una strategia di ripristino di emergenza si basa sulla capacità di creare il tuo carico di lavoro in un sito di ripristino se la tua sede principale non è disponibile per eseguire il carico di lavoro. Gli obiettivi di ripristino più comuni sono RTO e RPO, come discusso in REL13-BP01 Definizione degli obiettivi di ripristino in caso di downtime e perdita di dati.

Una strategia di ripristino di emergenza (DR) su più zone di disponibilità (AZ) all'interno di un singolo Regione AWS può offrire la mitigazione rispetto a eventi disastrosi come incendi, alluvioni e interruzioni gravi dell'energia. Se è un requisito implementare una protezione rispetto a un evento improbabile che impedisca al tuo carico di lavoro di poter essere eseguito in un determinato Regione AWS, puoi usare una strategia di ripristino di emergenza basata su più regioni.

Quando pianifichi una strategia di ripristino di emergenza su più regioni, devi scegliere una delle seguenti strategie. Sono elencate in ordine crescente di complessità e di costi e in ordine decrescente di RTO e RPO. La regione di ripristino si riferisce a una Regione AWS diversa da quella principale utilizzata per il tuo carico di lavoro.

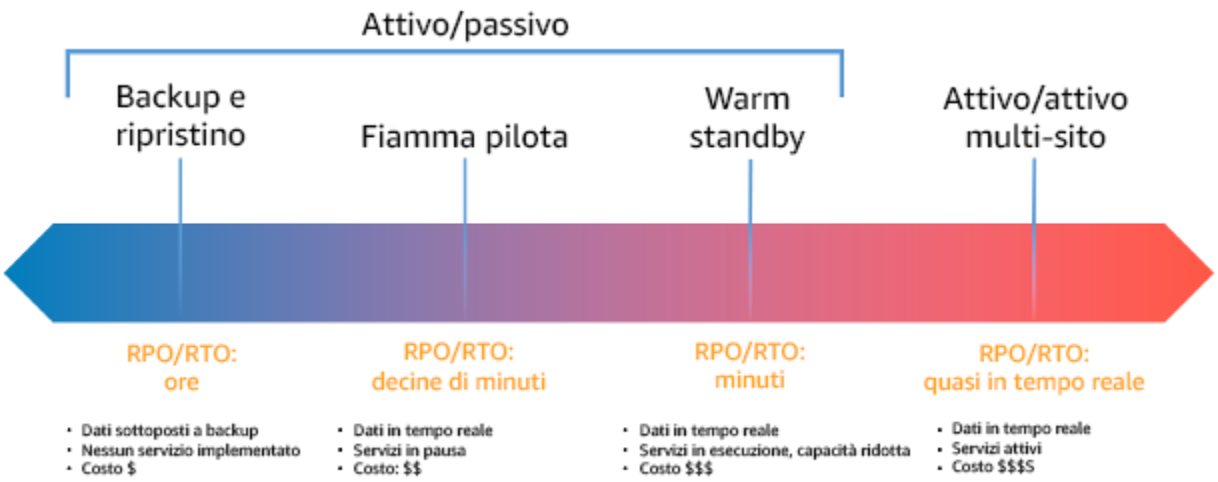


Figura 17: Strategie di ripristino di emergenza (DR)

- Backup e ripristino (RPO in poche ore, RTO in 24 ore o meno): esegui il backup dei dati e delle applicazioni nella regione di ripristino. Adottando backup continui o automatizzati otterrai un ripristino point-in-ime che può ridurre il valore dell'RPO fino a raggiungere in alcuni casi 5 minuti. Nel caso in cui si verifichi un disastro, distribuirai l'infrastruttura (usando l'infrastruttura come codice per ridurre l'RTO), distribuirai il codice e ripristinerai i dati del backup dopo un disastro nella regione di ripristino.
- Pilot light (RPO in minuti, RTO in decine di minuti): fornisci una copia dell'infrastruttura del carico di lavoro di base nella regione di ripristino. Replica i dati nella regione di ripristino e crea un backup in essa. Le risorse necessarie per supportare la replica dei dati e il backup, come database e archiviazione di oggetti, sono sempre attive. Altri elementi come i server applicativi o il calcolo serverless non vengono distribuiti, ma possono essere creati quando necessari con la configurazione e il codice applicativo richiesti.
- Warm standby (RPO in secondi, RTO in minuti): mantieni sempre una versione ridotta del carico di lavoro completamente funzionante in esecuzione nella regione di ripristino. I sistemi business critical sono completamente duplicati e sono sempre accesi, ma con un parco istanze ridimensionato. I dati vengono replicati e si trovano nella regione di ripristino. Quando viene il momento del ripristino, il sistema viene dimensionato rapidamente per gestire il carico di produzione. Più il Warm standby è dimensionato verso l'alto e più bassi saranno l'RTO e l'affidamento al piano di controllo. Quando il dimensionamento è completo ci troviamo nello Standby a caldo.

- Multi-regione (multi-sito) attivo-attivo (RPO vicino a zero, RTO uguale potenzialmente a zero): il carico di lavoro viene distribuito in più regioni Regioni AWS e serve attivamente il traffico da esse proveniente. Questa strategia comporta la sincronizzazione dei dati tra le regioni. È necessario evitare o gestire possibili conflitti causati da scritture sullo stesso record in due diverse repliche regionali, un'attività che potrebbe rivelarsi complessa. La replica dei dati è utile per la sincronizzazione dei dati e ti proteggerà da alcuni tipi di disastri, ma non dalla corruzione o dalla distruzione dei dati, a meno che la tua soluzione non includa opzioni per il ripristino point-in-time.

> **Note**
>
> La differenza tra Pilot Light e Warm Standby può talvolta essere difficile da comprendere. Entrambe prevedono un ambiente nella tua regione di ripristino con copie degli asset della tua regione principale. La differenza è che Pilot Light non può elaborare le richieste senza aver prima intrapreso altre azioni, mentre Warm Standby può gestire immediatamente il traffico (a livelli ridotti di capacità). Pilot Light ti richiederà di attivare i server, distribuire possibilmente un'infrastruttura aggiuntiva (non di base) e aumentare il dimensionamento, mentre Warm Standby richiede solo di aumentare il dimensionamento (tutto è già stato distribuito ed è in esecuzione). Scegli tra queste opzioni in base alle tue esigenze di RTO e RPO.

Risultato desiderato:

Per ogni carico di lavoro esiste una strategia di ripristino di emergenza definita e implementata che consente a quel carico di lavoro di raggiungere gli obiettivi di ripristino. Le strategie di ripristino di emergenza tra carichi di lavoro utilizzano modelli riutilizzabili (come strategie descritte in precedenza),

Anti-pattern comuni:

- Implementazione di procedure di ripristino incoerenti per carichi di lavoro con obiettivi di ripristino simili.
- Implementazione di una strategia di ripristino di emergenza ad-hoc quando si verifica un disastro.
- Assenza di un piano per il ripristino di emergenza.
- Dipendenza dalle operazioni del piano di controllo durante il ripristino.

Vantaggi dell'adozione di questa best practice:

- L'utilizzo di strategie di ripristino definite consente di utilizzare strumenti e procedure di test comuni.
- L'utilizzo di strategie di ripristino definite consente una condivisione più efficiente delle conoscenze tra i team e un'implementazione più facile del ripristino di emergenza sui carichi di lavoro proprietari.

Livello di rischio associato se questa best practice non fosse adottata: Alto

- Senza una strategia di ripristino di emergenza pianificata, implementata e testata, è poco probabile riuscire a raggiungere gli obiettivi di ripristino in caso di eventi disastrosi.

Guida all'implementazione

Per ognuno di questi passaggi guarda i dettagli qui di seguito.

1. Definisci una strategia di ripristino di emergenza in linea con i requisiti di ripristino di questo carico di lavoro.
2. Esamina i modelli con cui la strategia di ripristino di emergenza selezionata può essere implementata.
3. Valuta le risorse del tuo carico di lavoro e quale sarà la loro configurazione nella regione di ripristino prima del failover (durante la normale operatività).
4. Stabilisci e implementa le modalità con cui preparerai la tua regione al failover nel momento in cui sarà necessario (durante un evento disastroso).
5. Stabilisci e implementa le modalità con cui reindirizzerai il traffico al failover nel momento in cui sarà necessario (durante un evento disastroso).
6. Progetta un piano per il failback del carico di lavoro.

Passaggi dell'implementazione

1. Definisci una strategia di ripristino di emergenza in linea con i requisiti di ripristino di questo carico di lavoro.

Scegliere una strategia di ripristino di emergenza significa raggiungere un compromesso tra la riduzione dei tempi di inattività e della perdita di dati (RTO e RPO) e costi e complessità di implementazione. Dovresti evitare di implementare una strategia che sia più severa del necessario, in quanto questo comporterebbe costi aggiuntivi.

Ad esempio, nel diagramma seguente, l'azienda ha stabilito l'RTO massimo concesso e il limite di spesa per la strategia di ripristino del servizio. Considerati gli obiettivi dell'azienda, le strategie di ripristino di emergenza Pilot Light o Warm Standby soddisfano i criteri sui costi e l'RTO.

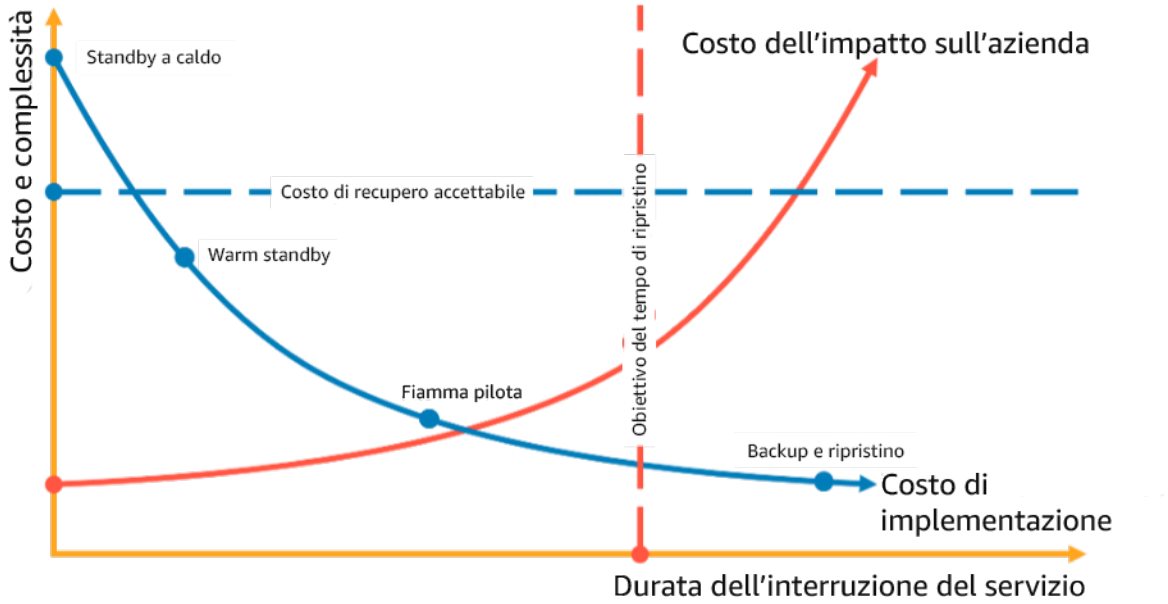


Figure 18: Scegliere una strategia di ripristino di emergenza in base all'RTO e ai costi

Per saperne di più consulta il piano di continuità aziendale (BCP).

2. Esamina i modelli con cui la strategia di ripristino di emergenza selezionata può essere implementata.

Questo passaggio consiste nel capire come implementare la strategia selezionata. Le strategie vengono spiegate con Regioni AWS come siti principali e di ripristino. Tuttavia, puoi anche decidere di utilizzare le zone di disponibilità in una singola regione come strategia di ripristino di emergenza, utilizzando aspetti di più strategie.

Nei passaggi successivi a questo, applicherai la strategia per il tuo carico di lavoro specifico.

Backup e ripristino

Backup e ripristino è la strategia meno complessa da implementare, ma richiederà più tempo e impegno per ripristinare il carico di lavoro, generando così valori RTO e RPO più elevati. È buona pratica creare sempre backup dei dati e copiarli in un altro sito (ad esempio, un altro Regione AWS).

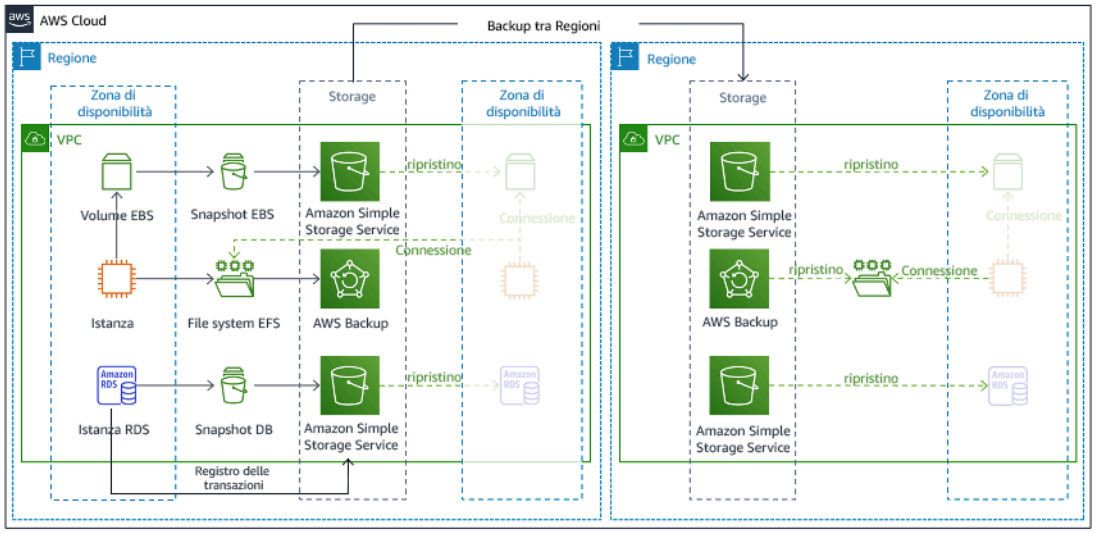


Figura 19: architettura di backup e ripristino

Per maggiori dettagli su questa strategia consulta [Disaster Recovery (DR) Architecture on AWS, Part II: Backup and Restore with Rapid Recovery (Architettura di ripristino di emergenza (DR) su AWS, Parte II: backup e ripristino con recupero rapido)].

Pilot light

Con l'approccio Pilot light , replichi i dati dalla tua regione principale alla regione di ripristino. Le risorse di base utilizzate per l'infrastruttura del carico di lavoro vengono distribuite nella regione di ripristino; tuttavia sono comunque necessarie risorse aggiuntive ed eventuali dipendenze per rendere funzionale questo stack. Ad esempio, nella Figura 20, non vengono distribuite istanze di calcolo.

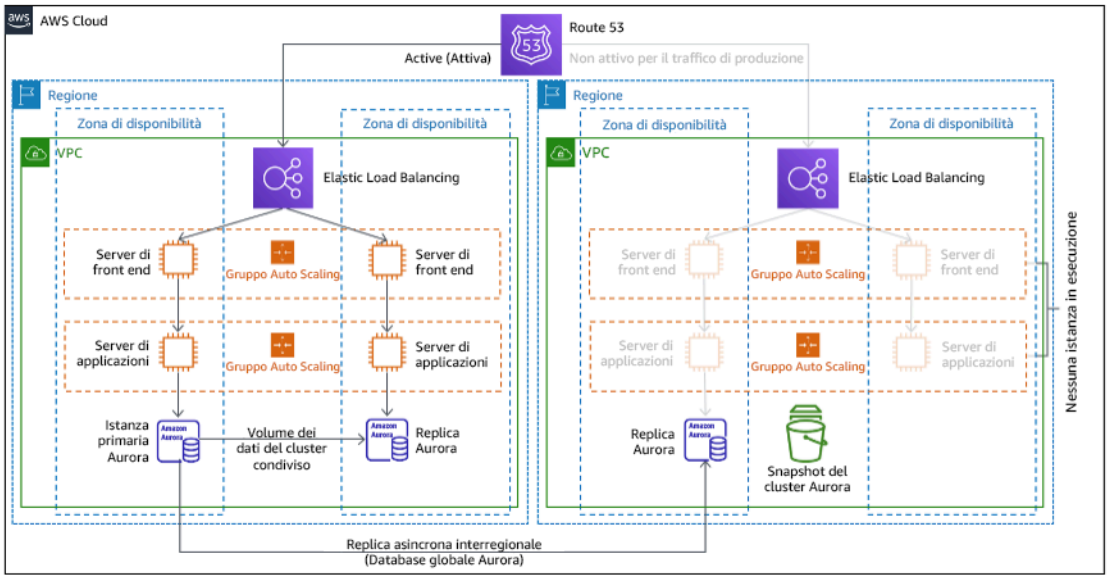


Figura 20: architettura pilot light

Per maggiori dettagli su questa strategia consulta Disaster Recovery (DR) Architecture on AWS, Part III: Pilot Light and Warm Standby (Architettura di ripristino di emergenza (DR) su AWS, Parte III: Pilot Light e Warm Standby).

Warm standby

L'approccio warm standby implica la verifica della presenza di una copia ridotta, ma comunque funzionale, dell'ambiente di produzione in un'altra regione. Questo approccio estende il concetto di Pilot Light e diminuisce il tempo di ripristino, poiché il carico di lavoro è sempre attivo in un'altra regione. Se la regione di ripristino ha raggiunto il massimo della capacità, allora viene definita come Standby a caldo.

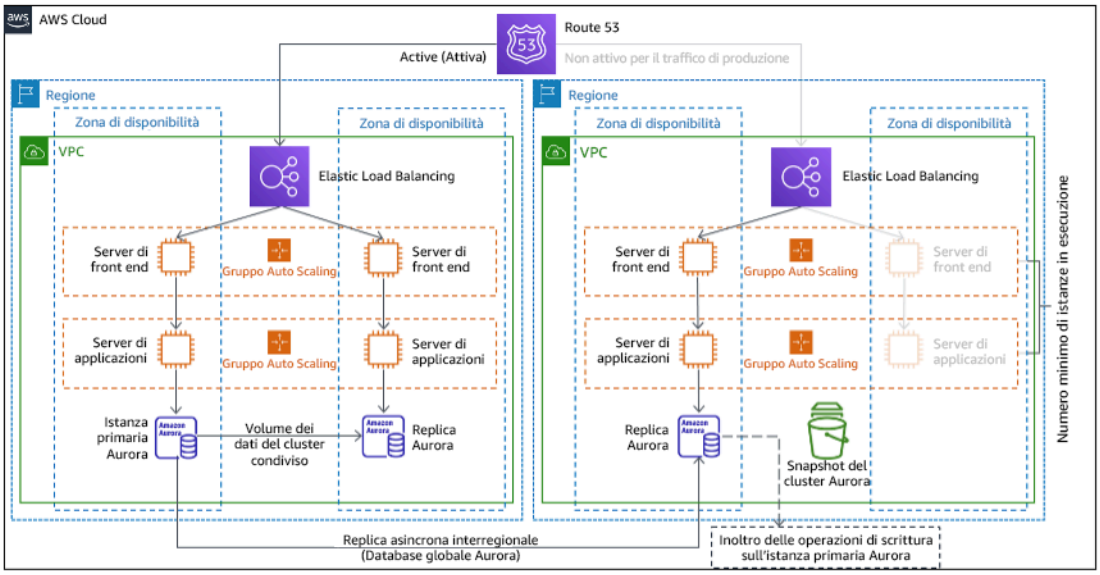


Figure 21: Architettura Warm Standby

Se si utilizza Warm Standby o Pilot Light è necessario un aumento delle risorse nella regione di ripristino. Per garantire che la capacità sia disponibile quando necessario, valuta l'uso di prenotazioni delle capacità per le istanze EC2. Se utilizzi AWS Lambda, la concorrenza fornita può garantire gli ambienti di esecuzione, in modo che siano pronti a rispondere immediatamente ai richiami della funzione.

Per maggiori dettagli su questa strategia consulta Disaster Recovery (DR) Architecture on AWS, Part III: Pilot Light and Warm Standby (Architettura di ripristino di emergenza (DR) su AWS, Parte III: Pilot Light e Warm Standby).

Attivo/attivo multi-sito

Puoi eseguire il carico di lavoro simultaneamente in più regioni come parte di una strategia attivo/attivo multi-sito . La strategia attivo/attivo multi-sito serve il traffico da tutte le regioni in cui è distribuita. I clienti possono selezionare questa strategia per motivi diversi dal ripristino di emergenza. Può essere utilizzata per aumentare la disponibilità o nella distribuzione di un carico di lavoro a un pubblico globale (per posizionare l'endpoint più vicino agli utenti e/o per distribuire stack localizzati al pubblico di quella regione). Come strategia di ripristino di emergenza, se il carico di lavoro non può essere supportato in una delle Regioni AWS in cui è stato distribuito, allora quella regione viene evacuata e le regioni rimanenti vengono utilizzate per garantire la disponibilità. Attivo/attivo multi-sito è la strategia di ripristino operativamente più complessa e dovrebbe essere selezionata solo quando lo richiedono i requisiti aziendali.

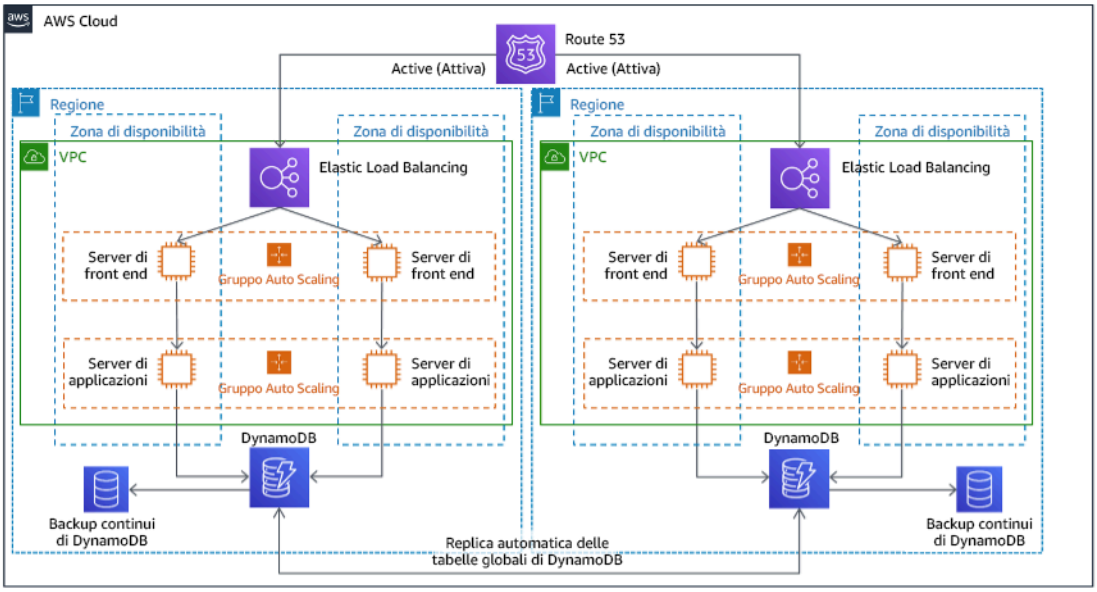


Figura 22: Architettura attivo/attivo multi-sito

Per maggiori dettagli su questa strategia consulta Disaster Recovery (DR) Architecture on AWS, Part IV: Multi-site Active/Active (Architettura di ripristino di emergenza su AWS, parte IV: attiva/attiva multi-sito).

Procedure aggiuntive per la protezione dei dati

Con tutte le strategie devi anche mitigare un disastro relativo ai dati. La replica continua dei dati ti proteggerà da alcuni tipi di disastri, ma non dalla corruzione o dalla distruzione dei dati, a meno che la tua soluzione non includa opzioni per il ripristino point-in-time o il controllo delle versioni dei dati archiviati. Devi anche creare un backup dei dati replicati nel sito di ripristino per creare backup point-in-time in aggiunta alle repliche.

Utilizzo di più zone di disponibilità all'interno di una singola Regione AWS

Quando si usano più zone di disponibilità all'interno di un'unica regione, l'implementazione della strategia di ripristino di emergenza usa più elementi delle strategie precedenti. Per prima cosa devi creare un'architettura con disponibilità elevata (HA), usando più zone di disponibilità come mostrato nella Figura 23. Questa architettura utilizza un approccio attivo/attivo multi-sito, poiché le istanze Amazon EC2 ed Elastic Load Balancer hanno risorse distribuite in più zone di disponibilità che gestiscono attivamente le richieste. L'architettura dimostra anche lo standby a caldo e se l'istanza Amazon RDS primaria fallisce (o la zona di disponibilità stessa fallisce), l'istanza in standby viene promossa a principale.

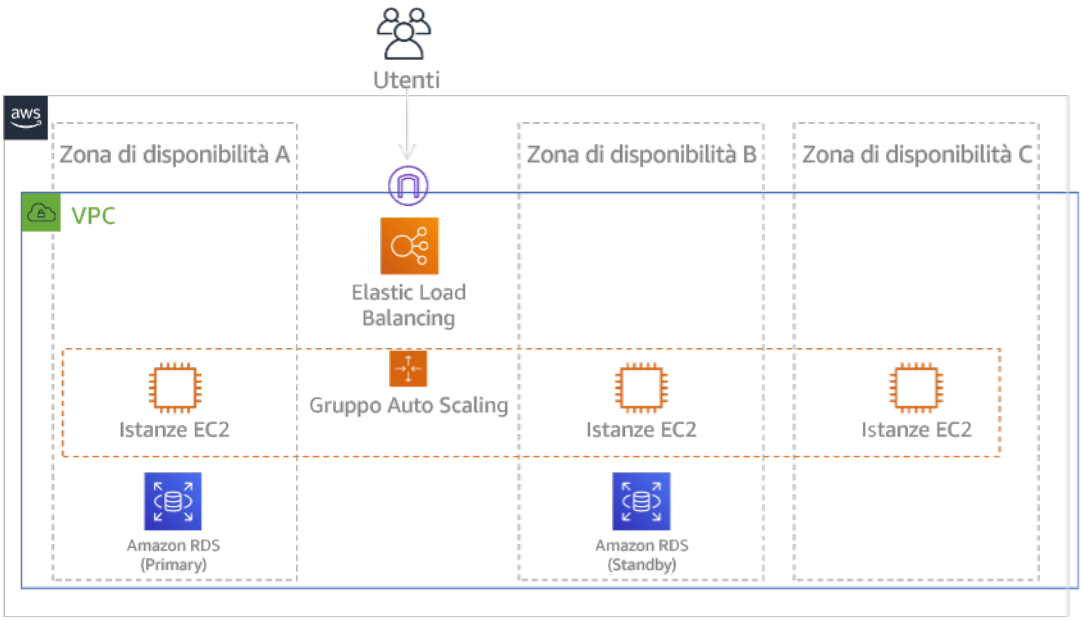


Figura 23: Architettura con più zone di disponibilità

Oltre a questa architettura HA, devi aggiungere i backup di tutti i dati richiesti per eseguire il tuo carico di lavoro. Questo aspetto è importante soprattutto per i dati limitati a una singola zona come i volumi Amazon EBS oppure i cluster Amazon Redshift. Se fallisce una zona di disponibilità, dovrai ripristinare i dati in un'altra zona di disponibilità. Laddove possibile, devi anche copiare i backup di dati su un'altra Regione AWS come livello di protezione aggiuntivo.

Un approccio alternativo meno comune a una strategia di ripristino di emergenza con una singola regione e più zone di disponibilità è illustrata nel post del blog, Building highly resilient applications using Amazon Route 53 Application Recovery Controller, Part 1: Single-Region stack (Creazione di applicazioni altamente resilienti con Amazon Route 53 Application Recovery Controller, parte 1: stack a singola regione). In questo caso la strategia adottata è quella di garantire il più possibile l'isolamento tra le zone di disponibilità, ossia come le regioni operano. Usando questa strategia alternativa puoi scegliere un approccio attivo/attivo o attivo/passivo.

Nota: alcuni carichi di lavoro hanno requisiti normativi di residenza dei dati. Se questo si applica a un carico di lavoro in una località che attualmente ha solo una Regione AWS, la multi-regione non soddisferà i requisiti aziendali. Le strategie con più zone di disponibilità offrono una buona protezione dalla maggior parte dei disastri.

3. Valuta le risorse del tuo carico di lavoro e quale sarà la loro configurazione nella regione di ripristino prima del failover (durante la normale operatività).

Per infrastrutture e risorse AWS usa l'infrastruttura come codice come AWS CloudFormation o strumenti di terze parti come Hashicorp Terraform. Per distribuire in più account e regioni con una singola operazione puoi usare AWS CloudFormation StackSets. Per le strategie multi-sito attivo/attivo e standby a caldo, l'infrastruttura distribuita nella tua regione di ripristino ha le stesse risorse della regione principale. Per le strategie Pilot Light e Warm Standby l'infrastruttura distribuita richiederà azioni aggiuntive per essere pronta per la produzione. Con l'utilizzo dei parametri di CloudFormation e della logica condizionale, puoi verificare se uno stack distribuito è attivo o in standby con un singolo modello. Un esempio di tale modello CloudFormation è incluso in questo post del blog.

Tutte le strategie di ripristino di emergenza richiedono un backup delle origini dei dati all'interno della Regione AWS e una copia di tali backup nella regione di ripristino. AWS Backup offre una visualizzazione centralizzata dove puoi configurare, pianificare e monitorare i backup di queste risorse. Per Pilot Light, Warm Standby e Multi-sito attivo/attivo, you should also replicate data from the primary devi anche replicare i dati dalla regione principale alle risorse di dati nella regione di ripristino, come istanze DB Amazon Relational Database Service (Amazon RDS) o tabelle Amazon DynamoDB . Queste risorse di dati sono pertanto attive e pronte per servire le richieste nella regione di ripristino.

Per saperne di più su come i servizi AWS operano nelle regioni, guarda questa serie di blog su Creazione di un'applicazione multiregione con i servizi AWS.

4. Stabilisci e implementa le modalità con cui preparerai la tua regione al failover nel momento in cui sarà necessario (durante un evento disastroso).

Per la strategia attivo/attivo multi-sito, il failover significa evacuare una regione e affidarsi alle regioni attive rimanenti. In generale, tali regioni sono pronte per accettare il traffico. Per le strategie Pilot Light e Warm Standby, le azioni di ripristino devono distribuire le risorse mancanti, come le istanze EC2 nella Figura 20, oltre ad risorse mancanti aggiuntive.

Per tutte le strategie precedenti potresti dover promuovere istanze di database i sola lettura a istanze di lettura/scrittura principali.

Per il backup e il ripristino, il ripristino dei dati dai backup crea risorse per tali dati, come volumi EBS, istanze DB RDS e tabelle DynamoDB. Devi anche ripristinare l'infrastruttura e distribuire il codice. Puoi usare AWS Backup per ripristinare i dati nella regione di ripristino. Consulta REL09-BP01 Identificazione e backup di tutti i dati che richiedono un backup o riproduzione dei dati dalle origini per ulteriori dettagli. Ricreare l'infrastruttura significa anche creare risorse come istanze EC2

oltre a Amazon Virtual Private Cloud (Amazon VPC), sottoreti e i gruppi di sicurezza necessari. Puoi automatizzare gran parte del processo di ripristino. Per scoprire come guarda questo post del blog.

5. Stabilisci e implementa le modalità con cui reindirizzerai il traffico al failover nel momento in cui sarà necessario (durante un evento disastroso).

Questa operazione di failover può essere avviata automaticamente o manualmente. Il failover avviato automaticamente in base a controlli dell'integrità o allarmi deve essere usato con attenzione, poiché un failover non necessario (falso allarme) comporta dei costi in termini di non disponibilità e perdita dei dati. Pertanto si usa spesso il failover avviato manualmente. In questo caso, devi comunque automatizzare i passaggi del failover, in modo che l'avvio manuale si limiti al clic su un pulsante.

Esistono diverse opzioni di gestione del traffico da considerare quando si usano i servizi AWS. Un'opzione consiste nell'utilizzare Amazon Route 53. Con Amazon Route 53 puoi associare più endpoint IP in una o più Regioni AWS con un nome di dominio Route 53. Per implementare un failover avviato manualmente puoi usare Amazon Route 53 Application Recovery Controller, che offre un'API del piano dati altamente disponibile per reindirizzare il traffico alla regione di ripristino. Nella fase di implementazione del failover, usa le operazioni di piano dati ed evita quelle del piano di controllo come descritto in REL11-BP04 Fare affidamento al piano dati invece che al piano di controllo durante il ripristino.

Per saperne di più su questa e su altre opzioni consulta questa sezione del whitepaper sul Ripristino di emergenza.

6. Progetta un piano per il failback del carico di lavoro.

Si parla di failback quando un'operazione del carico di lavoro torna alla regione principale, dopo che un vento disastroso è diminuito di intensità. Il provisioning di infrastruttura e codice alla regione principale in genere segue gli stessi passaggi usati inizialmente, affidandosi all'infrastruttura come codice e alle pipeline di distribuzione del codice. La sfida del failback è il ripristino dei data store e la garanzia della loro coerenza con la regione di ripristino attiva.

Nello stato di failover i database nella regione di ripristino sono attivi e hanno dati aggiornati. L'obiettivo è eseguire una nuova sincronizzazione tra la regione di ripristino e la regione principale, per garantire il suo aggiornamento.

Alcuni servizi AWS eseguono questa operazione in automatico. Se si utilizzano tabelle globali Amazon DynamoDB, anche se la tabella nella regione principale era diventata non disponibile,

quando torna di nuovo online, DynamoDB ripristina la propagazione di scritture in sospeso. Se si utilizzano Database globale Amazon Aurora e failover pianificato gestito, viene mantenuta la topologia di replica esistente del database globale Aurora. Pertanto, l'istanza precedente in lettura/scrittura nella regione principale diventa una replica e riceve gli aggiornamenti dalla regione di ripristino.

Nei casi in cui questo non è automatico devi ristabilire il database nella regione principale come replica del database nella regione di ripristino. In molti casi questo comporterà l'eliminazione del database principale precedente e la creazione di nuove repliche. Ad esempio, per istruzioni su come procedere con il Database globale Amazon Aurora in caso di failover non pianificato , consulta questa scheda: Failback di un database globale.

Dopo un failover, se puoi proseguire l'esecuzione nella tua regione di ripristino, valuta la possibilità di farlo nella tua regione principale. Compieresti comunque tutte le operazioni precedenti per trasformare la precedente regione principale in una regione di ripristino. Alcune organizzazioni eseguono una rotazione pianificata, scambiando periodicamente le regioni principale e di ripristino (ad esempio, ogni tre mesi).

Tutti i passaggi richiesti per failover e failback devono essere inseriti in un playbook disponibile a tutti i membri del team, sottoposto periodicamente a revisione.

Livello di impegno per il piano di implementazione: elevato

Risorse

Best practice correlate:

- the section called “REL09-BP01 Identificazione e backup di tutti i dati che richiedono un backup o riproduzione dei dati dalle origini”
- the section called “REL11-BP04 Fare affidamento al piano dati invece che al piano di controllo durante il ripristino”
- the section called “REL13-BP01 Definizione degli obiettivi di ripristino in caso di downtime e perdita di dati”

Documenti correlati:

- AWS Architecture Blog: Disaster Recovery Series
- Ripristino di emergenza dei carichi di lavoro su AWS: ripristino nel cloud (whitepaper di AWS)
- Opzioni di ripristino di emergenza nel cloud

- Build a serverless multi-region, active-active backend solution in an hour
- Multi-region serverless backend — reloaded
- RDS: creazione di una replica di lettura in una regione AWS diversa
- Route 53: configurazione del failover DNS
- S3: replica tra regioni
- Che cos'è AWS Backup?
- What is Route 53 Application Recovery Controller? (Che cos'è Amazon Route 53 Application Recovery Controller?)
- AWS Elastic Disaster Recovery (Ripristino di emergenza elastico AWS)
- HashiCorp Terraform: inizia subito - AWS
- Partner APN: partner che possono assistere con disaster recovery
- Marketplace AWS: prodotti utilizzabili per il disaster recovery

Video correlati:

- Ripristino di emergenza di carichi di lavoro su AWS
- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Modelli di architettura per applicazioni attive-attive multiregione) (ARC209-R2)
- Get Started with AWS Elastic Disaster Recovery | Amazon Web Services (Nozioni di base sul ripristino di emergenza elastico AWS | Amazon Web Services)

Esempi correlati:

- AWS Well-Architected Labs - Ripristino di emergenza - Serie di workshop che illustrano le strategie di ripristino di emergenza

REL13-BP03 Esecuzione di test sull'implementazione del ripristino di emergenza per convalidare l'implementazione

Testa con regolarità il failover nella tua sede di ripristino per verificare la correttezza delle operazioni e l'allineamento ai valori RPO e RTO.

Un modello da evitare è lo sviluppo di percorsi di ripristino eseguiti raramente. Ad esempio, è possibile che si disponga di un archivio dati secondario utilizzato per query di sola lettura. Quando

scrivi in un archivio dati e quello principale ha un guasto, puoi eseguire il failover verso l'archivio dati secondario. Se non testi frequentemente questo failover, è possibile che i presupposti relativi alle funzionalità dell'archivio dati secondario non siano corretti. La capacità dell'archivio dati secondario, che potrebbe essere stata sufficiente durante l'ultimo test, potrebbe non essere più in grado di tollerare il carico in questo scenario. La nostra esperienza ha dimostrato che l'unico ripristino da errore che funziona è il percorso sottoposto a frequenti test. Per questo è preferibile avere un numero ridotto di percorsi di ripristino. Puoi stabilire dei modelli di ripristino e testarli regolarmente. Se disponi di un percorso di ripristino complesso o critico, devi comunque riprodurre regolarmente il guasto specifico in produzione per convincerti che il percorso di ripristino funzioni. Nell'esempio appena discusso, è necessario eseguire il failover regolarmente in standby, indipendentemente dalle necessità.

Anti-pattern comuni:

- Non eseguire mai failover di prova in produzione.

Vantaggi dell'adozione di questa best practice: Testare regolarmente il piano di disaster recovery assicura che funzioni quando necessario e che il tuo team sappia come eseguire la strategia.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

- Progetta i carichi di lavoro per il ripristino. Testa con regolarità se l'informatica orientata al ripristino (ROC, Recovery Oriented Computing) identifica le caratteristiche nei sistemi che migliorano il ripristino. Queste caratteristiche sono: isolamento e ridondanza, capacità a livello di sistema di ripristinare le modifiche, capacità di monitorare e determinare lo stato, capacità di fornire diagnostica, ripristino automatizzato, progettazione modulare e possibilità di riavvio. Esegui il percorso di ripristino per assicurarti di poter realizzare il ripristino nel tempo specificato allo stato specificato. Usa i tuoi runbook durante questo ripristino per documentare i problemi e trovare le loro soluzioni prima del test successivo.
  - Il progetto di informatica orientata al ripristino Berkeley/Stanford
- Usa il ripristino di emergenza CloudEndure per implementare e testare la tua strategia di ripristino di emergenza.
  - Testing the Disaster Recovery Solution with CloudEndure
  - CloudEndure Disaster Recovery
  - Ripristino di emergenza CloudEndure in AWS

Risorse

Documenti correlati:

- Partner APN: partner che possono assistere con disaster recovery
- AWS Architecture Blog: Disaster Recovery Series
- Marketplace AWS: prodotti utilizzabili per il disaster recovery
- CloudEndure Disaster Recovery
- Ripristino di emergenza dei carichi di lavoro su AWS: ripristino nel cloud (whitepaper di AWS)
- Testing the Disaster Recovery Solution with CloudEndure
- Il progetto di informatica orientata al ripristino Berkeley/Stanford
- Che cos'è AWS Fault Injection Simulator?

Video correlati:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Modelli di architettura per applicazioni attive-attive multiregione) (ARC209-R2)
- AWS re:Invent 2019: Backup-and-restore and disaster-recovery solutions with AWS (STG208)

Esempi correlati:

- AWS Well-Architected Labs - Test di resilienza

REL13-BP04 Gestione della deviazione di configurazione nel sito o nella Regione del ripristino di emergenza

Assicurati che l'infrastruttura, i dati e la configurazione soddisfino le esigenze del sito o nella Regione del ripristino di emergenza. Ad esempio, controlla che le AMI e le quote di servizio siano aggiornate.

AWS Config monitora e registra in modo continuo le configurazioni delle risorse AWS. È in grado di rilevare le deviazioni e attivare AWS Systems Manager Automation per risolverle e attivare allarmi. AWS CloudFormation è inoltre in grado di rilevare le deviazioni negli stack distribuiti.

Anti-pattern comuni:

- Non eseguire aggiornamenti nelle sedi di ripristino, quando esegui modifiche di configurazione o di infrastruttura nelle tue sedi principali.

- Ignorare le limitazioni potenziali (ad esempio le differenze di servizio) nelle sedi di disaster recovery e principali.

Vantaggi dell'adozione di questa best practice: Assicurarsi che l'ambiente di disaster recovery sia coerente con quello esistente garantisce il ripristino completo.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Assicurati che le tue pipeline di distribuzione riforniscano sia i siti principali che di backup. Le pipeline per la distribuzione di applicazioni in produzione devono essere distribuite in tutte le posizioni della strategia di disaster recovery specificate, inclusi gli ambienti di sviluppo e test.
- Abilitazione di AWS Config per monitorare le potenziali posizioni di deviazione. Utilizza le regole AWS Config per creare sistemi in grado di applicare le strategie di disaster recovery e generare avvisi quando rilevano una deviazione.
  - Correzione di risorse AWS non conformi in base alle regole di Regole di AWS Config
  - AWS Systems Manager Automation
- Utilizza AWS CloudFormation per distribuire la tua infrastruttura. AWS CloudFormation è in grado di rilevare le deviazioni tra ciò che i modelli di CloudFormation specificano e ciò che viene effettivamente distribuito.
  - AWS CloudFormation: rilevamento delle deviazioni su un intero stack CloudFormation

Risorse

Documenti correlati:

- Partner APN: partner che possono assistere con disaster recovery
- AWS Architecture Blog: Disaster Recovery Series
- AWS CloudFormation: rilevamento delle deviazioni su un intero stack CloudFormation
- Marketplace AWS: prodotti utilizzabili per il disaster recovery
- AWS Systems Manager Automation
- Ripristino di emergenza dei carichi di lavoro su AWS: ripristino nel cloud (whitepaper di AWS)
- In che modo è possibile implementare una soluzione di gestione della configurazione dell'infrastruttura in AWS?

- Correzione di risorse AWS non conformi in base alle regole di Regole di AWS Config

Video correlati:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Modelli di architettura per applicazioni attive-attive multiregione) (ARC209-R2)

REL13-BP05 Automatizzazione del ripristino

Utilizza AWS o strumenti di terze parti per automatizzare il ripristino del sistema e instradare il traffico verso il sito o la Regione del ripristino di emergenza.

In base ai controlli di integrità configurati, i servizi AWS, come Elastic Load Balancing e AWS Auto Scaling, possono distribuire il carico a zone di disponibilità integre, mentre i servizi, come Amazon Route 53 e AWS Global Accelerator, instradano il carico a Regioni AWS integre. Amazon Route 53 Application Recovery Controller aiuta a gestire e coordinare il failover utilizzando i controlli di disponibilità e le funzionalità di controlli di routing. Queste funzionalità monitorano continuamente la capacità dell'applicazione di riprendersi dai guasti e permettono di controllarne il ripristino delle applicazioni su più Regioni AWS, zone di disponibilità e on-premise.

Per i carichi di lavoro su data center fisici o virtuali o cloud privati, Ripristino di emergenza elastico AWS, disponibile tramite Marketplace AWS, consente alle organizzazioni di organizzare una strategia di ripristino di emergenza su AWS. CloudEndure supporta, inoltre, il ripristino di emergenza tra Regioni e zone di disponibilità in AWS.

Anti-pattern comuni:

- L'implementazione di failover e failback automatici identici può causare flapping quando si verifica un errore.

Vantaggi dell'adozione di questa best practice: Il ripristino automatico riduce i tempi di ripristino eliminando la possibilità di errori manuali.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Automatizzazione dei percorsi di ripristino. Per tempi di ripristino brevi, non è possibile servirsi del giudizio umano e dell'azione per scenari di disponibilità elevata. Il sistema dovrebbe ripristinarsi automaticamente in ogni situazione.
  - Usa il ripristino di emergenza CloudEndure per failover e failback automatizzati. Il ripristino di emergenza CloudEndure replica in modo continuo le macchine (tra cui sistema operativo, configurazione dello stato del sistema, database, applicazioni e file) in un'area di gestione temporanea a basso costo nell'Account AWS di destinazione e nella Regione preferita. In caso di emergenza, è possibile indicare a CloudEndure Disaster Recovery di avviare automaticamente migliaia di macchine nello stato di provisioning completo in pochi minuti.
    - Performing a Disaster Recovery Failover and Failback
    - CloudEndure Disaster Recovery

Risorse

Documenti correlati:

- Partner APN: partner che possono assistere con disaster recovery
- AWS Architecture Blog: Disaster Recovery Series
- Marketplace AWS: prodotti utilizzabili per il disaster recovery
- AWS Systems Manager Automation
- Ripristino di emergenza CloudEndure in AWS
- Ripristino di emergenza dei carichi di lavoro su AWS: ripristino nel cloud (whitepaper di AWS)

Video correlati:

- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (Modelli architetturali per applicazioni attive-attive su più Regioni) (ARC209-R2)

Efficienza delle prestazioni

Argomenti

- Selezione
- Revisione

- Monitoraggio
- Compromessi

Selezione

Domande

- PERF 1 In che modo selezioni l'architettura più performante?
- PERF 2 In che modo selezioni la tua soluzione di calcolo?
- PERF 3 In che modo selezioni la soluzione di storage?
- PERF 4 In che modo selezioni la soluzione di database?
- PERF 5 In che modo configuri la soluzione di rete?

PERF 1 In che modo selezioni l'architettura più performante?

Spesso sono necessari molteplici approcci per ottenere prestazioni ottimali in un carico di lavoro. I sistemi Well-Architected utilizzano soluzioni multiple e funzionalità diverse per migliorare le prestazioni.

Best practice

- PERF01-BP01 Identificazione dei servizi e delle risorse disponibili
- PERF01-BP02 Definizione di un processo per le scelte architetturali
- PERF01-BP03 Considerazione dei requisiti di costo nelle decisioni
- PERF01-BP04 Utilizzo di policy o architetture di riferimento
- PERF01-BP05 Utilizzo delle linee guida del fornitore di servizi cloud o di un partner appropriato
- PERF01-BP06 Benchmarking dei carichi di lavoro esistenti
- PERF01-BP07 Esecuzione di un test di carico sul carico di lavoro

PERF01-BP01 Identificazione dei servizi e delle risorse disponibili

Scopri tutte le informazioni sull'ampia gamma di servizi e risorse disponibili nel cloud. Identifica quali servizi e opzioni di configurazione sono pertinenti per il tuo carico di lavoro e studia come utilizzarli per raggiungere prestazioni ottimali.

Se stai valutando un carico di lavoro esistente, devi generare un inventario delle varie risorse dei servizi che utilizza. Tale inventario ti aiuta a valutare quali componenti possono essere sostituiti con servizi gestiti e tecnologie più recenti.

Anti-pattern comuni:

- Utilizzi il cloud come data center in co-location.
- Utilizzi lo storage condiviso per tutto ciò che necessita di storage persistente.
- Non utilizzi l'auto scaling.
- Se necessario, utilizzi tipi di istanze strettamente correlate ai tuoi standard attuali, ma più grandi.
- Distribuisci e gestisci le tecnologie disponibili come servizi gestiti.

Vantaggi dell'adozione di questa best practice: Prendendo in considerazione i servizi con cui non hai familiarità, puoi ridurre notevolmente il costo dell'infrastruttura e l'impegno necessario per mantenere i servizi. Distribuendo nuovi servizi e funzionalità potresti accelerare l'immissione sul mercato.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Inventario del software e dell'architettura del carico di lavoro per i servizi correlati: raccogli un inventario del carico di lavoro e scegli la categoria di prodotti su cui ottenere ulteriori informazioni. Identifica componenti del carico di lavoro che possano essere sostituiti con servizi gestiti per aumentare le prestazioni e ridurre la complessità operativa.

Risorse

Documenti correlati:

- Centro di progettazione AWS
- AWS Partner Network
- Portfolio di soluzioni AWS
- Portale del sapere AWS

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- La mia architettura

Esempi correlati:

- Esempi di AWS
- Esempi di SDK AWS

PERF01-BP02 Definizione di un processo per le scelte architetturali

Affidati all'esperienza e alle competenze interne in materia di cloud o utilizza risorse esterne, come casi d'uso pubblicati, documentazione pertinente o whitepaper, per definire un processo per scegliere risorse e servizi. È necessario definire un processo che incoraggi la sperimentazione e il benchmarking con i servizi che potrebbero essere utilizzati nel tuo carico di lavoro.

Durante lo studio degli scenari utente critici per la tua architettura, devi includere i requisiti relativi alle prestazioni, specificando ad esempio la rapidità con il quale deve essere eseguito ogni scenario. Per questi scenari critici, devi implementare percorsi utente con script aggiuntivi per chiarire esattamente quali sono le loro prestazioni rispetto ai requisiti.

Anti-pattern comuni:

- Ritieni che l'architettura corrente diventi statica e non venga aggiornata nel corso del tempo.
- Introduci modifiche all'architettura nel tempo senza giustificazioni.

Vantaggi dell'adozione di questa best practice: Un processo definito per apportare modifiche all'architettura, ti consente di utilizzare i dati raccolti per influenzare la progettazione del carico di lavoro nel corso del tempo.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Scelta di un approccio architetturale: identifica il tipo di architettura che soddisfa i tuoi requisiti prestazionali. Identifica i vincoli, come il mezzo per la distribuzione (desktop, Web, dispositivo mobile, IoT), i requisiti legacy e le integrazioni. Identifica le opportunità di riutilizzo, incluso il refactoring. Consulta altri team, diagrammi architetturali e risorse come AWS Solutions Architect, architetture di riferimento AWS e Partner AWS per scegliere un'architettura.

Definisci i requisiti prestazionali: utilizza l'esperienza del cliente per identificare i parametri più importanti. Per ciascuno di essi, identifica l'obiettivo, l'approccio per la misurazione e la priorità. Definisci l'esperienza cliente. Documenta l'esperienza prestazionale richiesta dai clienti, incluso

come i clienti giudicano le performance del carico di lavoro. Assegna le priorità a problemi riguardanti l'esperienza per i casi utente critici. Includi requisiti prestazionali e implementa percorsi utente con script per chiarire esattamente quali sono le loro prestazioni rispetto ai requisiti.

Risorse

Documenti correlati:

- Centro di progettazione AWS
- AWS Partner Network
- Portfolio di soluzioni AWS
- Knowledge Center di AWS

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- La mia architettura

Esempi correlati:

- Esempi di AWS
- Esempi di SDK AWS

PERF01-BP03 Considerazione dei requisiti di costo nelle decisioni

I carichi di lavoro spesso hanno requisiti di costo per il funzionamento. Utilizza i controlli dei costi interni per selezionare le dimensioni e i tipi di risorse in base alle necessità previste in termini di risorse.

Determina quali componenti del carico di lavoro possono essere sostituiti con dei servizi completamente gestiti, come i database gestiti, le cache in memoria e i servizi ETL. La riduzione del carico di lavoro operativo consente di concentrare le risorse sui risultati aziendali.

Per le best practice relative ai requisiti di costo, consulta la sezione Risorse convenienti del Whitepaper sul principio dell'ottimizzazione dei costi.

Anti-pattern comuni:

- Utilizzi una sola famiglia di istanze.

- Ometti di valutare le soluzioni con licenza rispetto alle soluzioni open-source
- Utilizzi solo lo storage a blocchi.
- Implementa software comune su istanze EC2 ed Amazon EBS o volumi temporanei disponibili come servizio gestito.

Vantaggi dell'adozione di questa best practice: Tenendo in considerazione i costi quando effettui le selezioni ti consentirà di abilitare altri investimenti.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

Ottimizzazione dei componenti del carico di lavoro per ridurre i costi: dimensiona correttamente i componenti del carico di lavoro e consenti l'elasticità per ridurre i costi e massimizzare l'efficienza dei componenti. Determina quali componenti del carico di lavoro possono essere sostituiti con servizi completamente gestiti, laddove appropriato, ad esempio database gestiti, cache in memoria e proxy inversi.

Risorse

Documenti correlati:

- Centro di progettazione AWS
- AWS Partner Network
- Portfolio di soluzioni AWS
- Portale del sapere AWS
- AWS Compute Optimizer

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- La mia architettura
- Ottimizzare le prestazioni e i costi dell'elaborazione AWS (CMP323-R1)

Esempi correlati:

- Esempi di AWS

- Esempi di SDK AWS
- Rightsizing with Compute Optimizer and Memory utilization enabled (Dimensionamento corretto con Compute Optimizer e l'utilizzo della memoria abilitati)
- AWS Compute Optimizer Demo code (Codice dimostrativo di AWS Compute Optimizer)

PERF01-BP04 Utilizzo di policy o architetture di riferimento

Massimizza le prestazioni e l'efficienza valutando le policy interne e le architetture di riferimento esistenti e sfrutta la tua analisi per selezionare servizi e configurazioni per il carico di lavoro.

Anti-pattern comuni:

- Consenti l'utilizzo di una vasta gamma di tecnologie che possono influire sui costi di gestione della tua azienda.

Vantaggi dell'adozione di questa best practice: La definizione di una policy per la scelta dell'architettura, della tecnologia e del fornitore consentirà di prendere decisioni rapidamente.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

Distribuzione del carico di lavoro utilizzando policy esistenti o architetture di riferimento: integra i servizi nella distribuzione cloud, quindi utilizza i test delle prestazioni per assicurarti di continuare a soddisfare i requisiti prestazionali.

Risorse

Documenti correlati:

- Centro di progettazione AWS
- AWS Partner Network
- Portfolio di soluzioni AWS
- Knowledge Center di AWS

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- La mia architettura

Esempi correlati:

- Esempi di AWS
- Esempi di SDK AWS

PERF01-BP05 Utilizzo delle linee guida del fornitore di servizi cloud o di un partner appropriato

Utilizza le risorse del fornitore di servizi cloud, come solutions architect, servizi professionali o un partner appropriato per orientare le tue decisioni. Queste risorse possono aiutarti a rivedere e migliorare l'architettura per ottenere prestazioni ottimali.

Contatta AWS per ricevere assistenza quando ti occorrono ulteriori indicazioni o informazioni sui prodotti. Gli AWS Solutions Architect e AWS Professional Services forniscono linee guida per l'implementazione della soluzione. I Partner AWS mettono a disposizione la propria competenza su AWS per aiutarti ad assicurare alla tua azienda agilità ed innovazione.

Anti-pattern comuni:

- AWS viene utilizzato come provider di data center comune.
- I servizi AWS vengono utilizzati in modo diverso rispetto alla loro progettazione iniziale.

Vantaggi dell'adozione di questa best practice: Grazie alla consulenza con il tuo fornitore o partner potrai prendere decisioni con fiducia.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

Contatta le risorse AWS per assistenza: i Solutions Architect AWS e gli AWS Professional Services forniscono indicazioni per l'implementazione delle soluzioni. I partner APN mettono a disposizione la propria conoscenza di AWS per aiutarti ad assicurare alla tua azienda agilità e innovazione.

Risorse

Documenti correlati:

- Centro di progettazione AWS
- AWS Partner Network
- Portfolio di soluzioni AWS

- Knowledge Center di AWS

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- La mia architettura

Esempi correlati:

- Esempi di AWS
- Esempi di SDK AWS

PERF01-BP06 Benchmarking dei carichi di lavoro esistenti

Esegui il benchmarking delle prestazioni di un carico di lavoro esistente per comprendere le sue prestazioni sul cloud. Utilizza i dati raccolti da questi benchmark per orientare le decisioni architetturali.

Utilizza test sintetici di benchmarking e monitoraggio degli utenti reali per generare dati sulle prestazioni dei componenti durante il carico di lavoro. Di solito, i benchmark sono più rapidi da configurare rispetto ai test di carico e vengono utilizzati per valutare la tecnologia di un componente specifico. Il benchmarking viene spesso utilizzato all'inizio di un nuovo progetto, quando non è ancora disponibile una soluzione completa da sottoporre a test di carico.

Puoi creare benchmark personalizzati, oppure utilizzare un test standard di settore, come TPC-DS , per confrontare i carichi di lavoro del data warehousing. I benchmark di settore sono utili quando devi confrontare ambienti diversi. Quelli personalizzati, invece, sono indicati per analizzare tipi specifici di operazioni che prevedi di eseguire nell'architettura.

In fase di benchmarking, è importante effettuare delle operazioni preliminari sull'ambiente di test al fine di garantire la validità dei risultati. Devi eseguire lo stesso benchmark più volte, per assicurarti di avere acquisito ogni eventuale variazione nel corso del tempo.

Dal momento che, di solito, l'esecuzione dei benchmark è più rapida di quella dei test di carico, il benchmarking può essere utilizzato sin dalle prime fasi della pipeline di distribuzione, così da fornire al team feedback più rapidi sulle deviazioni delle prestazioni. Quando valuti un cambiamento significativo in un componente o servizio, i benchmark possono essere un modo rapido per verificare se l'impegno necessario per apportare la modifica sia giustificato. L'utilizzo del benchmarking in

combinazione con i test di carico è importante perché questi ultimi forniscono indicazioni sulle prestazioni del carico di lavoro in fase di produzione.

Anti-pattern comuni:

- Fai affidamento su valori di riferimento comuni che non sono indicativi delle caratteristiche del carico di lavoro.
- L'unico punto di riferimento è dato dal feedback e dalle percezioni dei clienti.

Vantaggi dell'adozione di questa best practice: Il benchmarking dell'implementazione corrente ti consente di misurare il miglioramento delle prestazioni.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

Monitoraggio delle prestazioni durante lo sviluppo: implementa processi che forniscono visibilità sulle prestazioni durante l'evoluzione del carico di lavoro.

Integrazione nella pipeline di distribuzione: esegui automaticamente test di carico nella pipeline di distribuzione. Confronta i risultati dei test con soglie e KPI predefiniti per assicurarti di poter continuare a soddisfare i requisiti delle prestazioni.

Test dei percorsi utente: utilizza versioni sintetiche o purificate dei dati di produzione, rimuovendo le informazioni sensibili o che permettono l'identificazione degli utenti, per eseguire i test di carico. Verifica l'intera architettura utilizzando percorsi utente riprodotti o già programmati su tutta l'applicazione in larga scala.

Monitoraggio degli utenti reali: utilizza il RUM CloudWatch come aiuto per raccogliere e visualizzare i dati lato cliente riguardanti le prestazioni dell'applicazione. Utilizza quindi i dati per definire i benchmark prestazionali per gli utenti reali.

Risorse

Documenti correlati:

- Centro di progettazione AWS
- AWS Partner Network
- Portfolio di soluzioni AWS
- Knowledge Center di AWS

- RUM Amazon CloudWatch
- Amazon CloudWatch Synthetics

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- La mia architettura
- Optimize applications through Amazon CloudWatch RUM (Ottimizzazione delle applicazioni tramite il RUM Amazon CloudWatch)
- Demo di Amazon CloudWatch Synthetics

Esempi correlati:

- Esempi di AWS
- Esempi di SDK AWS
- Test di carico distribuito
- Measure page load time with Amazon CloudWatch Synthetics (Misurare il tempo di caricamento della pagina con Amazon CloudWatch Synthetics)
- Amazon CloudWatch RUM Web Client (Client Web RUM Amazon CloudWatch)

PERF01-BP07 Esecuzione di un test di carico sul carico di lavoro

Distribuisci l'architettura del carico di lavoro più recente nel cloud utilizzando tipologie e dimensioni di risorse diverse. Monitora la distribuzione per acquisire parametri delle prestazioni che identificano colli di bottiglia o capacità in eccesso. Utilizza queste informazioni sulle prestazioni per progettare o migliorare la tua architettura e la selezione delle risorse.

I test di carico utilizzano il carico di lavoro effettivo per consentirti di osservare le prestazioni dell'intera soluzione in un ambiente di produzione. Occorre eseguire i test di carico tramite versioni sintetiche o purificate dei dati di produzione (rimuovendo le informazioni sensibili o che permettono l'identificazione degli utenti). Utilizza percorsi utente riprodotti o già programmati su tutto il carico di lavoro su vasta scala verificando l'intera architettura. Esegui automaticamente test di carico come parte della pipeline di distribuzione e confronta i risultati con KPI e soglie predefiniti. In questo, modo puoi continuare a raggiungere le prestazioni richieste.

Anti-pattern comuni:

- Vengono testate le singole parti del carico di lavoro, ma non l'intero carico di lavoro.
- Il test di carico viene eseguito su un'infrastruttura diversa dall'ambiente di produzione.
- Esegui i test di carico solo per il carico previsto e non oltre, per prevedere dove si potrebbero riscontrare problemi futuri.
- Esegui test di carico senza informare AWS Support; il test viene, quindi, bloccato perché appare come un evento Denial of Service.

Vantaggi dell'adozione di questa best practice: Misurando le prestazioni in un test di carico, potrai vedere dove avrà luogo l'impatto con l'aumento del carico. In questo modo puoi anticipare le modifiche necessarie prima che influiscano sul carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

Convalida dell'approccio con test di carico: esegui test di carico di un proof-of-concept per scoprire se l'approccio soddisfa i tuoi requisiti prestazionali. Puoi utilizzare i servizi AWS per eseguire ambienti in ambito di produzione e sottoporre l'architettura a test. Dal momento che paghi l'ambiente di test solo quando ti serve, puoi effettuare test su scala completa a un costo estremamente ridotto rispetto all'uso di un ambiente in locale.

Parametri di monitoraggio: Amazon CloudWatch può raccogliere i parametri per tutte le risorse dell'architettura. Puoi anche raccogliere e pubblicare parametri personalizzati per ottenere parametri aziendali o derivati. Utilizza CloudWatch o soluzioni di terze parti per impostare allarmi che indicano il superamento delle soglie.

Test su scala: i test di carico utilizzano il carico di lavoro effettivo, così da poter osservare le prestazioni dell'intera soluzione in un ambiente di produzione. Puoi utilizzare i servizi AWS per eseguire ambienti in ambito di produzione e sottoporre l'architettura a test. Dal momento che paghi l'ambiente di test solo quando ti serve, puoi effettuare test su scala completa a un costo inferiore rispetto all'uso di un ambiente in locale. Sfrutta i vantaggi offerti dal Cloud AWS per testare il carico di lavoro e scoprire dove la scalabilità non è possibile o se non è lineare. Ad esempio, usa le istanze Spot per generare carichi a costi ridotti e rilevare i colli di bottiglia prima che si verifichino in produzione.

Risorse

Documenti correlati:

- AWS CloudFormation
- Building AWS CloudFormation Templates using CloudFormer (Creazione di modelli AWS CloudFormation tramite CloudFormer)
- Usare Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Test di carico distribuito in AWS

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- Optimize applications through Amazon CloudWatch RUM (Ottimizzazione delle applicazioni tramite Amazon CloudWatch RUM)
- Demo di Amazon CloudWatch Synthetics

Esempi correlati:

- Test di carico distribuito in AWS

PERF 2 In che modo selezioni la tua soluzione di calcolo?

La soluzione di calcolo ottimale per un determinato carico di lavoro varia in base alla progettazione dell'applicazione, ai modelli di utilizzo e alle impostazioni di configurazione. Le architetture possono utilizzare diverse soluzioni di elaborazione per vari componenti e consentire funzioni diverse per migliorare le prestazioni. Selezionare la soluzione di calcolo sbagliata per un'architettura può portare a una riduzione dell'efficienza delle prestazioni.

Best practice

- PERF02-BP01 Valutazione delle opzioni di calcolo disponibili
- PERF02-BP02 Identificazione delle opzioni di configurazione di calcolo disponibili
- PERF02-BP03 Raccolta dei parametri relativi al calcolo
- PERF02-BP04 Definizione della configurazione richiesta in base al corretto dimensionamento
- PERF02-BP05 Utilizzo dell'elasticità disponibile delle risorse
- PERF02-BP06 Rivalutazione delle esigenze di elaborazione sulla base dei parametri

PERF02-BP01 Valutazione delle opzioni di calcolo disponibili

Comprendi in che modo l'utilizzo di diverse opzioni di calcolo, come istanze, container e funzioni, porta vantaggi al tuo carico di lavoro.

Risultato desiderato: comprendendo tutte le opzioni di calcolo disponibili, saprai quali sono le opportunità per migliorare le prestazioni, ma anche per ridurre i costi infrastrutturali inutili e l'impegno operativo necessario per mantenere il carico di lavoro. Puoi anche accelerare l'immissione sul mercato quando distribuisci nuovi servizi e funzionalità.

Anti-pattern comuni:

- Utilizzo, in un carico di lavoro post-migrazione, della stessa soluzione di calcolo utilizzata on-premise.
- Poca conoscenza delle soluzioni di cloud computing e di come queste migliorino le prestazioni di calcolo.
- Dimensionamento in eccesso della soluzione di calcolo per soddisfare i requisiti di dimensionamento o prestazioni, quando il passaggio a una nuova soluzione di calcolo soddisferebbe le caratteristiche del tuo carico di lavoro in modo più preciso.

Vantaggi dell'adozione di questa best practice: L'identificazione dei requisiti di calcolo e la valutazione delle soluzioni di calcolo disponibili consente a stakeholder e team di progettazione di comprendere i benefici e i limiti che l'utilizzo della soluzione scelta comporta. La soluzione di calcolo scelta deve soddisfare i criteri prestazionali del carico di lavoro. I criteri essenziali includono le esigenze di calcolo, gli schemi di traffico, gli schemi di accesso ai dati e i requisiti di latenza.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Analizza e comprendi le soluzioni di virtualizzazione, containerizzazione e gestione che portano benefici al carico di lavoro e soddisfano i tuoi requisiti prestazionali. Un carico di lavoro può contenere più tipi di soluzioni di calcolo, ognuna contraddistinta da caratteristiche diverse. In base alla dimensione del carico di lavoro e ai requisiti di calcolo, è possibile selezionare e configurare una soluzione di calcolo adatta alle tue esigenze. Il Cloud Architect deve conoscere i vantaggi e gli svantaggi legati a istanze, container e funzioni. I passaggi seguenti aiutano a capire come selezionare la soluzione di calcolo più adatta per le caratteristiche del tuo carico di lavoro e i tuoi requisiti prestazionali.

Tipo	Server	Container	Funzione
Servizio AWS	Le istanze dei server virtuali Amazon Elastic Compute Cloud (Amazon EC2)	Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	AWS Lambda
Caratteristiche chiave	Dispone di opzioni dedicate per requisiti di licenza hardware, opzioni di collocazione e una vasta gamma di famiglie di istanze diverse basate su parametri di calcolo	Implementazione facile, ambienti coerenti, esecuzione su istanze EC2, scalabilità	Runtime rapido (meno di 15 minuti), memoria e CPU massime non ai livelli di altri servizi, livello hardware gestito, scalabilità fino a milioni di richieste simultanee
Casi d'uso comuni	Migrazioni con rehosting (lift and shift), applicazione monolitica, ambienti ibridi, applicazioni aziendali	Microservizi, ambienti ibridi,	Microservizi, applicazioni basate su eventi

Passaggi dell'implementazione:

1. Seleziona la posizione in cui deve risiedere la soluzione di calcolo valutando the section called “PERF05-BP06 Scelta della posizione del carico di lavoro in base ai requisiti di rete”. La posizione comporterà dei limiti ai tipi di soluzione di calcolo disponibili.
2. Identifica il tipo di soluzione di calcolo che può funzionare con i requisiti della posizione e dell'applicazione
   a. Le istanze dei server virtuali Amazon Elastic Compute Cloud (Amazon EC2) sono disponibili in una vasta gamma di famiglie e dimensioni. Offrono un'ampia scelta di funzionalità, fra cui SSD

e GPU. Le istanze EC2 presentano la più grande flessibilità nella scelta dell'istanza. Quando avvii un'istanza EC2, il tipo specificato determina l'hardware dell'istanza stessa. Ciascun tipo di istanza offre varie capacità di elaborazione, memoria e storage. I tipi di istanza sono raggruppati in famiglie di istanze in base alle loro funzionalità. Sono casi d'uso tipici: l'esecuzione di applicazioni aziendali, il calcolo ad alte prestazioni (HPC), l'addestramento e l'implementazione di applicazioni di machine learning e l'esecuzione di applicazioni native per il cloud.

b. Amazon Elastic Container Service (Amazon ECS) è un servizio di orchestrazione dei container completamente gestito che consente di eseguire e gestire automaticamente i container su un cluster di istanze EC2 o istanze serverless utilizzando AWS Fargate. Puoi utilizzare Amazon ECS con altri servizi come Amazon Route 53, Secrets Manager, AWS Identity and Access Management (IAM) e Amazon CloudWatch. È consigliabile utilizzare Amazon ECS se la tua è un'applicazione containerizzata e il tuo team di progettazione preferisce i container Docker.

c. Amazon Elastic Kubernetes Service (Amazon EKS) è un servizio Kubernetes completamente gestito. Puoi scegliere di eseguire i cluster EKS utilizzando AWS Fargate, eliminando, così, la necessità di effettuare il provisioning e di gestire i server. La gestione di Amazon EKS è semplificata grazie alle integrazioni con servizi AWS come Amazon CloudWatch, gruppi con scalabilità automatica, AWS Identity and Access Management (IAM) e Amazon Virtual Private Cloud (VPC). Con i container devi utilizzare parametri di calcolo per selezionare il tipo più idoneo per il tuo carico di lavoro, proprio come li utilizzi per selezionare i tipi di istanza EC2 o AWS Fargate. Amazon EKS è consigliato se l'applicazione è containerizzata e i team di progettazione preferiscono i container Docker.

d. Puoi utilizzare AWS Lambda per eseguire codice a supporto delle opzioni consentite per runtime, memoria e CPU. È sufficiente caricare il codice: AWS Lambda gestirà tutto il necessario per eseguire e dimensionare il codice. È possibile configurare il codice perché si attivi automaticamente da altri servizi AWS o richiamarlo direttamente. Lambda è consigliato per architetture a microservizi con esecuzione breve e sviluppate per il cloud.

3. Dopo aver sperimentato la tua nuova soluzione di calcolo, pianifica la migrazione e convalida i parametri prestazionali. Si tratta di un processo continuo: consulta the section called “PERF02-BP04 Definizione della configurazione richiesta in base al corretto dimensionamento”.

Livello di impegno per il piano di implementazione: In caso di spostamento del carico di lavoro da una soluzione di calcolo a un'altra, può essere richiesto un moderato livello di impegno per riprogettare l'applicazione.

Risorse

Documenti correlati:

- Elaborazione in cloud con AWS
- Tipi di istanze EC2
- Controllo dello stato del processore per l'istanza EC2
- Container EKS: nodi worker EKS
- Container Amazon ECS: Istanze di container di Amazon ECS
- Funzioni: configurazione della funzione Lambda
- Prescriptive Guidance for Containers (Guida prescrittiva per i container)
- Prescriptive Guidance for Serverless (Guida prescrittiva per serverless)

Video correlati:

- How to choose compute option for startups (Come scegliere un'opzione di calcolo per le start-up)
- Ottimizzare le prestazioni e i costi dell'elaborazione AWS (CMP323-R1)
- Amazon EC2 foundations (CMP211-R2)
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Distribuire le inferenze ML ad alte prestazioni con AWS Inferentia (CMP324-R1)
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (Calcolo migliore, più veloce, più economico: ottimizzazione dei costi di Amazon EC2) (CMP202-R1)

Esempi correlati:

- Migrazione di un'applicazione Web verso i container
- Esecuzione di un "Hello, World!" serverless

PERF02-BP02 Identificazione delle opzioni di configurazione di calcolo disponibili

Ogni soluzione di calcolo offre opzioni e configurazioni utili per supportare le caratteristiche del tuo carico di lavoro. Scopri in che modo le varie opzioni completano il tuo carico di lavoro e quali opzioni di configurazione sono le migliori per la tua applicazione. Esempi di tali opzioni includono la famiglia di istanze, le dimensioni, le caratteristiche (GPU, I/O), il bursting, i timeout, le dimensioni delle funzioni, le istanze di container e la simultaneità.

Risultato desiderato: le caratteristiche del carico di lavoro, tra cui CPU, velocità di trasmissione effettiva della rete, GPU, IOPS, schemi di traffico e di accesso ai dati, vengono documentate e utilizzate per configurare la soluzione di calcolo più adatta a tali caratteristiche. Tutti questi parametri, più altri personalizzati specifici del carico di lavoro, vengono registrati, monitorati e, quindi, utilizzati per ottimizzare la configurazione di calcolo perché risponda in modo ideale ai requisiti.

Anti-pattern comuni:

- Utilizzo della stessa soluzione di calcolo utilizzata on-premise.
- Mancata analisi delle opzioni di calcolo o della famiglia di istanze perché soddisfino le caratteristiche del carico di lavoro.
- Sovradimensionamento del calcolo per garantire capacità di bursting.
- Utilizzi più piattaforme di gestione del calcolo per lo stesso carico di lavoro.

Vantaggi dell'adozione di questa best practice: Acquisisci familiarità con le offerte di calcolo AWS per poter stabilire qual è la soluzione corretta per ogni carico di lavoro. Dopo aver selezionato le offerte di calcolo per il carico di lavoro, puoi verificare rapidamente se soddisfano in modo ideale le relative esigenze. Una soluzione di calcolo ottimizzata per soddisfare le caratteristiche del carico di lavoro migliorerà le prestazioni, ridurrà i costi e garantirà maggiore affidabilità.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Se per il carico di lavoro è stata utilizzata la stessa opzione di calcolo per oltre quattro settimane e sai già che le caratteristiche resteranno uguali in futuro, puoi utilizzare AWS Compute Optimizer che ti fornirà un suggerimento in base alle caratteristiche di calcolo emerse. Se non puoi avvalerti di AWS Compute Optimizer per mancanza di parametri, perché il tipo di istanza non è supportato o perché prevedi cambiamenti nelle caratteristiche, dovrai eseguire una previsione dei parametri in base ai test di carico e alla sperimentazione.

Passaggi dell'implementazione:

1. Il carico di lavoro è in esecuzione su istanze o container EC2 con tipo di avvio EC2?
   a. Per il tuo carico di lavoro è possibile utilizzare GPU per migliorare le prestazioni?
      i. Le istanze di calcolo accelerate sono istanze basate su GPU che forniscono le migliori prestazioni per addestramento di machine learning, inferenza e calcolo ad alte prestazioni.
   b. Nel carico di lavoro sono in esecuzione applicazioni di inferenza di machine learning?

i. AWS Inferentia (Inf1) - Le istanze Inf1 sono create per supportare le applicazioni di inferenza di machine learning. Utilizzando le istanze Inf1, i clienti possono eseguire applicazioni di inferenza di machine learning su larga scala, tra cui riconoscimento di immagini, riconoscimento vocale, elaborazione del linguaggio naturale, personalizzazione e rilevamento di attività fraudolente. Puoi creare un modello in uno dei framework di machine learning più utilizzati, ad esempio TensorFlow, PyTorch o MXNet, e utilizzare istanze GPU per addestrare il modello. Dopo avere addestrato il modello di machine learning per soddisfare i requisiti, puoi distribuirlo su istanze Inf1 utilizzando AWS Neuron, un Software Development Kit (SDK) specializzato composto da un compilatore, runtime e strumenti di profilatura che ottimizzano le prestazioni di inferenza di machine learning dei chip Inferentia.

c. Il tuo carico di lavoro si integra con l'hardware a basso livello per migliorare le prestazioni?

   i. Field Programmable Gate Arrays (FPGA) - Gli FPGA ti consentono di ottimizzare i carichi di lavoro tramite l'esecuzione personalizzata con accelerazione hardware per quelli più impegnativi. Puoi definire gli algoritmi sfruttando i linguaggi di programmazione generale come C o Go oppure linguaggi orientati all'hardware come Verilog o VHDL.

d. Disponi di almeno quattro settimane di parametri e puoi già prevedere che gli schemi di traffico e i parametri resteranno pressoché invariati in futuro?

   i. Utilizzo Compute Optimizer per avere un suggerimento di machine learning riguardo alla configurazione più adatta alle tue caratteristiche di calcolo.

e. Le prestazioni del tuo carico di lavoro sono limitate dai parametri della CPU?

   i. Le istanze ottimizzate per il calcolo sono ideali per i carichi di lavoro che richiedono processori a elevate prestazioni.

f. Le prestazioni del tuo carico di lavoro sono limitate dai parametri della memoria?

   i. Le istanze ottimizzate per la memoria offrono grandi quantità di memoria per carichi di lavoro intensivi in questo senso.

g. Le prestazioni del tuo carico di lavoro sono limitate da IOPS?

   i. Le istanze ottimizzate per l'archiviazione sono progettate per carichi di lavoro che richiedono un accesso frequente e sequenziale in lettura e scrittura (IOPS) all'archiviazione locale.

h. Le caratteristiche del tuo carico di lavoro rappresentano un insieme equilibrato di esigenze relativo a tutti i parametri?

   i. La CPU del tuo carico di lavoro deve eseguire il bursting per gestire picchi di traffico?

      A. Le istanze a prestazioni espandibili sono simili alle istanze ottimizzate per il calcolo, ma offrono anche possibilità di espansione oltre la linea di base per la CPU identificata in un'istanza ottimizzata per il calcolo.

    ii. Le istanze per uso generico offrono un bilanciamento di tutte le caratteristiche a supporto di molti carichi di lavoro diversi.

i. La tua istanza di calcolo è in esecuzione su Linux e limitata dalla velocità di trasmissione effettiva della rete sulla scheda dell'interfaccia di rete?

    i. Revisione la domanda sulle prestazioni n. 5, best practice 2: Valutazione delle funzionalità di rete disponibili per trovare il tipo e la famiglia di istanze ideali per soddisfare le tue esigenze prestazionali.

j. Il tuo carico di lavoro ha bisogno di istanze coerenti e prevedibili in una specifica zona di disponibilità per cui puoi impegnarti per un anno?

    i. le istanze riservate confermano le prenotazioni della capacità in una specifica zona di disponibilità. Queste istanze sono ideali per chi necessita di potenza di calcolo in una specifica zona di disponibilità.

k. Il tuo carico di lavoro include licenze che richiedono hardware dedicato?

    i. Host dedicati supportano le licenze software esistenti e ti aiutano a soddisfare i requisiti di conformità.

l. La tua soluzione di calcolo si espande e richiede elaborazione sincrona?

    i. Le istanze on demand ti consentono di utilizzare la capacità di calcolo su base oraria o al secondo, senza impegni a lungo termine, e sono ideali per il bursting oltre le esigenze di base per le prestazioni.

m. La tua soluzione di calcolo è stateless, tollerante ai guasti e asincrona?

    i. Istanze Spot ti consentono di sfruttare la capacità non utilizzata nell'istanza per i carichi di lavoro stateless e tolleranti ai guasti.

2. Esegui container su Fargate?

    a. Le prestazioni della tua attività sono limitate dalla memoria o dalla CPU?

        i. Utilizza lo strumento per i suggerimenti sulla dimensione dell'attività per regolare la memoria o la CPU.

    b. Le prestazioni sono influenzate da picchi nello schema di traffico?

        i. Utilizza lo strumento per i suggerimenti Auto Scaling per trovare la configurazione adatta ai tuoi schemi di traffico.

3. La tua soluzione di calcolo è su Lambda?

    a. Disponi di almeno quattro settimane di parametri e puoi già prevedere che gli schemi di traffico e i parametri resteranno pressoché invariati in futuro?

   i. Utilizzo Compute Optimizer per avere un suggerimento di machine learning riguardo alla configurazione più adatta alle tue caratteristiche di calcolo.

b. Non hai parametri a sufficienza per utilizzare AWS Compute Optimizer?
   i. Se non hai parametri a sufficienza per utilizzare Compute Optimizer, scegli AWS Lambda Power Tuning per selezionare la configurazione migliore.

c. Le prestazioni della tua funzione sono limitate dalla memoria o dalla CPU?
   i. Configura la memoria Lambda perché soddisfi i parametri relativi alle tue esigenze prestazionali.

d. La tua funzione va in timeout durante l'esecuzione?
   i. Modifica le impostazioni di timeout.

e. Le prestazioni della tua funzione sono limitate da picchi di attività e simultaneità?
   i. Configura le impostazioni di simultaneità perché soddisfino le tue esigenze prestazionali.

f. La tua funzione è in esecuzione asincrona e riscontra errori nei nuovi tentativi?
   i. Configura l'età massima dell'evento e il limite massimo di nuovi tentativi nelle impostazioni della configurazione asincrona .

Livello di impegno per il piano di implementazione:

Per attuare questa best practice è necessario conoscere caratteristiche e parametri attuali del calcolo. Raccogliere tali parametri, definire una linea di base e, quindi, utilizzare i parametri per identificare l'opzione ideale per il calcolo richiede un livello di impegno da basso a moderato . La convalida migliore passa attraverso test di carico e sperimentazioni.

Risorse

Documenti correlati:

- Elaborazione in cloud con AWS
- AWS Compute Optimizer
- Tipi di istanze EC2
- Controllo dello stato del processore per l'istanza EC2
- Container EKS: nodi worker EKS
- Container Amazon ECS: Istanze di container di Amazon ECS
- Funzioni: configurazione della funzione Lambda

Video correlati:

- Amazon EC2 foundations (CMP211-R2)
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- Ottimizzare le prestazioni e i costi dell'elaborazione AWS (CMP323-R1)

Esempi correlati:

- Rightsizing with Compute Optimizer and Memory utilization enabled (Dimensionamento corretto con Compute Optimizer e l'utilizzo della memoria abilitati)
- AWS Compute Optimizer Demo code (Codice dimostrativo di AWS Compute Optimizer)

PERF02-BP03 Raccolta dei parametri relativi al calcolo

Per comprendere le prestazioni delle tue risorse di calcolo, è necessario registrare e tracciare l'utilizzo di vari sistemi. Questi dati possono essere utilizzati per determinare in modo più accurato i requisiti delle risorse.

I carichi di lavoro possono generare grandi volumi di dati quali parametri, registri ed eventi. Stabilisci se il servizio di archiviazione, monitoraggio e osservazione esistente è in grado di gestire i dati generati. Identifica i parametri che rispecchiano l'utilizzo delle risorse e che è possibile raccogliere, aggregare e correlare su un'unica piattaforma. Tali parametri devono essere rappresentativi di tutte le risorse, le applicazioni e i servizi del carico di lavoro; potrai così avere facilmente una visione globale del sistema e identificare con rapidità le opportunità di miglioramento delle prestazioni e i problemi.

Risultato desiderato: tutti i parametri relativi alle risorse di calcolo vengono identificati, raccolti, aggregati e correlati su un'unica piattaforma con conservazione implementata per supportare gli obiettivi operativi e di costi.

Anti-pattern comuni:

- Utilizzi solo i file di log manuali per la ricerca dei parametri.
- Pubblichi i parametri solo negli strumenti interni.
- Utilizzi solo i parametri predefiniti registrati dal software di monitoraggio selezionato.
- Rivedi i parametri solo quando c'è un problema.

Vantaggi dell'adozione di questa best practice: Per monitorare le prestazioni dei carichi di lavoro, devi registrare più parametri delle prestazioni in un periodo di tempo. Tali parametri ti consentono di rilevare anomalie nelle prestazioni e ti aiutano a valutare le prestazioni rispetto ai parametri aziendali, così da garantire il rispetto delle esigenze del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

Identifica, raccogli, aggrega e correla i parametri relativi al calcolo. L'utilizzo di servizi come Amazon CloudWatch possono rendere più rapida e facile da gestire l'implementazione. Oltre ai parametri predefiniti registrati, identifica e tieni traccia di altri parametri a livello di sistema nel tuo carico di lavoro. Registra dati quali utilizzo della CPU, memoria, I/O del disco e parametri di rete in entrata e in uscita per ottenere informazioni più dettagliate sui livelli di utilizzo o colli di bottiglia. Questi dati sono cruciali per capire come si comporta il carico di lavoro e come viene utilizzata la soluzione di calcolo. Utilizza tali parametri come parte di un approccio basato sui dati per ottimizzare e ottimizzare le risorse del tuo carico di lavoro.

Passaggi dell'implementazione:

1. Quali sono i parametri della soluzione di calcolo che è importante monitorare?
    a. Parametri predefiniti di EC2
    b. Parametri predefiniti di Amazon ECS
    c. Parametri predefiniti di EKS
    d. Parametri predefiniti di Lambda
    e. Parametri di EC2 relativi a memoria e disco
2. Attualmente dispongo di una soluzione di registrazione e monitoraggio approvata?
    a. Amazon CloudWatch
    b. AWS Distro for OpenTelemetry
    c. Amazon Managed Service for Prometheus
3. Ho identificato e configurato le policy di conservazione dei dati in modo che corrispondano ai miei obiettivi operativi e di sicurezza?
    a. Conservazione dei dati predefinita per i parametri CloudWatch
    b. Conservazione dei dati predefinita per CloudWatch Logs
4. Come distribuisci gli agenti per l'aggregazione di parametri e registri?
    a. Automazione di AWS Systems Manager

b. OpenTelemetry Collector

Livello di impegno per il piano di implementazione: è richiesto un livello di impegno medio per identificare, monitorare, raccogliere, aggregare e correlare i parametri di tutte le risorse di calcolo.

Risorse

Documenti correlati:

- Documentazione di Amazon CloudWatch
- Raccolta di parametri e registri da istanze Amazon EC2 e da server on-premise con l'agente di CloudWatch
- Accesso a Amazon CloudWatch Logs per AWS Lambda
- Utilizzo di CloudWatch Logs con istanze di container
- Pubblicazione di parametri personalizzati
- AWS Answers: Centralized Logging (AWS Answers: registrazione centralizzata)
- Servizi AWS che pubblicano parametri CloudWatch
- Monitoraggio di Amazon EKS su AWS Fargate

Video correlati:

- Application Performance Management on AWS (Gestione delle prestazioni delle applicazioni su AWS)
- Creazione di un piano di monitoraggio

Esempi correlati:

- Level 100: Monitoring with CloudWatch Dashboards (Livello 100: Monitoraggio con i pannelli di controllo CloudWatch)
- Level 100: Monitoring Windows EC2 instance with CloudWatch Dashboards (Monitoraggio dell'istanza EC2 di Windows con i pannelli di controllo CloudWatch)
- Level 100: Monitoring an Amazon Linux EC2 instance with CloudWatch Dashboards (Monitoraggio dell'istanza EC2 di Amazon Linux con i pannelli di controllo CloudWatch)

PERF02-BP04 Definizione della configurazione richiesta in base al corretto dimensionamento

Analizza le varie caratteristiche di prestazione del tuo carico di lavoro e come queste sono correlate a memoria, rete e utilizzo della CPU. Utilizza questi dati per scegliere le risorse che meglio corrispondono al profilo del tuo carico di lavoro. Ad esempio, un carico di lavoro a memoria elevata, come un database, potrebbe essere servito meglio dalla famiglia di istanze r. Al contrario, un carico di lavoro con picchi di prestazioni può trarre maggiori vantaggi da un sistema di container elastici.

Anti-pattern comuni:

- Scegli l'istanza più grande disponibile per tutti i carichi di lavoro.
- Esegui la standardizzazione di tutti i tipi di istanze in un solo tipo per semplificare la gestione.

Vantaggi dell'adozione di questa best practice: Acquisire familiarità con le offerte di calcolo AWS ti consente di determinare la soluzione corretta per i vari carichi di lavoro. Dopo aver selezionato le varie offerte di calcolo per il carico di lavoro, puoi verificare rapidamente quali soddisfano le esigenze del tuo carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

Modifica della configurazione del carico di lavoro con un corretto dimensionamento: al fine di ottimizzare le prestazioni e l'efficienza complessiva, determina le risorse necessarie per il tuo carico di lavoro. Seleziona le istanze ottimizzate per la memoria per i sistemi che richiedono più memoria rispetto alla CPU, oppure istanze ottimizzate per il calcolo per i componenti che effettuano l'elaborazione dati che non necessita di molta memoria. Il dimensionamento consente al tuo carico di lavoro di funzionare nel miglior modo possibile utilizzando solo le risorse di cui necessita.

Risorse

Documenti correlati:

- AWS Compute Optimizer
- Elaborazione in cloud con AWS
- Tipi di istanze EC2
- Container ECS: Istanze di container di Amazon ECS
- Container EKS: nodi worker EKS

- Funzioni: configurazione della funzione Lambda
- Controllo dello stato del processore per l'istanza EC2

Video correlati:

- Amazon EC2 foundations (CMP211-R2)
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (Calcolo migliore, più veloce, più economico: ottimizzazione dei costi di Amazon EC2) (CMP202-R1)
- Distribuire le inferenze ML ad alte prestazioni con AWS Inferentia (CMP324-R1)
- Ottimizzare le prestazioni e i costi dell'elaborazione AWS (CMP323-R1)
- Powering next-gen Amazon EC2: Deep dive into the Nitro system
- How to choose compute option for startups (Come scegliere un'opzione di calcolo per le start-up)
- Ottimizzare le prestazioni e i costi dell'elaborazione AWS (CMP323-R1)

Esempi correlati:

- Rightsizing with Compute Optimizer and Memory utilization enabled (Dimensionamento corretto con Compute Optimizer e l'utilizzo della memoria abilitati)
- AWS Compute Optimizer Demo code (Codice dimostrativo di AWS Compute Optimizer)

PERF02-BP05 Utilizzo dell'elasticità disponibile delle risorse

Il cloud offre la flessibilità necessaria per espandere o ridurre le risorse in modo dinamico attraverso una serie di meccanismi per soddisfare i cambiamenti della domanda. Se combinato con parametri relativi all'elaborazione, un carico di lavoro può rispondere automaticamente a questi cambiamenti e utilizzare la gamma di risorse più opportuna per raggiungere il suo obiettivo.

La corrispondenza ottimale tra risorse fornite e domanda determina il costo più basso per il sistema. Tuttavia, sarà anche necessario tenere conto del tempo di provisioning e di eventuali errori delle singole risorse e pianificare risorse sufficienti di conseguenza. La domanda può essere fissa o variabile e richiede parametri e automazione, per garantire che la gestione non diventi troppo complicata e onerosa dal punto di vista economico.

Con AWS puoi adottare varie strategie per associare l'offerta alla domanda. Il whitepaper sul principio dell'ottimizzazione dei costi descrive come utilizzare i seguenti approcci ai costi:

- Approccio basato sulla domanda
- Approccio basato sui buffer
- Approccio basato sul tempo

Assicurati che le distribuzioni dei carichi di lavoro siano in grado di gestire eventi di dimensionamento. Crea scenari di test per eventi di ridimensionamento per garantire che il carico di lavoro si comporti come previsto.

Anti-pattern comuni:

- Reagisci agli allarmi e aumenti manualmente la capacità.
- Dopo un evento di dimensionamento, lasci una capacità aumentata anziché ridurre il dimensionamento.

Vantaggi dell'adozione di questa best practice: Configurare e testare l'elasticità del carico di lavoro è utile per risparmiare denaro, mantenere i benchmark prestazionali e migliorare l'affidabilità al variare del traffico. La maggior parte delle istanze non di produzione deve essere arrestata quando le istanze non vengono utilizzate. Anche se è possibile arrestare manualmente le istanze inutilizzate, questa operazione è impraticabile su più grandi scale. Puoi inoltre sfruttare l'elasticità basata sul volume, che ti consente di ottimizzare le prestazioni e i costi aumentando automaticamente il numero di istanze di calcolo durante i picchi di domanda e riducendo la capacità quando la domanda diminuisce.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

Utilizzo ottimale dell'elasticità: l'elasticità corrisponde all'offerta di risorse disponibili rispetto alla relativa domanda. Istanze, container e funzioni forniscono tutti meccanismi di elasticità, in combinazione con il ridimensionamento automatico o sotto forma di funzionalità del servizio. Utilizza l'elasticità all'interno della tua architettura per assicurarti di disporre della capacità sufficiente per rispettare i requisiti di prestazione a tutti i livelli di utilizzo. Verifica che i parametri per il dimensionamento delle risorse elastiche siano convalidati rispetto al tipo di carico di lavoro distribuito. Se distribuisci un'applicazione di transcodifica video, è previsto il 100% di utilizzo della CPU e non deve essere il parametro principale. In alternativa, puoi misurare la profondità della coda dei processi di transcodifica in attesa di dimensionare i tipi di istanze. Assicurati che le distribuzioni dei carichi di lavoro siano in grado di gestire eventi di dimensionamento. Ridimensionare in modo sicuro i componenti del carico di lavoro è fondamentale quanto ricalibrare le risorse quando richiesto dalla

domanda. Crea scenari di test per eventi di ridimensionamento per garantire che il carico di lavoro si comporti come previsto.

Risorse

Documenti correlati:

- Elaborazione in cloud con AWS
- Tipi di istanze EC2
- Container ECS: Istanze di container di Amazon ECS
- Container EKS: nodi worker EKS
- Funzioni: configurazione della funzione Lambda
- Controllo dello stato del processore per l'istanza EC2

Video correlati:

- Amazon EC2 foundations (CMP211-R2)
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (Calcolo migliore, più veloce, più economico: ottimizzazione dei costi di Amazon EC2) (CMP202-R1)
- Distribuire le inferenze ML ad alte prestazioni con AWS Inferentia (CMP324-R1)
- Ottimizzare le prestazioni e i costi dell'elaborazione AWS (CMP323-R1)
- Powering next-gen Amazon EC2: Deep dive into the Nitro system

Esempi correlati:

- Esempi di gruppo di Amazon EC2 Auto Scaling
- Esercitazioni su Amazon EFS

PERF02-BP06 Rivalutazione delle esigenze di elaborazione sulla base dei parametri

Utilizza i parametri a livello di sistema per identificare il comportamento e i requisiti del tuo carico di lavoro nel tempo. Valuta le esigenze del tuo carico di lavoro confrontando le risorse disponibili con tali requisiti e apporta modifiche al tuo ambiente di elaborazione per soddisfare al meglio il profilo del carico di lavoro. Ad esempio, nel corso del tempo si potrebbe osservare che un sistema utilizza molta più memoria di quanto si pensasse inizialmente, e trasferirlo a una famiglia o una dimensione di istanze diversa potrebbe migliorarne sia le prestazioni sia l'efficienza.

Anti-pattern comuni:

- Monitori solamente i parametri a livello di sistema per ottenere informazioni approfondite sul carico di lavoro.
- Pianifichi le tue esigenze di calcolo in base ai requisiti di picco del carico di lavoro.
- Dimensioni in eccesso la tua soluzione di calcolo per soddisfare i requisiti di dimensionamento o prestazioni quando il passaggio a una nuova soluzione di calcolo soddisferebbe le caratteristiche del tuo carico di lavoro

Vantaggi dell'adozione di questa best practice: Per ottimizzare le prestazioni e l'utilizzo delle risorse, è necessario disporre di una vista operativa unificata, di dati granulari in tempo reale e di un riferimento storico. Puoi creare pannelli di controllo automatici per visualizzare questi dati ed eseguire calcoli matematici sui parametri per ottenere informazioni operative e di utilizzo.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

Utilizzo di un approccio basato sui dati per ottimizzare le risorse: al fine di ottenere il massimo delle prestazioni e dell'efficienza, utilizza i dati raccolti nel tempo dal tuo carico di lavoro per mettere a punto e ottimizzare le risorse. Osserva le tendenze nell'utilizzo del tuo carico di lavoro delle risorse attuali e determina in che punto puoi effettuare cambiamenti per soddisfare al meglio le esigenze del tuo carico di lavoro. Quando le risorse sono sovra-impegnate, le prestazioni del sistema diminuiscono, sebbene il sottoutilizzo delle risorse le renda meno efficienti e i costi siano più elevati.

Risorse

Documenti correlati:

- Elaborazione in cloud con AWS
- AWS Compute Optimizer
- Elaborazione in cloud con AWS
- Tipi di istanze EC2
- Container ECS: Istanze di container di Amazon ECS
- Container EKS: nodi worker EKS
- Funzioni: configurazione della funzione Lambda
- Controllo dello stato del processore per l'istanza EC2

Video correlati:

- Amazon EC2 foundations (CMP211-R2)
- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (Calcolo migliore, più veloce, più economico: ottimizzazione dei costi di Amazon EC2) (CMP202-R1)
- Distribuire le inferenze ML ad alte prestazioni con AWS Inferentia (CMP324-R1)
- Ottimizzare le prestazioni e i costi dell'elaborazione AWS (CMP323-R1)
- Powering next-gen Amazon EC2: Deep dive into the Nitro system

Esempi correlati:

- Rightsizing with Compute Optimizer and Memory utilization enabled (Dimensionamento corretto con Compute Optimizer e l'utilizzo della memoria abilitati)
- AWS Compute Optimizer Demo code (Codice dimostrativo di AWS Compute Optimizer)

PERF 3 In che modo selezioni la soluzione di storage?

La soluzione di storage ottimale per un sistema varia in base a fattori quali: tipo di metodo di accesso (blocco, file od oggetto), schemi di accesso (casuali o sequenziali), throughput necessario, frequenza di accesso (online, offline, archivio), frequenza di aggiornamento (WORM, dinamico) e vincoli di disponibilità e durata. I sistemi Well-Architected utilizzano più soluzioni di storage e consentono funzionalità diverse per migliorare le prestazioni e utilizzare le risorse in modo efficiente.

Best practice

- PERF03-BP01 Identificazione delle caratteristiche e dei requisiti di archiviazione
- PERF03-BP02 Valutazione delle opzioni di configurazione disponibili
- PERF03-BP03 Decisioni basate su schemi e parametri di accesso

PERF03-BP01 Identificazione delle caratteristiche e dei requisiti di archiviazione

Identifica e documenta i requisiti di archiviazione dei carichi di lavoro e definisci le caratteristiche di archiviazione di ciascuna posizione. Le caratteristiche di archiviazione includono, ad esempio, accesso condivisibile, dimensioni dei file, tasso di crescita, velocità di trasmissione effettiva, IOPS, latenza, schemi di accesso e persistenza dei dati. Utilizza queste caratteristiche per valutare se i servizi di archiviazione di blocchi, file, oggetti o istanze rappresentano la soluzione più efficiente per gli specifici requisiti di archiviazione.

Risultato desiderato: individua e documenta i requisiti di archiviazione e valuta le soluzioni di archiviazione disponibili. In base alle caratteristiche di archiviazione chiave, il team saprà valutare in che modo i servizi di archiviazione selezionati ottimizzeranno le prestazioni dei carichi di lavoro. I criteri chiave includono gli schemi di accesso ai dati, il tasso di crescita, i requisiti di dimensionamento e i requisiti di latenza.

Anti-pattern comuni:

- Utilizzi un unico tipo di archiviazione, ad esempio Amazon Elastic Block Store (Amazon EBS), per tutti i carichi di lavoro.
- Ritieni che tutti i carichi di lavoro abbiano requisiti di prestazioni di accesso allo storage simili.

Vantaggi dell'adozione di questa best practice: la selezione della soluzione di archiviazione in base alle caratteristiche individuate e richieste contribuirà a migliorare le prestazioni dei carichi di lavoro, ridurre i costi e limitare lo sforzo operativo necessario alla gestione del carico di lavoro. Le prestazioni dei carichi di lavoro trarranno benefici dalla soluzione, configurazione e posizione del servizio di archiviazione.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

Identifica i parametri delle prestazioni di storage più importanti del carico di lavoro e implementa i miglioramenti come parte di un approccio basato sui dati, ricorrendo a benchmark o test di carico. Utilizza tali dati per identificare i punti in cui la soluzione di storage è vincolante ed esamina le opzioni di configurazione per migliorare la soluzione. Determina il tasso di crescita previsto per il carico di lavoro e scegli una soluzione di storage che soddisfi tali percentuali. Cerca le offerte di archiviazione AWS per determinare la soluzione di archiviazione più adatta per i vari requisiti a livello di carichi di lavoro. Il provisioning delle soluzioni di archiviazione in AWS aumenta le opportunità che hai di verificare le offerte di archiviazione e verificare se sono in linea con i requisiti dei carichi di lavoro.

Servizio AWS	Caratteristiche chiave	Casi d'uso comuni
Amazon S3	Durabilità del 99,999999999%, crescita illimitata, accessibilità indipendente dalla posizione, svariati modelli di costi basati su accesso e resilienza	Dati applicativi nativi del cloud, archiviazione e backup dei dati, analisi, data lake, hosting di siti web statici, dati IoT

Servizio AWS	Caratteristiche chiave	Casi d'uso comuni
Amazon S3 Glacier	Latenza da secondi a ore, crescita illimitata, costi più bassi, archiviazione a lungo termine	Archiviazione dei dati, archivi di file multimediali, conservazione a lungo termine dei backup.
Amazon EBS	Le dimensioni di archiviazione richiedono la gestione e il monitoraggio, una bassa latenza, l'archiviazione persistente, una durabilità dal 99,8% al 99,9%; la maggior parte dei tipi di volume è accessibile solo da un'istanza EC2.	Applicazioni COTS, applicazioni ad alta intensità di I/O, database relazionali e NoSQL, backup e ripristino
Archivio dell'istanza EC2	Dimensioni di archiviazione predefinite, latenza minima, non persistente, accessibile solo da un'istanza EC2	Applicazioni COTS, applicazioni ad alta intensità di I/O, datastore in memoria
Amazon EFS	Durabilità del 99,999999999%, crescita illimitata, accessibile da più servizi di calcolo	Applicazioni moderne che condividono file tra servizi di calcolo, archiviazione di file per il dimensionamento dei sistemi di gestione dei contenuti
Amazon FSx	Supporta 4 file system (NetApp, OpenZFS, Windows File Server e Amazon FSx for Lustre), archiviazione disponibile in base al file system specifico, accessibile da più servizi di calcolo	Carichi di lavoro nativi del cloud, espansione del cloud privato, carichi di lavoro migrati che richiedono un file system specifico, VMC, sistemi ERP, archiviazione e backup di file on-premise

Servizio AWS	Caratteristiche chiave	Casi d'uso comuni
AWS Snow Family	Dispositivi portatili, crittografia a 256 bit, endpoint NFS, calcolo su scheda, TB di archiviazione	Migrazione dei dati nel cloud, archiviazione, calcolo in condizioni on-premis e estreme, ripristino di emergenza, raccolta di dati remoti
AWS Storage Gateway	Fornisce accesso on-premis e a bassa latenza all'archiviazione supportata dal cloud, cache on-premise completamente gestita	Migrazioni di dati on-premis e nel cloud, popolamento dei data lake nel cloud da origini on-premise, condivisione avanzata dei file

Passaggi dell'implementazione:

1. Utilizza test di benchmarking o caricamento per raccogliere le caratteristiche chiave dei requisiti di archiviazione. Le caratteristiche chiave includono:
   a. Condivisibilità (quali componenti accedono a questo tipo di archiviazione)
   b. Tasso di crescita
   c. Throughput
   d. Latenza
   e. Dimensione I/O
   f. Durabilità
   g. Schemi di accesso (letture/scritture, frequenza, con picchi o costante)
2. Individua il tipo di soluzione di archiviazione che supporta le caratteristiche di archiviazione desiderate.
   a. Amazon S3 è un servizio di archiviazione di oggetti con scalabilità illimitata, elevata disponibilità e più opzioni di accessibilità. Il trasferimento di oggetto e l'accesso a oggetti in Amazon S3 possono utilizzare un servizio, ad esempio Transfer Acceleration oppure Access Points , per supportare la posizione, i requisiti di sicurezza e gli schemi di accesso. Utilizza le linee guida relative alle prestazioni di Amazon S3 per ottimizzare la configurazione di Amazon S3 e garantire la conformità ai requisiti relativi alle prestazioni dei carichi di lavoro.

b. Amazon S3 Glacier è una classe di archiviazione di Amazon S3 creata per l'archiviazione dei dati. Puoi scegliere tra tre diverse soluzioni di archiviazione, a partire da un accesso in millisecondi a un accesso di 5-12 ore con opzioni di costi e sicurezza diverse. Amazon S3 Glacier ti consente di rispettare i requisiti relativi alle prestazioni mediante l'implementazione di un ciclo di vita dei dati che supporta i requisiti aziendali e le caratteristiche dei dati.

c. Amazon Elastic Block Store (Amazon EBS) è un servizio di archiviazione a blocchi a elevate prestazioni progettato per Amazon Elastic Compute Cloud (Amazon EC2). Puoi scegliere tra soluzioni basate su SSD o HDD con caratteristiche diverse che danno priorità agli IOPS oppure alla velocità di trasmissione effettiva. I volumi EBS sono adatti per carichi di lavoro a prestazioni elevate, archiviazione principale per file system, database o applicazioni che possono accedere solo a sistemi di staging collegati.

d. L'archivio dell'istanza Amazon EC2 è simile ad Amazon EBS in quanto si collega a un'istanza Amazon EC2. Tuttavia, l'archivio dell'istanza è solo un'archiviazione temporanea che deve essere idealmente utilizzata come buffer, cache o altro contenuto temporaneo. Non è possibile scollegare un archivio dell'istanza in quanto tutti i dati andranno perduti in caso di arresto dell'istanza stessa. Gli archivi dell'istanza possono essere utilizzati per casi d'uso basati su prestazioni di I/O elevate e bassa latenza, in cui i dati non devono essere persistenti.

e. Amazon Elastic File System (Amazon EFS) è un file system montabile a cui possono accedere più tipi di soluzioni di calcolo. Amazon EFS aumenta o riduce automaticamente lo spazio di archiviazione, mentre le relative prestazioni sono ottimizzate in modo da offrire costantemente latenze basse. EFS dispone di due modalità di configurazione delle prestazioni, ovveroGeneral Purpose (Uso generico) e Max I/O (I/O max). La modalità General Purpose (Uso generico) è caratterizzata da una latenza di lettura con valori inferiori a millisecondi e da una latenza di scrittura con valori nell'ordine di millisecondi a una cifra. La funzionalità Max I/O (I/O max) può supportare migliaia di istanze di calcolo che richiedono un file system condiviso. Amazon EFS supporta due modalità di velocità di trasmissione effettiva, ovveroBursting (Espansione) e Provisioned (Allocato). Un carico di lavoro caratterizzato da uno schema di accesso a picchi trarrà vantaggio dalla modalità Bursting (Espansione), mentre un carico di lavoro costantemente elevato sarà più performante con la modalità Provisioned (Allocato).

f. Amazon FSx è stato sviluppato con le più recenti soluzioni di calcolo AWS per supportare i 4 file system più comunemente utilizzati: NetApp ONTAP, OpenZFS, Windows File Server e Lustre. Relativamente ad Amazon FSx , la latenza, la velocità di trasmissione effettiva e le operazioni di input/output al secondo (IOPS) variano a seconda del file system; è necessario considerare attentamente questi elementi quando si deve selezionare il file system in modo conforme ai requisiti dei carichi di lavoro.

   g. AWS Snow Family sono dispositivi di archiviazione e calcolo che supportano la migrazione di dati online e offline nel cloud, nonché l'archiviazione e il calcolo dei dati on-premise. I dispositivi AWS Snow supportano la raccolta e l'elaborazione di grandi quantità di dati on-premise, nonché il loro spostamento nel cloud. Sono disponibili numerose best practice documentate sulle prestazioni relative a numero, dimensioni e compressione dei file.
   h. AWS Storage Gateway fornisce alle applicazioni on-premise l'accesso all'archiviazione basata sul cloud. AWS Storage Gateway supporta più servizi di archiviazione cloud, tra cui Amazon S3, Amazon S3 Glacier, Amazon FSx e Amazon EBS. Supporta più protocolli, ad esempio iSCSI, SMB e NFS. Fornisce prestazioni a bassa latenza mediante la memorizzazione nella cache dei dati on-premise con accesso frequente e invia solo i dati modificati e compressi ad AWS.
3. Dopo aver sperimentato la nuova soluzione di archiviazione e avere individuato la configurazione ottimale, pianifica la migrazione e convalida i parametri prestazionali. Questo sarà un processo continuativo, che dovrà essere riesaminato al variare delle caratteristiche chiave o delle opzioni o dei servizi disponibili.

Livello di impegno per il piano di implementazione: in caso di spostamento del carico di lavoro da una soluzione di archiviazione a un'altra, può essere richiesto un moderato livello di impegno per riprogettare l'applicazione.

Risorse

Documenti correlati:

- Tipi di volume di Amazon EBS
- Storage Amazon EC2
- Amazon EFS: prestazioni di Amazon EFS
- Prestazioni di Amazon FSx for Lustre
- Amazon FSx for Windows File Server Performance (Prestazioni di Amazon FSx for Windows File Server)
- Amazon FSx for NetApp ONTAP performance (Prestazioni di Amazon FSx for NetApp ONTAP)
- Amazon FSx for OpenZFS performance (Prestazioni di Amazon FSx for OpenZFS)
- Amazon S3 Glacier: Amazon S3 Glacier Documentation (Documentazione di Amazon S3 Glacier)
- Amazon S3: considerazioni su velocità e prestazioni delle richieste
- Storage cloud con AWS
- AWS Snow Family

- Caratteristiche di I/O di EBS

Video correlati:

- Analisi approfondita di Amazon EBS (STG303-R1)
- Ottimizzazione delle prestazioni di storage con Amazon S3 (STG343)

Esempi correlati:

- Driver CSI di Amazon EFS
- Driver CSI di Amazon EBS
- Utility di Amazon EFS
- Amazon EBS Autoscale
- Esempi di Amazon S3
- Amazon FSx for Lustre Container Storage Interface (CSI) Driver (Driver CSI [Container Storage Interface] per Amazon FSx for Lustre)

PERF03-BP02 Valutazione delle opzioni di configurazione disponibili

Valuta le varie caratteristiche e opzioni di configurazione e il modo in cui sono correlate allo storage. Comprendi dove e come utilizzare Provisioned IOPS, SSD, storage magnetico, storage a oggetti, storage di archiviazione o storage temporaneo per ottimizzare lo spazio di storage e le prestazioni del tuo carico di lavoro.

Amazon EBS offre una gamma di opzioni che ti permettono di ottimizzare le prestazioni di storage e i costi del tuo carico di lavoro. Tali opzioni sono suddivise in due categorie principali: storage basato su SSD per i carichi di lavoro relativi alle transazioni, come database e volumi di avvio (le prestazioni dipendono principalmente dagli IOPS), e storage basato su HDD per i carichi di lavoro con elevati requisiti di throughput, come MapReduce e l'elaborazione dei log (le prestazioni dipendono principalmente dalla velocità in MB/s).

I volumi con SSD includono SSD con la capacità di IOPS allocata più elevata per carichi di lavoro transazionali sensibili alla latenza, e SSD generici che bilanciano prezzo e prestazioni per un'ampia gamma di dati transazionali.

Amazon S3 Transfer Acceleration consente il trasferimento rapido dei file su lunghe distanze tra il client e il tuo bucket S3. Transfer Acceleration sfrutta le edge location di Amazon CloudFront

distribuite a livello globale per instradare i dati attraverso percorsi di rete ottimizzati. Per i carichi di lavoro in un bucket S3 con richieste GET intensive, utilizza Amazon S3 con CloudFront. Quando si caricano file di grandi dimensioni, è possibile utilizzare il caricamento simultaneo di più parti per ottimizzare il throughput di rete.

Amazon Elastic File System (Amazon EFS) fornisce un file system NFS elastico semplice, scalabile e completamente gestito, da utilizzare in combinazione con i servizi Cloud AWS e le risorse on-premise. Per supportare un'ampia gamma di carichi di lavoro di storage nel cloud, Amazon EFS offre due modalità prestazionali: modalità di prestazioni generiche e modalità di prestazioni I/O massime. Sono disponibili anche due modalità di velocità di trasmissione effettiva tra cui scegliere per il file system: Bursting Throughput e Provisioned Throughput. Per determinare quali impostazioni utilizzare per il carico di lavoro, consulta la Guida per l'utente di Amazon EFS.

Amazon FSx offre quattro file system tra cui scegliere: Amazon FSx for Windows File Server per carichi di lavoro aziendali, Amazon FSx for Lustre per carichi di lavoro ad alte prestazioni, Amazon FSx for NetApp ONTAP per i diffusi file system ONTAP di NetApps e Amazon FSx for OpenZFS per server di file basati su Linux. FSx è basato su SSD ed è progettato per offrire prestazioni rapide, prevedibili, scalabili e costanti. I file system di Amazon FSx offrono elevate velocità di lettura e scrittura e l'accesso costante ai dati a bassa latenza. È possibile scegliere il livello di throughput desiderato per soddisfare le esigenze del carico di lavoro.

Anti-pattern comuni:

- Utilizzi un solo tipo di storage, ad esempio Amazon EBS, per tutti i carichi di lavoro.
- Utilizzi la capacità di IOPS allocata per tutti i carichi di lavoro senza test reali su tutti i livelli di archiviazione.
- Ritieni che tutti i carichi di lavoro abbiano requisiti di prestazioni di accesso allo storage simili.

Vantaggi dell'adozione di questa best practice: La valutazione di tutte le opzioni del servizio di storage può ridurre il costo dell'infrastruttura e l'impegno necessario per mantenere i carichi di lavoro. Può potenzialmente accelerare l'immissione sul mercato per la distribuzione di nuovi servizi e funzionalità.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

Identificazione delle caratteristiche dell'archiviazione: quando valuti una soluzione di archiviazione, determina quali caratteristiche sono necessarie, come la possibilità di condivisione, le dimensioni dei

file e della cache, la latenza, la velocità di trasmissione effettiva e la persistenza dei dati. Quindi, in base ai tuoi requisiti, scegli il servizio AWS più adatto alle tue esigenze.

Risorse

Documenti correlati:

- Storage cloud con AWS
- Tipi di volume di Amazon EBS
- Storage Amazon EC2
- Amazon EFS: prestazioni di Amazon EFS
- Prestazioni di Amazon FSx for Lustre
- Prestazioni di Amazon FSx for Windows File Server
- Amazon Glacier: documentazione di Amazon Glacier
- Amazon S3: considerazioni su velocità e prestazioni delle richieste
- Storage cloud con AWS
- Storage cloud con AWS
- Caratteristiche di I/O di EBS

Video correlati:

- Analisi approfondita di Amazon EBS (STG303-R1)
- Ottimizzazione delle prestazioni di storage con Amazon S3 (STG343)

Esempi correlati:

- Driver CSI di Amazon EFS
- Driver CSI di Amazon EBS
- Utility di Amazon EFS
- Amazon EBS Autoscale
- Esempi di Amazon S3

PERF03-BP03 Decisioni basate su schemi e parametri di accesso

Scegli i sistemi di storage in base agli schemi di accesso del carico di lavoro e configurali determinando il modo in cui il carico di lavoro accede ai dati. Aumenta l'efficienza dello storage scegliendo lo storage di oggetti anziché lo storage a blocchi. Configura le opzioni di storage in funzione dei tuoi schemi di accesso ai dati.

Il modo in cui accedi ai dati influisce sulle prestazioni della soluzione di storage. Seleziona la soluzione più adatta ai tuoi schemi di accesso. In alternativa, puoi modificarli affinché siano in linea con la soluzione di storage, allo scopo di ottimizzare le prestazioni.

Creare un array in configurazione RAID 0 ti permette di ottenere prestazioni più elevate per i file system di cui puoi effettuare il provisioning su un unico volume. Prendi in considerazione l'uso di RAID 0 quando le prestazioni I/O sono più importanti della tolleranza agli errori. Ad esempio, puoi usarlo con un database che viene utilizzato in modo intensivo e in cui la replica dei dati è già stata configurata separatamente.

Seleziona i parametri di storage appropriati per il carico di lavoro tra tutte le opzioni di storage utilizzate per il carico di lavoro. Quando utilizzi i file system che utilizzano crediti di burst, puoi creare degli allarmi che ti informano quando stai per avvicinarti ai limiti di credito. È necessario creare pannelli di controllo di storage per visualizzare lo stato generale dello storage del carico di lavoro.

Per i sistemi di storage di dimensioni fisse, come Amazon EBS o Amazon FSx, assicurati di monitorare la quantità di archiviazione utilizzata rispetto alle dimensioni complessive dell'archiviazione e di creare, se possibile, un'automazione per aumentarne le dimensioni quando si raggiunge una soglia

Anti-pattern comuni:

- Ritieni che le prestazioni di storage siano adeguate se i clienti non si lamentano.
- Utilizzi un solo livello di storage, presupponendo che tutti i carichi di lavoro rientrino in tale livello.

Vantaggi dell'adozione di questa best practice: Sono necessari una vista operativa unificata, dati granulari in tempo reale e un riferimento storico per ottimizzare le prestazioni e l'utilizzo delle risorse. Puoi creare pannelli di controllo e dati automatici con granularità di un secondo per eseguire calcoli parametrici sui dati e ottenere informazioni operative e di utilizzo per le tue esigenze di archiviazione.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

Ottimizzazione dell'utilizzo dell'archiviazione e degli schemi di accesso: scegli i sistemi di archiviazione sulla base degli schemi di accesso del tuo carico di lavoro e confrontando le caratteristiche delle opzioni di archiviazione disponibili. Determina quale sia il miglior posto per archiviare dati e che ti consentirà di rispettare i requisiti riducendo nel contempo il sovraccarico. Utilizza ottimizzazioni delle prestazioni e schemi di accesso mentre configuri e interagisci con i dati sulla base delle caratteristiche del tuo storage (ad esempio, separazione dei volumi o partizionamento dei dati).

Selezione dei parametri appropriati per le opzioni di archiviazione: assicurati di selezionare i parametri di rete adeguati al tuo carico di lavoro. Ogni opzione di storage offre vari parametri per monitorare le prestazioni del carico di lavoro nel corso del tempo. Assicurati di effettuare le misurazioni a fronte di qualsiasi parametro burst dell'archiviazione (ad esempio, il monitoraggio dei crediti burst per Amazon EFS). Per i sistemi di archiviazione a dimensione fissa, come Amazon Elastic Block Store o Amazon FSx, assicurati di monitorare la quantità di archiviazione utilizzata rispetto alle rispettive dimensioni complessive. Crea automazione, laddove possibile, per aumentare le dimensioni dello storage quando si raggiunge una soglia.

Parametri di monitoraggio: Amazon CloudWatch può raccogliere i parametri per tutte le risorse dell'architettura. Puoi anche raccogliere e pubblicare parametri personalizzati per ottenere parametri aziendali o derivati. Utilizza CloudWatch o soluzioni di terze parti per impostare allarmi che indicano il superamento delle soglie.

Risorse

Documenti correlati:

- Tipi di volume di Amazon EBS
- Storage Amazon EC2
- Amazon EFS: prestazioni di Amazon EFS
- Prestazioni di Amazon FSx for Lustre
- Prestazioni di Amazon FSx for Windows File Server
- Amazon Glacier: documentazione di Amazon Glacier
- Amazon S3: considerazioni su velocità e prestazioni delle richieste
- Storage cloud con AWS
- Caratteristiche di I/O di EBS

- Monitoring and understanding Amazon EBS performance using Amazon CloudWatch (Monitorare e comprendere le prestazioni di Amazon EBS tramite Amazon CloudWatch)

Video correlati:

- Analisi approfondita di Amazon EBS (STG303-R1)
- Ottimizzazione delle prestazioni di storage con Amazon S3 (STG343)

Esempi correlati:

- Driver CSI di Amazon EFS
- Driver CSI di Amazon EBS
- Utility di Amazon EFS
- Amazon EBS Autoscale
- Esempi di Amazon S3

PERF 4 In che modo selezioni la soluzione di database?

La soluzione di database ottimale per un determinato sistema può variare in base ai requisiti di disponibilità, coerenza, tolleranza della partizione, latenza, durata, scalabilità e capacità di query. Molti sistemi utilizzano diverse soluzioni di database per vari sottosistemi e consentono funzionalità differenti per migliorare le prestazioni. Selezionare la soluzione e le funzionalità del database sbagliate per un sistema può ridurre l'efficienza delle prestazioni.

Best practice

- PERF04-BP01 Comprensione delle caratteristiche dei dati
- PERF04-BP02 Valutazione delle opzioni disponibili
- PERF04-BP03 Raccolta e registrazione dei parametri delle prestazioni del database
- PERF04-BP04 Scelta dello spazio di archiviazione dei dati in base ai modelli di accesso
- PERF04-BP05 Ottimizzazione dello spazio di archiviazione dei dati in base ai modelli e ai parametri di accesso

PERF04-BP01 Comprensione delle caratteristiche dei dati

Scegli le soluzioni di gestione dei dati perché coincidano in modo ottimale con le caratteristiche, gli schemi di accesso e i requisiti dei set di dati del carico di lavoro. Nel selezionare e implementare una soluzione di gestione dei dati, è necessario assicurarsi che le caratteristiche relative a query, dimensionamento e archiviazione siano adeguate ai requisiti dei dati del carico di lavoro. Scopriamo come si abbinano le diverse opzioni di database e i modelli di dati, nonché quali opzioni di configurazione sono più adatte al tuo caso d'uso.

AWS offre diversi motori di database dedicati, tra cui database relazionali, a chiave-valore, di documento, in memoria, a grafo, di serie temporali e di libro mastro. Ogni soluzione di gestione dei dati offre soluzioni e configurazioni adatte a gestire i tuoi casi d'uso e modelli di dati. Per il tuo carico di lavoro può essere possibile utilizzare più soluzioni di database diverse in base alle caratteristiche dei dati. Selezionando le soluzioni di database più adatte a uno specifico problema, puoi allontanarti dall'idea di database monolitico, contraddistinta da tutte le limitazioni di approccio universale, e concentrarti sulla gestione dei dati per soddisfare le esigenze del cliente.

Risultato desiderato: le caratteristiche dei dati del carico di lavoro sono documentate in modo sufficientemente dettagliato da agevolare la selezione e la configurazione di soluzioni di database di supporto e offrono informazioni approfondite su possibili alternative.

Anti-pattern comuni:

- Non considerare modi per segmentare grandi set di dati in raccolte più piccole con caratteristiche simili, mancando così di cogliere l'opportunità di utilizzare più spesso i database dedicati, che coincidono meglio con le caratteristiche dei dati e della crescita.
- Non identificare in anticipo gli schemi di accesso ai dati, con conseguenti costose e complesse rilavorazioni in seguito.
- Limitare la crescita adottando strategie di archiviazione dei dati il cui dimensionamento non è rapido quanto necessario
- Scegliere un fornitore e un tipo di database per tutti i carichi di lavoro.
- Continuare a utilizzare una soluzione di database per via dell'esperienza e delle competenze interne relative a quel particolare tipo di soluzione.
- Continuare con una soluzione di database perché ha funzionato bene in un ambiente on-premise.

Vantaggi dell'adozione di questa best practice: È utile conoscere tutte le soluzioni di database AWS in modo da determinare la soluzione di database corretta per i vari carichi di lavoro. Dopo aver

selezionato la soluzione di database appropriata per il tuo carico di lavoro, sperimenta rapidamente ciascuna di queste offerte di database per determinare se continuano a soddisfare le esigenze del tuo carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alto

- I potenziali risparmi sui costi possono non essere identificati.
- Sicurezza dei dati non al livello richiesto.
- Accesso ai dati e prestazioni di archiviazione non ottimali.

Guida all'implementazione

Definisci le caratteristiche dei dati e gli schemi di accesso del tuo carico di lavoro. Esamina tutte le soluzioni di database disponibili per identificare quella più adatta ai requisiti dei tuoi dati. Per un dato carico di lavoro possono essere selezionati più database. Considera ogni servizio o gruppo di servizi e valutali singolarmente. Se identifichi possibili alternative nelle soluzioni di gestione per tutti i dati (o parte di essi), sperimenta implementazioni alternative che potrebbero portare benefici in termini di costi, sicurezza, prestazioni e affidabilità. Se viene adottato un nuovo approccio alla gestione dei dati, aggiorna la documentazione esistente.

Tipo	Servizi AWS	Caratteristiche chiave	Casi d'uso comuni
Relazionale	Amazon RDS, Amazon Aurora	Integrità referenziale, transazioni ACID, schema-on-write	ERP, CRM, software COTS (Commercial off-the-shelf)
Chiave-valore	Amazon DynamoDB	Elevata velocità di trasmissione effettiva, bassa latenza, scalabilità praticamente infinita	Carrelli (e-commerce), cataloghi di prodotti, applicazioni chat
Documento	Amazon DocumentDB	Archiviazione documenti JSON e query su qualsiasi attributo	Gestione dei contenuti (CMS), profili clienti, applicazioni per dispositivi mobili

Tipo	Servizi AWS	Caratteristiche chiave	Casi d'uso comuni
In memoria	Amazon ElastiCache, Amazon MemoryDB	Latenza in microsecondi	Caching, classifiche di gioco
Grafo	Amazon Neptune	Dati altamente relazionali in cui le loro relazioni sono significative	Social networks, motori di personalizzazione, rilevamento frodi
Serie temporali	Amazon Timestream	Dati in cui il tempo è la dimensione principale	DevOps, IoT, monitoraggio
Colonna ampia	Amazon Keyspaces	Carichi di lavoro Cassandra	Manutenzione di apparecchiature industriali, ottimizzazione dei percorsi
Di libri mastri	Amazon QLDB	Libro mastro delle modifiche immutabile e verificabile tramite crittografia	Sistemi di registro, assistenza sanitaria, catene di fornitura, istituzioni finanziarie

Passaggi dell'implementazione

1. Come sono strutturati i dati (ad esempio, sono non strutturati, a chiave-valore, semi-strutturati, relazionali)?
   a. Se i dati non sono strutturati, valuta un'archiviazione a oggetti come Amazon S3 o un database NoSQL come Amazon DocumentDB.
   b. Per i dati chiave-valore, valuta DynamoDB, ElastiCache per Redis oppure MemoryDB.
   c. Se i dati hanno una struttura relazionale, qual è il livello di integrità referenziale richiesto?
      i. Per i vincoli di chiave esterna, i database relazionali come Amazon RDS e Aurora possono fornire questo livello di integrità.

   ii. In genere, in un modello di dati NoSQL, i dati vengono denormalizzati in un singolo documento o in una raccolta di documenti da recuperare in un'unica richiesta, anziché essere uniti tra diversi documenti o tabelle.

2. È richiesta la conformità ACID (atomicità, coerenza, isolamento, durabilità)?
   a. Se sono necessarie proprietà ACID associate ai database relazionali, valuta un database relazionale come Amazon RDS e Aurora.
3. Qual è il modello di consistenza richiesto?
   a. Se la tua applicazione può tollerare la consistenza finale, valuta un'implementazione NoSQL. Esamina le altre caratteristiche per scegliere quale database NoSQL è il più appropriato.
   b. Se è necessaria un'elevata coerenza, puoi utilizzare le elevate coerenze di lettura con DynamoDB o un database relazionale come Amazon RDS.
4. Quali formati di query e risultati devono essere supportati (ad esempio SQL, CSV, Parque, Avro, JSON, ecc.)?
5. Quali sono i tipi di dati, le dimensioni dei campi e le quantità complessive presenti (ad esempio testo, numero, spaziale, serie temporali calcolate, binario o BLOB, documento)?
6. Come cambierà nel tempo l'archiviazione? In che modo questo avrà effetto sulla scalabilità?
   a. I database serverless come DynamoDB e Amazon Quantum Ledger Database si dimensioneranno automaticamente fino a uno spazio di archiviazione pressoché illimitato.
   b. Per i database relazionali sono previsti limiti superiori per l'archiviazione assegnata, al raggiungimento dei quali si rende spesso necessario partizionare orizzontalmente tali database tramite meccanismi quali lo sharding.
7. Qual è la proporzione di query in lettura rispetto alle quelle in scrittura? Il caching potrebbe probabilmente migliorare le prestazioni?
   a. I carichi di lavoro con molte operazioni di lettura traggono beneficio da un livello di caching, che può essere rappresentato da ElastiCache oppure DAX se il database è DynamoDB.
   b. È anche possibile passare le operazioni di lettura alle repliche di lettura con database relazionali come Amazon RDS.
8. Hanno priorità più elevata le operazioni di archiviazione e modifica OLTP, Online Transaction Processing) o quelle di recupero e report (OLAP - Online Analytical Processing)?
   a. Per l'elaborazione transazionale ad alta velocità di trasmissione effettiva, valuta un database NoSQL come DynamoDB o Amazon DocumentDB.
   b. Per le query analitiche, valuta un database colonnare come Amazon Redshift o la possibilità di esportare i dati su Amazon S3 ed eseguire l'analisi tramite Athena oppure QuickSight.

9. Quanto sono sensibili questi dati e quale livello di protezione e crittografia richiedono?
    a. Tutti i motori Amazon RDS e Aurora supportano la crittografia dei dati inattivi tramite AWS KMS. Microsoft SQL Server e Oracle supportano anche la tecnologia nativa Transparent Data Encryption (TDE) con l'uso di Amazon RDS.
    b. Per DynamoDB, puoi utilizzare il controllo granulare degli accessi con IAM per controllare chi ha accesso a quali dati a livello di chiave.
10. Che livello di durabilità è necessario per i dati?
    a. Aurora replica automaticamente i dati su tre zone di disponibilità all'interno di una Regione, il che significa che i dati sono altamente durevoli con minori probabilità di perdite.
    b. DynamoDB viene automaticamente replicato in più zone di disponibilità per offrire livelli elevati di disponibilità e durabilità dei dati.
    c. Amazon S3 offre il 99,999999999% di durabilità. Molti servizi di database, come Amazon RDS e DynamoDB, supportano l'esportazione di dati su Amazon S3 per la conservazione e l'archiviazione a lungo termine.
11. I requisiti in termini di Obiettivo del tempo di ripristino (RTO) e Obiettivo del punto di ripristino (RPO) influenzano la soluzione?
    a. Amazon RDS, Aurora, DynamoDB, Amazon DocumentDB e Neptune supportano tutti sia il ripristino point-in-time, sia il backup e il ripristino on-demand.
    b. In caso di requisiti di elevata disponibilità, è possibile replicare le tabelle DynamoDB a livello globale tramite la funzionalità Tabelle globali, mentre i cluster Aurora possono essere replicati su più Regioni grazie alla funzionalità Database globale. Inoltre, è possibile replicare i bucket S3 tra Regioni AWS grazie alla replica fra Regioni.
12. È presente il desiderio di abbandonare i motori di database commerciali/i costi di licenza?
    a. Valuta motori open-source come PostgreSQL e MySQL su Amazon RDS o Aurora.
    b. Sfrutta AWS DMS e AWS SCT per eseguire le migrazioni dai motori di database commerciali a quelli open-source
13. Quali sono le aspettative operative per il database? Il passaggio ai servizi gestiti è una priorità?
    a. Utilizzare Amazon RDS, invece di Amazon EC2, e scegliere DynamoDB o Amazon DocumentDB, invece di ospitare in autonomia un database NoSQL, riduce le spese operative.
14. Come avviene attualmente l'accesso al database? È solo un accesso da applicazione o sono presenti utenti Business Intelligence (BI) e altre applicazioni pronte all'uso connesse?
    a. Se fossero presenti dipendenze verso altri strumenti esterni, potresti dover mantenere la compatibilità con i database che essi supportano. Amazon RDS è completamente compatibile

con le diverse versioni dei motori che supporta, compresi Microsoft SQL Server, Oracle, MySQL e PostgreSQL.

15. Di seguito è riportato un elenco di possibili servizi di gestione dei dati e i loro possibili migliori utilizzi:

    a. I database relazionali memorizzano i dati con relazioni e schemi predefiniti. Questi database sono progettati per supportare le transazioni ACID (atomicità, coerenza, isolamento, durabilità) e per mantenere l'integrità referenziale e una solida coerenza dei dati. Molte applicazioni tradizionali, Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) ed e-commerce utilizzano database relazionali per archiviare i propri dati. Puoi eseguire molti di questi motori di database in Amazon EC2 oppure scegliere tra i servizi AWS di database gestiti: Amazon Aurora, Amazon RDSe Amazon Redshift.

    b. I database chiave-valore sono ottimizzati per schemi di accesso di uso comune, in genere per archiviare e recuperare grandi volumi di dati. Questi database offrono tempi di risposta rapidi, anche nel caso di volumi estremi di richieste simultanee. Le Web app dal traffico elevato, i sistemi di e-commerce e le applicazioni di gaming sono casi d'uso tipici dei database chiave-valore. In AWS, puoi utilizzare Amazon DynamoDB, un database completamente gestito, multi-regione, multi-master e durevole, con capacità integrate di sicurezza, backup e ripristino, oltre al caching in memoria per le applicazioni su scala Internet.

    c. I database in memoria vengono utilizzati per applicazioni che richiedono accesso in tempo reale ai dati, bassissima latenza ed elevatissima velocità di trasmissione effettiva. Archiviando i dati direttamente in memoria, questi database forniscono una latenza di microsecondi alle applicazioni per le quali la latenza di millisecondi non è sufficiente. Puoi utilizzare database in memoria per il caching delle applicazioni, la gestione delle sessioni, l'archiviazione delle sessioni di gioco e le applicazioni geospaziali. Amazon ElastiCache è un datastore in memoria completamente gestito, compatibile con Redis oppure Memcached. Se le applicazioni presentano anche requisiti più elevati in termini di durabilità, Amazon MemoryDB per Redis li offre ed è anche un servizio di database in memoria durevole e con prestazioni ad altissima velocità.

    d. Un database di documenti è progettato per archiviare dati semistrutturati come documenti di tipo JSON. Questi database aiutano gli sviluppatori a creare e aggiornare rapidamente applicazioni quali gestione di contenuti, cataloghi e profili utente. Amazon DocumentDB è un database di documenti veloce, scalabile, ad elevata disponibilità e completamente gestito, che supporta i carichi di lavoro MongoDB.

    e. Uno store colonnare è un tipo di database NoSQL. Utilizza tabelle, righe e colonne, ma a differenza di un database relazionale, i nomi e il formato delle colonne possono variare da riga a

riga all'interno della stessa tabella. In genere, gli store colonnari sono utilizzati nelle applicazioni industriali su larga scala per la manutenzione delle apparecchiature, la gestione delle flotte e l'ottimizzazione dei percorsi. Amazon Keyspaces (per Apache Cassandra) è un servizio di database colonnare gestito compatibile con Apache Cassandra, scalabile e altamente disponibile.

f. I database a grafo vengono implementati con le applicazioni che devono navigare ed eseguire query su milioni di relazioni tra set di dati a grafo altamente connessi, con una latenza misurata in millisecondi su larga scala. Molte aziende utilizzano database a grafo per il rilevamento di attività fraudolente, i social network e i motori di raccomandazione. Amazon Neptune è un servizio di database a grafo veloce, affidabile e completamente gestito che semplifica la creazione e l'esecuzione di applicazioni che funzionano con set di dati altamente connessi.

g. I database di serie temporali sono efficienti per raccogliere, sintetizzare e derivare informazioni approfondite dai dati che cambiano nel tempo. I database di serie temporali sono spesso utilizzati dalle applicazioni IoT, DevOps e dalla telemetria industriale. Amazon Timestream è un servizio di database di serie temporali veloce, scalabile e completamente gestito per le applicazioni IoT ed operative che semplifica la memorizzazione e l'analisi di trilioni di eventi al giorno.

h. I database di libri mastri forniscono un'autorità centralizzata e affidabile per mantenere un registro delle transazioni scalabile, immutabile e verificabile tramite crittografia per ogni applicazione. I database di libri mastri vengono utilizzati per sistemi di record, catena di fornitura, registrazioni e persino transazioni bancarie. Amazon Quantum Ledger Database (Amazon QLDB) è un database di libro mastro completamente gestito che fornisce un registro delle transazioni trasparente, immutabile e verificabile tramite crittografia, di proprietà di un'autorità centrale attendibile. Amazon QLDB tiene traccia di ogni modifica ai dati dell'applicazione e conserva una cronologia completa e verificabile delle modifiche nel corso del tempo.

Livello di impegno per il piano di implementazione: In caso di spostamento del carico di lavoro da una soluzione di database a un'altra, può essere richiesto un elevato livello di impegno per riprogettare i dati e l'applicazione.

Risorse

Documenti correlati:

- Database su cloud AWS
- Memorizzazione nella cache di database AWS

- Amazon DynamoDB Accelerator
- Best practice di Amazon Aurora
- Prestazioni di Amazon Redshift
- 10 suggerimenti prestazionali su Amazon Athena
- Best practice di Amazon Redshift Spectrum
- Best practice di Amazon DynamoDB
- Choose between EC2 and Amazon RDS (Scegliere tra EC2 e Amazon RDS)
- Best practice per l'implementazione di Amazon ElastiCache

Video correlati:

- Database dedicati di AWS (DAT209-L)
- Sfatiamo i miti sullo storage Amazon Aurora: come funziona realmente (DAT309-R)
- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Esempi correlati:

- Optimize Data Pattern Using Amazon Redshift Data Sharing (Ottimizzare lo schema dei dati con la condivisione dei dati di Amazon Redshift)
- Migrazioni dei database
- MS SQL Server - AWS Database Migration Service (DMS) Replication Demo (Demo di replica di AWS Database Migration Service, DMS)
- Database Modernization Hands On Workshop (Workshop pratico sulla modernizzazione dei database)
- Esempi di Amazon Neptune

PERF04-BP02 Valutazione delle opzioni disponibili

Prima di selezionare una soluzione di gestione dei dati, è importante capire come può migliorare le prestazioni e quali sono le opzioni disponibili per il database. Utilizza il test di carico per identificare i parametri del database più importanti per il tuo carico di lavoro. Nella valutazione delle opzioni per il database, tieni in considerazione diversi aspetti, come i gruppi di parametri, le opzioni di archiviazione, la memoria, il calcolo, la replica di lettura, l'eventuale coerenza, il pooling di

connessioni e le opzioni di caching. Esegui prove con le diverse opzioni di configurazione per migliorare i parametri.

Risultato desiderato: un carico di lavoro può avere una o più soluzioni di database in base al tipo di dati. Le funzionalità e i vantaggi del database corrispondono in maniera ottimale alle caratteristiche dei dati, agli schemi di accesso e ai requisiti del carico di lavoro. Per ottimizzare le prestazioni e i costi del database, devi valutare gli schemi di accesso ai dati per determinare le opzioni di database appropriate. Valuta i tempi di query accettabili per accertarti che le opzioni di database selezionate possano rispettare i requisiti.

Anti-pattern comuni:

- Non identificare gli schemi di accesso.
- Non conoscere le opzioni di configurazione della soluzione di gestione dei dati scelta.
- Basarsi soltanto sull'aumento delle dimensioni dell'istanza, senza tenere conto di altre opzioni di configurazione disponibili.
- Non testare le caratteristiche di scalabilità della soluzione scelta.

Vantaggi dell'adozione di questa best practice: Esplorare le opzioni di database e sperimentare con esse può consentire di ridurre il costo dell'infrastruttura, migliorare le prestazioni e la scalabilità e ridurre l'impegno richiesto per mantenere i carichi di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alto

- L'ottimizzazione per un database universale comporta inutili compromessi.
- Costi più elevati dovuti alla mancata corrispondenza fra la configurazione della soluzione di database e gli schemi di traffico.
- Possibili problemi operativi come conseguenza dei problemi di dimensionamento.
- Sicurezza dei dati non al livello richiesto.

Guida all'implementazione

Comprendi le caratteristiche dei dati del carico di lavoro per configurare al meglio le opzioni del database. Esegui test di carico per identificare i parametri prestazionali chiave e i colli di bottiglia. Utilizza queste caratteristiche e questi parametri per valutare le opzioni di database e sperimentare con diverse configurazioni.

Servizi AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Dimensionamento del calcolo	Aumento della dimensione dell'istanza, le istanze Aurora Serverless si dimensionano automaticamente in risposta ai cambiamenti di carico	Dimensionamento automatico in lettura/scrittura con la modalità capacità on demand o dimensionamento automatico della capacità assegnata in lettura/scrittura nella modalità capacità assegnata	Aumento della dimensione dell'istanza	Aumento della dimensione dell'istanza, aggiunta di nodi al cluster	Aumento della dimensione dell'istanza	Dimensionamento automatico per regolare la capacità	Dimensionamento automatico in lettura/scrittura con la modalità capacità on demand o dimensionamento automatico della capacità assegnata in lettura/scrittura nella modalità capacità assegnata	Dimensionamento automatico per regolare la capacità
Dimensionamento orizzontale delle letture	Tutti i motori supportano le repliche	Aumento delle unità di capacità in	Repliche di lettura	Repliche di lettura	Repliche di lettura. Supporta il	Dimensionamento automatico	Aumento delle unità di capacità in	Aumento del dimensionamento in

Servizi AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
	di lettura. Aurora supporta il dimensionamento automatico delle istanze con replica di lettura	lettura assegnate			dimensionamento automatico delle istanze con replica di lettura		lettura assegnate	risposta a limiti di simultaneità documentati

Servizi AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Dimensionamento orizzontale delle scritture	Aumento della dimensione dell'istanza, raggruppamento in batch delle scritture nell'applicazione o aggiunta di una coda davanti al database. Dimensionamento orizzontale tramite partizione a livello dell'applicazione	Aumento delle unità di capacità in scrittura assegnate. Garanzia di una chiave di partizione ottimale per evitare limitazioni in scrittura a livello di partizione	Aumento delle dimensioni dell'istanza principale	Utilizzo di Redis in modalità cluster per distribuire le scritture tra le partizioni	Aumento della dimensione dell'istanza	Le richieste di scrittura potrebbero subire limitazioni durante il dimensionamento. Se riscontri eccezioni di limitazione, continua a inviare dati ad almeno la stessa velocità di trasmissione effettiva per	Aumento delle unità di capacità in scrittura assegnate. Garanzia di una chiave di partizione ottimale per evitare limitazioni in scrittura a livello di partizione	Aumento del dimensionamento in risposta a limiti di simultaneità documentati

Servizi AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
	fra più istanze					attivare il dimensionamento automatico. Scritture in batch per ridurre le richieste di scrittura simultanee		
Configurazione del motore	Gruppi di parametri	Non applicabile	Gruppi di parametri	Gruppi di parametri	Gruppi di parametri	Non applicabile	Non applicabile	Non applicabile

Servizi AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Caching	Caching in memoria, configurabile tramite gruppi di parametri. Associazione a una cache dedicata come ElastiCache per Redis a cui assegnare le richieste per gli elementi con accesso più frequente	Cache DAX completamente gestita disponibile	Caching in memoria. Facoltativo: associazione a una cache dedicata come ElastiCache per Redis a cui assegnare le richieste per gli elementi con accesso più frequente	La funzione principale è il caching	Utilizzo della cache dei risultati delle query per memorizzare in cache il risultato delle query di sola lettura	Timestream dispone di due livelli di archiviazione, uno dei quali è in memoria e ad alte prestazioni	Implementazione di una cache dedicata separata come ElastiCache per Redis a cui assegnare le richieste per gli elementi con accesso più frequente	Non applicabile

Servizi AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Alta disponibilità/ripristino di emergenz	Per i carichi di lavoro, la configurazione consigliata prevede l'esecuzione di un'istanza in stand-by in una seconda zona di disponibilità, al fine di garantire la resilienza all'interno di una Regione. Per la	Disponibilità elevata all'interno di una Regione. Possibilità di replicare le tabelle tra Regioni grazie alle tabelle globali di DynamoD	Creazione di più istanze in diverse zone di disponibilità per una maggiore disponibilità. Possibilità di condividere snapshot tra Regioni e cluster grazie a DMS, per garantire funzionalità di replica tra Regioni/r	La configurazione consigliata per i cluster di produzione prevede la creazione di almeno un nodo in una zona di disponibilità secondaria. Per la replica di cluster tra Regioni è possibile utilizzare	Le repliche di lettura in altre zone di disponibilità servono come destinazioni di failover. È possibile condividere snapshot tra Regioni e replicare cluster tramite i flussi Neptune, che consentono di replicare dati tra	Disponibilità elevata all'interno di una Regione. I replica tra Regioni richiede lo sviluppo di un'applicazione personalizzata tramite SDK Timestream	Disponibilità elevata all'interno di una Regione. La replica tra Regioni richiede una logica applicativa personalizzata o strumenti di terze parti	Disponibilità elevata all'interno di una Regione. Per la replica tra Regioni, esporta i contenuti del registro Amazon QLDB in un bucket S3 e configura tale bucket per la replica tra Regioni.

Servizi AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
	resilienza tra Regioni è possibile utilizzare Aurora Global Database		ipristino di emergenz	ElastiCache Global Datastore.	due cluster in due diverse Regioni.			

Passaggi dell'implementazione

1. Quali opzioni di configurazione sono disponibili per i database selezionati?
    a. I gruppi di parametri per Amazon RDS e Aurora consentono di regolare le impostazioni comuni del database a livello di motore, ad esempio la memoria allocata per la cache o la regolazione del fuso orario del database
    b. Per i servizi di database assegnati, come Amazon RDS, Aurora, Neptune, Amazon DocumentDB e quelli implementati su Amazon EC2, è possibile modificare il tipo di istanza e l'archiviazione assegnata, nonché aggiungere repliche di lettura.
    c. DynamoDB consente di specificare due modalità di capacità: on-demand e assegnata. Per rispondere ai diversi carichi di lavoro, puoi passare da una modalità all'altra e aumentare in qualsiasi momento la capacità allocata nella modalità assegnata.
2. Il carico di lavoro comporta operazioni intensive in lettura o scrittura?
    a. Quali sono le soluzioni disponibili per eliminare il carico delle letture (repliche di lettura, caching, ecc.)?
        i. Per le tabelle DynamoDB, è possibile eliminare il carico delle letture grazie a DAX per il caching.

ii. Per i database relazionali è possibile creare un cluster ElastiCache per Redis e configurare l'applicazione perché legga prima dalla cache, passando poi al database se l'elemento richiesto non è presente.

iii. I database relazionali come Amazon RDS e Aurora, nonché i database assegnati NoSQL come Neptune e Amazon DocumentDB, supportano tutti l'aggiunta di repliche di lettura per eliminare il carico creato dalle parti di lettura nel carico di lavoro.

iv. I database serverless come DynamoDB si dimensionano automaticamente. Assicurati di avere abbastanza unità di capacità di lettura (RCU) assegnate per gestire il carico di lavoro.

b. Quali soluzioni sono disponibili per il dimensionamento delle operazioni in scrittura (sharding della chiave di partizione, introduzione di una coda, ecc.)?

i. Per i database relazionali, è possibile aumentare la dimensione dell'istanza per gestire un maggiore carico di lavoro o aumentare la capacità di IOPS allocata per gestire una maggiore velocità di trasmissione effettiva verso l'archiviazione sottostante.

- È anche possibile introdurre una coda davanti al database, invece di eseguire direttamente la scrittura su di esso. Questo schema consente di disaccoppiare l'acquisizione dal database e controllare il flusso, in modo che il database sia in grado di gestirlo.
- Raggruppare in batch le richieste di scrittura, anziché creare molte transazioni di breve durata, può aiutare a migliorare la velocità di trasmissione effettiva in database relazionali con un elevato volume in scrittura.

ii. I database serverless come DynamoDB possono dimensionare automaticamente la velocità di trasmissione effettiva in scrittura oppure è possibile regolare le unità di capacità in scrittura (WCU) assegnate, a seconda della modalità di capacità.

- È comunque possibile che si verifichino problemi con partizioni ad accesso frequente quando si raggiungono i limiti di velocità di trasmissione effettiva per una data chiave di partizione. Questo problema può essere arginato scegliendo una chiave di partizione con una distribuzione più uniforme o eseguendo lo sharding in lettura della chiave di partizione.

3. Qual è il picco di transazioni per secondo (TPS) attuale o previsto? Esegui un test con questo volume di traffico da solo e con l'aggiunta di un X% per comprendere le caratteristiche di dimensionamento.

a. È possibile utilizzare strumenti nativi come pg_bench for PostgreSQL per mettere sotto stress il database e comprendere dove si creano colli di bottiglia e quali sono le caratteristiche di dimensionamento.

   b. È utile acquisire il traffico in contesti simili alla produzione perché possa essere riprodotto per simulare condizioni reali in aggiunta ai carichi di lavoro sintetici.
4. Se utilizzi capacità di calcolo serverless o a dimensionamento elastico, testa l'impatto del suo dimensionamento sul database. Se necessario, prevedi una gestione o un pooling delle connessioni per ridurre l'impatto sul database.
   a. È possibile utilizzare RDS Proxy con Amazon RDS e Aurora per gestire le connessioni al database.
   b. I database serverless come DynamoDB non hanno connessioni associate, ma valuta la capacità assegnata e le policy di dimensionamento automatico per affrontare i picchi nel carico.
5. Se il carico è prevedibile, sono presenti picchi e periodi di inattività?
   a. In presenza di periodi di inattività, valuta la possibilità di ridurre la capacità assegnata o la dimensione dell'istanza durante questi momenti. Aurora Serverless V2 aumenterà o ridurrà automaticamente le dimensioni in base al carico.
   b. Valuta la possibilità di mettere in pausa o interrompere le istanze non di produzione al di fuori degli orari lavorativi.
6. Devi segmentare e suddividere i tuoi modelli di dati in base agli schemi di accesso e alle caratteristiche dei dati?
   a. Valuta la possibilità di utilizzare AWS DMS o AWS SCT per spostare i dati su altri servizi.

Livello di impegno per il piano di implementazione:

Per attuare questa best practice è necessario conoscere caratteristiche e parametri attuali dei dati. Raccogliere tali parametri, definire una linea di base e quindi utilizzare i parametri per identificare le opzioni ideali per la configurazione del database richiede un livello di impegno da basso a moderato livello di impegno La convalida migliore passa attraverso test di carico e sperimentazioni.

Risorse

Documenti correlati:

- Database su cloud AWS
- AWS Database Caching (Memorizzazione nella cache del database AWS)
- Amazon DynamoDB Accelerator
- Best practice di Amazon Aurora
- Prestazioni di Amazon Redshift

- 10 suggerimenti prestazionali su Amazon Athena
- Best practice di Amazon Redshift Spectrum
- Best practice di Amazon DynamoDB

Video correlati:

- AWS purpose-built databases (Database dedicati di AWS) (DAT209-L)
- Amazon Aurora storage demystified: How it all works (Sfatiamo i miti sullo storage Amazon Aurora: come funziona realmente)(DAT309-R)
- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Esempi correlati:

- Esempi di Amazon DynamoDB
- Esempi di migrazione di database con AWS
- Database Modernization Workshop (Workshop sulla modernizzazione dei database)
- Utilizzo dei parametri sul database di Amazon RDS per PostgreSQL

PERF04-BP03 Raccolta e registrazione dei parametri delle prestazioni del database

Per capire come si comportano i sistemi di gestione dei dati, è importante monitorare i parametri pertinenti. Questi parametri ti aiuteranno a ottimizzare le risorse di gestione dei dati, a garantire che i requisiti del carico di lavoro siano soddisfatti e ad avere una chiara panoramica sulle prestazioni del carico di lavoro. Utilizza strumenti, librerie e sistemi che registrano misure delle prestazioni relative alle prestazioni del database.

Esistono parametri relativi al sistema su cui è ospitato il database (ad esempio, CPU, spazio di archiviazione, memoria, IOPS) e parametri di accesso ai dati stessi (ad esempio, transazioni al secondo, velocità di esecuzione delle query, tempi di risposta, errori). Questi parametri devono essere facilmente accessibili a tutto il personale di supporto o operativo e devono avere un registro cronologico sufficiente per poter identificare tendenze, anomalie e colli di bottiglia.

Risultato desiderato: per monitorare le prestazioni dei carichi di lavoro del database, è necessario registrare più parametri delle prestazioni in un dato periodo di tempo. Ciò consente di rilevare le

anomalie e di misurare le prestazioni rispetto ai parametri aziendali, per garantire che le esigenze del carico di lavoro siano soddisfatte.

Anti-pattern comuni:

- Utilizzi solo i file di log manuali per la ricerca dei parametri.
- Pubblichi i parametri solo sugli strumenti interni utilizzati dal tuo team e non hai un quadro completo del carico di lavoro.
- Utilizzi solo i parametri predefiniti registrati dal software di monitoraggio selezionato.
- Rivedi i parametri solo quando c'è un problema.
- Monitori solo i parametri a livello di sistema, senza catturare l'accesso ai dati o i parametri di utilizzo.

Vantaggi dell'adozione di questa best practice: la definizione di una linea di base delle prestazioni aiuta a comprendere il comportamento normale e i requisiti dei carichi di lavoro. I modelli anomali possono essere identificati ed eliminati più rapidamente, per migliorare le prestazioni e l'affidabilità del database. La capacità del database può essere configurata per garantire costi ottimali senza compromettere le prestazioni.

Livello di rischio associato se questa best practice non fosse adottata: Alta

- L'incapacità di distinguere il livello di prestazioni fuori dalla norma da quello nella norma crea difficoltà nell'identificazione dei problemi e nel processo decisionale.
- I potenziali risparmi sui costi possono non essere identificati.
- Non verranno identificati modelli di crescita che possono comportare un degrado dell'affidabilità o delle prestazioni.

Guida all'implementazione

Identificare, raccogliere, aggregare e correlare i parametri relativi al database. I parametri devono includere sia il sistema sottostante che supporta il database sia i parametri del database. I parametri del sistema sottostante possono includere utilizzo della CPU, memoria, spazio di archiviazione su disco disponibile, I/O su disco e parametri di rete in entrata e in uscita, mentre i parametri del database possono includere transazioni al secondo, query principali, velocità media delle query, tempi di risposta, utilizzo degli indici, blocco delle tabelle, timeout delle query e numero di connessioni aperte. Questi dati sono cruciali per capire come si comporta il carico di lavoro e come viene utilizzata

la soluzione di database. Utilizza tali parametri come parte di un approccio basato sui dati per mettere a punto e ottimizzare le risorse del tuo carico di lavoro.

Passaggi dell'implementazione:

1. Quali parametri del database è importante monitorare?
    a. Monitoraggio di parametri in un'istanza Amazon RDS
    b. Monitoraggio con Performance Insights
    c. Monitoraggio avanzato
    d. Parametri di DynamoDB
    e. Monitoraggio di DynamoDB DAX
    f. Monitoring MemoryDB (Monitoraggio di MemoryDB)
    g. Monitoraggio di Amazon Redshift
    h. Timeseries metrics and dimensions (Parametri e dimensioni delle serie temporali)
    i. Parametri a livello di cluster per Aurora
    j. Monitoring Amazon Keyspaces (Monitoraggio di Amazon Keyspaces)
    k. Monitoring Amazon Neptune (Monitoraggio di Amazon Neptune)
2. Il monitoraggio del database può trarre vantaggio da una soluzione di machine learning che rileva anomalie operative e problemi di prestazioni?
    a. Amazon DevOps Guru per Amazon RDS offre visibilità sui problemi di prestazioni e fornisce suggerimenti per le azioni correttive.
3. Hai bisogno di dettagli a livello di applicazione sull'utilizzo di SQL?
    a. AWS X-Ray può essere inserito nell'applicazione per ottenere approfondimenti e incapsulare tutti i punti di dati per una singola query.
4. Disponi attualmente di una soluzione di registrazione e monitoraggio approvata?
    a. Amazon CloudWatch può raccogliere i parametri per tutte le risorse dell'architettura. Puoi anche raccogliere e pubblicare parametri personalizzati per ottenere parametri aziendali o derivati. Utilizza CloudWatch o soluzioni di terze parti per impostare allarmi che indicano il superamento delle soglie.
5. Hai identificato e configurato le policy di conservazione dei dati in modo che corrispondano ai miei obiettivi operativi e di sicurezza?
    a. Conservazione dei dati predefinita per i parametri CloudWatch
    b. Conservazione dei dati predefinita per CloudWatch Logs

Livello di impegno per il piano di implementazione: esiste un livello medio di impegno per identificare, monitorare, raccogliere, aggregare e correlare i parametri di tutte le risorse del database.

Risorse

Documenti correlati:

- AWS Database Caching (Memorizzazione nella cache del database AWS)
- 10 suggerimenti prestazionali su Amazon Athena
- Best practice con Amazon Aurora
- Amazon DynamoDB Accelerator
- Best practice di Amazon DynamoDB
- Best practice di Amazon Redshift Spectrum (Best practice per Amazon Redshift Spectrum)
- Prestazioni di Amazon Redshift
- Database su cloud AWS
- Amazon RDS Performance Insights

Video correlati:

- AWS purpose-built databases (Database dedicati di AWS) (DAT209-L)
- Amazon Aurora storage demystified: How it all works (Sfatiamo i miti sullo storage Amazon Aurora: come funziona realmente)(DAT309-R)
- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Esempi correlati:

- Level 100: Monitoring with CloudWatch Dashboards (Livello 100: Monitoraggio con i pannelli di controllo CloudWatch)
- AWS Dataset Ingestion Metrics Collection Framework (Framework di raccolta dei parametri di ingestione del set di dati AWS)
- Workshop di monitoraggio Amazon RDS

PERF04-BP04 Scelta dello spazio di archiviazione dei dati in base ai modelli di accesso

Utilizza gli schemi di accesso del carico di lavoro per decidere quali servizi e tecnologie utilizzare. Oltre ai requisiti non funzionali, come le prestazioni e il dimensionamento, i modelli di accesso influenzano pesantemente la scelta del database e delle soluzioni di archiviazione. La prima dimensione è data da necessità di transazioni, conformità ACID e letture coerenti. Non tutti i database supportano queste caratteristiche e la maggior parte dei database NoSQL fornisce un modello di consistenza eventuale. La seconda dimensione importante è la distribuzione delle scritture e delle letture nel tempo e nello spazio. Le applicazioni distribuite a livello globale devono considerare i modelli di traffico, la latenza e i requisiti di accesso per identificare la soluzione di archiviazione ottimale. Il terzo aspetto cruciale da scegliere è la flessibilità dei modelli di query, i modelli di accesso casuale e le query una tantum. Occorre inoltre tenere conto delle funzionalità di query altamente specializzate per l'elaborazione del testo e del linguaggio naturale, delle serie temporali e dei grafici.

Risultato desiderato: l'archiviazione di dati è stata selezionata in base a modelli di accesso ai dati identificati e documentati. Ciò potrebbe includere le query di lettura, scrittura e cancellazione più comuni, la necessità di calcoli e aggregazioni ad hoc, la complessità dei dati, l'interdipendenza dei dati e le esigenze di coerenza richieste.

Anti-pattern comuni:

- È sufficiente selezionare un solo fornitore di database per semplificare la gestione delle operazioni.
- Ritieni che gli schemi di accesso ai dati rimarranno coerenti nel tempo.
- Implementi transazioni complesse, rollback e logica di coerenza nell'applicazione.
- Il database è configurato per supportare un potenziale burst di traffico elevato, che fa sì che le risorse del database rimangano inattive per la maggior parte del tempo.
- Utilizzo di un database condiviso per usi transazionali e analitici.

Vantaggi dell'adozione di questa best practice: la selezione e l'ottimizzazione dell'archiviazione dei dati in base ai modelli di accesso contribuirà a ridurre la complessità dello sviluppo e a ottimizzare le opportunità di prestazione. Capire quando utilizzare le repliche di lettura, le tabelle globali, il partizionamento dei dati e la memorizzazione nella cache, ti aiuterà a ridurre i costi operativi e a effettuare il dimensionamento in base alle esigenze del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

Identifica e valuta il modello di accesso ai dati per selezionare la configurazione di archiviazione corretta. Ogni soluzione di database dispone di opzioni per configurare e ottimizzare la soluzione di archiviazione. Utilizza i parametri e i registri raccolti e sperimenta le opzioni per trovare la configurazione ottimale. Utilizza la tabella seguente per esaminare le opzioni di archiviazione per ogni servizio di database.

Servizi AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
Dimensionamento dello spazio di archiviazione	L'opzione di dimensionamento automatico dello spazio di archiviazione è disponibile per dimensionare automaticamente lo spazio di archiviazione allocato. La capacità	Dimensionamento automatico. Le tabelle non hanno vincoli di dimensione.	Opzione di dimensionamento automatico dello spazio di archiviazione disponibile per dimensionare lo spazio di archiviazione allocato	Spazio di archiviazione in memoria, legato al tipo o al numero di istanze	Opzione di dimensionamento automatico dello spazio di archiviazione disponibile per dimensionare automaticamente lo spazio di archiviazione allocato	Configura il periodo di conservazione per i livelli in memoria e magnetici nei giorni	Aumenta e diminuisce automaticamente lo spazio di archiviazione della tabella	Dimensionamento automatico. Le tabelle non hanno vincoli di dimensione.

Servizi AWS	Amazon RDS, Amazon Aurora	Amazon DynamoD	Amazon Document B	Amazon ElastiCache	Amazon Neptune	Amazon Timestream	Amazon Keyspace	Amazon QLDB
	di IOPS allocata può essere dimensionata anche indipendentemente dallo spazio di archiviazione allocato quando si sfruttano i tipi di spazio di archiviazione IOPS allocati.							

Passaggi dell'implementazione:

1. Identifica e documenta la crescita prevista dei dati e del traffico.
   a. Amazon RDS e Aurora supportano l'aumento automatico dello spazio di archiviazione fino ai limiti documentati. Oltre a questo, si può prendere in considerazione la transizione dei dati

più vecchi verso Amazon S3 per l'archiviazione, l'aggregazione dei dati storici per l'analisi o la scalabilità orizzontale tramite partizioni.

b. DynamoDB e Amazon S3 dimensioneranno automaticamente fino a raggiungere un volume di archiviazione quasi illimitato.

c. I database e le istanze Amazon RDS in esecuzione su EC2 possono essere ridimensionati manualmente e le istanze EC2 possono avere nuovi volumi EBS aggiunti in un secondo momento per ottenere ulteriore spazio di archiviazione.

d. I tipi di istanza possono essere modificati in base alle variazioni dell'attività. Ad esempio, puoi iniziare con un'istanza più piccola durante i test, per poi dimensionare l'istanza quando inizi a ricevere traffico di produzione verso il servizio. Aurora Serverless V2 si riduce orizzontalmente in modo automatico in risposta alle modifiche nel carico.

1. Documenta i requisiti relativi alle prestazioni normali e di picco (transazioni al secondo TPS e query al secondo QPS) e alla consistenza (ACID e consistenza eventuale).
2. Documenta gli aspetti di implementazione della soluzione e i requisiti di accesso al database (globale, Mult-AZ, replica in lettura, nodi di scrittura multipli)

Livello di impegno per il piano di implementazione: se non disponi di registri o parametri per la tua soluzione di gestione dei dati, devi completarli prima di identificare e documentare i modelli di accesso ai dati. Una volta compreso il modello di accesso ai dati, la selezione e la configurazione dello spazio di archiviazione dei dati è un basso livello di impegno

Risorse

Documenti correlati:

- AWS Database Caching (Memorizzazione nella cache del database AWS)
- 10 suggerimenti prestazionali su Amazon Athena
- Best practice di Amazon Aurora
- Amazon DynamoDB Accelerator
- Best practice di Amazon DynamoDB
- Best practice di Amazon Redshift Spectrum
- Prestazioni di Amazon Redshift
- Database su cloud AWS
- Tipi di archiviazione Amazon RDS

Video correlati:

- AWS purpose-built databases (Database dedicati di AWS) (DAT209-L)
- Amazon Aurora storage demystified: How it all works (Sfatiamo i miti sullo storage Amazon Aurora: come funziona realmente)(DAT309-R)
- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Esempi correlati:

- Experiment and test with Distributed Load Testing on AWS (Esperimenti e prove con i test di carico distribuiti su AWS)

PERF04-BP05 Ottimizzazione dello spazio di archiviazione dei dati in base ai modelli e ai parametri di accesso

Utilizza caratteristiche delle prestazioni e schemi di accesso che ottimizzano il modo in cui i dati vengono archiviati o interrogati al fine di ottenere le migliori prestazioni possibili. Misura il modo in cui le ottimizzazioni come l'indicizzazione, la distribuzione delle chiavi, la progettazione dei data warehouse o le strategie di memorizzazione nella cache influenzano le prestazioni del sistema o la sua efficienza nel complesso.

Anti-pattern comuni:

- Utilizzi solo i file di log manuali per la ricerca dei parametri.
- Pubblichi i parametri solo negli strumenti interni.

Vantaggi dell'adozione di questa best practice: Per assicurarti di soddisfare i parametri necessari per il carico di lavoro, devi monitorare i parametri delle prestazioni del database correlati alla lettura e alla scrittura. Puoi utilizzare questi dati per aggiungere nuove ottimizzazioni per le operazioni di lettura e scrittura al livello di storage dei dati.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

Ottimizzazione dello spazio di archiviazione di dati in base a parametri e modelli: i parametri segnalati per identificare le aree con prestazioni inferiori nel carico di lavoro e ottimizzare i componenti del

database. Ogni sistema del database ha caratteristiche diverse relative alle prestazioni che devono essere valutate, come il modo in cui i dati sono indicizzati, memorizzati nella cache o distribuiti in più sistemi. Misurazione dell'impatto delle ottimizzazioni.

Risorse

Documenti correlati:

- AWS Database Caching (Memorizzazione nella cache del database AWS)
- 10 suggerimenti prestazionali su Amazon Athena
- Best practice con Amazon Aurora
- Amazon DynamoDB Accelerator
- Best practice di Amazon DynamoDB
- Best practice di Amazon Redshift Spectrum (Best practice per Amazon Redshift Spectrum)
- Prestazioni di Amazon Redshift
- Database su cloud AWS
- Analisi delle anomalie delle prestazioni con DevOps Guru per RDS
- Modalità di capacità in lettura/scrittura per DynamoDB

Video correlati:

- AWS purpose-built databases (Database dedicati di AWS) (DAT209-L)
- Amazon Aurora storage demystified: How it all works (Sfatiamo i miti sullo storage Amazon Aurora: come funziona realmente)(DAT309-R)
- Amazon DynamoDB deep dive: Advanced design patterns (DAT403-R1)

Esempi correlati:

- Laboratori pratici per Amazon DynamoDB

PERF 5 In che modo configuri la soluzione di rete?

La soluzione di rete ottimale per un carico di lavoro varia in base a latenza, requisiti di throughput, jitter e larghezza di banda. I vincoli fisici, ad esempio le risorse utente o in locale, determinano le

opzioni di posizione. Questi vincoli possono essere compensati con le edge location o la collocazione delle risorse.

Best practice

- PERF05-BP01 In che modo la rete influisce sulle prestazioni
- PERF05-BP02 Valutazione delle funzionalità di rete disponibili
- PERF05-BP03 Scelta di una connettività dedicata o una VPN di dimensioni adeguate ai carichi di lavoro ibridi
- PERF05-BP04 Sfruttamento del bilanciamento del carico e dell'offloading della crittografia
- PERF05-BP05 Scelta dei protocolli di rete per migliorare le prestazioni
- PERF05-BP06 Scelta della posizione del carico di lavoro in base ai requisiti di rete
- PERF05-BP07 Ottimizzazione della configurazione di rete in base ai parametri

PERF05-BP01 In che modo la rete influisce sulle prestazioni

Analizza e comprendi come le decisioni relative alla rete influenzano le prestazioni di rete. La rete è responsabile della connettività tra componenti dell'applicazione, servizi cloud, reti edge e dati on-premise e quindi può avere un forte impatto sulle prestazioni dei carichi di lavoro. Oltre alle prestazioni del carico di lavoro, l'esperienza dell'utente è influenzata anche da latenza della rete, larghezza di banda, protocolli, posizione, congestione della rete, jitter, velocità di trasmissione effettiva e regole di instradamento.

Risultato desiderato: Disporre di un elenco documentato dei requisiti di rete del carico di lavoro, tra cui latenza, dimensione dei pacchetti, regole di instradamento, protocolli e modelli di traffico di supporto. Esaminare le soluzioni di rete disponibili e individuare il servizio che soddisfi le caratteristiche di rete del proprio carico di lavoro. Le reti basate sul cloud possono essere ricostruite rapidamente, quindi l'evoluzione dell'architettura di rete nel tempo è necessaria per migliorare l'efficienza delle prestazioni.

Anti-pattern comuni:

- Tutto il traffico passa attraverso i data center esistenti.
- Le sessioni Direct Connect vengono sovradimensionate senza considerare i requisiti di utilizzo effettivi.
- Quando si definiscono le soluzioni di rete, non si considerano le caratteristiche del carico di lavoro e l'overhead della crittografia.

- Per le soluzioni di rete nel cloud si utilizzano concetti e strategie on-premise.

Vantaggi dell'adozione di questa best practice: Comprendere l'impatto della rete sulle prestazioni del carico di lavoro ti aiuterà a identificare i potenziali colli di bottiglia, migliorare l'esperienza dell'utente, aumentare l'affidabilità e ridurre la manutenzione operativa al variare del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Identifica i parametri importanti delle prestazioni di rete del carico di lavoro e cattura le sue caratteristiche di rete. Definisci e documenta i requisiti come parte di un approccio basato sui dati, utilizzando valori di riferimento o test di carico. Utilizza tali dati per identificare i punti in cui la soluzione di rete è vincolante ed esamina le opzioni di configurazione che possono migliorare il carico di lavoro. Comprendi le caratteristiche e le opzioni di rete native per il cloud disponibili e come questa possono influire sulle prestazioni del carico di lavoro in base ai requisiti. Ogni funzionalità di rete presenta vantaggi e svantaggi e può essere configurata per soddisfare le caratteristiche del carico di lavoro e dimensionarsi in base alle singole esigenze.

Passaggi dell'implementazione:

1. Definizione e documentazione dei requisiti di prestazioni della rete:
    a. Includi i parametri come latenza di rete, larghezza di banda, protocolli, posizioni, modelli di traffico (picchi e frequenza), velocità di trasmissione effettiva, crittografia, ispezione e regole di instradamento.
2. Acquisisci le caratteristiche fondamentali della rete:
    a. Log di flusso VPC
    b. Parametri di AWS Transit Gateway
    c. Parametri AWS PrivateLink
3. Acquisisci le caratteristiche di rete dell'applicazione:
    a. Adattatore di rete elastico
    b. AWS App Mesh metrics (Parametri AWS App Mesh)
    c. Amazon API Gateway metrics (Parametri Amazon API Gateway)
4. Acquisisci le caratteristiche di rete edge:
    a. Amazon CloudFront metrics (Visualizzazione dei parametri di CloudFront e di Lambda@Edge)

   b. Amazon Route 53 metrics (Monitoraggio delle risorse con i controlli dell'integrità di Amazon Route 53 e Amazon CloudWatch)
   c. AWS Global Accelerator metrics (Parametri AWS Global Accelerator)
5. Acquisisci le caratteristiche di rete edge ibride:
   a. Parametri Direct Connect
   b. Parametri VPN da-sito-a-sito AWS
   c. Parametri VPN client AWS
   d. Parametri WAN Cloud AWS
6. Acquisisci le caratteristiche di rete di sicurezza:
   a. AWS Shield, WAF, and Network Firewall metrics (Parametri AWS Shield, WAF e firewall di rete)
7. Acquisisci parametri di prestazioni end-to-end con gli strumenti di tracciamento:
   a. AWS X-Ray
   b. Usare il RUM Amazon CloudWatch
8. Esegui il benchmark e testa le prestazioni della rete:
   a. Velocità di trasmissione effettiva del benchmark: alcuni fattori che possono influire sulle prestazioni della rete EC2 quando le istanze si trovano nella stessa VPC. Misura la larghezza di banda della rete tra le istanze EC2 Linux nello stesso VPC.
   b. Esegui test di carico per sperimentare soluzioni e opzioni di rete

Livello di impegno per il piano di implementazione: esiste un livello di impegno medio per documentare requisiti di rete del carico di lavoro, opzioni e soluzioni disponibili.

Risorse

Documenti correlati:

- Application Load Balancer
- Reti avanzate su Linux
- Reti avanzate su Windows
- Gruppi di collocamento
- Abilitazione delle reti avanzate con Elastic Network Adapter (ENA) sulle istanze Linux
- Network Load Balancer
- Nuovi prodotti di rete con AWS

- Transit Gateway
- Passaggio all'instradamento basato sulla latenza in Amazon Route 53
- Endpoint VPC
- Log di flusso VPC

Video correlati:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)
- Improve Global Network Performance for Applications
- EC2 Instances and Performance Optimization Best Practices
- Optimizing Network Performance for Amazon EC2 Instances
- Networking best practices and tips with the Well-Architected Framework (Best practice di rete e suggerimenti sul Framework Well-Architected)
- AWS networking best practices in large-scale migrations

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluzioni di sicurezza scalabili)
- AWS Networking Workshops (Workshop di rete AWS)

PERF05-BP02 Valutazione delle funzionalità di rete disponibili

Valuta le funzionalità di rete nel cloud che possono aumentare le prestazioni. Misura l'impatto di tali funzionalità attraverso test, parametri e analisi. Ad esempio, sfrutta le funzionalità a livello di rete disponibili per ridurre latenza, perdita di pacchetti o jitter.

Vengono creati numerosi servizi per migliorare le prestazioni e altre caratteristiche incluse nelle offerte per ottimizzare le prestazioni della rete. Sono disponibili servizi come AWS Global Accelerator e Amazon CloudFront per migliorare le prestazioni, mentre la maggior parte degli altri servizi include funzionalità di prodotto per l'ottimizzazione del traffico di rete. Rivedi le caratteristiche dei servizi, ad esempio le funzionalità di rete delle istanze EC2, i tipi di istanze di rete migliorate, le istanze ottimizzate per Amazon EBS, l'accelerazione del trasferimento Amazon S3 e CloudFront, per migliorare le prestazioni del carico di lavoro.

Risultato desiderato: hai documentato l'inventario dei componenti all'interno del carico di lavoro e hai identificato le configurazioni di rete per componente che ti aiuteranno a rispettare la conformità ai requisiti relativi alle prestazioni. Dopo aver valutato le caratteristiche della rete, hai sperimentato e misurato le metriche relative alle prestazioni per identificare come utilizzare le funzionalità disponibili.

Anti-pattern comuni:

- Hai inserito tutti i carichi di lavoro nella Regione AWS più vicina alla sede centrale e non in una Regione AWS vicina agli utenti finali.
- Mancata esecuzione del benchmarking delle prestazioni dei carichi di lavoro e continua valutazione delle prestazioni dei carichi di lavoro in base all'analisi di benchmark.
- Mancata revisione delle configurazioni dei servizi in base alle opzioni di miglioramento delle prestazioni.

Vantaggi dell'adozione di questa best practice: la valutazione di tutte le funzionalità e le opzioni del servizio consente di ridurre il costo dell'infrastruttura e l'impegno necessario per mantenere il carico di lavoro e aumentare l'assetto di sicurezza generale. Puoi utilizzare la struttura portante globale di AWS per garantire ai tuoi clienti la migliore esperienza di rete.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

Analizza quali opzioni di configurazione relative alla rete sono disponibili e come possono influire sul tuo carico di lavoro. La comprensione della modalità in cui tali opzioni interagiscono con la tua architettura e dell'impatto che avranno sia sulle prestazioni misurate sia sulle prestazioni percepite dagli utenti è fondamentale per ottimizzare le prestazioni.

Passaggi dell'implementazione:

1. Crea l'elenco dei componenti del carico di lavoro.
   a. Crea, gestisci e monitora la rete dell'organizzazione mediante Cloud AWS WAN.
   b. Ottieni maggiore visibilità a livello di rete mediante Network Manager. Utilizza uno strumento per database di gestione delle configurazioni (CMDB) esistente oppure uno strumento come AWS Config per creare un inventario del carico di lavoro e della relativa configurazione.
2. Se si tratta di un carico di lavoro esistente, individua e documenta l'analisi di benchmark per le metriche relative alle prestazioni, concentrandoti sui colli di bottiglia e sulle aree da migliorare. Le metriche relative alla rete a livello di prestazioni varieranno a seconda dei requisiti aziendali e delle

caratteristiche del carico di lavoro. Come punto di partenza, le seguenti metriche possono essere importanti per la revisione del carico di lavoro: larghezza di banda, latenza, perdita di pacchetti, jitter e ritrasmissioni.

3. Se si tratta di un nuovo carico di lavoro, esegui i test di carico per individuare eventuali colli di bottiglia relativi alle prestazioni.
4. Per tutti i colli di bottiglia di questo tipo riscontrati, esamina le opzioni di configurazione per le soluzioni in uso per individuare le opportunità di miglioramento delle prestazioni.
5. Se il percorso o gli instradamenti di rete non sono noti, utilizza Network Access Analyzer per individuarli.
6. Rivedi i protocolli di rete per ridurre ulteriormente la latenza.
   - PERF05-BP05 Scelta dei protocolli di rete per migliorare le prestazioni
7. Se utilizzi una AWS Site-to-Site VPN tra più posizioni per connetterti a una Regione AWS, rivedi le connessioni Site-to-Site VPN accelerate per individuare eventuali opportunità di miglioramento delle prestazioni di rete.
8. Quando il traffico del carico di lavoro interessa più account, valuta la topologia della rete e i servizi per ridurre la latenza.
   - Valuta i compromessi a livello di operazioni e prestazioni tra Peering VPC e AWS Transit Gateway in caso di connessione di più account. AWS Transit Gateway supporta una velocità di trasmissione effettiva della VPN sito-sito AWS con dimensionamento superiore al limite IPsec massimo mediante l'utilizzo di percorsi multipli. Il traffico tra Amazon VPC e AWS Transit Gateway rimane sulla rete AWS privata e non è esposto a Internet. AWS Transit Gateway semplifica le procedure di interconnessione di tutti i VPC, che possono interessare migliaia di Account AWS ed estendersi su reti on-premise. Condividi AWS Transit Gateway tra più account utilizzando la funzionalità Gestione degli accessi alle risorse. Per ottenere visibilità nel traffico di rete globale, utilizza Network Manager per avere una visione centralizzata delle metriche della rete.
9. Rivedi la collocazione degli utenti e riduci la distanza tra gli utenti e il carico di lavoro.
   a. AWS Global Accelerator è un servizio di rete che migliora le prestazioni del traffico degli utenti fino al 60% utilizzando l'infrastruttura di rete globale di Amazon Web Services. In caso di congestione di Internet, AWS Global Accelerator ottimizza il percorso verso la tua applicazione per ridurre la perdita di pacchetti, il jitter e la latenza in modo consistente. Fornisce inoltre indirizzi IP statici che semplificano lo spostamento degli endpoint tra zone di disponibilità o Regioni AWS senza la necessità di aggiornare la configurazione DNS o di modificare le applicazioni lato client.

b. Amazon CloudFront può migliorare le prestazioni della distribuzione e della latenza del carico di lavoro globalmente. CloudFront ha oltre 410 punti di presenza dislocati a livello globale che possono memorizzare i contenuti nella cache e ridurre la latenza per gli utenti finali.

c. Amazon Route 53 offre opzioni di instradamento basato sulla latenza, instradamento basato sulla geolocalizzazione, instradamento basato sulla geoprossimitàe instradamento basato su IP per aiutare a migliorare le prestazioni del carico di lavoro per un pubblico globale. Individua l'opzione di instradamento in grado di ottimizzare le prestazioni del carico di lavoro. A tale scopo, rivedi il traffico del carico di lavoro e la posizione degli utenti.

10. Valuta altre caratteristiche di Amazon S3 per migliorare le operazioni di input/output al secondo (IOPS) relative all'archiviazione.

a. Amazon S3 Transfer Acceleration è una funzione che consente agli utenti esterni di sfruttare i vantaggi delle ottimizzazioni di rete di CloudFront per il caricamento dei dati in Amazon S3. Ciò migliora le caratteristiche di trasferimento di grandi quantità di dati da posizioni remote prive di connettività dedicata al Cloud AWS.

b. I punti di accesso multi-regione in Amazon S3 rappresentano una funzionalità che replica i contenuti in più regioni e semplifica il carico di lavoro fornendo un punto di accesso. Quando viene utilizzato un punto di accesso multi-regione, puoi richiedere o scrivere dati in Amazon S3 con il servizio che identifica il bucket con latenza più bassa.

11. Rivedi la larghezza di banda della rete per la risorsa di calcolo in uso.

a. Le interfacce di rete elastica (ENA) utilizzate da istanze EC2, container e funzioni Lambda sono limitate in base ai flussi. Rivedi i gruppi di collocazione per ottimizzare la velocità di trasmissione effettiva EC2. Per evitare colli di bottiglia a livello di flusso, progetta l'applicazione in modo che utilizzi più flussi. Per monitorare le metriche di rete associate al calcolo e avere maggiore visibilità su di esse, utilizza le metriche CloudWatch e `ethtool`. `ethtool` è incluso nel driver ENA ed espone metriche relative alla rete aggiuntive che possono essere pubblicate come metrica personalizzata in CloudWatch.

b. Le istanze EC2 più recenti possono sfruttare le reti avanzate. Istanze EC2 della serie N, ad esempio `M5n` e `M5dn`, sfruttano la quarta generazione di schede Nitro per offrire fino a 100 Gbps di velocità di trasmissione effettiva di rete a una singola istanza. Queste istanze offrono quattro volte la larghezza di banda della rete e il processo di pacchetti rispetto alle istanze `M5` di base, e sono ideali per le applicazioni che fanno un uso intensivo della rete.

c. Gli adattatori ENA (Elastic Network Adapter) Amazon offrono un'ulteriore ottimizzazione grazie a una migliore velocità di trasmissione effettiva per le istanze all'interno di un gruppo di collocazione cluster.

d. L'adattatore EFA (Elastic Fabric Adapter) è un'interfaccia di rete per le istanze Amazon EC2 che consente di eseguire carichi di lavoro che richiedono elevati livelli di comunicazioni tra i nodi su vasta scala in AWS. Con EFA, le applicazioni High Performance Computing (HPC) che utilizzano le applicazioni Message Passing Interface (MPI) e le applicazioni Machine Learning (ML) che utilizzano NVIDIA Collective Communications Library (NCCL) possono ridimensionare le risorse fino a migliaia di CPU o GPU.
e. Le istanze ottimizzate per Amazon EBS utilizzano uno stack di configurazione ottimizzato e forniscono un'ulteriore capacità dedicata per incrementare l'I/O di Amazon EBS. Questa ottimizzazione fornisce le prestazioni migliori ai tuoi volumi EBS, riducendo al minimo i conflitti tra Amazon EBS I/O ed altro traffico dalla tua istanza.

Livello di impegno per il piano di implementazione:

Per definire questa best practice, devi saper individuare le opzioni per i componenti del carico di lavoro corrente che possono avere un impatto sulle prestazioni della rete. La raccolta di informazioni sui componenti, la valutazione delle opzioni di miglioramento della rete, la sperimentazione, l'implementazione e la documentazione dei miglioramenti sono operazioni caratterizzate da un livello di impegno da basso a moderato .

Risorse

Documenti correlati:

- Istanze ottimizzate per Amazon EBS
- Application Load Balancer
- Amazon EC2 instance network bandwidth (Larghezza di banda della rete per le istanze Amazon EC2)
- Reti avanzate su Linux
- Reti avanzate su Windows
- Gruppi di collocamento
- Abilitazione delle reti avanzate con Elastic Network Adapter (ENA) sulle istanze Linux
- Network Load Balancer
- Nuovi prodotti di rete con AWS
- AWS Transit Gateway
- Passaggio all'instradamento basato sulla latenza in Amazon Route 53

- Endpoint VPC
- Log di flusso VPC
- Building a cloud CMDB (Creazione di una strategia CMDB per il cloud)
- Scaling VPN throughput using AWS Transit Gateway (Dimensionamento della velocità di trasmissione effettiva della VPN mediante AWS Transit Gateway)

Video correlati:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)
- AWS Global Accelerator

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluzioni di sicurezza scalabili)
- AWS Networking Workshops (Workshop di rete AWS)

PERF05-BP03 Scelta di una connettività dedicata o una VPN di dimensioni adeguate ai carichi di lavoro ibridi

Quando è necessaria una rete comune per collegare risorse on-premise e cloud in AWS, assicurati di avere una larghezza di banda adeguata per soddisfare i requisiti di prestazione. Fai una stima dei requisiti di larghezza di banda e latenza per il carico di lavoro ibrido. Sulla base di questi numeri potrai stabilire i requisiti di dimensionamento per AWS Direct Connect o gli endpoint VPN.

Risultato desiderato: quando si implementa un carico di lavoro che necessita di connettività di rete ibrida, sono disponibili diverse opzioni di configurazione per la connettività, ad esempio Direct Connect o VPN gestite e non gestite. Seleziona il tipo di connessione appropriato per ogni carico di lavoro, assicurandoti di avere requisiti adeguati di larghezza di banda e crittografia tra la tua posizione e il cloud.

Anti-pattern comuni:

- Valutare le soluzioni VPN solo per i tuoi requisiti di crittografia di rete.
- Non valutare le opzioni di backup o di connettività parallele.

- Utilizzare configurazioni predefinite per router, tunnel e sessioni BGP.
- Non riuscire a comprendere o a identificare tutti i requisiti del carico di lavoro (esigenze di crittografia, protocollo, larghezza di banda e traffico).

Vantaggi dell'adozione di questa best practice: la scelta e la configurazione di soluzioni di rete ibride di dimensioni adeguate aumenterà l'affidabilità del carico di lavoro e massimizzerà le opportunità di prestazioni. Identificando i requisiti del carico di lavoro, pianificando in anticipo e valutando soluzioni ibride, ridurrai al minimo le costose modifiche alla rete fisica e i costi operativi, aumentando al contempo il time to market.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Sviluppa un'architettura di rete ibrida basata sui requisiti di larghezza di banda: fai una stima dei requisiti di larghezza di banda e latenza delle applicazioni ibride. In base ai requisiti di larghezza di banda, una singola connessione VPN o Direct Connect potrebbe non essere sufficiente, pertanto è necessario progettare una configurazione ibrida per abilitare il bilanciamento del carico del traffico su più connessioni. Può essere richiesta la connessione diretta, che offre prestazioni più prevedibili e costanti grazie alla connettività di rete privata. È ideale per carichi di lavoro di produzione che richiedono latenza costante e jitter vicino allo zero.

AWS Direct Connect offre una connettività dedicata all'ambiente AWS, da 50 Mbps fino a 10 Gbps. In questo modo potrai disporre di una latenza gestita e controllata, nonché di una larghezza di banda assegnata, in modo che le applicazioni siano in grado di connettersi facilmente ad altri ambienti con prestazioni ottimali. Affidandoti a uno dei partner di AWS Direct Connect, ottieni una connettività end-to-end da più ambienti, per una rete estesa con prestazioni costanti.

La VPN site-to-site AWS è un servizio VPN gestito per VPC. Quando si crea una connessione VPN, AWS fornisce tunnel a due endpoint VPN diversi. Con AWS Transit Gateway, puoi semplificare la connettività tra più VPC e collegarti a qualsiasi VPC connesso ad AWS Transit Gateway con una singola connessione VPN. AWS Transit Gateway ti consente inoltre di dimensionare le risorse oltre il limite di velocità di trasmissione effettiva VPN IPsec di 1,25 Gbps, abilitando il supporto di instradamento ECMP (equal cost multi-path) su più tunnel VPN.

Livello di impegno per il piano di implementazione: esiste un livello di impegno elevato per valutare le esigenze di carico di lavoro delle reti ibride e per implementare soluzioni di rete ibride.

Risorse

Documenti correlati:

- Network Load Balancer
- Nuovi prodotti di rete con AWS
- Transit Gateway
- Passaggio all'instradamento basato sulla latenza in Amazon Route 53
- Endpoint VPC
- Log di flusso VPC
- VPN site-to-site
- Building a Scalable and Secure Multi-VPC AWS Network Infrastructure
- Direct Connect
- Client VPN

Video correlati:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)
- AWS Global Accelerator
- Direct Connect
- Transit Gateway Connect
- Soluzioni VPN
- Security with VPN Solutions

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluzioni di sicurezza scalabili)
- AWS Networking Workshops (Workshop di rete AWS)

PERF05-BP04 Sfruttamento del bilanciamento del carico e dell'offloading della crittografia

Distribuisci il traffico tra varie risorse o servizi affinché il carico di lavoro possa trarre vantaggio dall'elasticità fornita dal cloud. Puoi anche utilizzare il bilanciamento del carico per la terminazione dell'offloading della crittografia al fine di migliorare le prestazioni, gestire e instradare il traffico in modo efficiente.

Quando si implementa un'architettura a dimensionamento orizzontale in cui si desidera utilizzare più istanze per il contenuto del servizio, è possibile utilizzare i load balancer all'interno del Amazon VPC. AWS offre modelli multipli per le tue applicazioni nel servizio ELB. Application Load Balancer è più adatto per il bilanciamento del carico del traffico HTTP e HTTPS e fornisce un instradamento avanzato delle richieste mirato all'erogazione di architetture applicative moderne, compresi microservizi e container.

Network Load Balancer è l'ideale per il bilanciamento del carico del traffico TCP, in cui sono richieste prestazioni elevatissime. È in grado di gestire milioni di richieste al secondo, mantenendo al contempo latenze ridottissime. Inoltre, è ottimizzato per la gestione degli schemi di traffico improvvisi e incostanti.

Elastic Load Balancing fornisce la gestione integrata dei certificati e la decrittografia SSL/TLS, offrendoti la flessibilità di gestire centralmente le impostazioni SSL del sistema di bilanciamento del carico e di sollevare il carico di lavoro dall'utilizzo intensivo della CPU.

Anti-pattern comuni:

- Instradi tutto il traffico Internet attraverso i sistemi di bilanciamento del carico esistenti.
- Utilizzi il bilanciamento del carico TCP generico e fai in modo che ogni nodo di calcolo gestisca la crittografia SSL.

Vantaggi dell'adozione di questa best practice: Un sistema di bilanciamento del carico gestisce il carico variabile del traffico dell'applicazione in una o più zone di disponibilità. I sistemi di bilanciamento del carico offrono l'elevata disponibilità, l'auto scaling e la solida sicurezza necessarie per rendere le applicazioni tolleranti ai guasti.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Utilizzo del load balancer appropriato per il carico di lavoro: seleziona il load balancer appropriato per il carico di lavoro. Per bilanciare il carico delle richieste HTTP, ti consigliamo Application Load

Balancer. Per il bilanciamento del carico dei protocolli di rete e di trasporto (livello 4 – TCP, UDP) e per applicazioni con prestazioni estreme e a bassa latenza, ti consigliamo Network Load Balancer. Application Load Balancers supporta HTTPS, mentre Network Load Balancer supporta l'offloading della crittografia TLS.

Abilitazione dell'offload della crittografia HTTPS o TLS: Elastic Load Balancing include la gestione integrata dei certificati, l'autenticazione degli utenti e la decrittografia SSL/TLS. Offre la flessibilità necessaria per gestire centralmente le impostazioni TLS e per sollevare le applicazioni dai carichi di lavoro con elevati requisiti di CPU. Esegui la crittografia di tutto il traffico HTTPS come parte della distribuzione del sistema di bilanciamento del carico.

Risorse

Documenti correlati:

- Istanze ottimizzate per Amazon EBS
- Application Load Balancer
- Reti avanzate su Linux
- Reti avanzate su Windows
- Gruppi di collocamento
- Abilitazione delle reti avanzate con Elastic Network Adapter (ENA) sulle istanze Linux
- Network Load Balancer
- Nuovi prodotti di rete con AWS
- Transit Gateway
- Passaggio all'instradamento basato sulla latenza in Amazon Route 53
- Endpoint VPC
- Log di flusso VPC

Video correlati:

- Connettività ad AWS e architetture di rete AWS ibride (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluzioni di sicurezza scalabili)
- AWS Networking Workshops (Workshop di rete AWS)

PERF05-BP05 Scelta dei protocolli di rete per migliorare le prestazioni

Prendi decisioni sui protocolli per la comunicazione tra sistemi e reti in base all'impatto sulle prestazioni del carico di lavoro.

Esiste una relazione tra latenza e larghezza di banda per ottenere il throughput desiderato. Se il trasferimento di file utilizza TCP, latenze più elevate ridurranno il throughput complessivo. Alcuni approcci risolvono questo problema con l'ottimizzazione TCP e i protocolli di trasferimento ottimizzati, altri adottano UDP.

Anti-pattern comuni:

- Puoi utilizzare il TCP per tutti i carichi di lavoro, indipendentemente dai requisiti prestazionali.

Vantaggi dell'adozione di questa best practice: La selezione del protocollo appropriato per la comunicazione tra i componenti del carico di lavoro permette di ottenere le migliori prestazioni per quel carico di lavoro. L'UDP senza connessione garantisce velocità elevata, ma non offre ritrasmissione o alta affidabilità. Il TCP è un protocollo completo, ma richiede un sovraccarico maggiore per l'elaborazione dei pacchetti.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

Ottimizza il traffico di rete: seleziona il protocollo appropriato per ottimizzare le prestazioni del carico di lavoro. Esiste una relazione tra latenza e larghezza di banda per ottenere il throughput desiderato. Se il trasferimento di file utilizza il TCP, latenze più elevate ridurranno il throughput complessivo. Alcuni approcci risolvono il problema della latenza con l'ottimizzazione TCP e i protocolli di trasferimento ottimizzati, altri utilizzano l'UDP.

Risorse

Documenti correlati:

- Istanze ottimizzate per Amazon EBS

- Application Load Balancer
- Reti avanzate su Linux
- Reti avanzate su Windows
- Gruppi di collocamento
- Abilitazione delle reti avanzate con Elastic Network Adapter (ENA) sulle istanze Linux
- Network Load Balancer
- Nuovi prodotti di rete con AWS
- Transit Gateway
- Passaggio all'instradamento basato sulla latenza in Amazon Route 53
- Endpoint VPC
- Log di flusso VPC

Video correlati:

- Connettività AWS e architetture di rete AWS ibride (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluzioni di sicurezza scalabili)
- AWS Networking Workshops (Workshop di rete AWS)

PERF05-BP06 Scelta della posizione del carico di lavoro in base ai requisiti di rete

Utilizza le opzioni di posizione nel cloud disponibili per ridurre la latenza di rete o migliorare il throughput. Utilizza Regioni AWS, zone di disponibilità, gruppi di collocazione e posizioni edge come AWS Outposts, Zone locali AWS e AWS Wavelength per ridurre la latenza di rete o migliorare la velocità di trasmissione effettiva.

L'infrastruttura del Cloud AWS è basata su Regioni e zone di disponibilità. Una regione è un'area fisica del mondo in cui si trovano diverse zone di disponibilità.

Le zone di disponibilità sono composte da uno o più data center singoli provvisti di alimentazione, rete e connettività ridondanti, ognuno in una struttura separata. Le zone di disponibilità consentono

di eseguire applicazioni e database in ambienti di produzione con disponibilità, tolleranza agli errori e scalabilità altrimenti impossibili da ottenere all'interno di un singolo data center.

Scegli la regione o le regioni appropriate per la tua distribuzione in base ad alcuni elementi chiave:

- Ubicazione degli utenti: scegliere una regione vicina agli utenti del tuo carico di lavoro garantisce una latenza minore durante il suo utilizzo.
- Ubicazione dei dati: per le applicazioni con elevati carichi di dati, il collo di bottiglia principale in termini di latenza è il trasferimento dei dati. Il codice dell'applicazione deve essere eseguito il più vicino possibile ai dati.
- Altri vincoli: considera vincoli quali la sicurezza e la conformità.

Amazon EC2 offre gruppi di collocazione per le reti. Un gruppo di collocazione è un raggruppamento logico di istanze per ridurre la latenza o aumentare l'affidabilità. L'utilizzo di gruppi di collocazione con tipi di istanza supportati e un Elastic Network Adapter (ENA) consente ai carichi di lavoro di partecipare a una rete a 25 Gbps a bassa latenza. I gruppi di collocazione sono consigliati per i carichi di lavoro che traggono beneficio da reti a bassa latenza, throughput di rete elevato o entrambi. Utilizzare i gruppi di collocazione consente di ridurre il jitter nelle comunicazioni di rete.

I servizi sensibili alla latenza vengono distribuiti all'edge tramite una rete globale di edge location. Tali edge location forniscono solitamente servizi come CDN (Content Delivery Network) e DNS (Domain Name System). Fornendo questi servizi nell'edge, possono rispondere con una latenza ridotta alle richieste di contenuti o risoluzione DNS. Inoltre, possono offrire servizi geografici come la geotargetizzazione dei contenuti (ossia fornire contenuti diversi in base alla posizione dell'utente finale) o l'instradamento basato sulla latenza, per indirizzare gli utenti alla regione più vicina (latenza minima).

Amazon CloudFront è una CDN globale che consente di accelerare i contenuti statici come le immagini, gli script ed i video, nonché quelli dinamici come le API o le applicazioni Web. Si basa su una rete globale di edge location che memorizzano in cache i contenuti e offrono una connettività di rete ad alte prestazioni agli utenti. Inoltre, CloudFront accelera diverse altre funzioni, come il caricamento dei contenuti e le applicazioni dinamiche. Ciò garantisce prestazioni migliori su tutte le applicazioni che gestiscono il traffico su Internet. Lambda@Edge è una funzionalità di Amazon CloudFront che consente di eseguire il codice più vicino agli utenti del carico di lavoro, migliorando le prestazioni e riducendo la latenza.

Amazon Route 53 è un servizio Web di DNS altamente scalabile e disponibile. È stato progettato per offrire agli sviluppatori e alle aziende un modo estremamente affidabile ed economico per

indirizzare gli utenti finali verso le applicazioni Internet, traducendo i nomi, come www.example.com, in indirizzi IP numerici, come 192.168.2.1, che i computer utilizzano per connettersi tra loro. Route 53 è pienamente conforme all'IPv6.

AWS Outposts è stato progettato per i carichi di lavoro che devono rimanere in sede a causa dei requisiti di latenza e che devono essere eseguiti senza problemi con il resto dei carichi di lavoro in AWS. Gli AWS Outposts sono rack di calcolo e di archiviazione completamente gestiti e configurabili costruiti con hardware progettato per AWS che consentono di eseguire il calcolo e l'archiviazione on-premise, collegandosi al contempo senza soluzione di continuità all'ampia gamma di servizi AWS nel cloud.

Zone locali AWS sono progettate per eseguire carichi di lavoro che richiedono una latenza di pochi millisecondi, come il rendering di video e le applicazioni desktop virtuali con uso intensivo di risorse grafiche. Le zone locali consentono di sfruttare tutti i vantaggi derivanti dalla disponibilità di risorse di calcolo e storage più vicine agli utenti finali.

AWS Wavelength è progettato per fornire applicazioni a bassissima latenza ai dispositivi 5G estendendo l'infrastruttura, i servizi, le API e gli strumenti AWS alle reti 5G. Wavelength incorpora spazio di archiviazione e calcolo all'interno delle reti 5G dei provider di telecomunicazioni per aiutare i carichi di lavoro 5G che richiedono una latenza a una cifra al millisecondo, come i dispositivi IoT, lo streaming di giochi, i veicoli autonomi e la produzione di media in diretta.

Usa i servizi edge per ridurre la latenza e abilitare la memorizzazione nella cache dei contenuti. Al fine di sfruttare tutti i vantaggi offerti da tali approcci, devi assicurarti di avere configurato correttamente il controllo cache sia per DNS sia per HTTP/HTTPS.

Anti-pattern comuni:

- Consolidi tutte le risorse del carico di lavoro in un'unica posizione geografica.
- Hai scelto la regione più vicina alla tua, ma non al carico di lavoro dell'utente finale.

Vantaggi dell'adozione di questa best practice: Assicurati che la tua rete sia disponibile ovunque desideri raggiungere i clienti. L'utilizzo della rete globale privata di AWS garantisce ai clienti l'esperienza di latenza più bassa implementando i carichi di lavoro nelle sedi a loro più vicine.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

Riduzione della latenza selezionando le posizioni corrette: identifica la posizione degli utenti e dei dati. Trai beneficio da Regioni AWS, zone di disponibilità, gruppi di collocazione e posizioni edge per ridurre la latenza.

Risorse

Documenti correlati:

- Istanze ottimizzate per Amazon EBS
- Application Load Balancer
- Reti avanzate su Linux
- Reti avanzate su Windows
- Gruppi di collocamento
- Abilitazione delle reti avanzate con Elastic Network Adapter (ENA) sulle istanze Linux
- Network Load Balancer
- Nuovi prodotti di rete con AWS
- Transit Gateway
- Passaggio all'instradamento basato sulla latenza in Amazon Route 53
- Endpoint VPC
- Log di flusso VPC

Video correlati:

- Connettività ad AWS e architetture di rete AWS ibride (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluzioni di sicurezza scalabili)
- AWS Networking Workshops (Workshop di rete AWS)

PERF05-BP07 Ottimizzazione della configurazione di rete in base ai parametri

Usa i dati raccolti e analizzati per prendere decisioni informate riguardo l'ottimizzazione della configurazione della tua rete. Misura l'impatto di tali cambiamenti e usa le misurazioni per prendere decisioni future.

Abilita i registri di flusso VPC per tutte le reti VPC in uso dal carico di lavoro. I log di flusso VPC sono una funzione che ti permette di acquisire le informazioni sul traffico IP da e per le interfacce di rete nel tuo VPC. I log di flusso VPC possono aiutare a svolgere una serie di compiti; ad esempio, permettono di scoprire perché uno specifico traffico non raggiunge un'istanza, il che, a sua volta, consente di diagnosticare le regole di sicurezza del gruppo troppo restrittive. Puoi utilizzare i log di flusso come strumento di sicurezza per monitorare il traffico che raggiunge l'istanza, per profilare il traffico di rete e per cercare comportamenti di traffico anomali.

Utilizza i parametri di rete per apportare modifiche alla configurazione di rete a mano a mano che il carico di lavoro si evolve. Le reti basate sul cloud possono essere ricostruite rapidamente, perciò, per mantenere l'efficienza delle prestazioni, l'architettura di rete deve evolvere nel tempo.

Anti-pattern comuni:

- Ritieni che tutti i problemi relativi alle prestazioni siano correlati all'applicazione.
- Testi le prestazioni di rete solo da una posizione vicina a quella in cui hai distribuito il carico di lavoro.

Vantaggi dell'adozione di questa best practice:per assicurarti di soddisfare i parametri necessari per il carico di lavoro, devi monitorare i parametri delle prestazioni di rete. Puoi acquisire informazioni sul traffico IP da e verso le interfacce di rete nel tuo VPC e utilizzare questi dati per aggiungere nuove ottimizzazioni o distribuire il carico di lavoro in nuove regioni geografiche.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

Abilita i registri di flusso VPC: i registri di flusso VPC ti consentono di acquisire le informazioni sul traffico IP da e per le interfacce di rete nel tuo VPC. I log di flusso VPC possono aiutare a svolgere una serie di compiti; ad esempio, permettono di scoprire perché uno specifico traffico non raggiunge un'istanza, consentendoti di diagnosticare le regole di sicurezza del gruppo troppo restrittive. Puoi utilizzare i log di flusso come strumento di sicurezza per monitorare il traffico che raggiunge l'istanza, per profilare il traffico di rete e per cercare comportamenti di traffico anomali.

Abilita i parametri appropriati per le opzioni di rete: assicurati di selezionare i parametri di rete adeguati al tuo carico di lavoro. Puoi abilitare i parametri per gateway VPC NAT, gateway di transito e tunnel VPN.

Risorse

Documenti correlati:

- Istanze ottimizzate per Amazon EBS
- Application Load Balancer
- Reti avanzate su Linux
- Reti avanzate su Windows
- Gruppi di collocamento
- Abilitazione delle reti avanzate con Elastic Network Adapter (ENA) sulle istanze Linux
- Network Load Balancer
- Nuovi prodotti di rete con AWS
- Transit Gateway
- Passaggio all'instradamento basato sulla latenza in Amazon Route 53
- Endpoint VPC
- Log di flusso VPC
- Monitoring your global and core networks with Amazon Cloudwatch metrics (Monitoraggio delle reti globali e core con i parametri di Amazon CloudWatch)
- Continuously monitor network traffic and resources (Monitoraggio costante di traffico e risorse di rete)

Video correlati:

- Connectivity to AWS and hybrid AWS network architectures (NET317-R1)
- Optimizing Network Performance for Amazon EC2 Instances (CMP308-R1)
- Monitoring and troubleshooting network traffic
- Simplify Traffic Monitoring and Visibility with Amazon VPC Traffic Mirroring

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions (AWS Transit Gateway e soluzioni di sicurezza scalabili)
- AWS Networking Workshops (Workshop di rete AWS)
- AWS Network Monitoring

Revisione

Domanda

- PERF 6 In che modo fai evolvere il carico di lavoro per sfruttare le nuove versioni?

PERF 6 In che modo fai evolvere il carico di lavoro per sfruttare le nuove versioni?

Quando si progettano carichi di lavoro, le opzioni tra cui scegliere sono limitate. Tuttavia, nel tempo diventano disponibili nuove tecnologie e nuovi approcci che potrebbero migliorare le prestazioni.

Best practice

- PERF06-BP01 Mantenersi aggiornati sui nuovi servizi e sulle nuove risorse
- PERF06-BP02 Definizione di un processo per migliorare le prestazioni del carico di lavoro
- PERF06-BP03 Evoluzione delle prestazioni del carico di lavoro nel corso del tempo

PERF06-BP01 Mantenersi aggiornati sui nuovi servizi e sulle nuove risorse

Valuta i modi per migliorare le prestazioni man mano che nuovi servizi, modelli di progettazione e offerte di prodotti diventano disponibili. Determina come possono migliorare le prestazioni o aumentare l'efficienza del carico di lavoro tramite una valutazione, una discussione interna o un'analisi esterna.

Definisci un processo per valutare gli aggiornamenti, le nuove funzioni e i servizi rilevanti per il tuo carico di lavoro. Ad esempio, crea un proof of concept che utilizza le nuove tecnologie o consultati con un gruppo interno. Quando provi nuove idee o servizi, esegui i test delle prestazioni per misurare l'impatto del carico di lavoro sulle prestazioni. Usa l'Infrastruttura come codice (IaC) e una cultura DevOps per sfruttare la capacità di testare nuove idee o tecnologie frequentemente con costi o rischi minimi.

Risultato desiderato: hai creato la documentazione per l'inventario dei componenti, lo schema di progettazione e le caratteristiche del carico di lavoro. Utilizzi tale documentazione per creare un

elenco di sottoscrizioni per informare il tuo team su aggiornamenti del servizio, funzionalità e nuovi prodotti. Hai identificato gli stakeholder dei componenti che valuteranno le nuove versioni e forniranno un suggerimento per l'impatto sull'azienda e le priorità.

Anti-pattern comuni:

- puoi esaminare nuove opzioni e servizi solo quando il carico di lavoro non soddisfa i requisiti di prestazione.
- Ritieni che tutte le nuove offerte di prodotti non siano utili per il tuo carico di lavoro.
- Scegli sempre di creare invece di acquistare quando migliori il tuo carico di lavoro.

Vantaggi dell'adozione di questa best practice: tenendo in considerazione le nuove offerte di servizi o prodotti, puoi migliorare le prestazioni e l'efficienza del tuo carico di lavoro, ridurre i costi dell'infrastruttura e contenere l'impegno richiesto per mantenere i servizi.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

Definisci un processo per valutare gli aggiornamenti, le nuove funzioni e i servizi AWS. Ad esempio, crea proof of concept che utilizzano le nuove tecnologie. Quando metti alla prova nuove idee o servizi, esegui test delle prestazioni per misurare l'impatto sull'efficienza o sulle prestazioni del carico di lavoro. Sfrutta la flessibilità offerta da AWS per condurre test frequenti su nuove idee e tecnologie con costi o rischi minimi.

Passaggi dell'implementazione

1. Documenta le soluzioni del tuo carico di lavoro. Usa la tua soluzione di database di gestione delle configurazioni (CMDB, Configuration Management Database) per documentare l'inventario e classificare i servizi e le dipendenze. Usa strumenti come AWS Config per ottenere l'elenco di tutti i servizi in AWS utilizzati dal tuo carico di lavoro.
2. Usa una strategia di applicazione dei tag per documentare i proprietari per ogni componente e categoria del carico di lavoro. Ad esempio, se stai attualmente utilizzando Amazon RDS come soluzione di database, rivolgiti all'amministratore di database (DBA) assegnato e documentato come proprietario per la valutazione e la ricerca di nuovi servizi e aggiornamenti.
3. Identifica le novità e le fonti di aggiornamento relative ai componenti del carico di lavoro. Nell'esempio Amazon RDS menzionato in precedenza, il proprietario della categoria deve iscriversi al blog Novità di AWS per i prodotti che corrispondono al componente del carico di lavoro. Puoi

iscriverti al feed RSS o gestire le tue sottoscrizioni e-mail. Monitora gli aggiornamenti al database Amazon RDS che utilizzi, le funzionalità introdotte, le istanze rilasciate e i nuovi prodotti come Amazon Aurora Serverless. Monitora i blog di settore, i prodotti e i fornitori su cui si basa il componente.

4. Documenta il processo di valutazione di aggiornamenti e nuovi servizi. Fornisci ai proprietari della categoria il tempo e lo spazio necessari per ricercare, testare, sperimentare e convalidare aggiornamenti e nuovi servizi. Fai riferimento ai requisiti aziendali e ai KPI documentati per stabilire la priorità dell'aggiornamento che avrà un impatto positivo sull'azienda.

Livello di impegno per il piano di implementazione: per adottare questa best practice devi essere a conoscenza dei componenti del carico di lavoro correnti, identificare i proprietari della categoria e individuare le origini degli aggiornamenti del servizio. È richiesto un livello di impegno basso per iniziare, ma è un processo continuo che potrebbe evolversi e migliorare nel tempo.

Risorse

Documenti correlati:

- Blog AWS
- Novità di AWS

Video correlati:

- Canale YouTube degli eventi AWS
- Canale YouTube dei Tech talk online di AWS
- Canale YouTube di Amazon Web Services

Esempi correlati:

- Github AWS
- AWS Skill Builder

PERF06-BP02 Definizione di un processo per migliorare le prestazioni del carico di lavoro

Definisci un processo per valutare i nuovi servizi, i modelli di progettazione, i tipi di risorse e le configurazioni man mano che diventano disponibili. Ad esempio, esegui test delle prestazioni esistenti sulle nuove offerte di istanze per determinare il loro potenziale per migliorare il carico di lavoro.

Le prestazioni del carico di lavoro presentano alcuni vincoli principali. Documentali, in modo da sapere quali tipi di innovazione potrebbero migliorare le prestazioni del carico di lavoro. Utilizza queste informazioni quando vieni a conoscenza di nuovi servizi e tecnologie, man mano che si rendono disponibili, in modo da identificare le soluzioni per ovviare ai vincoli o ai colli di bottiglia.

Anti-pattern comuni:

- Ritieni che la tua architettura attuale diventerà statica e non si aggiornerà mai nel corso del tempo.
- Introduci modifiche all'architettura nel tempo senza dei parametri che le giustifichino.

Vantaggi dell'adozione di questa best practice: Definire un processo per apportare modifiche all'architettura consente ai dati raccolti di influenzare la progettazione del carico di lavoro nel corso del tempo.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

Identificazione dei vincoli di prestazione chiave per il tuo carico di lavoro: documenta i vincoli di prestazione del carico di lavoro in modo da sapere quali tipi di innovazione migliorano le prestazioni del carico di lavoro.

Risorse

Documenti correlati:

- Blog AWS
- Novità di AWS

Video correlati:

- Canale YouTube degli eventi AWS
- Canale YouTube dei Tech talk online di AWS
- Canale YouTube di Amazon Web Services

Esempi correlati:

- Github AWS
- AWS Skill Builder

PERF06-BP03 Evoluzione delle prestazioni del carico di lavoro nel corso del tempo

Come organizzazione, utilizza le informazioni raccolte durante il processo di valutazione per gestire attivamente l'adozione di nuovi servizi o risorse quando diventano disponibili.

Utilizza le informazioni ottenute con la valutazione di nuovi servizi o tecnologie per gestire il cambiamento. Man mano che la tua azienda o il tuo carico di lavoro evolve, anche le prestazioni devono cambiare. Sfrutta i dati raccolti dai parametri riguardanti il carico di lavoro per valutare le aree in cui è possibile ottenere i miglioramenti più significativi in termini di efficienza o prestazioni e adotta in modo proattivo nuovi servizi e tecnologie per tenere il passo con la domanda.

Anti-pattern comuni:

- Ritieni che la tua architettura attuale diventerà statica e non si aggiornerà mai nel corso del tempo.
- Introduci modifiche all'architettura nel tempo senza dei parametri che le giustifichino.
- Cambi architettura solo perché viene utilizzata da tutto il resto del settore.

Vantaggi dell'adozione di questa best practice: Per ottimizzare le prestazioni e i costi del carico di lavoro, è necessario valutare tutti i software e i servizi disponibili per determinare quelli appropriati per il carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

Evolvi il carico di lavoro nel tempo: utilizza le informazioni ottenute con la valutazione di nuovi servizi o tecnologie per gestire il cambiamento. Man mano che la tua azienda o il tuo carico di lavoro evolve, anche le prestazioni devono cambiare. Sfrutta i dati raccolti dai parametri riguardanti il carico di lavoro per valutare le aree in cui è possibile ottenere i miglioramenti più significativi in termini di efficienza o prestazioni e adotta in modo proattivo nuovi servizi e tecnologie per tenere il passo con la domanda.

Risorse

Documenti correlati:

- Blog AWS
- Novità di AWS

Video correlati:

- Canale YouTube degli eventi AWS
- Canale YouTube dei Tech talk online di AWS
- Canale YouTube di Amazon Web Services

Esempi correlati:

- Github AWS
- AWS Skill Builder

Monitoraggio

Domanda

- PERF 7 In che modo monitori le risorse per garantirne le prestazioni?

PERF 7 In che modo monitori le risorse per garantirne le prestazioni?

Le prestazioni del sistema possono peggiorare nel tempo. Monitora le prestazioni del sistema per identificare l'eventuale riduzione delle prestazioni e rimediare a fattori interni o esterni, come il sistema operativo o il carico dell'applicazione.

Best practice

- PERF07-BP01 Registrazione dei parametri relativi alle prestazioni
- PERF07-BP02 Analisi dei parametri in caso di eventi o incidenti
- PERF07-BP03 Individuazione degli indicatori chiave di prestazione (KPI) per misurare le prestazioni del carico di lavoro
- PERF07-BP04 Utilizzo del monitoraggio per generare notifiche basate su allarmi
- PERF07-BP05 Analisi dei parametri a intervalli regolari
- PERF07-BP06 Monitoraggio e allarmi proattivi

PERF07-BP01 Registrazione dei parametri relativi alle prestazioni

Utilizza un servizio di monitoraggio e osservazione per registrare i parametri correlati alle prestazioni. Esempi di parametri sono le transazioni record del database, le query lente, la latenza di I/O, la velocità di trasmissione effettiva delle richieste HTTP, la latenza del servizio o altri dati chiave.

Identifica i parametri relativi alle prestazioni rilevanti per il tuo carico di lavoro e registrali. Questi dati sono importanti per riuscire a identificare quali componenti influiscono sulle prestazioni o sull'efficienza complessive del carico di lavoro.

Partendo dall'esperienza del cliente, identifica quali sono i parametri rilevanti. Per ciascuno di essi, identifica l'obiettivo, l'approccio per la misurazione e la priorità. Utilizza questi elementi per creare allarmi e notifiche per affrontare in modo proattivo i problemi correlati alle prestazioni.

Anti-pattern comuni:

- Monitori solo i parametri a livello di sistema operativo per ottenere informazioni approfondite sul carico di lavoro.
- Pianifichi le tue esigenze di calcolo in base ai requisiti di picco del carico di lavoro.

Vantaggi dell'adozione di questa best practice: Per ottimizzare le prestazioni e l'utilizzo delle risorse, è necessario disporre di una vista operativa unificata dei tuoi principali indicatori prestazionali. Puoi creare pannelli di controllo ed eseguire calcoli parametrici sui dati per ottenere informazioni operative e di utilizzo.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Identifica i parametri prestazionali rilevanti per il tuo carico di lavoro e registrali. Questi dati aiutano a identificare quali componenti influiscono sulle prestazioni o sull'efficienza complessive del carico di lavoro.

Identificazione dei parametri di prestazione: utilizza l'esperienza del cliente per identificare i parametri più importanti. Per ciascuno di essi, identifica l'obiettivo, l'approccio per la misurazione e la priorità. Utilizza questi punti dati per creare allarmi e notifiche per affrontare in modo proattivo i problemi correlati alle prestazioni.

Risorse

Documenti correlati:

- Documentazione di CloudWatch
- Raccolta di parametri e registri da istanze Amazon EC2 e da server on-premise con l'agente di CloudWatch

- Pubblicazione di parametri personalizzati
- Monitoraggio, registrazione e prestazioni – Partner APN
- X-Ray Documentation (Documentazione di X-Ray)
- Usare Amazon CloudWatch RUM

Video correlati:

- Elimina il caos: acquisisci visibilità e approfondimenti operativi (MGT301-R1)
- Application Performance Management on AWS (Gestione delle prestazioni delle applicazioni su AWS)
- Creazione di un piano di monitoraggio

Esempi correlati:

- Level 100: Monitoring with CloudWatch Dashboards (Livello 100: Monitoraggio con i pannelli di controllo CloudWatch)
- Level 100: Monitoring Windows EC2 instance with CloudWatch Dashboards (Livello 100: Monitoraggio dell'istanza EC2 di Windows con i pannelli di controllo CloudWatch)
- Level 100: Monitoring an Amazon Linux EC2 instance with CloudWatch Dashboards (Livello 100: Monitoraggio dell'istanza EC2 di Amazon Linux con i pannelli di controllo CloudWatch)

PERF07-BP02 Analisi dei parametri in caso di eventi o incidenti

In risposta a nel corso di un evento o un incidente, utilizza pannelli di controllo o report di monitoraggio per comprendere e diagnosticare l'impatto. Queste viste forniscono informazioni sulle parti del carico di lavoro le cui prestazioni non raggiungono i livelli previsti.

Durante lo studio dei casi utente critici per la tua architettura, includi i requisiti relativi alle prestazioni, specificando ad esempio con quale rapidità deve essere eseguito ogni scenario. Per questi scenari critici, implementa percorsi utente con script aggiuntivi per chiarire esattamente quali sono le loro prestazioni rispetto ai requisiti.

Anti-pattern comuni:

- Parti dal presupposto che gli eventi relativi alle prestazioni siano problemi unici e correlati solo ad anomalie.

- Valuti i parametri delle prestazioni esistenti solo quando rispondi a eventi relativi alle prestazioni.

Vantaggi dell'adozione di questa best practice: Per stabilire se il carico di lavoro funziona ai livelli previsti, è necessario rispondere agli eventi relativi alle prestazioni raccogliendo ulteriori dati dei parametri affinché siano analizzati. Questi dati vengono utilizzati per comprendere l'impatto dell'evento relativo alle prestazioni e suggerire modifiche per migliorare le prestazioni del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Assegnazione della priorità a problemi di esperienza per le storie utente critiche: quando scrivi storie critiche dell'utente per la tua architettura, includi i requisiti di prestazione, ad esempio specificando la velocità di esecuzione di ogni storia critica. Per questi scenari critici, implementa percorsi utente con script aggiuntivi per chiarire esattamente quali sono le prestazioni dei casi utente rispetto ai requisiti.

Risorse

Documenti correlati:

- Documentazione di CloudWatch
- Amazon CloudWatch Synthetics
- Monitoraggio, registrazione e prestazioni – Partner APN
- X-Ray Documentation (Documentazione di X-Ray)

Video correlati:

- Elimina il caos: acquisisci visibilità e approfondimenti operativi (MGT301-R1)
- Optimize applications through Amazon CloudWatch RUM (Ottimizzazione delle applicazioni tramite Amazon CloudWatch RUM)
- Demo di Amazon CloudWatch Synthetics

Esempi correlati:

- Measure page load time with Amazon CloudWatch Synthetics (Misurare il tempo di caricamento della pagina con Amazon CloudWatch Synthetics)
- Amazon CloudWatch RUM Web Client (Client Web Amazon CloudWatch RUM)

PERF07-BP03 Individuazione degli indicatori chiave di prestazione (KPI) per misurare le prestazioni del carico di lavoro

Individua gli indicatori chiave di prestazione (KPI) per misurare le prestazioni del carico di lavoro. Gli indicatori chiave di prestazione aiutano a misurare l'integrità di un carico di lavoro in relazione a un obiettivo aziendale. Gli indicatori chiave di prestazione consentono ai team aziendali e di ingegneri di allinearsi sulla misurazione degli obiettivi e delle strategie e sul modo in cui questi si combinano per produrre risultati aziendali. Gli indicatori chiave di prestazione devono essere rivisti quando cambiano gli obiettivi aziendali, le strategie o i requisiti degli utenti finali.

Ad esempio, il carico di lavoro di un sito Web può utilizzare il tempo di caricamento della pagina come indicazione delle prestazioni complessive. Questo parametro è uno dei molteplici dati che misurano l'esperienza dell'utente finale. Oltre a identificare le soglie di tempo di caricamento della pagina, è necessario documentare il risultato atteso o il rischio aziendale se le prestazioni non vengono rispettate. Un lungo tempo di caricamento della pagina si ripercuote direttamente sugli utenti finali, diminuisce la loro esperienza d'uso e può portare a una perdita di clienti. Quando definisci le soglie degli indicatori chiave di prestazione, devi combinare sia i benchmark di settore sia le aspettative degli utenti finali. Ad esempio, se l'attuale benchmark del settore prevede il caricamento di una pagina Web entro un periodo di tempo di due secondi, ma gli utenti finali si aspettano che la pagina Web venga caricata entro un periodo di tempo di un secondo, allora devi prendere in considerazione entrambi i dati al momento di stabilire l'indicatore chiave di prestazione (KPI). Un altro esempio di KPI può essere quello di soddisfare le esigenze di prestazioni interne. Puoi stabilire una soglia KPI per la generazione di report sulle vendite entro un giorno lavorativo dalla generazione dei dati di produzione. Questi report possono influenzare direttamente le decisioni quotidiane e i risultati aziendali.

Risultato desiderato: La definizione dei KPI coinvolge diversi dipartimenti e stakeholder. Il team deve valutare i KPI del carico di lavoro utilizzando dati granulari in tempo reale e dati storici di riferimento e creare pannelli di controllo che eseguano calcoli metrici sui dati KPI per ricavare informazioni operative e di utilizzo. I KPI devono essere documentati per spiegare i KPI e le soglie concordati che supportano gli obiettivi e le strategie aziendali e che sono mappati sui parametri da monitorare. I KPI identificano i requisiti di prestazioni, vengono rivisti intenzionalmente e sono frequentemente condivisi e compresi da tutti i team. I rischi e i compromessi sono chiaramente identificati e si comprende l'impatto sull'azienda in caso di mancato raggiungimento delle soglie KPI.

Anti-pattern comuni:

- Si monitorano i parametri a livello di sistema solo per avere una visione del carico di lavoro e non si comprendono gli impatti aziendali di tali parametri.

- Ritieni che i KPI siano già in fase di pubblicazione e condivisi come dati parametrici standard.
- I KPI vengono definiti ma non vengono condivisi con tutti i team.
- Non si definisce un KPI quantitativo e misurabile.
- Mancato allineamento dei KPI con obiettivi o strategie aziendali.

Vantaggi dell'adozione di questa best practice: L'identificazione di parametri specifici che rappresentano l'integrità del carico di lavoro aiuta ad allineare i team sulle loro priorità e a definire i risultati aziendali di successo. La condivisione di tali parametri con tutti i reparti fornisce visibilità e allineamento su soglie, aspettative e impatto aziendale.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Tutti i reparti e i team aziendali che hanno un impatto sull'integrità del carico di lavoro devono contribuire alla definizione dei KPI. Collaborazione, tempistiche, documentazione e informazioni relative ai KPI dell'organizzazione devono essere guidate da una sola persona. Questo responsabile unico spesso condividerà gli obiettivi e le strategie aziendali e assegnerà agli stakeholder aziendali i compiti di creare i KPI nei rispettivi reparti. Una volta definiti i KPI, il team operativo contribuirà spesso a definire i parametri che supporteranno e informeranno il successo dei diversi KPI. I KPI sono efficaci solo se tutti i membri del team che supportano un carico di lavoro sono consapevoli della loro esistenza.

Passaggi dell'implementazione

1. Identificare e documentare gli stakeholder aziendali.
2. Identificare gli obiettivi e le strategie aziendali.
3. Esaminare i KPI comuni del settore che si allineano agli obiettivi e alle strategie aziendali.
4. Esaminare le aspettative degli utenti finali sul carico di lavoro.
5. Definire e documentare i KPI che supportano gli obiettivi e le strategie aziendali.
6. Identificare e documentare le strategie di compromesso approvate per soddisfare i KPI.
7. Identificare e documentare i parametri che informeranno i KPI.
8. Identificare e documentare le soglie KPI per il livello di gravità o di allarme.
9. Identificare e documentare il rischio e l'impatto se il KPI non viene raggiunto.

10. Identificare la frequenza di revisione per KPI.
11. Comunicare la documentazione KPI a tutti i team che supportano il carico di lavoro.

Livello di impegno per la guida all'implementazione: La definizione e la comunicazione dei KPI comportano basso livello di impegno. Questo può essere fatto nell'arco di qualche settimana, incontrando gli stakeholder aziendali, rivedendo gli obiettivi, le strategie e i parametri del carico di lavoro.

Risorse

Documenti correlati:

- Documentazione di CloudWatch
- Monitoraggio, registrazione e prestazioni – Partner APN
- X-Ray Documentation (Documentazione di X-Ray)
- Utilizzo dei pannelli di controllo Amazon CloudWatch
- Amazon QuickSight KPIs (KPI di Amazon QuickSight)

Video correlati:

- AWS re:Invent 2019: Scaling up to your first 10 million users (Dimensionare le risorse per i primi 10 milioni di utenti) (ARC211-R)
- Elimina il caos: acquisisci visibilità e approfondimenti operativi (MGT301-R1)
- Creazione di un piano di monitoraggio

Esempi correlati:

- Creating a dashboard with Amazon QuickSight (Creazione di un pannello di controllo con Amazon QuickSight)

PERF07-BP04 Utilizzo del monitoraggio per generare notifiche basate su allarmi

Avvalendoti degli indicatori chiave di prestazione (KPI) relativi alle prestazioni che hai identificato, utilizza un sistema di monitoraggio che genera automaticamente allarmi quando queste misurazioni sono al di fuori dei limiti previsti.

Amazon CloudWatch può raccogliere i parametri per tutte le risorse dell'architettura. Puoi anche raccogliere e pubblicare parametri personalizzati per ottenere parametri aziendali o derivati. Utilizza CloudWatch o un servizio di monitoraggio di terze parti per configurare allarmi che si attivino al superamento delle soglie impostate; gli allarmi segnalano che un parametro si trova al di fuori dei limiti previsti.

Anti-pattern comuni:

- Affidi al personale il controllo dei parametri e la risposta quando si verifica un problema.
- Ti affidi esclusivamente a runbook operativi, quando potresti attivare flussi di lavoro serverless per svolgere la stessa attività.

Vantaggi dell'adozione di questa best practice: Puoi impostare allarmi e automatizzare le operazioni in base a soglie predefinite o ad algoritmi di machine learning che identificano comportamenti anomali nei parametri. Questi stessi allarmi possono anche attivare flussi di lavoro serverless, che possono modificare le caratteristiche prestazionali del carico di lavoro (ad esempio, aumentare la capacità di elaborazione, modificare la configurazione del database).

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

Parametri di monitoraggio: Amazon CloudWatch può raccogliere i parametri per tutte le risorse dell'architettura. Puoi raccogliere e pubblicare parametri personalizzati per ottenere parametri aziendali o derivati. Utilizza CloudWatch o un servizio di monitoraggio di terze parti per impostare allarmi che indicano quando le soglie vengono superate.

Risorse

Documenti correlati:

- Documentazione di CloudWatch
- Monitoraggio, registrazione e prestazioni – Partner APN
- X-Ray Documentation (Documentazione di X-Ray)
- Using Alarms and Alarm Actions in CloudWatch (Utilizzo degli allarmi e delle azioni di allarme in Amazon CloudWatch)

Video correlati:

- AWS re:Invent 2019: Scaling up to your first 10 million users (Dimensionare le risorse per i primi 10 milioni di utenti) (ARC211-R)
- Elimina il caos: acquisisci visibilità e approfondimenti operativi (MGT301-R1)
- Creazione di un piano di monitoraggio
- Using AWS Lambda with Amazon CloudWatch Events (Utilizzo di AWS Lambda con Amazon CloudWatch Events)

Esempi correlati:

- Cloudwatch Logs Customize Alarms (Allarmi personalizzabili per i registri CloudWatch)

PERF07-BP05 Analisi dei parametri a intervalli regolari

Come manutenzione ordinaria o in risposta a eventi o incidenti, esamina quali parametri vengono raccolti. Stabilisci quali di questi parametri sono fondamentali per risolvere i problemi e quali altri parametri aggiuntivi, se monitorati, possono contribuire a identificare, affrontare o prevenire i problemi.

nell'ambito della risposta a incidenti ed eventi, valuta quali parametri sono stati utili per affrontare il problema e quali sarebbero stati utili ma non sono attualmente misurati. Queste considerazioni ti aiuteranno a migliorare la qualità dei parametri raccolti, per prevenire o risolvere più rapidamente gli incidenti futuri.

Anti-pattern comuni:

- Lasci che i parametri rimangano in uno stato di allarme per un lungo periodo di tempo.
- Crei allarmi che non sono utilizzabili da un sistema di automazione.

Vantaggi dell'adozione di questa best practice: Esamina continuamente i parametri raccolti per verificare che identifichino, risolvano o prevengano adeguatamente i problemi. I parametri possono anche diventare obsoleti se lasciati in uno stato di allarme per un lungo periodo di tempo.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

Migliora costantemente la raccolta e il monitoraggio dei parametri: nell'ambito della risposta a incidenti ed eventi, valuta quali parametri sono stati utili per affrontare il problema e quali sarebbero

stati utili ma non sono attualmente misurati. Queste considerazioni ti aiuteranno a migliorare la qualità dei parametri raccolti, in modo da prevenire o risolvere più rapidamente gli incidenti futuri.

Risorse

Documenti correlati:

- Documentazione di CloudWatch
- Raccolta di parametri e registri da istanze Amazon EC2 e da server on-premise con l'agente di CloudWatch
- Monitoraggio, registrazione e prestazioni – Partner APN
- X-Ray Documentation (Documentazione di X-Ray)

Video correlati:

- Elimina il caos: acquisisci visibilità e approfondimenti operativi (MGT301-R1)
- Application Performance Management on AWS (Gestione delle prestazioni delle applicazioni su AWS)
- Creazione di un piano di monitoraggio

Esempi correlati:

- Creating a dashboard with Amazon QuickSight (Creazione di un pannello di controllo con Amazon QuickSight)
- Level 100: Monitoring with CloudWatch Dashboards (Livello 100: Monitoraggio con i pannelli di controllo CloudWatch)

PERF07-BP06 Monitoraggio e allarmi proattivi

Utilizza indicatori chiave di prestazioni (KPI), in combinazione con sistemi di monitoraggio e allarmi, per risolvere in modo proattivo i problemi correlati alle prestazioni. Laddove possibile, utilizza gli allarmi per attivare operazioni automatizzate per risolvere i problemi. Se non è possibile rispondere in modo automatizzato, inoltra l'allarme a coloro che possono intervenire. Ad esempio, puoi implementare un sistema in grado di prevedere i valori attesi per i KPI e di inviare allarmi qualora essi oltrepassino determinate soglie, oppure uno strumento che arresta o esegue automaticamente il rollback delle distribuzioni nel caso in cui i valori dei KPI si discostino dai valori attesi.

Implementa processi che forniscono visibilità sulle prestazioni durante l'esecuzione del carico di lavoro. Crea pannelli di controllo del monitoraggio e stabilisci norme di riferimento per le aspettative riguardanti le prestazioni, per determinare se il carico di lavoro ha prestazioni ottimali.

Anti-pattern comuni:

- Consenti solo al personale operativo di apportare modifiche operative al carico di lavoro.
- Lasci che tutti gli allarmi giungano direttamente al team operativo senza alcuna correzione proattiva.

Vantaggi dell'adozione di questa best practice: La correzione proattiva delle azioni di allarme consente al personale di supporto di concentrarsi sugli elementi che non sono attivabili automaticamente. In questo modo, il personale operativo non viene sovraccaricato da tutti gli allarmi e si concentra, invece, solo sugli allarmi critici.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

Monitoraggio delle prestazioni durante le operazioni: implementa processi che forniscono visibilità sulle prestazioni durante l'esecuzione del carico di lavoro. Crea pannelli di controllo di monitoraggio e stabilisci una baseline per le previsioni delle prestazioni.

Risorse

Documenti correlati:

- Documentazione di CloudWatch
- Monitoraggio, registrazione e prestazioni – Partner APN
- X-Ray Documentation (Documentazione di X-Ray)
- Using Alarms and Alarm Actions in CloudWatch (Utilizzo degli allarmi e delle azioni di allarme in Amazon CloudWatch)

Video correlati:

- Elimina il caos: acquisisci visibilità e approfondimenti operativi (MGT301-R1)
- Application Performance Management on AWS (Gestione delle prestazioni delle applicazioni su AWS)

- Creazione di un piano di monitoraggio
- Utilizzo di AWS Lambda con Amazon CloudWatch Events

Esempi correlati:

- Cloudwatch Logs Customize Alarms (Allarmi personalizzabili per i registri CloudWatch)

Compromessi

Domanda

- PERF 8 A quali compromessi ricorri per migliorare le prestazioni?

PERF 8 A quali compromessi ricorri per migliorare le prestazioni?

Quando si progettano soluzioni, determinare i compromessi ti consente di selezionare un approccio ottimale. Spesso è possibile migliorare le prestazioni accettando compromessi in termini di coerenza, durata e spazio a favore di tempo e latenza.

Best practice

- PERF08-BP01 Definizione delle aree in cui le prestazioni sono più importanti
- PERF08-BP02 Studio dei servizi e dei modelli di progettazione
- PERF08-BP03 Identificazione dell'impatto dei compromessi sui clienti e sull'efficienza
- PERF08-BP04 Misurazione dell'impatto dei miglioramenti delle prestazioni
- PERF08-BP05 Scelta di più strategie relative alle prestazioni

PERF08-BP01 Definizione delle aree in cui le prestazioni sono più importanti

Comprendi e identifica le aree in cui l'aumento delle prestazioni del carico di lavoro determinerà un impatto positivo sull'efficienza o sull'esperienza del cliente. Ad esempio, un sito web che ha una grande quantità di interazione con i clienti può trarre vantaggio dall'utilizzo dei servizi edge per spostare la distribuzione di contenuti più vicino ai clienti.

Risultato desiderato: aumenta l'efficienza delle prestazioni comprendendo l'architettura, gli schemi di traffico e gli schemi di accesso ai dati e identifica la latenza e i tempi di elaborazione. Identifica i potenziali colli di bottiglia che potrebbero influire sull'esperienza del cliente man mano che il carico di

lavoro aumenta. Quando identifichi queste aree, individua quale soluzione puoi distribuire per evitare tali problemi di prestazioni.

Anti-pattern comuni:

- Ritieni che i parametri di calcolo standard come `CPUUtilization` o la pressione della memoria siano sufficienti per rilevare i problemi di prestazioni.
- Utilizzi solo i parametri predefiniti registrati dal software di monitoraggio selezionato.
- Rivedi i parametri solo quando c'è un problema.

Vantaggi dell'adozione di questa best practice: l'individuazione delle aree critiche delle prestazioni consente ai proprietari del carico di lavoro di monitorare i KPI e dare priorità ai miglioramenti ad alto impatto.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

Configura il tracciamento end-to-end per identificare gli schemi di traffico, la latenza e le aree con prestazioni critiche. Monitora gli schemi di accesso ai dati per query lente o dati scarsamente frammentati e partizionati. Identifica le aree vincolate del carico di lavoro utilizzando il test o il monitoraggio del carico.

Passaggi dell'implementazione

1. Configura il monitoraggio end-to-end per acquisire tutti i componenti e i parametri del carico di lavoro.
   - Utilizza Amazon CloudWatch Real-User Monitoring (RUM) per acquisire i parametri delle prestazioni delle applicazioni da sessioni lato client e front-end di utenti reali.
   - Configura AWS X-Ray per tenere traccia del traffico nei livelli dell'applicazione e identificare la latenza tra componenti e dipendenze. Utilizza le mappe del servizio X-Ray per osservare le relazioni e la latenza tra i componenti del carico di lavoro.
   - Utilizza gli approfondimenti sulle prestazioni di Amazon Relational Database Service per osservare i parametri delle prestazioni del database e identificare le prestazioni da migliorare.
   - Utilizza il monitoraggio avanzato di Amazon RDS per osservare i parametri delle prestazioni del sistema operativo del database.
   - Raccogli i parametri CloudWatch per ogni componente e servizio del carico di lavoro e identifica quali parametri influiscono sull'efficienza delle prestazioni.

- Configura Amazon DevOps Guru per ulteriori approfondimenti e suggerimenti sulle prestazioni.

2. Esegui i test per generare parametri, identificare schemi di traffico, colli di bottiglia e aree con prestazioni critiche.
   - Configura i canary sintetici di CloudWatch per emulare le attività dell'utente basate su browser in modo programmatico usando processi `cron` o espressioni di valutazione per generare parametri coerenti nel tempo.
   - Utilizza la soluzione AWS Distributed Load Testing (Test di carico distribuito in AWS) per generare picchi di traffico o testare il carico di lavoro al tasso di crescita previsto.
3. Valuta i parametri e i dati di telemetria per identificare le aree critiche delle prestazioni. Esamina queste aree con il tuo team per determinare il monitoraggio e le soluzioni per evitare i colli di bottiglia.
4. Sperimenta i miglioramenti delle prestazioni e valuta tali modifiche con i dati.
   - Utilizza CloudWatch Evidently per testare i nuovi miglioramenti e l'impatto delle prestazioni sul carico di lavoro.

Livello di impegno per il piano di implementazione: per adottare questa best practice occorre esaminare i parametri end-to-end ed essere a conoscenza delle prestazioni attuali del carico di lavoro. È richiesto un livello di impegno moderato per configurare il monitoraggio end-to-end e identificare le aree con prestazioni critiche.

Risorse

Documenti correlati:

- Amazon Builders' Library
- X-Ray Documentation (Documentazione di X-Ray)
- Usare Amazon CloudWatch RUM
- Amazon DevOps Guru
- CloudWatch RUM e X-Ray

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- Demo di Amazon CloudWatch Synthetics

Esempi correlati:

- Measure page load time with Amazon CloudWatch Synthetics (Misurare il tempo di caricamento della pagina con Amazon CloudWatch Synthetics)
- Amazon CloudWatch RUM Web Client (Client Web Amazon CloudWatch RUM)
- SDK X-Ray per Node.js
- SDK X-Ray per Phyton
- SDK X-Ray per Java
- SDK X-Ray per .Net
- SDK X-Ray per Ruby
- Daemon X-Ray
- Test di carico distribuito in AWS

PERF08-BP02 Studio dei servizi e dei modelli di progettazione

Ricerca e analizza i vari servizi e modelli di progettazione che permettono di migliorare le prestazioni del carico di lavoro. Nell'ambito dell'analisi, identifica gli elementi sui quali potresti accettare compromessi per ottenere prestazioni più elevate. Ad esempio, l'utilizzo di un servizio di cache può contribuire a ridurre il carico dei sistemi di database. Tuttavia, la memorizzazione nella cache può introdurre ulteriore consistenza e richiede un notevole sforzo tecnico a livello di esecuzione dell'implementazione in modo conforme ai requisiti aziendali e alle aspettative dei clienti.

Risultato desiderato: la ricerca di modelli di progettazione ti permetterà di scegliere la progettazione dell'architettura che supporterà il sistema con le migliori prestazioni. Scopri quali sono le opzioni di configurazione relative alla rete disponibili e come possono influire sul carico di lavoro. L'ottimizzazione delle prestazioni del carico di lavoro dipende dalla comprensione della modalità in cui tali opzioni interagiscono con l'architettura, nonché dell'impatto che avranno sia sulle prestazioni misurate sia sulle prestazioni percepite dagli utenti finali.

Anti-pattern comuni:

- Ritieni che tutte le tradizionali strategie informatiche relative alle prestazioni dei carichi di lavoro siano le più adatte ai carichi di lavoro nel cloud.
- Crei e gestisci soluzioni di caching invece di utilizzare i servizi gestiti.
- Utilizzi lo stesso modello di progettazione per tutti i carichi di lavoro senza valutare quale modello ne migliorerà le prestazioni.

Vantaggi dell'adozione di questa best practice: La corretta selezione del modello di progettazione e dei servizi per il carico di lavoro ti consentirà di ottimizzare le prestazioni, migliorare l'eccellenza operativa e aumentare l'affidabilità. Il modello di progettazione corretto sarà conforme alle caratteristiche del carico di lavoro corrente e ti aiuterà ad adeguare il ridimensionamento in base alla crescita o alle modifiche future.

Livello di rischio associato se questa best practice non fosse adottata: alto

Guida all'implementazione

Scopri quali sono le opzioni di configurazione relative alla rete disponibili e come possono influire sul carico di lavoro. L'ottimizzazione delle prestazioni del tuo carico di lavoro dipende dalla comprensione della modalità in cui tali opzioni interagiscono con la tua architettura e dell'impatto che hanno sulle prestazioni misurate e sulle prestazioni percepite dagli utenti.

Passaggi dell'implementazione:

1. Valuta e rivedi i modelli di progettazione che potrebbero migliorare le prestazioni del carico di lavoro.
   a. La Amazon Builders' Library fornisce una descrizione dettagliata di come Amazon crea e gestisce la tecnologia. Questi articoli sono scritti dagli ingegneri senior di Amazon e coprono una varietà di argomenti, tra cui architettura, distribuzione di software e operazioni.
   b. Portfolio di soluzioni AWS è una raccolta di soluzioni pronte per la distribuzione in cui sono assemblati servizi, codice e configurazioni. Queste soluzioni sono state create da AWS e partner AWS sulla base di casi d'uso comuni e modelli di progettazione raggruppati per settore o tipo di carico di lavoro. Ad esempio, puoi configurare una soluzione di testing del carico distribuito per il carico di lavoro.
   c. Centro di progettazione AWS fornisce diagrammi dell'architettura di riferimento raggruppati per modello di progettazione, tipo di contenuto e tecnologia.
   d. Esempi di AWS è un repository GitHub pieno di esempi pratici relativi a modelli di architettura, soluzioni e servizi comuni. Viene aggiornato di frequente in base ai più recenti servizi ed esempi.
2. Migliora il carico di lavoro per definire ulteriormente i modelli di progettazione selezionati e utilizza le opzioni di configurazione dei servizi per ottimizzare le prestazioni del carico di lavoro.
   a. Offri percorsi di formazione al team interno utilizzando le risorse disponibili in AWS Skills Guild.
   b. Utilizza la AWS Partner Network per fornire competenze specifiche in modo rapido e potenziare la tua capacità di apportare miglioramenti.

Livello di impegno per il piano di implementazione: per definire questa best practice, devi conoscere i modelli di progettazione e i servizi in grado di migliorare le prestazioni del carico di lavoro. Dopo aver valutato i modelli di progettazione, l'implementazione di tali modelli richiede un elevato livello di impegno.

Risorse

Documenti correlati:

- Centro di progettazione AWS
- AWS Partner Network
- Portfolio di soluzioni AWS
- Knowledge Center di AWS
- Amazon Builders' Library
- Utilizzo della riduzione del carico per evitare sovraccarichi
- Sfide e strategie di caching

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- This is My Architecture (La mia architettura)

Esempi correlati:

- Esempi di AWS
- Esempi di SDK AWS

PERF08-BP03 Identificazione dell'impatto dei compromessi sui clienti e sull'efficienza

Quando valuti i miglioramenti correlati alle prestazioni, determina quali scelte avranno un impatto sui clienti e sull'efficienza del carico di lavoro. Ad esempio, se l'utilizzo di un datastore chiave-valore aumenta le prestazioni del sistema, è importante valutare in che modo la consistenza della sua natura finale influirà sui clienti.

Grazie ai parametri e al monitoraggio, identifica le aree del sistema in cui le prestazioni sono scarse. Stabilisci in che modo puoi apportare miglioramenti e quali compromessi comportano, oltre al loro

impatto sul sistema e sull'esperienza degli utenti. L'implementazione di cache di dati, ad esempio, può contribuire a migliorare notevolmente le prestazioni ma richiede una strategia ben definita sulle modalità e sui tempi di aggiornamento o di invalidamento dei dati che vi sono contenuti, per evitare che il sistema si comporti in modo non corretto.

Anti-pattern comuni:

- Ritieni che tutti i vantaggi prestazionali debbano essere implementati, anche se ci sono compromessi per l'implementazione, come la consistenza finale.
- Valuti di apportare modifiche ai carichi di lavoro solo quando un problema prestazionale ha raggiunto un punto critico.

Vantaggi dell'adozione di questa best practice: Quando valuti potenziali miglioramenti relativi alle prestazioni, è necessario decidere se i compromessi per le modifiche sono coerenti con i requisiti del carico di lavoro. In alcuni casi, potrebbe essere necessario implementare controlli aggiuntivi per compensare i compromessi.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Identifica i compromessi: usa i parametri e il monitoraggio per identificare le aree di scarso rendimento del tuo sistema. Determina come apportare miglioramenti e che tipo di impatto avranno i compromessi sul sistema e sull'esperienza utente. L'implementazione di cache di dati, ad esempio, può contribuire a migliorare notevolmente le prestazioni, ma richiede una strategia ben definita sulle modalità e sui tempi di aggiornamento o di invalidamento dei dati che vi sono contenuti, per evitare che il sistema si comporti in modo non corretto.

Risorse

Documenti correlati:

- Amazon Builders' Library
- Amazon QuickSight KPIs (KPI di Amazon QuickSight)
- RUM Amazon CloudWatch
- X-Ray Documentation (Documentazione di X-Ray)

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- Creazione di un piano di monitoraggio
- Optimize applications through Amazon CloudWatch RUM (Ottimizzazione delle applicazioni tramite RUM Amazon CloudWatch)
- Demo di Amazon CloudWatch Synthetics

Esempi correlati:

- Measure page load time with Amazon CloudWatch Synthetics (Misurare il tempo di caricamento della pagina con Amazon CloudWatch Synthetics)
- Amazon CloudWatch RUM Web Client (Client Web RUM Amazon CloudWatch)

PERF08-BP04 Misurazione dell'impatto dei miglioramenti delle prestazioni

Quando vengono apportate modifiche per migliorare le prestazioni, valuta i parametri e i dati raccolti. Utilizza queste informazioni per determinare l'impatto che il miglioramento delle prestazioni ha avuto sul carico di lavoro, sui suoi componenti e sui clienti. Queste misurazioni permettono di capire quali sono i miglioramenti ottenuti dai compromessi applicati e aiutano a stabilire se si sono verificati eventuali effetti collaterali negativi.

Un sistema Well-Architected si basa su una combinazione di varie strategie riguardanti le prestazioni. Identifica quale strategia determinerà il maggiore impatto positivo su un determinato hotspot o collo di bottiglia. Lo sharding dei dati su più sistemi di database relazionali, ad esempio, può migliorare il throughput complessivo pur continuando a supportare le transazioni e, con ogni shard, il caching può contribuire a ridurre il carico.

Anti-pattern comuni:

- Distribuisci e gestisci manualmente le tecnologie disponibili come servizi gestiti.
- Ti concentri su un solo componente, ad esempio la rete, nonostante più componenti possano essere utilizzati per aumentare le prestazioni del carico di lavoro.
- L'unico punto di riferimento è dato dal feedback e dalle percezioni dei clienti.

Vantaggi dell'adozione di questa best practice: Per implementare strategie prestazionali, è necessario selezionare più servizi e caratteristiche che, insieme, ti permetteranno di soddisfare i requisiti prestazionali del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Medium

Guida all'implementazione

Un sistema Well-Architected si basa su una combinazione di varie strategie riguardanti le prestazioni. Identifica quale strategia determinerà il maggiore impatto positivo su un determinato hotspot o collo di bottiglia. Lo sharding dei dati su più sistemi di database relazionali, ad esempio, può migliorare il throughput complessivo pur continuando a supportare le transazioni e, con ogni shard, il caching può contribuire a ridurre il carico.

Risorse

Documenti correlati:

- Amazon Builders' Library
- RUM Amazon CloudWatch
- Amazon CloudWatch Synthetics
- Distributed Load Testing on AWS (Test di carico distribuito in AWS)

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- Optimize applications through Amazon CloudWatch RUM (Ottimizzazione delle applicazioni tramite RUM Amazon CloudWatch)
- Demo di Amazon CloudWatch Synthetics

Esempi correlati:

- Measure page load time with Amazon CloudWatch Synthetics (Misurare il tempo di caricamento della pagina con Amazon CloudWatch Synthetics)
- Amazon CloudWatch RUM Web Client (Client Web RUM Amazon CloudWatch)
- Distributed Load Testing on AWS (Test di carico distribuito in AWS)

PERF08-BP05 Scelta di più strategie relative alle prestazioni

Se possibile, utilizza più strategie per migliorare le prestazioni. Scegli, ad esempio, strategie come la memorizzazione dei dati nella cache per evitare eccessive chiamate di rete o dei database, l'utilizzo

di repliche di lettura per i motori di database al fine di migliorare i tassi di lettura, lo sharding o la compressione dei dati, ove possibile, per ridurre i volumi, e il buffering e lo streaming dei risultati man mano che diventano disponibili per evitare blocchi.

Man mano che apporti modifiche al carico di lavoro, raccogli e valuta i parametri per stabilire l'impatto dei cambiamenti. Misura gli impatti sul sistema e sugli utenti finali per capire in che modo i compromessi adottati influiscono sul carico di lavoro. Adotta un approccio sistematico, come il test del carico, per valutare se i compromessi migliorano le prestazioni.

Anti-pattern comuni:

- Ritieni che le prestazioni del carico di lavoro siano adeguate se i clienti non si lamentano.
- Raccogli i dati sulle prestazioni solo dopo aver apportato modifiche relative alle prestazioni.

Vantaggi dell'adozione di questa best practice: Per ottimizzare le prestazioni e l'utilizzo delle risorse, è necessario disporre di una vista operativa unificata, di dati granulari in tempo reale e di un riferimento storico. Puoi creare pannelli di controllo ed eseguire calcoli parametrici sui dati per ottenere informazioni operative e di utilizzo per i tuoi carichi di lavoro man mano che cambiano nel corso del tempo.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

Utilizza un approccio basato sui dati per far evolvere la tua architettura: quando apporti modifiche al carico di lavoro, raccogli e valuta i parametri per determinare l'impatto di tali modifiche. Misura gli impatti sul sistema e sugli utenti finali per capire in che modo i compromessi adottati influiscono sul carico di lavoro. Adotta un approccio sistematico, come il test del carico, per valutare se i compromessi migliorano le prestazioni.

Risorse

Documenti correlati:

- Amazon Builders' Library
- Best practice per l'implementazione di Amazon ElastiCache
- AWS Database Caching (Memorizzazione nella cache del database AWS)
- Usare il RUM Amazon CloudWatch
- Distributed Load Testing on AWS (Test di carico distribuito in AWS)

Video correlati:

- Introducing The Amazon Builders' Library (DOP328)
- AWS purpose-built databases (Database dedicati di AWS) (DAT209-L)
- Optimize applications through Amazon CloudWatch RUM (Ottimizzazione delle applicazioni tramite RUM Amazon CloudWatch)

Esempi correlati:

- Measure page load time with Amazon CloudWatch Synthetics (Misurare il tempo di caricamento della pagina con Amazon CloudWatch Synthetics)
- Amazon CloudWatch RUM Web Client (Client Web RUM Amazon CloudWatch)
- Distributed Load Testing on AWS (Test di carico distribuito in AWS)

Ottimizzazione dei costi

Argomenti

- Implementazione della gestione finanziaria del cloud
- Consapevolezza delle spese e dell'utilizzo
- Risorse a costi contenuti
- Gestione delle risorse di domanda e offerta
- Ottimizzazione nel tempo

Implementazione della gestione finanziaria del cloud

Domanda

- COST 1 In che modo implementi la gestione finanziaria nel cloud?

COST 1 In che modo implementi la gestione finanziaria nel cloud?

L'implementazione della gestione finanziaria del cloud consente alle organizzazioni di conseguire un valore aggiunto e il successo finanziario ottimizzando i costi e l'utilizzo e ricalibrando le risorse in AWS.

Best practice

- COST01-BP01 Creazione di una funzione di ottimizzazione dei costi
- COST01-BP02 Definizione di una partnership tra team finanziari e tecnologici
- COST01-BP03 Definizione di budget e previsioni per il cloud
- COST01-BP04 Implementazione della consapevolezza dei costi nei processi dell'organizzazione
- COST01-BP05 Invio di report e notifiche sull'ottimizzazione dei costi
- COST01-BP06 Monitoraggio proattivo dei costi
- COST01-BP07 Mantenimento dell'aggiornamento sulle nuove versioni dei servizi

COST01-BP01 Creazione di una funzione di ottimizzazione dei costi

Crea un team (Ufficio aziendale per il cloud o Centro di eccellenza del Cloud) responsabile di stabilire e mantenere la consapevolezza dei costi in tutta l'organizzazione. Il team richiede collaboratori dai ruoli finanziari, tecnologici e aziendali in tutta l'organizzazione.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

Forma un team Ufficio aziendale per il cloud (CBO) o Centro di eccellenza del Cloud (CCoE) responsabile di stabilire e mantenere una cultura basata sulla consapevolezza dei costi per il cloud computing. Può trattarsi di un individuo esistente, di un team all'interno della tua organizzazione o di un nuovo team di stakeholder chiave della finanza, della tecnologia e dell'organizzazione di tutta l'organizzazione.

La funzione (individuo o team) stabilisce le priorità e dedica la parte prevista del proprio tempo alle attività di gestione e ottimizzazione dei costi. In un'organizzazione di dimensioni ridotte, la quantità di tempo dedicata dalla funzione potrebbe essere inferiore rispetto a quella dedicata da una funzione a tempo pieno in un'azienda di dimensioni maggiori.

La funzione richiede un approccio multidisciplinare, con capacità di gestione dei progetti, data science, analisi finanziaria e sviluppo di software o infrastruttura. La funzione può migliorare l'efficienza dei carichi di lavoro mediante l'ottimizzazione dei costi nell'ambito di tre diverse responsabilità:

- Team centralizzati: mediante i team designati, ad esempio operazioni finanziarie, ottimizzazione dei costi, CBO o CCOE, i clienti possono progettare e implementare meccanismi di governance e favorire le best practice a livello aziendale.

- Team decentralizzati: i team tecnologici vengono persuasi a eseguire le ottimizzazioni.
- Team ibridi: una combinazione di team centralizzati e decentralizzati può collaborare fattivamente per eseguire l'ottimizzazione dei costi.

La funzione può essere valutata in base alla sua capacità di eseguire e conseguire risultati rispetto agli obiettivi di ottimizzazione dei costi (ad esempio, parametri di efficienza del carico di lavoro).

È necessario garantire la sponsorizzazione esecutiva affinché questa funzione possa fare la differenza e ciò rappresenta un fattore chiave per il successo. Lo sponsor è considerato un sostenitore del consumo efficiente del cloud e fornisce alla funzione un supporto di escalation, per garantire che le attività di ottimizzazione dei costi vengano trattate con il livello di priorità definito dall'organizzazione. In caso contrario, le linee guida verranno ignorate e non verrà definita la priorità delle opportunità di riduzione dei costi. Insieme, lo sponsor e la funzione assicurano che l'organizzazione utilizzi il cloud in modo efficiente e continui a offrire valore aggiunto.

Se hai un piano di supporto Business, Enterprise-On-Ramp o Enterprise e hai bisogno di assistenza per creare il team o la funzione, contatta gli esperti di gestione finanziaria del cloud mediante il team del tuo account.

Passaggi dell'implementazione

- Definizione dei membri chiave: Devi assicurarti che tutte le parti rilevanti della tua organizzazione contribuiscano e partecipino alla gestione dei costi. I team più comuni all'interno delle organizzazioni includono in genere: team finanziari, proprietari di applicazioni o prodotti, team di gestione e team tecnici (DevOps). Alcuni sono coinvolti a tempo pieno (finanziari, tecnici), altri periodicamente in base alle esigenze. I singoli o i team che svolgono mansioni di gestione finanziaria del cloud in genere devono disporre del seguente set di competenze:
  - Competenze inerenti allo sviluppo di software, in caso di sviluppo di script e funzioni di automazione.
  - Competenze inerenti alla progettazione di infrastrutture, per implementare script o funzioni automazione e comprendere in che modo viene effettuato il provisioning di risorse e servizi.
  - Sesto senso per le operazioni: "gestione finanziaria del cloud" significa una presenza efficiente nel cloud mediante la misurazione, il monitoraggio, la modifica, la pianificazione e il dimensionamento dell'utilizzo efficiente del cloud stesso.
- Definizione di obiettivi e parametri: La funzione deve fornire valore all'organizzazione in modi diversi. Questi obiettivi sono definiti e si evolvono continuamente con l'evolversi dell'organizzazione. Tra le attività più comuni figurano la creazione e l'esecuzione di programmi

di formazione sull'ottimizzazione dei costi in tutta l'organizzazione, lo sviluppo di standard a livello aziendale, come monitoraggio ed elaborazione di report per l'ottimizzazione dei costi, e la definizione degli obiettivi di ottimizzazione dei carichi di lavoro. Inoltre, è necessario comunicare regolarmente all'organizzazione la relativa capacità di ottimizzazione dei costi.

Puoi definire gli indicatori chiave delle prestazioni (KPI) basati sui valori. I KPI possono fare riferimento ai costi o ai valori. Quando vengono definiti i KPI, è possibile calcolare il costo previsto in termini di efficienza e risultati aziendali previsti. I KPI basati sui valori associano le metriche relative a costi e utilizzo ai driver relativi al valore aziendale e ciò aiuta a razionalizzare le variazioni a livello di spesa AWS. Il primo passo per derivare i KPI basati sui valori è collaborare tra organizzazioni al fine di scegliere e concordare un set standard di KPI.

- Definizione di una regolare cadenza: il gruppo (team finanziario, tecnologico e aziendale) deve riunirsi regolarmente per rivedere le metriche e gli obiettivi. Una cadenza tipica implica la revisione dello stato dell'organizzazione, la revisione dei programmi attualmente in esecuzione e la revisione dei parametri finanziari e di ottimizzazione generali. Quindi, per i carichi di lavoro chiave, è opportuno elaborare report più dettagliati.

  Durante queste riunioni periodiche è possibile analizzare l'efficienza (costo) dei carichi di lavoro e il risultato aziendale. Ad esempio, un incremento del 20% dei costi di un carico di lavoro potrebbe essere determinato dall'aumento dell'utilizzo da parte dei clienti. In questo caso, l'incremento del 20% dei costi può essere interpretato come investimento. Questi incontri regolari possono aiutare i team a individuare i KPI basati sui valori in grado di garantire un "valore aggiunto" all'intera organizzazione.

Risorse

Documenti correlati:

- AWS CCOE Blog (Blog CCOE AWS)
- Creating Cloud Business Office (Creazione di un ufficio aziendale per il cloud)
- CCOE - Cloud Center of Excellence (CCoE - Centro di eccellenza del Cloud)

Video correlati:

- Vanguard CCOE Success Story (Storia di successo CCOE Vanguard)

Esempi correlati:

- Using a Cloud Center of Excellence (CCOE) to Transform the Entire Enterprise (Utilizzo di un Centro di eccellenza del Cloud [CCoE] per trasformare l'intera azienda)
- Building a CCOE to transform the entire enterprise (Creazione di un Centro di eccellenza del Cloud [CCoE] per trasformare l'intera azienda)
- 7 Pitfalls to Avoid When Building CCOE (7 errori da evitare durante la creazione di un Centro di eccellenza del Cloud [CCoE])

COST01-BP02 Definizione di una partnership tra team finanziari e tecnologici

Coinvolgi i team finanziari e tecnologici nelle discussioni su costi e utilizzo in tutte le fasi del tuo approccio al cloud. I team si riuniscono regolarmente e discutono argomenti quali obiettivi e target organizzativi, stato attuale di costi e utilizzo e pratiche finanziarie e contabili.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

i team tecnologici possono innovare più rapidamente nel cloud grazie a cicli di approvazione, approvvigionamento e distribuzione dell'infrastruttura più brevi. Può trattarsi di una novità per le organizzazioni finanziarie che in precedenza erano abituate a eseguire processi dispendiosi, in termini di tempo e di risorse, per acquistare e distribuire capitale in data center e ambienti locali, allocando i costi solo in fase di approvazione del progetto.

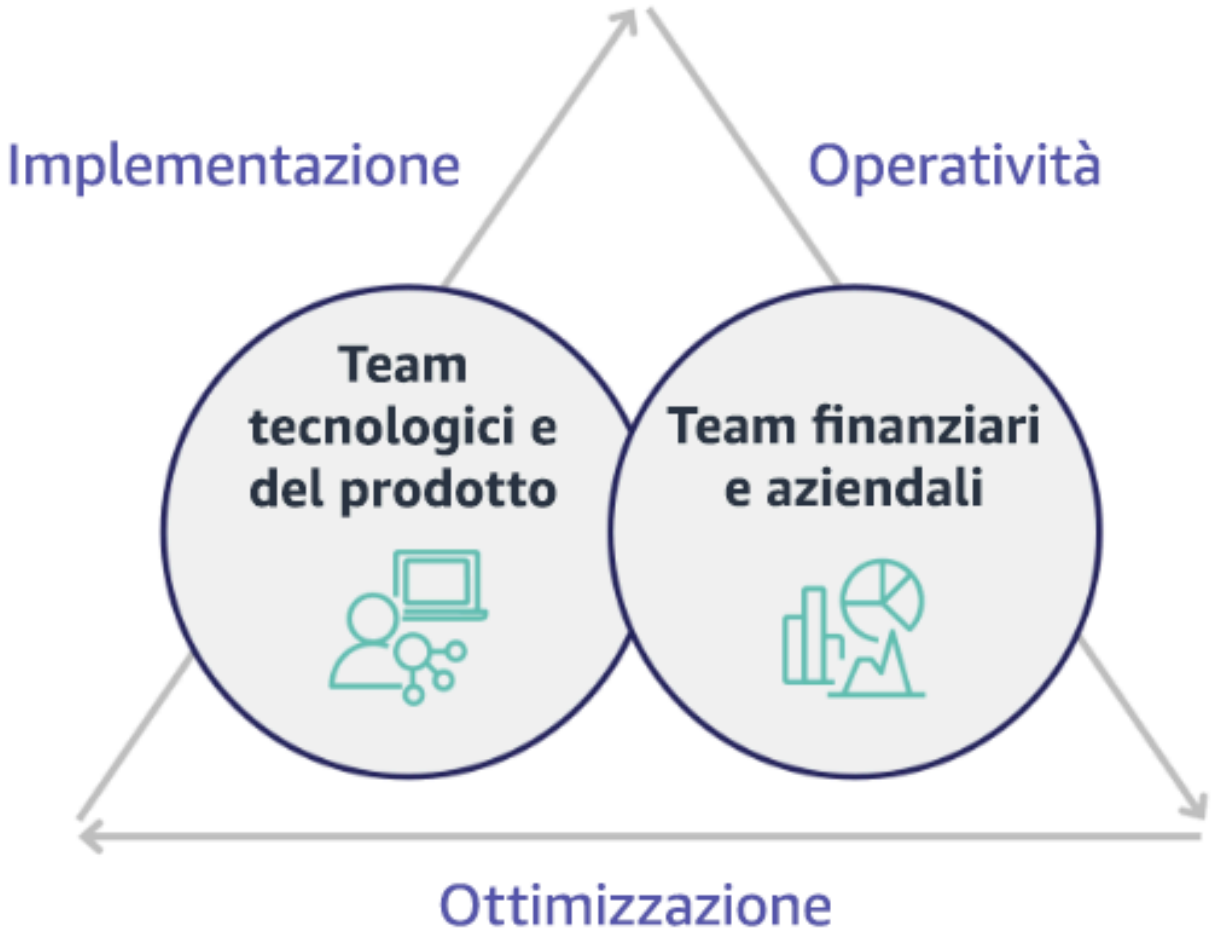
Dal punto di vista delle organizzazioni finanziarie e addette all'approvvigionamento, il processo di elaborazione del piano degli investimenti, della richiesta, dell'approvazione e dell'approvvigionamento degli investimenti e dell'installazione dell'infrastruttura fisica è stato interiorizzato e standardizzato da decenni:

- I team di progettazione o IT sono in genere i richiedenti
- I vari team finanziari fungono da approvatori e procuratori
- I team operativi assemblano, implementano e distribuiscono un'infrastruttura pronta all'uso

Con l'adozione del cloud, l'approvvigionamento e il consumo dell'infrastruttura non sono più vincolati da una catena di dipendenze. Nel modello cloud, i team tecnologici e del prodotto non sono più semplici "costruttori", ma anche operatori e proprietari dei loro prodotti, responsabili della maggior parte delle attività storicamente associate ai team finanziari e operativi, compresi l'approvvigionamento e l'implementazione.

Quanto in realtà è necessario per il provisioning delle risorse cloud è un account utente e il set appropriato di autorizzazioni, elementi questi che riducono i rischi IT e finanziari. Ciò significa che ai team basta un numero ridotto di clic o chiamate API per terminare le risorse cloud non necessarie o inattive. Ciò inoltre consente ai team tecnologici di velocizzare l'innovazione, grazie all'agilità e alla capacità di potenziare e quindi ridimensionare i vari progetti sperimentali. Se da un lato la natura

variabile del consumo del cloud può influenzare la prevedibilità dal punto di vista del processo di elaborazione del piano degli investimenti e delle previsioni, il cloud fornisce alle organizzazioni la capacità di ridurre il costo del provisioning eccessivo e contemporaneamente il costo delle opportunità associato a un provisioning insufficiente di carattere conservativo.



Stabilisci una collaborazione tra i principali stakeholder finanziari e tecnologici per creare una comprensione condivisa degli obiettivi organizzativi e sviluppare meccanismi che consentano il successo finanziario nel modello di spesa variabile del cloud computing. I team pertinenti all'interno della tua organizzazione devono essere coinvolti nelle discussioni su costi e utilizzo in tutte le fasi del tuo viaggio verso il cloud; tra di essi vi sono:

- Responsabili finanziari: CFO, controllori finanziari, pianificatori finanziari, analisti aziendali, approvvigionamento e selezione delle risorse e contabilità fornitori devono comprendere il modello di consumo del cloud, le opzioni di acquisto e il processo di fatturazione mensile. I team finanziari devono collaborare con i team tecnologici per creare e divulgare a livello aziendale una narrazione

del valore IT che aiuti i team aziendali a comprendere lo stretto legame tra spesa in tecnologie e risultati aziendali. In questo modo, la spesa tecnologica viene considerata non tanto come un costo, quanto piuttosto come un vero e proprio investimento. A causa delle differenze fondamentali tra il cloud (ad esempio il tasso di variazione dell'utilizzo, i prezzi a consumo o a scaglioni, i modelli di prezzo e le informazioni dettagliate su fatturazione e utilizzo) e le operazioni in locale, è essenziale che l'organizzazione finanziaria capisca in che modo l'utilizzo del cloud può influire sugli aspetti aziendali, tra cui processi di approvvigionamento, monitoraggio degli incentivi, allocazione dei costi e bilanci.

- Responsabili tecnologici: i responsabili tecnologici (inclusi i proprietari di prodotti e applicazioni) devono essere a conoscenza dei requisiti finanziari (ad esempio i vincoli di budget) e dei requisiti aziendali (ad esempio i contratti sul livello di servizio). In questo modo, il carico di lavoro può essere implementato in modo opportuno per raggiungere gli obiettivi desiderati dall'azienda.

La collaborazione tra finanza e tecnologia offre i seguenti vantaggi:

- I team finanziari e tecnologici hanno una visibilità quasi in tempo reale su costi e utilizzo.
- I team finanziari e tecnologici stabiliscono una procedura operativa standard per gestire le variazioni di spesa nel cloud.
- Gli stakeholder finanziari fungono da consulenti strategici per quanto riguarda il modo in cui il capitale viene utilizzato per acquistare sconti a fronte di impegni (ad esempio, istanze riservate o AWS Savings Plans) e il modo in cui il cloud viene utilizzato per far crescere l'organizzazione.
- I processi di approvvigionamento e di contabilità esistenti vengono applicati al cloud.
- I team finanziari e tecnologici collaborano per prevedere costi e utilizzo di AWS futuri, al fine di allineare e sviluppare i budget aziendali.
- La comunicazione all'interno dell'organizzazione migliora attraverso un linguaggio condiviso e una comprensione comune dei concetti finanziari.

Altri stakeholder all'interno della tua organizzazione che devono essere coinvolti nelle discussioni su costi e utilizzo includono:

- Proprietari delle unità aziendali: i proprietari delle unità aziendali devono comprendere il modello aziendale del cloud in modo da indirizzare l'operato delle unità aziendali e di tutta l'azienda. Questa conoscenza del cloud è fondamentale quando è necessario prevedere la crescita e l'utilizzo del carico di lavoro, ma anche quando si valutano le diverse opzioni di acquisto, come le istanze riservate o i Savings Plans.

- Team di progettazione: lo sviluppo di una partnership tra team finanziari e tecnologici è essenziale per la creazione di una cultura consapevole dei costi che incoraggi il coinvolgimento degli ingegneri nella gestione finanziaria del cloud. Uno dei problemi comuni dei professionisti della gestione finanziaria del cloud o delle operazioni e dei team finanziari è far capire agli ingegneri l'attività nel cloud nel suo complesso e implementare le azioni consigliate.
- Terze parti: se la tua organizzazione si avvale di terze parti (ad esempio, consulenti o strumenti), assicurati che esse siano allineate ai tuoi obiettivi finanziari e possano dimostrare sia l'allineamento, tramite i loro modelli di coinvolgimento, sia il ritorno sull'investimento (ROI). In genere, le terze parti contribuiscono alla creazione di report e all'analisi di eventuali carichi di lavoro da esse gestiti, e forniscono anche l'analisi dei costi relativi ai carichi di lavoro da esse progettati.

L'implementazione della gestione finanziaria del cloud e il conseguimento dei risultati richiedono la stretta collaborazione tra team finanziari, tecnologici e aziendali, nonché un cambiamento nel modo in cui la spesa cloud viene comunicata e valutata all'interno dell'organizzazione. Includi i team di progettazione in modo da renderli partecipi delle discussioni su costi e utilizzi in tutte le fasi e incoraggiali ad attenersi alle best practice e ad adottare le azioni concordate.

Passaggi dell'implementazione

- Definizione dei membri chiave: Verifica che tutti i membri rilevanti dei team finanziari e tecnologici partecipino alla partnership. I membri del team finanziario interessati saranno quelli che hanno a che fare con la fatturazione dei servizi cloud. In genere si tratta di CFO, controllori finanziari, pianificatori finanziari, analisti aziendali, addetti agli acquisti e al sourcing. I membri tecnologici sono in genere i proprietari di prodotti e applicazioni, manager tecnici e rappresentanti di tutti i team che si basano sul cloud. Altri membri possono includere i responsabili di unità aziendali, ad esempio il marketing che influenzerà l'utilizzo dei prodotti, e terze parti, come i consulenti, necessari per garantire l'allineamento agli obiettivi e meccanismi e per fornire assistenza nell'elaborazione dei report.
- Definizione degli argomenti oggetto della discussione: Definisci gli argomenti comuni tra i team o che necessitano di una comprensione condivisa. Segui il costo dal momento in cui viene creato, fino al pagamento della fattura. Prendi nota di tutti i membri coinvolti e dei processi organizzativi che devono essere applicati. Comprendi ogni fase o processo e le informazioni associate, come i modelli di prezzo disponibili, i prezzi a scaglioni, i modelli di sconto, il budget e i requisiti finanziari.
- Definizione di una regolare cadenza: per creare una partnership tra team finanziari e tecnologici, definisci la periodicità delle comunicazioni per creare e gestire l'allineamento. Il gruppo deve riunirsi regolarmente in base ai propri obiettivi e parametri. Una cadenza tipica implica la revisione dello

stato dell'organizzazione, la revisione dei programmi attualmente in esecuzione e la revisione dei parametri finanziari e di ottimizzazione generali. Quindi, per i carichi di lavoro chiave, è opportuno elaborare report più dettagliati.

Risorse

Documenti correlati:

- Blog delle novità di AWS

COST01-BP03 Definizione di budget e previsioni per il cloud

Adatta i processi di previsione e di budget organizzativi esistenti in modo che siano compatibili con la natura altamente variabile dei costi e dell'utilizzo del cloud. I processi devono essere dinamici, utilizzando algoritmi basati su tendenze o fattori chiave di business o una combinazione di entrambi.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

i clienti utilizzano il cloud per ottenere efficienza, velocità e agilità, determinando un'elevata variabilità in termini di costi e utilizzo. I costi possono diminuire in seguito all'aumento dell'efficienza del carico di lavoro o con la distribuzione di nuovi carichi di lavoro e funzionalità. È possibile rilevare l'incremento del costo all'aumentare dell'efficienza dei carichi di lavoro oppure quando vengono implementati nuovi carichi di lavoro e nuove caratteristiche. In alternativa, i carichi di lavoro possono essere dimensionati per servire un maggior numero di clienti, aumentando l'utilizzo e i costi del cloud. Le risorse sono ancora più accessibili di prima. L'elasticità del cloud va di pari passo con l'elasticità dei costi e delle previsioni. Gli attuali processi di creazione di budget dell'organizzazione devono essere modificati per incorporare questa variabilità.

Puoi adattare i processi di creazione dei budget e previsione esistenti per renderli più dinamici utilizzando un algoritmo basato su trend (utilizzando i costi storici come input), algoritmi basati su fattori chiavi di business (ad esempio, lanci di nuovi prodotti o espansione regionale) o una combinazione di trend e fattori chiave di business.

Puoi utilizzare Budget AWS per impostare budget personalizzati a un livello granulare specificando il periodo di tempo, la periodicità o l'importo (fisso o variabile), e aggiungendo filtri come servizi, regioni AWS e tag. Per essere sempre aggiornati in merito alle prestazioni dei budget esistenti, puoi creare e programmare report Budget AWS da inviare tramite e-mail a te e alle parti coinvolte

a cadenza regolare. Puoi anche creare avvisi Budget AWS basati sui costi effettivi, ovvero avvisi intrinsecamente reattivi, oppure sui costi previsti, ovvero avvisi che consentono di implementare tempestivamente azioni correttive a fronte di potenziali eventi di superamento dei costi. Riceverai un avviso al superamento o al previsto superamento dei costi o dell'utilizzo definiti nel budget.

AWS offre tutta la flessibilità necessaria per costruire processi di previsione e pianificazione delle spese in modo da garantire il costante aggiornamento sulla conformità o sul superamento dei costi rispetto ai limiti definiti per il budget.

Puoi utilizzare AWS Cost Explorer per elaborare previsioni relative ai costi per un intervallo di tempo futuro definito in base alle spese pregresse. Il motore di previsione di AWS Cost Explorer segmenta i dati storici in base ai tipi di addebito, ad esempio le istanze riservate, e utilizza una combinazione di machine learning e modelli basati su regole per elaborare previsioni di spesa per tutti i singoli tipi di addebito. Puoi infine utilizzare AWS Cost Explorer per prevedere i costi cloud giornalieri (fino a tre mesi) o mensili (fino a 12 mesi) in base agli algoritmi di machine learning applicati ai costi storici (basati su trend).

Dopo aver determinato la previsione basata su trend mediante Cost Explorer, puoi utilizzare AWS Pricing Calculator per calcolare la stima dei costi futuri e dei casi d'uso AWS in base all'utilizzo previsto (traffico, richieste al secondo, istanza Amazon Elastic Compute Cloud (Amazon EC2) richiesta e così via), nonché per supportare il processo di pianificazione delle modalità di spesa, ricerca delle opportunità di riduzione dei costi ed elaborazione di decisioni informate in caso di utilizzo di AWS.

Puoi utilizzare AWS Cost Anomaly Detection per evitare o ridurre gli imprevisti a livello di costi e ottimizzare il controllo senza rallentare i processi di innovazione. AWS Cost Anomaly Detection sfrutta tecnologie avanzate di machine learning per individuare spese anomale e cause principali in modo da rendere possibile la tempestiva adozione di misure correttive. Grazie a tre semplici passaggi, puoi creare una funzione di controllo contestualizzato personalizzato e ricevere avvisi quando viene rilevata una spesa anomala. Gli sviluppatori potranno dedicarsi a tempo pieno alle loro attività e lasciare ad AWS Cost Anomaly Detection il compito di monitorare la spesa e ridurre il rischio di imprevisti a livello di fatturazione.

Come indicato nella sezione Collaborazione tra finanza e tecnologia del Pilastro AWS Well-Architected relativo dell'ottimizzazione dei costi, è importante che esistano una partnership e opportunità di contatto tra IT, finanza e le altre parti coinvolte per garantire l'utilizzo da parte di tutti degli stessi strumenti e processi a garanzia del modello di consistenza. Nei casi in cui si rendano necessarie modifiche del budget, l'incremento della frequenza delle occasioni di contatto permette infatti di intervenire e reagire più tempestivamente.

Passaggi dell'implementazione

- Aggiornamento del budget esistente e dei processi di previsione: implementa l'utilizzo di tendenze, fattori chiave di business o una combinazione di entrambi nei processi di budget e previsione.
- Configurazione di avvisi e notifiche: utilizza gli avvisi Budget AWS e Cost Anomaly Detection.
- Esecuzione di revisioni a intervalli regolari con le parti coinvolte chiave, ad esempio le parti coinvolte nelle aree IT, finanza, piattaforma e altre aree strategiche dell'azienda, al fine di garantire l'allineamento alle modifiche a livello di direttive e utilizzi aziendali.

Risorse

Documenti correlati:

- AWS Cost Explorer
- Budget AWS
- AWS Pricing Calculator
- AWS Cost Anomaly Detection
- AWS License Manager

Esempi correlati:

- Launch: Usage-Based Forecasting now Available in AWS Cost Explorer (Lancio: previsioni basate sull'utilizzo ora disponibili in AWS Cost Explorer)
- AWS Well-Architected Labs - Cost and Usage Governance (Costi e governance)

COST01-BP04 Implementazione della consapevolezza dei costi nei processi dell'organizzazione

Implementa la consapevolezza dei costi e crea trasparenza e funzionalità di controllo in processi nuovi o esistenti che influiscono sull'utilizzo e sfrutta i processi esistenti per favorire la consapevolezza dei costi. Implementa la consapevolezza dei costi nella formazione dei dipendenti.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

La consapevolezza dei costi deve essere implementata nei processi organizzativi nuovi ed esistenti. Si tratta di un prerequisito fondamentale per altre best practice. È consigliabile riutilizzare e

modificare i processi esistenti, laddove possibile, riducendo al minimo l'impatto sull'agilità e sulla velocità. Comunica i costi del cloud ai team tecnologici e ai responsabili dei processi decisionali nei team aziendali e finanziari per accrescere la consapevolezza dei costi e definisci indicatori chiave delle prestazioni (KPI) per l'efficienza da segnalare alle parti coinvolte nelle varie aree finanziarie e aziendali. Le seguenti raccomandazioni aiuteranno a implementare la consapevolezza dei costi nel carico di lavoro:

- Verifica che la gestione delle modifiche includa una misurazione dei costi per quantificare l'impatto finanziario delle modifiche. Questo aiuta a risolvere in modo proattivo le problematiche relative ai costi nonché a evidenziare i risparmi ottenuti.
- Verifica che l'ottimizzazione dei costi sia un componente fondamentale delle tue capacità operative. Ad esempio, puoi sfruttare gli attuali processi di gestione degli incidenti per analizzare e identificare la causa principale di anomalie di costi e utilizzo o delle eccedenze di costo.
- Accelera la riduzione dei costi e la realizzazione del valore aggiunto attraverso l'automazione o l'utilizzo di strumenti. Quando valuti i costi dell'implementazione, includi nella valutazione un componente ROI per giustificare l'investimento di tempo o denaro.
- Assegna i costi del cloud mediante l'implementazione delle policy di showback/chargeback per la spesa cloud, compresa la spesa per opzioni di acquisto basate su impegno, servizi condivisi e acquisti su marketplace, a supporto di un consumo del cloud maggiormente consapevole dei costi.
- Estendi i programmi di formazione e sviluppo esistenti per includere la formazione sulla consapevolezza dei costi in tutta l'organizzazione, comprese attività di formazione continua e certificazione. In questo modo, creerai un'organizzazione in grado di gestire in modo autonomo i costi e l'utilizzo.
- Sfrutta i vantaggi degli strumenti nativi AWS gratuiti, come AWS Cost Anomaly Detection, Budget AWSe Report Budget AWS.

Quando le organizzazioni adottano in modo sistematico le best practice relative alla gestione finanziaria del cloud, questi comportamenti vengono inglobati nelle procedure di lavoro e nei processi decisionali. Ne risulterà una cultura basata su una maggiore consapevolezza dei costi, condivisa dagli sviluppatori che creano nuove applicazioni per il cloud e dai responsabili dell'area finanziaria che analizzano il ROI per questi nuovi investimenti a livello di cloud.

Passaggi dell'implementazione

- Identificazione dei processi organizzativi pertinenti: Ogni unità organizzativa esamina i propri processi e identifica i processi che influiscono su costi e utilizzo. Tutti i processi che determinano

la creazione o la cessazione di una risorsa devono essere inclusi nella revisione. Inoltre, individua i processi che possono supportare la consapevolezza dei costi nella tua azienda, ad esempio la gestione degli incidenti e la formazione.

- Definizione di una cultura consapevole dei costi autosufficiente: assicurati che tutte le parti coinvolte rilevanti siano concordi sulla causa della modifica e sull'impatto come costo in modo che abbiano la piena consapevolezza del costo del cloud. Ciò consentirà all'organizzazione di definire una cultura consapevole dei costi autosufficiente finalizzata all'innovazione.
- Aggiornamento dei processi con la consapevolezza dei costi: Ogni processo viene modificato in modo che ci sia una consapevolezza dei costi. Il processo potrebbe richiedere ulteriori controlli preliminari, ad esempio la valutazione dell'impatto dei costi, oppure controlli successivi che attestino il verificarsi dei cambiamenti previsti in termini di costi e utilizzo. I processi di supporto come la formazione e la gestione degli incidenti possono essere estesi per includere elementi relativi a costi e utilizzo.

Per ottenere assistenza, contatta gli esperti di gestione finanziaria del cloud mediante il team del tuo account oppure esplora le risorse e i documenti correlati elencati di seguito.

Risorse

Documenti correlati:

- Gestione finanziaria del cloud con AWS

Esempi correlati:

- Strategy for Efficient Cloud Cost Management (Strategia per un'efficiente gestione dei costi del cloud)
- Cost Control Blog Series #3: How to Handle Cost Shock (Blog relativo al controllo dei costi - Serie 3: Come gestire l'impatto dei costi)
- A Beginner’s Guide to AWS Cost Management (Guida per principianti alla AWS Cost Management)

COST01-BP05 Invio di report e notifiche sull'ottimizzazione dei costi

Configura Budget AWS e AWS Cost Anomaly Detection per fornire notifiche su costi e utilizzo rispetto agli obiettivi. Organizza riunioni regolari per analizzare l'efficienza dei costi del carico di lavoro e promuovere una cultura che ponga attenzione ai costi.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

è necessario rendicontare regolarmente l'ottimizzazione dei costi e dell'utilizzo all'interno dell'organizzazione. Puoi implementare sessioni dedicate per l'ottimizzazione dei costi o includere l'ottimizzazione dei costi nei regolari cicli di reporting operativi per i tuoi carichi di lavoro. Utilizza servizi e strumenti per individuare e implementare le opportunità di riduzione dei costi. AWS Cost Explorer fornisce pannelli di controllo e report. Puoi monitorare l'avanzamento dei costi e dell'utilizzo rispetto ai budget configurati con i Report Budget AWS.

Utilizza Budget AWS per configurare budget personalizzati al fine di tenere traccia dei costi e dell'utilizzo e reagire con tempestività agli avvisi ricevuti via e-mail o alle notifiche Amazon Simple Notification Service (Amazon SNS) in caso di superamento della soglia definita. Imposta il periodo di budget preferito su giornaliero, mensile, trimestrale o annuale e crea limiti di budget specifici per essere costantemente informato sui valori di utilizzo e sui costi effettivi o previsti rispetto alla soglia definita per il budget. Puoi anche configurare avvisi e operazioni da eseguire automaticamente o in base a un processo di approvazione a fronte di tali avvisi quando viene superato l'obiettivo del budget.

Implementa notifiche su costi e utilizzo in modo che si possa intervenire rapidamente in caso di variazioni impreviste di costi e utilizzo. AWS Cost Anomaly Detection consente di ridurre gli inconvenienti a livello di costi e migliorare il controllo senza rallentare il processo di innovazione. AWS Cost Anomaly Detection individua le spese anomale e le cause principali a favore della riduzione del rischio di imprevisti a livello di fatturazione. Grazie a tre semplici passaggi, puoi creare una funzione di controllo contestualizzato personalizzato e ricevere avvisi quando viene rilevata una spesa anomala.

Puoi anche utilizzare Amazon QuickSight con dati AWS Cost and Usage Report (CUR) per fornire funzionalità di reporting personalizzate con dati più granulari. Amazon QuickSight consente di programmare i report e ricevere via e-mail report periodici sui costi relativi all'utilizzo e sui costi storici o sulle opportunità di riduzione dei costi.

Utilizza AWS Trusted Advisor, che mette a disposizione linee guida per verificare se le risorse allocate sono conformi alle best practice AWS in relazione all'ottimizzazione dei costi.

Crea periodicamente report contenenti informazioni di primo piano relative a Savings Plans, istanze riservate e suggerimenti di ridimensionamento Amazon Elastic Compute Cloud (Amazon EC2) da AWS Cost Explorer per favorire la riduzione dei costi associati a carichi di lavoro con stato stazionario

e a risorse inattive e sottoutilizzate. Individua e ammortizza la spesa associata all'utilizzo non ottimale del cloud relativamente alle risorse implementate. Con utilizzo non ottimale del cloud si intende quando vengono create risorse con dimensioni errate oppure quando vengono rilevati modelli di utilizzo del cloud diversi da quanto previsto. Attieniti alle best practice AWS per ridurre i casi di utilizzo non ottimale e ottimizzare e ridurre i costi del cloud.

Genera regolarmente report per migliorare le opzioni di acquisto delle risorse al fine di ridurre il costo unitario dei carichi di lavoro. Le opzioni di acquisto quali, ad esempio, Savings Plans, istanze riservate o istanze spot Amazon EC2, offrono il massimo risparmio sui costi per carichi di lavoro con tolleranza ai guasti, consentendo alle parti coinvolte (proprietari di aziende, team finanziari e tecnologici) di venire coinvolti nelle discussioni di merito.

Condividi i report contenenti opportunità o annunci di nuovi rilasci a supporto della riduzione del costo totale di proprietà (TCO) del cloud. Adotta nuovi servizi, regioni, funzionalità, soluzioni o nuovi modi per migliorare ulteriormente la riduzione dei costi.

Passaggi dell'implementazione

- Configura Budget AWS: Configura Budget AWS su tutti gli account per il tuo carico di lavoro. Imposta un budget per la spesa complessiva dell'account e un budget per il carico di lavoro utilizzando i tag.
  - Well-Architected Labs: utilizzo di costi e governance
- Report sull'ottimizzazione dei costi: Configura un ciclo regolare per discutere e analizzare l'efficienza del carico di lavoro. Utilizzando i parametri stabiliti, segnala i parametri ottenuti e il costo sostenuto per ottenerli. Identifica e correggi eventuali tendenze negative e identifica le tendenze positive che puoi favorire in tutta l'organizzazione. L'elaborazione dei report deve coinvolgere i rappresentanti dei team applicativi e dei proprietari, dei team finanziari e di gestione.
  - AWS Well-Architected Labs: visualizzazione

Risorse

Documenti correlati:

- AWS Cost Explorer
- AWS Trusted Advisor
- Budget AWS
- Budget AWS Best Practices (Best practice per Budget AWS)

- Amazon CloudWatch
- AWS CloudTrail
- Amazon S3 Analytics (Analisi Amazon S3)
- AWS Cost and Usage Report

Esempi correlati:

- Well-Architected Labs: utilizzo di costi e governance
- AWS Well-Architected Labs: visualizzazione
- Key ways to start optimizing your AWS cloud costs (Principali soluzioni per iniziare a ottimizzare i costi del cloud AWS)

COST01-BP06 Monitoraggio proattivo dei costi

Implementa strumenti e pannelli di controllo per monitorare i costi in modo proattivo per il carico di lavoro. Rivedi regolarmente i costi utilizzando strumenti configurati o pronti all'uso e non limitarti a guardare solo i costi e le categorie quando ricevi le notifiche. Il monitoraggio e l'analisi proattivi dei costi aiutano a individuare i trend positivi e ti consente di promuoverli all'interno dell'organizzazione.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

si consiglia di monitorare i costi e l'utilizzo all'interno dell'organizzazione in modo proattivo, e non solo in caso di eccezioni o anomalie. I pannelli di controllo con un'elevata visibilità in tutto l'ufficio o l'ambiente di lavoro garantiscono che le persone chiave abbiano accesso alle informazioni di cui hanno bisogno e dimostrano l'attenzione che l'organizzazione presta all'ottimizzazione dei costi. I pannelli di controllo visibili consentono di promuovere attivamente i risultati positivi e di implementarli in tutta l'organizzazione.

Crea procedure giornaliere o frequenti che utilizzino AWS Cost Explorer o qualsiasi altro pannello di controllo, come Amazon QuickSight , per verificare i costi e analizzarli in modo proattivo. Analizza l'utilizzo e i costi dei servizi AWS a livello di account AWS, carico di lavoro o servizio AWS specifico in gruppo o mediante filtri e verifica che siano in linea con quanto previsto. Utilizza tag e granularità a livello orario o di risorsa per filtrare e individuare i costi ricorrenti relativi alle risorse di maggiore utilizzo. Puoi anche creare report personalizzati con il pannello di controllo Intelligence costi, una

soluzione Amazon QuickSight sviluppata dagli AWS Solutions Architect, e confrontare i budget con i costi e l'utilizzo effettivi.

Passaggi dell'implementazione

- Report sull'ottimizzazione dei costi: Configura un ciclo regolare per discutere e analizzare l'efficienza del carico di lavoro. Utilizzando i parametri stabiliti, segnala i parametri ottenuti e il costo sostenuto per ottenerli. Identifica e correggi eventuali tendenze negative e identifica le tendenze positive che puoi favorire in tutta l'organizzazione. L'elaborazione dei report deve coinvolgere i rappresentanti dei team applicativi e dei proprietari, dei team finanziari e di gestione.
- Creazione e abilitazione di Budget AWS con granularità giornaliera relativi a costi e utilizzo per adottare misure tempestive volte a impedire potenziali superamenti dei costi: Budget AWS consente di configurare notifiche di avviso per essere sempre informati se qualsiasi tipo di budget non è conforme alle soglie preconfigurate. Il modo migliore per utilizzare Budget AWS è configurare i costi e l'utilizzo previsti come limite in modo tale che qualsiasi superamento del budget possa essere considerato un superamento del limite di spesa.
- Creazione del AWS Cost Anomaly Detection per il monitoraggio dei costi: AWS Cost Anomaly Detection utilizza la tecnologia avanzata di machine learning per individuare le spese anomale e le cause principali in modo da garantire un intervento tempestivo. Ti consente di configurare funzionalità di monitoraggio dei costi che definiscono i segmenti di spesa da valutare, ad esempio singoli servizi AWS, account membro, tag di allocazione dei costi e categorie di costo, nonché di impostare quando, dove e come riceverai le notifiche di avviso. Per ciascuna funzionalità di monitoraggio, puoi associare più sottoscrizioni agli avvisi per proprietari di azienda e team tecnologici, inclusi un nome, una soglia relativa all'impatto dei costi e la frequenza di avviso (avvisi singoli, riepilogo giornaliero, riepilogo settimanale) per ciascuna sottoscrizione.
- Utilizzo di AWS Cost Explorer o integrazione dei dati AWS Cost and Usage Report (CUR) con i pannelli di controllo Amazon QuickSight per la visualizzazione dei costi dell'organizzazione: La funzionalità AWS Cost Explorer è caratterizzata da un'interfaccia di semplice utilizzo che consente di visualizzare, analizzare e gestire l'utilizzo e i costi AWS nel tempo. Il pannello di controllo Intelligence costi è personalizzabile e accessibile e consente di creare le basi di uno strumento di gestione e ottimizzazione dei costi personalizzato.

Risorse

Documenti correlati:

- Budget AWS

- AWS Cost Explorer
- Daily Cost and Usage Budgets (Budget per costi e utilizzo giornalieri)
- AWS Cost Anomaly Detection

Esempi correlati:

- AWS Well-Architected Labs: visualizzazione
- AWS Well-Architected Labs: visualizzazione avanzata
- Well-Architected Labs: Cloud Intelligence Dashboards (Pannelli di controllo Intelligence cloud)
- Well-Architected Labs: Cost Visualization (Visualizzazione dei costi)
- AWS Cost Anomaly Detection Alert with Slack (Avvisi AWS Cost Anomaly Detection con Slack)

COST01-BP07 Mantenimento dell'aggiornamento sulle nuove versioni dei servizi

Consultati regolarmente con gli esperti o con i partner AWS per valutare quali servizi e caratteristiche offrono un costo inferiore. Consulta i blog AWS e altre fonti di informazione.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

AWS continua ad aggiungere nuove caratteristiche in modo da consentirti di utilizzare le tecnologie più aggiornate a supporto di un più rapido processo di sperimentazione e innovazione. Potresti essere in grado di implementare nuovi servizi e funzionalità AWS per aumentare l'efficienza in termini di costi del carico di lavoro. Consulta regolarmente la pagina Gestione dei costi AWS, il Blog delle novità di AWS, il Blog sulla gestione dei costi AWSe Novità di AWS per informazioni su nuovi servizi e versioni di funzionalità. I post nella sezione Novità forniscono una breve panoramica di tutti gli annunci relativi a servizi AWS, funzionalità ed espansione delle regioni al momento del loro rilascio.

Passaggi dell'implementazione

- Iscriviti ai blog: Vai alle pagine dei blog AWS e iscriviti al Blog delle novità e ad altri blog di interesse. Puoi effettuare la registrazione nella pagina delle preferenze di comunicazione utilizzando il tuo indirizzo e-mail.
- Iscriviti alle novità di AWS: consulta regolarmente il Blog delle novità di AWS e Novità di AWS per informazioni su nuovi servizi e versioni di funzionalità. Iscriviti ai feed RSS oppure utilizza il tuo indirizzo e-mail per essere sempre aggiornato su annunci e nuovi rilasci.

- Segui le informazioni riportate nella sezione relativa alle riduzioni di prezzo AWS: con regolari riduzioni di prezzo su tutti i nostri servizi, AWS ha regolarmente offerto una maggiore efficienza economica ai nostri clienti acquisiti. Ad aprile 2022, AWS ha ridotto i prezzi 115 volte dal suo lancio nel 2006. Se hai ancora qualche dubbio in merito a decisioni commerciali da prendere a causa di questioni relative ai prezzi, puoi fare riferimento ai nuovi tariffari, che includono riduzioni dei prezzi e nuove integrazioni dei servizi. Puoi avere ulteriori informazioni sulle precedenti riduzioni dei prezzi, comprese quelle relative alle istanze Amazon Elastic Compute Cloud (Amazon EC2), nella categoria relativa alla riduzione dei prezzi del Blog delle novità di AWS.
- Eventi e incontri AWS: Partecipa al summit AWS locale e a qualsiasi incontro locale con altre organizzazioni della tua area. Se non riesci a partecipare dal vivo, prova ad accedere agli eventi virtuali per poter ascoltare gli esperti AWS e rimanere informato sui casi aziendali di altri clienti.
- Organizza riunioni con il team del tuo account: Pianifica una cadenza regolare di incontri con il team del tuo account, organizza riunioni con il team e discuti delle tendenze del settore e dei servizi AWS. Parla con gli account manager, i solutions architect e i team di supporto a te assegnati.

Risorse

Documenti correlati:

- Gestione dei costi AWS
- Novità di AWS
- Blog delle novità di AWS

Esempi correlati:

- Amazon EC2 – 15 Years of Optimizing and Saving Your IT Costs (15 anni di ottimizzazione e risparmio dei costi IT)
- AWS News Blog - Price Reduction (Blog delle novità di AWS - Riduzione dei prezzi)

Consapevolezza delle spese e dell'utilizzo

Domande

- COST 2 In che modo gestisci l'utilizzo?
- COST 3 In che modo monitori l'utilizzo e il costo?
- COST 4 In che modo ritiri le risorse?

COST 2 In che modo gestisci l'utilizzo?

Stabilisci policy e meccanismi per assicurarti di sostenere costi adeguati mentre raggiungi gli obiettivi. Utilizzando un approccio di controllo e bilanciamento reciproco, è possibile innovare senza spendere troppo.

Best practice

- COST02-BP01 Sviluppo di politiche basate sui requisiti dell'organizzazione
- COST02-BP02 Implementazione di obiettivi e target
- COST02-BP03 Implementazione di una struttura di account
- COST02-BP04 Implementazione di gruppi e ruoli
- COST02-BP05 Implementazione dei controlli di costo
- COST02-BP06 Monitoraggio del ciclo di vita del progetto

COST02-BP01 Sviluppo di politiche basate sui requisiti dell'organizzazione

Sviluppa politiche che definiscono come le risorse vengono gestite dalla tua organizzazione. Le policy devono coprire gli aspetti dei costi relativi alle risorse e ai carichi di lavoro, compresa la creazione, la modifica e la disattivazione nel ciclo di vita delle risorse.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

Comprendere i costi e i fattori chiave della tua organizzazione è fondamentale per gestire i costi e l'utilizzo in modo efficiente e per identificare le opportunità di riduzione dei costi. In genere, le organizzazioni gestiscono molteplici carichi di lavoro eseguiti da più team. Questi team possono trovarsi in diverse unità aziendali, ognuna con un proprio flusso di ricavi. La capacità di attribuire i costi delle risorse ai singoli proprietari del carico di lavoro, del prodotto o dell'organizzazione incoraggia un comportamento di utilizzo efficiente e contribuisce a ridurre gli sprechi. Un'attribuzione precisa dei costi ti consente di capire se le unità aziendali e i prodotti sono redditizi e ti aiuta a prendere decisioni più consapevoli in merito a dove allocare le risorse all'interno dell'azienda. La consapevolezza dell'utilizzo a tutti i livelli dell'organizzazione è fondamentale per promuovere il cambiamento, poiché la modifica dell'utilizzo determina variazioni dei costi. Prova a adottare una strategia versatile per acquisire consapevolezza delle tue spese.

il primo passo per attuare la governance consiste nell'utilizzare i requisiti della tua organizzazione per sviluppare policy per l'utilizzo del cloud. Queste policy definiscono il modo in cui l'organizzazione

utilizza il cloud e il modo in cui le risorse vengono gestite. Le politiche devono coprire tutti gli aspetti dei costi relativi alle risorse e ai carichi di lavoro correlati a costi o utilizzo, compresa la creazione, la modifica e la disattivazione durante il ciclo di vita della risorsa.

Le policy devono essere semplici, in modo che siano facilmente comprensibili e possano essere implementate in modo efficace in tutta l'organizzazione. Inizia con policy ampie e di alto livello, ad esempio in quale regione geografica è consentito l'utilizzo o l'ora del giorno in cui le risorse devono essere in esecuzione. Raffina gradualmente le policy per le varie unità organizzative e i diversi carichi di lavoro. Le policy comuni includono i servizi e le funzionalità che possono essere utilizzati (ad esempio, storage dalle prestazioni inferiori negli ambienti di test o sviluppo) e i tipi di risorse che possono essere utilizzati dai diversi gruppi (ad esempio, le dimensioni massime di una risorsa in un account di sviluppo possono essere impostate a medium).

Passaggi dell'implementazione

- Organizzare riunioni con i membri del team: Per sviluppare policy, richiedi a tutti i membri dei team della tua organizzazione di specificare i loro requisiti e di documentarli. Segui un approccio iterativo iniziando in modo generale e perfezionando continuamente le unità più piccole in ogni fase. I membri del team includono quelli con interesse diretto nel carico di lavoro, ad esempio unità organizzative o proprietari di applicazioni, nonché gruppi di supporto, come i team di sicurezza e i team finanziari.
- Definizione delle posizioni per il carico di lavoro: Definisci dove opera il carico di lavoro, incluso il paese e l'area all'interno del paese. Queste informazioni vengono utilizzate per la mappatura su Regioni AWS e sulle zone di disponibilità.
- Definizione e raggruppamento di servizi e risorse: Definisci i servizi richiesti dai carichi di lavoro. Per ogni servizio, specifica i tipi, la dimensione e il numero di risorse richieste. Definisci i gruppi per le risorse in base alla funzione, ad esempio i server di applicazioni o lo storage di database. Le risorse possono appartenere a più gruppi.
- Definizione e raggruppamento degli utenti per funzione: Definisci gli utenti che interagiscono con il carico di lavoro, concentrandoti su ciò che fanno e su come utilizzano il carico di lavoro, non su chi sono o sulla loro posizione nell'organizzazione. Raggruppa utenti o funzioni simili. Puoi utilizzare le policy gestite da AWS come guida di riferimento.
- Definizione delle operazioni: Utilizzando le posizioni, le risorse e gli utenti identificati in precedenza, definisci le azioni richieste da ciascuno di essi per ottenere i risultati del carico di lavoro durante il ciclo di vita (sviluppo, funzionamento e disattivazione). Identifica le operazioni in base ai gruppi, non ai singoli elementi nei gruppi, in ogni posizione. Inizia in generale con lettura o scrittura, quindi perfeziona le azioni specifiche per ciascun servizio.

- Definizione del periodo di revisione: I carichi di lavoro e i requisiti organizzativi possono cambiare nel corso del tempo. Definisci la pianificazione della revisione del carico di lavoro per assicurarti che sia allineata alle priorità organizzative.
- Documentazione delle policy: Assicurati che le policy definite siano accessibili come richiesto dalla tua organizzazione. Queste policy vengono utilizzate per implementare, mantenere e controllare l'accesso agli ambienti.

Risorse

Documenti correlati:

- Politiche gestite da AWS per le funzioni dell'attività
- Strategia di fatturazione con account multipli di AWS
- Operazioni, risorse e chiavi di condizione per i servizi AWS
- Prodotti cloud
- Controllo dell'accesso a Regioni AWS utilizzando le politiche IAM
- Regioni e zone di disponibilità dell'infrastruttura globale

COST02-BP02 Implementazione di obiettivi e target

Implementa obiettivi di costi e utilizzo per il carico di lavoro. Gli obiettivi forniscono indicazioni alla tua organizzazione su costi e utilizzo e i target forniscono risultati misurabili per i tuoi carichi di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

sviluppa obiettivi e target di costi e utilizzo per la tua organizzazione. Gli obiettivi forniscono all'organizzazione linee guida e indicazioni sui risultati previsti. I target forniscono i risultati specifici e misurabili da raggiungere. Ad esempio, un obiettivo potrebbe essere: l'utilizzo della piattaforma deve aumentare in modo significativo, implicando solamente un aumento minore (non lineare) dei costi. Un esempio di target invece potrebbe essere: un aumento del 20% dell'utilizzo della piattaforma, con un aumento dei costi inferiore al 5%. Un altro obiettivo comune è che i carichi di lavoro devono essere più efficienti ogni 6 mesi. Il target associato a tale obiettivo è che il costo per output del carico di lavoro deve diminuire del 5% ogni 6 mesi.

Un obiettivo comune per i carichi di lavoro nel cloud è l'incremento dell'efficienza del carico di lavoro, ossia la riduzione del costo per il risultato aziendale del carico di lavoro nel corso del tempo.

Si consiglia di implementare questo obiettivo per tutti i carichi di lavoro e di stabilire, inoltre, un target come l'aumento dell'efficienza del 5% ogni 6-12 mesi. Questo può essere ottenuto nel cloud attraverso la creazione di capacità per l'ottimizzazione dei costi e tramite il rilascio di nuovi servizi e loro funzionalità.

Passaggi dell'implementazione

- Definisci i livelli di utilizzo previsti: Concentrati sui livelli di utilizzo da cui iniziare. Coinvolgi i proprietari dell'applicazione, i team di marketing e i team aziendali a livello più ampio per capire quali saranno i livelli di utilizzo previsti per il carico di lavoro. Considera in che modo cambierà la domanda dei clienti nel corso del tempo e se ci saranno modifiche dovute ad aumenti stagionali o campagne di marketing.
- Definisci le risorse e i costi del carico di lavoro: Con i livelli di utilizzo definiti, quantifica le modifiche nelle risorse del carico di lavoro necessarie per soddisfare questi livelli di utilizzo. Potresti dover aumentare le dimensioni o il numero di risorse per un componente del carico di lavoro, aumentare il trasferimento dei dati o modificare i componenti del carico di lavoro in un servizio diverso a un livello specifico. Specifica quali saranno i costi in ciascuno di questi punti principali e quali saranno le variazioni dei costi in caso di variazioni di utilizzo.
- Definisci gli obiettivi aziendali: Prendendo l'output dalle variazioni previste in termini di utilizzo e costi, combinalo con le modifiche previste nella tecnologia o in qualsiasi programma in esecuzione e sviluppa obiettivi per il carico di lavoro. Gli obiettivi devono considerare l'utilizzo, il costo e la relazione tra i due. Assicurati che siano disponibili programmi a livello di organizzazione, ad esempio lo sviluppo di competenze come la formazione e l'istruzione, se ci sono variazioni previste dei costi senza variazioni di utilizzo.
- Definisci i target: Per ciascuno degli obiettivi definiti, specifica un target misurabile. Se l'obiettivo è aumentare l'efficienza nel carico di lavoro, il target quantificherà il miglioramento, generalmente espresso in risultati aziendali per dollaro speso, e il momento in cui sarà efficace.

Risorse

Documenti correlati:

- Politiche gestite da AWS per le funzioni dell'attività
- Strategia di fatturazione con account multipli di AWS
- Controllo dell'accesso a Regioni AWS utilizzando le politiche IAM

COST02-BP03 Implementazione di una struttura di account

Implementa una struttura di account che si adatta alla tua organizzazione. Questo aiuta a ripartire e gestire i costi in tutta la tua organizzazione.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

AWS presenta una struttura degli account ad albero, con un account principale (il padre, precedentemente detto account di pagamento) e vari account collegati (i figli, noti come account membri). Una best practice è di avere sempre almeno un account principale con un account membro, indipendentemente dalle dimensioni o dall'utilizzo dell'organizzazione. Tutte le risorse del carico di lavoro devono risiedere solo all'interno degli account membri.

Non esiste una risposta giusta o sbagliata in merito al numero di account AWS che bisognerebbe creare. Valuta i tuoi modelli operativi e di costo attuali e futuri per assicurarti che la struttura dei tuoi account AWS rispecchi quella della tua organizzazione. Alcune aziende creano molteplici account AWS per motivi aziendali, ad esempio:

- È richiesto l'isolamento amministrativo e/o fiscale e di fatturazione tra unità aziendali o centri di costo o carichi di lavoro specifici.
- Le restrizioni dei servizi AWS sono impostate in modo che risultino specifiche per determinati carichi di lavoro.
- Esiste un requisito per l'isolamento e la separazione tra carichi di lavoro e risorse.

All'interno di AWS Organizations, la fatturazione consolidata crea il costrutto tra uno o più account membri e l'account di gestione. Gli account membri consentono di isolare e distinguere i costi e l'utilizzo per gruppi. Una pratica comune è quella di avere account membri separati per ciascuna unità aziendale (come finanza, marketing e vendite), per il ciclo di vita di ciascun ambiente (come sviluppo, test e produzione) o per ciascun carico di lavoro (carico di lavoro a, b e c), e poi aggregare questi account membri tramite la fatturazione consolidata.

La fatturazione consolidata consente di accorpare i pagamenti di più account membri AWS sotto un unico account principale e, al tempo stesso, di fornire comunque visibilità all'attività di ciascun account membro. Poiché i costi e l'utilizzo vengono aggregati nell'account di gestione, questo consente di massimizzare gli sconti per volume di servizio e di massimizzare l'utilizzo degli sconti a fronte di impegni (Savings Plans e istanze riservate) per ottenere gli sconti più elevati.

AWS Control Tower può impostare e configurare rapidamente più account AWS, garantendo una governance in linea con i requisiti della tua organizzazione.

Passaggi dell'implementazione

- Definisci i requisiti di separazione: I requisiti di separazione sono una combinazione di più fattori, tra cui sicurezza, affidabilità e costrutti finanziari. Analizza ciascun fattore in ordine e specifica se il carico di lavoro o l'ambiente del carico di lavoro deve essere separato da altri carichi di lavoro. La sicurezza garantisce il rispetto dei requisiti di accesso e dei dati. L'affidabilità garantisce la gestione dei limiti, in modo che gli ambienti e i carichi di lavoro non influiscano sugli altri. I costrutti finanziari garantiscono separazione finanziaria e responsabilità rigorose. Esempi comuni di separazione sono i carichi di lavoro di produzione e test eseguiti in account separati o l'utilizzo di un account separato in modo che i dati di fatturazione possano essere forniti a un'organizzazione di terze parti.
- Definisci i requisiti di raggruppamento: I requisiti per il raggruppamento non sostituiscono i requisiti di separazione, ma vengono utilizzati a supporto della gestione. Raggruppa ambienti o carichi di lavoro simili che non richiedono separazione. Un esempio di questo è costituito dal raggruppamento di più ambienti di test o sviluppo da uno o più carichi di lavoro.
- Definisci la struttura dell'account: Utilizzando queste separazioni e questi raggruppamenti, specifica un account per ogni gruppo e assicurati che i requisiti di separazione siano mantenuti. Questi account sono i tuoi account membri o collegati. Raggruppando questi account membri in un unico account di gestione/di pagamento, puoi combinare l'utilizzo, che consente maggiori sconti per volume su tutti gli account e fornisce una singola fattura per tutti gli account. È possibile separare i dati di fatturazione e fornire a ciascun account membro una visualizzazione individuale dei dati di fatturazione. Se un account membro non deve avere i dati di utilizzo o di fatturazione visibili a qualsiasi altro account, oppure se è necessaria una fattura separata da parte di AWS, definisci più account di gestione/di pagamento. In questo caso, ogni account membro ha il proprio account di gestione/di pagamento. Le risorse devono sempre essere collocate negli account membri o collegati. Gli account di gestione/di pagamento devono essere utilizzati solo per la gestione.

Risorse

Documenti correlati:

- Politiche gestite da AWS per le funzioni dell'attività
- Strategia di fatturazione con account multipli di AWS
- Controllo dell'accesso a Regioni AWS utilizzando le politiche IAM
- AWS Control Tower

- AWS Organizations
- Fatturazione consolidata

Esempi correlati:

- Divisione della CUR e condivisione dell'accesso

COST02-BP04 Implementazione di gruppi e ruoli

Implementa gruppi e ruoli che si allineino alle tue policy e controlla chi può creare, modificare o ritirare istanze e risorse in ogni gruppo. Ad esempio, implementa gruppi di sviluppo, test e produzione. Questo si applica ai servizi AWS e a soluzioni di terze parti.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Dopo avere sviluppato le policy, è possibile creare gruppi logici e ruoli degli utenti all'interno dell'organizzazione. In questo modo puoi assegnare le autorizzazioni e controllare l'utilizzo. Inizia con i raggruppamenti di persone di alto livello. Generalmente questi seguono la definizione delle unità organizzative e dei ruoli lavorativi (ad esempio, amministratore di sistema nel reparto IT o controllore finanziario). I gruppi raggruppano persone che eseguono attività simili e necessitano di un accesso simile. I ruoli definiscono che cosa un gruppo deve fare. Ad esempio, un amministratore di sistema nel reparto IT deve disporre di un accesso che permetta di creare tutte le risorse, mentre un membro del team di analisi ha la necessità di creare soltanto risorse di analisi.

Passaggi dell'implementazione

- Implementa i gruppi: Utilizzando i gruppi di utenti definiti nelle policy dell'organizzazione, implementa i gruppi corrispondenti, se necessario. Fai riferimento al pilastro della sicurezza per le best practice su utenti, gruppi e autenticazione.
- Implementa ruoli e politiche: Utilizzando le operazioni definite nelle policy dell'organizzazione, crea i ruoli e le policy di accesso richiesti. Fai riferimento al pilastro della sicurezza per le best practice su ruoli e policy.

Risorse

Documenti correlati:

- Politiche gestite da AWS per le funzioni dell'attività
- Strategia di fatturazione con account multipli di AWS
- Controllo dell'accesso a Regioni AWS utilizzando le politiche IAM
- Principio della sicurezza Well-Architected

Esempi correlati:

- Well-Architected Labs: Identità e accesso base

COST02-BP05 Implementazione dei controlli di costo

Implementa controlli basati sulle policy dell'organizzazione e gruppi e ruoli definiti. Questi garantiscono che i costi siano sostenuti solo in base ai requisiti dell'organizzazione, ad esempio, controllano l'accesso alle regioni o ai tipi di risorse con le policy AWS Identity and Access Management (IAM).

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Un primo passo comune per implementare i controlli dei costi consiste nell'impostare le notifiche quando si verificano eventi di costi o utilizzo al di fuori delle policy. In questo modo, puoi agire rapidamente e verificare se è necessaria un'azione correttiva, senza limitare o influire negativamente sui carichi di lavoro o sulle nuove attività. Dopo avere appreso i limiti del carico di lavoro e dell'ambiente, puoi applicare la governance. In AWS, le notifiche vengono effettuate con Budget AWS, che consente di definire un budget mensile per i costi, l'utilizzo e gli sconti a fronte di impegni di AWS (Savings Plans e istanze riservate). Puoi creare budget a livello di costo aggregato (ad esempio, tutti i costi) o a un livello più granulare, nel quale includi solo dimensioni specifiche come account membri, servizi, tag o zone di disponibilità.

In secondo luogo, puoi applicare le politiche di governance in AWS tramite AWS Identity and Access Management (IAM) e le policy di controllo dei servizi (SCP) di AWS Organizations. IAM consente di gestire in modo sicuro l'accesso ai servizi e alle risorse AWS. Utilizzando IAM, puoi controllare chi può creare e gestire le risorse AWS, il tipo di risorse che possono essere create e dove possono essere create. Ciò riduce al minimo la creazione di risorse che non sono necessarie. Utilizza i ruoli e i gruppi creati in precedenza e assegna le policy IAM per garantire l'utilizzo corretto. Le SCP offrono il controllo centralizzato sul numero massimo di autorizzazioni disponibili per tutti gli account nella tua organizzazione, assicurando che i tuoi account rimangano entro le linee guida di controllo degli

accessi. Le SCP sono disponibili soltanto in un'organizzazione con tutte le funzionalità abilitate e possono essere configurate in modo da rifiutare o consentire operazioni agli account membri per impostazione predefinita. Consulta il whitepaper sul principio della sicurezza secondo il Canone di architettura per ulteriori dettagli sull'implementazione della gestione degli accessi.

La governance può essere implementata anche tramite la gestione delle Service Quotas. Assicurandoti che le quote di servizio siano impostate con spese minime e siano gestite in modo accurato, puoi ridurre al minimo la creazione di risorse che non rientrano nei requisiti della tua organizzazione. Per ottenere questo risultato, devi comprendere la velocità con cui i tuoi requisiti possono cambiare, valutare i progetti in corso (sia la creazione sia la disattivazione di risorse) e considerare la velocità con cui è possibile implementare le modifiche alle quote. Service Quotas possono essere utilizzate per aumentare le quote all'occorrenza.

Passaggi dell'implementazione

- Implementa le notifiche sulla spesa: Utilizzando le policy dell'organizzazione definite, crea dei budget AWS per inviare notifiche quando la spesa ricade al di fuori delle policy. Configura più budget dei costi, uno per ogni account, che invia una notifica sulla spesa complessiva dell'account. Quindi configura budget di costo aggiuntivi all'interno di ciascun account per unità più piccole all'interno dell'account. Queste unità variano a seconda della struttura dell'account. Alcuni esempi comuni sono Regioni AWS, carichi di lavoro (tramite i tag) o servizi AWS. Assicurati di configurare un elenco di distribuzione e-mail come destinatario per le notifiche e non un account e-mail di un singolo. Puoi configurare un budget effettivo per quando un importo viene superato oppure utilizzare un budget previsto per la notifica dell'utilizzo previsto.
- Implementa i controlli sull'utilizzo: Utilizzando le policy dell'organizzazione definite, implementa policy e ruoli IAM per specificare quali azioni possono eseguire gli utenti e quali non possono eseguire. In una policy AWS possono essere incluse più policy organizzative. Nello stesso modo in cui hai definito le policy, inizia in modo generale e quindi applica controlli più dettagliati a ogni fase. Anche le restrizioni dei servizi sono un controllo efficace sull'utilizzo. Implementa le restrizioni dei servizi corrette su tutti gli account.

Risorse

Documenti correlati:

- Politiche gestite da AWS per le funzioni dell'attività
- Strategia di fatturazione con account multipli di AWS
- Controllo dell'accesso a Regioni AWS utilizzando le politiche IAM

Esempi correlati:

- Well-Architected Labs: costi e governance
- Well-Architected Labs: costi e governance

COST02-BP06 Monitoraggio del ciclo di vita del progetto

Rileva, misura e controlla il ciclo di vita di progetti, team e ambienti per evitare di usare risorse non necessarie e pagare per esse.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

assicurati di tenere traccia dell'intero ciclo di vita del carico di lavoro. In questo modo, quando i carichi di lavoro o i componenti del carico di lavoro non sono più necessari, potrai disattivarli o modificarli. Ciò si rivela particolarmente utile quando rilasci nuovi servizi o funzionalità. I carichi di lavoro e i componenti esistenti possono essere mostrati come in uso, ma devono essere disattivati e reindirizzare i clienti al nuovo servizio. Presta attenzione alle diverse fasi dei carichi di lavoro: quando un carico di lavoro arriva in produzione, gli ambienti precedenti possono essere disattivati o notevolmente ridotti in termini di capacità fino a quando non sono nuovamente necessari.

AWS offre una serie di servizi di gestione e governance utilizzabili per il monitoraggio del ciclo di vita delle entità. Puoi utilizzare AWS Config oppure AWS Systems Manager per fornire un inventario dettagliato delle risorse e della configurazione AWS. Si consiglia di integrare questi servizi con i sistemi di gestione di progetti o asset esistenti per tenere traccia dei progetti e dei prodotti attivi all'interno della tua organizzazione. La combinazione del tuo sistema attuale con l'ampia gamma di eventi e parametri forniti da AWS ti consentirà di ottenere una panoramica degli eventi del ciclo di vita significativi e di gestire le risorse in modo proattivo per ridurre i costi non necessari.

Consulta lo whitepaper sul principio dell'eccellenza operativa secondo il Canone di architettura per ulteriori dettagli sull'implementazione del monitoraggio del ciclo di vita delle entità.

Passaggi dell'implementazione

- Esegui le revisioni del carico di lavoro: Secondo quanto definito dalle policy dell'organizzazione, controlla i progetti esistenti. Lo sforzo per l'audit deve essere proporzionale al rischio, al valore o al costo approssimativo per l'organizzazione. Le aree chiave da includere nell'audit sono il rischio di incidente o interruzione per l'organizzazione, il valore o contributo all'organizzazione

(misurato in fatturato o reputazione del marchio), il costo del carico di lavoro (misurato come costo totale delle risorse e costi operativi) e l'utilizzo del carico di lavoro (misurato in numero di risultati dell'organizzazione per unità di tempo). Se queste aree cambiano durante il ciclo di vita, sono necessarie modifiche al carico di lavoro, ad esempio la disattivazione completa o parziale.

Risorse

Documenti correlati:

- AWS Config
- AWS Systems Manager
- Politiche gestite da AWS per le funzioni dell'attività
- Strategia di fatturazione con account multipli di AWS
- Controllo dell'accesso a Regioni AWS utilizzando le politiche IAM

COST 3 In che modo monitori l'utilizzo e il costo?

Stabilisci policy e procedure per monitorare e allocare i costi in modo appropriato. Ciò ti consente di misurare e migliorare l'efficienza in termini di costi del carico di lavoro.

Best practice

- COST03-BP01 Configurazione di fonti di informazione dettagliate
- COST03-BP02 Identificazione delle categorie di attribuzione dei costi
- COST03-BP03 Definizione dei parametri dell'organizzazione
- COST03-BP04 Configurazione degli strumenti di fatturazione e di gestione dei costi
- COST03-BP05 Aggiunta di informazioni sull'organizzazione a costi e utilizzo
- COST03-BP06 Allocazione dei costi in base ai parametri del carico di lavoro

COST03-BP01 Configurazione di fonti di informazione dettagliate

Configura i report su costi e utilizzo di AWS e la granularità oraria di Cost Explorer per fornire informazioni dettagliate su costi e utilizzo. Configura il carico di lavoro per far sì che le voci di log vengano registrate per ogni risultato aziendale distribuito.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

Abilita la granularità oraria in AWS Cost Explorer e crea un AWS Cost and Usage Report (CUR). Queste origini dati forniscono la visualizzazione più accurata dei costi e dell'utilizzo dell'intera organizzazione. Il CUR fornisce una granularità di utilizzo giornaliera o oraria, tariffe, costi e attributi di utilizzo per tutti i servizi AWS addebitati. Tutte le dimensioni possibili sono incluse nel CUR, tra cui applicazione di tag, posizione, attributi di risorsa e ID account.

Configura il CUR con le seguenti personalizzazioni:

- Inclusione degli ID risorsa
- Aggiornamento automatico del CUR
- Granularità oraria
- Controllo delle versioni: sovrascrittura del report esistente
- Integrazione dei dati: Amazon Athena (formato Parquet e compressione)

Utilizzo AWS Glue per preparare i dati per l'analisi e Amazon Athena per eseguire l'analisi dei dati, utilizzando SQL per eseguire query sui dati. Puoi anche utilizzare Amazon QuickSight per creare visualizzazioni personalizzate e complesse e distribuirle in tutta l'organizzazione.

Passaggi dell'implementazione

- Configura il report su costi e utilizzo: Utilizzando la console di fatturazione, configura almeno un report costi e utilizzo. Configura un report con granularità oraria che include tutti gli identificatori e gli ID risorsa. Puoi anche creare altri report con granularità diverse per fornire informazioni di riepilogo di livello superiore.
- Configura la granularità oraria in Cost Explorer: Utilizzando la console di fatturazione, abilita i dati orari e i dati a livello di risorsa.

> **Note**
>
> L'attivazione di questa funzionalità comporta dei costi. Per i dettagli fare riferimento al listino prezzi.

- Configura la registrazione dell'applicazione: Verifica che l'applicazione registri ogni risultato aziendale che distribuisce in modo che possa essere monitorato e misurato. Assicurati che la granularità di questi dati sia almeno oraria affinché possa essere abbinata ai dati relativi a costi

e utilizzo. Consulta il Whitepaper sul pilastro dell'eccellenza operativa secondo il framework well-architected per ulteriori dettagli su registrazione e monitoraggio.

Risorse

Documenti correlati:

- Configurazione dell'account AWS
- AWS Cost and Usage Report (CUR)
- AWS Glue
- Amazon QuickSight
- Gestione dei costi AWS
- Applicazione di tag alle risorse AWS
- Analisi dei costi con Budget AWS
- Analisi dei costi con Cost Explorer
- Gestione del report su costi e utilizzo di AWS
- Whitepaper sul pilastro dell'eccellenza operativa secondo il framework well-architected

Esempi correlati:

- Configurazione dell'account AWS

COST03-BP02 Identificazione delle categorie di attribuzione dei costi

Identifica le categorie dell'organizzazione che possono essere utilizzate per allocare i costi all'interno della tua organizzazione.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

collabora con il tuo team finanziario e altri stakeholder per comprendere i requisiti di allocazione dei costi all'interno della tua organizzazione. I costi del carico di lavoro devono essere allocati per tutto il ciclo di vita, inclusi sviluppo, test, produzione e disattivazione. Comprendi in che modo i costi sostenuti per formazione, sviluppo del personale e creazione di idee sono attribuiti all'interno dell'organizzazione. Questo può essere utile per allocare correttamente gli account utilizzati per questo scopo ai budget di formazione e sviluppo, anziché ai budget generici dei costi IT.

Passaggi dell'implementazione

- Definisci le categorie dell'organizzazione: Organizza riunioni con le parti interessate per definire categorie che riflettano la struttura e i requisiti della tua organizzazione. Queste saranno mappate direttamente sulla struttura delle categorie finanziarie esistenti, ad esempio unità aziendale, budget, centro di costi o reparto. Osserva i risultati che il cloud offre per la tua azienda, ad esempio la formazione o l'istruzione, poiché anche queste sono categorie organizzative. A una risorsa possono essere assegnate più categorie e una risorsa può essere in più categorie diverse, quindi definisci tutte le categorie necessarie.
- Definisci le categorie funzionali: Organizza riunioni con le parti interessate per definire categorie che riflettano le funzioni che hai all'interno della tua azienda. Si può trattare del carico di lavoro o dei nomi delle applicazioni e il tipo di ambiente, ad esempio produzione, test o sviluppo. A una risorsa possono essere assegnate più categorie e una risorsa può essere in più categorie diverse, quindi definisci tutte le categorie necessarie.

Risorse

Documenti correlati:

- Applicazione di tag alle risorse AWS
- Analisi dei costi con Budget AWS
- Analisi dei costi con Cost Explorer
- Gestione del report su costi e utilizzo di AWS

COST03-BP03 Definizione dei parametri dell'organizzazione

Definisci i parametri dell'organizzazione necessari per questo carico di lavoro. I parametri esemplificativi di un carico di lavoro sono i report dei clienti prodotti o le pagine web scaricate dai clienti.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

comprendi in che modo viene misurato l'output del carico di lavoro rispetto al successo aziendale. Ogni carico di lavoro ha in genere un piccolo set di output principali che indicano le prestazioni. Se disponi di un carico di lavoro complesso con molti componenti, puoi dare priorità alle voci dell'elenco o definire e monitorare i parametri per ogni componente. Collabora con i tuoi team per capire quali

parametri utilizzare. Questa unità verrà utilizzata per comprendere l'efficienza del carico di lavoro o il costo per ciascun output aziendale.

Passaggi dell'implementazione

- Definisci i risultati del carico di lavoro: Organizza riunioni con le parti interessate dell'azienda e definisci i risultati del carico di lavoro. Si tratta di una misura principale dell'utilizzo da parte dei clienti e devono essere parametri aziendali e non parametri tecnici. Deve esserci un piccolo numero di parametri di alto livello (meno di cinque) per carico di lavoro. Se il carico di lavoro produce più risultati per diversi casi d'uso, raggruppali in un singolo parametro.
- Definisci i risultati dei componenti del carico di lavoro: Facoltativamente, se disponi di un carico di lavoro grande e complesso oppure puoi suddividere facilmente il carico di lavoro in componenti (ad esempio microservizi) con input e output ben definiti, definisci i parametri per ogni componente. Lo sforzo deve riflettere il valore e il costo del componente. Inizia con i componenti più grandi e punta ai componenti più piccoli.

Risorse

Documenti correlati:

- Applicazione di tag alle risorse AWS
- Analisi dei costi con Budget AWS
- Analisi dei costi con Cost Explorer
- Gestione del report su costi e utilizzo di AWS

COST03-BP04 Configurazione degli strumenti di fatturazione e di gestione dei costi

Configura AWS Cost Explorer e Budget AWS in linea con le policy della tua organizzazione.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

per modificare utilizzo e costi, ogni persona nell'organizzazione deve avere accesso alle informazioni relative a costi e utilizzo. È consigliabile che tutti i carichi di lavoro e i team dispongano dei seguenti strumenti configurati quando utilizzano il cloud:

- Report: riepilogano tutte le informazioni su costi e utilizzo
- Notifiche: avvisano quando il costo o l'utilizzo non rientra nei limiti definiti.

- Stato attuale: configura un pannello di controllo che mostra i livelli correnti di costi e utilizzo. Il pannello di controllo deve essere disponibile in un luogo altamente visibile all'interno dell'ambiente di lavoro (simile a un pannello di controllo delle operazioni).
- Tendenze: offri la possibilità di mostrare la variabilità dei costi e dell'utilizzo nel periodo di tempo richiesto e con la granularità richiesta.
- Previsioni: offri la possibilità di mostrare i costi futuri stimati.
- Monitoraggio: mostra i costi e l'utilizzo attuali rispetto a obiettivi o target stabiliti.
- Analisi: offri ai membri del team la possibilità di eseguire analisi personalizzate e approfondite fino alla granularità oraria, con tutte le dimensioni possibili.

Puoi utilizzare gli strumenti nativi di AWS, ad esempio AWS Cost Explorer, Budget AWSe Amazon Athena con Amazon QuickSight per fornire questa funzionalità. Puoi anche utilizzare strumenti di terze parti, tuttavia devi assicurarti che i costi di tali strumenti apportino un valore aggiunto effettivo alla tua organizzazione.

Passaggi dell'implementazione

- Crea un gruppo di ottimizzazione dei costi: Configura l'account e crea un gruppo che abbia accesso ai report costi e utilizzo richiesti. Questo gruppo deve includere i rappresentanti di tutti i team che possiedono o gestiscono un'applicazione. In questo modo ogni team ha accesso alle informazioni relative ai propri costi e utilizzo.
- Configura Budget AWS: Configura Budget AWS su tutti gli account per il tuo carico di lavoro. Imposta un budget per la spesa complessiva dell'account e un budget per il carico di lavoro utilizzando i tag.
- Configura AWS Cost Explorer: Configura AWS Cost Explorer per il carico di lavoro e gli account. Crea un pannello di controllo per il carico di lavoro che monitori la spesa complessiva e i parametri di utilizzo chiave per il carico di lavoro.
- Configura gli strumenti avanzati: Facoltativamente, puoi creare uno strumento personalizzato per la tua organizzazione che fornisca dettagli e granularità aggiuntivi. Puoi implementare le funzionalità di analisi avanzata utilizzando Amazon Athenae i pannelli di controllo utilizzando Amazon QuickSight.

Risorse

Documenti correlati:

- Applicazione di tag alle risorse AWS
- Analisi dei costi con Budget AWS
- Analisi dei costi con Cost Explorer
- Gestione del report su costi e utilizzo di AWS

Esempi correlati:

- Well-Architected Labs: configurazione dell'account AWS
- Well-Architected Labs: visualizzazione della fatturazione
- Well-Architected Labs: utilizzo di costi e governance
- Well-Architected Labs: analisi di costi e utilizzo
- Well-Architected Labs: visualizzazione di costi e utilizzo

COST03-BP05 Aggiunta di informazioni sull'organizzazione a costi e utilizzo

Definisci uno schema di applicazione di tag basato sull'organizzazione, attributi del carico di lavoro e categorie di allocazione dei costi. Implementa l'applicazione di tag su tutte le risorse. Utilizza Cost Categories per raggruppare i costi e l'utilizzo in base agli attributi dell'organizzazione.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Implementazione dei tag in AWS per aggiungere informazioni sull'organizzazione alle risorse, che verranno, quindi, integrate alle informazioni su costi e utilizzo. Un tag è una coppia chiave-valore: la chiave è definita e deve essere univoca all'interno dell'organizzazione, mentre il valore è univoco per un gruppo di risorse. Ad esempio, una coppia chiave-valore può essere costituita da ambiente (chiave) e produzione (valore). Tutte le risorse nell'ambiente di produzione avranno questa coppia chiave-valore. L'applicazione di tag consente di categorizzare e monitorare i costi con informazioni significative e rilevanti sull'organizzazione. Puoi applicare tag che rappresentano categorie dell'organizzazione (ad esempio, centri di costo, nomi di applicazioni, progetti o proprietari) e identificano carichi di lavoro e rispettive funzionalità (come test o produzione) per attribuire i costi e l'utilizzo all'interno dell'organizzazione.

Quando applichi i tag alle tue risorse AWS (come le istanze Amazon Elastic Compute Cloud o i bucket Amazon Simple Storage Service) e li attivi, AWS aggiunge queste informazioni ai report su

costi e utilizzo. Puoi creare report e condurre analisi su risorse con tag e senza tag per incrementare la conformità con le policy di gestione dei costi interne e garantire un'attribuzione accurata.

La creazione e l'implementazione di uno standard per l'applicazione di tag AWS tra gli account dell'organizzazione ti consente di gestire e amministrare gli ambienti AWS in modo coerente e uniforme. Utilizza policy di tag in AWS Organizations per definire regole su come i tag possono essere applicati alle risorse AWS nei tuoi account in AWS Organizations. Le policy di tag consentono di adottare con facilità un approccio standardizzato per l'applicazione di tag alle risorse AWS.

AWS Tag Editor consente di aggiungere, eliminare e gestire tag di più risorse.

AWS Cost Categories consente di assegnare ai tuoi costi significati per l'organizzazione, senza necessità di applicare tag alle risorse. Puoi mappare le informazioni su costi e utilizzo attribuendole a strutture organizzative interne univoche. Puoi definire regole di categoria per mappare e categorizzare i costi utilizzando le dimensioni di fatturazione, ad esempio account e tag. Questo offre un altro livello di funzionalità di gestione oltre all'applicazione di tag. Puoi anche mappare account e tag specifici attribuendoli a più progetti.

Passaggi dell'implementazione

- Definisci uno schema di applicazione dei tag: Riunisci tutte le parti interessate di tutta l'azienda per definire uno schema. Questo generalmente include i ruoli tecnici, finanziari e di gestione. Definisci un elenco di tag che tutte le risorse devono avere, nonché un elenco di tag che le risorse dovrebbero avere. Verifica che i nomi e i valori dei tag siano coerenti all'interno dell'organizzazione.
- Applica i tag alle risorse: Utilizzando le categorie di attribuzione dei costi definite, posiziona i tag su tutte le risorse nei carichi di lavoro in base alle categorie. Utilizza strumenti come l'interfaccia a riga di comando (CLI), Tag Editor o Systems Manager per aumentare l'efficienza.
- Implementazione di Cost Categories: Puoi creare delle Cost Categories senza implementare l'applicazione dei tag. Cost Categories utilizza le dimensioni di costo e utilizzo esistenti. Crea regole di categoria dallo schema e implementale in Cost Categories.
- Automatizza l'applicazione di tag: Per verificare di mantenere elevati livelli di applicazione di tag tra tutte le risorse, automatizza l'applicazione di tag in modo che le risorse siano contrassegnate automaticamente al momento della creazione. Utilizza le funzionalità all'interno del servizio o servizi come AWS CloudFormation per assicurarti che le risorse siano contrassegnate al momento della creazione. Puoi anche creare un microservizio personalizzato che scansioni periodicamente il carico di lavoro e rimuova le risorse non contrassegnate, l'ideale per ambienti di test e sviluppo.
- Monitoraggio ed elaborazione di report sull'applicazione di tag: Per verificare di mantenere elevati livelli di applicazione di tag nella tua organizzazione, segnala e monitora i tag tra i tuoi carichi di

lavoro. Puoi utilizzare AWS Cost Explorer per visualizzare il costo delle risorse con tag e senza tag oppure utilizzare servizi come Tag Editor. Verifica regolarmente il numero di risorse senza tag e aggiungi i tag fino a raggiungere il livello desiderato di applicazione di tag.

Risorse

Documenti correlati:

- Applicazione di tag alle risorse con AWS CloudFormation
- AWS Cost Categories
- Applicazione di tag alle risorse AWS
- Amazon EC2 e Amazon EBS aggiungono il supporto per l'applicazione di tag alle risorse al momento della creazione
- Analisi dei costi con Budget AWS
- Analisi dei costi con Cost Explorer
- Gestione del report su costi e utilizzo di AWS

COST03-BP06 Allocazione dei costi in base ai parametri del carico di lavoro

Alloca i costi del carico di lavoro in base ai parametri o ai risultati aziendali per misurare l'efficienza dei costi del carico di lavoro. Implementa un processo per analizzare il Report costi e utilizzo AWS con Amazon Athena, che può fornire informazioni approfondite e funzionalità di chargeback.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

ottimizzare i costi significa conseguire i risultati aziendali al prezzo più basso, e implica l'allocazione dei costi del carico di lavoro in base ai parametri di quest'ultimo (misurati in termini di efficienza). Monitora i parametri del carico di lavoro definiti tramite file di log o altre funzionalità di monitoraggio dell'applicazione. Combina questi dati con i costi del carico di lavoro, che possono essere ottenuti osservando i costi con un determinato valore di tag o ID account. Si consiglia di eseguire questa analisi a livello orario. L'efficienza cambia in genere se disponi di alcuni componenti di costo statico (ad esempio, un database back-end in esecuzione 24 ore su 24, 7 giorni su 7) con un tasso di richiesta variabile (ad esempio, picchi di utilizzo tra le 9:00 e le 17:00, con poche richieste di notte). Comprendere la relazione tra i costi statici e i costi variabili ti aiuterà a rendere più mirate le tue attività di ottimizzazione.

Passaggi dell'implementazione

- Alloca i costi ai parametri del carico di lavoro: Utilizzando i parametri e l'applicazione di tag definiti configurati, crea un parametro che combini l'output del carico di lavoro e il costo del carico di lavoro. Utilizza i servizi di analisi come Amazon Athena e Amazon QuickSight per creare un pannello di controllo in grado di visualizzare l'efficienza del carico di lavoro complessivo e di ogni suo componente.

Risorse

Documenti correlati:

- Applicazione di tag alle risorse AWS
- Analisi dei costi con Budget AWS
- Analisi dei costi con Cost Explorer
- Gestione del report su costi e utilizzo di AWS

COST 4 In che modo ritiri le risorse?

Implementa il controllo del cambiamento e la gestione delle risorse dall'inizio del progetto alla fine del ciclo di vita. In questo modo, puoi chiudere o interrompere le risorse non utilizzate per ridurre gli sprechi.

Best practice

- COST04-BP01 Monitoraggio delle risorse durante il loro ciclo di vita
- COST04-BP02 Implementazione di un processo di disattivazione
- COST04-BP03 Disattivazione delle risorse
- COST04-BP04 Disattivazione automatica delle risorse

COST04-BP01 Monitoraggio delle risorse durante il loro ciclo di vita

Definisci e implementa un metodo per monitorare le risorse e le loro associazioni con i sistemi durante il loro ciclo di vita. Puoi usare l'applicazione di tag per identificare il carico di lavoro o la funzione della risorsa.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

ritira le risorse dei carichi di lavoro che non sono più necessarie. Un esempio comune sono le risorse utilizzate per i test: dopo il completamento dei test, le risorse possono essere rimosse. Monitorare le risorse con i tag (ed esecuzione di report su tali tag) ti aiuterà a identificare gli asset da disattivare. L'utilizzo dei tag è un modo efficace per monitorare le risorse: puoi etichettare la risorsa con la relativa funzione o con una data nota in cui può essere disattivata. Puoi quindi eseguire i report su questi tag. Valori di esempio per l'applicazione di tag alla funzionalità sono `test funzionalità-X` per identificare lo scopo della risorsa in termini di ciclo di vita del carico di lavoro.

Passaggi dell'implementazione

- Implementa uno schema di applicazione di tag: Implementa uno schema di applicazione di tag che identifichi il carico di lavoro a cui appartiene la risorsa, verificando che tutte le risorse all'interno del carico di lavoro siano contrassegnate di conseguenza.
- Implementa il monitoraggio del throughput o dell'output del carico di lavoro. Implementa il monitoraggio o gli allarmi del throughput del carico di lavoro, attivandoli per richieste di input o completamenti di output. Configuralo per fornire notifiche quando le richieste o gli output del carico di lavoro scendono a zero, indicando che le risorse del carico di lavoro non sono più utilizzate. Incorpora un fattore temporale se il carico di lavoro scende periodicamente a zero in condizioni normali.

Risorse

Documenti correlati:

- AWS Auto Scaling
- AWS Trusted Advisor
- Applicazione di tag alle risorse AWS
- Pubblicazione di parametri personalizzati

COST04-BP02 Implementazione di un processo di disattivazione

Implementa un processo per identificare e disattivare le risorse orfane.

Livello di rischio associato se questa best practice non fosse adottata: Alta

Guida all'implementazione

implementa un processo standardizzato in tutta l'organizzazione per identificare e rimuovere le risorse inutilizzate. Il processo deve definire la frequenza di esecuzione della ricerca e i processi per rimuovere la risorsa al fine di garantire che tutti i requisiti dell'organizzazione siano soddisfatti.

Passaggi dell'implementazione

- Crea e implementa un processo di disattivazione: Collaborando con sviluppatori e proprietari del carico di lavoro, crea un processo di disattivazione per il carico di lavoro e le relative risorse. Il processo deve coprire il metodo per verificare se il carico di lavoro è in uso e anche se ciascuna delle risorse del carico di lavoro è in uso. Il processo dovrebbe, inoltre, coprire le fasi necessarie per disattivare la risorsa rimuovendola dal servizio e garantendo allo stesso tempo la conformità a qualsiasi requisito normativo. Sono inoltre coperte tutte le risorse associate, ad esempio le licenze o lo storage collegato. Infine, il processo dovrebbe inviare una notifica ai proprietari del carico di lavoro indicando che il processo di disattivazione è stato eseguito.

Risorse

Documenti correlati:

- AWS Auto Scaling
- AWS Trusted Advisor

COST04-BP03 Disattivazione delle risorse

Disattivazione delle risorse attivate da eventi come audit periodici o modifiche relative all'utilizzo. La disattivazione viene in genere eseguita periodicamente ed è manuale o automatizzata.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

la frequenza e lo sforzo di ricerca delle risorse inutilizzate dovrebbero riflettere i risparmi potenziali, pertanto un account con costi contenuti deve essere analizzato con una frequenza minore rispetto a un account che ha costi maggiori. Gli eventi di ricerca e disattivazione possono essere attivati da modifiche di stato nel carico di lavoro, ad esempio il termine del ciclo di vita di un prodotto o la sua sostituzione. Le ricerche e gli eventi di disattivazione possono anche essere attivati da eventi esterni, ad esempio cambiamenti nelle condizioni di mercato o cessazione del prodotto.

Passaggi dell'implementazione

- Disattiva le risorse: Utilizzando il processo di disattivazione, disattiva tutte le risorse identificate come orfane.

Risorse

Documenti correlati:

- AWS Auto Scaling
- AWS Trusted Advisor

COST04-BP04 Disattivazione automatica delle risorse

Progetta il tuo carico di lavoro in modo da gestire con eleganza l'interruzione delle risorse, identificando e disattivando le risorse non critiche, le risorse non necessarie o quelle a basso utilizzo.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

utilizza l'automazione per ridurre o rimuovere i costi associati al processo di ritiro. Progettare il carico di lavoro per eseguire automaticamente la disattivazione ridurrà i costi complessivi del carico di lavoro durante il suo ciclo di vita. Puoi utilizzare AWS Auto Scaling per eseguire il processo di ritiro. Puoi anche implementare un codice personalizzato utilizzando un' API o SDK per ritirare automaticamente le risorse del carico di lavoro.

Passaggi dell'implementazione

- Implementa AWS Auto Scaling: Configura le risorse supportate con AWS Auto Scaling.
- Configura CloudWatch per terminare le istanze: Le istanze possono essere configurate affinché terminino in base agli allarmi CloudWatch. Utilizzando i parametri del processo di disattivazione, implementa un allarme con un'operazione Amazon Elastic Compute Cloud (Amazon EC2). Verifica l'operazione in un ambiente non di produzione prima di eseguire il roll out.
- Implementa del codice all'interno del carico di lavoro: Puoi utilizzare l'SDK AWS o la AWS CLI per disattivare le risorse del carico di lavoro. Implementa il codice all'interno dell'applicazione che si integra con AWS e interrompe o rimuove le risorse che non vengono più utilizzate.

Risorse

Documenti correlati:

- AWS Auto Scaling
- AWS Trusted Advisor
- Crea allarmi per arrestare, terminare, riavviare o recuperare un'istanza
- Nozioni di base su Amazon EC2 Auto Scaling

Risorse a costi contenuti

Domande

- COST 5 In che modo valuti i costi quando selezioni i servizi?
- COST 6 In che modo raggiungi gli obiettivi di costo quando selezioni il tipo, le dimensioni e il numero delle risorse?
- COST 7 In che modo impieghi i modelli di prezzo per ridurre i costi?
- COST 8 In che modo pianifichi i costi per il trasferimento dei dati?

COST 5 In che modo valuti i costi quando selezioni i servizi?

Amazon EC2, Amazon EBS e Amazon S3 sono servizi AWS di base. I servizi gestiti, come Amazon RDS e Amazon DynamoDB, sono servizi AWS di livello superiore o di livello applicativo. Selezionando i blocchi predefiniti e i servizi gestiti appropriati, è possibile ottimizzare questo carico di lavoro per i costi. Ad esempio, utilizzando i servizi gestiti, puoi ridurre o eliminare gran parte dei costi generali amministrativi e operativi, liberandotene per lavorare su applicazioni e attività correlate al tuo business.

Best practice

- COST05-BP01 Identificazione dei requisiti dell'organizzazione sui costi
- COST05-BP02 Analisi di tutti i componenti di questo carico di lavoro
- COST05-BP03 Esecuzione di un'analisi accurata di ciascun componente
- COST05-BP04 Selezione di software con licenze convenienti
- COST05-BP05 Selezione dei componenti del carico di lavoro per ottimizzare i costi in linea con le priorità dell'organizzazione
- COST05-BP06 Esecuzione di un'analisi dei costi per diversi valori di utilizzo nel tempo

COST05-BP01 Identificazione dei requisiti dell'organizzazione sui costi

Lavora con i membri del team per definire il bilanciamento tra l'ottimizzazione dei costi e altri pilastri, come le prestazioni e l'affidabilità, per questo carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

al momento di selezionare i servizi per un carico di lavoro, è fondamentale comprendere le priorità dell'organizzazione. Assicurati che vi sia equilibrio tra i costi e gli altri principi del canone di architettura, ad esempio prestazioni e affidabilità. Un carico di lavoro completamente ottimizzato per i costi è la soluzione più in linea con i requisiti della tua organizzazione, e non necessariamente quella con il costo più basso. Interpella tutti i team all'interno della tua organizzazione, quali quelli dedicati ai prodotti, di business, tecnici e finanziari, per raccogliere il maggior numero di informazioni.

Passaggi dell'implementazione

- Identifica i requisiti dell'organizzazione sui costi: Organizza riunioni con i membri dei team della tua organizzazione, tra cui i team di gestione dei prodotti, i team proprietari delle applicazioni, i team operativi e di sviluppo, i team di gestione e finanziari. Assegna priorità ai principi del canone di architettura per questo carico di lavoro e per i relativi componenti, l'output è un elenco ordinato dei principi. Puoi anche aggiungere un elemento di ponderazione a ciascun elemento per indicare il livello di attenzione aggiuntiva assegnato a un principio o quanto è simile il livello di attenzione assegnato a due principi.

Risorse

Documenti correlati:

- Calcolatore del costo totale di proprietà (TCO) di AWS
- Classi di archiviazione di Amazon S3
- Prodotti cloud

COST05-BP02 Analisi di tutti i componenti di questo carico di lavoro

Verifica che ogni componente del carico di lavoro venga analizzato, indipendentemente dalle dimensioni attuali o dai costi correnti. L'attività di revisione deve riflettere i potenziali benefici, come i costi correnti e quelli previsti.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

esegui un'analisi approfondita su tutti i componenti del carico di lavoro. Assicurati che il costo dell'analisi e il potenziale risparmio nel carico di lavoro durante il ciclo di vita siano in equilibrio. Devi individuare l'impatto attuale e il potenziale impatto futuro del componente. Ad esempio, se il costo della risorsa proposta è di 10 USD al mese e sotto i carichi previsti non supererà i 15 USD al mese, la spesa di una giornata di impegno per ridurre i costi del 50% (5 USD al mese) potrebbe superare il potenziale vantaggio per tutta la durata del sistema. L'utilizzo di una stima basata sui dati, più rapida ed efficiente, fornirà il migliore risultato complessivo per questo componente.

I carichi di lavoro possono cambiare nel corso del tempo e il giusto set di servizi potrebbe non essere ottimale se l'architettura o l'utilizzo del carico di lavoro cambiano. L'analisi per la selezione dei servizi deve integrare gli stati del carico di lavoro e i livelli di utilizzo attuali e futuri. Implementare un servizio in funzione dello stato o dell'utilizzo futuro del carico di lavoro può ridurre i costi complessivi, riducendo o rimuovendo lo sforzo necessario per apportare modifiche future.

AWS Cost Explorer e il AWS Cost and Usage Report (CUR) possono analizzare i costi di un proof of concept (PoC) o di un ambiente in esecuzione. Puoi anche utilizzare AWS Pricing Calculator per stimare i costi del carico di lavoro.

Passaggi dell'implementazione

- Elenca i componenti del carico di lavoro: Crea l'elenco di tutti i componenti del carico di lavoro. Tale elenco viene utilizzato come verifica per controllare che ogni componente sia stato analizzato. Gli sforzi sostenuti devono riflettere la criticità del carico di lavoro secondo quanto definito dalle priorità dell'organizzazione. Raggruppare le risorse in modo funzionale, ad esempio in base allo storage del database di produzione, migliora l'efficienza se sono presenti più database.
- Assegna le priorità all'elenco dei componenti: Prendi l'elenco dei componenti e assegna le priorità in ordine di sforzo. Tale assegnazione viene in genere eseguita in ordine dal componente più costoso a quello meno costoso o in base alla criticità definita dalle priorità dell'organizzazione.
- Esegui l'analisi: Per ogni componente dell'elenco, esamina le opzioni e i servizi disponibili e scegli l'opzione che si allinea meglio alle priorità dell'organizzazione.

Risorse

Documenti correlati:

- AWS Pricing Calculator
- AWS Cost Explorer
- Classi di archiviazione di Amazon S3
- Prodotti cloud

COST05-BP03 Esecuzione di un'analisi accurata di ciascun componente

Considera il costo complessivo per l'organizzazione di ogni componente. Considera il costo di proprietà totale tenendo conto dei costi operativi e di gestione, soprattutto quando si utilizzano i servizi gestiti. L'attività di revisione deve riflettere i potenziali benefici, ad esempio il tempo speso per l'analisi dovrebbe essere proporzionale al costo dei componenti.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Considera il risparmio in termini di tempo, che consentirà al tuo team di concentrarsi sull'eliminazione del debito tecnico, sull'innovazione e sulle funzionalità che offrono un valore aggiunto. Ad esempio, potresti avere bisogno di eseguire il rehosting (lift and shift) del tuo ambiente on-premise nel cloud il più rapidamente possibile ed eseguire l'ottimizzazione in un secondo momento. Vale la pena soffermarsi sul risparmio che puoi ottenere usando i servizi gestiti che rimuovono o riducono i costi di licenza. i servizi gestiti eliminano l'onere operativo e amministrativo legato alla manutenzione di un servizio, consentendoti di concentrarti sull'innovazione. Inoltre, poiché i servizi gestiti operano su scala cloud, possono offrire un costo inferiore per transazione o servizio.

Solitamente, i servizi gestiti presentano attributi che puoi impostare per garantire la capacità necessaria. Devi impostare e monitorare questi attributi in modo che la tua capacità in eccesso sia mantenuta al minimo e le prestazioni siano massimizzate. Puoi modificare gli attributi di AWS Managed Services utilizzando la AWS Management Console o le API e gli SDK AWS per allineare le risorse necessarie con le variazioni della domanda. Ad esempio, puoi aumentare o diminuire il numero di nodi di un cluster Amazon EMR (o di un cluster Amazon Redshift) per ridimensionarlo.

Puoi anche unire più istanze in una risorsa AWS per ottenere una densità di utilizzo più elevata. Ad esempio, puoi effettuare il provisioning di diversi database più piccoli su una singola istanza database Amazon Relational Database Service (Amazon RDS). Quando l'utilizzo si intensifica, puoi migrare uno dei database su un'istanza database Amazon RDS dedicata utilizzando un processo di generazione dello snapshot e ripristino.

Quando predisponi carichi di lavoro su servizi gestiti, devi comprendere i requisiti inerenti alla modifica della capacità del servizio. Tali requisiti solitamente riguardano il tempo, l'impegno e qualunque impatto sul normale funzionamento del carico di lavoro. La risorsa predisposta deve offrire il tempo necessario per l'applicazione delle modifiche, pertanto procurati i mezzi necessari a tal fine. L'impegno costante richiesto per modificare i servizi può essere ridotto praticamente a zero grazie alle API e agli SDK integrati nel sistema, nonché grazie a strumenti di monitoraggio come Amazon CloudWatch.

Amazon RDS, Amazon Redshifte Amazon ElastiCache forniscono un servizio di database gestito. Amazon Athena, Amazon EMRe Amazon OpenSearch Service forniscono un servizio di analisi gestito.

AMS è un servizio che gestisce l'infrastruttura AWS per conto di clienti e partner aziendali. Offre un ambiente sicuro e conforme su cui è possibile implementare i carichi di lavoro. AMS utilizza modelli operativi cloud aziendali dotati di automazione per consentirti di soddisfare i requisiti aziendali, di passare più rapidamente al cloud e di ridurre i costi di gestione correnti.

Passaggi dell'implementazione

- Esegui un'analisi accurata: Utilizzando l'elenco dei componenti, analizza ogni componente dalla priorità più alta alla priorità più bassa. Per la priorità più alta e i componenti più costosi, esegui analisi aggiuntive e valuta tutte le opzioni disponibili e il loro impatto a lungo termine. Per i componenti con priorità più bassa, valuta se le modifiche relative all'utilizzo hanno un impatto sulla priorità del componente, quindi esegui un'analisi dello sforzo appropriato.

Risorse

Documenti correlati:

- Calcolatore del costo totale di proprietà (TCO) di AWS
- Classi di archiviazione di Amazon S3
- Prodotti cloud

COST05-BP04 Selezione di software con licenze convenienti

Il software open source elimina i costi di licenza del software, che contribuiscono in modo significativo ai costi dei carichi di lavoro. Nei casi in cui il software con licenza sia obbligatorio, evita le licenze

legate ad attributi arbitrari, ad esempio CPU, e cerca le licenze legate all'output o ai risultati. Il costo di queste licenze si ridimensiona in base ai vantaggi che offrono.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Utilizzando software open source è possibile eliminare il costo delle licenze software. Con l'aumentare delle dimensioni del carico di lavoro, l'impatto sui costi può essere significativo. Misura i vantaggi di usare software con licenza in rapporto ai costi totali per assicurarti di disporre di un carico di lavoro che sia il più ottimizzato possibile. Crea modelli per le eventuali modifiche alla licenza e il relativo impatto sui costi del carico di lavoro. Se un fornitore modifica il costo della licenza del database, valuta come questo incide sull'efficienza complessiva del carico di lavoro. Effettua un'analisi dello storico dei prezzi dei tuoi fornitori per scoprire le tendenze dei cambiamenti relativi alle licenze dei loro prodotti. I costi delle licenze possono anche essere adattati indipendentemente dal throughput o dall'utilizzo, come nel caso delle licenze che si adattano in base all'hardware (licenze legate alle CPU). È necessario evitare queste licenze poiché i costi possono aumentare rapidamente senza che vi siano vantaggi correlati.

Passaggi dell'implementazione

- Analisi delle opzioni di licenza: Esamina i termini di licenza del software disponibile. Cerca le versioni open source che dispongono delle funzionalità necessarie e considera se i vantaggi del software con licenza superano i costi. Le condizioni vantaggiose allineano il costo del software ai vantaggi che offre.
- Analizza i fornitori del software: Esamina tutte le modifiche ai prezzi o alle licenze effettuate dal fornitore. Identifica eventuali modifiche non allineate ai risultati, ad esempio termini punitivi per l'esecuzione su hardware o piattaforme di fornitori specifici. Inoltre, verifica il modo in cui eseguono gli audit e le sanzioni a cui potresti andare incontro.

Risorse

Documenti correlati:

- Calcolatore del costo totale di proprietà (TCO) di AWS
- Classi di archiviazione di Amazon S3
- Prodotti cloud

COST05-BP05 Selezione dei componenti del carico di lavoro per ottimizzare i costi in linea con le priorità dell'organizzazione

Tieni in considerazione il costo nella selezione di tutti i componenti. Questo include l'utilizzo di servizi a livello applicativo e gestiti come Amazon Relational Database Service (Amazon RDS), Amazon DynamoDB, Amazon Simple Notification Service (Amazon SNS) e Amazon Simple Email Service (Amazon SES) per ridurre il costo complessivo dell'organizzazione. Utilizza funzioni serverless e container per l'elaborazione, come AWS Lambda, Amazon Simple Storage Service (Amazon S3) per i siti web statici e Amazon Elastic Container Service (Amazon ECS). Riduci al minimo i costi di licenza utilizzando software open source o software che non prevedono tariffe di licenza, come ad esempio Amazon Linux per carichi di lavoro di calcolo, oppure esegui la migrazione dei database su Amazon Aurora.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

puoi utilizzare servizi serverless o a livello di applicazione, ad esempio AWS Lambda, Amazon Simple Queue Service (Amazon SQS), Amazon SNSe Amazon SES. Questi servizi eliminano la necessità di gestire una risorsa e forniscono funzioni di esecuzione del codice, servizi di accodamento e consegna dei messaggi. L'altro vantaggio è che le prestazioni e i costi vengono adattati in base all'utilizzo, garantendo l'allocazione e l'attribuzione dei costi in modo efficiente.

Per ulteriori informazioni su Serverless, consulta il whitepaper su Serverless Application Lens Well-Architected.

Passaggi dell'implementazione

- Seleziona ciascun servizio per ottimizzare i costi: Utilizzando l'elenco e l'analisi prioritari, seleziona ogni opzione che fornisce la migliore corrispondenza con le priorità dell'organizzazione.

Risorse

Documenti correlati:

- Calcolatore del costo totale di proprietà (TCO) di AWS
- Classi di archiviazione di Amazon S3
- Prodotti cloud

COST05-BP06 Esecuzione di un'analisi dei costi per diversi valori di utilizzo nel tempo

I carichi di lavoro possono cambiare nel corso del tempo. Alcuni servizi o funzionalità sono più convenienti a diversi livelli di utilizzo. Eseguendo l'analisi su ogni componente nel tempo e all'utilizzo previsto, il carico di lavoro rimane conveniente per tutta la sua durata.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Quando AWS rilascia nuovi servizi e funzionalità, è possibile che i servizi ottimali per il carico di lavoro cambino. Tale cambiamento comporta un impegno, che dovrebbe essere commensurato ai vantaggi potenziali. La frequenza di revisione del carico di lavoro dipende dai requisiti dell'organizzazione. Se si tratta di un carico di lavoro con costi importanti, una rapida implementazione dei nuovi servizi massimizzerà i risparmi sui costi, e in tal caso una revisione più frequente può risultare vantaggiosa. Un altro stimolo importante per la revisione è il cambiamento dei modelli di utilizzo. Se si verificassero notevoli cambiamenti nell'utilizzo, ciò potrebbe indicare un maggiore vantaggio dei servizi alternativi. Ad esempio, per velocità di trasferimento dei dati più elevate, un servizio di connessione diretta può risultare più economico di una VPN e garantire la connettività richiesta. Prevedi il potenziale impatto dei cambiamenti del servizio, così da monitorare le variazioni a livello di utilizzo e implementare con anticipo i servizi più convenienti.

Passaggi dell'implementazione

- Definisci i modelli di utilizzo previsti: Collaborando con la tua organizzazione, ad esempio con i proprietari di prodotti e marketing, documenta quali sono i modelli di utilizzo previsti per il carico di lavoro.
- Esegui l'analisi dei costi in base all'utilizzo previsto: Utilizzando i modelli di utilizzo definiti, esegui l'analisi in ciascuno di questi punti. Lo sforzo di analisi dovrebbe riflettere il potenziale risultato. Ad esempio, se la variazione dell'utilizzo è elevata, è necessario eseguire un'analisi accurata per verificare eventuali costi e modifiche.

Risorse

Documenti correlati:

- Calcolatore del costo totale di proprietà (TCO) di AWS
- Classi di archiviazione di Amazon S3

- Prodotti cloud

COST 6 In che modo raggiungi gli obiettivi di costo quando selezioni il tipo, le dimensioni e il numero delle risorse?

Assicurati di scegliere la dimensione e il numero delle risorse appropriati per l'attività in questione. Selezionando il tipo, le dimensioni e il numero più convenienti, riduci al minimo gli sprechi.

Best practice

- COST06-BP01 Esecuzione della modellazione dei costi
- COST06-BP02 Selezione di tipo, dimensione e numero di risorse sulla base dei dati
- COST06-BP03 Selezione automatica del tipo e della dimensione della risorsa in base ai parametri

COST06-BP01 Esecuzione della modellazione dei costi

Identifica i requisiti dell'organizzazione ed esegui la modellizzazione dei costi del carico di lavoro e di ciascuno dei suoi componenti. Esegui attività di analisi comparativa per il carico di lavoro in base ai diversi carichi previsti e confronta i costi. L'impegno di modellazione deve riflettere il potenziale risultato. Ad esempio, il tempo speso dovrebbe essere proporzionale al costo dei componenti.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

esegui la modellazione dei costi per il tuo carico di lavoro e ciascuno dei suoi componenti per stabilire il giusto equilibrio tra le risorse e la dimensione appropriata per ogni risorsa sulla base di un determinato livello di prestazioni. Esegui attività di analisi comparativa per il carico di lavoro in base ai diversi carichi previsti e confronta i costi. L'impegno di modellazione deve riflettere il potenziale risultato. Ad esempio, il tempo speso dovrebbe essere proporzionale al costo dei componenti o ai risparmi previsti. Per le best practice fai riferimento alla sezione Revisione del Whitepaper sul principio dell'efficienza delle prestazioni.

AWS Compute Optimizer è in grado di supportare la modellazione dei costi per l'esecuzione di carichi di lavoro. Fornisce consigli di dimensionamento appropriato per le risorse di calcolo in base a una valutazione cronologica dell'utilizzo. Questa è la fonte di dati ideale per le risorse di calcolo perché è un servizio gratuito e utilizza il machine learning per generare più raccomandazioni a seconda dei livelli di rischio. Puoi anche utilizzare Amazon CloudWatch e Amazon CloudWatch Logs con log

personalizzati come fonti di dati per operazioni di dimensionamento appropriato per altri servizi e componenti del carico di lavoro.

Di seguito sono riportate le raccomandazioni per i parametri e i dati di modellazione dei costi:

- Il monitoraggio deve corrispondere in modo preciso all'esperienza degli utenti finali. Seleziona la granularità corretta per un dato periodo di tempo e scegli in modo ponderato il 99° percentile o quello massimo invece del valore medio.
- Seleziona la granularità corretta per il periodo di analisi richiesto per coprire tutti i cicli del carico di lavoro. Ad esempio, se esegui un'analisi di due settimane, potresti ignorare un ciclo mensile di utilizzo elevato, e questo potrebbe causare un provisioning insufficiente.

Passaggi dell'implementazione

- Esegui la modellazione dei costi: Implementa il carico di lavoro o una proof of concept in un account separato con i tipi di risorse e le dimensioni specifiche da testare. Esegui il carico di lavoro con i dati di test e registra i risultati di output, insieme ai dati relativi ai costi per il tempo in cui è stato eseguito il test. Quindi, reimplementa il carico di lavoro o modifica i tipi e le dimensioni delle risorse ed esegui nuovamente il test.

Risorse

Documenti correlati:

- AWS Auto Scaling
- Funzionalità di Amazon CloudWatch
- Ottimizzazione dei costi: dimensionamento appropriato di Amazon EC2
- AWS Compute Optimizer

COST06-BP02 Selezione di tipo, dimensione e numero di risorse sulla base dei dati

Seleziona la dimensione o il tipo di risorsa in base ai dati relativi al carico di lavoro e alle caratteristiche delle risorse come, ad esempio, elaborazione, memoria, throughput o scrittura intensiva. Questa selezione è tipicamente effettuata utilizzando una versione precedente (on-premise) del carico di lavoro, utilizzando la documentazione o altre fonti di informazione sul carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

Seleziona la dimensione o il tipo di risorsa in base al carico di lavoro e alle caratteristiche delle risorse come, ad esempio, elaborazione, memoria, throughput o scrittura intensiva. Questa selezione è tipicamente effettuata ricorrendo alla modellazione dei costi, utilizzando una versione precedente del carico di lavoro (ad esempio una versione on-premise), utilizzando la documentazione o altre fonti di informazione sul carico di lavoro (come whitepaper e soluzioni pubblicate).

Passaggi dell'implementazione

- Seleziona le risorse in base ai dati: Utilizzando i dati di modellazione dei costi, seleziona il livello di utilizzo previsto del carico di lavoro, quindi seleziona il tipo e le dimensioni della risorsa specificata.

Risorse

Documenti correlati:

- AWS Auto Scaling
- Funzionalità di Amazon CloudWatch
- Cost Optimization: EC2 Right Sizing

COST06-BP03 Selezione automatica del tipo e della dimensione della risorsa in base ai parametri

Utilizza i parametri del carico di lavoro in esecuzione per selezionare la dimensione e il tipo giusti per ottimizzare i costi. Esegui il provisioning in modo appropriato di throughput, dimensione e spazio di archiviazione per servizi quali Amazon Elastic Compute Cloud (Amazon EC2), Amazon DynamoDB, Amazon Elastic Block Store (Amazon EBS) (PIOPS), Amazon Relational Database Service (Amazon RDS), Amazon EMR e reti. Questa operazione può essere eseguita con un loop di feedback, ad esempio l'auto scaling o tramite codice personalizzato nel carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

crea un ciclo di feedback all'interno del carico di lavoro che utilizza i parametri attivi del carico di lavoro in esecuzione per apportare modifiche a tale carico di lavoro. Puoi utilizzare un servizio gestito, ad esempio AWS Auto Scaling, e configurarlo in modo che esegua per tuo conto le operazioni di dimensionamento corrette. AWS offre anche API, SDKe funzionalità che consentono alle risorse di essere modificate con il minimo sforzo. Puoi programmare un carico di lavoro per arrestare e

avviare un'istanza Amazon Elastic Compute Cloud (Amazon EC2) per consentire una modifica delle dimensioni o del tipo di istanza. Ciò offre i vantaggi del dimensionamento appropriato, eliminando al contempo quasi tutti i costi operativi necessari per apportare la modifica.

Alcuni servizi AWS integrano la selezione automatica del tipo o della dimensione, ad esempio Amazon Simple Storage Service(Amazon S3) Intelligent-Tiering. Basandosi sui modelli di utilizzo, Amazon S3 Intelligent-Tiering sposta automaticamente i dati tra due livelli di accesso: frequente e poco frequente.

Passaggi dell'implementazione

- Configura i parametri del carico di lavoro: Assicurati di acquisire i parametri chiave per il carico di lavoro. Questi parametri forniscono un'indicazione dell'esperienza del cliente, ad esempio l'output del carico di lavoro, e si allineano alle differenze tra tipi e dimensioni di risorse, come l'utilizzo di CPU e memoria.
- Visualizza le raccomandazioni per il dimensionamento appropriato: Utilizza le raccomandazioni per il dimensionamento appropriato in AWS Compute Optimizer per ottimizzare il carico di lavoro.
- Seleziona automaticamente il tipo e la dimensione della risorsa in base ai parametri: Utilizzando i parametri del carico di lavoro, seleziona manualmente o automaticamente le risorse del carico di lavoro. La configurazione di AWS Auto Scaling o l'implementazione di codice all'interno dell'applicazione può ridurre lo sforzo necessario in caso di modifiche frequenti e permettere di implementare potenzialmente eventuali modifiche più velocemente rispetto a un processo manuale.

Risorse

Documenti correlati:

- AWS Auto Scaling
- AWS Compute Optimizer
- Funzionalità di Amazon CloudWatch
- Configurazione di CloudWatch
- Pubblicazione di parametri personalizzati di CloudWatch
- Ottimizzazione dei costi: dimensionamento appropriato di Amazon EC2
- Nozioni di base su Amazon EC2 Auto Scaling
- Amazon S3 Intelligent-Tiering
- Avvia un'istanza EC2 utilizzando l'SDK

COST 7 In che modo impieghi i modelli di prezzo per ridurre i costi?

Utilizza il modello di prezzo più appropriato per le tue risorse per ridurre al minimo le spese.

Best practice

- COST07-BP01 Esecuzione di un'analisi del modello di prezzo
- COST07-BP02 Implementazione delle regioni in base al costo
- COST07-BP03 Selezione di contratti di terze parti con condizioni economicamente convenienti
- COST07-BP04 Implementazione di modelli di determinazione dei prezzi per tutti i componenti del carico di lavoro
- COST07-BP05 Esecuzione dell'analisi del modello di prezzo a livello di account principale

COST07-BP01 Esecuzione di un'analisi del modello di prezzo

Analizza ogni componente del carico di lavoro. Determina se il componente e le risorse saranno in esecuzione per periodi prolungati (per sconti a fronte di impegni) o dinamici e di breve durata (per istanze spot oppure on demand). Esegui un'analisi del carico di lavoro utilizzando la funzione di raccomandazione di AWS Cost Explorer.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

AWS offre diversi modelli di prezzo che ti consentono di pagare per le risorse nel modo più conveniente e adatto alle esigenze della tua organizzazione.

Passaggi dell'implementazione

- Esegui un'analisi degli sconti a fronte di impegni: Utilizzando Cost Explorer nel tuo account, esamina Savings Plans e le raccomandazioni relative alle istanze riservate. Per verificare di implementare le raccomandazioni corrette con gli sconti e i rischi richiesti, segui i corsi Well-Architected Labs.
- Analizza l'elasticità del carico di lavoro: Utilizza la granularità oraria in Cost Explorer o un pannello di controllo personalizzato. Analizza l'elasticità del carico di lavoro. Vai alla ricerca di modifiche regolari del numero di istanze in esecuzione. Le istanze in esecuzione per brevi periodi di tempo sono candiate per essere istanze Spot o serie di istanze Spot.
  - Well-Architected Lab: Cost Explorer

- Well-Architected Labs: visualizzazione dei costi

Risorse

Documenti correlati:

- Accesso alle raccomandazioni di istanza riservata
- Opzioni di acquisto dell'istanza

Video correlati:

- Risparmia fino al 90% ed esegui i carichi di lavoro di produzione su Spot

Esempi correlati:

- Well-Architected Lab: Cost Explorer
- Well-Architected Labs: visualizzazione dei costi
- Well-Architected Lab: modelli di prezzo

COST07-BP02 Implementazione delle regioni in base al costo

La determinazione dei prezzi delle risorse può essere diversa in ciascuna regione. La valutazione del costo specifico della regione garantisce il pagamento del prezzo complessivo più basso per questo carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

Quando progetti le tue soluzioni, una best practice da seguire è quella di cercare di posizionare le risorse di calcolo vicino agli utenti per offrire una latenza inferiore e una forte sovranità dei dati. Per i gruppi di pubblico globali dovresti usare più ubicazioni al fine di soddisfare queste esigenze. Dovresti selezionare la posizione geografica che ti consente di ridurre al minimo i costi.

L'infrastruttura Cloud AWS è basata su Regioni e zone di disponibilità. Una regione è un'area fisica del mondo in cui si trovano diverse zone di disponibilità. Le zone di disponibilità sono composte da uno o più data center singoli provvisti di alimentazione, rete e connettività ridondanti, ognuno in una struttura separata.

Ciascuna regione AWS opera all'interno di condizioni di mercato locali e la determinazione dei prezzi delle risorse è diversa in ciascuna regione. Scegli una regione specifica per gestire un componente o tutta la tua soluzione in modo da eseguirla al minor prezzo possibile a livello globale. Puoi utilizzare il AWS Pricing Calculator per stimare il costo del carico di lavoro in diverse regioni.

Passaggi dell'implementazione

- Verifica dei prezzi delle regioni: Analizza i costi del carico di lavoro nella regione corrente. A partire dai costi più elevati per servizio e tipo di utilizzo, calcola i costi in altre regioni disponibili. Se il risparmio previsto supera il costo di spostamento del componente o del carico di lavoro, esegui la migrazione alla nuova regione.

Risorse

Documenti correlati:

- Accesso alle raccomandazioni di istanza riservata
- Prezzi di Amazon EC2
- Opzioni di acquisto dell'istanza
- Tabella delle regioni

Video correlati:

- Risparmia fino al 90% ed esegui i carichi di lavoro di produzione su Spot

COST07-BP03 Selezione di contratti di terze parti con condizioni economicamente convenienti

Gli accordi e i termini convenienti assicurano che i costi di questi servizi siano ridimensionati in base ai vantaggi che offrono. Seleziona gli accordi e i prezzi che si ridimensionano quando forniscono ulteriori vantaggi alla tua organizzazione.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

quando utilizzi soluzioni o servizi di terze parti nel cloud, è importante che le strutture dei prezzi siano allineate ai risultati dell'ottimizzazione dei costi. I prezzi devono essere adattati in base ai risultati e al valore che forniscono. Un esempio di questo è un software che contempla una percentuale del risparmio che fornisce, più risparmi (come risultato) e più ti addebita. Gli accordi che si adeguano in

base alla fattura in genere non sono allineati all'ottimizzazione dei costi, a meno che non forniscano risultati per ogni parte della fattura specifica. Ad esempio, una soluzione che fornisce suggerimenti per Amazon Elastic Compute Cloud (Amazon EC2) e addebita una percentuale dell'intera fattura aumenterà se utilizzi altri servizi per i quali non fornisce alcun vantaggio. Un altro esempio è un servizio gestito che viene addebitato a una percentuale del costo delle risorse che vengono gestite. Una dimensione di istanza più grande potrebbe non richiedere necessariamente un maggiore impegno di gestione, ma comporterà un addebito superiore. Assicurati che queste disposizioni tariffarie dei servizi includano un programma di ottimizzazione dei costi o funzionalità di servizio volte a migliorare l'efficienza.

Passaggi dell'implementazione

- Analizza i contratti e le condizioni stabilite con le terze parti: Esamina i prezzi nei contratti di terze parti. Esegui la modellazione per diversi livelli di utilizzo e considera i nuovi costi, come il nuovo utilizzo del servizio o aumenti dei servizi attuali a causa della crescita del carico di lavoro. Decidi se i costi aggiuntivi forniscono i vantaggi necessari alla tua azienda.

Risorse

Documenti correlati:

- Accesso alle raccomandazioni di istanza riservata
- Opzioni di acquisto dell'istanza

Video correlati:

- Risparmia fino al 90% ed esegui i carichi di lavoro di produzione su Spot

COST07-BP04 Implementazione di modelli di determinazione dei prezzi per tutti i componenti del carico di lavoro

Le risorse in esecuzione in modo permanente devono utilizzare la capacità riservata, ad esempio Savings Plans o istanze riservate. La capacità a breve termine è configurata per usare le istanze Spot o la serie di istanze Spot. Le istanze on demand vengono utilizzate solo per carichi di lavoro a breve termine che non possono essere interrotti e che non durano abbastanza a lungo per la capacità riservata, tra il 25% e il 75% del periodo, a seconda del tipo di risorsa.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

considera i requisiti dei componenti del carico di lavoro e comprendi i potenziali modelli di prezzo. Definisci il requisito di disponibilità del componente. Determina se ci sono più risorse indipendenti che eseguono la funzione nel carico di lavoro e quali sono i requisiti del carico di lavoro nel corso del tempo. Confronta il costo delle risorse utilizzando il modello di prezzo on demand predefinito e altri modelli applicabili. Tieni conto di qualsiasi potenziale cambiamento nelle risorse o nei componenti del carico di lavoro.

Passaggi dell'implementazione

- Implementa i modelli di prezzo: Utilizzando i risultati dell'analisi acquista Savings Plans (SP), istanze riservate (RI) o implementa istanze Spot. Se è il tuo primo acquisto di istanze riservate (RI), scegli le prime 5 o 10 raccomandazioni nell'elenco, quindi monitora e analizza i risultati per il mese successivo o per i due mesi successivi. Acquista piccoli numeri di sconti a fronte di impegni a cicli regolari, ad esempio ogni due settimane o ogni mese. Implementa istanze Spot per carichi di lavoro che possono essere interrotti o che sono stateless.
- Ciclo di revisione del carico di lavoro: Implementa un ciclo di revisione per il carico di lavoro che analizzi in modo specifico la copertura del modello di prezzo. Quando il carico di lavoro ha la copertura necessaria, acquista ulteriori sconti a fronte di impegni ogni 2-4 settimane o al variare dell'utilizzo dell'organizzazione.

Risorse

Documenti correlati:

- Accesso alle raccomandazioni di istanza riservata
- Parco istanze EC2
- Come acquistare istanze riservate
- Opzioni di acquisto dell'istanza
- Istanze Spot

Video correlati:

- Risparmia fino al 90% ed esegui i carichi di lavoro di produzione su Spot

COST07-BP05 Esecuzione dell'analisi del modello di prezzo a livello di account principale

Utilizza Cost Explorer Savings Plans e le raccomandazioni sulle istanze riservate per eseguire analisi periodiche a livello di account di gestione per ottenere sconti a fronte di impegni.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

l'esecuzione della modellazione dei costi a intervalli regolari garantisce l'implementazione di opportunità di ottimizzazione su più carichi di lavoro. Ad esempio, se più carichi di lavoro utilizzano istanze on demand a livello aggregato, il rischio di modifica è inferiore e l'implementazione di uno sconto a fronte di impegni permetterà di raggiungere un costo complessivo inferiore. Si consiglia di eseguire l'analisi seguendo cicli regolari con una periodicità da due settimane a un mese. In questo modo è possibile effettuare acquisti in piccoli incrementi, così che la copertura dei modelli di prezzo evolva di pari passo con i carichi di lavoro e i relativi componenti.

Utilizza lo strumento per i suggerimenti AWS Cost Explorer per trovare opportunità di sconti a fronte di impegni.

Per trovare opportunità per i carichi di lavoro Spot, utilizza una visualizzazione oraria dell'utilizzo complessivo e cerca periodi regolari di variazione dell'utilizzo o di elasticità.

Passaggi dell'implementazione

- Esegui un'analisi degli sconti a fronte di impegni: Utilizzando Cost Explorer nel tuo account, esamina Savings Plans e le raccomandazioni relative alle istanze riservate. Per verificare di implementare le raccomandazioni corrette con gli sconti e i rischi richiesti, segui i Well-Architected Labs.

Risorse

Documenti correlati:

- Accesso alle raccomandazioni di istanza riservata
- Opzioni di acquisto dell'istanza

Video correlati:

- Risparmia fino al 90% ed esegui i carichi di lavoro di produzione su Spot

Esempi correlati:

- Well-Architected Lab: modelli di prezzo

COST 8 In che modo pianifichi i costi per il trasferimento dei dati?

Assicurati di pianificare e monitorare i costi di trasferimento dei dati in modo da poter prendere decisioni sull'architettura per ridurre al minimo i costi. Una modifica piccola ma efficace dell'architettura può ridurre drasticamente i costi operativi nel tempo.

Best practice

- COST08-BP01 Esecuzione della modellazione del trasferimento dei dati
- COST08-BP02 Selezione dei componenti per ottimizzare il costo di trasferimento dei dati
- COST08-BP03 Implementazione dei servizi per ridurre il costo di trasferimento dei dati

COST08-BP01 Esecuzione della modellazione del trasferimento dei dati

Raccogli i requisiti dell'organizzazione ed esegui la modellizzazione del trasferimento dei dati del carico di lavoro e di ciascuno dei suoi componenti. Questo identifica il punto di costo più basso per le sue attuali esigenze di trasferimento dei dati.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

comprendi dove avviene il trasferimento dei dati nel carico di lavoro, il costo del trasferimento e il relativo vantaggio. In questo modo puoi prendere una decisione consapevole quando si tratta di modificare o accettare una decisione relativa all'architettura. Ad esempio, potresti disporre di una configurazione con più zone di disponibilità dove replichi i dati tra le varie zone di disponibilità. Puoi modellare il costo della struttura e decidere che questo sia un costo accettabile (simile a quello del calcolo e dello storage in entrambe le zone di disponibilità) per ottenere l'affidabilità e la resilienza richieste.

Modella i costi in base a livelli differenti di utilizzo. L'utilizzo del carico di lavoro può cambiare nel corso del tempo e servizi differenti possono risultare più convenienti a livelli differenti.

Utilizza AWS Cost Explorer o AWS Cost and Usage Report (CUR) per comprendere e modellare i costi di trasferimento dei dati. Configura un proof of concept (PoC) o testa il carico di lavoro ed esegui

un test con un carico simulato realistico. Puoi modellare i costi in base alle diverse esigenze di carico di lavoro.

Passaggi dell'implementazione

- Calcola i costi di trasferimento dei dati: Utilizza le pagine dei prezzi di AWS e calcola i costi di trasferimento dei dati per il carico di lavoro. Calcola i costi di trasferimento dei dati a diversi livelli di utilizzo, ipotizzando incrementi e riduzioni dell'utilizzo del carico di lavoro. Nei casi in cui sono disponibili più opzioni per l'architettura del carico di lavoro valuta i costi di ogni opzione per il confronto.
- Collega i costi ai risultati: Per ogni costo di trasferimento dei dati sostenuto, specifica il risultato ottenuto per il carico di lavoro. Se si tratta di un trasferimento tra componenti potrebbe trattarsi di una necessità di disaccoppiamento, se si tratta di un trasferimento tra zone di disponibilità potrebbe trattarsi di una necessità di ridondanza.

Risorse

Documenti correlati:

- Soluzioni di memorizzazione nella cache AWS
- Prezzi di AWS
- Prezzi di Amazon EC2
- Prezzi di Amazon VPC
- Distribuisci contenuti più rapidamente con Amazon CloudFront

COST08-BP02 Selezione dei componenti per ottimizzare il costo di trasferimento dei dati

Tutti i componenti sono selezionati e l'architettura è progettata per ridurre i costi di trasferimento dei dati. Questo include l'utilizzo di componenti come l'ottimizzazione WAN e le configurazioni Multi-AZ

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Una progettazione basata sul trasferimento dei dati ti assicura la massima riduzione dei costi di trasferimento dei dati. Potrebbe implicare l'uso di reti di distribuzione di contenuti per posizionare i dati vicino agli utenti, oppure l'uso di collegamenti di rete dedicati dalle tue sedi ad AWS. Puoi anche

utilizzare l'ottimizzazione WAN e l'ottimizzazione delle applicazioni per ridurre la quantità di dati trasferiti tra i componenti.

Passaggi dell'implementazione

- Seleziona i componenti per il trasferimento dei dati: Utilizzando la modellazione per il trasferimento dei dati, concentrati su dove si trovano i costi di trasferimento dei dati più elevati o dove sarebbero se l'utilizzo del carico di lavoro cambiasse. Individua architetture alternative o componenti aggiuntivi che eliminano o riducono la necessità di trasferimento dei dati o ne riducono i costi.

Risorse

Documenti correlati:

- Soluzioni di memorizzazione nella cache AWS
- Distribuisci contenuti più rapidamente con Amazon CloudFront

COST08-BP03 Implementazione dei servizi per ridurre il costo di trasferimento dei dati

Implementa i servizi per ridurre il costo di trasferimento dei dati, ad esempio utilizzando una rete di distribuzione di contenuti (CDN) come Amazon CloudFront per distribuire i contenuti agli utenti finali, gestendo livelli di cache utilizzando Amazon ElastiCache o sfruttando AWS Direct Connect invece della VPN per la connettività verso AWS.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Amazon CloudFront è una rete globale di distribuzione di contenuti che trasferisce i dati con una latenza ridotta e una velocità di trasferimento elevata. Cattura i dati nelle edge location di tutto il mondo, riducendo così il carico sulle tue risorse. Utilizzando CloudFront puoi ridurre l'impegno amministrativo legato alla distribuzione dei contenuti per numeri elevati di utenti a livello globale, con una latenza minima.

AWS Direct Connect ti consente di creare una connessione di rete dedicata ad AWS. In questo modo puoi ridurre i costi di rete, aumentare la larghezza di banda e offrire un'esperienza di rete più costante rispetto alle connessioni Internet.

AWS VPN consente di stabilire una connessione sicura e privata tra la rete privata e la rete globale AWS. È ideale per piccoli uffici o partner aziendali perché offre una connettività semplice e rapida, ed è un servizio completamente gestito ed elastico.

Endpoint VPC consentono la connettività tra i servizi AWS su reti private e possono essere utilizzati per ridurre i costi di trasferimento di dati pubblici e dei costi dei gateway NAT . Endpoint VPC gateway non prevedono tariffe orarie e supportano Amazon Simple Storage Service (Amazon S3) e Amazon DynamoDB. Gli endpoint VPC dell'interfaccia sono forniti da AWS PrivateLink e prevedono una tariffa oraria e un costo di utilizzo per GB.

Passaggi dell'implementazione

- Implementa i servizi: Utilizzando la modellazione del trasferimento dati, osserva dove si trovano i costi più elevati e i flussi di volume più elevati. Esamina i servizi AWS e valuta se esiste un servizio che riduce o rimuove il trasferimento, in particolare nell'ambito delle reti e della distribuzione di contenuti. Individua anche servizi di cache in cui si verifica un accesso ripetuto ai dati o in cui sono presenti grandi quantità di dati.

Risorse

Documenti correlati:

- AWS Direct Connect
- Esplora i prodotti AWS
- Soluzioni di memorizzazione nella cache AWS
- Amazon CloudFront
- Distribuisci contenuti più rapidamente con Amazon CloudFront

Gestione delle risorse di domanda e offerta

Domanda

- COST 9 Come gestisci la domanda e fornisci le risorse?

COST 9 Come gestisci la domanda e fornisci le risorse?

Per avere un carico di lavoro con costo e prestazioni bilanciate, assicurati che venga utilizzato tutto ciò per cui paghi ed evita le istanze molto sottoutilizzate. Un parametro di utilizzo distorto, in qualsiasi

delle suddette direzioni, ha un impatto negativo sull'organizzazione, sia per i costi operativi (basse prestazioni a causa di un utilizzo eccessivo) che per le spese inerenti a AWS sprecate (a causa di un provisioning eccessivo).

Best practice

- COST09-BP01 Analisi della domanda del carico di lavoro
- COST09-BP02 Implementazione di un buffer o del throttling per gestire la domanda
- COST09-BP03 Fornitura dinamica delle risorse

COST09-BP01 Analisi della domanda del carico di lavoro

Analizza la domanda del carico di lavoro nel tempo. Verifica che l'analisi copra l'andamento stagionale e rappresenti accuratamente le condizioni operative per l'intera durata del carico di lavoro. L'attività di analisi deve riflettere i potenziali benefici, ad esempio che il tempo speso sia proporzionale al costo del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

è importante conoscere i requisiti del carico di lavoro. I requisiti dell'organizzazione devono indicare i tempi di risposta del carico di lavoro per le richieste. Il tempo di risposta può essere utilizzato per determinare se la domanda è gestita o se l'offerta di risorse cambierà per soddisfare la domanda.

L'analisi deve includere la prevedibilità e la ripetibilità della domanda, la velocità di variazione della domanda e la quantità di variazione della domanda. Assicurati che l'analisi venga eseguita per un periodo sufficientemente lungo da incorporare qualsiasi variazione stagionale, ad esempio l'elaborazione di fine mese o i picchi legati alle festività.

Assicurati che le attività di analisi siano commensurate ai potenziali vantaggi dell'implementazione del dimensionamento. Osserva il costo totale previsto del componente, ed eventuali aumenti o riduzioni di utilizzo e costi durante il ciclo di vita del carico di lavoro.

Puoi utilizzare AWS Cost Explorer oppure Amazon QuickSight con AWS Cost and Usage Report (CUR) o i log dell'applicazione per eseguire un'analisi visiva della domanda del carico di lavoro.

Passaggi dell'implementazione

- Analizza i dati del carico di lavoro esistenti: Analizza i dati provenienti dal carico di lavoro esistente, dalle versioni precedenti del carico di lavoro o dai modelli di utilizzo previsti. Utilizza i file di log

e i dati di monitoraggio per ottenere informazioni su come i clienti utilizzano il carico di lavoro. I parametri tipici sono la domanda effettiva in termini di richieste al secondo, le volte in cui il tasso delle richieste cambia o si trova a livelli diversi e il tasso di variazione delle richieste. Assicurati di analizzare un ciclo completo del carico di lavoro, raccogliendo dati per eventuali variazioni stagionali, ad esempio eventi di fine mese o di fine anno. L'attività che emerge dall'analisi deve riflettere le caratteristiche del carico di lavoro. L'impegno maggiore dovrebbe riguardare i carichi di lavoro di alto valore che presentano le maggiori variazioni della domanda. Il minimo impegno dovrebbe riguardare carichi di lavoro di basso valore che hanno variazioni minime nella domanda. I parametri più comuni per il valore sono rischio, conoscenza del marchio, ricavi o costi del carico di lavoro.

- Esegui previsioni dell'influenza dei fattori esterni: Incontra i membri del team di tutta l'organizzazione che possono influenzare o modificare la domanda del carico di lavoro. I team più comuni sono le vendite, il marketing o il business development. Collabora con loro per conoscere i cicli secondo cui operano e se ci sono eventi che potrebbero modificare la domanda del carico di lavoro. Prevedi la richiesta del carico di lavoro con questi dati.

Risorse

Documenti correlati:

- AWS Auto Scaling
- AWS Instance Scheduler
- Nozioni di base su Amazon SQS
- AWS Cost Explorer
- Amazon QuickSight

COST09-BP02 Implementazione di un buffer o del throttling per gestire la domanda

Buffering e throttling modificano la domanda sul carico di lavoro, attenuando eventuali picchi. Implementa il throttling quando i client eseguono nuovi tentativi. Implementa il buffering per archiviare la richiesta e rinviare l'elaborazione a un secondo momento. Verifica che le esecuzioni di throttling e buffering siano progettate in modo che i client ricevano una risposta nel tempo richiesto.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Throttling: se l'origine della richiesta dispone di funzionalità di ripetizione dei tentativi, è possibile implementare il throttling. Il throttling indica alla sorgente che, se non è in grado di soddisfare la richiesta all'ora corrente, dovrebbe riprovare più tardi. La sorgente attenderà per un determinato periodo di tempo e quindi riproverà a effettuare la richiesta. L'implementazione del throttling ha il vantaggio di limitare la quantità massima di risorse e i costi del carico di lavoro. In AWS puoi utilizzare Amazon API Gateway per implementare il throttling. Consulta il whitepaper sul principio dell'affidabilità secondo il Canone di architettura per ulteriori dettagli sull'implementazione del throttling.

Basato su buffer: analogamente al throttling, il buffering rinvia l'elaborazione delle richieste, consentendo alle applicazioni eseguite a velocità diverse di comunicare in modo efficace. Un approccio basato sul buffering impiega una coda per l'accettazione dei messaggi (unità di lavoro) dai produttori. I messaggi vengono letti ed elaborati dai consumatori e ciò consente ai messaggi di essere eseguiti alla velocità che soddisfa i requisiti aziendali del consumatore stesso. Non devi preoccuparti del fatto che i produttori debbano gestire i problemi legati al throttling, come la durabilità e la contropressione dei dati (per cui i produttori rallentano per adeguarsi alla velocità dei consumatori).

Su AWS puoi scegliere fra più servizi per l'implementazione di una strategia di buffering. Amazon Simple Queue Service(Amazon SQS) un servizio gestito che offre code che consentono a un singolo consumatore di leggere singoli messaggi. Amazon Kinesis offre un flusso che consente a più consumatori di leggere gli stessi messaggi.

Durante la progettazione con un approccio basato sul buffering, assicurati di progettare il carico di lavoro per soddisfare la richiesta nel tempo richiesto e di essere in grado di gestire le richieste duplicate per il lavoro.

Passaggi dell'implementazione

- Analizza i requisiti del client: Analizza le richieste del client per determinare se sono in grado di eseguire nuovi tentativi. Per i client che non possono eseguire nuovi tentativi, è necessario implementare i buffer. Analizza la domanda complessiva, la velocità di modifica e il tempo di risposta richiesto per determinare le dimensioni del throttling o del buffer richiesto.
- Implementa un buffer o il throttling: Implementa un buffer o un throttling nel carico di lavoro. Una coda come Amazon Simple Queue Service (Amazon SQS) può offrire un buffer ai componenti del carico di lavoro. Amazon API Gateway può fornire una funzionalità di throttling ai componenti del carico di lavoro.

Risorse

Documenti correlati:

- AWS Auto Scaling
- AWS Instance Scheduler
- Amazon API Gateway
- Amazon Simple Queue Service
- Nozioni di base su Amazon SQS
- Amazon Kinesis

COST09-BP03 Fornitura dinamica delle risorse

Le risorse sono fornite in modo pianificato. La pianificazione può essere basata sulla domanda, ad esempio tramite l'auto scaling, oppure sul tempo, quando la domanda è prevedibile e le risorse sono fornite in base al tempo. Questi metodi comportano il minor numero possibile di sovra o sotto-provisioning.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Puoi utilizzare AWS Auto Scalingo incorporare il dimensionamento nel codice con API o SDK AWS. Ciò riduce i costi complessivi del carico di lavoro rimuovendo i costi operativi dall'apportare manualmente modifiche al tuo ambiente e può essere eseguito molto più rapidamente. In questo modo è possibile garantire che le risorse del carico di lavoro soddisfino al meglio la domanda, in qualsiasi momento.

Fornitura basata sulla domanda: sfrutta l'elasticità del cloud per fornire risorse in grado di soddisfare le mutevoli esigenze. Sfrutta API o funzionalità dei servizi per modificare in modo programmatico e dinamico la quantità di risorse del cloud nella tua architettura. Ciò ti consente di dimensionare i componenti nella tua architettura e aumentare automaticamente il numero di risorse durante i picchi di domanda per mantenere le prestazioni, nonché diminuire la capacità quando la domanda cala in modo da ridurre i costi.

AWS Auto Scaling ti aiuta a regolare la capacità per mantenere prestazioni stabili e prevedibili al minor costo possibile. Si tratta di un servizio completamente gestito e gratuito che si integra con

istanze e serie di istanze Spot Amazon Elastic Compute Cloud (Amazon EC2), Amazon Elastic Container Service (Amazon ECS), Amazon DynamoDB e Amazon Aurora.

Auto Scaling fornisce il rilevamento automatico delle risorse per aiutare a trovare risorse nel carico di lavoro che possono essere configurate, dispone di strategie di dimensionamento integrate per ottimizzare le prestazioni, i costi o un equilibrio tra i due e fornisce il dimensionamento predittivo per aiutare a risolvere i picchi ricorrenti con regolarità.

Auto Scaling può implementare il ridimensionamento manuale, programmato o in base alla domanda. Puoi anche utilizzare le metriche e gli allarmi di Amazon CloudWatch per attivare eventi di dimensionamento per il tuo carico di lavoro. I parametri tipici possono essere parametri standard di Amazon EC2, ad esempio l'utilizzo della CPU, il throughput di rete e la latenza di richiesta/risposta osservata da Elastic Load Balancing(ELB) . Quando possibile, è consigliabile utilizzare un parametro indicativo dell'esperienza del cliente. In genere si tratta di un parametro personalizzato che potrebbe avere origine dal codice dell'applicazione all'interno del carico di lavoro.

Quando prevedi una strategia basata sulla domanda in un progetto, tieni presenti due considerazioni principali. In primo luogo, devi capire con quale velocità è necessario predisporre le nuove risorse. In secondo luogo, devi capire che la dimensione del margine tra domanda e risorse fornite cambierà. Devi prepararti ad affrontare le variazioni nella domanda, nonché le risorse insufficienti.

ELB consente di ricalibrare le risorse distribuendo la domanda su più risorse. Man mano che implementi altre risorse, le aggiungi al load balancer per soddisfare la domanda. Elastic Load Balancing dispone di supporto per istanze Amazon EC2, container, indirizzi IP e funzioni AWS Lambda.

Fornitura basata sul tempo: una strategia basata sul tempo allinea la capacità delle risorse alla domanda, che è prevedibile o ben definita nel tempo. In genere questa strategia non dipende dai livelli di utilizzo delle risorse. Una strategia basata sul tempo assicura che le risorse siano disponibili nel momento esatto in cui vengono richieste e possano essere fornite senza ritardi dovuti alle procedure di avvio e ai controlli di sistema o di coerenza. Attraverso una strategia basata sul tempo puoi fornire risorse aggiuntive o incrementare la capacità nei periodi più intensi.

Puoi utilizzare Auto Scaling pianificato per implementare un approccio in base al tempo. I carichi di lavoro possono essere programmati per eseguire il dimensionamento in determinati momenti (ad esempio, all'inizio dell'orario di lavoro), garantendo quindi la disponibilità delle risorse all'arrivo degli utenti on demand.

Puoi anche sfruttare API e SDK AWS e AWS CloudFormation per predisporre e ritirare automaticamente interi ambienti quando ne hai bisogno. Questa strategia risulta particolarmente

adatta per gli ambienti di sviluppo o di prova che operano solo in determinati orari di lavoro o periodi di tempo.

Puoi usare le API per dimensionare le risorse all'interno di un ambiente (dimensionamento verticale). Ad esempio, potresti dimensionare verticalmente un carico di lavoro di produzione modificando la dimensione o la classe dell'istanza. Ciò è possibile interrompendo e avviando l'istanza e selezionando una dimensione o classe diversa. Questa tecnica può essere applicata anche ad altre risorse, come gli Elastic Volumes Amazon Elastic Block Store (Amazon EBS), che possono essere modificati per aumentarne le dimensioni, regolarne le prestazioni (IOPS) o modificare il tipo di volume durante l'utilizzo.

Quando prevedi una strategia basata sul tempo in un progetto, tieni presenti due considerazioni principali. In primo luogo, che livello di coerenza presenta il modello di utilizzo? In secondo luogo, qual è l'impatto se il modello cambia? Puoi migliorare l'accuratezza delle previsioni monitorando i tuoi carichi di lavoro e utilizzando la business intelligence. Se noti cambiamenti significativi nel modello di utilizzo, puoi modificare i tempi per assicurarti che la copertura sia fornita.

Passaggi dell'implementazione

- Configura la pianificazione in base al tempo: Per le variazioni prevedibili della domanda, il dimensionamento basato sul tempo può fornire il numero corretto di risorse in modo tempestivo. Inoltre è utile se la creazione e la configurazione delle risorse non sono abbastanza veloci da rispondere alle variazioni della domanda. Utilizzando l'analisi del carico di lavoro, configura il dimensionamento pianificato utilizzando AWS Auto Scaling.
- Configura il dimensionamento automatico: Per configurare il dimensionamento in base ai parametri del carico di lavoro attivi, utilizza Amazon Auto Scaling. Utilizza l'analisi e configura l'auto scaling per attivare i livelli di risorse corretti e assicurati che il carico di lavoro si ridimensioni nel tempo richiesto.

Risorse

Documenti correlati:

- AWS Auto Scaling
- AWS Instance Scheduler
- Nozioni di base su Amazon EC2 Auto Scaling
- Nozioni di base su Amazon SQS

- Dimensionamento programmato per Amazon EC2 Auto Scaling

Ottimizzazione nel tempo

Domanda

- COST 10 In che modo valuti i nuovi servizi?

COST 10 In che modo valuti i nuovi servizi?

Poiché AWS rilascia nuovi servizi e caratteristiche, è buona prassi rivedere le decisioni correnti sull'architettura per garantire che continuino a essere le più convenienti.

Best practice

- COST10-BP01 Sviluppo di un processo di revisione del carico di lavoro
- COST10-BP02 Valutazione e analisi regolare del carico di lavoro

COST10-BP01 Sviluppo di un processo di revisione del carico di lavoro

Sviluppa un processo che definisca i criteri e il processo per la revisione del carico di lavoro. L'impegno analitico deve riflettere il potenziale risultato. Ad esempio, i carichi di lavoro principali o i carichi di lavoro con un valore superiore al 10% della fattura sono analizzati trimestralmente, mentre i carichi di lavoro inferiori al 10% sono analizzati annualmente.

Livello di rischio associato se questa best practice non fosse adottata: Alto

Guida all'implementazione

Per far sì che il carico di lavoro sia sempre efficiente in termini di costi, devi analizzarlo regolarmente per stabilire se ci sono opportunità di implementare nuovi servizi, funzionalità e componenti. Per garantire costi complessivi ridotti, il processo deve essere proporzionale al potenziale risparmio. Ad esempio, i carichi di lavoro che rappresentano il 50% della spesa complessiva devono essere esaminati con maggiore regolarità e più nel dettaglio rispetto ai carichi di lavoro che rappresentano il 5% della spesa complessiva. Prendi in considerazione qualsiasi fattore esterno o volatilità. Se il carico di lavoro serve una determinata area geografica o un segmento di mercato e viene previsto un cambiamento in tale area, revisioni più frequenti possono portare a risparmi sui costi. Un altro fattore in fase di revisione è rappresentato dall'impegno necessario per implementare le modifiche. Se i test

e la convalida delle modifiche comportassero costi significativi, le revisioni dovrebbero essere meno frequenti.

Prendi in considerazione il costo nel lungo termine della manutenzione di componenti e risorse obsoleti e legacy, e dell'impossibilità di implementare in essi nuove funzionalità. L'attuale costo del test e della convalida potrebbe superare il vantaggio auspicato. Tuttavia, nel corso del tempo, il costo di apportare modifiche potrebbe crescere in modo significativo all'aumentare del divario tra il carico di lavoro e le tecnologie attuali, generando costi ancora maggiori. Ad esempio, il costo del passaggio a un nuovo linguaggio di programmazione potrebbe attualmente non risultare conveniente. Tuttavia, nel giro di cinque anni, il costo del personale qualificato per tale linguaggio potrebbe aumentare e, a causa dell'aumento del carico di lavoro, potresti dover trasferire un sistema ancora più grande al nuovo linguaggio, richiedendo sforzi ancora maggiori rispetto a prima.

Suddividi il carico di lavoro in componenti, assegna un costo ai componenti (una stima è sufficiente) e quindi elenca i fattori (ad esempio, impegno richiesto e mercati esterni) accanto a ciascun componente. Utilizza questi indicatori per determinare una frequenza di revisione per ogni carico di lavoro. Ad esempio, potresti avere i server web come un costo elevato, con un impegno di modifica ridotto e fattori esterni elevati, e da questo potrebbe derivare un'alta frequenza di revisione. Un database centrale può essere un costo medio, con un impegno di modifica elevato e un basso fattore esterno, e da questo potrebbe derivare una frequenza di revisione media.

Passaggi dell'implementazione

- Definisci la frequenza di revisione: Definisci la frequenza con cui il carico di lavoro e i relativi componenti devono essere revisionati. Si tratta di una combinazione di fattori e può variare da carico di lavoro a carico di lavoro all'interno dell'organizzazione, ma può anche variare tra i componenti del carico di lavoro. Fattori più comuni sono: l'importanza per l'organizzazione misurata in termini di fatturato o marchio, il costo totale di esecuzione del carico di lavoro (inclusi costi operativi e delle risorse), la complessità del carico di lavoro, la facilità di implementazione di una modifica, eventuali accordi di licenza software e l'eventuale aumento dei costi di licenza dovuti a licenze punitive in seguito a una modifica. I componenti possono essere definiti a livello funzionale o tecnico come server Web e database, oppure come risorse di calcolo e storage. Equilibra i fattori di conseguenza e prevedi un periodo per il carico di lavoro e i relativi componenti. Puoi decidere di esaminare l'intero carico di lavoro ogni 18 mesi, esaminare i server Web ogni 6 mesi, il database ogni 12 mesi, l'elaborazione e lo storage a breve termine ogni 6 mesi e lo storage a lungo termine ogni 6 mesi.
- Definisci l'accuratezza della revisione: Stabilisci quanto impegno deve essere impiegato per la revisione dei componenti o dell’intero carico di lavoro. Come per la frequenza di revisione, si tratta

di un equilibrio tra più fattori. Puoi decidere di dedicare una settimana all'analisi del database e quattro ore alla revisione dello storage.

Risorse

Documenti correlati:

- Blog delle novità di AWS
- Tipi di cloud computing
- Novità di AWS

COST10-BP02 Valutazione e analisi regolare del carico di lavoro

I carichi di lavoro esistenti vengono regolarmente analizzati per ogni processo definito.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Per ottenere i vantaggi offerti dai nuovi servizi e dalle nuove funzionalità di AWS, devi eseguire il processo di revisione sui carichi di lavoro e implementare nuovi servizi e funzionalità in base alle esigenze. Ad esempio, puoi esaminare i carichi di lavoro e sostituire il componente di messaggistica con Amazon Simple Email Service (Amazon SES). Ciò elimina il costo di gestione e manutenzione di un parco istanze, fornendo al contempo tutte le funzionalità a un costo ridotto.

Passaggi dell'implementazione

- Esamina regolarmente il carico di lavoro: Utilizzando il processo definito, esegui le revisioni con la frequenza specificata. Verifica di dedicare la quantità di impegno necessaria per ciascun componente. Questo processo è simile a quello di progettazione iniziale in cui hai selezionato i servizi per l'ottimizzazione dei costi. Analizza i servizi e i vantaggi che porterebbero, il costo del tempo necessario per la modifica, non solo i vantaggi a lungo termine.
- Implementa nuovi servizi: Se in seguito all'analisi ritieni di dover implementare modifiche, esegui innanzitutto una baseline del carico di lavoro per scoprire il costo corrente per ogni output. Implementa le modifiche, quindi esegui un'analisi per verificare il nuovo costo per ogni output.

Risorse

Documenti correlati:

- Blog delle novità di AWS
- Tipi di cloud computing
- Novità di AWS

Sostenibilità

Argomenti

- Selezione delle regioni
- Modelli di comportamento degli utenti
- Modelli di software e architetture
- Modelli di dati
- Modelli hardware
- Processo di sviluppo e implementazione

Selezione delle regioni

Domanda

- SUS 1 In che modo selezioni le Regioni per sostenere i tuoi obiettivi di sostenibilità?

SUS 1 In che modo selezioni le Regioni per sostenere i tuoi obiettivi di sostenibilità?

Scegli le Regioni in cui implementerai i tuoi carichi di lavoro, tenendo presenti sia i requisiti aziendali sia gli obiettivi di sostenibilità.

Best practice:

SUS01-BP01 Scegli le Regioni vicino ai progetti di energia rinnovabile di Amazon e le Regioni in cui la griglia presenta un'intensità di emissione di anidride carbonica nota inferiore a quella di altre sedi (o Regioni).

Scegli le Regioni vicino ai progetti di energia rinnovabile di Amazon e le Regioni in cui la griglia presenta un'intensità di emissione di anidride carbonica nota inferiore a quella di altre sedi (o Regioni).

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

Scegli le Regioni vicino ai progetti di energia rinnovabile di Amazon e le Regioni in cui la griglia presenta un'intensità di emissione di anidride carbonica nota inferiore a quella di altre sedi (o Regioni).

Risorse

Documenti correlati:

- Amazon Around the Globe
- Metodologia delle energie rinnovabili
- Cosa considerare quando si seleziona una Regione per i propri carichi di lavoro

Modelli di comportamento degli utenti

Domanda

- SUS 2 In che modo sfrutti i modelli di comportamento degli utenti per sostenere i tuoi obiettivi di sostenibilità?

SUS 2 In che modo sfrutti i modelli di comportamento degli utenti per sostenere i tuoi obiettivi di sostenibilità?

Il modo in cui gli utenti utilizzano i tuoi carichi di lavoro e altre risorse può aiutarti a identificare i miglioramenti da implementare per raggiungere gli obiettivi di sostenibilità. Dimensiona l'infrastruttura in modo che si adegui continuamente al carico degli utenti e implementa solo le risorse minime richieste per supportare gli utenti. Allinea i livelli di servizio alle esigenze dei clienti. Posiziona le risorse in modo da limitare la rete richiesta per il consumo da parte degli utenti. Elimina risorse esistenti non utilizzate. Identifica le risorse create non utilizzate e smetti di generarle. Offri ai membri del tuo team dispositivi in grado di soddisfare le loro esigenze con un impatto ridotto in termini di sostenibilità.

Best practice:

SUS02-BP01 Dimensionamento dell'infrastruttura in base al carico degli utenti

Identifica i periodi di utilizzo assente o ridotto e riduci le risorse per evitare capacità in eccesso e migliorare il livello di efficienza.

Anti-pattern comuni:

- Mancato dimensionamento dell'infrastruttura in base al carico degli utenti.
- Costante dimensionamento manuale dell'infrastruttura.
- Dopo un evento di dimensionamento, lasci una capacità aumentata anziché ridurre il dimensionamento.

Vantaggi dell'adozione di questa best practice: la configurazione e il testing dell'elasticità del carico di lavoro ti consentirà di ridurre l'impatto ambientale del carico di lavoro, risparmiare denaro e mantenere i benchmark prestazionali. Puoi sfruttare i vantaggi dell'elasticità nel cloud per dimensionare automaticamente la capacità durante o dopo i picchi dei carichi degli utenti per essere sicuro di utilizzare solo il numero esatto di risorse e soddisfare le esigenze dei clienti.

Livello di rischio associato se questa best practice non fosse adottata: medio

Guida all'implementazione

- L'elasticità corrisponde all'offerta di risorse disponibili rispetto alla relativa domanda. Istanze, container e funzioni forniscono tutti meccanismi di elasticità, in combinazione con il ridimensionamento automatico o sotto forma di funzionalità del servizio. Utilizza l'elasticità nell'architettura in uso per assicurarti di ridurre il dimensionamento del carico di lavoro in modo semplice e rapido durante un periodo di basso carico degli utenti:
  - Utilizza Amazon EC2 Auto Scaling per verificare che sia disponibile il numero corretto di istanze Amazon EC2 per gestire il carico degli utenti dell'applicazione.
  - Utilizza Application Auto Scaling per dimensionare automaticamente le risorse per servizi AWS diversi da Amazon EC2, ad esempio funzioni Lambda o servizi Amazon Elastic Container Service (Amazon ECS).
  - Utilizza Kubernetes Cluster Autoscaler per dimensionare automaticamente i cluster Kubernetes su AWS.
- Verifica che le metriche per il dimensionamento verticale o orizzontale siano convalidate in base al tipo di carico di lavoro implementato. Se distribuisci un'applicazione di transcodifica video, è previsto il 100% di utilizzo della CPU e non deve essere il parametro principale. Puoi utilizzare una metrica personalizzata (ad esempio, l'utilizzo della memoria) per la policy di dimensionamento, se necessario. Per scegliere la metrica corretta, consulta le linee guida seguenti per Amazon EC2:
  - La metrica deve essere una metrica di utilizzo valida e descrivere il livello di impiego di un'istanza.

  - Il valore della metrica deve aumentare o diminuire proporzionalmente in base al numero di istanze nel gruppo con Auto Scaling.
- Utilizza il dimensionamento dinamico invece del dimensionamento manuale per il gruppo con Auto Scaling in uso. È consigliabile utilizzare le policy di dimensionamento del monitoraggio degli obiettivi nel dimensionamento dinamico.
- Verifica che le implementazioni dei carichi di lavoro siano in grado di gestire eventi di dimensionamento verticale/orizzontale. Crea scenari di test per eventi di ridimensionamento per garantire che il carico di lavoro si comporti come previsto. Puoi utilizzare la cronologia delle attività per testare e verificare un'attività di dimensionamento per un gruppo con Auto Scaling.
- Analizza il tuo carico di lavoro per individuare modelli prevedibili e dimensionare le tue risorse in modo proattivo, anticipando variazioni nella domanda previste e pianificate. Utilizza il dimensionamento predittivo con Amazon EC2 Auto Scaling per eliminare la necessità di sottoporre la capacità a eccessive prove.

Risorse

Documenti correlati:

- Nozioni di base su Amazon EC2 Auto Scaling
- Dimensionamento predittivo per EC2, alimentato dal machine learning
- Analizza il comportamento degli utenti tramite Amazon OpenSearch Service, Amazon Data Firehose e Kibana
- Che cos'è Amazon CloudWatch?
- Che cos'è AWS X-Ray?
- Log di flusso VPC
- Monitoraggio del carico del database con Performance Insights su Amazon RDS
- Introducing Native Support for Predictive Scaling with Amazon EC2 Auto Scaling (Introduzione al supporto nativo per il dimensionamento predittivo con Amazon EC2 Auto Scaling)
- How to create an Amazon EC2 Auto Scaling policy based on a memory utilization metric (Linux) (Come creare una policy an Amazon EC2 Auto Scaling basata sulla metrica di utilizzo della memoria [Linux])
- Introducing Karpenter - An Open-Source, High-Performance Kubernetes Cluster Autoscaler (Introduzione a Karpenter - Kubernetes Cluster Autoscaler, uno strumento open source a elevate prestazioni)

Video correlati:

- Better, faster, cheaper compute: Cost-optimizing Amazon EC2 (Calcolo migliore, più veloce, più economico: ottimizzazione dei costi di Amazon EC2) (CMP202-R1)

Esempi correlati:

- Lab: Amazon EC2 Auto Scaling Group Examples (Laboratorio: Esempi di gruppi Amazon EC2 Auto Scaling)
- Lab: Implement Autoscaling with Karpenter (Laboratorio: Implementazione del dimensionamento automatico con Karpenter)

SUS02-BP02 Allineamento degli SLA agli obiettivi di sostenibilità

Definisci e aggiorna gli Accordi sul Livello di Servizio (SLA), come la disponibilità o i periodi di conservazione dei dati, per ridurre il numero di risorse richieste a supporto dei carichi di lavoro, senza per questo venire meno ai requisiti di business.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Definisci SLA che supportano i tuoi obiettivi di sostenibilità e, al tempo stesso, soddisfano gli altri requisiti aziendali.
- Ridefinisci gli SLA per soddisfare i requisiti di business e non per superarli.
- Accetta dei compromessi che riducano l'impatto in termini di sostenibilità in modo significativo in cambio di una diminuzione accettabile dei livelli di servizio.
- Usa modelli di progettazione che danno la priorità a funzioni strategiche per la tua azienda e consentono livelli di servizio inferiori (in tema di obiettivi per tempi di risposta o di ripristino) per funzioni non critiche.

Risorse

Documenti correlati:

- Contratti sul livello di servizio (SLA) AWS
- L’importanza del contratto sul livello di servizi (SLA) per i provider SaaS

Video correlati:

- Building Sustainably on AWS (Creare sostenibilità su AWS)

SUS02-BP03 Interruzione della creazione e della manutenzione di risorse inutilizzate

Analizza le risorse delle applicazioni (come report precompilati, set di dati e immagini statiche) e i modelli di accesso alle risorse per identificare ridondanze, sottoutilizzi e obiettivi potenziali di disattivazione. Consolida le risorse generate con contenuti ridondanti (come, ad esempio, report mensili con set di dati e output comuni o in sovrapposizione) per eliminare le risorse utilizzate per la duplicazione degli output. Disattiva le risorse non utilizzate (come, ad esempio, immagini di prodotto non più in vendita) per liberare le risorse usate e ridurre il numero di risorse sfruttate per supportare il carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Gestisci le risorse statiche ed elimina quelle che non sono più considerate necessarie.
- Gestisci le risorse generate, interrompi la loro generazione ed elimina le risorse che non sono più necessarie.
- Analizza le risorse generate in sovrapposizione per rimuovere le elaborazioni ridondanti.
- Istruisci le terze parti affinché smettano di produrre e di archiviare per tuo conto risorse gestite non più necessarie.
- Istruisci le terze parti e invitale a consolidare le risorse ridondanti prodotte per tuo conto.

Risorse

Documenti correlati:

- Optimizing your AWS Infrastructure for Sustainability, Part II: Storage (Ottimizzazione dell'infrastruttura AWS per la sostenibilità, Parte II: Achiviazione)

Video correlati:

- Building Sustainably on AWS (Creare sostenibilità su AWS)

SUS02-BP04 Ottimizzazione del posizionamento geografico dei carichi di lavoro in base alle posizioni degli utenti

Analizza i modelli di accesso alla rete per capire da quali aree geografiche si connettono i tuoi clienti. Seleziona le Regioni e i servizi per ridurre la distanza che il traffico di rete deve percorrere e diminuire così le risorse totali di rete richieste per supportare il carico di lavoro.

Anti-pattern comuni:

- Selezione della regione del carico di lavoro in base alla propria collocazione.

Vantaggi dell'adozione di questa best practice: il posizionamento di un carico di lavoro in prossimità dei relativi clienti garantisce la latenza più bassa e la contemporanea riduzione dello trasferimento dei dati nella rete e dell'impatto ambientale.

Livello di rischio associato se questa best practice non fosse adottata: medio

Guida all'implementazione

- Seleziona le regioni appropriate per l'implementazione del carico di lavoro in base ai seguenti elementi chiave:
  - Obiettivo di sostenibilità definito: come illustrato in Selezione delle regioni.
  - Ubicazione dei dati: Per le applicazioni a uso intensivo di dati, ad esempio applicazioni di big data e machine learning, il codice dell'applicazione deve essere eseguito il più vicino possibile ai dati.
  - Ubicazione degli utenti: per le applicazioni per gli utenti, scegli una regione vicina alla base di clientela del carico di lavoro.
  - Altri vincoli: considera vincoli quali la sicurezza e la conformità, come illustrato in Cosa considerare quando si seleziona una Regione per i propri carichi di lavoro.
- Utilizza Zone locali AWS per eseguire carichi di lavoro quali, ad esempio, applicazioni di rendering video e applicazioni desktop virtuale a uso intensivo di grafica. Le zone locali consentono di sfruttare i vantaggi derivanti dalla disponibilità di risorse di calcolo e archiviazione più vicine agli utenti finali.
- Utilizza la memorizzazione nella cache locale o soluzioni di memorizzazione nella cache AWS per le risorse di frequente utilizzo per migliorare le prestazioni, ridurre lo spostamento dei dati e minimizzare l'impatto ambientale.

    - Utilizza Amazon CloudFront per memorizzare nella cache contenuti statici come immagini, script e video, nonché contenuti dinamici come risposte API o applicazioni Web.
    - Utilizza Amazon ElastiCache per memorizzare nella cache i contenuti per le applicazioni Web.
    - Utilizza DynamoDB Accelerator per aggiungere accelerazione in memoria alle tabelle DynamoDB.
- Utilizza servizi in grado di supportarti nell'esecuzione del codice in posizioni più vicine agli utenti del carico di lavoro:
    - Utilizza Lambda@Edge per operazioni a uso intensivo di risorse di calcolo eseguite quando gli oggetti non si trovano nella cache.
    - Utilizza Funzioni Amazon CloudFront per casi d'uso semplici, ad esempio manipolazioni di risposte o richieste HTTP(s) che possono essere eseguite da funzioni di breve durata.
    - Utilizza AWS IoT Greengrass per eseguire la memorizzazione nella cache di risorse di calcolo, messaggistica e dati per i dispositivi connessi.
- Utilizza il pooling delle connessioni per consentire il loro riutilizzo e ridurre le risorse richieste.
- Utilizza archivi di dati distribuiti che non si affidano a connessioni persistenti e aggiornamenti sincroni per garantire coerenza e servire le popolazioni regionali.
- Sostituisci la capacità di rete statica preassegnata con una capacità dinamica condivisa e condividi l'impatto in termini di sostenibilità della capacità di rete con altri sottoscrittori.

Risorse

Documenti correlati:

- Optimizing your AWS Infrastructure for Sustainability, Part III: Networking (Ottimizzazione dell'infrastruttura AWS per la sostenibilità, parte III: reti)
- Documentazione su Amazon ElastiCache
- Che cos'è Amazon CloudFront?
- Caratteristiche principali di Amazon CloudFront
- Lambda@Edge
- CloudFront Functions (Funzioni CloudFront)
- AWS IoT Greengrass

Video correlati:

- Building Sustainably on AWS (Creare sostenibilità su AWS)

Esempi correlati:

- AWS Networking Workshops (Workshop di rete AWS)

SUS02-BP05 Ottimizzazione delle risorse dei membri del team in base alle attività eseguite

Ottimizza le risorse fornite ai membri del team per ridurre al minimo l'impatto sulla sostenibilità e supportare al tempo stesso le loro esigenze. Esegui ad esempio operazioni complesse, come rendering e compilazione, su desktop cloud condivisi altamente utilizzati, invece che su sistemi per utenti singoli, sottoutilizzati e con un alto dispendio energetico.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Effettua il provisioning di workstation e altri dispositivi in linea con il modo in cui vengono utilizzati.
- Usa desktop virtuali e lo streaming di applicazioni per limitare gli aggiornamenti e i requisiti dei dispositivi.
- Trasferisci i processori o le attività a uso intensivo della memoria nel cloud.
- Valuta l'impatto di processi e sistemi sul ciclo di vita dei tuoi dispositivi e seleziona soluzioni che riducono al minimo i requisiti per la sostituzione dei dispositivi, pur continuando a soddisfare i requisiti di business.
- Implementa la gestione remota dei dispositivi per ridurre gli spostamenti aziendali.

Risorse

Documenti correlati:

- Che cos'è Amazon WorkSpaces?
- Documentazione su Amazon AppStream 2.0
- NICE DCV
- Gestione dei gruppi di nodi AWS Systems Manager Manager

Video correlati:

- Building Sustainably on AWS (Creare sostenibilità su AWS)

Modelli di software e architetture

Domanda

- SUS 3 In che modo sfrutti i modelli di software e architetture per sostenere i tuoi obiettivi di sostenibilità?

SUS 3 In che modo sfrutti i modelli di software e architetture per sostenere i tuoi obiettivi di sostenibilità?

Implementa modelli per eseguire lo smoothing del carico e garantire un utilizzo elevato e coerente delle risorse implementate per ridurre al minimo il loro consumo. In seguito alle modifiche nei comportamenti degli utenti nel tempo, alcuni componenti potrebbero diventare inattivi per mancanza di utilizzo. Rivedi modelli e architetture per consolidare i componenti sottoutilizzati e aumentare l'uso complessivo. Ritira i componenti che non sono più necessari. Analizza le prestazioni dei componenti dei tuoi carichi di lavoro e ottimizza quelli che usano la maggior quantità di risorse. Identifica i dispositivi che i clienti utilizzano per accedere ai servizi e implementa modelli in grado di ridurre al minimo la necessità di aggiornamenti dei dispositivi.

Best practice:

SUS03-BP01 Ottimizzazione del software e architetture per processi asincroni e pianificati

Usa progettazioni e architetture software efficienti per ridurre al minimo le risorse medie richieste per unità di lavoro. Implementa meccanismi che generano un utilizzo uniforme dei componenti per ridurre le risorse inattive tra le attività e diminuire l'impatto di picchi di carico.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Accoda le richieste che non necessitano di un'elaborazione immediata.
- Aumenta la serializzazione per diminuire l'utilizzo nella pipeline.
- Modifica la capacità dei singoli componenti per evitare la presenza di risorse inattive in attesa di input.
- Crea buffer e stabilisci limiti di velocità per uniformare il consumo di servizi esterni.
- Usa l'hardware più efficiente disponibile sul mercato per le tue ottimizzazioni software.

- Utilizza architetture basate su code, gestione di pipeline e lavoratori istanza on demand per ottimizzare l'utilizzo dell'elaborazione in batch.
- Pianifica le attività per evitare picchi di carico e conflitti delle risorse dovute a esecuzioni simultanee.
- Elabora i processi nei momenti del giorno in cui l'intensità di emissione di anidride carbonica per l'alimentazione è più bassa.

Risorse

Documenti correlati:

- What is Amazon Simple Queue Service? (Che cos'è Amazon Simple Queue Service?)
- Che cos'è Amazon MQ?
- Dimensionamento basato su Amazon SQS
- What is AWS Step Functions? (Cos'è AWS Step Functions?)
- What is AWS Lambda? (Che cos'è AWS Lambda?)
- Utilizzo di AWS Lambda con Amazon SQS
- What is Amazon EventBridge? (Che cos'è Amazon EventBridge?)

Video correlati:

- Building Sustainably on AWS (Creare sostenibilità su AWS)
- Moving to event-driven architectures (Passaggio ad architetture basate su eventi)

SUS03-BP02 Rimozione o rifattorizzazione dei componenti dei carichi di lavoro con un utilizzo ridotto o assente

Monitora l'attività dei carichi di lavoro per individuare i cambiamenti che si verificano nel tempo nell'utilizzo dei singoli componenti. Elimina i componenti non utilizzati e non più necessari e rifattorizza quelli con scarso utilizzo per limitare lo spreco di risorse.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Analizza il carico (utilizzando indicatori come il flusso delle transazioni e le chiamate API) sui componenti funzionali per individuare i componenti non utilizzati o sottoutilizzati.

- Ritira i componenti non più necessari.
- Rifattorizza i componenti sottoutilizzati.
- Consolida i componenti sottoutilizzati con altre risorse per promuovere un utilizzo efficiente.

Risorse

Documenti correlati:

- Che cos'è AWS X-Ray?
- Che cos'è Amazon CloudWatch?
- Usare ServiceLens per monitorare l'integrità delle applicazioni
- Automated Cleanup of Unused Images in Amazon ECR

Video correlati:

- Building Sustainably on AWS (Creare sostenibilità su AWS)

SUS03-BP03 Ottimizzazione delle aree di codice che consumano la maggior parte del tempo o delle risorse

Monitora l'attività dei carichi di lavoro per individuare i componenti delle applicazioni che usano la maggior parte delle risorse. Ottimizza il codice eseguito all'interno di questi componenti per ridurre l'utilizzo delle risorse e massimizzare al tempo stesso le prestazioni.

Livello di rischio associato se questa best practice non fosse adottata: Bassa

Guida all'implementazione

- Monitora le prestazioni in funzione dell'utilizzo delle risorse per individuare i componenti da ottimizzare che presentano maggiori requisiti di risorse per unità di lavoro.
- Utilizza un profiler di codice per identificare le aree di codice che utilizzano la maggior parte del tempo o delle risorse e trasformale in obiettivi di ottimizzazione.
- Sostituisci gli algoritmi con versioni più efficienti che producono lo stesso risultato.
- Utilizza l'accelerazione hardware per migliorare l'efficienza di blocchi di codice con tempi di esecuzione lunghi.
- Usa il sistema operativo e il linguaggio di programmazione più efficienti per il carico di lavoro.

- Rimuovi l'ordinamento e la formattazione non necessari.
- Usa modelli di trasferimento dei dati che riducono la quantità di risorse utilizzate in base alla frequenza con cui si verificano modifiche dei dati e al modo in cui vengono consumati. Invia ad esempio informazioni sulla modifica dello stato a un client, così eviterà di usare risorse per effettuare richieste e ricevere messaggi irrilevanti come "Nessuna modifica".

Risorse

Documenti correlati:

- Che cos'è Amazon CloudWatch?
- What is Amazon CodeGuru Profiler? (Che cos'è Amazon CodeGuru Profiler?)
- Istanze FPGA
- SDK AWS su Strumenti per creare su AWS

Video correlati:

- Building Sustainably on AWS (Creare sostenibilità su AWS)

SUS03-BP04 Ottimizzazione dell'impatto su dispositivi e apparecchiature dei clienti

Identifica i dispositivi e le attrezzature che i tuoi clienti usano per accedere ai tuoi servizi, il loro ciclo di vita atteso e l'impatto finanziario e di sostenibilità che deriva dalla loro sostituzione. Implementa modelli e architetture software per ridurre al minimo la necessità dei clienti di sostituire dispositivi e aggiornare attrezzature. Implementa ad esempio nuove caratteristiche usando un codice compatibile con versioni di hardware e sistemi operativi precedenti o gestisci la dimensione dei payload in modo che non superino la capacità di archiviazione del dispositivo target.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Inventario dei dispositivi utilizzati dai clienti.
- Esegui i test usando device farm gestite con set di hardware rappresentativi per misurare l'impatto delle tue modifiche e iterare lo sviluppo per ottimizzare i dispositivi supportati.

- Prendi in considerazione la larghezza di banda e la latenza della rete durante la creazione di payload e implementa funzionalità che consentano alle tue applicazioni di lavorare bene anche in presenza di una larghezza di banda ridotta e di link ad alta latenza.
- Elabora in anticipo i payload di dati per ridurre i requisiti di elaborazione in locale e limitare i requisiti relativi al trasferimento di dati.
- Esegui attività a elevata intensità computazionale lato server (come, ad esempio, il rendering delle immagini) oppure usa lo streaming delle applicazioni per migliorare l'esperienza utente sui dispositivi di versioni precedenti.
- Esegui la segmentazione e la paginazione dell'output, soprattutto per le sessioni interattive, per gestire i payload e limitare i requisiti di archiviazione in locale.

Risorse

Documenti correlati:

- Che cos'è AWS Device Farm?
- Documentazione su Amazon AppStream 2.0
- NICE DCV
- Documentazione su Amazon Elastic Transcoder

Video correlati:

- Building Sustainably on AWS (Creare sostenibilità su AWS)

SUS03-BP05 Uso dei modelli e le architetture software che meglio supportano l'accesso ai dati e i modelli di archiviazione

Scopri come i dati vengono utilizzati all'interno del tuo carico di lavoro, consumati dagli utenti, trasferiti e archiviati. Seleziona tecnologie che ti consentono di ridurre l'elaborazione dei dati e i requisiti di archiviazione.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Analizza gli accessi ai dati e i modelli di archiviazione.

- Archivia i file di dati in formati di file efficienti, ad esempio Parquet, per evitare elaborazioni non necessarie (durante ad esempio l'esecuzione di analisi dei dati) e per ridurre l'archiviazione totale assegnata.
- Sfrutta le tecnologie che lavorano in modo nativo con i dati compressi.
- Utilizza il motore del database che meglio supporta il modello di query dominante.
- Gestisci gli indici di database per verificare che le progettazioni degli indici siano in grado di supportare un'esecuzione efficiente delle query.
- Seleziona protocolli di rete che riducano la quantità di capacità di rete utilizzata.

Risorse

Documenti correlati:

- Formati file di supporto alla compressione di Athena
- COPY dai formati dei dati in colonne con Amazon Redshift
- Convertire il formato dei record di input in Firehose
- Opzioni di formato per input e output ETL in AWS Glue
- Migliora le prestazioni delle query su Amazon Athena con una conversione ai formati in colonne
- caricamento di file di dati compressi da Amazon S3 con Amazon Redshift
- Monitoraggio del carico del database con Performance Insights su Amazon Aurora
- Monitoraggio del carico del database con Performance Insights su Amazon RDS
- AWS IoT FleetWise

Video correlati:

- Building Sustainably on AWS (Creare sostenibilità su AWS)

Modelli di dati

Domanda

- SUS 4 In che modo sfrutti i modelli di accesso e di utilizzo dei dati per sostenere i tuoi obiettivi di sostenibilità?

SUS 4 In che modo sfrutti i modelli di accesso e di utilizzo dei dati per sostenere i tuoi obiettivi di sostenibilità?

Implementa procedure di gestione dei dati per ridurre l'archiviazione assegnata richiesta per supportare il carico di lavoro e le risorse necessarie per l'uso correlato. Analizza i tuoi dati e usa tecnologie e configurazioni di archiviazione che meglio supportano il valore aziendale dei dati e il modo in cui vengono utilizzati. Esegui il ciclo di vita dei dati su un'archiviazione più efficiente e meno performante al diminuire dei requisiti ed elimina i dati che non sono più necessari.

Best practice:

SUS04-BP01 Implementazione di una policy di classificazione dei dati

Classifica i dati per comprenderne il significato in favore dei risultati aziendali. Usa queste informazioni per stabilire quando trasferire i dati in un'archiviazione più efficiente dal punto di vista energetico o eliminarli in totale sicurezza.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Determina i requisiti di distribuzione, conservazione ed eliminazione dei tuoi dati.
- Usa l'assegnazione di tag su volumi e oggetti per registrare i metadati usati e stabilire come vengono gestiti, inclusa la loro classificazione.
- Verifica periodicamente il tuo ambiente per individuare dati non classificati e privi di tag e quindi taggare e classificare i dati in maniera adeguata.

Risorse

Documenti correlati:

- Processo di classificazione dei dati
- Utilizzo di Cloud AWS per supportare la classificazione dei dati
- Policy di tag di AWS Organizations

SUS04-BP02 Utilizzo di tecnologie che supportano l'accesso ai dati e i modelli di archiviazione

Usa l'archiviazione in grado di supportare al meglio il modo in cui viene effettuato l'accesso ai dati e come vengono archiviati per ridurre la quantità di risorse assegnate e supportare al tempo stesso il

tuo carico di lavoro. I dispositivi allo stato solido (SSD) utilizzano ad esempio l'energia in modo più intensivo rispetto ai drive magnetici e dovrebbero essere usati solo per casi d'uso di dati attivi. Usa storage di classe di archiviazione ad alta efficienza energetica per i dati ad accesso infrequente.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Monitora i modelli di accesso ai dati.
- Migra i dati alla tecnologia appropriata in base al modello di accesso.
- Migra i dati di archiviazione a servizi progettati per questo scopo.

Risorse

Documenti correlati:

- Tipi di volume di Amazon EBS
- Archivio dell'istanza Amazon EC2
- Piano intelligente Amazon S3
- Utilizzo delle classi di storage di Amazon S3
- Che cos'è Amazon CloudWatch?
- Che cos'è Amazon S3 Glacier?

Video correlati:

- Architectural Patterns for Data Lakes on AWS (Modelli architetturali per i data lake su AWS)

SUS04-BP03 Utilizzo delle policy del ciclo di vita per eliminare i dati non necessari

Gestisci il ciclo di vita di tutti i tuoi dati e applica in automatico cronologie di eliminazione per ridurre i requisiti totali di archiviazione del tuo carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Definisci le policy del ciclo di vita per tutti i tipi di classificazione dei dati.
- Imposta policy automatizzate per il ciclo di vita per applicare le regole correlate.

- Elimina volumi e snapshot inutilizzati.
- Aggrega i dati ove applicabile in base alle regole del ciclo di vita.

Risorse

Documenti correlati:

- Policy per la gestione del ciclo di vita di Amazon ECR
- Gestione del ciclo di vita di Amazon EFS
- Piano intelligente Amazon S3
- Valutazione delle risorse con Regole di AWS Config
- Gestione del ciclo di vita dell'archiviazione su Amazon S3
- Policy per la gestione del ciclo di vita degli oggetti in AWS Elemental MediaStore

Video correlati:

- Gestione del ciclo di vita di Amazon S3

SUS04-BP04 Riduzione del provisioning eccessivo nell'archiviazione a blocchi

Per ridurre la quantità totale di archiviazione assegnata, crea un'archiviazione a blocchi con l'allocazione di dimensioni in base al carico di lavoro. Usa i volumi elastici per espandere l'archiviazione all'aumentare dei dati senza dover ridimensionare l'archiviazione collegata alle risorse di calcolo. Esamina regolarmente i volumi elastici e riduci i volumi con un provisioning eccessivo per adattarli alla dimensione corrente dei dati.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Monitora l'utilizzo dei volumi di dati.
- Utilizza volumi elastici e servizi di dati a blocchi gestiti per automatizzare l'allocazione di archivi aggiuntivi man mano che i dati persistenti aumentano.
- Imposta i livelli target di utilizzo per i volumi di dati e ridimensiona i volumi al di fuori degli intervalli previsti.
- Dimensiona i volumi di sola lettura per adattarli ai dati.

- Migra i dati su archivi oggetti per evitare il provisioning di capacità eccessive da dimensioni di volumi fisse su archiviazioni a blocchi.

Risorse

Documenti correlati:

- Volumi elastici Amazon EBS
- Documentazione di Amazon FSx
- Che cos'è Amazon CloudWatch?
- Che cos'è Amazon Elastic File System?

SUS04-BP05 Eliminazione dei dati ridondanti o non necessari

Duplica i dati solo quando è necessario per ridurre la quantità totale di archiviazione utilizzata. Utilizza tecnologie di backup che deduplicano i dati a livello di file e blocco. Limita l'uso di configurazioni Redundant Array of Independent Drives (RAID), ad eccezione dei casi in cui sono richieste per soddisfare gli SLA.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Utilizza meccanismi che possono deduplicare i dati a livello di blocco e oggetto.
- Utilizza la tecnologia di backup in grado di effettuare backup incrementali e deduplicare i dati a livello di blocco, file e oggetto.
- Usa RAID solo quando richiesto per soddisfare gli SLA.
- Centralizza i registri e traccia i dati, deduplica le voci di registro identiche e stabilisci meccanismi per ottimizzarne la verbosità quando necessario.
- Popola in anticipo le cache solo quando è necessario.
- Definisci il monitoraggio e l'automazione della cache per ridimensionarla in base alle esigenze.
- Rimuovi le implementazioni e le risorse obsolete dagli archivi di oggetti e dalle cache edge durante l'invio di nuove versioni del carico di lavoro.

Risorse

Documenti correlati:

- Snapshot Amazon EBS
- Modifica la conservazione dei dati di registro in CloudWatch Logs
- Deduplicazione dei dati su Amazon FSx per Windows File Server
- Caratteristiche di Amazon FSx per ONTAP, inclusa la deduplicazione dei dati
- Invalidazione dei file su Amazon CloudFront
- Utilizzo di AWS Backup per il backup e il ripristino dei file system di Amazon EFS
- Che cos'è Amazon CloudWatch Logs?
- Lavorare con i backup su Amazon RDS

Esempi correlati:

- Laboratorio: Optimize Data Pattern Using Amazon Redshift Data Sharing

SUS04-BP06 Utilizzo di file system condivisi o archiviazione di oggetti per accedere a dati comuni

Adotta l'archiviazione condivisa e singole fonti di verità per evitare la duplicazione dei dati e ridurre i requisiti di archiviazione complessiva del tuo carico di lavoro. Recupera i dati dall'archiviazione condivisa solo in base alle esigenze. Scollega volumi non utilizzati per rendere disponibili più risorse.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Esegui la migrazione dei dati nell'archiviazione condivisa quando i dati hanno più consumer.
- Recupera i dati dall'archiviazione condivisa solo in base alle esigenze.
- Elimina i dati secondo quanto richiesto dai tuoi modelli di utilizzo e implementa funzionalità di durata (TTL) per gestire i dati nella cache.
- Distacca i volumi dai client che non li utilizzano attivamente.

Risorse

Documenti correlati:

- Amazon FSx
- Strategie di cache
- Che cos'è Amazon Elastic File System?

- Che cos'è Amazon S3?

SUS04-BP07 Riduzione al minimo dello spostamento di dati tra reti

Usa un'archiviazione condivisa e accedi ai dati da archivi regionali per contenere le risorse di rete totali necessarie per supportare i trasferimenti dei dati per il carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Archivia i dati il più vicino possibile al consumer.
- Esegui la partizione dei servizi consumati a livello regionale in modo che i dati specifici della Regione siano archiviati nella Regione in cui sono usati.
- Utilizza la duplicazione a livello di blocco anziché la duplicazione a livello di file o oggetto durante la copia delle modifiche nella rete.
- Comprimi i dati prima di spostarli sulla rete.

Risorse

Documenti correlati:

- Optimizing your AWS Infrastructure for Sustainability, Part III: Networking
- Infrastruttura globale di AWS
- Funzionalità principali di Amazon CloudFront incluso CloudFront Global Edge Network
- Compressione delle richieste HTTP in Amazon OpenSearch Service
- Compressione intermedia dei dati con Amazon EMR
- Caricamento di file di dati compressi da Amazon S3 a Amazon Redshift
- Distribuzione dei file compressi con Amazon CloudFront

SUS04-BP08 Backup dei dati solo quando sono difficili da ricreare

Per ridurre al minimo l'uso delle risorse di archiviazione, esegui il backup solo dei dati che abbiano un valore aziendale o siano considerati necessari per soddisfare requisiti di conformità. Esamina le policy di backup ed escludi l'archiviazione temporanea che non offre valore in uno scenario di ripristino.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Utilizza la classificazione dei dati per stabilire i dati per i quali è necessario eseguire il backup.
- Escludi i dati che possono essere ricreati facilmente.
- Escludi dati temporanei dai backup.
- Escludi copie secondarie dei dati, a meno che il tempo necessario per ripristinare tali dati da una posizione comune superi i contratti sul livello di servizio (SLA).

Risorse

Documenti correlati:

- Using AWS Backup to back up and restore Amazon EFS file systems
- Snapshot Amazon EBS
- Lavorare con i backup su Amazon Relational Database Service

Modelli hardware

Domanda

- SUS 5 In che modo la gestione dell'hardware e le procedure di utilizzo sostengono i tuoi obiettivi di sostenibilità?

SUS 5 In che modo la gestione dell'hardware e le procedure di utilizzo sostengono i tuoi obiettivi di sostenibilità?

Cerca opportunità per ridurre l'impatto dei carichi di lavoro in termini di sostenibilità apportando modifiche alle tue prassi di gestione hardware. Riduci la quantità di hardware necessaria per il provisioning e l'implementazione e seleziona l'hardware più efficiente per il singolo carico di lavoro.

Best practice:

SUS05-BP01 Utilizzo della quantità minima di hardware per soddisfare le esigenze aziendali

Le funzionalità del cloud consentono di apportare modifiche frequenti alle implementazioni dei carichi di lavoro. Aggiorna i componenti distribuiti man mano che le tue esigenze cambiano.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Abilita il dimensionamento orizzontale e utilizza l'automazione per aumentare orizzontalmente i carichi man mano che crescono e ridurli orizzontalmente man mano che diminuiscono.
- Dimensiona usando i piccoli incrementi per carichi di lavoro variabili.
- Allinea il dimensionamento a modelli di utilizzo ciclici (ad esempio, un sistema di buste paga con attività intense di elaborazione bisettimanali) al variare dei carichi giornalieri, settimanali, mensili o annuali.
- Negozia SLA che consentano una riduzione temporanea della capacità quando l'automazione implementa risorse di sostituzione.

Risorse

Documenti correlati:

- Documentazione di AWS Compute Optimizer
- Uso di Lambda: ottimizzazione delle performance
- Documentazione sulla scalabilità automatica

SUS05-BP02 Utilizzo di tipi di istanze con il minimo impatto

Monitora costantemente il rilascio di nuovi tipi di istanza e sfrutta le migliorie in tema di efficienza energetica, inclusi i tipi di istanza progettati per supportare carichi di lavoro specifici, come la formazione del machine learning, le inferenze e la transcodifica dei video.

Anti-pattern comuni:

- Utilizzi una sola famiglia di istanze.
- Utilizzi solo istanze x86.
- Specifichi un tipo di istanza nella configurazione Amazon EC2 Auto Scaling.
- Utilizzi istanze AWS in un modo per il quale non sono state progettate, ad esempio utilizzi istanze ottimizzate per il calcolo per un carico di lavoro a uso intensivo della memoria.
- Non valuti regolarmente l'uso di nuovi tipi di istanza.
- Non segui i consigli ricevuti dagli strumenti di dimensionamento AWS, ad esempio AWS Compute Optimizer.

Vantaggi dell'adozione di questa best practice: l'uso di risorse energeticamente efficienti e di dimensioni corrette ti consente di ridurre in modo considerevole l'impatto ambientale e i costi del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: basso

Guida all'implementazione

- Esplora e approfondisci i tipi di istanza in grado di ridurre l'impatto ambientale del carico di lavoro.
  - Iscriviti a Novità di AWS per rimanere aggiornato sulle più recenti tecnologie e istanze AWS.
  - Approfondisci i vari tipi di istanza AWS.
  - Impara a conoscere le istanze basate su AWS Graviton, che offrono le migliori prestazioni per watt di energia utilizzato in Amazon EC2 guardando re:Invent 2020 - Deep dive on AWS Graviton2 processor-powered Amazon EC2 instances (re:Invent 2020 - Approfondimenti relativi alle istanze Amazon EC2 con tecnologia basata su processi AWS Graviton2) e Approfondisci le istanze AWS Graviton3 e Amazon EC2 C7g.
- Pianifica la transizione del carico di lavoro a tipi di istanza caratterizzati da un minore impatto.
  - Definisci un processo per valutare nuove caratteristiche o istanze per il carico di lavoro. Sfrutta l'agilità del cloud per testare in modo semplice e rapido in che modo i nuovi tipi di istanza possono migliorare la sostenibilità ambientale del carico di lavoro. Utilizza metriche proxy per misurare la quantità di risorse necessarie per completare un'unità di lavoro.
  - Se possibile, modifica il carico di lavoro in modo che funzioni con diversi numeri di CPU e quantità di memoria diverse per massimizzare la scelta del tipo di istanza.
  - Valuta l'ipotesi di trasferire il carico di lavoro in istanze basate su Graviton per migliorare l'efficienza e le prestazioni del carico di lavoro (consulta AWS Graviton Fast Start e AWS Graviton2 for ISVs (AWS Graviton2 per fornitori di software indipendente [ISV])). Tieni presente le considerazioni relative alla transizione dei carichi di lavoro in istanze Amazon Elastic Compute Cloud basate su AWS Graviton.
  - Valuta l'ipotesi di selezionare l'opzione AWS Graviton quando utilizzi i servizi gestiti da AWS.
  - Esegui la migrazione del carico di lavoro nelle regioni che offrono istanze con il minor impatto in termini di sostenibilità e che contemporaneamente soddisfano i requisiti aziendali.
  - Per i carichi di lavoro di machine learning, utilizza istanze Amazon EC2, che sono basate su chip Amazon Machine Learning personalizzati come AWS Trainium, AWS Inferentiae Amazon EC2 DL1.
  - Utilizza l' inferenza con funzione di suggerimento Amazon SageMaker per dimensionare l'endpoint dell'inferenza ML.

    - Per i carichi di lavoro con transcodifica video in tempo reale, utilizza istanze Amazon EC2 VT1.
    - Per carichi di lavoro con picchi (carichi di lavoro con requisiti non frequenti di capacità aggiuntiva), utilizza istanze espandibili di prestazioni.
    - Per carichi di lavoro stateless e con tolleranza ai guasti, usa le istanze Spot Amazon EC2 per aumentare l'utilizzo complessivo del cloud e ridurre l'impatto di sostenibilità delle risorse inutilizzate.
- Esegui e ottimizza l'istanza del carico di lavoro.
    - Per i carichi di lavoro effimeri, valuta le metriche Amazon CloudWatch dell'istanza , ad esempio `CPUUtilization` , per verificare se l'istanza è inattiva o sottoutilizzata.
    - Per i carichi di lavoro stabili, esegui controlli con gli strumenti di dimensionamento AWS come AWS Compute Optimizer a intervalli regolari per individuare eventuali opportunità di ottimizzazione e ridimensionamento delle istanze.

Risorse

Documenti correlati:

- Optimizing your AWS Infrastructure for Sustainability, Part I: Compute (Ottimizzazione dell'infrastruttura AWS per la sostenibilità, Parte I: Calcolo)
- AWS Graviton Processor
- AWS Inferentia
- AWS Trainium
- Amazon EC2 DL1
- Istanze espandibili di prestazioni di Amazon EC2
- Parchi istanze di prenotazione della capacità di Amazon EC2
- Serie di istanze Spot Amazon EC2
- Istanze Spot Amazon EC2
- Istanze Amazon EC2 VT1
- Tipi di istanza Amazon EC2
- AWS Compute Optimizer
- Funzioni: configurazione della funzione Lambda

Video correlati:

- Deep dive on AWS Graviton2 processer-powered Amazon EC2 instances (Approfondimenti relativi alle istanze Amazon EC2 con tecnologia basata su processi AWS Graviton2)
- Approfondisci le istanze AWS Graviton3 e Amazon EC2 C7g

Esempi correlati:

- Laboratorio: Suggerimenti per il dimensionamento
- Lab: Rightsizing with Compute Optimizer (Laboratorio: Dimensionamento con Compute Optimizer)
- Lab: Optimize Hardware Patterns and Observe Sustainability KPIs (Laboratorio: Ottimizzazione dei modelli hardware e conformità con gli indicatori KPI di sostenibilità)

SUS05-BP03 Utilizzo dei servizi gestiti

I servizi gestiti consentono di affidare ad AWS la responsabilità di mantenere un utilizzo medio alto e un'ottimizzazione della sostenibilità dell'hardware implementato. Utilizza i servizi gestiti per distribuire l'impatto della sostenibilità dei servizi su tutti i tenant relativi, riducendo così il singolo contributo.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Migra dai servizi autogestiti ai servizi gestiti. Per esempio, usa istanze gestite Amazon Relational Database Service (Amazon RDS) invece di mantenere le istanze Amazon RDS su Amazon Elastic Compute Cloud (Amazon EC2)o utilizza servizi di container gestiti, come AWS Fargate, invece di implementare un'infrastruttura di container proprietaria.

Risorse

Documenti correlati:

- AWS Fargate
- Amazon DocumentDB
- Amazon Elastic Kubernetes Service (EKS)
- Amazon Managed Streaming for Apache Kafka (Amazon MSK)
- Amazon Redshift
- Amazon Relational Database Service (RDS)

SUS05-BP04 Ottimizzazione dell'utilizzo delle GPU

Le Graphics Processing Unit (GPU) possono comportare un uso energetico intensivo e molti carichi di lavoro delle GPU sono altamente variabili, come il rendering, la transcodifica e la formazione e la modellazione del machine learning. Esegui le istanze GPU solo per il tempo necessario e disattivale automaticamente quando non occorrono per ridurre la quantità di risorse utilizzate.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Usa le GPU solo per le attività dove risultano più efficienti rispetto ad alternative basate sulla CPU.
- Utilizza l'automazione per rilasciare le istanze GPU non in uso.
- Usa un'accelerazione grafica flessibile al posto di istanze GPU dedicate.
- Sfrutta soluzioni hardware ad hoc specifiche per il carico di lavoro.

Risorse

Documenti correlati:

- Calcolo accelerato
- AWS Inferentia
- AWS Trainium
- Calcoli accelerati per istanze EC2
- Istanze Amazon EC2 VT1
- Amazon Elastic Graphics

Processo di sviluppo e implementazione

Domanda

- SUS 6 In che modo i processi di sviluppo e implementazione adottati supportano i tuoi obiettivi di sostenibilità?

SUS 6 In che modo i processi di sviluppo e implementazione adottati supportano i tuoi obiettivi di sostenibilità?

Cerca opportunità per ridurre l'impatto di sostenibilità apportando modifiche alle tue prassi di sviluppo, test e implementazione.

Best practice:

SUS06-BP01 Adozione di metodi che consentano di introdurre rapidamente migliorie in tema di sostenibilità

Testa e convalida potenziali miglioramenti prima di distribuirli in produzione. Tieni in considerazione il costo dei test quando calcoli il potenziale vantaggio futuro di un miglioramento. Sviluppa metodi di test a basso costo per consentire la distribuzione di piccoli miglioramenti.

Livello di rischio associato se questa best practice non fosse adottata: Medio

Guida all'implementazione

- Aggiungi i requisiti per la sostenibilità al tuo processo di sviluppo.
- Permetti alle risorse di lavorare in parallelo per sviluppare, testare e distribuire miglioramenti della sostenibilità.
- Testa e convalida potenziali miglioramenti all'impatto sulla sostenibilità prima di implementarli in produzione.
- Testa i potenziali miglioramenti utilizzando i componenti rappresentativi minimi realizzabili.
- Implementa miglioramenti di sostenibilità testati in produzione non appena diventano disponibili.

Risorse

Documenti correlati:

- AWS offre soluzioni di sostenibilità

Esempi correlati:

- Laboratorio: Turning cost & usage reports into efficiency reports

SUS06-BP02 Aggiornamento del carico di lavoro

Sistemi operativi, librerie e applicazioni aggiornati possono incidere sull'efficienza dei carichi di lavoro e facilitano l'adozione delle tecnologie più efficienti. Il software aggiornato potrebbe anche includere funzionalità per misurare in modo più accurato l'impatto in termini di sostenibilità del carico di lavoro, poiché i fornitori offrono caratteristiche per raggiungere i propri obiettivi di sostenibilità.

Anti-pattern comuni:

- Ritieni che l'architettura attuale diventerà statica e non verrà mai aggiornata nel corso del tempo.
- Non disponi di sistemi né esegui regolarmente una valutazione per la compatibilità di software e pacchetti aggiornati con il carico di lavoro.
- Introduci modifiche all'architettura nel tempo senza giustificazioni.

Vantaggi dell'adozione di questa best practice: la definizione di un processo per garantire il costante aggiornamento del carico di lavoro ti consentirà di adottare nuove caratteristiche e funzionalità, risolvere i problemi e migliorare l'efficienza del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: basso

Guida all'implementazione

- Definisci un processo e una pianificazione per valutare nuove caratteristiche o istanze per il carico di lavoro. Sfrutta l'agilità del cloud per testare in modo semplice e rapido il modo in cui le nuove funzionalità possono migliorare il carico di lavoro nei seguenti ambiti:
  - Riduzione dell'impatto a livello di sostenibilità.
  - Raggiungimento di maggiore efficienza in termini di prestazioni.
  - Eliminazione delle barriere finalizzata a un miglioramento pianificato.
  - Miglioramento della capacità di misurare e gestire l'impatto a livello di sostenibilità.
- Esegui l'inventario del software e dell'architettura e identifica i componenti che richiedono un aggiornamento. Puoi utilizzare AWS Systems Manager Inventory per raccogliere i metadati relativi a sistema operativo (SO), applicazioni e istanze dalle istanze Amazon EC2 per avere una panoramica immediata su quali istanze stanno eseguendo il software e le configurazioni richieste dalle policy software e quali istanze devono essere aggiornate.
- Individua le modalità di aggiornamento dei componenti del carico di lavoro.
  - Gestisci gli aggiornamenti ad Amazon Machine Images (AMI) per immagini Linux o Windows Server utilizzando EC2 Image Builder.

  - Devi utilizzare Amazon Elastic Container Registry (Amazon ECR) con la pipeline esistente per gestire le immagini di Amazon Elastic Container Service (Amazon ECS) e gestire le immagini di Amazon Elastic Kubernetes Service.
  - AWS Lambda include funzionalità di gestione delle versioni.
- Utilizza l'automazione del processo di aggiornamento per ridurre il livello di impegno per distribuire le nuove funzionalità e limitare gli errori causati dai processi manuali. Utilizza strumenti come AWS Systems Manager Patch Manager per automatizzare il processo di aggiornamento del sistema e pianificare l'attività utilizzando le finestre di manutenzione di AWS Systems Manager.

Risorse

Documenti correlati:

- Centro di progettazione AWS
- Novità di AWS
- Strumenti per sviluppatori in AWS
- AWS Systems Manager Patch Manager

Esempi correlati:

- Well-Architected Labs - Inventory and Patch Management (Gestione di inventario e patch)
- Laboratorio: AWS Systems Manager

SUS06-BP03 Aumento dell'utilizzo degli ambienti di costruzione

Utilizza l'automazione e l'infrastruttura come codice per rendere operativi gli ambienti di preproduzione quando necessario e dismetterli quando non vengono utilizzati. Un modello comune consiste nel pianificare periodi di disponibilità che coincidano con l'orario di lavoro dei membri del team incaricati dello sviluppo. L'ibernazione è uno strumento utile per preservare lo stato e portare rapidamente le istanze online solo quando necessario. Utilizza tipi di istanze con capacità di espansione, istanze Spot, servizi di database elastici, container e altre tecnologie per allineare la capacità di sviluppo e test all'uso.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

- Utilizza l'automazione per ottimizzare l'utilizzo degli ambienti di sviluppo e test.
- Utilizza l'automazione per gestire il ciclo di vita degli ambienti di sviluppo e test.
- Utilizza ambienti rappresentativi minimi realizzabili per lo sviluppo e il test di potenziali miglioramenti.
- Utilizza istanzee on demand per integrare i dispositivi per gli sviluppatori.
- Utilizza l'automazione per massimizzare l'efficienza delle risorse di compilazione.
- Utilizza i tipi di istanze con capacità di espansione, istanze Spot e altre tecnologie per allineare la capacità di compilazione all'uso.
- Adotta servizi cloud nativi per un accesso sicuro alle shell delle istanze invece di implementare parchi istanze di host bastion.

Risorse

Documenti correlati:

- AWS Systems Manager Session Manager
- Istanze espandibili di prestazioni di Amazon EC2
- Che cos'è AWS CloudFormation?

SUS06-BP04 Utilizzo di device farm gestite per i test

Le device farm gestiste distribuiscono l'impatto di sostenibilità della produzione di hardware e dell'utilizzo delle risorse su più tenant. Le device farm gestite offrono diversi tipi di dispositivi e consentono di supportare hardware meno diffusi e di generazioni precedenti e di evitare l'impatto sulla sostenibilità dei clienti dovuti ad aggiornamenti dei dispositivi non necessari.

Livello di rischio associato se questa best practice non fosse adottata: Basso

Guida all'implementazione

Esegui i test usando device farm gestite con set di hardware rappresentativi per misurare l'impatto delle tue modifiche e iterare lo sviluppo per ottimizzare i dispositivi supportati.

Risorse

Documenti correlati:

- What is AWS Device Farm? (Che cos'è AWS Device Farm?)

Avvisi

I clienti sono responsabili della propria valutazione autonoma delle informazioni contenute in questo documento. Questo documento: (a) è solo a scopo informativo, (b) mostra le offerte e le pratiche attuali dei prodotti AWS, soggette a modifiche senza preavviso, e (c) non crea alcun impegno o garanzia da parte di AWS e dei suoi affiliati, fornitori o licenziatari. I prodotti o servizi di AWS vengono forniti "così come sono", senza garanzie, rappresentazioni o condizioni di nessun tipo, sia espresse che implicite. Le responsabilità e gli obblighi di AWS verso i propri clienti sono disciplinati dagli accordi AWS e il presente documento non fa parte né modifica alcun accordo tra AWS e i suoi clienti.