Framework AWS Well-Architected

Il principio dell'efficienza delle prestazioni

Il principio dell'efficienza delle prestazioni: Framework AWS Well-Architected

Table of Contents

Pilastro dell'efficienza delle prestazioni - Framework AWS Well-Architected

Data di pubblicazione: 27 giugno 2024 (Revisioni del documento)

Riassunto

Questo whitepaper è incentrato sul principio dell'efficienza delle prestazioni del Framework AWS Well-Architected. Lo scopo di questo documento è fornire linee guida che aiutino i clienti a utilizzare le risorse cloud in modo efficiente, al fine di soddisfare i requisiti aziendali e mantenere tale efficienza man mano che la domanda cambia e le tecnologie si evolvono.

Introduzione

Al Framework AWS Well-Architected aiuta a comprendere i pro e i contro delle decisioni prese durante la creazione dei carichi di lavoro 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. Il canone permette di misurare in modo coerente le architetture secondo le best practice e di identificare le aree da migliorare. Disporre di carichi di lavoro ben architettati aumenta notevolmente la probabilità di successo aziendale.

Il Canone si basa su sei principi:

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

Questo whitepaper tratta dell'applicazione del principio dell'efficienza delle prestazioni ai carichi di lavoro. Nei tradizionali ambienti in locale, raggiungere prestazioni durature e di alto livello può essere difficoltoso. L'utilizzo dei principi contenuti in questo documento ti aiuterà a creare architetture in AWS in grado di offrire prestazioni efficienti e sostenibili nel tempo. Le linee guida e le best practice

contenute in questo documento sono suddivise in cinque aree di interesse chiave che operano da principi guida per la creazione di soluzioni cloud in AWS efficienti in termini di prestazioni. Queste aree di interesse sono:

- Scelta dell'architettura
- Calcolo e hardware
- Gestione dati
- Reti e distribuzione di contenuti
- Processo e cultura

Questo documento è rivolto a chi ricopre ruoli nell'ambito della tecnologia, ad esempio ai Chief Technology Officer (CTO), ai progettisti, agli sviluppatori e ai membri dei team operativi. Dopo avere letto questo documento, comprenderai le best practice di AWS e le strategie da utilizzare durante la progettazione di architetture di un ambiente cloud dalle prestazioni elevate.

Efficienza delle prestazioni

Il principio dell'efficienza delle prestazioni si concentra sull'utilizzo efficiente delle risorse di elaborazione per soddisfare i requisiti, e sulla modalità di mantenimento di tale efficienza all'evolversi delle esigenze e delle tecnologie.

Argomenti

Principi di progettazione

I seguenti principi di progettazione possono aiutarti a raggiungere e mantenere carichi di lavoro efficienti 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: Implementare 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 approccio orientato alla meccanica: Utilizza l'approccio tecnologico che meglio si allinea ai tuoi obiettivi. Ad esempio, prendi in considerazione gli schemi di accesso ai dati quando scegli una strategia basata su database o archiviazione per il tuo carico di lavoro.

Definizione

Concentrati sulle seguenti aree per ottenere l'efficienza delle prestazioni nel cloud:

- Scelta dell'architettura
- Calcolo e hardware
- Gestione dati
- Reti e distribuzione di contenuti
- Processo e cultura

Adotta un approccio basato sui dati per creare un'architettura ad alte prestazioni. 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 del cloud AWS. 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.

Scelta dell'architettura

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 diverse configurazioni e tipologie, 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.

Questa area di interesse offre linee guida e best practice su come selezionare risorse cloud e modelli di architettura efficienti e ad alte prestazioni.

Best practice

- PERF01-BP01 Informazioni e identificazione dei servizi e delle funzionalità cloud disponibili
- PERF01-BP02 Utilizzo delle indicazioni del provider cloud o di un partner appropriato per conoscere gli schemi di architettura e le best practice
- PERF01-BP03 Influenza dei costi nelle decisioni sull'architettura
- PERF01-BP04 Valutazione dell'influenza dei compromessi sui clienti e sull'efficienza dell'architettura
- PERF01-BP05 Uso delle policy e delle architetture di riferimento
- PERF01-BP06 Uso del benchmarking per guidare le decisioni sull'architettura
- PERF01-BP07 Uso di un approccio basato sui dati per le scelte dell'architettura

PERF01-BP01 Informazioni e identificazione dei servizi e delle funzionalità cloud disponibili

Informati continuamente e identifica i servizi e le configurazioni disponibili che ti aiutano a prendere le decisioni giuste sull'architettura e a migliorare l'efficienza delle prestazioni dei carichi di lavoro.

Anti-pattern comuni:

- Utilizzi il cloud come data center in co-location.

- Non stai modernizzando la tua applicazione con la migrazione al cloud.
- Stai solo usando un tipo di archiviazione per tutte le cose che devono essere conservate in modo persistente.
- 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 nuovi servizi e configurazioni, puoi migliorare notevolmente le prestazioni, ridurre i costi e ottimizzare le attività necessarie per mantenere il carico di lavoro. Puoi anche accelerare il time-to-value per i prodotti abilitati al cloud.

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

Guida all'implementazione

AWS rilascia continuamente nuovi servizi e funzionalità in grado di migliorare le prestazioni e ridurre i costi dei carichi di lavoro del cloud. Rimanere aggiornati su questi nuovi servizi e funzionalità è fondamentale per mantenere l'efficacia delle prestazioni nel cloud. La modernizzazione dell'architettura dei carichi di lavoro consente inoltre di accelerare la produttività, promuovere l'innovazione e sbloccare ulteriori opportunità di crescita.

Passaggi dell'implementazione

- Esegui l'inventario del software e dell'architettura del carico di lavoro per i servizi correlati. Determina su quale categoria di prodotti ottenere ulteriori informazioni.
- Esplora le offerte AWS per individuare e conoscere i servizi e le opzioni di configurazione pertinenti che possono aiutarti a migliorare le prestazioni e ridurre i costi e la complessità operativa.
  - Cloud Amazon Web Services
  - Academy AWS
  - Novità di AWS
  - Blog AWS
  - AWS Skill Builder
  - Eventi e webinar AWS
  - AWS Training e certificazioni
  - Canale YouTube di AWS

  - Workshop di AWS
  - Community AWS
- Usa gli ambienti sandbox non di produzione per comprendere e sperimentare nuovi servizi senza incorrere in costi aggiuntivi.
- Scopri servizi e funzionalità cloud sempre nuovi.

Risorse

Documenti correlati:

- Panoramica di Amazon Web Services
- Caratteristiche di Amazon EC2
- Impara passo per passo con il Programma di apprendimento dei Partner AWS
- Formazione e certificazione AWS
- My learning path to become an AWS solutions architect
- Centro di progettazione AWS
- AWS Partner Network
- Portfolio di soluzioni AWS
- Knowledge Center di AWS
- Costruisci applicazioni moderne su AWS

Video correlati:

- AWS re:Invent 2023 - What's new with Amazon EC2
- AWS re:Invent 2022 - Reduce your operational and infrastructure costs with Amazon ECS
- AWS re:Invent 2023 - Build with the efficiency, agility & innovation of the cloud with AWS
- AWS re:Invent 2022 - Deploy ML models for inference at high performance and low cost
- La mia architettura

Esempi correlati:

- Esempi di AWS
- Esempi di SDK AWS

PERF01-BP02 Utilizzo delle indicazioni del provider cloud o di un partner appropriato per conoscere gli schemi di architettura e le best practice

Usa le risorse aziendali del cloud come documentazione, solutions architect, servizi professionali o partner appropriati per guidare le tue decisioni sull'architettura. Queste risorse ti aiutano a rivedere e migliorare l'architettura per ottenere prestazioni ottimali.

Anti-pattern comuni:

- AWS è usato come un comune provider di servizi cloud.
- I servizi AWS vengono utilizzati in modo diverso rispetto alla loro progettazione iniziale.
- Le indicazioni vengono seguite senza considerare il contesto aziendale.

Vantaggi dell'adozione di questa best practice: avvalersi della guida di un provider di servizi cloud o di un partner appropriato può aiutarti a fare le scelte giuste per l'architettura del tuo carico di lavoro e darti fiducia nelle tue decisioni.

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

Guida all'implementazione

AWS offre un'ampia scelta di linee guida, documentazione e risorse che possono aiutarti a creare e gestire i carichi di lavoro del cloud in modo efficiente. La documentazione AWS fornisce esempi di codice, esercitazioni e spiegazioni dettagliate sui servizi. Oltre alla documentazione, AWS offre programmi di formazione e certificazione, solutions architect e servizi professionali che i clienti possono usare per esplorare diversi aspetti dei servizi cloud e implementare un'architettura cloud efficiente su AWS.

Sfrutta queste risorse per ottenere approfondimenti sulle informazioni e sulle best practice preziose per risparmiare tempo e ottenere risultati migliori nel Cloud AWS.

Passaggi dell'implementazione

- Consulta la documentazione e le linee guida AWS e segui le best practice. Queste risorse possono aiutarti a scegliere e configurare i servizi in modo efficace e a ottenere prestazioni migliori.
  - Documentazione di AWS (come guide utente e white paper)

  - Blog AWS
  - AWS Training e certificazioni
  - Canale YouTube di AWS
- Partecipa agli eventi per i partner AWS (come summit AWS a livello mondiale, gruppi di utenti di AWS re:Invent e workshop) per apprendere dagli esperti AWS le best practice per l'utilizzo dei servizi AWS.
  - Impara passo per passo con il Programma di apprendimento dei Partner AWS
  - Eventi e webinar AWS
  - Workshop di AWS
  - Community AWS
- 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.
- utilizza AWS Support se hai bisogno di supporto tecnico per utilizzare un servizio in modo efficace. I nostri piani di assistenza sono pensati per offrirti il giusto mix di strumenti e competenze in modo da poter conseguire il successo con AWS ottimizzando le prestazioni, gestendo i rischi e tenendo sotto controllo i costi.

Risorse

Documenti correlati:

- Centro di progettazione AWS
- AWS Partner Network
- Portfolio di soluzioni AWS
- Knowledge Center di AWS
- Supporto Enterprise di AWS

Video correlati:

- La mia architettura
- AWS re:Invent 2023 - Advanced event-driven patterns with Amazon EventBridge

- AWS re:Invent 2023 - Implementing distributed design patterns on AWS
- AWS re:Invent 2023 - Application architecture as code

Esempi correlati:

- Esempi di AWS
- Esempi di SDK AWS
- AWS Analytics Reference Architecture

PERF01-BP03 Influenza dei costi nelle decisioni sull'architettura

Tieni conto dei costi nelle decisioni sull'architettura per migliorare l'utilizzo delle risorse e l'efficienza delle prestazioni del tuo carico di lavoro cloud. Quando si è consapevoli delle implicazioni dei costi del carico di lavoro cloud, è più probabile che si utilizzino risorse efficienti e si riducano le procedure inutili.

Anti-pattern comuni:

- Utilizzi una sola famiglia di istanze.
- Ometti di valutare le soluzioni con licenza rispetto alle soluzioni open-source.
- Non definisci le policy del ciclo di vita dell'archiviazione.
- Non esamini i nuovi servizi e funzionalità del Cloud AWS.
- Utilizzi solo lo storage a blocchi.

Vantaggi dell'adozione di questa best practice: La contabilizzazione dei costi nel processo decisionale consente di utilizzare risorse più efficienti ed esplorare altri investimenti.

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

Guida all'implementazione

L'ottimizzazione dei carichi di lavoro in base ai costi può migliorare l'utilizzo delle risorse ed evitare sprechi nel carico di lavoro cloud. Tenere conto dei costi nelle decisioni sull'architettura di solito include il corretto dimensionamento dei componenti del carico di lavoro e l'abilitazione dell'elasticità, comportando una migliore efficienza delle prestazioni del carico di lavoro cloud.

Passaggi dell'implementazione

- Stabilisci gli obiettivi di costo, come i limiti del budget, per il tuo carico di lavoro cloud.
- Identifica i componenti chiave, come istanze e archiviazione, che determinano il costo del carico di lavoro. Puoi utilizzare AWS Pricing Calculator e AWS Cost Explorer per identificare i principali fattori di costo del carico di lavoro.
- Comprensione dei modelli di prezzo nel cloud, ad esempio istanze on-demand, riservate, Savings Plans e spot.
- Utilizza Migliori pratiche di ottimizzazione dei costi di Well-Architected per ottimizzare questi componenti chiave in termini di costi.
- Monitora e analizza continuamente i costi per identificare le opportunità di ottimizzazione dei costi nel tuo carico di lavoro.
  - utilizza Budget AWS per ricevere gli avvisi per i costi inaccettabili.
  - utilizza AWS Compute Optimizer oppure AWS Trusted Advisor per ottenere suggerimenti sull'ottimizzazione dei costi.
  - utilizza Rilevamento delle anomalie dei costi AWS per rilevare in modo automatico le anomalie dei costi e analizzare la causa principale.

Risorse

Documenti correlati:

- Che cos'è la Gestione costi e fatturazione AWS?
- Ottimizzazione dei costi con AWS
- Choosing an AWS cost management strategy
- A Beginner’s Guide to AWS Cost Management
- A Detailed Overview of the Cost Intelligence Dashboard
- Centro di progettazione AWS
- Portfolio di soluzioni AWS
- Knowledge Center di AWS

Video correlati:

- La mia architettura

- AWS re:Invent 2023 - What's new with AWS cost optimization
- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2023 - AWS storage cost-optimization best practices
- AWS re:Invent 2023 - Optimize costs in your multi-account environments

Esempi correlati:

- AWS Compute Optimizer Demo code (Codice dimostrativo di AWS Compute Optimizer)
- Cost Optimization Workshop
- Cloud Financial Management Technical Implementation Playbooks
- Startup optimization: Tuning application performance for maximum efficiency
- Serverless Optimization Workshop (Performance and Cost)
- Scaling cost effective architectures

PERF01-BP04 Valutazione dell'influenza dei compromessi sui clienti e sull'efficienza dell'architettura

Quando valuti i miglioramenti correlati alle prestazioni, determina quali scelte hanno 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 finale intrinseca di questo cambiamento avrà un impatto sui clienti.

Anti-pattern comuni:

- Ritieni che tutti i vantaggi prestazionali debbano essere implementati, anche se ci sono compromessi per l'implementazione.
- 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 si valutano potenziali miglioramenti relativi alle prestazioni, è necessario decidere se i compromessi per le modifiche sono accettabili 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: alto

Guida all'implementazione

Identifica le aree critiche della tua architettura in termini di prestazioni e impatto sui clienti. 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.

Passaggi dell'implementazione

- Comprendi i requisiti del tuo carico di lavoro e gli accordi sul livello di servizio (SLA).
- Definisci chiaramente i fattori di valutazione. I fattori possono riguardare il costo, l'affidabilità, la sicurezza e le prestazioni del carico di lavoro.
- Seleziona l'architettura e i servizi in grado di soddisfare le tue esigenze.
- Conduci sperimentazioni e proof of concept (POC) per valutare i fattori di compromesso, l'impatto sui clienti e l'efficienza dell'architettura. Di solito, i carichi di lavoro altamente disponibili, performanti e sicuri consumano più risorse cloud offrendo al contempo una esperienza cliente migliore. Comprendi i compromessi in termini di complessità, prestazioni e costi del tuo carico di lavoro. In genere, dare la priorità a due fattori va a scapito del terzo.

Risorse

Documenti correlati:

- Amazon Builders' Library
- Amazon QuickSight KPIs (KPI di Amazon QuickSight)
- RUM Amazon CloudWatch
- X-Ray Documentation (Documentazione di X-Ray)
- Understand resiliency patterns and trade-offs to architect efficiently in the cloud

Video correlati:

- Optimize applications through Amazon CloudWatch RUM (Ottimizzazione delle applicazioni tramite RUM Amazon CloudWatch)
- AWS re:Invent 2023 - Capacity, availability, cost efficiency: Pick three

- AWS re:Invent 2023 - Advanced integration patterns & trade-offs for loosely coupled systems

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)

PERF01-BP05 Uso delle policy e delle architetture di riferimento

Utilizza le policy interne e le architetture di riferimento esistenti per la selezione dei servizi e delle configurazioni per essere più efficiente nella progettazione e nell'implementazione del carico di lavoro.

Anti-pattern comuni:

- Usi una vasta gamma di tecnologie che possono influire sul sovraccarico 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

Avere politiche interne nella selezione delle risorse e dell'architettura fornisce standard e linee guida da seguire quando si effettuano scelte architettoniche. Queste linee guida semplificano il processo decisionale nella scelta del servizio cloud giusto e possono contribuire a migliorare l'efficienza delle prestazioni. Distribuisci il carico di lavoro utilizzando policy o architetture di riferimento. Integra i servizi nell'implementazione cloud, quindi utilizza i test delle prestazioni per verificare che i requisiti prestazionali siano sempre rispettati.

Passaggi dell'implementazione

- Comprendi chiaramente i requisiti del tuo carico di lavoro cloud.
- Rivedi le policy interne ed esterne per identificare quelle più pertinenti.

- Utilizza le architetture di riferimento appropriate fornite dalle best practice AWS o di settore.
- Crea un contesto composto da policy, standard, architetture di riferimento e linee guida prescrittive per situazioni comuni. In questo modo i tuoi team possono muoversi più velocemente. Personalizza le risorse per il tuo settore verticale, se applicabile.
- Convalida queste policy e architetture di riferimento per il tuo carico di lavoro in ambienti di sperimentazione (sandbox).
- Rimani aggiornato con gli standard e gli aggiornamenti AWS del settore per assicurarti che le tue policy e le architetture di riferimento ottimizzino il carico di lavoro cloud.

Risorse

Documenti correlati:

- Centro di progettazione AWS
- AWS Partner Network
- Portfolio di soluzioni AWS
- Knowledge Center di AWS
- AWS Architecture Blog

Video correlati:

- La mia architettura
- AWS re:Invent 2022 - Accelerate value for your business with SAP & AWS reference architecture

Esempi correlati:

- Esempi di AWS
- Esempi di SDK AWS

PERF01-BP06 Uso del benchmarking per guidare le decisioni sull'architettura

Esegui il benchmark delle prestazioni di un carico di lavoro esistente per comprendere le prestazioni sul cloud e guidare le decisioni sull'architettura basate sui dati.

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'attuale implementazione consente di misurare i miglioramenti delle prestazioni.

Livello di rischio associato alla mancata adozione di questa best practice: medio

Guida all'implementazione

Utilizza test sintetici di benchmarking per valutare le 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 test di benchmarking personalizzati oppure utilizzare i test standard del settore, come TPC-DS, per effettuare un'analisi comparativa dei carichi di lavoro. 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. Dovrai eseguire lo stesso benchmark più volte, per verificare 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.

Passaggi dell'implementazione

- Pianifica e definisci:

    - Definisci gli obiettivi, la baseline, gli scenari di test, le metriche, ad esempio l'utilizzo della CPU, la latenza o il throughput, e i KPI per il tuo benchmark.
    - Concentrati sui requisiti degli utenti in termini di esperienza utente e su fattori come i tempi di risposta e l'accessibilità.
    - Individua uno strumento di benchmark adatto al tuo carico di lavoro. Puoi utilizzare i servizi AWS, come Amazon CloudWatch, o uno strumento di terze parti compatibile con il carico di lavoro.
- Configura ed esegui l'instrumentazione:
    - Imposta il tuo ambiente e configura le risorse.
    - Implementa il monitoraggio e la registrazione per acquisire i risultati dei test.
- Esegui i test di benchmark e monitora:
    - Esegui i test di benchmark e monitora i parametri durante il test.
- Analizza e documenta:
    - Documenta il processo di benchmark e gli esiti.
    - Analizza i risultati per identificare i colli di bottiglia, le tendenze e le aree di miglioramento.
    - Usa i risultati dei test per prendere decisioni sull'architettura e modificare il carico di lavoro. Questa operazione può includere la modifica dei servizi o l'adozione di nuove funzionalità.
- Ottimizza e ripeti:
    - Modifica le configurazioni e le allocazioni delle risorse in base ai tuoi benchmark.
    - Ripeti il test del carico di lavoro dopo i cambiamenti per convalidare i miglioramenti.
    - Documenta le informazioni e ripeti il processo per identificare altre aree di miglioramento.

Risorse

Documenti correlati:

- Centro di architettura AWS
- AWS Partner Network
- Biblioteca di soluzioni AWS
- Centro conoscenze AWS
- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Genomics workflows, Part 5: automated benchmarking

- Benchmark and optimize endpoint deployment in Amazon SageMaker JumpStart

Video correlati:

- AWS re:Invent 2023 - Benchmarking AWS Lambda cold starts
- Benchmarking stateful services in the cloud
- La mia architettura
- Optimize applications through Amazon CloudWatch RUM
- Demo di Amazon CloudWatch Synthetics

Esempi correlati:

- Esempi AWS
- Esempi di SDKAWS
- Test di carico distribuito
- Measure page load time with Amazon CloudWatch Synthetics
- Amazon CloudWatch RUM Web Client

PERF01-BP07 Uso di un approccio basato sui dati per le scelte dell'architettura

Definisci un approccio chiaro e basato sui dati per le scelte dell'architettura e verificare che vengano utilizzati i servizi e le configurazioni cloud corretti per soddisfare le tue esigenze aziendali specifiche.

Anti-pattern comuni:

- Ritieni che l'architettura corrente diventi statica e non venga aggiornata nel corso del tempo.
- Le tue scelte dell'architettura si basano su ipotesi e presupposizioni.
- Introduci modifiche all'architettura nel tempo senza giustificazioni.

Vantaggi dell'adozione di questa best practice: Con un approccio ben definito per le scelte dell'architettura, utilizzi i dati per influenzare la progettazione del carico di lavoro e prendere decisioni informate nel tempo.

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

Guida all'implementazione

Affidati all'esperienza e alle competenze interne in materia di cloud o utilizza risorse esterne, come casi d'uso pubblicati o whitepaper, per scegliere risorse e servizi per la tua architettura. È necessario definire con cura un processo che incoraggi la sperimentazione e il benchmarking con i servizi che possono essere utilizzati nel carico di lavoro.

I backlog dei carichi di lavoro critici devono consistere non solo in storie che offrono funzionalità rilevanti per l'azienda e gli utenti, ma anche in storie tecniche che definiscono la presentazione dell'architettura per il carico di lavoro. Questa presentazione include i nuovi progressi tecnologici e i nuovi servizi e li adotta sulla base di dati e giustificazioni adeguate. Verifica che l'architettura sia a prova di futuro e non diventi obsoleta.

Passaggi dell'implementazione

- Interagisci con i principali stakeholder per definire i requisiti del carico di lavoro, comprese le prestazioni, la disponibilità e le considerazioni sui costi. Includi fattori quali il numero di utenti e il modello di utilizzo del tuo carico di lavoro.
- Crea una presentazione dell'architettura o un backlog tecnologico a cui venga assegnata la priorità insieme al backlog funzionale.
- Valuta e identifica i diversi servizi cloud (per maggiori dettagli, consulta PERF01-BP01 Informazioni e identificazione dei servizi e delle funzionalità cloud disponibili).
- Esplora i diversi modelli di architettura, come microservizi o serverless, che soddisfano i tuoi requisiti di prestazioni (per maggiori dettagli, consulta PERF01-BP02 Utilizzo delle indicazioni del provider cloud o di un partner appropriato per conoscere gli schemi di architettura e le best practice).

- Consulta altri team, diagrammi di architettura e risorse, come AWS Solution Architects, Centro di progettazione AWSe AWS Partner Network, per aiutarti a scegliere l'architettura giusta per il tuo carico di lavoro.

- Definisci i parametri, come la velocità di trasmissione effettiva e il tempo di risposta, che possono aiutarti a valutare le prestazioni del tuo carico di lavoro.
- Sperimenta e utilizza i parametri definiti per convalidare le prestazioni dell'architettura selezionata.

- Monitora continuamente e apporta le modifiche necessarie per mantenere ottimali le prestazioni della tua architettura.
- Documenta l'architettura e le decisioni selezionate come riferimento per aggiornamenti e apprendimenti futuri.
- Rivedi e aggiorna continuamente l'approccio di selezione dell'architettura in base agli apprendimenti, alle nuove tecnologie e ai parametri che indicano un problema o un cambiamento necessario nell'approccio attuale.

Risorse

Documenti correlati:

- Portfolio di soluzioni AWS
- Knowledge Center di AWS
- Architectural Patterns to Build End-to-End Data Driven Applications on AWS

Video correlati:

- La mia architettura
- AWS re:Invent 2021 - Data-driven enterprise: Going from vision to value
- AWS re:Invent 2022 - Delivering sustainable, high-performing architectures
- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results

Esempi correlati:

- Esempi di AWS
- Esempi di SDK AWS

Calcolo e hardware

La soluzione ottimale di elaborazione per un determinato sistema potrebbe variare in base alla progettazione dell'applicazione, ai modelli di utilizzo e alle impostazioni di configurazione. Le architetture possono utilizzare diverse soluzioni di calcolo per vari componenti e impiegare funzionalità diverse per migliorare le prestazioni. Selezionare la soluzione di elaborazione sbagliata per un'architettura può ridurre l'efficienza delle prestazioni.

Questa area di interesse offre linee guida e best practice su come identificare e ottimizzare le opzioni di calcolo per ottenere prestazioni di calcolo nel cloud efficienti.

Best practice

- PERF02-BP01 Selezione delle migliori opzioni di elaborazione per il carico di lavoro
- PERF02-BP02 Identificazione delle funzionalità e configurazione di calcolo disponibili
- PERF02-BP03 Raccolta dei parametri relativi al calcolo
- PERF02-BP04 Configurazione e dimensionamento corretto delle risorse di elaborazione
- PERF02-BP05 Dimensionamento dinamico delle risorse di elaborazione
- PERF02-BP06 Uso di acceleratori di elaborazione ottimizzati basati su hardware

PERF02-BP01 Selezione delle migliori opzioni di elaborazione per il carico di lavoro

La selezione dell'opzione di elaborazione più appropriata per il carico di lavoro consente di migliorare le prestazioni, ridurre i costi non necessari dell'infrastruttura e diminuire le attività operative richieste per mantenere il carico di lavoro.

Anti-pattern comuni:

- Si utilizza la stessa opzione di elaborazione utilizzata in locale.
- Non si conoscono le opzioni, le funzionalità e le soluzioni di cloud computing e come queste migliorino le prestazioni di elaborazione.
- Si dimensiona in eccesso l'opzione di elaborazione per soddisfare i requisiti di dimensionamento o prestazioni, quando il passaggio a una nuova opzione di elaborazione soddisferebbe le caratteristiche del carico di lavoro in modo più preciso.

Vantaggi dell'adozione di questa best practice: identificando i requisiti di elaborazione e valutando le opzioni disponibili è possibile rendere il carico di lavoro più efficiente in termini di risorse.

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

Guida all'implementazione

Per ottimizzare i carichi di lavoro cloud e ottenere prestazioni efficienti, è importante selezionare le opzioni di elaborazione più appropriate per il tuo caso d'uso e i requisiti di prestazioni. AWS offre una varietà di opzioni di elaborazione che soddisfano diversi carichi di lavoro nel cloud. Ad esempio, è possibile utilizzare Amazon EC2 per avviare e gestire server virtuali, AWS Lambda per eseguire codice senza dover effettuare il provisioning o gestire server, Amazon ECS o Amazon EKS per eseguire e gestire container oppure AWS Batch per elaborare grandi volumi di dati in parallelo. In base alle tue esigenze di dimensionamento ed elaborazione, scegli e configura la soluzione di elaborazione ottimale per la tua situazione. Puoi anche prendere in considerazione l'utilizzo di più tipi di soluzioni di elaborazione in un unico carico di lavoro in quanto ognuna ha i suoi vantaggi e svantaggi.

I passaggi seguenti ti guidano nella selezione delle opzioni di elaborazione giuste per soddisfare le caratteristiche del carico di lavoro e i requisiti prestazionali.

Passaggi dell'implementazione

- Comprendi i requisiti di elaborazione del tuo carico di lavoro. I requisiti essenziali da considerare includono le esigenze di elaborazione, gli schemi di traffico, gli schemi di accesso ai dati, le esigenze di dimensionamento e i requisiti di latenza.
- Scopri le diverse opzioni di elaborazione disponibili per il tuo carico di lavoro in AWS (come descritto in PERF01-BP01 Informazioni e identificazione dei servizi e delle funzionalità cloud disponibili. Ecco alcune importanti opzioni di elaborazione AWS, le caratteristiche e i casi d'uso più comuni:

AWS service	Key characteristics	Common use cases
Amazon Elastic Compute Cloud (Amazon EC2)	Has dedicated option for hardware, license requirements, large selection of different instance families, processor	Lift and shift migrations, monolithic application, hybrid environments, enterprise applications

AWS service	Key characteristics	Common use cases
	types and compute accelerators	
Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	Easy deployment, consistent environments, scalable	Microservices, hybrid environments
AWS Lambda	Elaborazione serverless service that runs code in response to events and automatically manages the underlying compute resources.	Microservices, event-driven applications
AWS Batch	Efficiently and dynamically provisions and scales Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS), and AWS Fargate compute resources, with an option to use On-Demand or Spot Instances based on your job requirements	HPC, train ML models
Amazon Lightsail	Preconfigured Linux and Windows application for running small workloads	Simple web applications, custom website

- Valuta i costi (come la tariffa oraria o il trasferimento dei dati) e il sovraccarico di gestione (come l'applicazione di patch e il dimensionamento) associati a ciascuna opzione di elaborazione.
- Esegui esperimenti e benchmarking in un ambiente non di produzione per identificare quale opzione di elaborazione può soddisfare al meglio i requisiti del tuo carico di lavoro.

- Dopo aver sperimentato e identificato la tua nuova soluzione di calcolo, pianifica la migrazione e convalida i parametri prestazionali.
- Utilizza gli strumenti di monitoraggio AWS come Amazon CloudWatch e i servizi di ottimizzazione come AWS Compute Optimizer per ottimizzare continuamente le risorse di calcolo in base a modelli di utilizzo reali.

Risorse

Documenti correlati:

- Elaborazione nel cloud con AWS
- Tipi di istanza - Amazon EC2
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Amazon ECS Containers: Amazon ECS Container Instances
- Configurazione della funzione Lambda
- Prescriptive Guidance for Containers (Guida prescrittiva per i container)
- Prescriptive Guidance for Serverless (Guida prescrittiva per serverless)

Video correlati:

- AWS re:Invent 2023 - AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 - New Amazon Elastic Compute Cloud generative AI capabilities in AMS
- AWS re:Invent 2023 - What’s new with Amazon Elastic Compute Cloud
- AWS re:Invent 2023 - Smart savings: Amazon Elastic Compute Cloud cost-optimization strategies
- AWS re:Invent 2021 - Powering next-gen Amazon Elastic Compute Cloud: Deep dive on the Nitro System
- AWS re:Invent 2019 - Optimize performance and cost for your AWS compute
- AWS re:Invent 2019 - Amazon Elastic Compute Cloud foundations
- AWS re:Invent 2022 - Deploy ML models for inference at high performance and low cost
- AWS re:Invent 2019 - Optimize performance and cost for your AWS compute
- Amazon EC2 foundations
- Deploy ML models for inference at high performance and low cost

Esempi correlati:

- Migrating the Web application to containers
- Esecuzione di un "Hello, World!" serverless
- Amazon EKS Workshop
- Amazon EC2 Workshop
- Efficient and Resilient Workloads with Amazon Elastic Compute Cloud Auto Scaling
- Migrating to AWS Graviton with Container Services

PERF02-BP02 Identificazione delle funzionalità e configurazione di calcolo disponibili

Comprendi le opzioni e le funzionalità di configurazione disponibili per il tuo servizio di elaborazione in modo da fornire la giusta quantità di risorse e migliorare l'efficienza delle prestazioni.

Anti-pattern comuni:

- Non valuti le opzioni di elaborazione o le famiglie di istanze disponibili rispetto alle caratteristiche del carico di lavoro.
- Esegui un provisioning eccessivo delle risorse di elaborazione per soddisfare i requisiti di picco della domanda.

Vantaggi dell'adozione di questa best practice: acquisisci familiarità con le funzionalità e le configurazioni di elaborazione di AWS in modo da poter utilizzare una soluzione di elaborazione ottimizzata per soddisfare le caratteristiche e le esigenze del carico di lavoro.

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

Guida all'implementazione

Ogni soluzione di elaborazione ha disponibili configurazioni e funzionalità specifiche per supportare caratteristiche e requisiti diversi del carico di lavoro. Scopri in che modo puoi completare al meglio 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. 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 per scoprire se la tua attuale opzione di elaborazione è adatta ai carichi di lavoro dal punto di vista della CPU e della memoria.

Passaggi dell'implementazione

1. Comprendi i requisiti del carico di lavoro, come CPU, memoria e latenza.
2. Consulta la documentazione e le best practice AWS per scoprire le opzioni di configurazione consigliate che possono contribuire a migliorare le prestazioni dell'elaborazione. Ecco alcune opzioni di configurazione chiave da considerare:

Opzione di configurazione	Esempi
Tipo di istanza	• Le istanze ottimizzate per il calcolo sono l'ideale per i carichi di lavoro che richiedono un rapporto vCPU/memoria molto elevato. • Le istanze ottimizzate per la memoria offrono grandi quantità di memoria per carichi di lavoro intensivi in questo senso. • 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.
Modello di prezzi	• 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. • Savings Plans offrono risparmi significativi rispetto alle istanze on demand in cambio dell'impegno a utilizzare una quantità specifica di potenza di elaborazione per un periodo di uno o tre anni.

Opzione di configurazione	Esempi
	• istanze spot consentono di sfruttare la capacità inutilizzata delle istanze con uno sconto per i carichi di lavoro stateless e tolleranti ai guasti.
Auto Scaling	Utilizza Auto Scaling configurazione per abbinare le risorse di elaborazione ai modelli di traffico.
Valutazione	• utilizza Compute Optimizer per ricevere un efficace suggerimento di machine learning riguardo alla configurazione più adatta alle tue caratteristiche di elaborazione. • utilizza AWS Lambda Power Tuning per selezionare la configurazione migliore per la tua funzione Lambda.
Acceleratori di calcolo basati su hardware	• Le istanze a calcolo accelerato eseguono funzioni come l'elaborazione grafica o la corrispondenza di schemi di dati in modo più efficiente rispetto alle alternative basate sulla CPU. • Per i carichi di lavoro di machine learning, sfrutta l'hardware specifico per il tuo carico di lavoro, come ad esempio AWS Trainium, AWS Inferentiae Amazon EC2 DL1

Risorse

Documenti correlati:

- Elaborazione in cloud con AWS
- Tipi di istanza Amazon EC2

- Controllo degli stati del processore dell'istanza Amazon EC2
- Amazon EKS Container: nodi worker di Amazon EKS
- Container Amazon ECS: Istanze di container di Amazon ECS
- Funzioni: configurazione della funzione Lambda

Video correlati:

- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations
- AWS re:Invent 2022 – https://www.youtube.com/watch?v=5B4-s_ivn1o

Esempi correlati:

- Codice dimostrativo di Compute Optimizer
- Workshop sulle istanze spot Amazon EC2
- Efficient and Resilient Workloads with Amazon EC2 AWS Auto Scaling
- Workshop per sviluppatori Graviton
- AWS for Microsoft workloads immersion day
- AWS for Linux workloads immersion day
- AWS Compute Optimizer Demo code (Codice dimostrativo di AWS Compute Optimizer)
- Workshop su Amazon EKS

PERF02-BP03 Raccolta dei parametri relativi al calcolo

Registra e monitora i parametri relativi all'elaborazione per comprendere meglio le prestazioni delle tue risorse di elaborazione e migliorarne le prestazioni e l'utilizzo.

Anti-pattern comuni:

- Utilizzi solo i file di log manuali per la ricerca dei parametri.
- Utilizzi solo i parametri predefiniti registrati dal software di monitoraggio.
- Rivedi i parametri solo quando c'è un problema.

Vantaggi dell'adozione di questa best practice: la raccolta dei parametri relativi alle prestazioni ti aiuta ad allineare le prestazioni delle applicazioni ai requisiti aziendali per garantire il rispetto delle esigenze dei carichi di lavoro. Può anche aiutarti a migliorare costantemente le prestazioni e l'utilizzo delle risorse del tuo carico di lavoro.

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

Guida all'implementazione

I carichi di lavoro del cloud possono generare grandi volumi di dati quali parametri, log ed eventi. Nel Cloud AWS, la raccolta dei parametri è un passaggio cruciale per migliorare la sicurezza, l'efficienza in termini di costi, le prestazioni e la sostenibilità. AWS fornisce un'ampia gamma di parametri relativi alle prestazioni utilizzando servizi di monitoraggio, come Amazon CloudWatch per fornirti approfondimenti preziosi. Parametri quali l'utilizzo della CPU, l'utilizzo della memoria, l'I/O del disco e il traffico di rete in entrata e in uscita possono fornire approfondimenti sui livelli di utilizzo o sui colli di bottiglia delle prestazioni. Utilizza tali parametri come parte di un approccio basato sui dati per ottimizzare e ottimizzare le risorse del tuo carico di lavoro. L'ideale sarebbe raccogliere tutti i parametri relativi alle tue risorse di elaborazione in un'unica piattaforma con policy di conservazione implementate per supportare costi e obiettivi operativi.

Passaggi dell'implementazione

1. Identifica quali parametri relativi alle prestazioni sono rilevanti per il tuo carico di lavoro. Raccogli i parametri sull'utilizzo delle risorse e sul modo in cui opera il tuo carico di lavoro nel cloud (come il tempo di risposta e la velocità di trasmissione effettiva).
   a. Parametri predefiniti di Amazon EC2
   b. Parametri predefiniti di Amazon ECS
   c. Parametri predefiniti di Amazon EKS
   d. Parametri predefiniti di Lambda
   e. Parametri di memoria e del disco di Amazon EC2
2. Scegli e configura la soluzione di registrazione e monitoraggio giusta per il tuo carico di lavoro.
   a. Osservabilità nativa di AWS

   b. AWS Distro for OpenTelemetry
   c. Amazon Managed Service for Prometheus
3. Definisci il filtro e l'aggregazione richiesti per i parametri in base ai requisiti del tuo carico di lavoro.
   a. Quantify custom application metrics with Amazon CloudWatch Logs and metric filters
   b. Collect custom metrics with Amazon CloudWatch strategic tagging
4. Configura le policy di conservazione dei dati per i parametri in modo che corrispondano ai tuoi obiettivi operativi e di sicurezza.
   a. Conservazione dei dati predefinita per i parametri CloudWatch
   b. Conservazione dei dati predefinita per CloudWatch Logs
5. Se necessario, crea allarmi e notifiche per i parametri in modo da rispondere in modo proattivo ai problemi relativi alle prestazioni.
   a. Create alarms for custom metrics using Amazon CloudWatch anomaly detection
   b. Create metrics and alarms for specific web pages with Amazon CloudWatch RUM
6. Usa l'automazione per implementare gli agenti di aggregazione di parametri e log.
   a. Automazione AWS Systems Manager
   b. OpenTelemetry Collector

Risorse

Documenti correlati:

- Monitoraggio e osservabilità
- Best practices: implementing observability with AWS
- 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:

- AWS re:Invent 2023 – [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 – Implementing application observability
- AWS re:Invent 2023 – Building an effective observability strategy
- AWS re:Invent 2023 – Seamless observability with AWS Distro for OpenTelemetry
- Application Performance Management on AWS

Esempi correlati:

- AWS for Linux Workloads Immersion Day- Amazon CloudWatch
- Monitoring Amazon ECS clusters and containers
- Monitoring with Amazon CloudWatch dashboards
- Workshop su Amazon EKS

PERF02-BP04 Configurazione e dimensionamento corretto delle risorse di elaborazione

Configura e dimensiona correttamente le risorse di elaborazione per soddisfare i requisiti di prestazioni del carico di lavoro ed evitare un utilizzo insufficiente o eccessivo delle risorse.

Anti-pattern comuni:

- Ignori i requisiti di prestazioni del carico di lavoro, con il risultato del provisioning eccessivo o insufficiente delle risorse di elaborazione.
- Scegli semplicemente l'istanza più grande o più piccola disponibile per tutti i carichi di lavoro.
- Usi una sola famiglia di istanze per semplificare la gestione.
- Ignori i suggerimenti di AWS Cost Explorer o Compute Optimizer per il corretto dimensionamento.
- Non rivaluti il carico di lavoro in base all'idoneità dei nuovi tipi di istanza.
- Certifichi solo un numero limitato di configurazioni di istanza per l'organizzazione.

Vantaggi dell'adozione di questa best practice: il corretto dimensionamento delle risorse di elaborazione garantisce un funzionamento ottimale nel cloud evitando il provisioning eccessivo o

insufficiente delle risorse. Il corretto dimensionamento delle risorse di elaborazione comporta in genere prestazioni ottimali e una migliore esperienza cliente, riducendo al contempo i costi.

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

Guida all'implementazione

Il dimensionamento corretto consente alle organizzazioni di gestire la propria infrastruttura cloud in modo efficiente ed economico, rispettando al contempo le esigenze aziendali. Un provisioning eccessivo delle risorse cloud può comportare costi aggiuntivi, mentre un provisioning insufficiente può comportare prestazioni scadenti e un'esperienza negativa per il cliente. AWS fornisce strumenti come AWS Compute Optimizer e AWS Trusted Advisor che utilizzano dati storici per fornire consigli per dimensionare correttamente le risorse di elaborazione.

Passaggi dell'implementazione

- Scegli il tipo di istanza più adatto alle tue esigenze:
  - Come faccio a scegliere il tipo di istanza Amazon EC2 appropriato per il mio carico di lavoro?
  - Selezione del tipo di istanza basata sugli attributi per il parco istanze Amazon EC2
  - Create an Auto Scaling group using attribute-based instance type selection
  - Optimizing your Kubernetes compute costs with Karpenter consolidation
- 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 e agli obiettivi di prestazioni.
- Monitora l'utilizzo delle risorse con gli strumenti di monitoraggio di AWS come Amazon CloudWatch.
- Seleziona la configurazione corretta per la risorsa di elaborazione.
  - Per i carichi di lavoro effimeri, valuta le metriche Amazon CloudWatch dell'istanza , ad esempio `CPUUtilization` per identificare se l'istanza è sottoutilizzata o sovrautilizzata.
  - Per i carichi di lavoro stabili, esegui i controlli con gli strumenti di ridimensionamento di AWS, come AWS Compute Optimizer e AWS Trusted Advisor a intervalli regolari per individuare le opportunità di ottimizzazione e ridimensionamento della risorsa di elaborazione.
- Esegui il test delle modifiche apportate alla configurazione in un ambiente non di produzione prima di implementarle in un ambiente live.
- Rivaluta costantemente nuove offerte di elaborazione e confrontale con le esigenze del carico di lavoro.

Risorse

Documenti correlati:

- Elaborazione in cloud con AWS
- Tipi di istanza Amazon EC2
- Container Amazon ECS: Istanze di container di Amazon ECS
- Amazon EKS Container: nodi worker di Amazon EKS
- Funzioni: configurazione della funzione Lambda
- Controllo degli stati del processore dell'istanza Amazon EC2

Video correlati:

- Amazon EC2 foundations
- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 – What’s new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations

Esempi correlati:

- AWS Compute Optimizer Demo code (Codice dimostrativo di AWS Compute Optimizer)
- Workshop su Amazon EKS
- Right-sizing recommendations

PERF02-BP05 Dimensionamento dinamico delle risorse di elaborazione

Sfrutta l'elasticità del cloud per dimensionare dinamicamente le risorse di elaborazione per soddisfare le tue esigenze ed evitare un provisioning eccessivo o insufficiente per il tuo carico di lavoro.

Anti-pattern comuni:

- Risposta agli allarmi aumentando manualmente la capacità.
- Utilizzi le stesse linee guida per il dimensionamento (generalmente infrastruttura statica) di quelle on-premise.
- Mantenimento della maggiore capacità dopo un evento di dimensionamento, senza ripristinare quella originale.

Vantaggi dell'adozione di questa best practice: La configurazione e il test dell'elasticità delle risorse di elaborazione possono aiutarti a risparmiare denaro, mantenere i benchmark delle prestazioni e migliorare l'affidabilità al variare del traffico.

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

Guida all'implementazione

AWS offre la flessibilità necessaria per dimensionare le risorse in modo dinamico attraverso una varietà di meccanismi di dimensionamento per soddisfare le variazioni della domanda. In combinazione con i parametri relativi all'elaborazione, il dimensionamento dinamico consente ai carichi di lavoro di rispondere automaticamente alle modifiche e utilizzare il set ottimale di risorse di elaborazione per raggiungere l'obiettivo.

Puoi adottare varie strategie di approccio per associare l'offerta di risorse alla domanda.

- Approccio al tracciamento degli obiettivi: monitora il parametro di dimensionamento e aumenta o diminuisci automaticamente la capacità in base alle esigenze.
- Dimensionamento predittivo: dimensiona in previsione delle tendenze giornaliere e settimanali.
- Approccio basato sulla pianificazione: imposta il tuo programma di dimensionamento in base alle variazioni di carico prevedibili.
- Scalabilità del servizio: scegli i servizi (come quelli serverless) che si dimensionano automaticamente per progettazione.

Assicurati che le distribuzioni dei carichi di lavoro siano in grado di gestire eventi di dimensionamento.

Passaggi dell'implementazione

- Istanze di elaborazione, container e funzioni forniscono tutti meccanismi di elasticità, in combinazione con il dimensionamento automatico o sotto forma di funzionalità del servizio. Ecco alcuni esempi di meccanismi di dimensionamento automatico:

Meccanismo di scalabilità automatica	Dove usare
Amazon EC2 Auto Scaling	Per assicurarti di avere il numero corretto di Amazon EC2 istanze disponibili per gestire il carico utente per la tua applicazione.
Application Auto Scaling	per dimensionare automaticamente le risorse per servizi AWS diversi da Amazon EC2, ad esempio AWS Lambda funzioni o Amazon Elastic Container Service (Amazon ECS) servizi.
Kubernetes Cluster Autoscaler/Karpenter	Per dimensionare automaticamente i cluster Kubernetes.

- Si parla spesso di dimensionamento con servizi di elaborazione come le istanze Amazon EC2 o le funzioni AWS Lambda. Assicurati di considerare anche la configurazione di servizi non di elaborazione come AWS Glue per soddisfare la domanda.
- Verifica che i parametri per il dimensionamento corrispondano alle caratteristiche del carico di lavoro da implementare. Se distribuisci un'applicazione di transcodifica video, è previsto il 100% di utilizzo della CPU e non deve essere il parametro principale. Utilizza la profondità della coda dei processi di transcodifica. Puoi utilizzare una metrica personalizzata per la tua politica di scalabilità, 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.
- Assicurati di utilizzare 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 entrambi gli eventi di dimensionamento (aumento e riduzione). Ad esempio, puoi usare la cronologia delle attività per verificare un'attività di dimensionamento per un gruppo 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. Con il

dimensionamento predittivo puoi eliminare la necessità di offrire capacità in eccedenza. Per ulteriori informazioni, consulta Dimensionamento predittivo con Amazon EC2 Auto Scaling.

Risorse

Documenti correlati:

- Elaborazione in cloud con AWS
- Tipi di istanza Amazon EC2
- Container Amazon ECS: Istanze di container di Amazon ECS
- Amazon EKS Container: nodi worker di Amazon EKS
- Funzioni: configurazione della funzione Lambda
- Controllo degli stati del processore dell'istanza Amazon EC2
- Deep Dive on Amazon ECS Cluster Auto Scaling
- Introducing Karpenter – An Open-Source High-Performance Kubernetes Cluster Autoscaler

Video correlati:

- AWS re:Invent 2023 – AWS Graviton: The best price performance for your AWS workloads
- AWS re:Invent 2023 – New Amazon EC2 generative AI capabilities in AWS Management Console
- AWS re:Invent 2023 – What's new with Amazon EC2
- AWS re:Invent 2023 – Smart savings: Amazon EC2 cost-optimization strategies
- AWS re:Invent 2021 – Powering next-gen Amazon EC2: Deep dive on the Nitro System
- AWS re:Invent 2019 – Amazon EC2 foundations

Esempi correlati:

- Esempi di gruppo di Amazon EC2 Auto Scaling
- Workshop su Amazon EKS
- Scale your Amazon EKS workloads by running on IPv6

PERF02-BP06 Uso di acceleratori di elaborazione ottimizzati basati su hardware

Usa gli acceleratori hardware per eseguire determinate funzioni in modo più efficiente rispetto alle alternative basate sulla CPU.

Anti-pattern comuni:

- Nel carico di lavoro non hai confrontato un'istanza generica con un'istanza dedicata in grado di offrire prestazioni più elevate e costi inferiori.
- Usi gli acceleratori di calcolo basati su hardware per attività in cui sono più efficienti le alternative basate su CPU.
- Utilizzo delle GPU non monitorato.

Vantaggi dell'adozione di questa best practice: utilizzando gli acceleratori basati su hardware, come le unità di elaborazione grafica (GPU) e gli FPGA (Field Programmable Gate Array), è possibile eseguire determinate funzioni di elaborazione in modo più efficiente.

Livello di rischio associato alla mancata adozione di questa best practice: medio

Guida all'implementazione

Le istanze a calcolo accelerato forniscono l'accesso agli acceleratori di calcolo basati su hardware, come GPU e FPGA. Questi acceleratori hardware eseguono alcune funzioni, come l'elaborazione grafica o la rilevazione della corrispondenza dei modelli di dati, in modo più efficiente rispetto alle alternative basate su CPU. Molti carichi di lavoro accelerati, come il rendering grafico, la transcodifica e il machine learning, sono altamente variabili in termini di utilizzo di risorse. Esegui questo hardware solo per il tempo necessario e disattivalo con l'automazione quando non serve per migliorare l'efficienza complessiva delle prestazioni.

Passaggi dell'implementazione

- Identifica quali istanze a calcolo accelerato possono soddisfare i tuoi requisiti.
- Per i carichi di lavoro di machine learning, sfrutta l'hardware specifico per le tue esigenze, ad esempio AWS Trainium, AWS Inferentia e Amazon EC2 DL1. Le istanze AWS Inferentia come Inf2 offrono fino al 50% in più di prestazioni/watt rispetto alle istanze Amazon EC2 paragonabili.

- Raccogli i parametri di utilizzo delle istanze a calcolo accelerato. Ad esempio, puoi utilizzare l'agente CloudWatch per raccogliere parametri come `utilization_gpu` e `utilization_memory` per le GPU come mostrato in Raccolta dei parametri delle GPU NVIDIA con Amazon CloudWatch.
- Ottimizza il codice, il funzionamento della rete e le impostazioni degli acceleratori hardware per garantire il pieno utilizzo dell'hardware sottostante.
  - Ottimizza l impostazioni delle GPU
  - Monitoraggio e ottimizzazione delle GPU nell'AMI per il Deep Learning
  - Ottimizzazione dell'I/O per la messa a punto delle prestazioni delle GPU dedicate all'addestramento del deep learning in Amazon SageMaker
- Utilizzate le librerie e i driver per GPU più recenti e performanti.
- Utilizza l'automazione per rilasciare le istanze GPU non in uso.

Risorse

Documenti correlati:

- Working with GPUs on Amazon Elastic Container Service
- Istanze GPU
- Istanze con AWS Trainium
- Istanze con AWS Inferentia
- Let’s Architect! Architecting with custom chips and accelerators

- Calcolo accelerato
- Istanze Amazon EC2 VT1
- How do I choose the appropriate Amazon EC2 instance type for my workload?
- Choose the best AI accelerator and model compilation for computer vision inference with Amazon SageMaker

Video correlati:

- AWS re:Invent 2021 - How to select Amazon Elastic Compute Cloud GPU instances for deep learning

- AWS re:Invent 2022 - [NEW LAUNCH!] Introducing AWS Inferentia2-based Amazon EC2 Inf2 instances
- AWS re:Invent 2022 - Accelerate deep learning and innovate faster with AWS Trainium
- AWS re:Invent 2022 - Deep learning on AWS with NVIDIA: From training to deployment

Esempi correlati:

- Amazon SageMaker and NVIDIA GPU Cloud (NGC)
- Use SageMaker with Trainium and Inferentia for optimized deep learning training and inferencing workloads
- Optimizing NLP models with Amazon Elastic Compute Cloud Inf1 instances in Amazon SageMaker

Gestione dati

La soluzione ottimale per la gestione dei dati in un sistema specifico varia in base al tipo di dati (blocco, file o oggetto), agli schemi di accesso (casuali o sequenziali), alla velocità di trasmissione effettiva necessaria, alla frequenza di accesso (online, offline, archivio), alla frequenza di aggiornamento (WORM, dinamico) e ai vincoli di disponibilità e durata. I carichi di lavoro Well-Architected utilizzano archivi dati appositamente progettati che impiegano diverse funzionalità per migliorare le prestazioni.

Quest'area di interesse offre linee guida e best practice per ottimizzare l'archiviazione dei dati, i modelli di spostamento e accesso e l'efficienza delle prestazioni dell'archiviazione di dati.

Best practice

- PERF03-BP01 Uso di un archivio dati dedicato che supporta al meglio i requisiti di accesso e archiviazione dei dati
- PERF03-BP02 Valutazione delle opzioni di configurazione disponibili per datastore
- PERF03-BP03 Raccolta e registrazione dei parametri delle prestazioni del datastore
- PERF03-BP04 Implementazione di strategie per migliorare le prestazioni delle query nel datastore
- PERF03-BP05 Implementa modelli di accesso ai dati che utilizzano la memorizzazione nella cache

PERF03-BP01 Uso di un archivio dati dedicato che supporta al meglio i requisiti di accesso e archiviazione dei dati

Comprendi le caratteristiche dei dati (come la condivisione, le dimensioni, la dimensione della cache, gli schemi di accesso, la latenza, la velocità di trasmissione effettiva e la persistenza dei dati) per selezionare i data store (archiviazione o database) dedicati per il tuo carico di lavoro.

Anti-pattern comuni:

- Continui a utilizzare un datastore per via dell'esperienza e delle competenze interne relative a quel particolare tipo di soluzione di database.
- Ritieni che tutti i carichi di lavoro abbiano requisiti di accesso e archiviazione dei dati simili.
- Non hai implementato un catalogo di dati per eseguire l'inventario dei tuoi asset.

Vantaggi dell'adozione di questa best practice: la comprensione delle caratteristiche e dei requisiti dei dati ti consente di determinare la tecnologia di archiviazione più efficiente e performante appropriata per le esigenze del tuo carico di lavoro.

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

Guida all'implementazione

Quando selezioni e implementi l'archiviazione di dati, assicurati che le caratteristiche di query, dimensionamento e archiviazione supportino i requisiti dei dati del carico di lavoro. AWS fornisce numerose tecnologie di database e archiviazione di dati, tra cui archiviazione a blocchi, archiviazione di oggetti, archiviazione di streaming, file system, database relazionali, chiave-valore, di documenti, 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. Comprendendo le caratteristiche e i requisiti dei dati, puoi abbandonare la tecnologia di archiviazione monolitica e gli approcci restrittivi e validi per tutti, per concentrarti sulla gestione dei dati in modo appropriato.

Passaggi dell'implementazione

- Esegui un inventario dei vari tipi di dati esistenti nel tuo carico di lavoro.
- Comprendi e documenta le caratteristiche e i requisiti dei dati, tra cui:
  - Tipo di dati (non strutturati, semi-strutturati, relazionali)
  - Volume e crescita dei dati
  - Durabilità dei dati: persistenti, effimeri, transitori
  - Requisiti ACID (atomicità, coerenza, isolamento, durabilità)
  - Schemi di accesso ai dati (con uso intensivo di lettura o scrittura)
  - Latenza
  - Throughput
  - IOPS (operazioni di input/output al secondo)
  - Periodo di conservazione dei dati
- Scopri i diversi archivi di dati (servizi di database e archiviazione) disponibili per il carico di lavoro AWS che possono soddisfare le caratteristiche dei tuoi dati, come descritto in PERF01-BP01 Informazioni e identificazione dei servizi e delle funzionalità cloud disponibili. Alcuni esempi di tecnologie di archiviazione AWS e delle loro caratteristiche chiave sono:

Tipo	Servizi AWS	Caratteristiche chiave
Object storage	Amazon S3	Unlimited scalability, high availability, and multiple options for accessibility. Transferring and accessing objects in and out of Amazon S3 can use a service, such as Accelerazione del trasferimento or Punti di accesso, to support your location, security needs, and access patterns.
Archiving storage	Amazon S3 Glacier	Built for data archiving.
Streaming storage	Amazon Kinesis Amazon Managed Streaming for Apache Kafka (Amazon MSK)	Efficient ingestion and storage of streaming data.
Shared file system	Amazon Elastic File System (Amazon EFS)	File system montabile a cui è possibile accedere da più tipi di soluzioni di calcolo.

Tipo	Servizi AWS	Caratteristiche chiave
Shared file system	Amazon FSx	Built on the latest AWS compute solutions to support four commonly used file systems: NetApp ONTAP, OpenZFS, Windows File Server, and Lustre. Amazon FSx , la latenza, la velocità di trasmissione effettiva e le operazioni di input/output al secondo (IOPS) vary per file system and should be considered when selecting the right file system for your workload needs.
Block storage	Amazon Elastic Block Store (Amazon EBS)	Scalable, high-performance block-storage service designed for Amazon Elastic Compute Cloud (Amazon EC2). Amazon EBS includes SSD-backed storage for transactional, IOPS-intensive workloads and HDD-backed storage for throughput-intensive workloads.

Tipo	Servizi AWS	Caratteristiche chiave
Relational database	Amazon Aurora, Amazon RDS, Amazon Redshift.	Designed to support ACID (atomicity, consistency, isolation, durability) transactions, and maintain referential integrity and strong data consistency. Many traditional applications, enterprise resource planning (ERP), customer relationship management (CRM), and ecommerce use relational databases to store their data.
Key-value database	Amazon DynamoDB	Optimized for common access patterns, typically to store and retrieve large volumes of data. High-traffic web apps, ecommerce systems, and gaming applications are typical use-cases for key-value databases.
Document database	Amazon DocumentDB	Designed to store semi-structured data as JSON-like documents. These databases help developers build and update applications such as content management, catalogs, and user profiles quickly.

Tipo	Servizi AWS	Caratteristiche chiave
In-memory database	Amazon ElastiCache , Amazon MemoryDB per Redis	Used for applications that require real-time access to data, lowest latency and highest throughput. You may use in-memory databases for application caching, session management, gaming leaderboards, low latency ML feature store, microserv ices messaging system, and a high-throughput streaming mechanism
Graph database	Amazon Neptune	Used for applications that must navigate and query millions of relationships between highly connected graph datasets with milliseco nd latency at large scale. Many companies use graph databases for fraud detection , social networking, and recommendation engines.
Time Series database	Amazon Timestream	Used to efficiently collect, synthesize, and derive insights from data that changes over time. IoT applications, DevOps, and industrial telemetry can utilize time-series databases.

Tipo	Servizi AWS	Caratteristiche chiave
Wide column	Amazon Keyspaces (per Apache Cassandra)	Uses tables, rows, and columns, but unlike a relational database, the names and format of the columns can vary from row to row in the same table. You typically see a wide column store in high scale industrial apps for equipment maintenance, fleet management, and route optimization.
Ledger	Amazon Quantum Ledger Database (Amazon QLDB)	Provides a centralized and trusted authority to maintain a scalable, immutable, and cryptographically verifiabl e record of transactions for every application. We see ledger databases used for systems of record, supply chain, registrations, and even banking transactions.

- Se stai creando una piattaforma dati, sfrutta la moderna architettura dei dati AWS per integrare data lake, data warehouse e archivi di dati dedicati.
- Le domande chiave da porsi quando si sceglie un data store per il carico di lavoro sono le seguenti:

Question	Things to consider
How is the data structured?	• Se i dati non sono strutturati, prendi in considerazione un archivio di oggetti come Amazon S3 o un database NoSQL come Amazon DocumentDB

Question	Things to consider
	• Per i dati chiave-valore, valuta DynamoDB, Amazon ElastiCache for Redis o Amazon MemoryDB for Redis
What level of referential integrity is required?	• Per i vincoli di chiave esterna, i database relazionali come Amazon RDS e Aurora possono fornire questo livello di integrità. • 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.
Is ACID (atomicity, consistency, isolation, durability) compliance required?	• Se sono necessarie proprietà ACID associate ai database relazionali, valuta un database relazionale come Amazon RDS e Aurora. • Se è necessaria un'elevata consistenza per il database NoSQL, puoi utilizzare l'elevata consistenza di lettura di DynamoDB.
How will the storage requirements change over time? How does this impact scalability?	• Database serverless come DynamoDB e Amazon Quantum Ledger Database (Amazon QLDB) possono dimensionarsi dinamicamente. • Per i database relazionali sono previsti limiti massimi per l'archiviazione assegnata, al raggiungimento dei quali si rende spesso necessario partizionare orizzontalmente tali database tramite meccanismi quali lo sharding.

Question	Things to consider
What is the proportion of read queries in relation to write queries? Would caching be likely to improve performance?	• I carichi di lavoro con molte operazioni di lettura possono trarre vantaggio da un livello di caching, ad esempio ElastiCache o DAX se il database è DynamoDB. • È anche possibile passare le operazioni di lettura alle repliche di lettura con database relazionali come Amazon RDS.
Does storage and modification (OLTP - Online Transaction Processing) or retrieval and reporting (OLAP - Online Analytical Processing) have a higher priority?	• Per un'elaborazione transazionale letta così com'è ad alta velocità di trasmissione effettiva, prendi in considerazione un database NoSQL come DynamoDB. • Per schemi di lettura complessi con velocità di trasmissione effettiva elevata (come il join) con un uso coerente di Amazon RDS. • Per le query analitiche, prendi in considerazione un database colonnare come Amazon Redshift o l'esportazione dei dati in Amazon S3 e l'esecuzione di analisi utilizzando Athena o Amazon QuickSight.

Question	Things to consider
What level of durability does the data require?	• 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. • DynamoDB viene automaticamente replicato in più zone di disponibilità per offrire livelli elevati di disponibilità e durabilità dei dati. • 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.
Is there a desire to move away from commercial database engines or licensing costs?	• Valuta motori open-source come PostgreSQL e MySQL su Amazon RDS o Aurora. • Usa AWS Database Migration Service e AWS Schema Conversion Tool per eseguire le migrazioni dai motori di database commerciali a quelli open-source.
What is the operational expectation for the database? Is moving to managed services a primary concern?	• Utilizzare Amazon RDS, invece di Amazon EC2, e scegliere DynamoDB o Amazon DocumentDB anziché ospitare in autonomia un database NoSQL, riduce le spese operative.

Question	Things to consider
How is the database currently accessed? Is it only application access, or are there business intelligence (BI) users and other connected off-the-shelf applications?	• Se fossero presenti dipendenze verso altri strumenti esterni, potresti dover mantenere la compatibilità con i database che essi supportano. Amazon RDS è completam ente compatibile con le diverse versioni dei motori che supporta, compresi Microsoft SQL Server, Oracle, MySQL e PostgreSQL.

- Esegui esperimenti e benchmarking in un ambiente non di produzione per identificare quale datastore può soddisfare al meglio i requisiti del tuo carico di lavoro.

Risorse

Documenti correlati:

- Tipi di volume Amazon EBS
- Archiviazione Amazon EC2
- Amazon EFS: Amazon EFS Performance
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance
- DocumentazioneAmazon S3 Glacier: S3 Glacier
- Amazon S3: Request Rate and Performance Considerations
- Archiviazione nel cloud in AWS
- Caratteristiche e monitoraggio degli I/O - Amazon EBS
- Database su AWS Cloud
- AWS Database Caching
- DynamoDB Accelerator
- Best practice con Amazon Aurora
- Prestazioni Amazon Redshift
- Amazon Athena top 10 performance tips
- Amazon Redshift Spectrum best practices
- Amazon DynamoDB best practices

- Choose between Amazon EC2 and Amazon RDS
- Best practice e strategie di caching - Amazon ElastiCache

Video correlati:

- AWS re:Invent 2023: Improve Amazon Elastic Block Store efficiency and be more cost-efficient
- AWS re:Invent 2023: Optimizing storage price and performance with Amazon Simple Storage Service
- AWS re:Invent 2023: Building and optimizing a data lake on Amazon Simple Storage Service
- AWS re:Invent 2022: Building modern data architectures on AWS
- AWS re:Invent 2022: Building data mesh architectures on AWS
- AWS re:Invent 2023: Deep dive into Amazon Aurora and its innovations
- AWS re:Invent 2023: Advanced data modeling with Amazon DynamoDB
- AWS re:Invent 2022: Modernize apps with purpose-built databases
- Amazon DynamoDB deep dive: Advanced design patterns

Esempi correlati:

- AWS Purpose Built Databases Workshop
- Databases for Developers
- AWS Modern Data Architecture Immersion Day
- Build a Data Mesh on AWS
- Esempi di Amazon S3
- Optimize Data Pattern using Amazon Redshift Data Sharing
- Migrazioni dei database
- MS SQL Server - AWS Database Migration Service (AWS DMS) Replication Demo
- Database Modernization Hands On Workshop (Workshop pratico sulla modernizzazione dei database)
- Amazon Neptune Esempi

PERF03-BP02 Valutazione delle opzioni di configurazione disponibili per datastore

Comprendi e valuta le varie funzionalità e opzioni di configurazione disponibili per i tuoi datastore per ottimizzare lo spazio di archiviazione e le prestazioni per il tuo 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.
- Non conosci le opzioni di configurazione della soluzione di gestione dei dati scelta.
- Ti basi soltanto sull'aumento delle dimensioni dell'istanza, senza tenere conto di altre opzioni di configurazione disponibili.
- Non esegui il test delle caratteristiche di dimensionamento del tuo datastore.

Vantaggi dell'adozione di questa best practice: l'esplorazione e la sperimentazione delle configurazioni dei datastore ti consentono di ridurre il costo dell'infrastruttura, migliorare le prestazioni e diminuire le attività richieste per mantenere i carichi di lavoro.

Livello di rischio associato alla mancata adozione di questa best practice: medio

Guida all'implementazione

Un carico di lavoro può utilizzare uno o più datastore in base ai requisiti di archiviazione e accesso ai dati. Per ottimizzare prestazioni, efficienza e costi, è necessario valutare gli schemi di accesso ai dati per determinare le configurazioni appropriate del datastore. Nella valutazione delle opzioni di datastore, prendi in considerazione vari aspetti come le opzioni di archiviazione, la memoria, l'elaborazione, la replica di lettura, i requisiti di coerenza, il pool di connessioni e le opzioni di caching. Esegui esperimenti con queste diverse opzioni di configurazione per migliorare i parametri di efficienza delle prestazioni.

Passaggi dell'implementazione

- Esamina le configurazioni correnti (come il tipo di istanza, la dimensione di archiviazione o la versione del motore di database) del tuo datastore.

- Consulta la documentazione e le best practice AWS per scoprire le opzioni di configurazione consigliate che possono contribuire a migliorare le prestazioni del datastore. Le principali opzioni da considerare per il datastore sono le seguenti:

Configuration option	Examples
Offloading reads (like read replicas and caching)	• Per le tabelle DynamoDB, è possibile rimuovere le operazioni di lettura grazie a DAX per il caching. • Puoi creare un cluster Amazon ElastiCac he for Redis e configurare l'applicazione in modo che legga prima dalla cache e quindi passi al database se l'elemento richiesto non è presente. • I database relazionali come Amazon RDS e Aurora, nonché i database NoSQL allocati, come Neptune e Amazon DocumentDB, supportano tutti l'aggiunta di repliche di lettura per rimuovere le operazioni di lettura del carico di lavoro. • I database serverless come DynamoDB si dimensionano automaticamente. Assicurat i di avere abbastanza unità di capacità di lettura (RCU) assegnate per gestire il carico di lavoro.

Configuration option	Examples
Scaling writes (like partition key sharding or introducing a queue)	• 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. • 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 tuttavia possibile che si verifichino problemi con le partizioni hot quando si raggiungono i limiti di velocità di trasmissione effettiva per una determinata chiave di partizione. Questo problema può essere arginato scegliendo una chiave di partizione con una distribuzione più uniforme o

Configuration option	Examples
	eseguendo lo sharding in lettura della chiave di partizione.
Policies to manage the lifecycle of your datasets	• È possibile utilizzare Amazon S3 Lifecycle per gestire gli oggetti durante il loro ciclo di vita. Se gli schemi di accesso sono sconosciuti, mutevoli o imprevedibili, puoi usare Amazon S3 Intelligent-Tiering, che monitora gli schemi di accesso e sposta automaticamente gli oggetti che non hanno fatto registrare accessi a costi contenuti . Puoi sfruttare i parametri di Amazon S3 Storage Lens per identificare le opportunità di ottimizzazione e le lacune nella gestione del ciclo di vita. • La gestione del ciclo di vita di Amazon EFS gestisce automaticamente l'archiviazione dei file per i tuoi file system.
Connection management and pooling	• È possibile utilizzare Server proxy per Amazon RDS con Amazon RDS e Aurora per gestire le connessioni al database. • I database serverless come DynamoDB non hanno connessioni associate, ma valuta la capacità assegnata e le policy di dimension amento automatico per affrontare i picchi nel carico.

- Esegui esperimenti e benchmarking in un ambiente non di produzione per identificare quale opzione di configurazione può soddisfare i requisiti del tuo carico di lavoro.
- Dopo aver sperimentato, pianifica la migrazione e convalida i parametri delle prestazioni.
- Usa gli strumenti AWS per il monitoraggio, come Amazon CloudWatch, e l'ottimizzazione, come Amazon S3 Storage Lens, per ottimizzare continuamente il tuo datastore utilizzando schemi di utilizzo reali.

Risorse

Documenti correlati:

- Archiviazione nel cloud in AWS
- Tipi di volume Amazon EBS
- Archiviazione Amazon EC2
- Amazon EFS: Amazon EFS Performance
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance
- DocumentazioneAmazon S3 Glacier: S3 Glacier
- Amazon S3: Request Rate and Performance Considerations
- Caratteristiche e monitoraggio degli I/O - Amazon EBS
- Database su AWS Cloud
- AWS Database Caching
- DynamoDB Accelerator
- Best practice con Amazon Aurora
- Prestazioni Amazon Redshift
- Amazon Athena top 10 performance tips
- Amazon Redshift Spectrum best practices
- Amazon DynamoDB best practices

Video correlati:

- AWS re:Invent 2023: Improve Amazon Elastic Block Store efficiency and be more cost-efficient
- AWS re:Invent 2023: Optimize storage price and performance with Amazon Simple Storage Service
- AWS re:Invent 2023: Building and optimizing a data lake on Amazon Simple Storage Service
- AWS re:Invent 2023: What's new with AWS file storage
- AWS re:Invent 2023: Dive deep into Amazon DynamoDB

Esempi correlati:

- AWS Purpose Built Databases Workshop
- Databases for Developers
- AWS Modern Data Architecture Immersion Day
- Amazon EBS Autoscale
- Esempi di Amazon S3
- Amazon DynamoDB Examples
- AWS Database migration samples
- Database Modernization Workshop (Workshop sulla modernizzazione dei database)
- Working with parameters on your Amazon RDS for Postgress DB

PERF03-BP03 Raccolta e registrazione dei parametri delle prestazioni del datastore

Tieni traccia e registra i parametri delle prestazioni pertinenti per il tuo datastore per capire l'andamento delle prestazioni delle soluzioni di gestione dei dati. Questi parametri possono aiutarti a ottimizzare il tuo datastore, verificare che i requisiti del carico di lavoro siano rispettati e fornire una panoramica chiara sull'andamento delle prestazioni del carico di lavoro.

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 acquisire i parametri di accesso ai dati o di utilizzo.

Vantaggi dell'adozione di questa best practice: la definizione di una linea di base delle prestazioni ti aiuta a comprendere il comportamento normale e i requisiti dei carichi di lavoro. Gli schemi anomali possono essere identificati ed eliminati più rapidamente, per migliorare le prestazioni e l'affidabilità del datastore.

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

Guida all'implementazione

Per monitorare le prestazioni dei datastore, devi registrare più parametri delle prestazioni in un periodo di tempo. Ciò consente di rilevare le anomalie e di misurare le prestazioni rispetto ai parametri aziendali, per verificare che le esigenze del carico di lavoro siano rispettate.

I parametri devono includere sia il sistema sottostante che supporta il datastore 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, percentuale di riscontri nella cache e parametri di rete in entrata e in uscita, mentre i parametri del datastore 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 l'andamento del carico di lavoro e come viene utilizzata la soluzione di gestione dei dati. Utilizza tali parametri come parte di un approccio basato sui dati per mettere a punto e ottimizzare le risorse del tuo carico di lavoro.

Utilizza strumenti, librerie e sistemi che registrano misure delle prestazioni relative alle prestazioni del database.

Passaggi dell'implementazione

1. Determina i principali parametri delle prestazioni da monitorare per il tuo datastore.
    a. Parametri e dimensioni - Amazon S3
    b. Metriche di monitoraggio per un'istanza di Amazon RDS
    c. Monitoraggio del carico del database con Performance Insights su Amazon RDS
    d. Panoramica del monitoraggio avanzato
    e. Parametri e dimensioni - DynamoDB
    f. Monitoraggio di DynamoDB Accelerator
    g. Monitoraggio di Amazon MemoryDB for Redis con Amazon CloudWatch
    h. Quali parametri è opportuno monitorare?
    i. Monitoraggio delle prestazioni del cluster Amazon Redshift
    j. Parametri e dimensioni - Timestream
    k. Metriche di Amazon CloudWatch per Amazon Aurora
    l. Registrazione e monitoraggio in Amazon Keyspaces (for Apache Cassandra)
    m. Monitoraggio delle risorse di Amazon Neptune

2. Utilizza una soluzione di registrazione e monitoraggio approvata per raccogliere queste metriche. 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.
3. Verifica se il monitoraggio dei datastore può trarre vantaggio da una soluzione di machine learning che rileva le anomalie delle prestazioni.
   a. Amazon DevOps Guru per Amazon RDS offre visibilità sui problemi di prestazioni e fornisce suggerimenti per le azioni correttive.
4. Configura la conservazione dei dati nella soluzione di monitoraggio e registrazione per soddisfare i tuoi obiettivi operativi e di sicurezza.
   a. Conservazione dei dati predefinita per i parametri CloudWatch
   b. Conservazione dei dati predefinita per CloudWatch Logs

Risorse

Documenti correlati:

- AWS Database Caching (Memorizzazione nella cache del database AWS)
- 10 suggerimenti prestazionali su Amazon Athena
- Best practice con Amazon Aurora
- 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 re:Invent 2022 - Performance monitoring with Amazon RDS and Aurora, featuring Autodesk
- Database Performance Monitoring and Tuning with Amazon DevOps Guru for Amazon RDS
- AWS re:Invent 2023 - What’s new with AWS file storage

- AWS re:Invent 2023 - Dive deep into Amazon DynamoDB
- AWS re:Invent 2023 - Building and optimizing a data lake on Amazon S3
- AWS re:Invent 2023 - What's new with AWS file storage
- AWS re:Invent 2023 - Dive deep into Amazon DynamoDB
- Best Practices for Monitoring Redis Workloads on Amazon ElastiCache

Esempi correlati:

- AWS Dataset Ingestion Metrics Collection Framework (Framework di raccolta dei parametri di ingestione del set di dati AWS)
- Workshop di monitoraggio Amazon RDS
- AWS Purpose Built Databases Workshop

PERF03-BP04 Implementazione di strategie per migliorare le prestazioni delle query nel datastore

Implementa le strategie per ottimizzare i dati e migliorare le query sui dati in modo da consentire una maggiore scalabilità e prestazioni più efficienti per il tuo carico di lavoro.

Anti-pattern comuni:

- Non suddividi i dati in partizioni nel tuo datastore.
- I dati vengono archiviati in un solo formato di file nel tuo datastore.
- Non usi gli indici nel tuo datastore.

Vantaggi dell'adozione di questa best practice: L'ottimizzazione delle prestazioni dei dati e delle query si traduce in maggiore efficienza, costi inferiori e migliore esperienza utente.

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

Guida all'implementazione

L'ottimizzazione dei dati e delle query è un aspetto critico dell'efficienza delle prestazioni in un datastore, poiché influisce sulle prestazioni e sulla reattività dell'intero carico di lavoro cloud. Le query non ottimizzate possono comportare un maggiore utilizzo delle risorse e rallentamenti, riducendo così l'efficienza complessiva di un datastore.

L'ottimizzazione dei dati include diverse tecniche per garantire prestazioni efficienti per l'archiviazione e l'accesso ai dati. Ciò aiuta anche a migliorare le prestazioni delle query in un datastore. Le strategie chiave includono il partizionamento, la compressione e la denormalizzazione dei dati, che contribuiscono a ottimizzare i dati sia per l'archiviazione che per l'accesso.

Passaggi dell'implementazione

- Esamina e analizza le query sui dati critiche che vengono eseguite nel tuo datastore.
- Individua le query lente del tuo datastore e utilizza i piani di query per comprenderne lo stato attuale.
  - Analisi del piano di query in Amazon Redshift
  - Using EXPLAIN and EXPLAIN ANALYZE in Athena
- Implementa le strategie per migliorare le prestazioni delle query. Alcune strategie chiave sono:
  - Usando un formato di file colonnare (come Parquet o ORC).
  - Compressione dei dati nel datastore per ridurre lo spazio di archiviazione e il funzionamento di I/O.
  - Partizionamento dei dati per suddividere i dati in parti più piccole e ridurre i tempi di analisi dei dati.
    - Partizionamento dei dati in Athena
    - Partizioni e distribuzione dei dati
  - L'indicizzazione dei dati sulle colonne comuni della query.
  - Usa le viste materializzate per le domande frequenti.
    - Comprensione delle viste materializzate
    - Creazione di viste materializzate in Amazon Redshift
  - Scegli l'operazione di unione corretta per la query. Quando unisci due tabelle, specifica la tabella più grande sul lato sinistro dell'unione e la tabella più piccola sul lato destro.
  - La soluzione di caching distribuita migliora la latenza e riduce il numero di operazioni di I/O del database.
  - La manutenzione regolare, ad esempio l'esecuzione di statistiche.
- La sperimentazione e i test delle strategie in un ambiente non di produzione.

Risorse

Documenti correlati:

- Best practice con Amazon Aurora
- Prestazioni di Amazon Redshift
- 10 suggerimenti prestazionali su Amazon Athena
- AWS Database Caching (Memorizzazione nella cache del database AWS)
- Best practice per l'implementazione di Amazon ElastiCache
- Partizionamento dei dati in Athena

Video correlati:

- AWS re:Invent 2023 - AWS storage cost-optimization best practices
- AWS re:Invent 2022 - Performance monitoring with Amazon RDS and Aurora, featuring Autodesk
- Optimize Amazon Athena Queries with New Query Analysis Tools

Esempi correlati:

- Amazon S3 Select - Querying data without servers or databases
- AWS Purpose Built Databases Workshop

PERF03-BP05 Implementa modelli di accesso ai dati che utilizzano la memorizzazione nella cache

Implementa modelli di accesso che possano trarre vantaggio dalla memorizzazione dei dati nella cache per il recupero rapido dei dati a cui si accede di frequente.

Anti-pattern comuni:

- Memorizzare nella cache dati che cambiano in maniera frequente.
- Fare affidamento sui dati memorizzati nella cache come se fossero archiviati in modo duraturo e sempre disponibili.
- Non tenere conto della coerenza dei dati memorizzati nella cache.
- Non monitorare l'efficienza dell'implementazione della cache.

Vantaggi dell'adozione di questa best practice: L'archiviazione dei dati in una cache può migliorare la latenza di lettura, la velocità effettiva di lettura, l'esperienza utente e l'efficienza complessiva, oltre a ridurre i costi.

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

Guida all'implementazione

Una cache è un componente software o hardware progettato per archiviare dati in modo che le richieste future degli stessi dati possano essere soddisfatte più velocemente o in modo più efficiente. I dati memorizzati in una cache possono essere ricostruiti in caso di perdita, ripetendo un calcolo precedente o recuperandolo da un altro datastore.

La memorizzazione dei dati nella cache può essere una delle strategie più efficaci per migliorare le prestazioni complessive delle applicazioni e ridurre il carico sulle origini dati primarie sottostanti. I dati possono essere memorizzati nella cache a diversi livelli dell'applicazione, ad esempio all'interno dell'applicazione che effettua chiamate remote, operazione nota come memorizzazione nella cache lato client, o utilizzando un servizio secondario veloce per l'archiviazione dei dati, operazione nota come memorizzazione nella cache remota.

Memorizzazione nella cache lato client

Con la memorizzazione nella cache lato client, ogni client (un'applicazione o un servizio che interroga il datastore di backend) può archiviare localmente i risultati delle proprie query uniche per un periodo di tempo specificato. Ciò può ridurre il numero di richieste a un datastore attraverso la rete perché viene controllata prima la cache del client locale. Se questa non contiene risultati, l'applicazione può interrogare il datastore e archiviare tali risultati localmente. Questo modello consente a ciascun client di archiviare i dati nella sede più vicina possibile (il client stesso), garantendo così la latenza più bassa possibile. I client possono inoltre continuare a eseguire query quando il datastore di backend non è disponibile, aumentando la disponibilità dell'intero sistema.

Uno svantaggio di questo approccio è che quando sono coinvolti più client, potrebbero archiviare localmente gli stessi dati memorizzati nella cache. Ciò si traduce in un utilizzo duplicato dell'archiviazione e nell'incoerenza dei dati tra questi client. Può accadere che un client memorizzi nella cache i risultati di una query e un minuto dopo un altro client esegua la stessa query ottenendo un risultato diverso.

Memorizzazione nella cache remota

Per risolvere il problema della duplicazione dei dati tra client, utilizza un servizio esterno veloce o una cache remotaper archiviare i dati sottoposti a query. Anziché controllare un datastore locale, ogni client controllerà la cache remota prima di interrogare il datastore di backend. Questa strategia consente di ottenere risposte più coerenti tra i client, una migliore efficienza dei dati archiviati e un volume maggiore di dati memorizzati nella cache, perché lo spazio di archiviazione si dimensiona in maniera indipendente dai client.

Lo svantaggio di una cache remota è che l'intero sistema può registrare una latenza più elevata, poiché è necessario un hop di rete aggiuntivo per controllare la cache remota. Per migliorare la latenza, è possibile utilizzare la memorizzazione nella cache lato client insieme alla memorizzazione nella cache remota, eseguendo così una memorizzazione nella cache su più livelli.

Passaggi dell'implementazione

1. Identifica database, API e servizi di rete che potrebbero trarre vantaggio dalla memorizzazione nella cache. I candidati migliori per la memorizzazione nella cache sono i servizi che presentano carichi di lavoro di lettura elevati, un rapporto lettura/scrittura elevato o che sono costosi da dimensionare.
   - Memorizzazione nella cache del database
   - Abilita la memorizzazione nella cache dell'API per migliorare la velocità di risposta
2. Identifica il tipo di strategia di memorizzazione nella cache più adatto al tuo modello di accesso.
   - Strategie di cache
   - Soluzioni di memorizzazione nella cache AWS
3. Seguici Best practice di memorizzazione nella cache per il tuo datastore.
4. Configura una strategia di invalidazione della cache per tutti i dati, ad esempio un TTL (Time-to-live), che permetta di bilanciare attualità dei dati e riduzione della pressione sul datastore di backend.
5. Abilita funzionalità quali tentativi di connessione automatici, backoff esponenziale, timeout lato client e pool di connessioni nel client, se disponibili, che possono migliorare prestazioni e affidabilità.
   - Best practice: client Redis e Amazon ElastiCache for Redis
6. Monitora la percentuale di riscontri nella cache con un obiettivo dell'80% o superiore. Valori inferiori possono indicare una dimensione della cache insufficiente o un modello di accesso che non sfrutta la memorizzazione nella cache.
   - Quali parametri è opportuno monitorare?

- Best practices for monitoring Redis workloads on Amazon ElastiCache
- Monitoring best practices with Amazon ElastiCache for Redis using Amazon CloudWatch

7. Implementa la replica dei dati per distribuire il carico delle letture su più istanze e migliorare le prestazioni e la disponibilità di lettura dei dati.

Risorse

Documenti correlati:

- Using the Amazon ElastiCache Well-Architected Lens
- Monitoring best practices with Amazon ElastiCache for Redis using Amazon CloudWatch
- Quali parametri è opportuno monitorare?
- Performance at Scale with Amazon ElastiCache whitepaper
- Sfide e strategie di caching

Video correlati:

- Amazon ElastiCache Learning Path
- Design for success with Amazon ElastiCache best practices
- AWS re:Invent 2020 - Design for success with Amazon ElastiCache best practices
- AWS re:Invent 2023 - [LAUNCH] Introducing Amazon ElastiCache Serverless
- AWS re:Invent 2022 - 5 great ways to reimagine your data layer with Redis
- AWS re:Invent 2021 - Deep dive on Amazon ElastiCache for Redis

Esempi correlati:

- Boosting MySQL database performance with Amazon ElastiCache for Redis

Reti e distribuzione di contenuti

La soluzione di rete ottimale per un carico di lavoro varia in base a latenza, requisiti di velocità di trasmissione effettiva, 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.

In AWS, le reti sono virtualizzate e vengono fornite in molti tipi e configurazioni diversi. In questo modo puoi soddisfare le tue esigenze di rete più facilmente. AWS offre caratteristiche di prodotto (ad esempio reti avanzate, istanze Amazon EC2 ottimizzate per la rete, accelerazione del trasferimento Amazon S3 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.

Questa area di interesse offre linee guida e best practice per progettare, configurare e gestire soluzioni di rete e distribuzione di contenuti nel cloud in maniera efficiente.

Best practice

- PERF04-BP01 In che modo la rete influisce sulle prestazioni
- PERF04-BP02 Valuta le funzionalità di rete disponibili
- PERF04-BP03 Scegli la connettività dedicata o la VPN appropriata per il tuo carico di lavoro
- PERF04-BP04 Utilizzo del bilanciamento del carico per distribuire il traffico su più risorse
- PERF04-BP05 Scelta dei protocolli di rete per migliorare le prestazioni
- PERF04-BP06 Scelta della posizione del carico di lavoro in base ai requisiti di rete
- PERF04-BP07 Ottimizzazione della configurazione di rete in base alle metriche

PERF04-BP01 In che modo la rete influisce sulle prestazioni

Analizza e comprendi in che modo le decisioni correlate alla rete influiscono sul carico di lavoro per fornire prestazioni efficienti e una migliore esperienza utente.

Anti-pattern comuni:

- Tutto il traffico passa attraverso i data center esistenti.

- Si instrada tutto il traffico attraverso i firewall centrali anziché utilizzare strumenti di sicurezza di rete nativi del cloud.
- Si effettua il provisioning delle connessioni AWS Direct Connect senza comprendere gli effettivi requisiti di utilizzo.
- 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 aiuta 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: alto

Guida all'implementazione

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 può essere influenzata anche da latenza della rete, larghezza di banda, protocolli, posizione, congestione della rete, jitter, velocità di trasmissione effettiva e regole di instradamento.

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.

Passaggi dell'implementazione:

1. Definisci e documenta i requisiti di prestazioni di rete, tra cui metriche 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. Scopri i principali servizi di rete AWS come VPC, AWS Direct Connect, Elastic Load Balancing (ELB)e Amazon Route 53.
3. Acquisisci le seguenti caratteristiche di rete fondamentali:

Caratteristiche	Strumenti e metriche
Caratteristiche fondamentali della rete	• Log di flusso VPC • Log di flusso AWS Transit Gateway • Metriche AWS Transit Gateway • Parametri AWS PrivateLink
Caratteristiche di rete dell'applicazione	• Elastic Fabric Adapter • Metriche AWS App Mesh • Parametri Amazon API Gateway
Caratteristiche della rete edge	• Parametri Amazon CloudFront • Parametri Amazon Route 53 • Metriche AWS Global Accelerator
Caratteristiche della rete ibrida	• Metriche AWS Direct Connect • Metriche AWS Site-to-Site VPN • Metriche AWS Client VPN • Parametri WAN Cloud AWS
Caratteristiche della sicurezza di rete	• Metriche AWS Shield, AWS WAF e AWS Network Firewall
Caratteristiche del tracciamento	• AWS X-Ray • VPC Reachability Analyzer • Network Access Analyzer • Amazon Inspector • Usare il RUM Amazon CloudWatch

4. Esegui il benchmark e testa le prestazioni della rete:

    a. Esegui il benchmark della velocità di trasmissione effettiva della rete, poiché alcuni fattori possono influire sulle prestazioni della rete Amazon EC2 quando le istanze si trovano nello stesso VPC. Misura la larghezza di banda della rete tra le istanze Amazon EC2 Linux nello stesso VPC.

b. Esegui test di carico per sperimentare soluzioni e opzioni di rete.

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

Video correlati:

- AWS re:Invent 2023 - AWS networking foundations
- AWS re:Invent 2023 - What can networking do for your application?
- AWS re:Invent 2023 - Advanced VPC designs and new capabilities
- AWS re:Invent 2023 - A developer’s guide to cloud networking
- AWS re:Invent 2019 - Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2019 - Optimizing Network Performance for Amazon EC2 Instances
- AWS Summit Online - Improve Global Network Performance for Applications
- AWS re:Invent 2020 - Networking best practices and tips with the Well-Architected Framework
- AWS re:Invent 2020 - AWS networking best practices in large-scale migrations

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions

- AWS Networking Workshops
- Hands-on Network Firewall Workshop
- Observing and Diagnosing your Network on AWS
- Finding and addressing Network Misconfigurations on AWS

PERF04-BP02 Valuta le 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, distanza di rete o jitter.

Anti-pattern comuni:

- Rimani all'interno di una regione perché è lì che si trova fisicamente la tua sede centrale.
- Utilizzi i firewall anziché i gruppi di sicurezza per filtrare il traffico.
- Interrompi TLS per l'ispezione del traffico anziché affidarti a gruppi di sicurezza, policy degli endpoint e altre funzionalità native del cloud.
- Utilizzi solo la segmentazione basata su sottoreti anziché i gruppi di sicurezza.

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. La struttura portante globale di AWS ti aiuta a fornire ai tuoi clienti la migliore esperienza di rete.

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

Guida all'implementazione

AWS offre servizi come AWS Global Accelerator e Amazon CloudFront , i quali possono contribuire a migliorare le prestazioni della rete, mentre la maggior parte dei servizi AWS include funzionalità di prodotto (come l'Accelerazione del trasferimento Amazon S3) per ottimizzare il traffico di rete.

Analizza quali opzioni di configurazione relative alla rete sono disponibili e come possono influire sul tuo carico di lavoro. L'ottimizzazione delle prestazioni dipende dalla comprensione del modo in cui queste opzioni interagiscono con l'architettura e dall'impatto che hanno sulle prestazioni misurate e sull'esperienza utente.

Passaggi dell'implementazione

- Crea l'elenco dei componenti del carico di lavoro.
  - Prendi in considerazione l'utilizzo di Cloud AWS WAN per creare, gestire e monitorare la rete dell'organizzazione durante la creazione di una rete globale unificata.
  - Monitora le tue reti globali e principali con le metriche di Amazon CloudWatch Logs. Sfrutta Amazon CloudWatch RUM, che fornisce approfondimenti utili per identificare, comprendere e migliorare l'esperienza digitale degli utenti.
  - Visualizza la latenza di rete aggregata tra Regioni AWS e zone di disponibilità, nonché all'interno di ciascuna zona di disponibilità, utilizzando AWS Network Manager , che ti permette di ottenere informazioni dettagliate sul modo in cui le prestazioni delle applicazioni variano in base alle prestazioni della rete AWS sottostante.
  - Utilizza uno strumento per database di gestione delle configurazioni (CMDB) esistente oppure un servizio come AWS Config per creare un inventario del carico di lavoro e della relativa configurazione.
- 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.
- Se si tratta di un nuovo carico di lavoro, esegui i test di carico per individuare eventuali colli di bottiglia relativi alle prestazioni.
- 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. Consulta le seguenti opzioni e funzionalità di rete fondamentali:

Opportunità di miglioramento	Soluzione
Percorso o instradamenti di rete	Utilizza Network Access Analyzer per identific are percorsi o instradamenti.
Protocolli di rete	Consulta PERF04-BP05 Scelta dei protocolli di rete per migliorare le prestazioni

Opportunità di miglioramento	Soluzione
Topologia di rete	Valuta i compromessi a livello di operazioni e prestazioni tra Peering VPC e AWS Transit Gateway quando si collegano più account. AWS Transit Gateway semplifica il modo in cui interconnetti tutti i VPC, che possono essere distribuiti su migliaia di Account AWS e in reti on-premise. Condividi AWS Transit Gateway tra più account utilizzando la funzionalità AWS Resource Access Manager. Consulta PERF04-BP03 Scegli la connettività dedicata o la VPN appropriata per il tuo carico di lavoro

Opportunità di miglioramento	Soluzione
Servizi di rete	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 AWS. Amazon CloudFront può migliorare le prestazioni della distribuzione dei contenuti del tuo carico di lavoro e la latenza a livello globale. Utilizza Lambda@edge per eseguire funzioni di personalizzazione dei contenuti che CloudFront distribuisce più vicino agli utenti, ridurre la latenza e migliorare le prestazioni. 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. Rivedi il traffico del carico di lavoro e la posizione dell'utente quando il carico di lavoro è distribuito a livello globale per individuare quale opzione di instradamento è in grado di ottimizzare le prestazioni del carico di lavoro.

Opportunità di miglioramento	Soluzione
Funzionalità delle risorse di archiviazione	L'Accelerazione del trasferimento Amazon S3 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. 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-reg ione, puoi richiedere o scrivere dati in Amazon S3 con il servizio che identifica il bucket con latenza più bassa.

Opportunità di miglioramento	Soluzione
Funzionalità delle risorse di calcolo	Le interfacce di rete elastiche (ENA) utilizzate da istanze Amazon EC2, container e funzioni Lambda sono limitate in base al flusso. 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 visibilit à su di esse, utilizza le metriche CloudWatch ed ethtool. Il comando `ethtool` è incluso nel driver ENA e permette di utilizzare metriche relative alla rete aggiuntive che possono essere pubblicate come metrica personali zzata in CloudWatch. Gli adattatori elastici di rete (ENA) Amazon offrono un'ulteriore ottimizzazione grazie a una migliore velocità di trasmissione effettiva per le istanze all'interno di un gruppo di collocazione cluster. Elastic Fabric Adapter (EFA) 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. 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.

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
- Passaggio all'instradamento basato sulla latenza in Amazon Route 53
- Endpoint VPC
- Log di flusso VPC

Video correlati:

- AWS re:Invent 2023 – Ready for what's next? Designing networks for growth and flexibility
- AWS re:Invent 2023 – Advanced VPC designs and new capabilities
- AWS re:Invent 2023 – A developer's guide to cloud networking
- AWS re:Invent 2022 – Dive deep on AWS networking infrastructure
- AWS re:Invent 2019 – Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2018 – Optimizing Network Performance for Amazon EC2 Instances
- AWS Global Accelerator

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions
- AWS Networking Workshops
- Observing and diagnosing your network
- Finding and addressing network misconfigurations on AWS

PERF04-BP03 Scegli la connettività dedicata o la VPN appropriata per il tuo carico di lavoro

Quando hai bisogno di una connettività ibrida per connettere risorse on-premise e cloud, assicurati di avere una larghezza di banda adeguata per soddisfare i tuoi requisiti di prestazione. Fai una stima dei requisiti di larghezza di banda e latenza per il carico di lavoro ibrido. I valori calcolati determineranno le tue esigenze di dimensionamento.

Anti-pattern comuni:

- Valutazione delle soluzioni VPN solo per i tuoi requisiti di crittografia di rete.
- Non vengono valutate opzioni di backup o di connettività ridondante.
- Non è possibile 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 selezione e la configurazione di soluzioni di connettività appropriate migliorano l'affidabilità del carico di lavoro e massimizzano le prestazioni. L'identificazione di requisiti del carico di lavoro, la pianificazione anticipata e la valutazione di soluzioni ibride ti permetteranno di ridurre al minimo le costose modifiche alla rete fisica e i costi operativi, migliorando al contempo il time-to-value.

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

Guida all'implementazione

Sviluppa un'architettura di rete ibrida basata sui requisiti di larghezza di banda. AWS Direct Connect ti consente di connettere la tua rete on-premise in modo privato con AWS. È utile quando hai bisogno di larghezza di banda elevata, bassa latenza e di mantenere le prestazioni coerenti. Una connessione VPN permette di connettersi in modo sicuro su Internet. Viene utilizzata quando è necessaria solo una connessione temporanea, quando il costo è un fattore importante o come misura di contingenza in attesa che venga stabilita una connettività di rete fisica resiliente mentre AWS Direct Connect è in uso.

Se i tuoi requisiti di larghezza di banda sono elevati, potresti prendere in considerazione l'utilizzo di più AWS Direct Connect o di servizi di VPN. Il traffico può essere bilanciato in termini di carico tra i servizi, ma il bilanciamento del carico tra AWS Direct Connect e VPN è sconsigliato a causa delle differenze di latenza e larghezza di banda.

Passaggi dell'implementazione

1. Calcola i requisiti di larghezza di banda e latenza delle tue app esistenti.
   a. Per i carichi di lavoro esistenti che vengono spostati in AWS, utilizza i dati raccolti dai sistemi di monitoraggio di rete interni.
   b. Per i carichi di lavoro nuovi o esistenti per i quali non sono disponibili dati di monitoraggio, consulta i proprietari dei prodotti per definire metriche sulle prestazioni adeguate e offrire un'esperienza utente soddisfacente.
2. Scegli una connessione dedicata o una VPN come opzione di connettività. A seconda di tutti i requisiti del carico di lavoro (esigenze di crittografia, larghezza di banda e traffico), puoi scegliere AWS Direct Connect o AWS VPN (o entrambi). Il diagramma seguente può aiutarti a scegliere il tipo di connessione appropriato.
   a. AWS Direct Connect fornisce connettività dedicata all'ambiente AWS da 50 Mbps fino a 100 Gbps, utilizzando connessioni dedicate od ospitate. In questo modo, disporrai di latenza gestita e controllata, nonché di larghezza di banda assegnata, in modo che il carico di lavoro possa connettersi con efficienza ad altri ambienti. Ricorrendo a partner AWS Direct Connect, otterrai connettività end-to-end da più ambienti, per una rete estesa con prestazioni coerenti. AWS permette di dimensionare la larghezza di banda di connessione Direct Connect usando connettività nativa a 100 Gbps, gruppi di aggregazione di collegamenti (LAG, Link Aggregation Group) o instradamento ECMP (Equal-Cost Multipath) con BGP.
   b. AWS Site-to-Site VPN offre un servizio VPN gestito che supporta il protocollo IPSec (Internet Protocol security). Quando viene creata una connessione VPN, ogni connessione include due tunnel per la disponibilità elevata.
3. Consulta la documentazione AWS per scegliere l'opzione di connettività appropriata:
   a. Se decidi di utilizzareAWS Direct Connect, seleziona la larghezza di banda appropriata per la tua connettività.
   b. Se utilizzi una AWS Site-to-Site VPN tra più posizioni per connetterti a una Regione AWS, prova a utilizzare una connessione Site-to-Site VPN accelerata per migliorare le prestazioni della rete.
   c. Se il progetto di rete è costituito da una connessione VPN IPSec tramite AWS Direct Connect, prendi in considerazione l'utilizzo di VPN con indirizzo IP privato per migliorare la sicurezza e ottenere la segmentazione. La VPN sito-sito AWS con indirizzo IP privato viene implementata sull'interfaccia virtuale di transito (VIF).
   d. AWS Direct Connect SiteLink consente di creare connessioni ridondanti e a bassa latenza tra i data center in tutto il mondo inviando dati lungo il percorso più veloce tra sedi di AWS Direct Connect, bypassando Regioni AWS.

4. Convalida la configurazione della connettività prima di eseguire l'implementazione in produzione. Esegui test di sicurezza e prestazioni per assicurarti di soddisfare i requisiti di larghezza di banda, affidabilità, latenza e conformità.
5. Monitora regolarmente le prestazioni e l'utilizzo della connettività e ottimizzali, se necessario.

Diagramma di flusso per le prestazioni deterministiche.

Risorse

Documenti correlati:

- Nuovi prodotti di rete con AWS
- AWS Transit Gateway
- Endpoint VPC
- Creazione di un'infrastruttura di rete AWS scalabile e sicura con più VPC
- Client VPN

Video correlati:

- AWS re:Invent 2023 – Building hybrid network connectivity with AWS
- AWS re:Invent 2023 – Secure remote connectivity to AWS
- AWS re:Invent 2022 – Optimizing performance with Amazon CloudFront
- AWS re:Invent 2019 – Connectivity to AWS and hybrid AWS network architectures
- AWS re:Invent 2020 – AWS Transit Gateway Connect

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions
- AWS Networking Workshops

PERF04-BP04 Utilizzo del bilanciamento del carico per distribuire il traffico su più risorse

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, l'affidabilità e gestire e instradare il traffico in modo efficiente.

Anti-pattern comuni:

- Scelta del tipo di sistema di bilanciamento del carico senza tenere conto dei requisiti del carico di lavoro.
- Mancato utilizzo delle funzionalità del sistema di bilanciamento del carico per l'ottimizzazione delle prestazioni.
- Esposizione diretta del carico di lavoro a Internet senza un sistema di bilanciamento del carico.
- 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 bilanciatore del carico gestisce il carico variabile del traffico dell'applicazione in una o più zone di disponibilità e fornisce alta disponibilità, dimensionamento automatico e un migliore utilizzo del carico di lavoro.

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

Guida all'implementazione

I sistemi di bilanciamento del carico operano come punto di ingresso per il carico di lavoro, dal quale distribuiscono il traffico alle destinazioni di backend, come istanze di calcolo o container per migliorarne l'utilizzo.

La scelta del tipo corretto di sistema di bilanciamento del carico è il primo passaggio per ottimizzare l'architettura. Per iniziare, elenca le caratteristiche del carico di lavoro, tra cui protocollo (TCP, HTTP, TLS o WebSocket), tipo di destinazione (istanze, container o servizi serverless), requisiti dell'applicazione (connessioni a esecuzione prolungata, autenticazione utente o persistenza) e ubicazione (regione, zona locale, Outpost o isolamento zonale).

AWS fornisce più modelli per consentire alle tue applicazioni di utilizzare il bilanciamento del carico. Application Load Balancer è l'ideale per il bilanciamento del carico del traffico HTTP e HTTPS, nonché offre l'instradamento avanzato delle richieste, dedicato alla distribuzione delle architetture applicative moderne, fra cui 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 bilanciatore del carico e di sollevare il carico di lavoro dall'utilizzo intensivo della CPU.

Dopo aver scelto il sistema di bilanciamento del carico appropriato, puoi iniziare a utilizzarne le funzionalità per ridurre la quantità di attività che deve svolgere il backend per distribuire il traffico.

Ad esempio, usando un Application Load Balancer (ALB) e un Network Load Balancer (NLB), puoi eseguire l'offload della crittografia SSL/TLS, il che costituisce un'opportunità per evitare il completamento dell'handshake TLS a elevato utilizzo di CPU da parte delle destinazioni e migliorare anche la gestione dei certificati.

Se configurato nel sistema di bilanciamento del carico, l'offload SSL/TLS diventa responsabile della crittografia del traffico da e verso i client, distribuendo il traffico non crittografato ai backend, liberando le risorse backend e migliorando il tempo di risposta per i client.

L'Application Load Balancer può anche distribuire traffico HTTP/2 senza che questo debba essere supportato nelle destinazioni. Questa semplice decisione può migliorare il tempo di risposta dell'applicazione, in quanto HTTP/2 usa connessioni TCP in modo più efficiente.

Nel definire l'architettura, è bene tenere conto dei requisiti di latenza del carico di lavoro. Ad esempio, se un'applicazione è sensibile alla latenza, è possibile scegliere di usare un Network Load Balancer, che offre latenze estremamente ridotte. In alternativa, è possibile decidere di avvicinare il carico di lavoro ai clienti utilizzando Application Load Balancer in zone locali AWS o anche in AWS Outposts.

Un altro aspetto di cui tenere conto per i carichi di lavoro sensibili alla latenza è il bilanciamento del carico tra zone. Con il bilanciamento del carico tra zone, ogni nodo del sistema di bilanciamento del carico distribuisce il traffico tra le destinazioni registrate in tutte le zone di disponibilità autorizzate.

Usa Auto Scaling integrato con il sistema di bilanciamento del carico. Uno degli aspetti principali di un sistema con prestazioni efficienti riguarda il dimensionamento corretto delle risorse backend.

A questo scopo, puoi utilizzare integrazioni dei sistemi di bilanciamento del carico per le risorse di destinazione backend. Usando l'integrazione dei sistemi di bilanciamento del carico con gruppi con Auto Scaling, le destinazioni vengono aggiunte o rimosse nel e dal sistema di bilanciamento del carico in base alle esigenze, in risposta al traffico in ingresso. I bilanciatori del carico possono integrarsi anche con Amazon ECS e Amazon EKS per i carichi di lavoro distribuiti in container.

- Amazon ECS - Service load balancing
- Application load balancing on Amazon EKS
- Network load balancing on Amazon EKS

Passaggi dell'implementazione

- Definisci i tuoi requisiti di bilanciamento del carico, tra cui volume di traffico, disponibilità e scalabilità delle applicazioni.
- Scegli il tipo di sistema di bilanciamento del carico giusto per la tua applicazione.
  - Usa un Application Load Balancer per carichi di lavoro HTTP/HTTPS.
  - Use un Network Load Balancer per carichi di lavoro non HTTP in esecuzione su TCP o UDP.
  - Usa una combinazione dei due sistemi (un ALB come destinazione di un NLB) se vuoi usufruire delle funzionalità di entrambi i prodotti. Ad esempio, puoi scegliere questa opzione per usare gli indirizzi IP statici del NLB insieme all'instradamento basato su intestazione HTTP dell'ALB oppure se vuoi esporre il carico di lavoro HTTP a un AWS PrivateLink.
  - Per un confronto completo dei bilanciatori del carico, consulta la tabella di confronto dei prodotti ELB.
- Se possibile, utilizza l'offload SSL/TLS.
  - Configura gli ascoltatori HTTPS/TLS con un Application Load Balancer e un Network Load Balancer integrati con AWS Certificate Manager.
  - Alcuni carichi di lavoro possono richiedere la crittografia end-to-end per motivi di conformità. In questo caso, è necessario consentire la crittografia nelle destinazioni.
  - Per le best practice per la sicurezza, consulta SEC09-BP02 Applicazione della crittografia dei dati in transito.
- Seleziona l'algoritmo di instradamento corretto (solo ALB)
  - L'algoritmo di instradamento può fare la differenza per quanto riguarda l'uso corretto delle destinazioni backend e, di conseguenza, l'impatto sulle prestazioni. Ad esempio, l'ALB offre due opzioni per gli algoritmi di instradamento:

  - Numero minimo di richieste in sospeso: usa questa opzione per ottenere una migliore distribuzione del carico nelle destinazioni back-end nei casi in cui le richieste per l'applicazione variano per complessità o le destinazioni variano per capacità di elaborazione.
  - Round robin: usa questa opzione quando le richieste e le destinazioni sono simili o se devi distribuire equamente le richieste tra le destinazioni.
- Valuta se usare l'isolamento tra zone o quello zonale.
  - Disattiva l'isolamento tra zone (usando l'isolamento zonale) per migliorare la latenza e in caso di errori di zona. È disattivato per impostazione predefinita nel NLB, mentre nell'ALB puoi disattivarlo per ogni gruppo di destinazioni.
  - Attiva l'isolamento tra zone per ottenere disponibilità e flessibilità maggiori. Per impostazione predefinita, l'isolamento tra zone è disattivato per l'ALB, mentre nel NLB puoi attivarlo per ogni gruppo di destinazioni.
- Attiva keep-alive HTTP per i carichi di lavoro HTTP (solo ALB). Con questa funzionalità, il sistema di bilanciamento del carico può riutilizzare le connessioni backend fino allo scadere del timeout del keep-alive, migliorando la richiesta HTTP e il tempo di risposta e riducendo anche l'utilizzo delle risorse nelle destinazioni backend. Per informazioni sulla configurazione per Apache e Nginx, consulta Quali sono le impostazioni ottimali per utilizzare Apache o NGINX come server di backend per ELB?
- Attiva il monitoraggio del tuo sistema di bilanciamento del carico.
  - Attiva i log di accesso per l'Application Load Balancer e il Network Load Balancer.
  - I campi principali da considerare per l'ALB sono `request_processing_time`, `request_processing_time` e `response_processing_time`.
  - I campi principali da considerare per il NLB sono `connection_time` e `tls_handshake_time`.
  - Preparati a eseguire query sui log quando necessario. Puoi usare Amazon Athena per eseguire query su log dell'ALB e log del NLB.
  - Crea gli allarmi per le metriche relative alle prestazioni come `TargetResponseTime` per l'ALB.

Risorse

Documenti correlati:

- Confronti di prodotti ELB
- Infrastruttura globale di AWS

- Improving Performance and Reducing Cost Using Availability Zone Affinity
- Step by step for Log Analysis with Amazon Athena
- Querying Application Load Balancer logs
- Monitor your Application Load Balancers
- Monitor your Network Load Balancer
- Use Elastic Load Balancing to distribute traffic across the instances in your Auto Scaling group

Video correlati:

- AWS re:Invent 2023: What can networking do for your application?
- AWS re:Inforce 20: How to use Elastic Load Balancing to enhance your security posture at scale
- AWS re:Invent 2018: Elastic Load Balancing: Deep Dive and Best Practices
- AWS re:Invent 2021 - How to choose the right load balancer for your AWS workloads
- AWS re:Invent 2019: Get the most from Elastic Load Balancing for different workloads

Esempi correlati:

- Gateway Load Balancer
- CDK and AWS CloudFormation samples for Log Analysis with Amazon Athena

PERF04-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 la velocità di trasmissione desiderata. Se per il trasferimento di file viene usato il protocollo TCP, latenze più elevate molto probabilmente ridurranno la velocità di trasmissione effettiva complessiva. Alcuni approcci risolvono questo problema tramite l'ottimizzazione del TCP e l'utilizzo di protocolli di trasferimento ottimizzati, un altro prevede l'utilizzo del protocollo UDP (User Datagram Protocol).

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 verifica del protocollo appropriato per la comunicazione tra utenti e componenti del carico di lavoro contribuisce a migliorare l'esperienza utente complessiva per le applicazioni. Ad esempio, 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: medio

Guida all'implementazione

Se hai la possibilità di scegliere protocolli diversi per la tua applicazione e hai esperienza in questo campo, ottimizza l'applicazione e l'esperienza dell'utente finale utilizzando un protocollo diverso. Tieni conto che questo approccio presenta notevoli difficoltà e dovrebbe essere tentato solo dopo aver ottimizzato l'applicazione in altri modi.

Un aspetto principale per il miglioramento delle prestazioni del carico di lavoro consiste nell'identificare i requisiti di latenza e velocità di trasmissione effettiva e quindi scegliere i protocolli di rete che ottimizzano le prestazioni.

Quando valutare se usare TCP

TCP permette la trasmissione affidabile dei dati e può essere usato per la comunicazione tra i componenti del carico di lavoro quando l'affidabilità e la garanzia di trasmissione dei dati sono due aspetti importanti. Molte applicazioni Web usano protocolli basati su TCP, come HTTP e HTTPS, per aprire socket TCP per la comunicazione tra i componenti dell'applicazione. Il TCP viene comunemente usato per il trasferimento di dati di posta elettronica e di file, in quanto è un meccanismo di trasferimento semplice e affidabile tra i componenti dell'applicazione. L'uso di TLS con TCP può aggiungere un certo sovraccarico alla comunicazione, il che produce maggiore latenza e velocità di trasmissione effettiva inferiore, ma presenta come vantaggio una maggiore sicurezza. Il sovraccarico è dovuto prevalentemente al processo di handshake, il cui completamento può richiedere diversi round trip. Al termine del processo di handshake, il sovraccarico dovuto alla crittografia e alla decrittografia dei dati è relativamente ridotto.

Quando valutare se usare UDP

UDP è un protocollo di tipo connection-less (senza connessione) e di conseguenza è ideale per applicazioni che necessitano di una trasmissione veloce ed efficiente, ad esempio per i log, il monitoraggio e i dati VoIP. Valuta se usare UDP anche se vi sono componenti del carico di lavoro che rispondono a piccole query provenienti da grandi quantità di client per garantire prestazioni

ottimali del carico di lavoro. Datagram Transport Layer Security (DTLS) è l'equivalente UDP di Transport Layer Security (TLS). Quando viene usato DTLS con UDP, il sovraccarico è dovuto alla crittografia e alla decrittografia dei dati, in quanto il processo di handshake è semplificato. DTLS aggiunge anche un piccolo sovraccarico ai pacchetti UDP, perché include altri campi per indicare i parametri di sicurezza e rilevare la manomissione.

Quando valutare se usare SRD

SRD (Scalable Reliable Datagram) è un protocollo di trasporto di rete ottimizzato per carichi di lavoro a velocità di trasmissione effettiva elevata grazie alla sua capacità di bilanciare il carico del traffico tra più percorsi e di recuperare rapidamente dalla perdita di pacchetti e da errori di collegamento. Di conseguenza, SRD è ideale per carichi di lavoro di calcolo ad alte prestazioni (HPC) che richiedono comunicazioni tra nodi di calcolo a velocità di trasmissione effettiva elevata e a bassa latenza. Possono essere incluse attività di elaborazione in parallelo come la simulazione, la modellazione e l'analisi dei dati che implicano il trasferimento di grandi quantità di dati tra nodi.

Passaggi dell'implementazione

1. Utilizza AWS Global Accelerator e AWS Transfer Family per migliorare la velocità di trasmissione effettiva delle applicazioni di trasferimento di file online. Il servizio AWS Global Accelerator ti permette di ottenere latenza inferiore tra i dispositivi client e il carico di lavoro in AWS. Con AWS Transfer Family puoi usare protocolli basati su TCP come SFTP (Secure Shell File Transfer Protocol) ed FTPS (File Transfer Protocol over SSL) per dimensionare e gestire i trasferimenti di file in servizi di archiviazione AWS in tutta sicurezza.
2. Usa la latenza di rete per determinare se TCP sia il protocollo appropriato per la comunicazione tra componenti del carico di lavoro. Se la latenza di rete tra l'applicazione client e il server è elevata, il processo di handshake a tre vie tramite TCP può richiedere tempo, influendo sulla velocità di risposta dell'applicazione. Per misurare la latenza di rete, puoi usare, ad esempio, le metriche TTFB (tempo di acquisizione al primo byte) e RTT (tempo di andata e ritorno). Se il tuo carico di lavoro fornisce agli utenti contenuti dinamici, prendi in considerazione l'utilizzo di Amazon CloudFront, che stabilisce una connessione persistente a ogni origine per il contenuto dinamico in modo da eliminare il tempo di configurazione della connessione, che altrimenti rallenterebbe ogni richiesta client.
3. L'uso di TLS con TCP o UDP può causare maggiore latenza e minore velocità di trasmissione effettiva per il carico di lavoro a causa dell'impatto della crittografia e della decrittografia. Per carichi di lavoro di questo tipo, prendi in considerazione l'offload SSL/TLS in Elastic Load Balancing per migliorare le prestazioni permettendo al sistema di bilanciamento del carico di gestire la crittografia e la decrittografia SSL/TLS invece di predisporre a questo scopo istanze

back-end. In questo modo, puoi ridurre l'utilizzo della CPU sulle istanze back-end, migliorando le prestazioni e aumentando la capacità.

4. Utilizza il Network Load Balancer (NLB) per implementare servizi basati sul protocollo UDP, tra cui autenticazione e autorizzazione, registrazione, DNS, IoT e streaming di contenuti multimediali, in modo da migliorare le prestazioni e l'affidabilità del carico di lavoro. L'NLB distribuisce il traffico UDP in ingresso tra più destinazioni, permettendo di aumentare o ridurre orizzontalmente il carico di lavoro, incrementare la capacità e diminuire il sovraccarico su un'unica destinazione.
5. Per i tuoi carichi di lavoro HPC (calcolo ad alte prestazioni), prendi in considerazione l'utilizzo della funzionalità Adattatore elastico di rete (ENA) Express , che usa il protocollo SRD per migliorare le prestazioni di rete fornendo una larghezza di banda a flusso singolo più elevata (25 Gbps) e una latenza di coda inferiore (99,9 percentile) per il traffico di rete tra istanze EC2.
6. Utilizza l'Application Load Balancer (ALB) per instradare il traffico gRPC (Remote Procedure Call) tra componenti del carico di lavoro o tra client e servizi gRPC e per bilanciarne il carico. gRPC usa il protocollo HTTP/2 basato su TCP per il trasporto e fornisce vantaggi in termini di prestazioni, tra cui un impatto di rete minore, la compressione, la serializzazione binaria efficiente, il supporto per diversi linguaggi e lo streaming bidirezionale.

Risorse

Documenti correlati:

- How to route UDP traffic into Kubernetes
- 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
- Passaggio all'instradamento basato sulla latenza in Amazon Route 53
- Endpoint VPC

Video correlati:

- AWS re:Invent 2022 – Scaling network performance on next-gen Amazon Elastic Compute Cloud instances
- AWS re:Invent 2022 – Application networking foundations

Esempi correlati:

- AWS Transit Gateway and Scalable Security Solutions
- AWS Networking Workshops

PERF04-BP06 Scelta della posizione del carico di lavoro in base ai requisiti di rete

Valuta le opzioni per il posizionamento delle risorse in modo da diminuire la latenza di rete e migliorare la velocità di trasmissione effettiva, fornendo un'esperienza utente ottimale attraverso la riduzione dei tempi di caricamento delle pagine e di trasferimento dei dati.

Anti-pattern comuni:

- Consolidamento di tutte le risorse del carico di lavoro in un'unica posizione geografica.
- Scelta della regione più vicina alla propria posizione, ma non al carico di lavoro dell'utente finale.

Vantaggi dell'adozione di questa best practice: l'esperienza utente è fortemente condizionata dalla latenza tra utente e applicazione. Utilizzando le Regioni AWS appropriate e una rete globale AWS privata, puoi ridurre la latenza e offrire un'esperienza migliore agli utenti remoti.

Livello di rischio associato alla mancata adozione di questa best practice: medio

Guida all'implementazione

Le risorse, ad esempio le istanze Amazon EC2, vengono posizionate in zone di disponibilità all'interno delle Regioni AWS, in zone locali AWS, in AWS Outposts o in zone AWS Wavelength. La scelta della posizione influisce sulla latenza di rete e sulla velocità di trasmissione effettiva dall'ubicazione di un utente specifico. È anche possibile usare i servizi edge, quali Amazon CloudFront e AWS Global Accelerator, per migliorare le prestazioni della rete memorizzando i contenuti nella cache delle posizioni edge o fornendo agli utenti il percorso ottimale del carico di lavoro tramite la rete globale AWS.

Amazon EC2 offre gruppi di collocazione per le reti. Un gruppo di collocazione è un raggruppamento logico di istanze per ridurre la latenza. L'utilizzo di gruppi di collocazione con tipi di istanza supportati e un Adattatore elastico di rete (ENA) consente ai carichi di lavoro di partecipare a una rete a 25 Gbps a bassa latenza e a jitter ridotto. 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.

I servizi sensibili alla latenza vengono forniti nelle posizioni edge utilizzando una rete AWS globale, ad esempio Amazon CloudFront. Tali posizioni edge forniscono solitamente servizi come rete di distribuzione di contenuti (CDN) e sistema dei nomi di dominio (DNS). 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).

Usa i servizi edge per ridurre la latenza e abilitare la memorizzazione nella cache dei contenuti. Configura correttamente il controllo cache sia per DNS sia per HTTP/HTTPS al fine di sfruttare tutti i vantaggi offerti da tali approcci.

Passaggi dell'implementazione

- Acquisisci informazioni sul traffico IP in entrata e in uscita dalle interfacce di rete.
  - Log del traffico IP tramite log di flusso VPC
  - Come viene conservato l’indirizzo IP del client in AWS Global Accelerator
- Analizza i modelli di accesso alla rete nel tuo carico di lavoro per capire come gli utenti usano la tua applicazione.
  - Usa strumenti di monitoraggio, come Amazon CloudWatch e AWS CloudTrail, per raccogliere i dati sull'attività della rete.
  - Analizza i dati per identificare il modello di accesso alla rete.
- Seleziona regioni appropriate per l'implementazione del carico di lavoro in base ai seguenti elementi chiave:
  - Dove si trovano i tuoi 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.
  - Dove si trovano i tuoi utenti: per le applicazioni rivolte agli utenti, scegli una o più regioni vicine agli utenti del tuo carico di lavoro.

  - Altre limitazioni: considera le limitazioni relative a costi e conformità, come spiegato nel post What to Consider when Selecting a Region for your Workloads..
- Usa le zone locali AWS per eseguire carichi di lavoro come il rendering di video. Le zone locali consentono di sfruttare i vantaggi derivanti dalla disponibilità di risorse di calcolo e archiviazione più vicine agli utenti finali.
- Usa AWS Outposts per carichi di lavoro che devono rimanere in locale, ma vuoi che vengano eseguiti in modo ottimale con il resto degli altri carichi di lavoro in AWS.
- Applicazioni come quelle di streaming di video live ad alta risoluzione, audio ad alta fedeltà o realtà aumentata o realtà virtuale (AR/VR) richiedono latenza bassissima per i dispositivi 5G. Per applicazioni di questo tipo, prendi in considerazione AWS Wavelength. AWS Wavelength incorpora servizi di calcolo e archiviazione AWS in reti 5G, fornendo un'infrastruttura di edge computing per dispositivi mobili per lo sviluppo, l'implementazione e il dimensionamento di applicazioni a latenza bassissima.
- Usa la cache locale o le Soluzioni per la acche di AWS per le risorse di frequente utilizzo per migliorare le performance, ridurre lo spostamento dei dati e minimizzare l'impatto ambientale.

Service	When to use
Amazon CloudFront	Usa per memorizzare nella cache contenuti statici come immagini, script e video, nonché contenuti dinamici come risposte API o applicazioni Web.
Amazon ElastiCache	Usa per memorizzare nella cache i contenuti per le applicazioni Web.
DynamoDB Accelerator	Usa 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, come i seguenti:

Service	When to use
Lambda@edge	Usa per operazioni a uso intensivo di risorse di calcolo eseguite quando gli oggetti non si trovano nella cache.
Funzioni Amazon CloudFront	Usa per casi d'uso semplici, ad esempio manipolazioni di risposte o richieste HTTP(s) che possono essere avviate da funzioni di breve durata.
AWS IoT Greengrass	Usa per eseguire la memorizzazione nella cache di risorse di calcolo, messaggistica e dati per i dispositivi connessi.

- Alcune applicazioni richiedono punti di ingresso fissi o prestazioni più elevate attraverso la riduzione della latenza di ricezione del primo byte e l'instabilità e l'aumento della velocità di trasmissione effettiva. Queste applicazioni possono trarre vantaggio da servizi di rete che forniscono indirizzi IP anycast statici e terminazione TCP in posizioni edge. AWS Global Accelerator può migliorare le prestazioni per le applicazioni fino al 60% e offre un failover rapido per architetture in più regioni. AWS Global Accelerator fornisce indirizzi IP anycast statici che fungono da punto di ingresso fisso per le applicazioni ospitate in una o più Regioni AWS. Questi indirizzi IP permettono l'ingresso del traffico nella rete AWS globale più vicina possibile agli utenti. AWS Global Accelerator riduce il tempo di configurazione della connessione iniziale stabilendo una connessione TCP tra il client e la posizione edge di AWS più vicina al client. Riesamina l'uso di AWS Global Accelerator per migliorare le prestazioni dei carichi di lavoro TCP/UDP e fornire il rapido failover per architetture in più regioni.

Risorse

Best practice correlate:

- COST07-BP02 Implementazione delle regioni in base al costo
- COST08-BP03 Implementazione dei servizi per ridurre il costo di trasferimento dei dati
- REL10-BP01 Implementazione del carico di lavoro in diversi luoghi
- REL10-BP02 Selezione delle posizioni appropriate per la tua implementazione multiposizione

- SUS01-BP01 Scelta della Regione in base alle esigenze aziendali e agli obiettivi di sostenibilità.
- SUS02-BP04 Ottimizzazione del posizionamento geografico dei carichi di lavoro in base ai requisiti di rete
- SUS04-BP07 Riduzione al minimo dello spostamento di dati tra reti

Documenti correlati:

- Infrastruttura globale AWS
- Zone locali AWS e AWS Outposts, scelta della giusta tecnologia per un carico di lavoro edge
- Gruppi di collocazione
- Zone locali AWS
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Video correlati:

- Video di presentazione delle zone locali AWS
- AWS Outposts: Overview and How it Works
- AWS re:Invent 2023 - A migration strategy for edge and on-premises workloads
- AWS re:Invent 2021: AWS Outposts: Spostamento dell'esperienza AWS in un ambiente on-premise
- AWS re:Invent 2020: AWS Wavelength: esecuzione di app con latenza bassissima nell'edge 5G
- AWS re:Invent 2022: Zone locali AWS: creazione di applicazioni per una posizione edge distribuita
- AWS re:Invent 2021: Creazione di siti Web a bassa latenza con Amazon CloudFront
- AWS re:Invent 2022: Miglioramento delle prestazioni e della disponibilità con AWS Global Accelerator
- AWS re:Invent 2022: Creazione di una rete WAN usando AWS

- AWS re:Invent 2020: Gestione del traffico globale con Amazon Route 53

Esempi correlati:

- AWS Global Accelerator Custom Routing Workshop
- Gestione delle riscritture e dei reindirizzamenti usando funzioni di edge computing

PERF04-BP07 Ottimizzazione della configurazione di rete in base alle metriche

Usa i dati raccolti e analizzati per prendere decisioni informate riguardo l'ottimizzazione della configurazione della tua rete.

Anti-pattern comuni:

- Ritieni che tutti i problemi relativi alle prestazioni siano correlati all'applicazione.
- Verifica delle prestazioni di rete solo da una posizione vicina a quella in cui è stato distribuito il carico di lavoro.
- Uso di configurazioni predefinite per tutti i servizi di rete.
- Provisioning in eccesso di risorse di rete per fornire capacità sufficiente.

Vantaggi dell'adozione di questa best practice: la raccolta delle metriche necessarie per la rete AWS e l'implementazione di strumenti di monitoraggio di rete permettono di identificare le prestazioni di rete e ottimizzare le configurazioni di rete.

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

Guida all'implementazione

Il monitoraggio del traffico da e verso VPC, sottoreti o interfacce di rete è essenziale per identificare come utilizzare risorse di rete AWS e ottimizzare le configurazioni di rete. Usando i seguenti strumenti di rete AWS, puoi esaminare ulteriormente le informazioni sull'utilizzo del traffico, sull'accesso alla rete e sui log.

Passaggi dell'implementazione

- Identifica le metriche delle prestazioni fondamentali da raccogliere, come la latenza o la perdita di pacchetti. AWS fornisce diversi strumenti che possono aiutarti a raccogliere queste metriche. Usando i seguenti strumenti, puoi esaminare ulteriormente le informazioni sull'utilizzo del traffico, sull'accesso alla rete e sui log:

Strumento AWS	Dove usare
Amazon VPC IP Address Manager.	Utilizza IPAM per pianificare, seguire e monitorare gli indirizzi IP per i carichi di lavoro AWS e on-premise. Si tratta di una best practice per ottimizzare l'utilizzo e l'allocazione degli indirizzi IP.
Log di flusso VPC	Usa log di flusso VPC per acquisire informazioni dettagliate sul traffico da e verso le interfacce di rete nei VPC. Con i log di flusso VPC puoi diagnosticare regole dei gruppi di sicurezza eccessivamente restrittive o permissive e determinare la direzione del traffico da e verso le interfacce di rete.
Log di flusso AWS Transit Gateway	Utilizza i log di flusso AWS Transit Gateway per acquisire informazioni sul traffico IP in entrata e in uscita dai gateway di transito.
Registrazione di query DNS	Registra le informazioni sulle query DNS pubbliche o private ricevute da Route 53. Con i log DNS puoi ottimizzare le configurazioni DNS identificando il dominio e il sottodomi nio richiesto o le posizioni edge Route 53 che hanno risposto a query DNS.

Strumento AWS	Dove usare
Reachability Analyzer	Reachability Analyzer ti aiuta a effettuare l'analisi e il debug della raggiungibilità della rete. Reachability Analyzer è uno strumento di analisi della configurazione che permette di eseguire test di connettività tra una risorsa di origine e una risorsa di destinazione nei VPC. Questo strumento permette di verificare che la configurazione di rete corrisponda alla connettività desiderata.
Network Access Analyzer	Network Access Analyzer ti aiuta a definire l'accesso alla rete per le tue risorse. Puoi usare Network Access Analyzer per specifica re i requisiti di accesso alla rete e identificare i potenziali percorsi di rete che non li soddisfan o. Ottimizzando la configurazione di rete corrispondente, puoi determinare e verificare lo stato della rete e indicare se la rete su AWS soddisfa i requisiti di conformità.
Amazon CloudWatch	Utilizza Amazon CloudWatch e attiva le metriche appropriate per le opzioni di rete. Assicurati di scegliere le metriche di rete corrette per il carico di lavoro. Ad esempio, puoi attivare le metriche per l'utilizzo degli indirizzi di rete del VPC, il gateway NAT del VPC, AWS Transit Gateway, il tunnel VPN, AWS Network Firewall, Elastic Load Balancing e AWS Direct Connect. Il monitorag gio continuo delle metriche è una procedura utile per osservare e identificare lo stato e l'utilizzo della rete che semplifica l'ottimiz zazione della configurazione di rete in base alle osservazioni.

Strumento AWS	Dove usare
AWS Network Manager	Con AWS Network Manager puoi monitorar e le prestazioni in tempo reale e storiche della rete globale AWS per scopi operativi e di pianificazione. Network Manager fornisce una latenza di rete aggregata tra Regioni AWS e zone di disponibilità e all'interno di ciascuna zona di disponibilità, permettendoti di comprendere meglio in che modo le prestazio ni delle applicazioni si relazionano con le prestazioni della rete AWS sottostante.
Amazon CloudWatch RUM	Usa Amazon CloudWatch RUM per raccoglie re le metriche che ti consentono di ottenere approfondimenti utili per identificare, comprendere e migliorare l'esperienza utente.

- Identifica i top talker e gli schemi di traffico delle applicazioni utilizzando VPC e i log di flusso di AWS Transit Gateway.
- Valuta e ottimizza la tua attuale architettura di rete, inclusi VPC, sottoreti e routing. Ad esempio, puoi valutare come i diversi VPC per il peering o AWS Transit Gateway possono aiutarti a migliorare la rete nella tua architettura.
- Valuta i percorsi di routing nella tua rete per verificare che venga sempre utilizzato il percorso più breve tra le destinazioni. Network Access Analyzer può aiutarti a farlo.

Risorse

Documenti correlati:

- Registrazione delle query DNS pubbliche
- Che cos'è IPAM?
- What is Reachability Analyzer?
- What is Network Access Analyzer?
- Parametri di CloudWatch per i VPC

- Ottimizzazione delle prestazioni e riduzione dei costi per l'analisi della rete con log di flusso VPC in formato Apache Parquet
- Monitoring your global and core networks with Amazon CloudWatch metrics
- Continuously monitor network traffic and resources

Video correlati:

- AWS re:Invent 2023 – A developer's guide to cloud networking
- AWS re:Invent 2023 – Ready for what’s next? Designing networks for growth and flexibility
- AWS re:Invent 2023 – Advanced VPC designs and new capabilities
- AWS re:Invent 2022 – Dive deep on AWS networking infrastructure
- AWS re:Invent 2020 – Networking best practices and tips with the AWS Well-Architected Framework
- AWS re:Invent 2020 – Monitoring and troubleshooting network traffic

Esempi correlati:

- AWS Networking Workshops
- AWS Network Monitoring
- Observing and diagnosing your network on AWS
- Finding and addressing network misconfigurations on AWS

Processo e cultura

Durante la fase di progettazione dei carichi di lavoro, esistono principi e pratiche che è possibile adottare per gestire al meglio carichi di lavoro cloud efficienti e ad alte prestazioni. Questa area di interesse offre le best practice per aiutarti ad adottare una cultura che promuova l'efficienza delle prestazioni dei carichi di lavoro cloud.

Per sviluppare questa cultura, considera questi principi chiave:

- Scrittura dell'infrastruttura come codice: definisci la tua infrastruttura come codice utilizzando approcci come i modelli AWS CloudFormation. L'uso dei modelli ti consente di collocare la tua infrastruttura nel controllo sorgente, insieme al codice e alle configurazioni dell'applicazione. Ciò ti premette di applicare le stesse procedure di sviluppo software all'infrastruttura, in modo da accelerare l'iterazione.
- Pipeline di distribuzione: usa una pipeline di integrazione continua/distribuzione continua (CI/CD) (ad esempio repository del codice sorgente, sistemi di sviluppo, distribuzione e automazione dei test) per distribuire la tua infrastruttura. Ciò ti consente di effettuare l'implementazione in modo ripetibile, coerente ed economicamente vantaggioso nel corso dell'iterazione.
- Parametri ben definiti: configura e monitora le metriche per raccogliere gli indicatori chiave di prestazione (KPI). Ti consigliamo di adottare parametri tecnici e aziendali. Per i siti Web o le app mobili, le metriche principali sono il tempo di acquisizione al primo byte o il rendering. Gli altri parametri generalmente validi includono il numero di thread, il tasso di raccolta di dati superflui e gli stati di attesa. I parametri aziendali, come il costo cumulativo aggregato per richiesta, possono indicarti due modi per ridurre i costi. Valuta attentamente il modo in cui prevedi di interpretare i parametri. Ad esempio, potresti scegliere il 99° percentile o quello massimo anziché il valore medio.
- Automatizza i test delle prestazioni: nell'ambito del processo di implementazione, avvia automaticamente i test delle prestazioni dopo che quelli dall'esecuzione più rapida hanno dato esito positivo. L'automazione deve creare un nuovo ambiente, configurare le condizioni iniziali come i dati del test ed eseguire una serie di benchmark e test di carico. I risultati dei test devono essere confrontati con la build, in modo da monitorare le variazioni delle prestazioni nel corso del tempo. Per i test di lunga durata, puoi inserirli nella pipeline in maniera asincrona rispetto al resto della build. In alternativa, puoi eseguire i test delle prestazioni negli orari notturni, tramite le istanze Spot di Amazon EC2.
- Generazione del carico: crea una serie di script di test che replichino percorsi utente sintetici o pre-registrati. Tali script devono essere idempotenti e non accoppiati. Inoltre, potrebbe essere

necessario includere script preliminari per ottenere risultati validi. Testa gli script il più possibile, per assicurarti che replichino le abitudini di utilizzo in produzione. Puoi usare soluzioni software o SaaS (Software-as-a-Service) per generare il carico. Prendi in considerazione l'utilizzo di soluzioni Marketplace AWS e istanze spot , dal momento che potrebbero rappresentare metodi economicamente vantaggiosi per la generazione del carico.

- Visibilità delle prestazioni: i parametri principali devono essere visibili dal team, in particolar modo quelli relativi a ciascuna versione della build. Ciò ti consente di rilevare tendenze positive o negative rilevanti nel corso del tempo. Dovresti anche visualizzare i parametri sul numero di errori o eccezioni per assicurarti di testare un sistema funzionante.
- Visualizzazione: sfrutta le tecniche di visualizzazione che indicano in modo chiaro i punti in cui si verificano problemi di prestazioni, hot spot, stati di attesa o utilizzo ridotto. Sovrapponi i parametri delle prestazioni ai diagrammi architetturali: i grafici delle chiamate o il codice possono aiutarti a individuare più rapidamente i problemi.
- Processo di revisione regolare: le prestazioni scarse delle architetture sono in genere il risultato di un processo di revisione delle prestazioni inesistente o incompleto. Se la tua architettura offre prestazioni insufficienti, l'implementazione di un processo di revisione delle prestazioni ti consente di favorire il miglioramento delle iterazioni.
- Ottimizzazione continua: adotta una cultura per ottimizzare continuamente l'efficienza delle prestazioni del tuo carico di lavoro cloud.

Best practice

- PERF05-BP01 Individuazione degli indicatori chiave di prestazioni (KPI) per misurare l'integrità e le prestazioni del carico di lavoro
- PERF05-BP02 Uso di soluzioni di monitoraggio per comprendere le aree in cui le prestazioni sono più critiche
- PERF05-BP03 Definizione di un processo per migliorare le prestazioni del carico di lavoro
- PERF05-BP04 Esecuzione del test del carico di lavoro
- PERF05-BP05 Uso dell'automazione per risolvere in modo proattivo i problemi relativi alle prestazioni
- PERF05-BP06 Aggiornamento continuo del carico di lavoro e dei servizi
- PERF05-BP07 Analisi dei parametri a intervalli regolari

PERF05-BP01 Individuazione degli indicatori chiave di prestazioni (KPI) per misurare l'integrità e le prestazioni del carico di lavoro

Individua gli indicatori chiave di prestazione (KPI) per misurare le prestazioni del carico di lavoro. I KPI consentono di misurare l'integrità e le prestazioni di un carico di lavoro correlato a un obiettivo aziendale.

Anti-pattern comuni:

- Monitori i parametri a livello di sistema solo per avere una visione del carico di lavoro e non valuti gli impatti aziendali di tali parametri.
- Ritieni che i KPI siano già in fase di pubblicazione e condivisi come dati parametrici standard.
- Non definisci un KPI quantitativo e misurabile.
- Non esegui l'allineamento dei KPI a obiettivi o strategie aziendali.

Vantaggi dell'adozione di questa best practice: l'individuazione di KPI specifici che rappresentino l'integrità e le prestazioni del carico di lavoro aiuta ad allineare i team alle priorità e a definire risultati aziendali ottimali. 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: elevato

Guida all'implementazione

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 fattori si combinano per produrre risultati aziendali. Ad esempio, il carico di lavoro di un sito Web può utilizzare il tempo di caricamento della pagina come indicazione delle prestazioni complessive. Questa metrica sarebbe uno dei molteplici punti dati che misurano l'esperienza dell'utente. Oltre a identificare le soglie di tempo di caricamento della pagina, è necessario documentare il risultato atteso o il rischio aziendale se le prestazioni ideali 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).

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 e includere le soglie che supportano gli obiettivi e le strategie aziendali e che sono mappati sui parametri da monitorare. Gli indicatori chiave di prestazione devono essere riesaminati quando cambiano gli obiettivi aziendali, le strategie o i requisiti degli utenti finali.

Passaggi dell'implementazione

- Individua le parti interessate: identifica e documenta le principali parti interessate aziendali, compresi i team di sviluppo e operativi.
- Definisci gli obiettivi: collabora con queste parti interessate per definire e documentare gli obiettivi del carico di lavoro. Considera gli aspetti critici relativi alle prestazioni dei carichi di lavoro, come il throughput, i tempi di risposta e i costi, nonché gli obiettivi aziendali, come la soddisfazione degli utenti.
- Esamina le best practice del settore: esamina le best practice del settore per individuare i KPI pertinenti in linea con gli obiettivi del carico di lavoro.
- Individua le metriche: identifica le metriche che sono allineate agli obiettivi del carico di lavoro e che possono aiutarti a misurare le prestazioni e gli obiettivi aziendali. Stabilisci i KPI in base a queste metriche, ad esempio le misurazioni del tempo medio di risposta o del numero di utenti simultanei.
- Definisci e documenta i KPI: utilizza le best practice del settore e gli obiettivi del carico di lavoro per stabilire i valori dei KPI del carico di lavoro. Utilizza queste informazioni per impostare soglie dei KPI per livello di gravità o allarme. Individua e documenta il rischio e l'impatto se il KPI non viene raggiunto.
- Implementa il monitoraggio: utilizza gli strumenti di monitoraggio come Amazon CloudWatch o AWS Config per raccogliere le metriche e misurare i KPI.
- Comunica visivamente i KPI: utilizza gli strumenti della dashboard, come Amazon QuickSight, per visualizzare e comunicare i KPI alle parti interessate.
- Analizza e ottimizza: esamina e analizza regolarmente i KPI per identificare le aree del carico di lavoro che devono essere migliorate. Collabora con le parti interessate per implementare i miglioramenti.
- Riesamina e perfeziona: revisiona regolarmente le metriche e i KPI per valutarne l'efficacia, soprattutto quando cambiano gli obiettivi aziendali o le prestazioni del carico di lavoro.

Risorse

Documenti correlati:

- Documentazione CloudWatch
- Monitoraggio, registrazione di log e prestazioni - AWS Partner
- AWS observability tools
- The Importance of Key Performance Indicators (KPIs) for Large-Scale Cloud Migrations
- How to track your cost optimization KPIs with the KPI Dashboard
- DocumentazioneX-Ray
- Using Amazon CloudWatch dashboards
- Amazon QuickSight KPIs

Video correlati:

- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2023 - Manage resource lifecycle events at scale with AWS Health
- AWS re:Invent 2023 - Performance & efficiency at Pinterest: Optimizing the latest instances
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2023 - Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2023 - Scaling on AWS for the first 10 million users
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- Creating an Effective Metrics Strategy for Your Business | AWS Events

Esempi correlati:

- Creating a dashboard with Amazon QuickSight

PERF05-BP02 Uso di soluzioni di monitoraggio per comprendere le aree in cui le prestazioni sono più critiche

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.

Anti-pattern comuni:

- Ritieni che i parametri di calcolo standard, ad esempio l'utilizzo della CPU o il carico della memoria, siano sufficienti per rilevare 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: la comprensione delle aree critiche delle prestazioni aiuta i proprietari dei carichi di lavoro a monitorare i KPI e a dare priorità ai miglioramenti ad alto impatto.

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

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.

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. Dopo aver identificato queste aree, individua quale soluzione puoi implementare per evitare tali problemi di prestazioni.

Passaggi dell'implementazione

- Configura il monitoraggio end-to-end per acquisire tutti i componenti e i parametri del carico di lavoro. Ecco alcuni esempi di soluzioni di monitoraggio su AWS.

Service	Where to use
Amazon CloudWatch Real-User Monitoring (RUM)	To capture application performance metrics from real user client-side and frontend sessions.
AWS X-Ray	To trace traffic through the application layers and identify latency between components and dependencies. Use X-Ray service maps to see relationships and latency between workload components.
Amazon Relational Database Service Performance Insights	To view database performance metrics and identify performance improvements.
Monitoraggio dei parametri del sistema operativo con il monitoraggio avanzato - Amazon RDS	To view database OS performance metrics.
Amazon DevOps Guru	To detect abnormal operating patterns so you can identify operational issues before they impact your customers.

- Esegui i test per generare parametri, identificare schemi di traffico, colli di bottiglia e aree con prestazioni critiche. Ecco alcuni esempi di come eseguire i test:
  - Configura i canary Synthetics di CloudWatch per simulare le attività degli utenti basate sul browser in modo programmatico utilizzando espressioni di valutazione o processi CRON di Linux per generare parametri coerenti nel tempo.
  - Utilizza la soluzione per il test di carico distribuito AWS per generare picchi di traffico o testare il carico di lavoro al tasso di crescita previsto.
- 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.
- Sperimenta i miglioramenti delle prestazioni e valuta tali modifiche con i dati. Ad esempio, puoi usare CloudWatch Evidently per testare i nuovi miglioramenti e l'impatto sulle prestazioni del tuo carico di lavoro.

Risorse

Documenti correlati:

- What's new in AWS Observability at re:Invent 2023
- Amazon Builders' Library
- DocumentazioneX-Ray
- Amazon CloudWatch RUM
- Amazon DevOps Guru

Video correlati:

- AWS re:Invent 2023 - [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 - Implementing application observability
- AWS re:Invent 2023 - Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2022 - The Amazon Builders’ Library: 25 years of Amazon operational excellence
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- Visual Monitoring of Applications with Amazon CloudWatch Synthetics

Esempi correlati:

- Measure page load time with Amazon CloudWatch Synthetics
- Amazon CloudWatch RUM Web Client
- X-Ray SDK for Python
- Distributed Load Testing on AWS

PERF05-BP03 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.

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 dei parametri che le giustifichino.

Vantaggi dell'adozione di questa best practice: definendo il processo per apportare modifiche all'architettura, puoi utilizzare i dati raccolti per influenzare la progettazione del carico di lavoro nel tempo.

Livello di rischio associato alla mancata adozione di questa best practice: medio

Guida all'implementazione

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.

Determina i principali vincoli riguardanti le prestazioni del carico di lavoro. Documenta i vincoli prestazionali del carico di lavoro in modo da sapere quali tipi di innovazione potrebbero migliorare le prestazioni del carico di lavoro.

Passaggi dell'implementazione

- Individua i KPI: determina i KPI delle prestazioni del carico di lavoro come indicato in PERF05-BP01 Individuazione degli indicatori chiave di prestazioni (KPI) per misurare l'integrità e le prestazioni del carico di lavoro per definire la baseline del carico di lavoro.
- Implementa il monitoraggio: utilizza gli strumenti di osservabilità AWS per raccogliere le metriche sulle prestazioni e misurare i KPI.
- Esegui l'analisi: conduci un'analisi approfondita per individuare le aree del carico di lavoro, ad esempio la configurazione e il codice applicativo, con prestazioni insufficienti, come indicato in

PERF05-BP02 Uso di soluzioni di monitoraggio per comprendere le aree in cui le prestazioni sono più critiche. Usa i tuoi strumenti di analisi e prestazioni per individuare la strategia di miglioramento delle prestazioni.

- Convalida i miglioramenti: utilizza ambienti sandbox o di pre-produzione per convalidare l'efficacia delle strategie di miglioramento.
- Apporta le modifiche: implementa le modifiche in produzione e monitora continuamente le prestazioni del carico di lavoro. Documenta i miglioramenti e comunica i risultati alle parti interessate.
- Riesamina e perfeziona: revisiona regolarmente il processo di miglioramento delle prestazioni per identificare le aree da potenziare.

Risorse

Documenti correlati:

- Blog AWS
- Novità di AWS
- AWS Skill Builder

Video correlati:

- AWS re:Invent 2022 - Delivering sustainable, high-performing architectures
- AWS re:Invent 2023 - Optimize cost and performance and track progress toward mitigation
- AWS re:Invent 2022 - AWS optimization: Actionable steps for immediate results
- AWS re:Invent 2022 - Optimize your AWS workloads with best-practice guidance

Esempi correlati:

- AWS Github

PERF05-BP04 Esecuzione del test del carico di lavoro

Esegui il test del carico di lavoro per verificare che sia in grado di gestire la produzione e individuare eventuali colli di bottiglia nelle prestazioni.

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 il test di carico senza consultare la policy di test di Amazon EC2 e presentare un modulo di invio di eventi simulati. Ciò comporta la mancata esecuzione del test, poiché sembra un evento di negazione del servizio.

Vantaggi dell'adozione di questa best practice: misurando le prestazioni con un test di carico puoi osservare dove avrà luogo l'impatto dell'aumento del carico. In questo modo puoi anticipare le modifiche necessarie prima che influiscano sul carico di lavoro.

Livello di rischio associato alla mancata adozione di questa best practice: basso

Guida all'implementazione

Il test di carico nel cloud è un processo per misurare le prestazioni del carico di lavoro in condizioni realistiche e con il carico degli utenti previsto. Questo processo prevede il provisioning di un ambiente cloud simile a quello di produzione, l'utilizzo di strumenti di test di carico per generare il carico e l'analisi dei parametri per valutare la capacità del carico di lavoro di gestire un carico realistico. 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). Esegui automaticamente test di carico come parte della pipeline di distribuzione e confronta i risultati con KPI e soglie predefiniti. Questo processo ti consente di ottenere le prestazioni richieste.

Passaggi dell'implementazione

- Definisci gli obiettivi del test: individua gli aspetti prestazionali del carico di lavoro che desideri valutare, come il throughput e il tempo di risposta.
- Seleziona uno strumento di test: scegli e configura lo strumento di test più adatto al tuo carico di lavoro.
- Configura l'ambiente: configura l'ambiente di test in base all'ambiente di produzione. Puoi utilizzare i servizi AWS per eseguire ambienti in ambito di produzione e sottoporre l'architettura a test.

- Implementa il monitoraggio: utilizza gli strumenti di monitoraggio, ad esempio Amazon CloudWatch, per raccogliere le metriche delle risorse della tua architettura. Puoi anche raccogliere e pubblicare metriche personalizzate.
- Definisci gli scenari: stabilisci gli scenari e i parametri del test di carico, come la durata del test e il numero di utenti.
- Conduci il test di carico: esegui gli scenari di test su larga scala. 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.
- Analizza i risultati del test: analizza i risultati per individuare i colli di bottiglia delle prestazioni e le aree di miglioramento.
- Documenta e condividi gli esiti: crea i documenti per comunicare i risultati e le raccomandazioni. Condividi queste informazioni con le parti interessate per aiutarle a prendere decisioni informate sulle strategie di ottimizzazione delle prestazioni.
- Itera in modo continuo: i test di carico devono essere eseguiti a cadenza regolare, soprattutto dopo una modifica o un aggiornamento del sistema.

Risorse

Documenti correlati:

- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Distributed Load Testing on AWS

Video correlati:

- AWS Summit ANZ 2023: Accelerate with confidence through AWS Distributed Load Testing
- AWS re:Invent 2022 - Scaling on AWS for your first 10 million users
- Solving with AWS Solutions: Distributed Load Testing

- AWS re:Invent 2021 - Ottimizzare le applicazioni con gli approfondimenti degli utenti finali con Amazon CloudWatch RUM
- Demo di Amazon CloudWatch Synthetics

Esempi correlati:

- Distributed Load Testing on AWS

PERF05-BP05 Uso dell'automazione per risolvere in modo proattivo i problemi relativi alle prestazioni

Utilizza indicatori chiave di prestazioni (KPI), in combinazione con sistemi di monitoraggio e allarmi, per risolvere in modo proattivo i problemi correlati alle prestazioni.

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 team 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 alla mancata adozione di questa best practice: basso

Guida all'implementazione

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 gli indicatori chiave di prestazioni (KPI) e di inviare allarmi qualora essi oltrepassino determinate soglie, oppure uno strumento che arresta o esegue automaticamente il rollback delle implementazioni 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.

Passaggi dell'implementazione

- Individua il flusso di lavoro della risoluzione: identifica e comprendi il problema delle prestazioni che può essere risolto automaticamente. Utilizza soluzioni di monitoraggio AWS come Amazon CloudWatch o AWS X-Ray per comprendere meglio la causa principale del problema.
- Definisci il processo di automazione: crea un processo di risoluzione dettagliato che possa essere utilizzato per risolvere automaticamente il problema.
- Configura l'evento di avvio: definisci l'evento che avvia automaticamente il processo di risoluzione. Ad esempio, è possibile definire un trigger per riavviare automaticamente un'istanza quando raggiunge una determinata soglia di utilizzo della CPU.
- Automatizza la risoluzione: utilizza i servizi e le tecnologie AWS per automatizzare il processo di risoluzione. Ad esempio, l'automazione AWS Systems Manager fornisce un modo sicuro e dimensionabile per automatizzare il processo di risoluzione. Assicurati di utilizzare la logica di risoluzione automatica per annullare le modifiche se non risolvono correttamente il problema.
- Esegui il test del flusso di lavoro: esegui il test del processo di risoluzione automatizzato in un ambiente di pre-produzione.
- Implementa il flusso di lavoro: implementa la risoluzione automatizzata nell'ambiente di produzione.
- Sviluppa un playbook: crea e documenta un playbook che delinei i passaggi per il piano di risoluzione, inclusi gli eventi di avvio, la logica di risoluzione e le azioni intraprese. Assicurati di fornire la giusta preparazione alle parti interessate per aiutarle a rispondere efficacemente agli eventi di risoluzione automatizzati.
- Rivedi e perfeziona: valuta regolarmente l'efficacia del flusso di lavoro di risoluzione automatizzato. Modifica gli eventi di avvio e la logica di risoluzione, se necessario.

Risorse

Documenti correlati:

- Documentazione CloudWatch
- Monitoraggio, registrazione di log e prestazioni - Partner AWS Partner Network
- DocumentazioneX-Ray
- Using Alarms and Alarm Actions in CloudWatch
- Build a Cloud Automation Practice for Operational Excellence: Best Practices from AWS Managed Services

- Automate your Amazon Redshift performance tuning with automatic table optimization

Video correlati:

- AWS re:Invent 2023 - Strategies for automated scaling, remediation, and smart self-healing
- AWS re:Invent 2023 - [LAUNCH] Application monitoring for modern workloads
- AWS re:Invent 2023 - Implementing application observability
- AWS re:Invent 2021 - Intelligently automating cloud operations
- AWS re:Invent 2022 - Setting up controls at scale in your AWS environment
- AWS re:Invent 2022 - Automating patch management and compliance using AWS
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- AWS re:Invent 2023 - Take a load off: Diagnose & resolve performance issues with Amazon RDS
- AWS re:Invent 2021 -{New Launch} Automatically detect and resolve issues with Amazon DevOps Guru
- AWS re:Invent 2023 - Centralize your operations

Esempi correlati:

- CloudWatch Logs Customize Alarms

PERF05-BP06 Aggiornamento continuo del carico di lavoro e dei servizi

Rimani aggiornato sui nuovi servizi cloud per adottare funzionalità efficienti, rimuovere i problemi e migliorare l'efficienza complessiva delle prestazioni del tuo carico di lavoro.

Anti-pattern comuni:

- Ritieni che l'architettura corrente diventi statica e non venga 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.

Vantaggi dell'adozione di questa best practice: stabilendo un processo per rimanere aggiornati su nuovi servizi e offerte, è possibile adottare nuove capacità e funzionalità, risolvere problemi e migliorare le prestazioni dei carichi di lavoro.

Livello di rischio associato alla mancata adozione di questa best practice: basso

Guida all'implementazione

Valuta i modi per migliorare le prestazioni man mano che nuovi servizi, modelli di progettazione e funzionalità 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.

Passaggi dell'implementazione

- Esegui l'inventario del carico di lavoro: redigi l'inventario del software e dell'architettura del carico di lavoro e identifica i componenti che richiedono un aggiornamento.
- Individua le origini di aggiornamento: identifica le origini di notizie e aggiornamenti relative ai componenti del carico di lavoro. Ad esempio, puoi iscriverti al blog Novità di AWS per scoprire i prodotti che corrispondono al tuo componente del carico di lavoro. Puoi iscriverti al feed RSS o gestire le tue sottoscrizioni e-mail.
- Definisci una pianificazione degli aggiornamenti: stabilisci la pianificazione per valutare nuovi servizi e funzionalità per il carico di lavoro.
  - Puoi usare 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.
- Valuta il nuovo aggiornamento: individua le modalità di aggiornamento dei componenti del 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 per ottenere efficienza delle prestazioni.
- Usa l'automazione: utilizza l'automazione del processo di aggiornamento per ridurre il livello di impegno per implementare le nuove funzionalità e limitare gli errori causati dai processi manuali.
  - Puoi usare CI/CD per aggiornare automaticamente le AMI, le immagini di container e altri artefatti relativi alla tua applicazione cloud.

  - Puoi usare strumenti come AWS Systems Manager Patch Manager per automatizzare il processo degli aggiornamenti di sistema e pianificare le attività tramite Finestre di manutenzione AWS Systems Manager.
- Documenta il processo: crea i documenti per il processo di valutazione degli aggiornamenti e dei nuovi servizi. Fornisci ai proprietari 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.

Risorse

Documenti correlati:

- Blog AWS
- Novità di AWS
- Implementing up-to-date images with automated EC2 Image Builder pipelines

Video correlati:

- AWS re:Inforce 2022 - Automating patch management and compliance using AWS
- All Things Patch: AWS Systems Manager | AWS Events

Esempi correlati:

- Inventory and Patch Management
- One Observability Workshop

PERF05-BP07 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.

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 individuano, risolvano o prevengano correttamente 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 alla mancata adozione di questa best practice: medio

Guida all'implementazione

Migliora continuamente 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. Questo metodo ti aiuterà a migliorare la qualità dei parametri raccolti, in modo da prevenire o risolvere più rapidamente gli incidenti futuri.

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.

Passaggi dell'implementazione

- Definisci le metriche: individua le metriche prestazionali critiche da monitorare affinché siano allineate all'obiettivo del carico di lavoro, ad esempio il tempo di risposta e l'utilizzo delle risorse.
- Stabilisci le baseline: imposta una baseline e il valore desiderabile per ogni metrica. La baseline deve fornire i punti di riferimento per identificare deviazioni o anomalie.
- Imposta una cadenza: stabilisci una cadenza, ad esempio settimanale o mensile, per la revisione delle metriche critiche.
- Individua i problemi relativi alle prestazioni: durante ogni revisione, valuta le tendenze e le deviazioni dai valori della baseline. Cerca eventuali rallentamenti o anomalie nelle prestazioni. Per i problemi identificati, esegui un'analisi approfondita delle cause principali per comprendere il motivo più importante alla base del problema.
- Individua le azioni correttive: utilizza l'analisi per individuare le azioni correttive, come l'ottimizzazione dei parametri, la correzione di bug e il dimensionamento delle risorse.
- Documenta gli esiti: crea i documenti per comunicare gli esiti, inclusi i problemi identificati, le cause principali e le azioni correttive.

- Itera e migliora: valuta e perfeziona continuamente il processo di revisione delle metriche. Usa le indicazioni apprese dalla revisione precedente per migliorare il processo nel tempo.

Risorse

Documenti correlati:

- Documentazione CloudWatch
- Collect metrics and logs from Amazon EC2 Instances and on-premises servers with the CloudWatch Agent
- Query your metrics with CloudWatch Metrics Insights
- Monitoraggio, registrazione di log e prestazioni - Partner AWS Partner Network
- DocumentazioneX-Ray

Video correlati:

- AWS re:Invent 2022 - Setting up controls at scale in your AWS environment
- AWS re:Invent 2022 - How Amazon uses better metrics for improved website performance
- AWS re:Invent 2023 - Building an effective observability strategy
- AWS Summit SF 2022 - Full-stack observability and application monitoring with AWS
- AWS re:Invent 2023 - Take a load off: Diagnose & resolve performance issues with Amazon RDS

Esempi correlati:

- Creating a dashboard with Amazon QuickSight
- CloudWatch Dashboards

Conclusione

Raggiungere e mantenere l'efficienza delle prestazioni richiede un approccio basato sui dati. Devi prendere in considerazione in modo attivo gli schemi di accesso e i compromessi che ti permetteranno di ottimizzare ulteriormente le prestazioni. I processi di revisione basati su benchmark e test di carico ti permettono di selezionare i tipi di risorse e le configurazioni più adatte. Trattare l'infrastruttura come codice ti permetterà di fare evolvere l'architettura in modo rapido e sicuro, mentre potrai utilizzare i dati per prendere decisioni circostanziate in merito all'architettura stessa. Adoperare una combinazione di monitoraggio attivo e passivo ti aiuterà a mantenere costanti le prestazioni dell'architettura nel corso del tempo.

AWS si impegna sempre ad aiutarti a realizzare infrastrutture che siano efficienti dal punto di vista delle prestazioni e in grado di garantire valore all'azienda. Utilizza gli strumenti e le tecniche illustrati in questo documento per avere successo.

Collaboratori

Hanno contribuito alla stesura di questo documento:

- Sam Mokhtari, Senior Efficiency Lead Solutions Architect, Amazon Web Services
- Josh Hart, Solutions Architect, Amazon Web Services
- Richard Trabing, Solutions Architect, Amazon Web Services
- Brett Looney, Principal Solutions Architect, Amazon Web Services
- Nina Vogl, Principal Solutions Architect, Amazon Web Services
- Eric Pullen, Solutions Architect, Amazon Web Services
- Julien Lépine, Specialist SA Manager, Amazon Web Services
- Ronnen Slasky, Solutions Architect, Amazon Web Services

Approfondimenti

Per ulteriori informazioni, consulta le seguenti risorse:

- Framework AWS Well-Architected
- Centro di progettazione AWS

Revisioni del documento

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

Modifica	Descrizione	Data
Whitepaper aggiornato	Best practice aggiornate con nuova guida all'imple mentazione.	June 27, 2024
Aggiornamento importante e ristrutturazione	Il pilastro è stato ristrutturato in modo da avere cinque aree di best practice (3 in meno rispetto a prima). Il contenuto è stato raggruppato nelle cinque aree e aggiornato. Le nuove aree di best practice sono Scelta dell'architettura, Elaborazione e hardware, Gestione dati, Reti e distribuz ione di contenutie Processo e cultura.	October 3, 2023
Aggiornamento di minore entità	Rimuovere linguaggio non inclusivo.	April 13, 2023
Aggiornamenti per il nuovo canone	Best practice aggiornate con prontuario e nuove best practice aggiunte.	April 10, 2023
Whitepaper aggiornato	Best practice aggiornate con nuova guida all'imple mentazione.	December 15, 2022
Whitepaper aggiornato	Ampliamento delle best practice e aggiunta dei piani di miglioramento.	October 20, 2022

Aggiornamento di minore entità	Rimozione del linguaggio non inclusivo.	April 22, 2022
Aggiornamento di minore entità	Aggiunta del pilastro della sostenibilità all'introduzione.	December 2, 2021
Aggiornamenti di minore entità	Link aggiornati.	March 10, 2021
Aggiornamenti di minore entità	È stato modificato il timeout di AWS Lambda a 900 secondi e corretto il nome di Amazon Keyspaces (for Apache Cassandra).	October 5, 2020
Aggiornamento di minore entità	Correzione di un link danneggiato.	July 15, 2020
Aggiornamenti per il nuovo canone	Revisione e aggiornamento importanti dei contenuti	July 8, 2020
Whitepaper aggiornato	Aggiornamento minore per la correzione di problemi grammaticali	July 1, 2018
Whitepaper aggiornato	Aggiornamento del whitepaper per rispecchiare le modifiche apportate a AWS	November 1, 2017
Pubblicazione originale	Pubblicazione del Pilastro dell'efficienza delle prestazioni - Framework AWS Well-Architected.	November 1, 2016

AWS Glossary

For the latest AWS terminology, see the AWS glossary in the Glossario AWS Reference.