Framework AWS Well-Architected

Pilastro dell'efficienza delle prestazioni

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

Table of Contents

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

Data di pubblicazione: 6 novembre 2024 (Revisioni del documento)

Il presente whitepaper riguarda il pilastro dell'efficienza delle prestazioni del Framework AWS Well-Architected. Inoltre, fornisce indicazioni per aiutarti i clienti ad applicare le best practice per la progettazione, la distribuzione e la manutenzione degli AWS ambienti.

Introduzione

Il 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 Framework permette di misurare in modo coerente le architetture secondo le best practice e identificare le aree di miglioramento. Disporre di carichi di lavoro ben progettati aumenta notevolmente la probabilità di successo aziendale.

Il Framework si basa su sei pilastri:

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

Il presente whitepaper tratta dell'applicazione dei principi del pilastro dell'efficienza delle prestazioni ai carichi di lavoro. Nei tradizionali ambienti on-premises, 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. Linee guida e best practice contenute nel presente documento sono suddivise in cinque aree di interesse chiave, che costituiscono 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 dei dati
- Reti e distribuzione di contenuti
- Processo e cultura

Questo documento è rivolto ai ruoli nell'ambito della tecnologia, ad esempio a Chief Technology Officer (CTO), progettisti, sviluppatori e 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 ad alte prestazioni.

Efficienza delle prestazioni

Il pilastro dell'efficienza delle prestazioni include la capacità di utilizzare in modo efficiente le risorse nel cloud per soddisfare i requisiti in termini di prestazione e di mantenere tale efficienza a fronte al cambiamento della domanda e all'evoluzione 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, No SQL database, transcodifica multimediale e apprendimento automatico 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.
- Diventa globale in pochi minuti: l'implementazione del carico di lavoro in più AWS regioni del mondo ti consente di fornire una latenza inferiore e un'esperienza migliore per i tuoi clienti a costi minimi.
- Utilizza 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 serverless 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 più di frequente: le risorse virtuali e automatizzabili ti permettono di portare a termine velocemente i test comparativi utilizzando diversi tipi di istanze, storage o configurazioni.

- Prendi in considerazione la comprensione meccanica: sfrutta la strategia tecnologica più adatta 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 dei 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 regolarmente le tue scelte, ti assicurerai di sfruttare i vantaggi del cloud in continua evoluzione. 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 forniscono 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 on-premises. 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 Fattore di costo nelle decisioni architettoniche
- PERF01-BP04 Valutazione dell’influenza dei compromessi sui clienti e sull’efficienza dell’architettura
- PERF01-BP05 Usa politiche e 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: elevato

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.
  - Amazon Web Services Cloud
  - AWS Academy
  - Novità di AWS
  - Blog AWS
  - AWS Skill Builder
  - Eventi e webinar AWS
  - AWS Training e certificazioni
  - Canale YouTube di AWS

  - Workshop AWS
  - Community AWS
- Usa Amazon Q per ricevere informazioni e consigli pertinenti sui servizi.
- 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:

- Overview of 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
- AWS Architecture Center
- AWS Partner Network
- AWS Biblioteca di soluzioni di
- Centro conoscenze 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
- This is my Architecture

Esempi correlati:

- AWS Esempi di

- AWS Esempi di SDK

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 whitepaper)
  - 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 AWS
  - Community AWS
- Contatta AWS per ricevere assistenza quando ti occorrono ulteriori indicazioni o informazioni sui prodotti. AWS I Solutions Architect e i servizi professionali di AWS forniscono indicazioni per l'implementazione delle soluzioni. AWS I partner mettono a disposizione la propria conoscenza di AWS per aiutarti ad assicurare alla tua azienda agilità e innovazione.
- Usa Supporto se hai bisogno di supporto tecnico per utilizzare un servizio in modo efficace. I nostri piani di supporto 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:

- AWS Architecture Center
- AWS Partner Network
- Biblioteca di soluzioni di AWS
- Centro conoscenze di AWS
- Supporto AWS Enterprise

Video correlati:

- This is my Architecture
- 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:

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

PERF01-BP03 Fattore di costo nelle decisioni architettoniche

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 recensisci i nuovi servizi e funzionalità di. 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 usare Calcolatore dei prezzi AWS e AWS Cost Explorer per identificare i principali fattori di costo del carico di lavoro.
- Esamina i modelli di prezzo nel cloud, ad esempio istanze on-demand, riservate, Savings Plans e istanze spot.
- Segui le best practice per l'ottimizzazione dei costi di Well-Architected per ottimizzare questi componenti principali in termini di costi.
- Monitora e analizza continuamente i costi per identificare le opportunità di ottimizzazione dei costi nel tuo carico di lavoro.
  - Usa Budget AWS per ricevere gli avvisi per i costi inaccettabili.
  - Usa AWS Compute Optimizer o AWS Trusted Advisor per ottenere suggerimenti sull'ottimizzazione dei costi.
  - Usa AWS Cost Anomaly Detection per rilevare in modo automatico le anomalie dei costi e analizzare la causa principale.

Risorse

Documenti correlati:

- Che cos'è AWS Billing and Cost Management?
- Ottimizzazione dei costi con AWS
- Scelta di una strategia di gestione dei AWS costi
- Una guida per principianti alla gestione AWS dei costi
- A Detailed Overview of the Cost Intelligence Dashboard

- AWS Architecture Center
- Biblioteca di soluzioni di AWS
- Centro conoscenze di AWS

Video correlati:

- This is my Architecture
- AWS re:Invent 2023 - Cosa c'è di nuovo con l'ottimizzazione dei costi AWS
- AWS re:Invent 2023 - Ottimizza costi e prestazioni e monitora i progressi verso la mitigazione
- AWS re:Invent 2023 - best practice per l'ottimizzazione dei costi di storage AWS
- AWS re:Invent 2023 - Ottimizza i costi nei tuoi ambienti con più account

Esempi correlati:

- AWS Compute Optimizer Codice demo
- 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: elevato

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 i contratti 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.
- Effettua 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
- KPI di Quick

- Amazon CloudWatch RUM
- Documentazione di X-Ray
- Understand resiliency patterns and trade-offs to architect efficiently in the cloud

Video correlati:

- Optimize applications through Amazon CloudWatch RUM
- 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:

- Misurazione dei tempi di caricamento delle pagine con Amazon CloudWatch Synthetics
- Client Web Amazon CloudWatch RUM

PERF01-BP05 Usa politiche e architetture di riferimento

Utilizza le policy interne e le architetture di riferimento esistenti per la selezione dei servizi e delle configurazioni per una maggiore efficienza 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 consente di prendere decisioni rapidamente.

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

Guida all'implementazione

Avere policy 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. Implementi 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 sandbox.
- Resta up-to-date conforme agli standard e agli AWS aggiornamenti del settore per assicurarti che le tue policy e le architetture di riferimento contribuiscano a ottimizzare il carico di lavoro sul cloud.

Risorse

Documenti correlati:

- AWS Architecture Center
- AWS Partner Network
- Biblioteca di soluzioni di AWS
- Centro conoscenze di AWS
- AWS Blog di architettura

Video correlati:

- This is my Architecture
- AWS re:Invent 2022 - Accelera il valore della tua azienda con SAP un'architettura di riferimento AWS

Esempi correlati:

- Esempi AWS
- AWS SDKEsempi

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: misurazione dei miglioramenti in termini di prestazioni grazie al benchmarking dell'implementazione attuale.

Livello di rischio associato se questa best practice non fosse adottata: 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 i tuoi test di benchmarking personalizzati oppure utilizzare test standard del settore, come TPC-DS, per il benchmark 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 implementazione, 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 tuo carico di lavoro.
- Configura ed esegui l'instrumentazione:
  - Imposta il tuo ambiente e configura le risorse.
  - Implementa il monitoraggio e la creazione di log 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:

- AWS Architecture Center
- AWS Partner Network
- Biblioteca di soluzioni di AWS

- Centro conoscenze di 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
- This is my Architecture
- Optimize applications through Amazon CloudWatch RUM
- Demo of Amazon CloudWatch Synthetics

Esempi correlati:

- AWS Esempi di
- Esempi di SDK AWS
- Test del carico distribuito
- Misurazione dei tempi di caricamento delle pagine con Amazon CloudWatch Synthetics
- Client Web Amazon CloudWatch RUM

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 supposizioni.
- 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 le principali parti interessate 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 ulteriori 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 architetturali e risorse, come AWS Solution Architect, il Centro di architettura AWS e AWS Partner Network, per scegliere l'architettura più adatta al tuo carico di lavoro.

- Definisci i parametri, come il throughput 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:

- Biblioteca di soluzioni di AWS
- Centro conoscenze di AWS
- Architectural Patterns to Build End-to-End Data Driven Applications on AWS

Video correlati:

- This is my Architecture
- 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:

- AWS Esempi di
- Esempi di SDK AWS

Calcolo e hardware

La soluzione ottimale in termini di calcolo per un determinato carico di lavoro 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 calcolo 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 al fine di 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 Raccogli metriche relative 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 on-premises.
- Non si conoscono le opzioni, le funzionalità e le soluzioni di cloud computing e come queste migliorino le prestazioni di elaborazione.
- Si effettua il provisioning eccessivo dell'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 allocare o gestire server, Amazon ECS o Amazon EKS per eseguire e gestire container o 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 i vari servizi di elaborazione AWS per il tuo carico di lavoro. Per ulteriori informazioni, consulta 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:

AWSServizio	Caratteristiche chiave	Casi di utilizzo comune
Amazon Elastic Compute Cloud (Amazon EC2)	Dispone di un'opzione dedicata per hardware, requisiti di licenza, ampia selezione di diverse famiglie di istanze, tipi di processori e acceleratori di elaborazione	Migrazioni con rehosting (lift and shift), applicazione monolitica, ambienti ibridi, applicazioni aziendali

AWSServizio	Caratteristiche chiave	Casi di utilizzo comune
Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS)	Implementazione semplice, ambienti coerenti, scalabile	Microservizi, ambienti ibridi
AWS Lambda	Servizio di elaborazione serverless che esegue il codice in risposta agli eventi e gestisce automaticamente le risorse di elaborazione sottostanti.	Microservizi, applicazioni basate su eventi
AWS Batch	Procede ad allocare e scalare in modo efficient e e dinamico le risorse di elaborazione di di Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS e AWS Fargate, con la possibilità di utilizzare istanze spot o on-demand in base ai requisiti del tuo lavoro	HPC, addestramento dei modelli di ML
Amazon Lightsail	Applicazione Linux e Windows preconfigurata per l'esecuzione di piccoli carichi di lavoro	Applicazioni Web semplici, sito Web personalizzato

- 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 elaborazione in base a modelli di utilizzo reali.

Risorse

Documenti correlati:

- Elaborazione in cloud con AWS
- Tipi di istanza di Amazon EC
- Amazon EKS Containers: Amazon EKS Worker Nodes
- Amazon ECS Containers: Amazon ECS Container Instances
- Funzioni: configurazione della funzione Lambda
- Prescriptive Guidance for Containers
- Prescriptive Guidance for 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
- Workshop su Amazon EKS
- Workshop su Amazon EC2
- 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 calcolo in modo da fornire la giusta quantità di risorse e migliorare l'efficienza delle prestazioni.

Anti-pattern comuni:

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

Vantaggi dell'adozione di questa best practice: acquisisci familiarità con le funzionalità e le configurazioni di calcolo di AWS in modo da poter utilizzare una soluzione di calcolo 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 calcolo 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 calcolo è adatta ai carichi di lavoro dal punto di vista della CPU e della memoria.

Passaggi dell'implementazione

- Comprendi i requisiti del carico di lavoro, come CPU, memoria e latenza.
- Consulta la documentazione e le best practice AWS per scoprire le opzioni di configurazione consigliate che possono contribuire a migliorare le prestazioni di calcolo. 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'archivi azione sono progettate per carichi di lavoro che richiedono un accesso frequente e sequenziale in lettura e scrittura (IOPS) all'archiviazione locale.
Modello tariffario	• Le istanza 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
	• Le istanze spot ti consentono di sfruttare la capacità inutilizzata delle istanze con uno sconto per i carichi di lavoro stateless e tolleranti ai guasti.
Auto Scaling	Usa la configurazione Auto Scaling per abbinare le risorse di calcolo ai modelli di traffico.
Dimensionamento	• Usa Compute Optimizer per ricevere un efficace suggerimento di machine learning riguardo alla configurazione più adatta alle tue caratteristiche di elaborazione. • Usa AWS Lambda Power Tuning per selezionare la configurazione migliore per la tua funzione Lambda.
Acceleratori di calcolo basati su hardware	• Le istanza 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 AWS Trainium, AWS Inferentia e Amazon EC2 DL1.

Risorse

Documenti correlati:

- Elaborazione in cloud con AWS
- Tipi di istanza di Amazon EC
- Processor State Control for Your Amazon EC2 Instance

- Amazon EKS Containers: Amazon EKS Worker Nodes
- Amazon ECS Containers: Amazon ECS Container Instances
- 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 Console di gestione AWS
- 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 – Optimizing Amazon EKS for performance and cost on AWS

Esempi correlati:

- Codice dimostrativo di Compute Optimizer
- Workshop relativo alle 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
- Codice dimostrativo AWS Compute Optimizer
- Workshop su Amazon EKS

PERF02-BP03 Raccogli metriche relative 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.
- Revisione dei 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: elevato

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 delle metriche è un passaggio fondamentale per migliorare la sicurezza, l'efficienza dei costi, le prestazioni e la sostenibilità. AWS fornisce un'ampia gamma di metriche relative alle prestazioni utilizzando servizi di monitoraggio come Amazon CloudWatch per fornirti informazioni preziose. Metriche come CPU l'utilizzo, l'utilizzo della memoria, l'I/O del disco e la rete in entrata e in uscita possono fornire informazioni 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

- 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 il throughput).
  - Metriche EC2 predefinite di Amazon
  - Metriche ECS predefinite di Amazon
  - Metriche EKS predefinite di Amazon
  - Parametri predefiniti di Lambda
  - Parametri EC2 della memoria e del disco di Amazon
- Scegli e configura la soluzione di registrazione e monitoraggio giusta per il tuo carico di lavoro.
  - AWS native Observability
  - AWS Distro per OpenTelemetry

  - Amazon Managed Service per Prometheus
- Definisci il filtro e l'aggregazione richiesti per i parametri in base ai requisiti del tuo carico di lavoro.
  - Quantifica i parametri delle applicazioni personalizzate con Amazon CloudWatch Logs e filtri metrici
  - Raccogli metriche personalizzate con il tagging CloudWatch strategico di Amazon
- Configura le policy di conservazione dei dati per i parametri in modo che corrispondano ai tuoi obiettivi operativi e di sicurezza.
  - Conservazione dei dati predefinita per le metriche CloudWatch
  - Conservazione dei dati predefinita per i registri CloudWatch
- Se necessario, crea allarmi e notifiche per i parametri in modo da rispondere in modo proattivo ai problemi relativi alle prestazioni.
  - Crea allarmi per metriche personalizzate utilizzando il rilevamento delle anomalie di Amazon CloudWatch
  - Crea metriche e allarmi per pagine Web specifiche con Amazon CloudWatch RUM
- Usa l'automazione per implementare gli agenti di aggregazione di parametri e log.
  - AWS Systems Manager automazione
  - OpenTelemetryCollezionista

Risorse

Documenti correlati:

- Monitoraggio e osservabilità
- Migliori pratiche: implementazione dell'osservabilità con AWS
- CloudWatch Documentazione Amazon
- Raccogli metriche e log EC2 dalle istanze Amazon e dai server locali con l'agente CloudWatch
- Accesso ad Amazon CloudWatch Logs per AWS Lambda
- Utilizzo dei CloudWatch log con istanze di container
- Publish custom metrics
- AWS Answers: Centralized Logging
- AWS Servizi che pubblicano metriche CloudWatch
- Monitoraggio di Amazon EKS su AWS Fargate

Video correlati:

- AWS re:Invent 2023 — [LAUNCH] Monitoraggio delle applicazioni per carichi di lavoro moderni
- AWS re:Invent 2023 — Implementazione dell'osservabilità delle applicazioni
- AWS re:Invent 2023 — Creazione di una strategia di osservabilità efficace
- AWS re:Invent 2023 — Osservabilità senza interruzioni con Distro per AWS OpenTelemetry
- Gestione delle prestazioni delle applicazioni su AWS

Esempi correlati:

- AWS per Linux Workload Immersion Day- Amazon CloudWatch
- Monitoraggio di ECS cluster e container Amazon
- Monitoraggio con CloudWatch dashboard Amazon
- EKSWorkshop Amazon

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. Il provisioning eccessivo di risorse cloud può tradursi in costi aggiuntivi, mentre il provisioning insufficiente può comportare prestazioni non soddisfacenti e un'esperienza negativa per il cliente. AWS offre strumenti come AWS Compute Optimizer e AWS Trusted Advisor che sfruttano dati cronologici per fornire consigli sul corretto dimensionamento delle 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 appropriata per il mio carico di lavoro?
  - Selezione del tipo di istanza basata su attributi per Amazon EC2 Fleet
  - 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 carichi di lavoro effimeri, valuta i parametri dell'istanza di Amazon CloudWatch, ad esempio `CPUUtilization`, per identificare se l'istanza è sovra o sottoutilizzata.
  - Per i carichi di lavoro stabili, esegui i controlli con gli strumenti di ridimensionamento corretto di AWS, come AWS Compute Optimizer e AWS Trusted Advisor a intervalli regolari per individuare le opportunità di ottimizzazione e ridimensionamento corretto 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 di Amazon EC
- 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 Console di gestione AWS
- 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:

- Codice dimostrativo AWS Compute Optimizer
- Workshop su Amazon EKS
- Right-sizing recommendations

PERF02-BP05 Dimensionamento dinamico delle risorse di elaborazione

Sfrutta l'elasticità del cloud per scalare 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-premises.
- Dopo un evento di dimensionamento, lasci una capacità aumentata anziché ridurre il dimensionamento.

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: elevato

Guida all'implementazione

AWS offre la flessibilità necessaria per scalare 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: procedi a ridurre orizzontalmente 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 implementazioni dei carichi di lavoro siano in grado di gestire eventi che prevedono l'aumentare verticalmente e il ridurre verticalmente.

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 usarlo
Amazon EC2 Auto Scaling	Assicura di disporre del numero corretto di istanze Amazon EC2 disponibili per gestire il carico dell’applicazione.
Application Auto Scaling	Dimensiona in automatico le risorse per singoli servizi AWS oltre Amazon EC2, ad esempio, funzioni AWS Lambda o servizi Amazon Elastic Container Service (Amazon ECS).
Kubernetes Cluster Autoscaler/Karpenter	Dimensiona automaticamente i cluster Kubernetes.

- Si parla spesso di dimensionamento con servizi di calcolo come le istanze Amazon EC2 o le funzioni AWS Lambda. Assicurati di considerare anche la configurazione di servizi non di calcolo come AWS Glue per soddisfare la domanda.
- Verifica che i parametri per il dimensionamento corrispondano alle caratteristiche del carico di lavoro da implementare. Se implementi 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. Se necessario, puoi utilizzare una metrica personalizzata per la tua policy di dimensionamento. 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 del parametro deve aumentare e diminuire in proporzione al numero di istanze nel gruppo con scalabilità automatica.
- Assicurati di utilizzare il dimensionamento dinamico anziché il dimensionamento manuale per il tuo gruppo Auto Scaling. 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 le attività di ridimensionamento 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 di Amazon EC
- 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 per uso generico 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 le serie di porte programmabili sul campo (FPGA), è possibile eseguire determinate funzioni di elaborazione in modo più efficiente.

Livello di rischio associato se questa best practice non fosse adottata: 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 le istanza a calcolo accelerato in grado di soddisfare i tuoi requisiti.
- Per i carichi di lavoro di machine learning, sfrutta l'hardware specifico per il tuo carico di lavoro, come AWS Trainium, AWS Inferentia e Amazon EC2 DL1. AWS Le istanze Inferentia come le

istanze Inf2 offrono fino al 50% in più di prestazioni per 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 acquisire metriche quali `utilization_gpu` e `utilization_memory` per le tue GPU, come illustrato in Collect NVIDIA GPU metrics with Amazon CloudWatch.
- Ottimizza il codice, il funzionamento della rete e le impostazioni degli acceleratori hardware per garantire il pieno utilizzo dell'hardware sottostante.
  - Ottimizza le impostazioni GPU
  - Monitoraggio e ottimizzazione delle GPU nell'AMI per il deep learning
  - Optimizing I/O for GPU performance tuning of deep learning training in Amazon SageMaker
- Utilizza 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:

- Utilizzo di GPU su 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
- Amazon EC2 VT1 Instances
- Come faccio a scegliere il tipo di istanza Amazon EC2 appropriata per il mio carico di lavoro?
- 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 dei 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), al throughput necessario, 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 Implementazione di 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, il throughput 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 di 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 tue esigenze del 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 di libro mastro, relazionali, chiave-valore, di documenti, in memoria, a grafo, di serie temporali. 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 archiviazione e database) disponibili per il carico di lavoro in AWS che possono soddisfare le caratteristiche dei tuoi dati (come illustrato 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	Services	Caratteristiche chiave
Archiviazione di oggetti	Amazon S3	Scalabilità illimitata, alta disponibilità e molteplic i opzioni di accessibilità. L'accesso a oggetti e il relativo trasferimento da e verso Amazon S3 può utilizzare un servizio, come Transfer Accelerat ion o Punti di accesso, per supportare la posizione, le esigenze di sicurezza e i modelli di accesso.
Archiviazione	Amazon Glacier	Progettato per l'archiviazione dei dati.
Archiviazione in streaming	Amazon Kinesis Streaming gestito da Amazon per Apache Kafka (Amazon MSK)	Acquisizione e archiviazione efficienti dei dati in streaming.
File system condiviso	Amazon Elastic File System (Amazon EFS)	File system montabile a cui è possibile accedere da più tipi di soluzioni di calcolo.

Tipo	Services	Caratteristiche chiave
File system condiviso	Amazon FSx	Sviluppato con le più recenti soluzioni di calcolo AWS per supportare i 4 file system più comunemente utilizzati: NetApp ONTAP, OpenZFS, Windows File Server e Lustre. Latenza, throughput e IOPS di Amazon FSx variano a seconda del file system; è necessario considerare attentamente questi elementi quando si deve selezionare il file system in modo conforme ai requisiti dei carichi di lavoro.
Storage a blocchi	Amazon Elastic Block Store (Amazon EBS)	Servizio di storage a blocchi scalabile e a elevate prestazioni progettato per Amazon Elastic Compute Cloud (Amazon EC2). Amazon EBS include storage su SSD per carichi di lavoro transazionali e intensivi dal punto di vista dell'IOPS, oltre a storage su HDD per carichi di lavoro con throughput intensivo.

Tipo	Services	Caratteristiche chiave
Database relazionale	Amazon Aurora, Amazon RDS, Amazon Redshift.	Progettati per supportare le transazioni ACID (atomicità, coerenza, isolamento, durabilità) e per mantenere l'integrità referenziale e una solida coerenza dei dati. Molte applicazioni tradizionali, Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) ed e-commerce utilizzano database relazionali per archiviare i propri dati.
Database chiave-valore	Amazon DynamoDB	Ottimizzato per schemi di accesso di uso comune, in genere per archiviare e recuperare grandi volumi di dati. Le app Web dal traffico elevato, i sistemi di e-commerce e le applicazioni di videogiochi sono casi d'uso tipici dei database chiave-valore.
Database di documenti	Amazon DocumentDB	Progettato per archiviare dati semistrutturati come documenti simili a JSON. Questi database aiutano gli sviluppatori a creare e aggiornare rapidamente applicazioni quali gestione di contenuti, cataloghi e profili utente.

Tipo	Services	Caratteristiche chiave
Database in memoria	Amazon ElastiCache, Amazon MemoryDB per Redis	Vengono utilizzati per applicazioni che richiedon o accesso in tempo reale ai dati, bassissima latenza ed elevatissimo throughput. È possibile utilizzare database in memoria per la memorizza zione nella cache delle applicazioni, la gestione delle sessioni, la classifica dei giochi, l'archivio delle caratteri stiche ML a bassa latenza, il sistema di messaggistica dei microservizi e un meccanism o di streaming a elevato throughput.
Database a grafo	Amazon Neptune	Utilizzato con le applicazi oni che devono navigare ed eseguire query su milioni di relazioni tra set di dati a grafo altamente connessi, con una latenza misurata in milliseco ndi su larga scala. Molte aziende utilizzano database a grafo per il rilevamen to di attività fraudolente, i social network e i motori di raccomandazione.

Tipo	Services	Caratteristiche chiave
Database di serie temporali	Amazon Timestream	Utilizzato per raccogliere, sintetizzare e derivare in modo efficiente approfond imenti dai dati che cambiano nel tempo. I database di serie temporali sono spesso utilizzati dalle applicazioni IoT, DevOps e dalla telemetri a industriale.
Colonna ampia	Amazon Keyspaces (per Apache Cassandra)	Utilizza tabelle, righe e colonne, ma a differenza di un database relaziona le, i nomi e il formato delle colonne possono variare da riga a riga all'interno della stessa tabella. In genere, gli store colonnari sono utilizzati nelle applicazioni industriali su larga scala per la manutenzione delle apparecchiature, la gestione delle flotte e l'ottimizzazione dei percorsi.

Tipo	Services	Caratteristiche chiave
Di libri mastri	Database Amazon Quantum Ledger (Amazon QLDB)	Fornisce un'autorità centralizzata e affidabile per mantenere un registro delle transazioni scalabile , immutabile e verificabile tramite crittografia per ogni applicazione. I database di libri mastri vengono utilizzati per sistemi di record, catena di fornitura, registrazioni e persino transazioni bancarie.

- Per una piattaforma dati, sfrutta l'architettura dei dati moderna di AWS per integrare data lake, data warehouse e archivi dati appositamente progettati.
- Le domande chiave da porsi quando si sceglie un data store per il carico di lavoro sono le seguenti:

Domanda	Aspetti da considerare
Come sono strutturati i dati?	• Se i dati non sono strutturati, prendi in considerazione un archivio di oggetti, come Amazon S3, o un database NoSQL, come Amazon DocumentDB • Per i dati di tipo chiave-valore, valuta DynamoDB, Amazon ElastiCache (Redis OSS) o Amazon MemoryDB
Quale livello di integrità referenziale è richiesto ?	• Per i vincoli di chiave esterna, i database relazionali come Amazon RDS e Aurora possono fornire livello di integrità richiesto. • 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.

Domanda	Aspetti da considerare
È richiesta la conformità ACID (atomicità, coerenza, isolamento, durabilità)?	• Se sono necessarie proprietà ACID associate ai database relazionali, valuta un database relazionale come Amazon RDS e Aurora. • Se è necessaria un'elevata coerenza per i database NoSQL, puoi utilizzare le elevate consistenza di lettura con DynamoDB.
Come cambierà nel tempo l'archiviazione? In che modo questo avrà effetto sulla scalabilità?	• I database serverless, come DynamoDB e i Database Amazon Quantum Ledger (Amazon QLDB) offrono la scalabilità dinamica. • Per i database relazionali sono previsti limiti massimi per l'archiviazione allocata, al raggiungimento dei quali si rende spesso necessario partizionare orizzontalmente tali database tramite meccanismi quali la partizione.
Qual è la proporzione di query in lettura rispetto alle quelle in scrittura? Il caching potrebbe probabilmente migliorare le prestazioni?	• Per i carichi di lavoro gravosi in termini di lettura, può essere utile un livello di memorizzazione nella cache, come 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.

Domanda	Aspetti da considerare
Hanno priorità più elevata le operazioni di archiviazione e modifica OLTP, Online Transaction Processing) o quelle di recupero e report (OLAP - Online Analytical Processing)?	• Per un'elaborazione transazionale letta così com'è a elevato throughput, prendi in considerazione un database NoSQL come DynamoDB. • Per schemi di lettura complessi con throughput elevato (come il join) con un uso coerente di Amazon RDS. • Per le query analitiche, prendi in considera zione un database a colonne, come Amazon Redshift, o l'esportazione dei dati su Amazon S3, nonché l'esecuzione di analisi mediante Athena o Amazon QuickSight.
Che livello di durabilità è necessario per i dati?	• 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 durabilit à 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.

Domanda	Aspetti da considerare
È presente il desiderio di abbandonare i motori di database commerciali o i costi di licenza?	• Valuta motori open-source come PostgreSQL e MySQL su Amazon RDS o Aurora. • Sfrutta AWS Database Migration Service e AWS Schema Conversion Tool per eseguire le migrazioni dai motori di database commerciali a quelli open-source
Quali sono le aspettative operative per il database? Il passaggio ai servizi gestiti è una priorità?	• Utilizzare Amazon RDS, invece di Amazon EC2, e scegliere DynamoDB o Amazon DocumentDB, invece di ospitare in autonomia un database NoSQL, riduce le spese operative.
Come avviene attualmente l'accesso al database? È solo un accesso da applicazione o sono presenti utenti Business Intelligence (BI) e altre applicazioni pronte all'uso connesse?	• In presenza di dipendenze da strumenti esterni, potresti dover mantenere la compatibilità con i database che supportano. Amazon RDS è del tutto compatibile con le varie versioni dei motori che supporta, tra cui 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
- Amazon Glacier: documentazione di Amazon Glacier

- Amazon S3: considerazioni su velocità e prestazioni delle richieste
- Archiviazione nel cloud in AWS
- Amazon EBS I/O Characteristics
- Database su cloud con AWS
- AWS Database Caching
- DynamoDB Accelerator
- Best practice di Amazon Aurora
- Prestazioni di 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 per l'implementazione di 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
- Amazon S3 Examples
- 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
- Esempi di Amazon Neptune

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: esplorare le configurazioni del datastore e sperimentare con esse può consentire di ridurre il costo dell'infrastruttura, migliorare le prestazioni e ridurre l'impegno richiesto per mantenere i carichi di lavoro.

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

Guida all’implementazione

Un carico di lavoro può utilizzare uno o più datastore in base ai requisiti di archiviazione di dati e relativo accesso. 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 documentazione e 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:

Opzione di configurazione	Esempi
Riduzione del carico delle letture (come le repliche di lettura e la memorizzazione nella cache)	• Per le tabelle DynamoDB, è possibile eliminare il carico delle letture grazie a DAX per la memorizzazione nella cache. • Puoi creare un cluster Amazon ElastiCache (Redis OSS) 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 eliminare il carico creato dalle parti di lettura nel carico di lavoro. • I database serverless come DynamoDB si dimensionano automaticamente. Assicurat i di avere abbastanza unità di capacità di lettura (RCU) allocate per gestire il carico di lavoro.

Opzione di configurazione	Esempi
Dimensionamento delle scritture (come la partizione delle chiavi di partizione o l'introduzione di una coda)	• 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 un maggior throughput 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 il throughput in database relazionali con un elevato volume in scrittura. • I database serverless come DynamoDB possono dimensionare automaticamente il throughput in scrittura oppure è possibile regolare le unità di capacità in scrittura (WCU) allocate, a seconda della modalità di capacità. • È tuttavia possibile che si verifichino problemi con le partizioni hot quando si raggiungono i limiti di throughput per una determinata chiave di partizione. Questo problema può essere arginato scegliendo una chiave di partizione con una distribuzione più uniforme o eseguendo lo sharding in lettura della chiave di partizione.

Opzione di configurazione	Esempi
Policy per gestire il ciclo di vita dei set di dati	• Puoi utilizzare il ciclo di vita Amazon S3 per gestire gli oggetti durante il loro ciclo di vita. In caso di schemi di accesso sconosciuti, mutevoli o imprevedibili, puoi utilizzare il Piano intelligente Amazon S3, che monitora gli schemi di accesso e sposta in automatico gli oggetti che non hanno fatto registrare accessi a livelli di accessi più economici. Sfrutta i parametri di Amazon S3 Storage Lens per individuare opportunità di ottimizzazione e lacune nella gestione del ciclo di vita. • La gestione del ciclo di vita di Amazon EFS gestisce automaticamente l'archiviazione di file a costi contenuti per i file system.
Gestione e pooling delle connessioni	• 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 dimensionamento 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 gli esperimenti, pianifica la migrazione e convalida i parametri delle prestazioni.
- Usa strumenti di monitoraggio AWS (come Amazon CloudWatch) e 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 EC
- Amazon EFS: Amazon EFS Performance
- Amazon FSx for Lustre Performance
- Amazon FSx for Windows File Server Performance
- Amazon Glacier: documentazione di Amazon Glacier
- Amazon S3: considerazioni su velocità e prestazioni delle richieste
- Amazon EBS I/O Characteristics
- Database su cloud con AWS
- AWS Database Caching
- DynamoDB Accelerator
- Best practice di Amazon Aurora
- Prestazioni di 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
- Dimensionamento automatico di Amazon EBS
- Amazon S3 Examples
- Esempi di Amazon DynamoDB
- Esempi di migrazione di database con AWS
- 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.
- Utilizzo solo dei parametri predefiniti registrati dal software di monitoraggio selezionato.
- Revisione dei 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: elevato

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

- Determina i principali parametri delle prestazioni da monitorare per il tuo datastore.
  - Parametri e dimensioni di Amazon S3
  - Monitoraggio di parametri in un'istanza Amazon RDS
  - Monitoring DB load with Performance Insights on Amazon RDS
  - Panoramica sul monitoraggio avanzato
  - DynamoDB Metrics and dimensions
  - Monitoraggio di DynamoDB Accelerator
  - Monitoring Amazon MemoryDB with Amazon CloudWatch
  - Quali parametri è opportuno monitorare?
  - Monitoring Amazon Redshift cluster performance
  - Timestream metrics and dimensions
  - Amazon CloudWatch metrics for Amazon Aurora
  - Creazione di log e monitoraggio in Amazon Keyspaces (per Apache Cassandra)
  - Monitoring Amazon Neptune Resources

- 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 quando le soglie vengono superate.
- Verifica se il monitoraggio dei datastore può trarre vantaggio da una soluzione di machine learning che rileva le anomalie delle prestazioni.
  - Amazon DevOps Guru per Amazon RDS offre visibilità sui problemi di prestazioni e fornisce suggerimenti per le azioni correttive.
- Configura la conservazione dei dati nella soluzione di monitoraggio e registrazione per soddisfare i tuoi obiettivi operativi e di sicurezza.
  - Conservazione dei dati predefinita per i parametri CloudWatch
  - Conservazione dei dati predefinita per i parametri CloudWatch Logs

Risorse

Documenti correlati:

- AWS Database Caching
- Amazon Athena top 10 performance tips
- Amazon Aurora best practices
- DynamoDB Accelerator
- Amazon DynamoDB best practices
- Amazon Redshift Spectrum best practices
- Prestazioni di Amazon RedShift
- Database su cloud AWS
- Approfondimenti sulle prestazioni di Amazon RDS

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:

- Framework di raccolta dei parametri di ingestione del set di dati AWS
- Workshop relativo al monitoraggio di 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 di dati e 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 di dati e il relativo accesso. 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. Ecco alcune strategie chiave:
  - Utilizzo di 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
    - Partitions and data distribution
  - Indicizzazione dei dati sulle colonne comuni della query.
  - Uso delle viste materializzate per le domande frequenti.
    - Understanding materialized views
    - Creating materialized views in Amazon Redshift
  - Scelta dell'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.
  - Miglioramento della latenza e riduzione del numero di operazioni di I/O del database grazie alla soluzione di cache distribuita.
  - Manutenzione regolare, ad esempio vacuum, reindicizzazione ed esecuzione di statistiche.
- La sperimentazione e i test delle strategie in un ambiente non di produzione.

Risorse

Documenti correlati:

- Best practice di Amazon Aurora
- Prestazioni di Amazon RedShift
- Amazon Athena top 10 performance tips
- AWS Database Caching
- 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:

- AWS Purpose Built Databases Workshop

PERF03-BP05 Implementazione di 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, il throughput, 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. È possibile memorizzare i dati nella cache a più livelli dell'applicazione, ad esempio all'interno dell'applicazione che effettua chiamate remote, operazione nota come memorizzazione nella cache lato client, o mediante 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

Come soluzione al problema della duplicazione dei dati tra client, è possibile utilizzare un servizio esterno veloce o la memorizzazione nella cache remota per 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

- 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.
  - Database Caching
  - Abilitazione della memorizzazione nella cache dell'API per migliorare la velocità di risposta
- Identifica il tipo di strategia di memorizzazione nella cache più adatto al tuo modello di accesso.
  - Caching strategies
  - AWS Caching Solutions
- Attieniti alle best practice sulla memorizzazione nella cache per il tuo archivio dati.
- 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.
- 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 practices: Redis clients and Amazon ElastiCache (Redis OSS)
- 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.
  - Which metrics should I monitor?
  - Best practices for monitoring Redis workloads on Amazon ElastiCache
  - Monitoring best practices with Amazon ElastiCache (Redis OSS) using Amazon CloudWatch
- Implementa la replica dei dati per eliminare il carico delle letture per più istanze e migliorare prestazioni e disponibilità della lettura dei dati.

Risorse

Documenti correlati:

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

Esempi correlati:

- Boosting MySQL database performance with Amazon ElastiCache (Redis OSS)

Reti e distribuzione di contenuti

La soluzione di rete ottimale per un carico di lavoro varia in base a latenza, requisiti di throughput, jitter e larghezza di banda. I vincoli fisici, ad esempio le risorse utente o on-premises, determinano le opzioni di posizione. Questi vincoli possono essere compensati con le posizioni edge 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 funzionalità 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 VPC Amazon, 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 Scegli la 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-premises.

Vantaggi dell'adozione di questa best practice: la comprensione dell'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: elevato

Guida all’implementazione

La rete è responsabile della connettività tra componenti dell'applicazione, servizi cloud, reti edge e dati on-premises 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, throughput e regole di instradamento.

Predisponi 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. Esamina le soluzioni di rete disponibili e individua 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:

- 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), throughput, crittografia, ispezione e regole di instradamento.
- Scopri i principali servizi di rete AWS come VPC, AWS Direct Connect, Elastic Load Balancing (ELB) e Amazon Route 53.
- 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 • AWS Transit Gateway Parametri di • AWS PrivateLink Parametri di
Caratteristiche di rete dell'applicazione	• Elastic Fabric Adapter • AWS App Mesh Parametri di • Parametri per Gateway Amazon API
Caratteristiche della rete edge	• Parametri di Amazon CloudFront • Parametri di Amazon Route 53 • AWS Global Accelerator Parametri di
Caratteristiche della rete ibrida	• Direct Connect Parametri di • AWS Site-to-Site VPN Parametri di • AWS Client VPN Parametri di • Parametri WAN Cloud AWS
Caratteristiche della sicurezza di rete	• Parametri AWS Shield, AWS WAF e AWS Network Firewall
Caratteristiche del tracciamento	• AWS X-Ray • VPC Reachability Analyzer • Strumento di analisi degli accessi alla rete • Amazon Inspector • Amazon CloudWatch RUM

- Esegui il benchmark e testa le prestazioni della rete:
  - Esegui il benchmark del throughput 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.
  - Effettua test di carico per sperimentare soluzioni e opzioni di rete.

Risorse

Documenti correlati:

- Application Load Balancer
- Reti avanzate EC2 su Linux
- Reti avanzate EC2 su Windows
- Gruppi di collocamento EC
- Abilitazione delle reti avanzate con l'Adattatore elastico di rete (ENA) sulle istanze Linux
- Network Load Balancer
- Prodotti di rete con AWS
- Gateway di transito
- Transitioning to latency-based routing 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
- Workshop sulle reti AWS
- 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 aumentare le prestazioni del carico di lavoro, ridurre il costo dell'infrastruttura, diminuire il livello di impegno richiesto per mantenere il tuo carico di lavoro 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: elevato

Guida all'implementazione

AWS offre servizi come AWS Global Accelerator e Amazon CloudFront per migliorare le prestazioni di rete, mentre la maggior parte dei servizi AWS offre funzionalità di prodotto (come la funzionalità Amazon S3 Transfer Acceleration) per l'ottimizzazione del 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'uso della WAN di Cloud AWS 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 le zone di disponibilità, nonché all'interno di ciascuna zona di disponibilità, sfruttando AWS Network Manager per ottenere informazioni dettagliate sulla relazione fra le prestazioni delle applicazioni e quelle della rete AWS sottostante.
- Utilizza uno strumento esistente per il database di gestione della configurazione (CMDB) o 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 test di carico per individuare i colli di bottiglia delle prestazioni.
- Per tutti i colli di bottiglia di questo tipo individuati, 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	Usa lo Strumento di analisi degli accessi alla rete per identificare percorsi o instradamenti.
Protocolli di rete	Per informazioni, consultare PERF04-BP05 Scelta dei protocolli di rete per migliorare le prestazioni
Topologia di rete	Valuta i compromessi a livello di operazioni e prestazioni tra VPC Peering e AWS Transit Gateway quando si collegano più account.

Opportunità di miglioramento	Soluzione
	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-premises. Condividi AWS Transit Gateway tra più account utilizzando AWS Resource Access Manager. Per informazioni, consultare PERF04-BP 03 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. Usa 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 geoprossimità e instradamento basato su IP per migliorare le prestazioni del tuo 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	Amazon S3 Transfer Acceleration è una funzionalità 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 a Cloud AWS. I punti di accesso multi-regione di Amazon S3 rappresentano una funzionalità che replica i contenuti in più regioni e semplificano il carico di lavoro fornendo un punto di accesso. Quando viene utilizzato un punto di accesso multi-regione, puoi richiedere o scrivere dati in Amazon S3 con il servizio che identifica il bucket con latenza più bassa.

Opportunità di miglioramento	Soluzione
Funzionalità delle risorse di calcolo	Le interfacce di rete elastica (ENI) utilizzate da istanze Amazon EC2, container e funzioni Lambda sono limitate in base ai flussi. Rivedi i gruppi di collocazione per ottimizzare il throughput di rete di EC2. Per evitare colli di bottiglia a livello di flusso, progetta l'applica zione in modo che utilizzi più flussi. Per monitorare le metriche di rete associate al calcolo e avere maggiore visibilità su di esse, utilizza i parametri CloudWatch ed ethtool. Il comando `ethtool` è incluso nel driver ENA e permette di utilizzare parametri relativi alla rete aggiuntivi che possono essere pubblicati come parametri personalizzati in CloudWatch. Gli adattatori elastici di rete (ENA) Amazon offrono un'ulteriore ottimizzazione, migliorando il throughput per le tue istanze all'interno di un gruppo di posizionamento 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 interni su vasta scala su 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 EC2 su Linux
- Reti avanzate EC2 su Windows
- Gruppi di collocamento EC2
- Abilitazione delle reti avanzate con l'Adattatore elastico di rete (ENA) sulle istanze Linux
- Network Load Balancer
- 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
- Workshop sulle reti AWS
- 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-premises 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: elevato

Guida all’implementazione

Sviluppa un’architettura di rete ibrida basata sui tuoi requisiti di larghezza di banda. Direct Connect consente di connettere la rete on-premises in 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 Direct Connect è in uso.

Se i tuoi requisiti di larghezza di banda sono elevati, potresti prendere in considerazione l’utilizzo di più 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 Direct Connect e VPN è sconsigliato a causa delle differenze di latenza e larghezza di banda.

Passaggi dell’implementazione

- Calcola i requisiti di larghezza di banda e latenza delle tue app esistenti.
  - Per i carichi di lavoro esistenti che vengono spostati in AWS, utilizza i dati raccolti dai sistemi di monitoraggio di rete interni.
  - 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.
- 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 Site-to-Site VPN (o entrambi). Il diagramma seguente può aiutarti a scegliere il tipo di connessione appropriato.
  - 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.
  - AWS VPN Site-to-Site 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.
- Consulta la documentazione AWS per scegliere l’opzione di connettività appropriata:
  - Se decidi di utilizzare Direct Connect, seleziona la larghezza di banda appropriata per la tua connettività.
  - In caso di utilizzo di una AWS Site-to-Site VPN tra più posizioni per connetterti a una Regione AWS, utilizza una connessione Site-to-Site VPN accelerata per migliorare le prestazioni di rete.
  - Se la progettazione della rete è costituita 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. AWS La VPN Site-to-Site Site con indirizzo IP privato viene implementata su un’interfaccia virtuale di transito (VIF).
  - 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 AWS Direct Connect, bypassando Regioni AWS.

- 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.
- Monitora regolarmente le prestazioni e l'utilizzo della connettività e ottimizzali, se necessario.

Diagramma di flusso per le prestazioni deterministiche

Risorse

Documenti correlati:

- Prodotti di rete con AWS
- AWS Transit Gateway
- Endpoint VPC
- Realizzazione di un'infrastruttura di reti multi-VPC sicura e scalabile
- 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 Workshop sulle reti

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 bilanciatore del carico senza tenere conto dei requisiti del carico di lavoro.
- Mancato utilizzo delle funzionalità del bilanciatore del carico per l'ottimizzazione delle prestazioni.
- Esposizione diretta del carico di lavoro a Internet senza un bilanciatore del carico.
- Instradi tutto il traffico Internet attraverso i bilanciatori 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 consente 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 bilanciatori 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 bilanciatore 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 diversi modelli di bilanciamento del carico per le tue applicazioni. Application Load Balancer è l'ideale per il bilanciamento del carico del traffico HTTP e HTTPS. Inoltre, 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 bilanciatore 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 Application Load Balancer (ALB) e 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 bilanciatore 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.

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 Network Load Balancer, che offre latenze estremamente ridotte. In alternativa, è possibile decidere di avvicinare il carico di lavoro ai clienti sfruttando Application Load Balancer nelle zone locali AWS o addirittura 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 bilanciatore del carico distribuisce il traffico tra le destinazioni registrate in tutte le zone di disponibilità autorizzate.

Usa Auto Scaling integrato con il bilanciatore 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 bilanciatori del carico per le risorse di destinazione backend. Usando l'integrazione dei bilanciatori del carico con gruppi Auto Scaling, le destinazioni vengono aggiunte o rimosse nel e dal bilanciatore 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 carichi di lavoro distribuiti in container.

- Amazon ECS: bilanciamento del carico di servizio
- Bilanciamento del carico di applicazione su Amazon EKS
- Bilanciamento del carico di rete su 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 bilanciatore del carico giusto per la tua applicazione.
  - Utilizza Application Load Balancer per i carichi di lavoro HTTP/HTTPS.
  - Utilizza Network Load Balancer per carichi di lavoro non HTTP in esecuzione su TCP o UDP.
  - Usa una combinazione dei due sistemi (ALB come destinazione di NLB) per sfruttare le funzionalità di entrambi i prodotti. Ad esempio, puoi scegliere questa opzione se vuoi usare gli indirizzi IP statici dell'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 Application Load Balancer e 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 in materia di 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, 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 backend 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 domini con errori di zona. La funzione è disattivata per impostazione predefinita in NLB e in ALB è possibile disattivarla per gruppo di destinazione.
  - Attiva l'isolamento tra zone per ottenere disponibilità e flessibilità maggiori. L'isolamento tra zone è disattivato per impostazione predefinita in ALB e in NLB è possibile attivarlo per gruppo di destinazione.
- Attiva keep-alive HTTP per i carichi di lavoro HTTP (solo ALB). Con questa funzionalità, il bilanciatore 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 bilanciatore del carico.
  - Attiva i log di accesso per Application Load Balancer e 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 l'NLB sono `connection_time` e `tls_handshake_time`.
  - Preparati a eseguire query sui log quando necessario. Puoi usare Amazon Athena per eseguire query sui log ALB e sui log NLB.
  - Crea allarmi per metriche correlate alle prestazioni, come `TargetResponseTime` per ALB.

Risorse

Documenti correlati:

- Tabella di confronto dei 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 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 il throughput desiderato. Se per il trasferimento file si usa il protocollo TCP, latenze più elevate molto probabilmente ridurranno il throughput complessivo. Alcuni approcci risolvono questo problema tramite l'ottimizzazione del TCP e l'utilizzo di protocolli di trasferimento ottimizzati, ma una soluzione prevede l'utilizzo del protocollo User Datagram Protocol (UDP).

Anti-pattern comuni:

- Puoi utilizzare il TCP per tutti i carichi di lavoro, indipendentemente dai requisiti prestazionali.

Vantaggi dell’adozione di questa best practice: la verifica del protocollo adeguato 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 elevata 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 l’ottimizzazione dell’applicazione in altri modi.

Un aspetto fondamentale per il miglioramento delle prestazioni del tuo carico di lavoro consiste nell’identificare i requisiti in termini di latenza e throughput, quindi scegliere i protocolli di rete che ottimizzano le prestazioni.

Quando valutare se usare TCP

Il protocollo 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 throughput inferiore, ma presenta come vantaggio una maggiore sicurezza. Il sovraccarico è dovuto perlopiù 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 in presenza di 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). In caso di utilizzo di 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, poiché comprende 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 elevato throughput grazie alla sua capacità di bilanciatore del 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 throughput elevato 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

- Utilizzare i servizi AWS Global Accelerator e AWS Transfer Family per migliorare il throughput delle applicazioni di trasferimento file online. Il servizio AWS Global Accelerator ti permette di ottenere latenze inferiori 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) e FTPS (File Transfer Protocol over SSL) per scalare e gestire i trasferimenti file in servizi di archiviazione AWS in tutta sicurezza.
- 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 tempo di acquisizione al primo byte (TTFB) e tempo di andata e ritorno (RTT) Se il tuo carico di lavoro offre contenuti dinamici agli utenti, prendi in considerazione l'utilizzo di Amazon CloudFront, che stabilisce una connessione persistente a ciascuna origine per il contenuto dinamico in modo da eliminare il tempo di configurazione della connessione, che altrimenti rallenterebbe ogni richiesta client.
- L'uso di TLS con TCP o UDP può causare maggiore latenza e minore throughput per il carico di lavoro a causa dell'impatto della crittografia e della decrittografia. Per tali carichi di lavoro, prendi in considerazione l'offload SSL/TLS su Elastic Load Balancing per migliorare le prestazioni del carico di lavoro permettendo al bilanciatore del carico di gestire la crittografia e la decrittografia SSL/TLS invece di predisporre a questo scopo istanze backend. In questo modo, puoi ridurre l'utilizzo della CPU sulle istanze backend, migliorando le prestazioni e aumentando la capacità.

- Usa Network Load Balancer (NLB) per implementare servizi basati sul protocollo UDP, tra cui autenticazione e autorizzazione, log, DNS, IoT e streaming di contenuti multimediali, in modo da migliorare prestazioni e affidabilità del carico di lavoro. L'NLB distribuisce il traffico UDP in ingresso tra più destinazioni, permettendo di scalare orizzontalmente il carico di lavoro, incrementare la capacità e diminuire il sovraccarico su un'unica destinazione.
- Per i carichi di lavoro di calcolo ad alte prestazioni (HPC), prendi in considerazione l'utilizzo della funzionalità Adattatore elastico di rete (ENA) Express che sfrutta il protocollo SRD per migliorare le prestazioni di rete fornendo una maggiore larghezza di banda a flusso singolo (25 Gbps) e una latenza di coda inferiore (99,9 percentile) per il traffico di rete tra istanze EC2.
- Usa Application Load Balancer (ALB) per instradare e bilanciare 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 EC2 su Linux
- Reti avanzate EC2 su Windows
- Gruppi di collocamento EC2
- Abilitazione delle reti avanzate con l'Adattatore elastico di rete (ENA) sulle istanze Linux
- Network Load Balancer
- 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
- Workshop sulle reti AWS

PERF04-BP06 Scegli la 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 il throughput, 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 una rete globale appropriata Regioni AWS e AWS privata, è possibile ridurre la latenza e offrire un'esperienza migliore agli utenti remoti.

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

Guida all'implementazione

Le risorse, come le EC2 istanze Amazon, vengono collocate nelle Availability Zones within Regioni AWS, AWS Local Zones o AWS Wavelengthnelle zone. AWS Outposts La scelta della posizione influisce su latenza di rete e throughput dall'ubicazione di un utente specifico. I servizi edge come Amazon CloudFront AWS Global Acceleratorpossono essere utilizzati anche per migliorare le prestazioni di rete memorizzando nella cache i contenuti nelle edge location o fornendo agli utenti un percorso ottimale per il carico di lavoro attraverso la rete AWS globale.

Amazon EC2 fornisce gruppi di collocamento per il networking. Un gruppo di collocazione è un raggruppamento logico di istanze per ridurre la latenza. L'utilizzo di gruppi di collocamento con tipi di

istanze supportati e un Elastic Network Adapter (ENA) consente ai carichi di lavoro di partecipare a una rete a 25 Gbps a bassa latenza e con 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 sedi periferiche utilizzando una rete AWS globale, come Amazon. CloudFront Queste edge location forniscono in genere servizi come Content Delivery Network (CDN) e Domain Name System (). DNS Disponendo di questi servizi all'edge, i carichi di lavoro possono rispondere con bassa latenza alle richieste di contenuto o DNS risoluzione. 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 della cache per entrambi DNS eHTTP/HTTPSper ottenere il massimo vantaggio da questi approcci.

Passaggi dell'implementazione

- Acquisisci informazioni sul traffico IP in entrata e in uscita dalle interfacce di rete.
  - Registrazione del traffico IP utilizzando VPC Flow Logs
  - Come viene preservato 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.
  - Utilizza strumenti di monitoraggio, come Amazon CloudWatch e AWS CloudTrail, per raccogliere dati sulle attività di 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:
  - Ubicazione dei dati per le applicazioni a uso intensivo di dati, ad esempio applicazioni di big data e machine learning, il codice dell'applicazione dovrebbe essere eseguito il più vicino possibile ai dati.
  - Ubicazione degli utenti: per le applicazioni rivolte agli utenti, scegli una regione o più regioni vicine agli utenti del carico di lavoro.
  - Altri vincoli: prendi in considerazione vincoli come costi e conformità, come illustrato in What to Consider when Selecting a Region for your Workloads.

- Usa le zone locali AWS per eseguire carichi di lavoro come il rendering 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 on-premises, ma vuoi che vengano eseguiti in modo ottimale con il resto degli altri carichi di lavoro in AWS.
- Applicazioni come lo streaming video in diretta ad alta risoluzione, l'audio ad alta fedeltà e la realtà aumentata o la realtà virtuale (AR/VR) richiedono dispositivi 5G. ultra-low-latency Per tali applicazioni, considera. AWS Wavelength AWS Wavelength incorpora servizi di AWS elaborazione e archiviazione nelle reti 5G, fornendo un'infrastruttura di edge computing mobile per lo sviluppo, l'implementazione e la scalabilità delle applicazioni. ultra-low-latency
- Usa la cache locale o le soluzioni di caching AWS per i dati di frequente utilizzo per migliorare le performance, ridurre lo spostamento dei dati e minimizzare l'impatto ambientale.

Servizio	Quando usare
Amazon CloudFront	Utilizzalo per memorizzare nella cache contenuti statici come immagini, script e video, nonché contenuti dinamici come API risposte o applicazioni web.
Amazon ElastiCache	Usalo per memorizzare nella cache i contenuti per le applicazioni Web.
DynamoDB Accelerator	Usalo 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:

Servizio	Quando usare
Lambda@Edge	Usalo per operazioni a uso intensivo di risorse di calcolo eseguite quando gli oggetti non si trovano nella cache.
CloudFront Funzioni Amazon	Utilizzalo per casi d'uso semplici come richieste HTTP (s) o manipolazioni di risposte

Servizio	Quando usare
	che possono essere avviate da funzioni di breve durata.
AWS IoT Greengrass	Usale 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 del throughput. Queste applicazioni possono trarre vantaggio da servizi di rete che forniscono indirizzi IP anycast statici e TCP terminazioni in postazioni periferiche. AWS Global Acceleratorpossono migliorare le prestazioni delle applicazioni fino al 60% e fornire un failover rapido per architetture multiregionali. AWS Global Accelerator fornisce indirizzi IP anycast statici che fungono da punto di ingresso fisso per le applicazioni ospitate in una o più applicazioni. Regioni AWS Questi indirizzi IP consentono al traffico di entrare nella rete AWS globale il più vicino possibile agli utenti. AWS Global Accelerator riduce il tempo di configurazione iniziale della connessione stabilendo una TCP connessione tra il client e la AWS edge location più vicina al client. Rivedi l'utilizzo di AWS Global Accelerator per migliorare le prestazioni dei tuoi UDP carichi di lavoroTCP/e fornire un failover rapido per architetture multiregionali.

Risorse

Best practice correlate:

- COST07-BP02 Implementazione delle regioni in base ai costi
- COST08-BP03 Implementazione di servizi per ridurre i costi di trasferimento dei dati
- REL10-BP01 Implementa il carico di lavoro in più sedi
- REL10-BP02 Seleziona le posizioni appropriate per l'implementazione in più sedi
- SUS01-BP01 Scegli la regione in base ai requisiti aziendali e agli obiettivi di sostenibilità
- SUS02-BP04 Ottimizza il posizionamento geografico dei carichi di lavoro in base ai requisiti di rete
- SUS04-BP07 Riduci al minimo lo spostamento dei dati tra le reti

Documenti correlati:

- AWS Infrastruttura globale
- AWS Local Zones e AWS Outposts scelta della tecnologia giusta per il tuo carico di lavoro edge
- Placement groups
- AWS Local Zones
- AWS Outposts
- AWS Wavelength
- Amazon CloudFront
- AWS Global Accelerator
- AWS Direct Connect
- AWS Site-to-Site VPN
- Amazon Route 53

Video correlati:

- AWS Video esplicativo su Local Zones
- AWS Outposts: Overview and How it Works
- AWS re:Invent 2023 - Una strategia di migrazione per carichi di lavoro edge e locali
- AWS re:Invent 2021 -: Portare l'esperienza in sede AWS OutpostsAWS
- AWS re:Invent 2020:: Esegui app con latenza AWS Wavelength ultra bassa su 5G Edge
- AWS re:Invent 2022 - AWS Local Zones: creazione di applicazioni per un edge distribuito
- AWS re:Invent 2021 - Creazione di siti Web a bassa latenza con Amazon CloudFront
- AWS re:Invent 2022 - Migliora le prestazioni e la disponibilità con AWS Global Accelerator
- AWS re:Invent 2022 - Costruisci la tua rete WAN utilizzando AWS
- AWS re:Invent 2020: gestione globale del traffico con Amazon Route 53

Esempi correlati:

- AWS Global Accelerator Workshop sul routing personalizzato
- Handling Rewrites and Redirects using Edge Functions

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:

- Si ritiene 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 usarlo
Amazon VPC IP Address Manager.	Utilizza IPAM per pianificare, seguire e monitorare gli indirizzi IP per i carichi di lavoro AWS e on-premises. 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.
Log di query DNS	Crea log di 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 sottodominio richiesto o le posizioni edge Route 53 che hanno risposto a query DNS.
Reachability Analyzer	Con Reachability Analyzer puoi analizzare la raggiungibilità della rete ed eseguirne il debug. Reachability Analyzer è uno strumento di analisi della configurazione che consente di eseguire test di connettività tra una risorsa di origine e una risorsa di destinazione nei VPC. Lo strumento in questione consente di verificare la corrispondenza fra configurazione e connettività desiderata.

Strumento AWS	Dove usarlo
Strumento di analisi degli accessi alla rete	È possibile utilizzare lo Strumento di analisi degli accessi alla rete per comprendere l'accesso di rete alle risorse. Puoi usare lo Strumento di analisi degli accessi alla rete per specificare i requisiti di accesso alla rete e identificare i potenziali percorsi di rete che non li soddisfano. Ottimizzando la configura zione 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 opportune 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 monitoraggio continuo delle metriche è una procedura utile per osservare e identificare lo stato e l'utilizzo della rete che semplifica l'ottimizzazione della configurazione di rete in base alle osservazi oni.

Strumento AWS	Dove usarlo
AWS Network Manager	Grazie a AWS Network Manager, puoi monitorare le prestazioni in tempo reale e cronologiche 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'inter no di ciascuna zona di disponibilità, permetten doti di comprendere meglio la relazione fra prestazioni delle applicazioni e 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 instradamento nella tua rete per verificare che venga sempre utilizzato il percorso più breve tra le destinazioni. Lo Strumento di analisi degli accessi alla rete è utile in questa operazione.

Risorse

Documenti correlati:

- Public DNS query logging
- What is IPAM?
- What is Reachability Analyzer?
- What is Network Access Analyzer?
- CloudWatch metrics for your VPCs

- Optimize performance and reduce costs for network analytics with VPC Flow Logs in Apache Parquet format
- 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:

- Workshop sulle reti AWS
- 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:

- Infrastructure as code: definisci il tuo modello Infrastructure as code tramite approcci come i modelli di 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 implementazione: usa una pipeline di integrazione continua/implementazione 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 rimozione di oggetti inutili (garbage collection) 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 i percorsi utente sintetici o pre-registrati. Tali script devono essere idempotenti e non devono essere associati in coppie.

Inoltre, potrebbe essere necessario includere script preliminari per garantire risultati validi. Testa gli script il più possibile, per assicurarti che replichino le abitudini di utilizzo in produzione. Puoi usare soluzioni software as a service (SaaS) per generare il carico. Valuta se l'utilizzo delle soluzioni Marketplace AWS e le istanze spot possono essere modi convenienti per generare il 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.
- Revisione regolare dei processi: 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 Load Esegui un test del tuo carico di lavoro
- PERF05-BP05 Uso dell'automazione per risolvere in modo proattivo i problemi relativi alle prestazioni
- PERF05-BP06 Conserva il carico di lavoro e i servizi up-to-date
- 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 integrità e prestazioni di un carico di lavoro correlato a un obiettivo aziendale.

Anti-pattern comuni:

- Monitoraggio dei parametri a livello di sistema solo per avere una visione del carico di lavoro e mancata valutazione degli impatti aziendali di tali parametri.
- Si suppone che i KPI siano già in fase di pubblicazione e condivisi come dati parametrici standard.
- Mancata definizione di un KPI quantitativo e misurabile.
- Mancato allineamento dei KPI a obiettivi o strategie aziendali.

Vantaggi dell’adozione di questa best practice: l’individuazione di KPI specifici che rappresentino integrità e prestazioni del carico di lavoro aiuta ad allineare i team alle priorità e a definire risultati aziendali ottimali. La condivisione di tali metriche 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 in termini di 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 vari punti dati che misurano l’esperienza dell’utente. Oltre a identificare le soglie di tempo di caricamento della pagina, occorre documentare il risultato atteso o il rischio aziendale in caso di mancato raggiungimento delle prestazioni ideali. Un lungo tempo di caricamento della pagina si ripercuote direttamente sugli utenti finali, peggiora la loro esperienza d’uso e può portare a una perdita di clienti. Quando definisci le soglie degli indicatori chiave di prestazione, devi combinare benchmark di settore e 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 cronologici 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, mappati sui parametri da monitorare. Gli indicatori chiave di prestazione devono essere riesaminati in caso di cambiamento di obiettivi aziendali, strategie o requisiti degli utenti finali.

Passaggi dell'implementazione

- Identifica 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 di 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 in linea con gli obiettivi del carico di lavoro e in grado di aiutarti a misurare prestazioni e 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 fissare i valori dei KPI del carico di lavoro. Utilizza queste informazioni per impostare soglie dei KPI per livello di gravità o allarme. Identifica e documenta il rischio e l'impatto in caso di mancato raggiungimento del KPI.
- Implementa il monitoraggio: utilizza strumenti di monitoraggio, come Amazon CloudWatch o AWS Config, per la raccolta di metriche e la misurazione dei KPI.
- Comunica visivamente i KPI: utilizza strumenti del pannello di controllo, come Amazon Quick, per visualizzare e comunicare i KPI alle parti interessate.
- Analizza e ottimizza: esamina e analizza in modo regolare i parametri per individuare le aree del carico di lavoro da migliorare. Collabora con le parti interessate per implementare tali miglioramenti.
- Riesamina e perfeziona: rivedi con regolarità metriche e KPI per valutare la loro efficacia, soprattutto in caso di modifica di obiettivi aziendali o prestazioni del carico di lavoro.

Risorse

Documenti correlati:

- CloudWatch documentation
- Monitoring, Logging, and Performance AWS Partners
- 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
- Documentazione di X-Ray
- Using Amazon CloudWatch dashboards
- KPI di Quick

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:

- Creazione di un pannello di controllo con Quick

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:

- Si ritiene che i parametri di calcolo standard, ad esempio l'utilizzo della CPU o il carico della memoria, siano sufficienti per rilevare problemi di prestazioni.
- Utilizzo solo dei parametri predefiniti registrati dal software di monitoraggio selezionato.
- Revisione dei parametri solo quando c'è un problema.

Vantaggi dell'adozione di questa best practice: l'individuazione delle aree critiche delle prestazioni consente ai proprietari del carico di lavoro di monitorare i KPI e dare priorità ai miglioramenti ad alto impatto.

Livello di rischio associato se questa best practice non fosse adottata: 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 test o monitoraggio del carico.

Aumenta l'efficienza delle prestazioni esaminando 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.

Servizio	Dove usarlo
Monitoraggio degli utenti reali di Amazon CloudWatch (RUM)	Per acquisire i parametri delle prestazioni delle applicazioni da sessioni lato client e frontend di utenti reali.
AWS X-Ray	Per tenere traccia del traffico nei livelli dell'applicazione e identificare la latenza tra componenti e dipendenze. Utilizza le mappe del servizio X-Ray per osservare le relazioni e la latenza tra i componenti del carico di lavoro.
Informazioni dettagliate sulle prestazioni del servizio Amazon Relational Database	Per osservare i parametri delle prestazioni del database e identificare le prestazioni da migliorare.
Monitoraggio avanzato di Amazon RDS	Per osservare i parametri delle prestazioni del sistema operativo del database.
Amazon DevOps Guru	Per rilevare modelli operativi anomali in modo da poter identificare i problemi operativi prima che abbiano un impatto sui clienti.

- 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 di CloudWatch Synthetic per simulare le attività degli utenti basate sul browser in modo programmatico utilizzando espressioni della frequenza o processi CRON di Linux per generare parametri coerenti nel tempo.
  - Usa la soluzione Test di carico distribuito di AWS per generare picchi di traffico o testare il carico di lavoro al tasso di crescita previsto.
- Valuta parametri e 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 utilizzare CloudWatch Evidently per testare nuovi miglioramenti e gli impatti in termini di prestazioni sul tuo carico di lavoro.

Risorse

Documenti correlati:

- What's new in AWS Observability at re:Invent 2023
- Amazon Builders' Library
- Documentazione di X-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:

- Misurazione dei tempi di caricamento delle pagine con Amazon CloudWatch Synthetics
- Client Web Amazon CloudWatch RUM
- SDK X-Ray per Python
- Test del carico distribuito su 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:

- Si ritiene che l'architettura corrente diventi statica e non venga aggiornata nel corso del tempo.
- Introduzione di modifiche all'architettura nel tempo senza dei parametri che le giustifichino.

Vantaggi dell'adozione di questa best practice: definire un processo per apportare modifiche all'architettura consente ai dati raccolti di influenzare la progettazione del carico di lavoro nel corso del tempo.

Livello di rischio associato se questa best practice non fosse adottata: 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 o 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: stabilisci i KPI in termini di 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 come base il carico di lavoro.
- Implementa il monitoraggio: sfrutta gli strumenti di osservabilità AWS per raccogliere metriche delle prestazioni e misurare i KPI.
- Effettua analisi: conduci anali approfondite per individuare le aree (come la configurazione e il codice applicativo) del carico di lavoro 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 gli ambienti sandbox o di preproduzione per convalidare l'efficacia della strategia di miglioramento.
- Implementa le modifiche: implementa le modifiche nella produzione e monitora in modo continuo le prestazioni del carico di lavoro. Documenta i miglioramenti e comunica i risultati alle parti interessate.
- Riesamina e perfeziona: rivedi con regolarità il processo di miglioramento delle prestazioni per individuare le aree di miglioramento.

Risorse

Documenti correlati:

- AWS Blog
- 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:

- GitHub AWS

PERF05-BP04 Load Esegui un test del tuo carico di lavoro

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

Anti-pattern comuni:

- Test delle singole parti del carico di lavoro, ma non dell'intero carico di lavoro.
- Test di carico eseguito su un'infrastruttura diversa dall'ambiente di produzione.
- Test di carico eseguiti solo per il carico previsto e non oltre, per prevedere dove si potrebbero riscontrare problemi futuri.
- Esegui test di carico senza consultare la Amazon EC2 Testing Policy e inviare un modulo di invio di eventi simulati. Ciò comporta la mancata esecuzione del test, in quanto sembra un evento. denial-of-service

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

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

Guida all'implementazione

Il test di carico nel cloud è un processo volto a 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). Eseguite automaticamente i test di carico come parte della vostra pipeline di distribuzione e confrontate i risultati con soglie e soglie predefiniteKPIs. Questo processo ti consente di ottenere le prestazioni richieste.

Passaggi dell'implementazione

- Definisci gli obiettivi dei test: individua gli aspetti in termini di prestazione del carico di lavoro da valutare, come il throughput e il tempo di risposta.
- Seleziona uno strumento di test: scegli e configura lo strumento di test più adatto al carico di lavoro.
- Configura l'ambiente: configura l'ambiente di test in base al tuo ambiente di produzione. Puoi utilizzare AWS i servizi per eseguire ambienti su scala di produzione per testare la tua architettura.
- Implementa il monitoraggio: utilizza strumenti di monitoraggio come Amazon CloudWatch per raccogliere metriche tra le risorse della tua architettura. Puoi anche raccogliere e pubblicare metriche personalizzate.

- Definisci gli scenari definisci scenari e parametri del test di carico (come la durata del test e il numero di utenti).
- Esegui test di carico: effettua scenari di test su vasta scala. Approfittane Cloud AWS per testare il tuo carico di lavoro e scoprire dove non riesce a scalare o se è scalabile in modo non 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 dei test: analizza i risultati per individuare colli di bottiglia delle prestazioni e aree di miglioramento.
- Documenta e condividi gli esiti: documenta esiti e raccomandazioni e crea report al riguardo. Condividi queste informazioni con le parti interessate per aiutarle a prendere decisioni informate sulle strategie di ottimizzazione delle prestazioni.
- Effettua iterazioni continue: esegui con regolarità i test di carico, specie dopo una modifica o un aggiornamento del sistema.

Risorse

Documenti correlati:

- Amazon CloudWatch RUM
- Amazon CloudWatch Synthetics
- Test di carico distribuito su AWS

Video correlati:

- AWS Summit ANZ 2023: accelera con fiducia grazie ai test di carico AWS distribuiti
- AWS re:Invent 2022: scalabile AWS per i primi 10 milioni di utenti
- Soluzione con AWS soluzioni: test di carico distribuiti

- AWS re:Invent 2021 - Ottimizza le applicazioni attraverso approfondimenti sugli utenti finali con Amazon CloudWatch RUM
- Demo di Amazon CloudWatch Synthetics

Esempi correlati:

- Test di carico distribuito su 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:

- Solo il personale operativo è autorizzato ad apportare modifiche operative al carico di lavoro.
- Tutti gli allarmi giungono direttamente al team operativo senza alcuna correzione proattiva.

Vantaggi dell'adozione di questa best practice: la correzione proattiva delle azioni di allarme consente al personale di supporto di concentrarsi sugli elementi non attivabili in automatico. In questo modo, il personale operativo non viene sovraccaricato da tutti gli allarmi e si concentra, invece, solo sugli allarmi critici.

Livello di rischio associato se questa best practice non fosse adottata: 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 chi può 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 in automatico il rollback delle implementazioni in caso di discostamento dei KPI 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 in termini di prestazioni, per determinare se il carico di lavoro presenta prestazioni ottimali.

Passaggi dell'implementazione

- Identifica il flusso di correzione: individua e comprendi il problema delle prestazioni risolvibile 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 utilizzabile per risolvere in automatico il problema.
- Configura l'evento di avvio: configura l'evento per l'avvio automatico del 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 correzione: utilizza i servizi e le tecnologie AWS per automatizzare il processo di risoluzione. Ad esempio, AWS Systems Manager Automation fornisce un modo sicuro e scalabile per automatizzare il processo di risoluzione. Assicurati di utilizzare la logica di risoluzione automatica per annullare le modifiche se non risolvono correttamente il problema.
- Testa il flusso di lavoro: esegui il test del processo di risoluzione automatizzato in un ambiente di preproduzione.
- Implementa il flusso di lavoro: implementa la risoluzione automatizzata nell'ambiente di produzione.
- Sviluppa un playbook: predisponi e documenta un playbook che delinei le fasi del piano di risoluzione, inclusi eventi di avvio, logica di risoluzione e azioni intraprese. Assicurati di fornire la giusta preparazione alle parti interessate in modo che possano rispondere efficacemente agli eventi di risoluzione automatizzati.
- Esamina e perfeziona: valuta con regolarità l'efficacia del flusso di lavoro di risoluzione automatizzato. Modifica gli eventi di avvio e la logica di risoluzione, se necessario.

Risorse

Documenti correlati:

- CloudWatch Documentation
- Monitoraggio, registrazione di log e prestazioni: partner AWS Partner Network
- Documentazione di X-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 Conserva il carico di lavoro e i servizi up-to-date

Resta up-to-date su nuovi servizi e funzionalità cloud per adottare funzionalità efficienti, rimuovere problemi e migliorare l'efficienza complessiva delle prestazioni del tuo carico di lavoro.

Anti-pattern comuni:

- Si ritiene che l'architettura corrente diventi statica e non venga aggiornata nel corso del tempo.
- Non si dispone di sistemi né si esegue regolarmente una valutazione per la compatibilità di software e pacchetti aggiornati con il carico di lavoro.

Vantaggi derivanti dall'adozione di questa best practice: stabilendo un processo per rimanere aggiornato up-to-date su nuovi servizi e offerte, puoi adottare nuove funzionalità e funzionalità, risolvere problemi e migliorare le prestazioni del carico di lavoro.

Livello di rischio associato se questa best practice non fosse adottata: 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 in che modo possono migliorare le prestazioni

o aumentare l'efficienza del carico di lavoro tramite valutazioni, discussioni interne o analisi esterne. Definisci un processo per valutare gli aggiornamenti, le nuove funzionalità e i servizi pertinenti 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 test delle prestazioni per misurare l'impatto sulle prestazioni del carico di lavoro.

Passaggi dell'implementazione

- Esegui l'inventario del tuo carico di lavoro: esegui l'inventario di software e architettura del carico di lavoro e identifica i componenti da aggiornare.
- Identifica le origini dell'aggiornamento: identifica novità e origini dell'aggiornamento relative ai componenti del carico di lavoro. Ad esempio, puoi iscriverti al AWS blog What's New at per i prodotti che corrispondono al tuo componente di carico di lavoro. Puoi iscriverti al RSS feed o gestire le tue iscrizioni e-mail.
- Definisci un programma di aggiornamento: definisci un programma per valutare nuovi servizi e funzionalità per il tuo carico di lavoro.
  - Puoi utilizzare AWS Systems Manager Inventory per raccogliere i metadati del sistema operativo (OS), delle applicazioni e delle istanze dalle tue EC2 istanze Amazon e capire rapidamente quali istanze eseguono il software e le configurazioni richieste dalla tua politica 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.
- Utilizza l'automazione: sfrutta 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 utilizzare CI/CD per aggiornare AMIs automaticamente le immagini dei container e altri elementi relativi alla tua applicazione cloud.
  - È possibile utilizzare strumenti come AWS Systems Manager Patch Manager per automatizzare il processo di aggiornamento del sistema e pianificare l'attività utilizzando le finestre di manutenzione di AWS Systems Manager.
- Documenta il processo: documenta il tuo processo di valutazione di aggiornamenti e nuovi servizi. Fornisci ai proprietari il tempo e lo spazio necessari per ricercare, testare, sperimentare e convalidare aggiornamenti e nuovi servizi. Fate riferimento ai requisiti aziendali documentati e aiutateci KPIs a stabilire le priorità degli aggiornamenti che avranno un impatto positivo sull'azienda.

Risorse

Documenti correlati:

- Blog AWS
- Cosa c'è di nuovo con AWS
- Implementazione di up-to-date immagini con pipeline automatizzate di EC2 Image Builder

Video correlati:

- AWS RE:InForce 2022 - Automatizzazione della gestione e della conformità delle patch utilizzando AWS
- All Things Patch: | Eventi AWS Systems ManagerAWS

Esempi correlati:

- Inventory and Patch Management
- One Observability Workshop

PERF05-BP07 Analisi dei parametri a intervalli regolari

Nell'ambito della manutenzione ordinaria o in risposta a eventi o incidenti, esamina i parametri 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:

- Si lascia che i parametri rimangano in uno stato di allarme per un lungo periodo di tempo.
- Creazione di allarmi non utilizzabili da un sistema di automazione.

Vantaggi dell'adozione di questa best practice: esamina in modo continuo i parametri raccolti per verificare che identifichino, risolvano o prevengano adeguatamente i problemi. I parametri possono anche diventare obsoleti se lasciati in uno stato di allarme per un lungo periodo di tempo.

Livello di rischio associato se questa best practice non fosse adottata: 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 in modo più rapido 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, così da prevenire o risolvere più rapidamente gli incidenti futuri.

Passaggi dell’implementazione

- Definisci metriche: stabilisci metriche in termini di prestazioni critiche da monitorare, allineate all’obiettivo del carico di lavoro, incluse metriche quali il tempo di risposta e l’utilizzo delle risorse.
- Stabilisci una base: imposta un valore di base e auspicabile per ciascuna metrica. La base e deve fornire i punti di riferimento per identificare deviazioni o anomalie.
- Imposta una cadenza: imposta una cadenza (ad esempio, settimanale o mensile) per rivedere le metriche più critiche.
- Identifica i problemi di prestazioni: durante ogni revisione, valuta tendenze e deviazione dai valori di base. Cerca eventuali colli di bottiglia 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 identificare le azioni correttive, come l’ottimizzazione dei parametri, la correzione di bug e il dimensionamento delle risorse.
- Documenta gli esiti: documenta gli esiti, compresi i problemi identificati, le cause principali e le azioni correttive.
- Itera migliora: valuta e migliora continuamente il processo di revisione delle metriche. Usa le indicazioni apprese dalla revisione precedente per migliorare il processo nel tempo.

Risorse

Documenti correlati:

- CloudWatch Documentation

- 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
- Documentazione di X-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:

- Creazione di un pannello di controllo con Quick
- CloudWatch Dashboards

Conclusioni

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 aiuta a fare evolvere l'architettura in modo rapido e sicuro, mentre potrai utilizzare i dati per prendere decisioni informate 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 ad aiutarvi a creare architetture che funzionino in modo efficiente offrendo al contempo valore aziendale. Utilizza gli strumenti e le tecniche illustrati in questo documento per avere successo.

Collaboratori

Le seguenti persone e organizzazioni hanno contribuito a 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
- AWS Architecture Center

Revisioni del documento

Per ricevere una notifica sugli aggiornamenti del presente whitepaper, iscriviti al feed RSS.

Modifica	Descrizione	Data
Aggiornamento secondario alle best practice	PERF03-BP04 è stato aggiornato con nuovi suggerimenti sui servizi.	6 novembre 2024
Linee guida sulle best practice aggiornate	Diversi aggiornamenti di entità minore per tutto il pilastro.	27 giugno 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 selezione dell'architettura, calcolo e hardware, gestione dei dati, rete e distribuzione dei contenuti e processi e cultura.	3 ottobre 2023
Aggiornamento secondario	Rimozione del linguaggio non inclusivo.	13 aprile 2023
Aggiornamenti per il nuovo framework	Best practice aggiornate con prontuario e nuove best practice aggiunte.	10 aprile 2023
Aggiornamento del whitepaper	Best practice aggiornate con nuova guida all'implementazione.	15 dicembre 2022

Aggiornamento del whitepaper	Ampliamento delle best practice e aggiunta dei piani di miglioramento.	20 ottobre 2022
Aggiornamento secondario	Rimozione del linguaggio non inclusivo.	22 aprile 2022
Aggiornamenti minori	Link aggiornati.	10 marzo 2021
Aggiornamenti minori	Timeout AWS Lambda modificato in 900 secondi e corretto il nome di Amazon Keyspaces (per Apache Cassandra).	5 ottobre 2020
Aggiornamento secondario	Correzione di un link danneggiato.	15 luglio 2020
Aggiornamenti per il nuovo framework	Revisione e aggiornamento importanti dei contenuti	8 luglio 2020
Aggiornamento del whitepaper	Aggiornamento minore per la correzione di problemi grammaticali	1° luglio 2018
Aggiornamento del whitepaper	Aggiornamento del whitepaper per rispecchiare le modifiche apportate a AWS	1° novembre 2017
Pubblicazione iniziale	Pubblicazione del pilastro dell'efficienza delle prestazioni - Framework AWS Well-Architected.	1° novembre 2016

Note

I clienti sono responsabili della propria valutazione indipendente delle informazioni contenute nel presente documento. Questo documento: (a) è solo a scopo informativo, (b) rappresenta le offerte e le pratiche attuali di AWS prodotti, che sono soggette a modifiche senza preavviso, e (c) non crea alcun impegno o assicurazione da parte dei suoi affiliati, AWS fornitori o licenzianti. AWS i prodotti o i servizi sono forniti «così come sono» senza garanzie, dichiarazioni o condizioni di alcun tipo, esplicite o implicite. Le responsabilità e le responsabilità dei AWS propri clienti sono regolate da AWS accordi e il presente documento non fa parte di, né modifica, alcun accordo tra AWS e i suoi clienti.

AWS Glossario

Per la AWS terminologia più recente, consultate il AWS glossario nella sezione Reference.Glossario AWS