# Il principio dell'efficienza delle prestazioni

Canone di architettura AWS

*Luglio 2020*

# Avvisi

I clienti sono responsabili della propria valutazione autonoma delle informazioni contenute in questo documento. Questo documento: (a) è solo a scopo informativo, (b) mostra le offerte e le pratiche attuali dei prodotti AWS soggette a modifiche senza preavviso e (c) non crea alcun impegno o garanzia da parte di AWS e dei suoi affiliati, fornitori o licenziatari. I prodotti o servizi AWS sono forniti "così come sono" senza garanzie, dichiarazioni o condizioni di alcun tipo, sia esplicite che implicite. Le responsabilità e gli obblighi di AWS verso i propri clienti sono disciplinati dagli accordi AWS e il presente documento non fa parte né modifica alcun accordo tra AWS e i suoi clienti.

# Sommario

# Riassunto

Questo whitepaper si concentra sul principio dell'efficienza delle prestazioni del canone di architettura di Amazon Web Services (AWS). Fornisce istruzioni per aiutarti ad applicare best practice per la progettazione, la distribuzione e la manutenzione degli ambienti AWS.

Il principio dell'efficienza delle prestazioni riguarda le best practice per la gestione degli ambienti di produzione. Questo documento non illustra la progettazione e la gestione di ambienti e processi non di produzione, come l'integrazione continua o la distribuzione.

# Introduzione

Il canone di architettura AWS aiuta a comprendere i pro e i contro delle decisioni che vengono prese durante la progettazione dei carichi di lavoro in AWS. Utilizzando il canone, scoprirai le best practice architetturali per progettare e gestire carichi di lavoro affidabili, sicuri, efficienti e convenienti nel cloud. Il canone permette di misurare in modo coerente le architetture secondo le best practice e di identificare le aree da migliorare. Disporre di carichi di lavoro ben architettati aumenta notevolmente la probabilità di successo aziendale.

Il Canone si basa su cinque principi:

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

Il presente documento è incentrato sull'applicazione del principio dell'efficienza delle prestazioni ai carichi di lavoro. Nei tradizionali ambienti in locale, raggiungere prestazioni durature e di alto livello può essere difficoltoso. L'utilizzo dei principi contenuti in questo documento ti aiuterà a creare architetture in AWS in grado di offrire prestazioni efficienti e sostenute nel tempo.

Questo documento è rivolto a chi svolge ruoli tecnologici, ad esempio ai Chief Technology Officer (CTO), ai progettisti, agli sviluppatori e ai membri dei team operativi. Dopo avere letto questo documento, comprenderai le best practice di AWS e le strategie da utilizzare durante la progettazione di architetture di un ambiente cloud dalle prestazioni elevate. Il presente documento non fornisce dettagli sull'implementazione o sui modelli architetturali. Tuttavia, include riferimenti alle risorse in cui trovare tali informazioni.

# Efficienza delle prestazioni

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

## Principi di progettazione

I seguenti principi di progettazione possono aiutarti a raggiungere e mantenere carichi di lavoro efficienti nel cloud.

- **Estendi a tutti le tecnologie avanzate:** facilita l'implementazione di tecnologie avanzate da parte del tuo team delegando le attività complesse al tuo fornitore di cloud. Anziché chiedere al team IT di imparare come adottare e gestire una nuova tecnologia, valuta l'opportunità di utilizzare la tecnologia come servizio. Ad esempio, i database NoSQL, la transcodifica multimediale e il machine learning sono tutte tecnologie che richiedono competenze specialistiche. Nel cloud, tali tecnologie diventano servizi che il tuo team può semplicemente utilizzare mentre si concentra sullo sviluppo di un prodotto invece che sul provisioning e sulla gestione delle risorse.
- **Passa a una disponibilità a livello globale in pochi minuti:** distribuire il carico di lavoro in più regioni AWS in tutto il mondo ti consente di ridurre la latenza e di fornire un'esperienza migliore ai 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 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.
- **Privilegia un approccio orientato all'automazione:** sfrutta la strategia tecnologica più adatta ai tuoi obiettivi. Ad esempio, prendi in considerazione i modelli di accesso ai dati quando scegli una strategia basata su database o storage.

# Definizione

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

- Selezione
- Revisione
- Monitoraggio
- Compromessi

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

Rivedendo le tue decisioni a intervalli regolari, avrai la certezza di sfruttare le capacità in continua evoluzione di AWS Cloud. Il monitoraggio ti assicurerà di conoscere 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.

# Selezione

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

Le risorse AWS sono disponibili in numerose tipologie e configurazioni, il che semplifica la ricerca di un approccio che soddisfi appieno le tue esigenze. Inoltre, puoi trovare opzioni difficilmente raggiungibili con le infrastrutture in locale. Ad esempio, un servizio gestito come Amazon DynamoDB offre un database NoSQL interamente gestito, con una latenza di pochissimi millisecondi a prescindere dalle dimensioni.

## Selezione dell'architettura delle prestazioni

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

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

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

**Conosci i servizi e le risorse disponibili:** scopri tutte le informazioni sull'ampia gamma di servizi e risorse disponibili nel cloud. Identifica quali servizi e opzioni di configurazione sono pertinenti per il tuo carico di lavoro e studia come utilizzarli per raggiungere prestazioni ottimali.

Se stai valutando un carico di lavoro esistente, devi generare un inventario delle varie risorse di servizi che utilizza. Tale inventario ti aiuta a valutare quali componenti possono essere sostituiti con servizi gestiti e tecnologie più recenti.

**Definisci un processo per le scelte architettoniche:** affidati all'esperienza e alle competenze interne o utilizza risorse esterne, come casi d'uso pubblicati, documentazione rilevante o whitepaper, per definire un processo per scegliere risorse e servizi. È necessario definire un processo che incoraggi la sperimentazione e il benchmarking con i servizi che potrebbero essere utilizzati nel tuo carico di lavoro.

Durante lo studio degli scenari utente critici per la tua architettura, devi includere i requisiti relativi alle prestazioni, specificando ad esempio con quale rapidità deve essere eseguito ogni scenario. Per questi scenari critici, devi implementare percorsi utente con script aggiuntivi per chiarire esattamente quali sono le loro prestazioni rispetto ai requisiti.

**Tieni in considerazione i requisiti di costo nelle decisioni:** i carichi di lavoro spesso implicano dei requisiti di costo per il loro funzionamento. Utilizza i controlli dei costi interni per selezionare le dimensioni e i tipi di risorse in base alle necessità previste in termini di risorse.

Determina quali componenti del carico di lavoro possono essere sostituiti con servizi completamente gestiti, ad esempio database gestiti, cache in memoria e altri servizi. La riduzione del carico di lavoro operativo consente di concentrare le risorse sui risultati aziendali.

Per le best practice relative ai requisiti di costo, consulta la sezione *Risorse convenienti* del whitepaper Il principio dell'ottimizzazione dei costi.

**Utilizza policy o architetture di riferimento:** massimizza le prestazioni e l'efficienza valutando le policy interne e le architetture di riferimento esistenti e sfrutta la tua analisi per selezionare servizi e configurazioni per il carico di lavoro.

**Segui le indicazioni del tuo fornitore di servizi cloud o di un partner appropriato:** utilizza le risorse del fornitore di servizi cloud, come solutions architect, servizi professionali o un partner appropriato per orientare le tue decisioni. Queste risorse possono aiutarti a rivedere e migliorare l'architettura per ottenere prestazioni ottimali.

Contatta AWS per ricevere assistenza quando ti occorrono ulteriori indicazioni o informazioni sui prodotti. I solutions architect di AWS e i Servizi professionali di AWS forniscono indicazioni per l'implementazione delle soluzioni. I partner APN mettono a disposizione la propria conoscenza di AWS per aiutarti ad assicurare alla tua azienda agilità e innovazione

**Effettua il benchmarking dei carichi di lavoro esistenti:** effettua un'analisi comparativa delle prestazioni di un carico di lavoro esistente per comprenderne le prestazioni nel cloud. Utilizza i dati raccolti da questi benchmark per orientare le decisioni architetturali.

Utilizza il benchmarking con test sintetici per generare dati sulle prestazioni dei componenti del 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 effettuare un'analisi comparativa dei carichi di lavoro di data warehousing. I benchmark di settore sono utili quando devi confrontare ambienti diversi. Quelli personalizzati, invece, sono indicati per analizzare tipi specifici di operazioni che prevedi di eseguire nell'architettura.

In fase di benchmarking, è importante effettuare delle operazioni preliminari sull'ambiente di test al fine di garantire la validità dei risultati. Devi eseguire lo stesso benchmark più volte, per assicurarti di avere acquisito ogni eventuale variazione nel corso del tempo.

Dal momento che, di solito, l'esecuzione dei benchmark è più rapida di quella dei test di carico, il benchmarking può essere impiegato sin dalle prime fasi della pipeline di distribuzione per 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.

**Esegui un test di carico sul carico di lavoro:** distribuisci l'architettura del carico di lavoro più recente nel cloud utilizzando tipologie e dimensioni di risorse diverse. Monitora la distribuzione per acquisire parametri prestazionali che identificano colli di bottiglia o capacità in eccesso. Utilizza queste informazioni sulle prestazioni per progettare o migliorare la tua architettura e la selezione delle risorse.

I test di carico utilizzano il carico di lavoro *effettivo:* in questo modo puoi osservare le prestazioni dell'intera soluzione in un ambiente di produzione. Occorre eseguire i test di carico tramite versioni sintetiche o purificate dei dati di produzione (rimuovendo le informazioni sensibili o che permettono l'identificazione degli utenti). Utilizza percorsi utente riprodotti o già programmati su tutto il carico di lavoro su vasta scala verificando l'intera architettura. Esegui automaticamente test di carico come parte della pipeline di distribuzione e confronta i risultati con KPI e soglie predefiniti. In questo, modo puoi continuare a raggiungere le prestazioni richieste.

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 per configurare allarmi che segnalino la violazione delle soglie impostate, in modo da avvertire che i risultati di un test si trovano al di fuori delle prestazioni previste.

I servizi AWS ti consentono di eseguire ambienti in ambito di produzione per sottoporre l'architettura a test in maniera aggressiva. Dal momento che paghi l'ambiente di test solo quando ti serve, puoi effettuare test su scala completa a un costo estremamente ridotto rispetto all'uso di un ambiente in locale. Sfrutta i vantaggi offerti da AWS Cloud per testare il carico di lavoro e per verificare se non si ridimensiona o si ridimensiona in modo non lineare. Usa le istanze Amazon EC2 Spot per generare carichi a costi ridotti e rilevare gli ostacoli alle prestazioni prima che si verifichino in produzione.

Nei casi in cui l'esecuzione dei test di carico richieda molto tempo, puoi eseguirli in parallelo tramite copie multiple dell'ambiente di test. I costi resteranno simili, ma ridurrai i tempi necessari per il completamento dei test. (Eseguire un'istanza EC2 per 100 ore costa quanto eseguire 100 istanze per un'ora). Puoi ridurre i costi dei test anche tramite le istanze Spot, selezionando le regioni dai costi inferiori rispetto a quelle che usi per la produzione.

La posizione dei client per il test di carico deve rispecchiare la distribuzione geografica degli utenti finali.

## Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice di AWS in merito ai test di carico.

**Video**

- Presentazione di Amazon Builders' Library (DOP328)

**Documentazione**

- Centro architetturale AWS
- Ottimizzazione delle prestazioni di Amazon S3
- Prestazioni dei volumi di Amazon EBS
- AWS CodeDeploy
- AWS CloudFormation
- CloudFront per i test di carico
- Pannelli di controllo di AWS CloudWatch

# Selezione dell'architettura di elaborazione

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

**Valuta le opzioni di elaborazione disponibili:** studia e comprendi le caratteristiche di prestazione delle opzioni di elaborazione disponibili. Comprendi il modo in cui funzionano le istanze, i container e le funzioni e quali sono i vantaggi e gli svantaggi che comportano per il tuo carico di lavoro.

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

## Istanze

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

Le istanze dei server virtuali di Amazon Elastic Compute Cloud (Amazon EC2) sono disponibili in una varietà di famiglie e dimensioni. Offrono un'ampia gamma di funzionalità, fra cui SSD e GPU. Quando avvii un'istanza EC2, il tipo specificato determina l'hardware del computer host da utilizzare per l'istanza stessa. Ciascun tipo di istanza offre varie capacità di elaborazione, memoria e storage. I tipi di istanza sono raggruppati in famiglie di istanze in base alle loro funzionalità.

Puoi basarti sui dati per selezionare il tipo di istanza EC2 ottimale per il tuo carico di lavoro, scegliere le opzioni di rete e storage corrette e prendere in considerazione le impostazioni del sistema operativo in grado di migliorare le prestazioni del tuo carico di lavoro.

## Container

I container rappresentano un metodo di virtualizzazione del sistema operativo con cui puoi eseguire un'applicazione e le relative dipendenze in processi isolati dalle risorse.

Quando esegui container in AWS, puoi scegliere tra due opzioni. In primo luogo, devi scegliere se gestire o meno i server. AWS Fargate è un servizio di elaborazione serverless per container, oppure puoi scegliere Amazon EC2 se hai bisogno di controllare l'installazione, la configurazione e la gestione del tuo ambiente di elaborazione. In secondo luogo, scegli quale orchestratore di container utilizzare: Amazon Elastic Container Service (ECS) o Amazon Elastic Kubernetes Service (EKS).

Amazon Elastic Container Service (Amazon ECS) è un servizio di orchestrazione di container completamente gestito che consente di eseguire e gestire automaticamente i container su un cluster di istanze EC2 o istanze serverless utilizzando AWS Fargate. Amazon ECS può essere integrato in modo nativo con altri servizi quali Amazon Route 53, Secrets Manager, AWS Identity and Access Management (IAM) e Amazon CloudWatch.

Amazon Elastic Kubernetes Service (Amazon EKS) è un servizio Kubernetes completamente gestito. Puoi scegliere di eseguire i cluster EKS utilizzando AWS Fargate, eliminando la necessità di effettuare il provisioning e di gestire i server. EKS è profondamente integrato con servizi quali Amazon CloudWatch, gruppi Auto Scaling, AWS Identity and Access Management (IAM) e Amazon Virtual Private Cloud (VPC).

Quando usi i container, devi selezionare il tipo ottimale per il tuo carico di lavoro tramite i dati, esattamente come quando devi sceglierei tipi di istanze EC2 o AWS Fargate. Devi prendere in considerazione le opzioni di configurazione del container, come memoria, CPU e configurazione dei tenant. Per abilitare l'accesso di rete tra i servizi di container, valuta la possibilità di utilizzare una mesh di servizio come AWS App Mesh, che standardizza il modo in cui i servizi comunicano. La mesh di servizio offre un visibilità end-to-end e garantisce un'elevata disponibilità per le applicazioni.

## Funzioni

Le funzioni astraggono l'ambiente di esecuzione dal codice che desideri eseguire. Ad esempio, AWS Lambda ti permette di eseguire un codice senza eseguire un'istanza.

Puoi utilizzare AWS Lambda per eseguire codice per qualsiasi tipo di applicazione o servizio di back-end senza alcuna amministrazione. È sufficiente caricare il codice: AWS Lambda gestirà tutto il necessario per eseguire e ricalibrare il codice. Puoi configurare il tuo codice affinché venga eseguito automaticamente da altri servizi AWS, richiamarlo direttamente o utilizzarlo con Amazon API Gateway.

Amazon API Gateway è un servizio completamente gestito che permette agli sviluppatori di creare, pubblicare, mantenere, monitorare e proteggere le API su qualsiasi scala in modo facile. Puoi creare una API che agisce da "front-door" verso la funzione Lambda. API Gateway gestisce tutte le attività di accettazione ed elaborazione relative a centinaia di migliaia di chiamate API simultanee, inclusi gestione del traffico, controllo di accessi e autorizzazioni, monitoraggio e gestione delle versioni delle API.

Per garantire prestazioni ottimali con AWS Lambda, scegli la quantità di memoria più adatta alla tua funzione. In questo modo, verranno allocate in modo proporzionale la potenza della CPU e altre risorse utili. Ad esempio, se scegli 256 MB di memoria, la potenza della CPU allocata per la funzione Lambda sarà circa il doppio di quella assegnata richiedendone 128 MB. Puoi anche stabilire per quanto tempo ogni funzione dovrà essere eseguita (fino a un massimo di 300 secondi).

**Comprendi le opzioni di configurazione dell'elaborazione disponibili:** comprendi in che modo le varie opzioni completano il carico di lavoro e quali opzioni di configurazione sono più adatte per il tuo sistema. Esempi di tali opzioni includono la famiglia di istanze, le dimensioni, le caratteristiche (GPU, I/O), le dimensioni delle funzioni, le istanze di container, multi-tenancy o singola, e così via.

Quando si selezionano famiglie e tipi di istanze, è necessario considerare anche le opzioni di configurazione disponibili per soddisfare le esigenze del carico di lavoro:

- **Unità di elaborazione grafica (GPU) —**l'elaborazione generica sulle GPU ti consente di sviluppare applicazioni che sfruttano il vantaggio dell'elevato grado di parallelismo offerto dalle GPU stesse, grazie all'utilizzo di piattaforme come CUDA durante il processo di sviluppo. Inoltre, se l'applicazione richiede il rendering in 3D o la compressione video, le GPU consentono di sfruttare l'elaborazione e la codifica con accelerazione hardware, rendendo la tua applicazione ancora più efficiente.
- **Field Programmable Gate Arrays (FPGA) —** gli FPGA ti consentono di ottimizzare i carichi di lavoro tramite l'esecuzione personalizzata con accelerazione hardware per quelli più impegnativi. Puoi definire gli algoritmi sfruttando i linguaggi di programmazione generale come C o Go oppure linguaggi orientati all'hardware come Verilog o VHDL.
- **AWS Inferentia (Inf1) —** Le istanze Inf1 sono create per supportare le applicazioni di inferenza di machine learning. Utilizzando le istanze Inf1, i clienti possono eseguire applicazioni di inferenza di machine learning su larga scala, tra cui riconoscimento di immagini, riconoscimento vocale, elaborazione del linguaggio naturale, personalizzazione e rilevamento di attività fraudolente. Puoi creare un modello in uno dei framework di machine learning più utilizzati, ad esempio TensorFlow, PyTorch o MXNet, e utilizzare istanze GPU come P3 o P3dn per addestrare il modello. Dopo avere addestrato il modello di machine learning per soddisfare i requisiti, puoi distribuirlo su istanze Inf1 utilizzando AWS Neuron, un kit di sviluppo software (SDK) specializzato composto da un compilatore, runtime e strumenti di profilatura che ottimizzano le prestazioni di inferenza di machine learning dei chip Inferentia.
- **Famiglie di istanze compatibili con il bursting -** Le istanze compatibili con il bursting sono progettate per offrire prestazioni di base moderate e la possibilità di effettuare il bursting, fornendo prestazioni molto più elevate, in base alle esigenze del carico di lavoro. Tali istanze sono destinate a carichi di lavoro che non utilizzano la CPU al massimo in modo frequente o costante, ma che richiedono il bursting di tanto in tanto. Sono ideali per i carichi di lavoro generici, come i web server, gli ambienti per sviluppatori e database di piccole dimensioni. Queste istanze offrono crediti CPU utilizzabili quando l'istanza deve garantire prestazioni elevate. I crediti si accumulano quando l'istanza non li richiede.
- **Funzionalità di elaborazione avanzate -** Amazon EC2 offre l'accesso a funzioni di elaborazione avanzate, come la gestione dei registri degli stati C e P e il controllo della funzione turbo-boost dei processori. L'accesso al coprocessore consente l'offloading delle operazioni di crittografia tramite AES-NI o l'elaborazione avanzata mediante estensioni AVX.

Il sistema Nitro di AWS è una combinazione di hardware dedicato e hypervisor leggero che consente un'innovazione più rapida e una maggiore sicurezza. Utilizza i sistemi Nitro di AWS, quando disponibili, per utilizzare appieno le risorse di elaborazione e memoria dell'hardware host. Inoltre, le schede Nitro dedicate consentono prestazioni di rete ed EBS a elevata velocità, oltre all'accelerazione I/O.

**Raccogli i parametri relativi all'elaborazione:** uno dei modi migliori per comprendere le prestazioni dei tuoi sistemi è registrare e tracciare l'utilizzo effettivo delle varie risorse. Questi dati possono essere utilizzati per determinare in modo più accurato i requisiti delle risorse.

I carichi di lavoro (come quelli in esecuzione su architetture di microservizi) possono generare grandi volumi di dati sotto forma di parametri, log ed eventi. Stabilisci se il servizio di monitoraggio e osservazione esistente è in grado di gestire i dati generati. Amazon CloudWatch può essere utilizzato per raccogliere, accedere e correlare questi dati su un'unica piattaforma da tutte le risorse, le applicazioni e i servizi AWS in esecuzione su server AWS e locali, in modo da ottenere facilmente visibilità a livello di sistema e risolvere rapidamente i problemi.

**Stabilisci la configurazione richiesta in base al corretto dimensionamento:** analizza le varie caratteristiche di prestazione del tuo carico di lavoro e come queste sono correlate a memoria, rete e utilizzo della CPU. Utilizza questi dati per scegliere le risorse che meglio corrispondono al profilo del tuo carico di lavoro. Ad esempio, un carico di lavoro a memoria elevata, come un database, potrebbe essere servito meglio dalla famiglia di istanze r. Al contrario, un carico di lavoro con picchi di prestazioni può trarre maggiori vantaggi da un sistema di container elastici.

**Utilizza l'elasticità disponibile delle risorse:** il cloud offre la flessibilità necessaria per espandere o ridurre le risorse in modo dinamico attraverso una serie di meccanismi per soddisfare i cambiamenti della domanda. Se combinato con parametri relativi all'elaborazione, un carico di lavoro può rispondere automaticamente a questi cambiamenti e utilizzare la gamma di risorse più opportuna per raggiungere il suo obiettivo.

La corrispondenza ottimale tra risorse fornite e domanda determina il costo più basso per il sistema. Tuttavia, sarà necessario pianificare anche una fornitura sufficiente rispetto al tempo di provisioning e che tenga conto di eventuali errori delle singole risorse. La domanda può essere fissa o variabile e richiede parametri e automazione, per garantire che la gestione non diventi troppo complicata e onerosa dal punto di vista economico.

In AWS, puoi adottare varie strategie per associare la fornitura alla domanda. Il whitepaper Il pilastro dell'ottimizzazione dei costi descrive come utilizzare i seguenti approcci ai costi:

- Approccio basato sulla domanda
- Approccio basato sui buffer
- Approccio basato sul tempo

Devi assicurarti che le distribuzioni dei carichi di lavoro siano in grado di gestire eventi di dimensionamento. Crea scenari di test per eventi di ridimensionamento per garantire che il carico di lavoro si comporti come previsto.

**Rivaluta le esigenze di calcolo in base ai parametri:** utilizza i parametri a livello di sistema per identificare il comportamento e i requisiti del carico di lavoro nel corso del tempo. Valuta le esigenze del tuo carico di lavoro confrontando le risorse disponibili con tali requisiti e apporta modifiche al tuo ambiente di elaborazione per soddisfare al meglio il profilo del carico di lavoro. Ad esempio, nel corso del tempo si potrebbe osservare che un sistema utilizza molta più memoria di quanto si pensasse inizialmente, e trasferirlo a una famiglia o una dimensione di istanze diversa potrebbe migliorarne sia le prestazioni sia l'efficienza.

## Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice di AWS in merito all'elaborazione.

**Video**

- Nozioni di base di Amazon EC2 (CMP211-R2)
- Implementare la nuova generazione di Amazon EC2: analisi approfondita del sistema Nitro
- Distribuirei inferenze ML ad alte prestazioni con AWS Inferentia (CMP324-R1)
- Ottimizzare le prestazioni e i costi dell'elaborazione AWS (CMP323-R1)
- Un'elaborazione migliore, più veloce ed economica: ottimizzazione dei costi di Amazon EC2 (CMP202-R1)

**Documentazione**

- Istanze:
  - Tipi di istanza
  - Controllo dello stato del processore per l'istanza EC2
- Container EKS: nodi di lavoro EKS
- Container ECS: istanze di container di Amazon ECS
- Funzioni: configurazione della funzione Lambda

# Selezione dell'architettura di storage

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

In AWS, lo storage è virtualizzato e disponibile in una varietà di tipi. Ciò semplifica la scelta dei metodi di storage più adatti alle tue esigenze. Inoltre, vengono offerte opzioni di storage difficilmente disponibili nelle infrastrutture in locale. Ad esempio, Amazon S3 è progettato per offrire 11 nove (99,999999999%) di durabilità. Puoi anche passare dagli hard disk magnetici (HDD) alle unità a stato solido (SSD) e spostare facilmente le unità virtuali da un'istanza all'altra in pochi secondi.

Le prestazioni possono essere misurate osservando throughput, operazioni di input/output al secondo (IOPS) e latenza. Comprendere la relazione fra questi tipi di misurazione ti aiuta a scegliere la soluzione di storage più adatta.

| Storage | Servizi | Latenza | Throughput | Possibilità di condivisione |
|---|---|---|---|---|
| **A blocchi** | Amazon EBS, Store di istanze EC2 | Il valore più basso, coerente | Singolo | Montaggio su un'istanza EC2, copia tramite snapshot |
| **File system** | Amazon EFS, Amazon FSx | Valore basso, coerente | Multiplo | Molti client |
| **A oggetti** | Amazon S3 | Bassa latenza | Scala web | Molti client |
| **Archiviazione** | Amazon S3 Glacier | Da minuti a ore | Elevata | No |

Dal punto di vista della latenza, se l'accesso ai tuoi dati avviene da parte di una sola istanza, è consigliabile utilizzare lo storage a blocchi, come Amazon EBS. I file system distribuiti come Amazon EFS tendono a generare una latenza minima per ogni operazione con i file, pertanto è opportuno utilizzarli nel caso in cui più istanze richiedano l'accesso.

Amazon S3 dispone di funzioni in grado di ridurre la latenza e aumentare il throughput. Puoi usare la replica su più regioni (CRR) per offrire l'accesso ai dati a bassa latenza a varie aree geografiche.

Dal punto di vista del throughput, Amazon EFS supporta carichi di lavoro altamente in parallelo (ad esempio, con operazioni simultanee da più thread e istanze EC2). Ciò garantisce livelli elevati di throughput aggregato e operazioni al secondo. Per Amazon EFS, esegui un benchmark o un test di carico per selezionare la modalità delle prestazioni più adeguata.

**Comprendi le caratteristiche e i requisiti dello storage:** studia le diverse caratteristiche (ad esempio condivisibilità, dimensioni dei file, dimensioni della cache, schemi di accesso, latenza, throughput e persistenza dei dati) necessarie per selezionare i servizi più adatti al carico di lavoro, ad esempio storage a oggetti, storage a blocchi, storage a file o storage dell'istanza.

Determina il tasso di crescita previsto per il carico di lavoro e scegli una soluzione di storage che soddisfi tali percentuali. Le soluzioni di storage a oggetti e a file, come Amazon S3 e Amazon Elastic File System, consentono uno storage illimitato; Amazon EBS offre invece dimensioni di storage predefinite. I volumi elastici consentono di aumentare dinamicamente la capacità, ottimizzare le prestazioni e cambiare il tipo di qualsiasi volume di generazione nuovo o esistente, senza tempi di inattività o ripercussioni sulle prestazioni, ma richiedono modifiche al file system del sistema operativo.

**Valuta le opzioni di configurazione disponibili:** valuta le varie caratteristiche e opzioni di configurazione e la loro relazione con lo storage. Comprendi dove e come utilizzare Provisioned IOPS, SSD, storage magnetico, storage a oggetti, storage di archiviazione o storage temporaneo per ottimizzare lo spazio di storage e le prestazioni del tuo carico di lavoro.

Amazon EBS offre una gamma di opzioni che ti permettono di ottimizzare le prestazioni di storage e il costo del tuo carico di lavoro. Tali opzioni sono suddivise in due categorie principali: storage basato su SSD per i carichi di lavoro relativi alle transazioni, come database e volumi di avvio (le prestazioni dipendono principalmente dagli IOPS), e storage basato su HDD per i carichi di lavoro con elevati requisiti di throughput, come MapReduce e l'elaborazione dei log (le prestazioni dipendono principalmente dalla velocità in MB/s).

I volumi basati su SSD includono SSD Provisioned IOPS dalle prestazioni più elevate per i carichi di lavoro relativi alle transazioni e sensibili alla latenza, nonché SSD generici che garantiscono un rapporto prestazioni/prezzo equilibrato per una vasta gamma di dati sulle transazioni.

Amazon S3 Transfer Acceleration consente il trasferimento rapido dei file su lunghe distanze tra il client e il tuo bucket S3. Transfer Acceleration sfrutta le edge location di Amazon CloudFront distribuite a livello globale per instradare i dati attraverso percorsi di rete ottimizzati. Per i carichi di lavoro in un bucket S3 con richieste GET intensive, utilizza Amazon S3 con CloudFront. Quando si caricano file di grandi dimensioni, è possibile utilizzare il caricamento simultaneo di più parti per ottimizzare il throughput di rete.

Amazon Elastic File System (Amazon EFS) fornisce un file system NFS elastico semplice, scalabile e completamente gestito da utilizzare con i servizi AWS cloud e le risorse locali. Per supportare un'ampia gamma di carichi di lavoro di storage nel cloud, Amazon EFS offre due modalità prestazionali: modalità di prestazioni generiche e modalità di prestazioni I/O massime. Sono disponibili anche due modalità di throughput tra cui scegliere per il file system: Bursting Throughput e Provisioned Throughput. Per determinare quali impostazioni utilizzare per il carico di lavoro, consulta la Guida per l'utente di Amazon EFS.

Amazon FSx offre due file system tra cui scegliere: Amazon FSx for Windows File Server, per i carichi di lavoro aziendali, e Amazon FSx for Lustre, per i carichi di lavoro ad alte prestazioni. FSx è basato su SSD ed è progettato per offrire prestazioni rapide, prevedibili, scalabili

e costanti. I file system di Amazon FSx offrono elevate velocità di lettura e scrittura e l'accesso costante ai dati a bassa latenza. È possibile scegliere il livello di throughput desiderato per soddisfare le esigenze del carico di lavoro.

**Prendi decisioni in base a schemi di accesso e parametri:** scegli i sistemi di storage in base agli schemi di accesso del carico di lavoro e configurali determinando il modo in cui il carico di lavoro accede ai dati. Aumenta l'efficienza dello storage scegliendo lo storage a oggetti anziché lo storage a blocchi. Configura le opzioni di storage in funzione dei tuoi schemi di accesso ai dati.

Il modo in cui accedi ai dati influisce sulle prestazioni della soluzione di storage. Seleziona la soluzione più adatta ai tuoi schemi di accesso. In alternativa, puoi modificarli affinché siano in linea con la soluzione di storage, allo scopo di ottimizzare le prestazioni.

Creare un array RAID 0 (zero) ti permette di ottenere prestazioni più elevate per i file system di cui puoi effettuare il provisioning su un unico volume. Prendi in considerazione l'uso di RAID 0 quando le prestazioni I/O sono più importanti della tolleranza agli errori. Ad esempio, puoi usarlo con un database che viene utilizzato in modo intensivo e in cui la replica dei dati è già stata configurata separatamente.

Seleziona i parametri di storage appropriati per il carico di lavoro tra tutte le opzioni di storage utilizzate per il carico di lavoro. Quando utilizzi i file system che utilizzano crediti di burst, puoi creare degli allarmi che ti informano quando stai per avvicinarti ai limiti di credito. È necessario creare pannelli di controllo di storage per visualizzare lo stato generale dello storage del carico di lavoro.

Per i sistemi di storage di dimensioni fisse, ad esempio Amazon EBS o Amazon FSx, assicurati di monitorare la quantità di storage utilizzata rispetto alle dimensioni complessive dello storage e di creare, se possibile, un'automazione per aumentare le dimensioni dello storage quando si raggiunge una soglia.

## Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice di AWS in merito allo storage.

**Video**

- Analisi approfondita di Amazon EBS (STG303-R1)
- Ottimizzare le prestazioni di storage con Amazon S3 (STG343)

**Documentazione**

- Amazon EBS:
  - Amazon EC2 Storage
  - Tipi di volume Amazon EBS
  - Caratteristiche I/O
- Amazon S3: considerazioni su velocità e prestazioni delle richieste

- Amazon Glacier: documentazione di Amazon Glacier
- Amazon EFS: prestazioni di Amazon EFS
- Amazon FSx:
    - Prestazioni di Amazon FSx for Lustre
    - Prestazioni di Amazon FSx for Windows File Server

# Selezione dell'architettura del database

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

**Comprendi le caratteristiche dei dati:** studia le diverse caratteristiche dei dati nel carico di lavoro. Determina se il carico di lavoro necessita di transazioni, in che modo interagisce con i dati e quali sono le sue esigenze in termini di prestazioni. Utilizza tali dati per selezionare l'approccio di database con le prestazioni migliori per il tuo carico di lavoro (ad esempio storage con database relazionali, chiave-valore NoSQL, documento, colonnare, grafi, serie temporali o in memoria).

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

I database relazionali memorizzano i dati con schemi predefiniti e le relazioni tra di essi. Questi database sono progettati per supportare le transazioni ACID (atomicità, coerenza, isolamento, durabilità) e per mantenere l'integrità referenziale e una solida coerenza dei dati. Molte applicazioni tradizionali, Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) ed e-commerce utilizzano database relazionali per archiviare i propri dati. Puoi eseguire molti di questi motori di database in Amazon EC2 oppure scegliere uno dei servizi di database gestiti di AWS: Amazon Aurora, Amazon RDS e Amazon Redshift.

I database chiave-valore sono ottimizzati per schemi di accesso di uso comune, in genere per archiviare e recuperare grandi volumi di dati. Questi database offrono tempi di risposta rapidi, anche nel caso di volumi estremi di richieste simultanee.

Le applicazioni web a traffico elevato, i sistemi di e-commerce e le applicazioni di gaming sono casi d'uso tipici dei database chiave-valore. In AWS, è possibile utilizzare Amazon DynamoDB, un database completamente gestito, multi-regione, multi-master e durevole con capacità integrate di sicurezza, backup e ripristino e caching in memoria per applicazioni su scala Internet.

I database in memoria vengono utilizzati per applicazioni che richiedono l'accesso in tempo reale ai dati. Archiviando i dati direttamente in memoria, questi database forniscono una latenza di microsecondi alle applicazioni per le quali la latenza di millisecondi non è sufficiente. Puoi utilizzare database in memoria per il caching delle applicazioni, la gestione delle sessioni, l'archiviazione delle sessioni di gioco e le applicazioni geospaziali. Amazon ElastiCache è un datastore in memoria completamente gestito, compatibile con Redis o Memcached.

Un database a documento è progettato per archiviare dati semistrutturati come documenti di tipo JSON. Questi database aiutano gli sviluppatori a creare e aggiornare rapidamente applicazioni quali gestione di contenuti, cataloghi e profili utente. Amazon DocumentDB è un servizio di database a documento rapido, scalabile, a elevata disponibilità e completamente gestito che supporta i carichi di lavoro MongoDB.

Uno store colonnare è un tipo di database NoSQL. Utilizza tabelle, righe e colonne, ma a differenza di un database relazionale, i nomi e il formato delle colonne possono variare da riga a riga all'interno della stessa tabella. In genere, gli store colonnari sono utilizzati nelle applicazioni industriali su larga scala per la manutenzione delle apparecchiature, la gestione delle flotte e l'ottimizzazione dei percorsi. Amazon Managed Apache Cassandra Service è un servizio di database colonnare gestito compatibile con Apache Cassandra, scalabile e altamente disponibile.

I database a grafo sono destinati ad applicazioni che devono navigare ed eseguire query su milioni di relazioni tra set di dati a grafo altamente connessi, con una latenza di millisecondi su larga scala. Molte aziende utilizzano database a grafo per il rilevamento di attività fraudolente, i social network e i motori di raccomandazione. Amazon Neptune è un servizio di database a grafo veloce, affidabile e completamente gestito che semplifica la creazione e l'esecuzione di applicazioni che funzionano con set di dati altamente connessi.

I database di serie temporali raccolgono, sintetizzano e derivano informazioni approfondite dai dati che cambiano nel tempo. I database di serie temporali sono spesso utilizzati dalle applicazioni IoT, DevOps e dalla telemetria industriale. Amazon Timestream è un servizio di database di serie temporali veloce, scalabile e completamente gestito per le applicazioni IoT e operative che semplifica la memorizzazione e l'analisi di migliaia di miliardi di eventi al giorno.

I database di libri mastri forniscono un'autorità centralizzata e affidabile per mantenere un registro delle transazioni scalabile, immutabile e verificabile tramite crittografia per ogni applicazione. I database di libri mastri vengono utilizzati per sistemi di record, catena di fornitura, registrazioni e persino transazioni bancarie. Amazon Quantum Ledger Database (QLDB) è un database di libri mastri completamente gestito che fornisce un log delle transazioni trasparente, immutabile e verificabile tramite crittografia, di proprietà di un'autorità centrale attendibile. Amazon QLDB tiene traccia di ogni modifica ai dati dell'applicazione e conserva una cronologia completa e verificabile delle modifiche nel corso del tempo.

**Valuta le opzioni disponibili:** valuta i servizi e le opzioni di storage disponibili come parte del processo di selezione per i meccanismi di storage del tuo carico di lavoro. Comprendi come e quando utilizzare un determinato servizio o sistema per lo storage dei dati. Scopri le opzioni di configurazione disponibili in grado di ottimizzare le prestazioni o l'efficienza del database, ad esempio Provisioned IOPS, risorse di memoria ed elaborazione e memorizzazione nella cache.

Di solito, le soluzioni di database presentano opzioni di configurazione che ti consentono di effettuare l'ottimizzazione in base al tipo di carico di lavoro. Utilizzando benchmarking o test di carico, puoi identificare i parametri del database più importanti per il tuo carico di lavoro. Prendi in considerazione le opzioni di configurazione per la strategia di database che hai scelto, come l'ottimizzazione dello storage, le impostazioni al livello del database, la memoria e la cache.

Valuta le opzioni di database con memorizzazione nella cache per il carico di lavoro. I tre tipi più comuni di database con memorizzazione nella cache sono i seguenti:

- **Cache integrate nel database:** alcuni database (ad esempio Amazon Aurora) offrono una cache integrata, che viene gestita all'interno del motore del database e dispone di funzionalità di scrittura integrate.
- **Cache locali:** una cache locale archivia i dati utilizzati di frequente all'interno dell'applicazione. Questo velocizza il recupero dei dati ed elimina il traffico di rete associato al recupero dei dati, rendendo il recupero dei dati più rapido rispetto ad altre architetture di memorizzazione nella cache.
- **Cache remote:** le cache remote vengono memorizzate su server dedicati e in genere sono basate su store NoSQL chiave-valore, come Redis e Memcached. Forniscono fino a un milione di richieste al secondo per nodo di cache.

Per i carichi di lavoro di Amazon DynamodDB, DynamoDB Accelerator (DAX) fornisce una cache in memoria completamente gestita. DAX è una cache in memoria che offre un'elevata velocità di lettura per le tabelle su vasta scala tramite l'utilizzo di una cache in memoria completamente gestita. Utilizzando DAX, è possibile migliorare le prestazioni di lettura delle tabelle DynamoDB fino a 10 volte, riducendo il tempo necessario per le letture da millisecondi a microsecondi, anche in caso di milioni di richieste al secondo.

**Raccogli e registra i parametri delle prestazioni del database:** utilizza strumenti, librerie e sistemi che registrano le misurazioni delle prestazioni correlate alle prestazioni del database. Ad esempio, misura le transazioni al secondo, le query lente o la latenza di sistema introdotta durante l'accesso al database. Utilizza questi dati per comprendere le prestazioni dei sistemi di database.

Implementa tutti i parametri dell'attività del database che puoi raccogliere dal carico di lavoro. Questi parametri potrebbero essere pubblicati direttamente dal carico di lavoro o raccolti da un servizio di gestione delle prestazioni dell'applicazione. Puoi utilizzare AWS X-Ray per analizzare ed eseguire il debug di applicazioni distribuite di produzione, come quelle create utilizzando un'architettura di microservizi. Una traccia X-Ray può includere segmenti che incapsulano tutti i punti dati per un singolo componente. Ad esempio, quando l'applicazione effettua una chiamata a un database in risposta a una richiesta, crea un segmento per la richiesta con un sottosegmento che rappresenta la chiamata al database e il relativo risultato. Il sottosegmento può contenere dati quali la query, la tabella utilizzata, il timestamp e lo stato di errore. Dopo l'implementazione, devi abilitare degli allarmi per i parametri del database che indichino quando le soglie vengono superate.

**Scegli lo storage dei dati in base agli schemi di accesso:** utilizza gli schemi di accesso del carico di lavoro per decidere quali servizi e tecnologie utilizzare. Per esempio, utilizza un database relazionale per i carichi di lavoro che necessitano di transazioni, o uno store chiave-valore che fornisce un throughput maggiore ma anche una lettura finale consistente, ove applicabile.

**Ottimizza lo storage dei dati in base a schemi di accesso e parametri:** utilizza caratteristiche prestazionali e schemi di accesso che ottimizzano il modo in cui i dati vengono archiviati o interrogati per ottenere le migliori prestazioni possibili. Misura il modo in cui le ottimizzazioni come l'indicizzazione, la distribuzione delle chiavi, la progettazione dei data warehouse o le strategie di memorizzazione nella cache influenzano le prestazioni del sistema o la sua efficienza nel complesso.

### Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice di AWS in merito ai database.

**Video**

- Database dedicati di AWS (DAT209-L)
- Sfatiamo i miti sullo storage Amazon Aurora: come funziona realmente (DAT309-R)
- Analisi approfondita di Amazon DynamoDB: modelli di progettazione avanzati (DAT403-R1)

**Documentazione**

- Memorizzazione nella cache di database AWS
- Database nel cloud con AWS
- Best practice di Amazon Aurora
- Prestazioni di Amazon RedShift
- I 10 suggerimenti più importanti per le prestazioni di Amazon Athena
- Best practice per Amazon Redshift Spectrum
- Best practice per Amazon DynamoDB
- Amazon DynamoDB Accelerator

## Selezione dell'architettura di rete

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

In AWS, le reti sono virtualizzate e vengono fornite in una varietà di tipi e configurazioni. Ciò semplifica la scelta delle metodologie di rete più adatte alle tue esigenze. AWS offre funzionalità dei prodotti (ad esempio reti avanzate, istanze ottimizzate di Amazon EC2, Amazon S3 Transfer Acceleration e Amazon CloudFront dinamico) pensate per l'ottimizzazione

del traffico di rete. AWS offre anche funzionalità di rete (ad esempio instradamento della latenza di Amazon Route 53, endpoint VPC di Amazon, AWS Direct Connect e AWS Global Accelerator) per ridurre la distanza di rete o il jitter.

**Scopri in che modo le reti influiscono sulle prestazioni:** analizza e scopri in che modo le caratteristiche correlate alla rete influiscono sulle prestazioni del carico di lavoro. Ad esempio, la latenza di rete spesso influisce sull'esperienza utente, e non fornire una capacità di rete sufficiente può compromettere le prestazioni del carico di lavoro.

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

**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.

Molti servizi offrono in genere delle funzionalità per ottimizzare le prestazioni di rete. Prendi in considerazione funzionalità dei prodotti come la funzionalità di rete dell'istanza EC2, i tipi di istanze di rete avanzate, le istanze ottimizzate di Amazon EBS, Amazon S3 Transfer Acceleration e CloudFront dinamico per ottimizzare il traffico di rete.

AWS Global Accelerator è un servizio che migliora la disponibilità e le prestazioni globali delle applicazioni utilizzando la rete globale AWS. Ottimizza il percorso di rete sfruttando la vasta rete globale AWS, priva di congestioni. Fornisce indirizzi IP statici che semplificano lo spostamento degli endpoint tra zone di disponibilità o regioni AWS senza la necessità di aggiornare la configurazione DNS o di modificare le applicazioni lato client.

L'accelerazione dei contenuti di Amazon S3 è una funzione che consente agli utenti esterni di sfruttare i vantaggi delle ottimizzazioni di rete di CloudFront per il caricamento dei dati in Amazon S3. Ciò semplifica il trasferimento di grandi quantità di dati da posizioni remote prive di connettività dedicata ad AWS Cloud.

Inoltre, le istanze EC2 possono sfruttare le reti avanzate. Le istanze EC2 della serie N, ad esempio M5n e M5dn, sfruttano la quarta generazione di schede Nitro e dispositivi personalizzati Elastic Network Adapter (ENA) per offrire fino a 100 Gbps di throughput di rete a una singola istanza. Queste istanze offrono 4 volte la larghezza di banda di rete e il processo di pacchetti rispetto alle istanze M5 di base, e sono ideali per le applicazioni che fanno un uso intensivo della rete. I clienti possono anche abilitare Elastic Fabric Adapter (EFA) sulle istanze M5n e M5dn di determinate dimensioni per ottenere una latenza di rete bassa e uniforme.

Gli Amazon Elastic Network Adapter (ENA) offrono un'ulteriore ottimizzazione, con 20 Gbps di capacità di rete per le tue istanze all'interno di un unico gruppo di collocazione. 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 internodi su vasta scala su AWS. Con EFA, le applicazioni High Performance Computing (HPC) che utilizzano le applicazioni Message Passing Interface (MPI) e le applicazioni Machine Learning (ML) che utilizzano NVIDIA Collective Communications Library (NCCL) possono ridimensionare le risorse fino a migliaia di CPU o GPU.

Le istanze ottimizzate per Amazon EBS utilizzano uno stack di configurazione ottimizzato e forniscono un'ulteriore capacità dedicata ad Amazon EBS I/O. Questa ottimizzazione fornisce le prestazioni migliori ai tuoi volumi EBS, riducendo al minimo i conflitti tra Amazon EBS I/O e un traffico diverso dalla tua istanza.

L'instradamento basato sulla latenza (Latency-based routing o LBR) per Amazon Route 53 ti aiuta a migliorare le prestazioni della tua applicazione per un pubblico globale. La funzione LBR prevede l'instradamento dei tuoi clienti all'endpoint AWS (per istanze EC2, indirizzi IP elastici o sistemi di bilanciamento del carico ELB) che offre l'esperienza più rapida in base alle misurazioni delle prestazioni effettive delle varie regioni AWS in cui la tua applicazione viene eseguita.

Gli endpoint VPC di Amazon offrono una connettività affidabile ai servizi AWS (ad esempio Amazon S3) senza la necessità di un Internet gateway o di un'istanza Network Address Translation (NAT).

**Scegli una connettività dedicata o VPN di dimensioni adeguate per carichi di lavoro ibridi:** in presenza di un requisito di comunicazione in locale, assicurati di disporre di una larghezza di banda adeguata per le prestazioni del carico di lavoro. In base ai requisiti di larghezza di banda, una singola connessione dedicata o una singola VPN potrebbe non essere sufficiente, richiedendo pertanto l'abilitazione del bilanciamento del carico del traffico su più connessioni.

È necessario stimare la larghezza di banda e i requisiti di latenza per il carico di lavoro ibrido. Sulla base di questi numeri potrai stabilire i requisiti di dimensionamento per AWS Direct Connect o gli endpoint VPN.

AWS Direct Connect offre una connettività dedicata all'ambiente AWS, da 50 Mbps fino a 10 Gbps. In questo modo potrai disporre di una latenza gestita e controllata, nonché di una larghezza di banda assegnata, in modo che le applicazioni siano in grado di connettersi facilmente ad altri ambienti con prestazioni ottimali. Affidandoti a uno dei partner di AWS Direct Connect, ottieni una connettività end-to-end da più ambienti, per una rete estesa con prestazioni costanti.

AWS Site-to-Site VPN è un servizio VPN gestito per VPC. Quando viene creata una connessione VPN, AWS fornisce tunnel a due endpoint VPN diversi. Con AWS Transit Gateway, è possibile semplificare la connettività tra più VPC e collegarsi a qualsiasi VPC collegato ad AWS Transit Gateway con una singola connessione VPN. AWS Transit Gateway consente inoltre di ridimensionare le risorse oltre il limite di throughput VPN IPsec di 1,25 Gbps abilitando il supporto di routing ECMP (equal cost multi-path) su più tunnel VPN.

**Sfruttare il bilanciamento del carico e l'offloading della crittografia:** distribuisci il traffico tra più risorse o servizi per consentire al carico di lavoro di sfruttare l'elasticità offerta dal cloud.

Puoi anche utilizzare il bilanciamento del carico per la terminazione dell'offloading della crittografia al fine di migliorare le prestazioni, gestire e instradare il traffico in modo efficiente.

Quando implementi un'architettura scalabile in cui vuoi usare più istanze per i contenuti del servizio, puoi sfruttare i sistemi di bilanciamento del carico all'interno di Amazon VPC. AWS offre diversi modelli per le tue applicazioni nel servizio ELB. Application Load Balancer è l'ideale per il bilanciamento del carico del traffico HTTP e HTTPS: 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 sistema di bilanciamento del carico e di sollevare il carico di lavoro dall'utilizzo intensivo della CPU.

**Scegli i protocolli di rete per ottimizzare il traffico di rete:** prendi decisioni sui protocolli per la comunicazione tra sistemi e reti in base all'impatto sulle prestazioni del carico di lavoro.

Esiste una relazione tra latenza e larghezza di banda per ottenere il throughput desiderato. Se il trasferimento di file utilizza TCP, latenze più elevate ridurranno il throughput complessivo. Alcuni approcci risolvono questo problema con l'ottimizzazione TCP e i protocolli di trasferimento ottimizzati, altri adottano UDP.

**Scegli la posizione in base ai requisiti di rete:** utilizza le opzioni di posizione nel cloud disponibili per ridurre la latenza di rete o migliorare il throughput. Utilizza regioni AWS, zone di disponibilità, gruppi di collocazione e edge location come Outposts, Local Zones e Wavelength per ridurre la latenza di rete o migliorare il throughput.

L'infrastruttura di AWS Cloud è basata su regioni e zone di disponibilità. Una regione è un'area fisica del mondo in cui si trovano diverse zone di disponibilità.

Le zone di disponibilità sono composte da uno o più data center singoli provvisti di alimentazione, rete e connettività ridondanti, ognuno ubicato in una struttura separata. Le zone di disponibilità consentono di eseguire applicazioni e database in ambienti di produzione con disponibilità, tolleranza agli errori e scalabilità altrimenti impossibili da ottenere all'interno di un singolo data center.

Scegli la regione o le regioni appropriate per la tua distribuzione in base ad alcuni elementi chiave:

- **Ubicazione degli utenti**: scegliere una regione vicina agli utenti del tuo carico di lavoro garantisce una latenza minore durante il suo utilizzo.
- **Ubicazione dei dati**: per le applicazioni con elevati carichi di dati, il collo di bottiglia principale in termini di latenza è il trasferimento dei dati. Il codice dell'applicazione deve essere eseguito il più vicino possibile ai dati.

- **Altri vincoli**: prendi in considerazione limiti come sicurezza e conformità.

Amazon EC2 offre gruppi di collocazione per le reti. Un gruppo di collocazione è un raggruppamento logico delle istanze all'interno di una singola zona di disponibilità. L'utilizzo di gruppi di collocazione con tipi di istanza supportati e un Elastic Network Adapter (ENA) consente ai carichi di lavoro di partecipare a una rete a 25 Gbps a bassa latenza. I gruppi di collocazione sono consigliati per i carichi di lavoro che traggono beneficio da reti a bassa latenza, throughput di rete elevato o entrambi. Utilizzare i gruppi di collocazione consente di ridurre il jitter nelle comunicazioni di rete.

I servizi sensibili alla latenza vengono distribuiti all'edge tramite una rete globale di edge location. Tali edge location forniscono solitamente servizi come CDN (Content Delivery Network) e DNS (Domain Name System). Fornendo questi servizi nell'edge, possono rispondere con una latenza ridotta alle richieste di contenuti o risoluzione DNS. Inoltre, possono offrire servizi geografici come la geotargetizzazione dei contenuti (ossia fornire contenuti diversi in base alla posizione dell'utente finale) o l'instradamento basato sulla latenza, per indirizzare gli utenti alla regione più vicina (latenza minima).

Amazon CloudFront è una CDN globale che consente di accelerare i contenuti statici come le immagini, gli script e i video, nonché quelli dinamici come API o applicazioni web. Si basa su una rete globale di edge location che memorizzano in cache i contenuti e offrono una connettività di rete ad alte prestazioni agli utenti. Inoltre, CloudFront accelera diverse altre funzioni, come il caricamento dei contenuti e le applicazioni dinamiche. Ciò garantisce prestazioni migliori su tutte le applicazioni che gestiscono il traffico su Internet. Lambda@Edge è una funzionalità di Amazon CloudFront che consente di eseguire il codice più vicino agli utenti del carico di lavoro, migliorando le prestazioni e riducendo la latenza.

Amazon Route 53 è un servizio web di DNS cloud altamente scalabile e disponibile. È concepito per fornire a sviluppatori e aziende un modo estremamente affidabile ed economicamente vantaggioso per instradare gli utenti finali verso le applicazioni Internet, trasformando i nomi, come www.esempio.com, in indirizzi IP numerici, come 192.168.2.1, che i computer impiegano per collegarsi tra loro. Route 53 è completamente conforme con IPv6.

AWS Outposts è progettato per carichi di lavoro che, a causa dei requisiti di latenza, devono rimanere in locale, dove desideri che vengano eseguiti in modo ottimale con il resto degli altri carichi di lavoro in AWS. AWS Outposts sono rack di elaborazione e storage completamente gestiti e configurabili, creati con hardware progettato da AWS che consente di eseguire elaborazione e storage in locale, collegandosi senza soluzione di continuità all'ampia gamma di servizi AWS nel cloud.

Le zone locali di AWS sono un nuovo tipo di infrastruttura AWS progettata per eseguire carichi di lavoro che richiedono una latenza di pochi millisecondi, come il rendering di video e le applicazioni desktop virtuali con uso intensivo di grafica. Le zone locali consentono di sfruttare tutti i vantaggi derivanti dalla disponibilità di risorse di calcolo e storage più vicine agli utenti finali.

AWS Wavelength è progettato per fornire applicazioni a latenza estremamente bassa a dispositivi 5G estendendo infrastruttura, servizi, API e strumenti AWS alle reti 5G. Wavelength incorpora storage e calcolo all'interno delle reti 5G dei fornitori di telecomunicazioni a supporto del carico di lavoro 5G se richiede una latenza di pochi millisecondi, come dispositivi IoT, streaming di giochi, veicoli autonomi e produzione di contenuti multimediali in tempo reale.

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

**Ottimizza la configurazione di rete in base ai parametri:** utilizza i dati raccolti e analizzati per prendere decisioni informate sull'ottimizzazione della configurazione di rete. Misura l'impatto di tali cambiamenti e usa le misurazioni per prendere decisioni future.

Abilita i log di flusso VPC per tutte le reti VPC utilizzate dal carico di lavoro. I log di flusso VPC sono una funzione che ti permette di acquisire le informazioni sul traffico IP da e per le interfacce di rete nel tuo VPC. I log di flusso VPC possono aiutare a svolgere una serie di compiti; ad esempio, permettono di scoprire perché uno specifico traffico non raggiunge un'istanza, il che, a sua volta, consente di diagnosticare le regole di sicurezza del gruppo troppo restrittive. Puoi utilizzare i log di flusso come strumento di sicurezza per monitorare il traffico che raggiunge l'istanza, per profilare il traffico di rete e per cercare comportamenti di traffico anomali.

Utilizza i parametri di rete per apportare modifiche alla configurazione di rete a mano a mano che il carico di lavoro si evolve. Le reti basate sul cloud possono essere ricostruite rapidamente, perciò, per mantenere l'efficienza delle prestazioni, l'architettura di rete deve evolvere nel tempo.

## Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice di AWS in merito alle reti.

**Video**

- Connettività ad AWS e architetture di rete AWS ibride (NET317-R1)
- Ottimizzare le prestazioni di rete per le istanze Amazon EC2 (CMP308-R1)

**Documentazione**

- Transizione all'instradamento basato sulla latenza in Amazon Route 53
- Prodotti di rete con AWS
- EC2
    - Istanze ottimizzate per Amazon EBS
    - Reti avanzate EC2 su Linux

  - Reti avanzate EC2 su Windows
  - Gruppi di collocazione EC2
  - Abilitare reti avanzate con Elastic Network Adapter (ENA) sulle istanze Linux
- VPC
  - Transit Gateway
  - Endpoint VPC
  - Log di flusso VPC
- Elastic Load Balancer
  - Application Load Balancer
  - Network Load Balancer

# Revisione

Quando si progettano dei carichi di lavoro, le opzioni tra cui scegliere sono limitate. Tuttavia, nel tempo diventano disponibili nuove tecnologie e nuovi approcci che potrebbero migliorare le prestazioni. In AWS Cloud, sperimentare le nuove funzioni e i nuovi servizi è molto più semplice, dal momento che l'infrastruttura è costituita unicamente da codice.

Per adottare un approccio all'architettura basato sui dati, devi implementare un processo di revisione delle prestazioni che prenda in considerazione quanto segue:

- **Infrastruttura come codice:** definisci la tua infrastruttura come codice 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 distribuzione:** usa una pipeline di integrazione continua/distribuzione continua (CI/CD) (ad esempio repository del codice sorgente, sistemi di sviluppo, distribuzione e automazione dei test) per distribuire la tua infrastruttura. Ciò ti consente di effettuare la distribuzione in modo ripetibile, coerente ed economicamente vantaggioso nel corso dell'iterazione.
- **Parametri ben definiti:** configura i parametri e il monitoraggio per raccogliere gli indicatori chiave di prestazione (KPI). Ti consigliamo di adottare parametri tecnici e aziendali. Per i siti web o le app mobili, i parametri principali sono il tempo di acquisizione al primo byte o il rendering. Gli altri parametri generalmente validi includono il numero di thread, il tasso di raccolta di dati superflui e gli stati di attesa. I parametri aziendali, come il costo cumulativo aggregato per richiesta, possono indicarti due modi per ridurre i costi. Valuta attentamente il modo in cui prevedi di interpretare i parametri. Ad esempio, potresti scegliere il 99° percentile o quello massimo anziché il valore medio.
- **Automatizza i test delle prestazioni:** nell'ambito del processo di distribuzione, attiva 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 o SaaS (Software-as-a-Service) per generare il carico. Valuta se utilizzare le soluzioni AWS Marketplace 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.

Questo processo di revisione delle prestazioni può essere implementato come una semplice estensione della tua pipeline di distribuzione esistente. Inoltre, puoi farlo evolvere nel tempo, all'aumentare della complessità dei requisiti di test. Per le architetture future, puoi generalizzare il tuo approccio e riutilizzare lo stesso processo e gli stessi artefatti.

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

## Sviluppa il tuo carico di lavoro per trarre vantaggio dalle nuove versioni

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

**Rimani aggiornato su nuove risorse e servizi:** valuta i modi per migliorare le prestazioni man mano che nuovi servizi, modelli di progettazione e offerte di prodotti diventano disponibili. Studia come le novità potrebbero migliorare le prestazioni o aumentare l'efficienza del carico di lavoro tramite una valutazione ad hoc, una discussione interna o un'analisi esterna.

Stabilisci un processo per valutare gli aggiornamenti, le nuove funzioni e i servizi di AWS. Ad esempio, crea proof of concept che utilizzano le nuove tecnologie o consultati con un gruppo interno. Quando metti alla prova nuove idee o servizi, esegui test delle prestazioni per misurare l'impatto sull'efficienza o sulle prestazioni del carico di lavoro. Sfrutta la

flessibilità offerta da AWS per condurre test frequenti su nuove idee e tecnologie con costi o rischi minimi.

**Definisci un processo per migliorare le prestazioni del carico di lavoro:** definisci un processo per valutare i nuovi servizi, i modelli di progettazione, i tipi di risorse e le configurazioni man mano che diventano disponibili. Ad esempio, esegui test delle prestazioni esistenti sulle nuove offerte di istanze per determinare il loro potenziale per migliorare il carico di lavoro.

Le prestazioni del carico di lavoro presentano alcuni vincoli chiave. Documentali, in modo da sapere quali tipi di innovazione potrebbero migliorare le prestazioni del carico di lavoro. Utilizza queste informazioni quando vieni a conoscenza di nuovi servizi e tecnologie, man mano che si rendono disponibili, in modo da identificare le soluzioni per ovviare ai vincoli o ai colli di bottiglia.

**Sostieni l'evoluzione delle prestazioni dei carichi di lavoro nel tempo:** come organizzazione, utilizza le informazioni raccolte durante il processo di valutazione per gestire attivamente l'adozione di nuovi servizi o risorse quando diventano disponibili.

Utilizza le informazioni ottenute con la valutazione di nuovi servizi o tecnologie per favorire il cambiamento. Man mano che la tua azienda o il tuo carico di lavoro evolve, anche le prestazioni devono cambiare. Sfrutta i dati raccolti dai parametri riguardanti il carico di lavoro per valutare le aree in cui è possibile ottenere i miglioramenti più significativi in termini di efficienza o prestazioni e adotta in modo proattivo nuovi servizi e tecnologie per tenere il passo con la domanda.

## Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice di AWS in merito ai benchmark.

**Video**

- Canale YouTube di Amazon Web Services
- Canale YouTube dei Tech talk online di AWS
- Canale YouTube degli eventi AWS

# Monitoraggio

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

In AWS, il monitoraggio è composto da cinque fasi distinte, spiegate in maggiore dettaglio nel whitepaper dedicato al Principio dell'affidabilità:

1. **Generazione** - ambito di monitoraggio, parametri e soglie
2. **Aggregazione** - creazione di una visualizzazione completa da più origini
3. **Elaborazione e allarmi in tempo reale** - riconoscimento e risposta
4. **Storage** - policy di gestione e conservazione dei dati
5. **Analisi** - pannelli di controllo, reportistica e approfondimenti

CloudWatch è un servizio di monitoraggio per le risorse di AWS Cloud e i carichi di lavoro in esecuzione su AWS. Puoi usare CloudWatch per raccogliere e monitorare parametri e file di log, nonché per impostare allarmi. CloudWatch consente il monitoraggio di risorse AWS quali le istanze Amazon EC2 e le istanze database di Amazon RDS, nonché dei parametri personalizzati generati dai tuoi carichi di lavoro e dai tuoi servizi e dei file di log generati dalle tue applicazioni. CloudWatch può essere impiegato per ottenere visibilità a livello di sistema su utilizzo delle risorse, prestazioni delle applicazioni e stato di integrità operativa. Puoi utilizzare le informazioni ottenute per correggere rapidamente il funzionamento e mantenere le prestazioni del carico di lavoro sempre ottimali.

I pannelli di controllo di CloudWatch ti consentono di creare grafici riutilizzabili delle risorse di AWS e parametri personalizzati, in modo da monitorare lo stato operativo e individuare i problemi a colpo d'occhio.

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

Le soluzioni di monitoraggio sono di due tipi: monitoraggio attivo e monitoraggio passivo. Le soluzioni di monitoraggio attivo e passivo si integrano a vicenda, per offrirti una visione completa delle prestazioni del tuo carico di lavoro.

Il **monitoraggio attivo** simula l'attività degli utenti in percorsi utente soggetti a script nelle posizioni critiche del prodotto. Occorre eseguire il monitoraggio attivo in modo continuo, per testare le prestazioni e la disponibilità di un carico di lavoro. Il monitoraggio attivo integra quello passivo in quanto continuo, leggero e prevedibile. È eseguibile in tutti gli ambienti,

in particolare quelli di pre-produzione, per individuare i problemi o le prestazioni insufficienti prima che interessino gli utenti finali.

Di solito, il **monitoraggio passivo** viene utilizzato con i carichi di lavoro basati su web. Il monitoraggio passivo raccoglie i parametri delle prestazioni dal browser; i carichi di lavoro non basati sul web possono adottare un approccio simile. Puoi raccogliere i parametri di tutti gli utenti o di un loro sottoinsieme, le aree geografiche, i browser e i tipi di dispositivo. Puoi utilizzare il monitoraggio passivo per comprendere i problemi seguenti:

- **Prestazioni relative all'esperienza utente**: il monitoraggio passivo offre parametri relativi a ciò che viene sperimentato dagli utenti, per avere una visione continua dello stato di funzionamento della produzione e dell'impatto delle modifiche nel corso del tempo.
- **Variabilità delle prestazioni a livello geografico**: se un carico di lavoro ha un impatto a livello globale e gli utenti accedono all'applicazione da tutto il mondo, l'uso del monitoraggio passivo può consentirti di individuare i problemi di prestazioni che interessano gli utenti di una data area.
- **Impatto dell'utilizzo delle API**: i carichi di lavoro moderni utilizzano API interne e API di terze parti. Il monitoraggio passivo consente di osservare l'utilizzo delle API in modo da individuare i colli di bottiglia delle prestazioni, derivanti non solo dalle API interne ma anche dai fornitori di API di terzi.

CloudWatch offre la possibilità di monitorare e inviare allarmi sotto forma di notifica. Puoi usare l'automazione per correggere i problemi di prestazioni tramite l'attivazione di azioni mediante Amazon Kinesis, Amazon Simple Queue Service (Amazon SQS) e AWS Lambda.

## Monitora le tue risorse per assicurarti che abbiano le prestazioni previste

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

**Registra i parametri correlati alle prestazioni:** utilizza un servizio di monitoraggio e osservazione per registrare i parametri correlati alle prestazioni. Ad esempio, registra le transazioni di database, le query lente, la latenza I/O, il throughput delle richieste HTTP, la latenza del servizio o altri dati chiave.

Identifica i parametri relativi alle prestazioni rilevanti per il tuo carico di lavoro e registrali. Questi dati sono importanti per riuscire a identificare quali componenti influiscono sulle prestazioni o sull'efficienza complessive del carico di lavoro.

Partendo dall'esperienza del cliente, identifica quali sono i parametri rilevanti. Per ciascuno di essi, identifica l'obiettivo, l'approccio di misurazione e la priorità. Utilizza questi elementi per creare allarmi e notifiche per affrontare in modo proattivo i problemi correlati alle prestazioni.

**Analizza i parametri quando si verificano eventi o incidenti:** in risposta a nel corso di un evento o un incidente, utilizza pannelli di controllo o report di monitoraggio per comprendere e diagnosticare l'impatto. Queste viste forniscono informazioni sulle parti del carico di lavoro le cui prestazioni non raggiungono i livelli previsti.

Durante lo studio dei casi utente critici per la tua architettura, includi i requisiti relativi alle prestazioni, specificando ad esempio con quale rapidità deve essere eseguito ogni scenario. Per questi scenari critici, implementa percorsi utente con script aggiuntivi per chiarire esattamente quali sono le loro prestazioni rispetto ai requisiti.

**Stabilisci indicatori chiave di prestazione (KPI) per misurare le prestazioni del carico di lavoro:** identifica i KPI che indicano se le prestazioni del carico di lavoro sono quelle previste. Un carico di lavoro basato su API, ad esempio, può utilizzare la latenza di risposta complessiva come indicazione delle prestazioni complessive, mentre per un sito di e-commerce un KPI valido può essere il numero di acquisti andati a buon fine.

Documenta l'esperienza prestazionale richiesta dai clienti, incluso il modo in cui i clienti valutano le prestazioni del carico di lavoro. In base a questi requisiti, stabilisci i tuoi KPI chiave, che indicheranno quali sono le prestazioni complessive del sistema.

**Utilizza il monitoraggio per generare notifiche basate su allarme:** avvalendoti degli indicatori chiave di prestazione (KPI) relativi alle prestazioni che hai identificato, utilizza un sistema di monitoraggio che genera automaticamente allarmi quando queste misurazioni sono al di fuori dei limiti previsti.

Amazon CloudWatch può raccogliere i parametri per tutte le risorse dell'architettura. Puoi anche raccogliere e pubblicare parametri personalizzati per ottenere parametri aziendali o derivati. Utilizza CloudWatch o un servizio di monitoraggio di terze parti per configurare degli allarmi che si attivino al superamento delle soglie impostate; gli allarmi segnalano che un parametro si trova al di fuori dei limiti previsti.

**Esamina i parametri a intervalli regolari:** come manutenzione ordinaria o in risposta a eventi o incidenti specifici, esamina quali parametri vengono raccolti. Stabilisci quali di questi parametri sono fondamentali per risolvere i problemi e quali altri parametri aggiuntivi, se monitorati, potrebbero contribuire a identificare, affrontare o prevenire i problemi.

Nell'ambito della risposta a incidenti ed eventi, valuta quali parametri sono stati utili per risolvere il problema e quali sarebbero stati utili ma non sono attualmente monitorati. Queste considerazioni ti aiuteranno a migliorare la qualità dei parametri raccolti, per prevenire o risolvere più rapidamente gli incidenti futuri.

**Monitora e invia allarmi in modo proattivo:** utilizza indicatori chiave di prestazioni (KPI), in combinazione con sistemi di monitoraggio e allarmi, per risolvere in modo proattivo i problemi correlati alle prestazioni. Laddove possibile, utilizza gli allarmi per attivare operazioni automatizzate per risolvere i problemi. Se non è possibile rispondere in modo automatizzato, inoltra l'allarme a coloro che possono intervenire. Ad esempio, puoi implementare un sistema in grado di prevedere i valori attesi per i KPI e di inviare allarmi qualora essi oltrepassino determinate soglie, oppure uno strumento che arresta o esegue automaticamente il rollback delle distribuzioni nel caso in cui i valori dei KPI si discostino dai valori attesi.

Implementa dei processi che forniscono visibilità sulle prestazioni durante l'esecuzione del carico di lavoro. Crea pannelli di controllo del monitoraggio e stabilisci norme di riferimento per le aspettative riguardanti le prestazioni, per determinare se il carico di lavoro ha prestazioni ottimali.

## Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice di AWS per il monitoraggio e la promozione dell'efficienza delle prestazioni.

**Video**

- Elimina il caos: acquisisci visibilità e approfondimenti operativi (MGT301-R1)

**Documentazione**

- Documentazione di X-Ray
- Documentazione di CloudWatch

# Compromessi

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

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

AWS offre soluzioni di memorizzazione nella cache come Amazon ElastiCache, che offre un datastore o una cache in memoria, e Amazon CloudFront, che memorizza nella cache le copie dei tuoi contenuti statici in prossimità degli utenti finali. Amazon DynamoDB Accelerator (DAX) offre un livello di memorizzazione nella cache distribuito con read-through/write-through su Amazon DynamoBD, supportando la stessa API ma con una latenza di pochi millisecondi per le entità che si trovano nella cache.

## Accettare compromessi per migliorare le prestazioni

Quando si progettano soluzioni, prendere in considerazione attivamente i compromessi ti consente di selezionare un approccio ottimale. Spesso è possibile migliorare le prestazioni accettando compromessi in termini di coerenza, durata e spazio a favori di tempo e latenza. I compromessi possono aumentare la complessità della tua architettura e richiedere test di carico per garantire vantaggi misurabili.

**Comprendi le aree in cui le prestazioni sono più importanti:** comprendi e identifica le aree in cui l'aumento delle prestazioni del carico di lavoro determinerà un impatto positivo sull'efficienza o sull'esperienza del cliente. Ad esempio, un sito web che ha una grande quantità di interazione con i clienti può trarre vantaggio dall'utilizzo dei servizi edge per spostare la distribuzione di contenuti più vicino ai clienti.

**Studia i modelli e i servizi di progettazione:** ricerca e analizza i vari servizi e modelli di progettazione che permettono di migliorare le prestazioni del carico di lavoro. Nell'ambito dell'analisi, identifica gli elementi sui quali potresti accettare compromessi per ottenere prestazioni più elevate. Ad esempio, l'utilizzo di un servizio di cache può contribuire a ridurre il carico sui sistemi di database, tuttavia richiede una certa quantità di progettazione per l'implementazione di cache sicure o l'eventuale introduzione di consistenza finale in alcune aree.

Scopri quali sono le opzioni di configurazione relative alla rete disponibili e come possono influire sul carico di lavoro. L'ottimizzazione delle prestazioni del tuo carico di lavoro dipende dalla comprensione della modalità in cui tali opzioni interagiscono con la tua architettura e dell'impatto che avranno sia sulle prestazioni misurate sia sulle prestazioni percepite dagli utenti.

La Amazon Builders' Library fornisce ai lettori una descrizione dettagliata di come Amazon crea e gestisce la tecnologia. Questi articoli gratuiti sono scritti dagli ingegneri senior di Amazon e coprono una varietà di argomenti, tra cui architettura, distribuzione di software e operazioni. Ad esempio, puoi scoprire in che modo Amazon automatizza la distribuzione di software per raggiungere oltre 150 milioni di distribuzioni l'anno, oppure in che modo gli ingegneri di Amazon impiegano principi come lo shuffle sharding per creare sistemi resilienti con disponibilità e tolleranza ai guasti elevate.

**Analizza l'impatto dei compromessi sui clienti e sull'efficienza:** quando valuti i miglioramenti correlati alle prestazioni, determina quali scelte avranno un impatto sui clienti e sull'efficienza del carico di lavoro. Ad esempio, se l'utilizzo di un datastore chiave-valore aumenta le prestazioni del sistema, è importante valutare in che modo la consistenza della sua natura finale influirà sui clienti.

Attraverso i parametri e il monitoraggio, identifica le aree del sistema in cui le prestazioni sono scarse. Stabilisci in che modo puoi apportare miglioramenti e quali compromessi comportano, oltre al loro impatto sul sistema e sull'esperienza degli utenti. L'implementazione di cache di dati, ad esempio, può contribuire a migliorare notevolmente le prestazioni ma richiede una strategia ben definita sulle modalità e sui tempi di aggiornamento o di invalidamento dei dati che vi sono contenuti, per evitare che il sistema si comporti in modo non corretto.

**Misura l'impatto dei miglioramenti delle prestazioni:** quando vengono apportate modifiche per migliorare le prestazioni, valuta i parametri e i dati raccolti. Utilizza queste informazioni per determinare l'impatto che il miglioramento delle prestazioni ha avuto sul carico di lavoro, sui suoi componenti e sui clienti. Queste misurazioni permettono di capire quali sono i miglioramenti ottenuti dai compromessi applicati e aiutano a stabilire se si sono verificati eventuali effetti collaterali negativi.

Un sistema Well-Architected si basa su una combinazione di varie strategie riguardanti le prestazioni. Identifica quale strategia determinerà il maggiore impatto positivo su un determinato hotspot o collo di bottiglia. Lo sharding dei dati su più sistemi di database relazionali, ad esempio, può migliorare il throughput complessivo pur continuando a supportare le transazioni e, all'interno di ogni shard, la memorizzazione nella cache può contribuire a ridurre il carico.

**Utilizza un mix di strategie correlate alle prestazioni:** se possibile, utilizza più strategie per migliorare le prestazioni. Scegli, ad esempio, strategie come la memorizzazione dei dati nella cache per evitare eccessive chiamate di rete o dei database, l'utilizzo di repliche di lettura per i motori di database al fine di migliorare i tassi di lettura, lo sharding o la compressione dei dati, ove possibile, per ridurre i volumi, e il buffering e lo streaming dei risultati man mano che diventano disponibili per evitare blocchi.

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

### Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice di AWS in merito alla memorizzazione nella cache.

**Video**

- Presentazione di Amazon Builders' Library (DOP328)

**Documentazione**

- Amazon Builders' Library
- Best practice per l'implementazione di Amazon ElastiCache

# 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 permetterà di fare evolvere l'architettura in modo rapido e sicuro, mentre potrai utilizzare i dati per prendere decisioni circostanziate in merito all'architettura stessa. Adoperare una combinazione di monitoraggio attivo e passivo ti aiuterà a mantenere costanti le prestazioni dell'architettura nel corso del tempo.

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

# Collaboratori

Hanno contribuito alla stesura di questo documento:

- Eric Pullen, Performance Efficiency Lead Well-Architected, Amazon Web Services
- Philip Fitzsimons, Sr Manager Well-Architected, 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:

- Canone di architettura AWS

# Revisioni del documento

| Data | Descrizione |
|---|---|
| **Aprile 2020** | Aggiornamento principale a v2 |
| **Luglio 2018** | Aggiornamento minore per la correzione di problemi grammaticali |
| **Novembre 2017** | Aggiornamento del whitepaper per rispecchiare le modifiche apportate a AWS |
| **Novembre 2016** | Prima pubblicazione |