Il principio dell'ottimizzazione dei costi

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

Introduzione 1

Ottimizzazione dei costi 2

- Principi di progettazione 2
- Definizione 2

Esercizio della gestione finanziaria del cloud 3

- Proprietà funzionale 4
- Collaborazione tra finanza e tecnologia 4
- Budget e previsioni per il cloud 6
- Processi consapevoli dei costi 6
- Cultura consapevole dei costi 7
- Quantificare il valore aggiunto realizzato attraverso l'ottimizzazione dei costi 8

Consapevolezza delle spese e dell'utilizzo 9

- Governance 10
- Monitora i costi e l'utilizzo 13
- Disattiva le risorse 15

Convenienza delle risorse 17

- Valuta i costi al momento di selezionare i servizi 17
- Seleziona il tipo, le dimensioni e il numero di risorse in modo corretto 20
- Seleziona il modello di prezzo migliore 21
- Piano per il trasferimento dei dati 26

Gestione delle risorse di domanda e offerta 28

- Gestisci la domanda 29
- Fornitura dinamica 29

Apporta ottimizzazioni nel corso del tempo 31

- Valuta e implementa nuovi servizi 31

Conclusioni 33

Collaboratori 33

Approfondimenti 34

Revisioni del documento 34

Riassunto

Questo whitepaper tratta del principio dell'ottimizzazione dei costi 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.

Un carico di lavoro basato sull'ottimizzazione dei costi utilizzerà appieno tutte le risorse, raggiungerà un risultato al minor prezzo possibile e soddisferà i tuoi requisiti funzionali. Questo whitepaper fornisce linee guida approfondite per creare capacità all'interno dell'organizzazione, progettare il carico di lavoro, selezionare, configurare e gestire i servizi e applicare tecniche di ottimizzazione dei costi.

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

Questo documento si concentra sul principio dell'ottimizzazione dei costi e su come progettare carichi di lavoro usando servizi e risorse con l'utilizzo più efficace di servizi e risorse, per ottenere risultati aziendali al prezzo più basso.

Scoprirai come applicare alle tue soluzioni le best practice del principio dell'ottimizzazione dei costi. L'ottimizzazione dei costi può rivelarsi difficoltosa nelle soluzioni in locale tradizionali, poiché occorre prevedere le capacità e le esigenze aziendali future, portando avanti al tempo stesso complesse procedure di approvvigionamento. L'adozione delle prassi descritte in questo documento consentirà alla tua organizzazione di raggiungere i seguenti obiettivi:

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

Questo documento è rivolto a chi svolge ruoli tecnologici, ad esempio ai Chief Technology Officer (CTO), ai Chief Financial Officer (CFO), ai progettisti, agli sviluppatori, ai controller finanziari, ai pianificatori finanziari, agli analisti aziendali e ai membri del team operativo.
Il presente documento non fornisce dettagli sull'implementazione o sui modelli architetturali; tuttavia, include riferimenti alle risorse in cui trovare tali informazioni.

Ottimizzazione dei costi

L'ottimizzazione dei costi è un processo continuo di perfezionamento e miglioramento di un carico di lavoro durante tutto il suo ciclo di vita. Le pratiche di questo documento consentono di creare e gestire carichi di lavoro consapevoli dei costi, che consentono di ottenere risultati aziendali riducendo al minimo i costi e permettendo alla tua organizzazione di massimizzare il ritorno sugli investimenti.

Principi di progettazione

Considera i seguenti principi di progettazione per l'ottimizzazione dei costi:

Implementa la gestione finanziaria del cloud: per ottenere successo finanziario e accelerare la realizzazione del valore aggiunto nel cloud, è necessario investire nella gestione finanziaria del cloud. La tua organizzazione deve dedicare il tempo e le risorse necessarie per creare le capacità in questo nuovo dominio di gestione della tecnologia e dell'utilizzo. Analogamente alle tue capacità di sicurezza o operatività, devi creare capacità tramite lo sviluppo di competenze, programmi, risorse e processi, per diventare un'organizzazione efficiente in termini di costi.

Adotta un modello a consumo: paga solo le risorse di calcolo che utilizzi e incrementa o riduci l'utilizzo a seconda delle necessità aziendali. Ad esempio, gli ambienti di test e di sviluppo sono generalmente usati solo per otto ore al giorno durante la settimana lavorativa. Puoi sospendere queste risorse quando non le utilizzi, risparmiando potenzialmente il 75% dei costi (40 ore anziché 168).

Misura l'efficienza complessiva: misura il risultato aziendale del carico di lavoro e i costi associati alla sua fornitura. Utilizza questi dati per comprendere i vantaggi ottenuti dall'incremento dell'output, dall'aumento della funzionalità e dalla riduzione dei costi.

Smetti di spendere denaro per attività onerose e indifferenziate: AWS si occupa delle attività onerose dei data center come il racking, lo stacking e l'alimentazione dei server. Inoltre, elimina l'onere operativo della gestione di sistemi operativi e applicazioni con servizi gestiti. In questo modo, potrai concentrarti sui tuoi clienti e sui progetti aziendali anziché sull'infrastruttura IT.

Analizza e attribuisci le spese: il cloud ti aiuta a individuare con facilità e precisione l'utilizzo e i costi dei carichi di lavoro, permettendoti quindi di attribuire in modo trasparente i costi IT ai flussi di ricavi e ai singoli proprietari dei carichi di lavoro. Questo ti aiuta a misurare il ritorno sull'investimento (ROI) e offre ai proprietari del carico di lavoro la possibilità di ottimizzare le proprie risorse e ridurre i costi.

Definizione

Sono cinque le aree di interesse per l'ottimizzazione dei costi nel cloud:

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

Come per gli altri principi di base all'interno del canone di architettura, occorre considerare alcuni compromessi per ottimizzare i costi. Ad esempio, occorre valutare se ottimizzare la velocità di commercializzazione o i costi. In alcuni casi, è meglio ottimizzare la velocità di commercializzazione (arrivare sul mercato rapidamente, distribuire nuove funzionalità o semplicemente rispettare una scadenza) anziché investire nell'ottimizzazione anticipata dei costi.

Talvolta le decisioni di progettazione sono guidate dalla fretta invece che dallo studio dei dati, ed esiste sempre la tentazione di compensare in modo eccessivo anziché dedicare tempo all'esecuzione di benchmark per conseguire la distribuzione più conveniente. Questo potrebbe portare a distribuzioni con un provisioning eccessivo e sotto-ottimizzate. Tuttavia, potrebbe essere una scelta ragionevole quando è necessario trasferire le risorse dal tuo ambiente in locale al cloud ed eseguire l'ottimizzazione in un momento successivo.

Investire in anticipo la giusta quantità di energia in una strategia di ottimizzazione dei costi consente di realizzare i vantaggi economici del cloud in modo più rapido, assicurando il rispetto costante delle best practice ed evitando un over-provisioning superfluo. Le sezioni seguenti forniscono tecniche e best practice per l'implementazione iniziale e continua della gestione finanziaria del cloud e l'ottimizzazione dei costi dei carichi di lavoro.

Esercizio della gestione finanziaria del cloud

La gestione finanziaria del cloud (Cloud Financial Management o CFM) consente alle organizzazioni di conseguire un valore aggiunto e il successo finanziario ottimizzando i costi e l'utilizzo e ricalibrando le risorse in AWS.

Di seguito sono elencate le best practice per la gestione finanziaria nel cloud:

- Proprietà funzionale
- Collaborazione tra finanza e tecnologia
- Budget e previsioni per il cloud
- Processi consapevoli dei costi
- Cultura consapevole dei costi
- Quantificare il valore aggiunto realizzato attraverso l'ottimizzazione dei costi

Proprietà funzionale

Stabilisci una funzione di ottimizzazione dei costi: questa funzione è responsabile di stabilire e mantenere una cultura di consapevolezza dei costi. Può trattarsi di un individuo esistente, di un team all'interno della tua organizzazione o di un nuovo team di stakeholder chiave della finanza, della tecnologia e dell'organizzazione di tutta l'organizzazione.

La funzione (individuo o team) stabilisce le priorità e dedica la parte prevista del proprio tempo alle attività di gestione e ottimizzazione dei costi. In un'organizzazione di dimensioni ridotte, la quantità di tempo dedicata dalla funzione potrebbe essere inferiore rispetto a quella dedicata da una funzione a tempo pieno in un'azienda di dimensioni maggiori.

La funzione richiede un approccio multidisciplinare, con capacità di gestione dei progetti, data science, analisi finanziaria e sviluppo di software/infrastruttura. La funzione è in grado di migliorare l'efficienza dei carichi di lavoro eseguendo ottimizzazioni dei costi (approccio centralizzato), influenzando i team tecnologici nell'esecuzione di ottimizzazioni (approccio decentralizzato) o una combinazione di entrambi (approccio ibrido). La funzione può essere valutata in base alla sua capacità di eseguire e conseguire risultati rispetto agli obiettivi di ottimizzazione dei costi (ad esempio, parametri di efficienza del carico di lavoro).

È necessario tutelare questa funzione mediante uno sponsor a livello direttivo. Lo sponsor è considerato un sostenitore del consumo efficiente del cloud e fornisce alla funzione un supporto di escalation, per garantire che le attività di ottimizzazione dei costi vengano trattate con il livello di priorità definito dall'organizzazione. Insieme, lo sponsor e la funzione assicurano che l'organizzazione utilizzi il cloud in modo efficiente e continui a offrire valore aggiunto.

Collaborazione tra finanza e tecnologia

Stabilisci una collaborazione tra finanza e tecnologia: i team tecnologici possono innovare più rapidamente nel cloud grazie a cicli di approvazione, approvvigionamento e distribuzione dell'infrastruttura più brevi. Può trattarsi di una novità per le organizzazioni finanziarie che in precedenza erano abituate a eseguire processi dispendiosi, in termini di tempo e di risorse, per acquistare e distribuire capitale in data center e ambienti locali, allocando i costi solo in fase di approvazione del progetto.

Stabilisci una collaborazione tra i principali stakeholder finanziari e tecnologici per creare una comprensione condivisa degli obiettivi organizzativi e sviluppare meccanismi che consentano il successo finanziario nel modello di spesa variabile del cloud computing. I team pertinenti all'interno della tua organizzazione devono essere coinvolti nelle discussioni su costi e utilizzo in tutte le fasi del tuo viaggio verso il cloud; tra di essi vi sono:

- **Responsabili finanziari:** CFO, controllori finanziari, pianificatori finanziari, analisti aziendali, approvvigionamento e selezione delle risorse e contabilità fornitori devono comprendere il modello di consumo del cloud, le opzioni di acquisto e il processo di fatturazione mensile. A causa delle differenze fondamentali tra il cloud (ad esempio il tasso di variazione dell'utilizzo, i prezzi a consumo o a scaglioni, i modelli di prezzo e le informazioni dettagliate su fatturazione e utilizzo) e le operazioni in locale, è essenziale che l'organizzazione finanziaria capisca in che modo l'utilizzo del cloud può influire sugli aspetti aziendali, tra cui processi di approvvigionamento, monitoraggio degli incentivi, allocazione dei costi e bilanci.
- **Responsabili tecnologici:** i responsabili tecnologici (inclusi i proprietari di prodotti e applicazioni) devono essere a conoscenza dei requisiti finanziari (ad esempio i vincoli di budget) e dei requisiti aziendali (ad esempio i contratti sul livello di servizio). In questo modo, il carico di lavoro può essere implementato in modo opportuno per raggiungere gli obiettivi desiderati dall'azienda.

La collaborazione tra finanza e tecnologia offre i seguenti vantaggi:

- I team finanziari e tecnologici hanno una visibilità quasi in tempo reale su costi e utilizzo.
- I team finanziari e tecnologici stabiliscono una procedura operativa standard per gestire le variazioni di spesa nel cloud.
- Gli stakeholder finanziari fungono da consulenti strategici per quanto riguarda il modo in cui il capitale viene utilizzato per acquistare sconti a fronte di impegni (ad esempio, istanze riservate o AWS Savings Plans) e il modo in cui il cloud viene utilizzato per far crescere l'organizzazione.
- I processi di approvvigionamento e di contabilità esistenti vengono applicati al cloud.
- I team finanziari e tecnologici collaborano per prevedere costi e utilizzo di AWS futuri, al fine di allineare o creare i budget aziendali.
- La comunicazione all'interno dell'organizzazione migliora attraverso un linguaggio condiviso e una comprensione comune dei concetti finanziari.

Altri stakeholder all'interno della tua organizzazione che devono essere coinvolti nelle discussioni su costi e utilizzo includono:

- **Proprietari delle unità aziendali:** i proprietari delle unità aziendali devono comprendere il modello aziendale del cloud in modo da indirizzare l'operato delle unità aziendali e di tutta l'azienda. Questa conoscenza del cloud è fondamentale quando è necessario prevedere la crescita e l'utilizzo del carico di lavoro, ma anche quando si valutano le diverse opzioni di acquisto, come le istanze riservate o i Savings Plans.

- **Terze parti:** se la tua organizzazione si avvale di terze parti (ad esempio, consulenti o strumenti), assicurati che esse siano allineate ai tuoi obiettivi finanziari e possano dimostrare sia l'allineamento, tramite i loro modelli di coinvolgimento, sia il ritorno sull'investimento (ROI). In genere, le terze parti contribuiscono alla creazione di report e all'analisi di eventuali carichi di lavoro da esse gestiti, e forniscono anche l'analisi dei costi relativi ai carichi di lavoro da esse progettati.

Budget e previsioni per il cloud

Stabilisci budget e previsioni per il cloud: i clienti utilizzano il cloud per ottenere efficienza, velocità e agilità, determinando un'elevata variabilità in termini di costi e utilizzo. I costi possono diminuire in seguito all'aumento dell'efficienza del carico di lavoro o con la distribuzione di nuovi carichi di lavoro e funzionalità. In alternativa, i carichi di lavoro possono essere dimensionati per servire un maggior numero di clienti, aumentando l'utilizzo e i costi del cloud. Gli attuali processi di creazione di budget dell'organizzazione devono essere modificati per incorporare questa variabilità.

Puoi adattare i processi di creazione dei budget e previsione esistenti per renderli più dinamici utilizzando un algoritmo basato sui trend (utilizzando i costi storici come input), algoritmi basati su fattori chiavi di business (ad esempio, lanci di nuovi prodotti o espansione regionale) o una combinazione di trend e fattori chiave di business.

Puoi utilizzare AWS Cost Explorer per prevedere i costi cloud giornalieri (fino a 3 mesi) o mensili (fino a 12 mesi) in base agli algoritmi di machine learning applicati ai costi storici (basati sui trend).

Processi consapevoli dei costi

Implementa la consapevolezza dei costi nei processi organizzativi: la consapevolezza dei costi deve essere implementata in processi organizzativi nuovi ed esistenti. È consigliabile riutilizzare e modificare i processi esistenti, laddove possibile, riducendo al minimo l'impatto sull'agilità e la velocità. Le seguenti raccomandazioni aiuteranno a implementare la consapevolezza dei costi nel carico di lavoro:

- Assicurati che la gestione delle modifiche includa una misurazione dei costi per quantificare l'impatto finanziario delle modifiche. Questo aiuta a risolvere in modo proattivo le problematiche relative ai costi nonché a mettere in luce i risparmi ottenuti.
- Assicurati che l'ottimizzazione dei costi sia un componente fondamentale delle tue capacità operative. Ad esempio, puoi sfruttare gli attuali processi di gestione degli incidenti per analizzare e identificare la causa principale di anomalie di costi e utilizzo (eccedenze di costo).

- Accelera la riduzione dei costi e la realizzazione del valore aggiunto attraverso l'automazione o l'utilizzo di strumenti. Quando ponderi i costi dell'implementazione, includi nella valutazione un componente ROI per giustificare l'investimento di tempo o denaro.
- Estendi i programmi di formazione e sviluppo esistenti per includere la formazione sulla consapevolezza dei costi in tutta l'organizzazione, comprese attività di formazione continua e certificazione. In questo modo, creerai un'organizzazione in grado di gestire in modo autonomo i costi e l'utilizzo.

Fornisci report e notifiche sull'ottimizzazione dei costi e dell'utilizzo: è necessario rendicontare regolarmente l'ottimizzazione dei costi e dell'utilizzo all'interno dell'organizzazione. Puoi implementare sessioni dedicate per l'ottimizzazione dei costi o includere l'ottimizzazione dei costi nei regolari cicli di reporting operativi per i tuoi carichi di lavoro. AWS Cost Explorer fornisce pannelli di controllo e report. Puoi monitorare l'avanzamento dei costi e dell'utilizzo rispetto ai budget configurati con i report di Budget AWS.

Puoi anche utilizzare Amazon QuickSight con i dati dei report costi e utilizzo (Cost and Usage Report o CUR) per fornire report altamente personalizzati con dati più granulari.

Implementa notifiche su costi e utilizzo in modo che si possa intervenire rapidamente in caso di variazioni di costi e utilizzo. Budget AWS consente di inviare notifiche rispetto a target definiti. Ti consigliamo di configurare le notifiche sia per gli aumenti sia per le diminuzioni, sia sui costi sia sull'utilizzo relativi ai carichi di lavoro.

Monitora costi e utilizzo in modo proattivo: si consiglia di monitorare i costi e l'utilizzo all'interno dell'organizzazione in modo proattivo, e non solo in caso di eccezioni o anomalie. I pannelli di controllo con un'elevata visibilità in tutto l'ufficio o l'ambiente di lavoro garantiscono che le persone chiave abbiano accesso alle informazioni di cui hanno bisogno e dimostrano l'attenzione che l'organizzazione presta all'ottimizzazione dei costi. I pannelli di controllo visibili consentono di promuovere attivamente i risultati positivi e di implementarli in tutta l'organizzazione.

Cultura consapevole dei costi

Crea una cultura consapevole dei costi: implementa modifiche o programmi all'interno della tua organizzazione per creare una cultura consapevole dei costi. Si consiglia di iniziare in piccolo, per poi implementare programmi di grandi dimensioni e di vasta portata all'aumentare delle capacità e dell'utilizzo del cloud da parte dell'organizzazione.

Una cultura basata sui costi consente di ricalibrare l'ottimizzazione e la gestione finanziaria del cloud attraverso best practice eseguite in modo organico e decentralizzato all'interno di tutta l'organizzazione. Questo crea livelli elevati di capacità all'interno dell'organizzazione con uno sforzo minimo, qualcosa di analogo a un approccio centralizzato e dall'alto verso il basso.

Piccoli cambiamenti nella cultura possono avere un grande impatto sull'efficienza dei carichi di lavoro attuali e futuri. Esempi di questo tipo includono:

- Gamificare costi e utilizzo in tutta l'organizzazione. Questa operazione può essere eseguita tramite un pannello di controllo visibile pubblicamente o un report che confronta i costi e l'utilizzo normalizzati tra i team (ad esempio, i costi per carico di lavoro e i costi per transazione).
- Premiare l'efficienza dei costi. Ricompensa pubblicamente o privatamente i risultati di ottimizzazione dei costi volontari o non sollecitati e impara dagli errori per evitare di ripeterli in futuro.
- Crea requisiti organizzativi dall'alto verso il basso affinché i carichi di lavoro siano eseguiti nel rispetto di budget predefiniti.

Tieniti aggiornato sulle nuove versioni dei servizi: potresti essere in grado di implementare nuovi servizi e funzionalità AWS per aumentare l'efficienza in termini di costi nel tuo carico di lavoro. Consulta regolarmente il blog delle novità AWS, il blog sulla gestione dei costi AWS e la sezione Novità di AWS per informazioni su nuovi servizi e versioni di funzionalità.

Quantificare il valore aggiunto realizzato attraverso l'ottimizzazione dei costi

Quantifica il valore aggiunto realizzato attraverso l'ottimizzazione dei costi: oltre a rendicontare i risparmi derivanti dall'ottimizzazione dei costi, è consigliabile quantificare il valore aggiunto fornito. I vantaggi dell'ottimizzazione dei costi sono in genere quantificati in termini di costi inferiori per ottenere un risultato aziendale. Ad esempio, puoi quantificare la riduzione dei costi di Amazon Elastic Compute Cloud (Amazon EC2) on demand quando acquisti Savings Plans, che riduce i costi e mantiene i livelli di output del carico di lavoro. Puoi quantificare le riduzioni dei costi in AWS quando le istanze Amazon EC2 inattive vengono terminate o quando i volumi Amazon Elastic Block Store (Amazon EBS) scollegati vengono eliminati.

La quantificazione del valore aggiunto realizzato tramite l'ottimizzazione dei costi consente di comprendere l'intero set di vantaggi per la tua organizzazione. Poiché l'ottimizzazione dei costi è un investimento necessario, la quantificazione del valore aggiunto consente di spiegare il ritorno sull'investimento agli stakeholder. La quantificazione del valore aggiunto può aiutarti a ottenere maggiori consensi dagli stakeholder sugli investimenti futuri in materia di ottimizzazione dei costi, e fornisce un framework per misurare i risultati delle attività di ottimizzazione dei costi della tua organizzazione.

I vantaggi derivanti dall'ottimizzazione dei costi, tuttavia, vanno oltre la riduzione o l'eliminazione dei costi. Prendi in considerazione l'acquisizione di dati aggiuntivi per misurare i miglioramenti dell'efficienza e il valore aggiunto. Esempi di miglioramenti includono:

- **Esegui best practice di ottimizzazione dei costi:** ad esempio, la gestione del ciclo di vita delle risorse riduce i costi operativi e infrastrutturali e fornisce margini di tempo e budget non pianificati da dedicare alla sperimentazione. Questo aumenta l'agilità dell'organizzazione e svela nuove opportunità per la generazione di ricavi.
- **Implementa l'automazione:** ad esempio, prova Auto Scaling, che garantisce elasticità al minimo sforzo e aumenta la produttività del personale eliminando il lavoro di pianificazione manuale della capacità. Per ulteriori dettagli sulla resilienza operativa, consulta il whitepaper sul principio dell'affidabilità del canone di architettura.
- **Prevedi i costi AWS futuri:** la previsione consente agli stakeholder finanziari di stabilire le aspettative con altri soggetti interni ed esterni dell'organizzazione e aiuta a migliorare la prevedibilità finanziaria dell'organizzazione. AWS Cost Explorer può essere utilizzato per eseguire previsioni relative a costi e utilizzo.

Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice di AWS in merito al budget e alla previsione delle spese per il cloud.

- Rendicontare i parametri di budget con report di budget
- Previsioni con AWS Cost Explorer
- Training AWS
- AWS Certification
- Partner di AWS Cloud Management Tools

Consapevolezza delle spese e dell'utilizzo

Comprendere i costi e i fattori chiave della tua organizzazione è fondamentale per gestire i costi e l'utilizzo in modo efficiente e per identificare le opportunità di riduzione dei costi. In genere, le organizzazioni gestiscono molteplici carichi di lavoro eseguiti da più team. Questi team possono trovarsi in diverse unità aziendali, ognuna con un proprio flusso di ricavi. La capacità di attribuire i costi delle risorse ai singoli proprietari del carico di lavoro, del prodotto o dell'organizzazione incoraggia un comportamento di utilizzo efficiente e contribuisce a ridurre gli sprechi. Un'attribuzione precisa dei costi ti consente di capire se le unità aziendali e i prodotti sono redditizi e ti aiuta a prendere decisioni più consapevoli in merito a dove allocare le risorse all'interno dell'azienda. La consapevolezza dell'utilizzo a tutti i livelli dell'organizzazione è fondamentale per promuovere il cambiamento, poiché la modifica dell'utilizzo determina variazioni dei costi.

Prova a adottare una strategia versatile per acquisire consapevolezza delle tue spese. Il tuo team deve raccogliere i dati, analizzarli e poi creare dei report. I fattori chiave da considerare includono:

- Governance
- Monitora i costi e l'utilizzo
- Disattivazione

Governance

Per gestire i costi nel cloud, è necessario gestire l'utilizzo tramite le seguenti aree di governance:

Sviluppa policy organizzative: il primo passo per eseguire la governance consiste nell'utilizzare i requisiti della tua organizzazione per sviluppare policy per l'utilizzo del cloud. Queste policy definiscono il modo in cui l'organizzazione utilizza il cloud e il modo in cui le risorse vengono gestite. Le politiche devono coprire tutti gli aspetti dei costi relativi alle risorse e ai carichi di lavoro correlati a costi o utilizzo, compresa la creazione, la modifica e la disattivazione durante il ciclo di vita della risorsa.

Le policy devono essere semplici, in modo che siano facilmente comprensibili e possano essere implementate in modo efficace in tutta l'organizzazione. Inizia con policy ampie e di alto livello, ad esempio in quale regione geografica è consentito l'utilizzo o l'ora del giorno in cui le risorse devono essere in esecuzione. Raffina gradualmente le policy per le varie unità organizzative e i diversi carichi di lavoro. Le policy comuni includono i servizi e le funzionalità che possono essere utilizzati (ad esempio, lo storage ha prestazioni inferiori negli ambienti di test/sviluppo) e i tipi di risorse che possono essere utilizzati dai diversi gruppi (ad esempio, le dimensioni massime di una risorsa in un account di sviluppo sono medie).

Sviluppa obiettivi e target: sviluppa obiettivi e target di costi e utilizzo per la tua organizzazione. Gli obiettivi forniscono all'organizzazione linee guida e indicazioni sui risultati previsti. I target forniscono i risultati specifici e misurabili da raggiungere. Ad esempio, un obiettivo potrebbe essere: l'utilizzo della piattaforma deve aumentare in modo significativo, implicando solamente un aumento minore (non lineare) dei costi. Un esempio di target invece potrebbe essere: un aumento del 20% dell'utilizzo della piattaforma, con un aumento dei costi inferiore al 5%. Un altro obiettivo comune è che i carichi di lavoro devono essere più efficienti ogni 6 mesi. Il target associato a tale obiettivo è che il costo per output del carico di lavoro deve diminuire del 5% ogni 6 mesi.

Un obiettivo comune per i carichi di lavoro nel cloud è l'incremento dell'efficienza del carico di lavoro, ossia la riduzione del costo per il risultato aziendale del carico di lavoro nel corso del tempo. Si consiglia di implementare questo obiettivo per tutti i carichi di lavoro e di stabilire inoltre un target come l'aumento dell'efficienza del 5% ogni 6-12 mesi. Questo può essere ottenuto nel cloud attraverso la creazione di capacità per l'ottimizzazione dei costi e tramite il rilascio di nuovi servizi e loro funzionalità.

Struttura dell'account: AWS presenta una struttura degli account ad albero, con un account principale (il padre, comunemente noto come account di pagamento) e vari account collegati (i figli, noti come account membri). Una best practice è di avere sempre almeno un account principale con un account membro, indipendentemente dalle dimensioni o dall'utilizzo dell'organizzazione. Tutte le risorse del carico di lavoro devono risiedere solo all'interno degli account membri.

Non esiste una risposta unica in merito al numero di account AWS che dovresti avere. Valuta i tuoi modelli operativi e di costo attuali e futuri per assicurarti che la struttura dei tuoi account AWS rispecchi quella della tua organizzazione. Alcune aziende creano molteplici account AWS per motivi aziendali, ad esempio:

- È richiesto l'isolamento amministrativo e/o fiscale e di fatturazione tra unità aziendali o centri di costo o carichi di lavoro specifici.
- Le restrizioni dei servizi AWS sono impostate in modo che risultino specifiche per determinati carichi di lavoro.
- Esiste un requisito per l'isolamento e la separazione tra carichi di lavoro e risorse.

All'interno di AWS Organizations, la fatturazione consolidata crea il costrutto tra uno o più account membri e l'account principale. Gli account membri consentono di isolare e distinguere i costi e l'utilizzo per gruppi. Una pratica comune è quella di avere account membri separati per ciascuna unità aziendale (come finanza, marketing e vendite), per il ciclo di vita di ciascun ambiente (come sviluppo, test e produzione) o per ciascun carico di lavoro (carico di lavoro a, b e c), e poi aggregare questi account membri tramite la fatturazione consolidata.

La fatturazione consolidata consente di accorpare i pagamenti di più account membri AWS sotto un unico account principale, e al tempo stesso di fornire comunque visibilità all'attività di ciascun account membro. Poiché i costi e l'utilizzo vengono aggregati nell'account principale, questo consente di massimizzare gli sconti per volume di servizio e di massimizzare l'utilizzo degli sconti a fronte di impegni (Savings Plans e istanze riservate) per ottenere gli sconti più elevati.

AWS Control Tower può impostare e configurare rapidamente più account AWS, garantendo una governance in linea con i requisiti della tua organizzazione.

Gruppi e ruoli organizzativi: dopo avere sviluppato le policy, è possibile creare gruppi logici e ruoli degli utenti all'interno dell'organizzazione. In questo modo puoi assegnare le autorizzazioni e controllare l'utilizzo. Inizia con gruppi di alto livello di persone, in genere seguendo le unità organizzative e i ruoli lavorativi (ad esempio, amministratore di sistema nel reparto IT o controllore finanziario). I gruppi raggruppano persone che eseguono attività simili e necessitano di un accesso simile. I ruoli definiscono che cosa un gruppo deve fare. Ad esempio, un amministratore di sistema nel reparto IT deve disporre di un accesso che permetta di creare tutte le risorse, mentre un membro del team di analisi ha la necessità di creare soltanto risorse di analisi.

Controlli - Notifiche: un primo passo comune per implementare i controlli dei costi consiste nell'impostare delle notifiche quando si verificano eventi di costi o utilizzo al di fuori delle policy. In questo modo, puoi agire rapidamente e verificare se è necessaria un'azione correttiva, senza limitare o influire negativamente sui carichi di lavoro o sulle nuove attività. Dopo avere appreso i limiti del carico di lavoro e dell'ambiente, puoi applicare la governance. In AWS, le notifiche vengono effettuate con Budget AWS, che consente di definire un budget mensile per i costi, l'utilizzo e gli sconti a fronte di impegni di AWS (Savings Plans e istanze riservate). Puoi creare budget a livello di costo aggregato (ad esempio, tutti i costi) o a un livello più granulare, nel quale includi solo dimensioni specifiche come account membri, servizi, tag o zone di disponibilità. Puoi anche collegare ai tuo budget delle notifiche tramite e-mail, che si attivano quando i tuoi costi attuali o previsti superano una soglia percentuale da te definita.

Controlli - Applicazione: in secondo luogo, è possibile applicare le policy di governance in AWS tramite AWS Identity and Access Management (IAM) e le policy di controllo dei servizi (SCP) di AWS Organizations. IAM consente di gestire in modo sicuro l'accesso ai servizi e alle risorse AWS. Utilizzando IAM, puoi controllare chi può creare e gestire le risorse AWS, il tipo di risorse che possono essere create e dove possono essere create. Ciò riduce al minimo la creazione di risorse che non sono necessarie. Utilizza i ruoli e i gruppi creati in precedenza e assegna policy IAM per garantire l'utilizzo corretto. Le SCP offrono il controllo centralizzato sul numero massimo di autorizzazioni disponibili per tutti gli account nella tua organizzazione, assicurando che i tuoi account rimangano entro le linee guida di controllo degli accessi. Le SCP sono disponibili soltanto in un'organizzazione con tutte le funzionalità abilitate e possono essere configurate in modo da rifiutare o consentire operazioni agli account membri per impostazione predefinita. Consulta il whitepaper sul principio della sicurezza del canone di architettura per ulteriori dettagli sull'implementazione della gestione degli accessi.

Controlli - Quote di servizio: la governance può essere implementata anche tramite la gestione delle quote di servizio. Assicurandoti che le quote di servizio siano impostate con spese minime e siano gestite in modo accurato, puoi ridurre al minimo la creazione di risorse che non rientrano nei requisiti della tua organizzazione. Per ottenere questo risultato, devi comprendere la velocità con cui i tuoi requisiti possono cambiare, comprendere i progetti in corso (sia la creazione sia la disattivazione di risorse) e considerare la velocità con cui è possibile implementare le modifiche alle quote. Le quote di servizio possono essere utilizzate per aumentare le quote all'occorrenza.

Il servizio di gestione costi AWS è integrato dal servizio AWS Identity and Access Management (IAM). Puoi utilizzare il servizio IAM insieme a quello di gestione costi per controllare l'accesso ai tuoi dati finanziari e agli strumenti AWS nella console di fatturazione.

Monitora il ciclo di vita del carico di lavoro: assicurati di tenere traccia dell'intero ciclo di vita del carico di lavoro. In questo modo, quando i carichi di lavoro o i componenti del carico di lavoro non sono più necessari, potrai disattivarli o modificarli. Ciò si rivela particolarmente utile quando rilasci nuovi servizi o funzionalità. I carichi di lavoro e i componenti esistenti possono essere mostrati come in uso, ma devono essere disattivati e reindirizzare i clienti al nuovo servizio.

Presta attenzione alle diverse fasi dei carichi di lavoro: quando un carico di lavoro arriva in produzione, gli ambienti precedenti possono essere disattivati o notevolmente ridotti in termini di capacità fino a quando non sono nuovamente necessari.

AWS offre una serie di servizi di gestione e governance utilizzabili per il monitoraggio del ciclo di vita delle entità. Puoi utilizzare AWS Config o AWS Systems Manager per fornire un inventario dettagliato delle risorse e della configurazione AWS. Si consiglia di integrare questi servizi con i sistemi di gestione di progetti o asset esistenti per tenere traccia dei progetti e dei prodotti attivi all'interno della tua organizzazione. La combinazione del tuo sistema attuale con l'ampia gamma di eventi e parametri forniti da AWS ti consentirà di ottenere una panoramica degli eventi del ciclo di vita significativi e di gestire le risorse in modo proattivo per ridurre i costi non necessari.

Consulta il whitepaper sul principio dell'eccellenza operativa del canone di architettura per ulteriori dettagli sull'implementazione del monitoraggio del ciclo di vita delle entità.

Monitora i costi e l'utilizzo

Consenti ai team di intervenire sui costi e sull'utilizzo tramite una visibilità dettagliata del carico di lavoro. L'ottimizzazione dei costi inizia dalla comprensione granulare dei costi e dell'utilizzo, dalla possibilità di modellare e prevedere spese, utilizzo e funzionalità futuri e dall'implementazione di meccanismi sufficienti per allineare i costi e l'utilizzo agli obiettivi della tua organizzazione. Di seguito sono elencate le aree necessarie per monitorare i costi e l'utilizzo:

Configura origini dati dettagliate: abilita la granularità oraria in Cost Explorer e crea un report costi e utilizzo (CUR). Queste origini dati forniscono la visualizzazione più accurata dei costi e dell'utilizzo dell'intera organizzazione. Il CUR fornisce una granularità di utilizzo giornaliera o oraria, tariffe, costi e attributi di utilizzo per tutti i servizi AWS addebitati. Tutte le dimensioni possibili sono incluse nel CUR, tra cui applicazione di tag, posizione, attributi di risorsa e ID account.

Configura il CUR con le seguenti personalizzazioni:

- Inclusione degli ID risorsa
- Aggiornamento automatico del CUR
- Granularità oraria
- Versioni multiple: sovrascrivi il report esistente
- Integrazione dei dati: Athena (formato Parquet e compressione)

Utilizza AWS Glue per preparare i dati per l'analisi e Amazon Athena per eseguire l'analisi dei dati, utilizzando SQL per eseguire query sui dati. Puoi anche utilizzare Amazon QuickSight per creare visualizzazioni personalizzate e complesse e distribuirle in tutta l'organizzazione.

Identifica le categorie di attribuzione dei costi: collabora con il tuo team finanziario e altri stakeholder per comprendere i requisiti di allocazione dei costi all'interno della tua

organizzazione. I costi del carico di lavoro devono essere allocati per tutto il ciclo di vita, inclusi sviluppo, test, produzione e disattivazione. Comprendi in che modo i costi sostenuti per formazione, sviluppo del personale e creazione di idee sono attribuiti all'interno dell'organizzazione. Questo può essere utile per allocare correttamente gli account utilizzati per questo scopo ai budget di formazione e sviluppo, anziché ai budget generici dei costi IT.

Stabilisci i parametri del carico di lavoro: comprendi in che modo viene misurato l'output del carico di lavoro rispetto al successo aziendale. Ogni carico di lavoro ha in genere un piccolo set di output principali che indicano le prestazioni. Se disponi di un carico di lavoro complesso con molti componenti, puoi dare priorità alle voci dell'elenco o definire e monitorare i parametri per ogni componente. Collabora con i tuoi team per capire quali parametri utilizzare. Questa unità verrà utilizzata per comprendere l'efficienza del carico di lavoro o il costo per ciascun output aziendale.

Assegna a costi e utilizzo un significato per l'organizzazione: applica dei tag in AWS per aggiungere informazioni sull'organizzazione alle tue risorse, che verranno quindi aggiunte alle informazioni su costi e utilizzo. Un tag è una coppia chiave-valore: la chiave è definita e deve essere univoca all'interno dell'organizzazione, mentre il valore è univoco per un gruppo di risorse. Ad esempio, una coppia chiave-valore può essere costituita da ambiente (chiave) e produzione (valore). Tutte le risorse nell'ambiente di produzione avranno questa coppia chiave-valore. L'applicazione di tag consente di categorizzare e monitorare i costi con informazioni significative e rilevanti sull'organizzazione. Puoi applicare tag che rappresentano categorie dell'organizzazione (ad esempio centri di costo, nomi di applicazioni, progetti o proprietari) e identificano carichi di lavoro e rispettive funzionalità (ad esempio test o produzione) per attribuire i costi e l'utilizzo all'interno dell'organizzazione.

Quando applichi i tag alle tue risorse AWS (come le istanze EC2 o i bucket Amazon S3) e li attivi, AWS aggiunge queste informazioni ai report costi e utilizzo. Puoi creare report e condurre analisi su risorse con tag e senza tag per incrementare la conformità con le policy di gestione dei costi interne e garantire un'attribuzione accurata.

La creazione e l'implementazione di uno standard per l'applicazione di tag AWS tra gli account della tua organizzazione ti consentiranno di gestire e amministrare i tuoi ambienti AWS in modo coerente e uniforme. Utilizza le policy di tag in AWS Organizations per definire regole su come i tag possono essere applicati alle risorse AWS nei tuoi account in AWS Organizations. Le policy di tag consentono di adottare con facilità un approccio standardizzato per l'applicazione di tag alle risorse AWS.

AWS Tag Editor consente di aggiungere, eliminare e gestire tag di più risorse.

AWS Cost Categories consente di assegnare ai tuoi costi significati per l'organizzazione, senza necessità di applicare tag alle risorse. Puoi mappare le informazioni su costi e utilizzo attribuendole a strutture organizzative interne univoche. Puoi definire regole di categoria per mappare e categorizzare i costi utilizzando le dimensioni di fatturazione, ad esempio account e tag. Questo offre un altro livello di funzionalità di gestione oltre all'applicazione di tag. Puoi anche mappare account e tag specifici attribuendoli a più progetti.

Configura gli strumenti di fatturazione e ottimizzazione dei costi: per modificare l'utilizzo e modificare i costi, ogni persona nella tua organizzazione deve avere accesso alle informazioni relative a costi e utilizzo. È consigliabile che tutti i carichi di lavoro e i team dispongano dei seguenti strumenti configurati quando utilizzano il cloud:

- **Report:** riepiloga tutte le informazioni su costi e utilizzo.
- **Notifiche:** fornisci notifiche quando il costo o l'utilizzo non rientra nei limiti definiti.
- **Stato attuale:** configura un pannello di controllo che mostra i livelli correnti di costi e utilizzo. Il pannello di controllo deve essere disponibile in un luogo altamente visibile all'interno dell'ambiente di lavoro (simile a un pannello di controllo delle operazioni).
- **Tendenze:** offri la possibilità di mostrare la variabilità dei costi e dell'utilizzo nel periodo di tempo richiesto e con la granularità richiesta.
- **Previsioni:** offri la possibilità di mostrare i costi futuri stimati.
- **Monitoraggio:** mostra i costi e l'utilizzo attuali rispetto a obiettivi o target stabiliti.
- **Analisi:** offri ai membri del team la possibilità di eseguire analisi personalizzate e approfondite fino alla granularità oraria, con tutte le dimensioni possibili.

Per fornire queste funzionalità, puoi utilizzare strumenti nativi di AWS come AWS Cost Explorer, Budget AWS e Amazon Athena con QuickSight. Puoi anche utilizzare strumenti di terze parti, tuttavia devi assicurarti che i costi di tali strumenti forniscano un valore effettivo alla tua organizzazione.

Alloca i costi in base ai parametri del carico di lavoro: ottimizzare i costi significa conseguire i risultati aziendali al prezzo più basso, e implica l'allocazione dei costi del carico di lavoro in base ai parametri del carico di lavoro (misurati in termini di efficienza del carico di lavoro). Monitora i parametri del carico di lavoro definiti tramite file di log o altre funzionalità di monitoraggio dell'applicazione. Combina questi dati con i costi del carico di lavoro, che possono essere ottenuti osservando i costi con un determinato valore di tag o ID account. Si consiglia di eseguire questa analisi a livello orario. L'efficienza cambia in genere se disponi di alcuni componenti di costo statico (ad esempio, un database back-end in esecuzione 24 ore su 24, 7 giorni su 7) con un tasso di richiesta variabile (ad esempio, picchi di utilizzo tra le 9:00 e le 17:00, con poche richieste di notte). Comprendere la relazione tra i costi statici e i costi variabili ti aiuterà a rendere più mirate le tue attività di ottimizzazione.

Disattiva le risorse

Quando gestisci una serie di progetti, di dipendenti e di risorse tecnologiche nel tempo, sarai in grado di individuare le risorse che non sono più utilizzate o i progetti accantonati che non hanno più un proprietario.

Monitora le risorse nel loro ciclo di vita: disattiva le risorse dei carichi di lavoro che non sono più necessarie. Un esempio comune sono le risorse utilizzate per i test: dopo il completamento dei test, le risorse possono essere rimosse. Monitorare le risorse con i tag (ed esecuzione di report su tali tag) ti aiuterà a identificare gli asset da disattivare. L'utilizzo dei tag è un modo efficace per monitorare le risorse: puoi etichettare la risorsa con la relativa funzione o con una data nota in cui può essere disattivata. Puoi quindi eseguire i report su questi tag. Esempi di valori per l'applicazione di tag relativi alle funzionalità sono "test funzionalitàX" per identificare lo scopo della risorsa in termini di ciclo di vita del carico di lavoro.

Implementa un processo di disattivazione: implementa un processo standardizzato in tutta l'organizzazione per identificare e rimuovere le risorse inutilizzate. Il processo deve definire la frequenza di esecuzione della ricerca e i processi per rimuovere la risorsa al fine di garantire che tutti i requisiti dell'organizzazione siano soddisfatti.

Disattiva le risorse: la frequenza e lo sforzo di ricerca delle risorse inutilizzate dovrebbero riflettere i risparmi potenziali, pertanto un account con costi contenuti deve essere analizzato con una frequenza minore rispetto a un account che ha costi maggiori. Gli eventi di ricerca e disattivazione possono essere attivati da modifiche di stato nel carico di lavoro, ad esempio il termine del ciclo di vita di un prodotto o la sua sostituzione. Le ricerche e gli eventi di disattivazione possono anche essere attivati da eventi esterni, ad esempio cambiamenti nelle condizioni di mercato o cessazione del prodotto.

Disattiva le risorse in modo automatico: utilizza l'automazione per ridurre o rimuovere i costi associati al processo di disattivazione. Progettare il carico di lavoro per eseguire automaticamente la disattivazione ridurrà i costi complessivi del carico di lavoro durante il suo ciclo di vita. Puoi utilizzare AWS Auto Scaling per eseguire il processo di disattivazione. Puoi anche implementare un codice personalizzato utilizzando un'API o SDK per disattivare automaticamente le risorse del carico di lavoro.

Risorse

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

- Strategie di applicazione di tag AWS
- Attivazione dei tag per l'allocazione dei costi in base all'utente
- Fatturazione e gestione costi AWS
- Blog sulla gestione dei costi
- Strategia di fatturazione con account multipli di AWS
- SDK AWS e strumenti

- Best practice per l'applicazione di tag
- Well-Architected Labs - Nozioni di base dei costi
- Well-Architected Labs - Consapevolezza nella spesa

Convenienza delle risorse

Utilizzare le risorse, le configurazioni e i servizi adeguati per i tuoi carichi di lavoro è fondamentale per ridurre i costi. Considera i seguenti aspetti durante la creazione di risorse convenienti:

- Valuta i costi al momento di selezionare i servizi
- Seleziona il tipo, la dimensione e il numero di risorse corretti
- Seleziona il migliore modello di prezzo
- Pianifica il trasferimento dei dati

Puoi ricorrere agli AWS Solutions Architect, alle soluzioni AWS, alle architetture di riferimento AWS e ai partner APN per scegliere l'architettura in base a ciò che hai appreso.

Valuta i costi al momento di selezionare i servizi

Identifica i requisiti dell'organizzazione: al momento di selezionare i servizi per un carico di lavoro, è fondamentale comprendere le priorità dell'organizzazione. Assicurati che vi sia equilibrio tra i costi e gli altri principi del canone di architettura, ad esempio prestazioni e affidabilità. Un carico di lavoro completamente ottimizzato per i costi è la soluzione più in linea con i requisiti della tua organizzazione, e non necessariamente quella con il costo più basso. Interpella tutti i team all'interno della tua organizzazione, quali prodotti, business, tecnici e finanziari, per raccogliere il maggior numero di informazioni.

Analizza tutti i componenti del carico di lavoro: esegui un'analisi completa su tutti i componenti del carico di lavoro. Assicurati che il costo dell'analisi e il potenziale risparmio nel carico di lavoro durante il ciclo di vita siano in equilibrio. Devi individuare l'impatto attuale e il potenziale impatto futuro del componente. Ad esempio, se il costo della risorsa proposta è di 10 USD al mese e sotto i carichi previsti non supererà i 15 USD al mese, la spesa di una giornata di impegno per ridurre i costi del 50% (5 USD al mese) potrebbe superare il potenziale vantaggio per tutta la durata del sistema. L'utilizzo di una stima basata sui dati, più rapida ed efficiente, fornirà il migliore risultato complessivo per questo componente.

Dato che i carichi di lavoro possono cambiare nel corso del tempo, il giusto set di servizi potrebbe non essere ottimale se l'architettura o l'utilizzo del carico di lavoro cambiano. L'analisi per la selezione dei servizi deve integrare gli stati del carico di lavoro e i livelli di utilizzo attuali e futuri. Implementare un servizio in funzione dello stato o dell'utilizzo futuro del carico di lavoro può ridurre i costi complessivi, riducendo o rimuovendo lo sforzo necessario per apportare modifiche future.

AWS Cost Explorer e il CUR possono analizzare i costi di un proof of concept (PoC) o di un ambiente in esecuzione. Puoi anche utilizzare il Calcolatore di costo mensile AWS o il Calcolatore di prezzi AWS per stimare i costi del carico di lavoro.

Servizi gestiti: i servizi gestiti eliminano l'onere operativo e amministrativo legato alla manutenzione di un servizio, consentendoti di concentrarti sull'innovazione. Inoltre, poiché i servizi gestiti operano su scala cloud, possono offrire un costo inferiore per transazione o servizio.

Considera il risparmio in termini di tempo, che consentirà al tuo team di concentrarsi sull'eliminazione del debito tecnico, sull'innovazione e sulle funzionalità che offrono un valore aggiunto. Ad esempio, potresti avere bisogno di trasferire il tuo ambiente locale nel cloud il più rapidamente possibile ed eseguire l'ottimizzazione in un secondo momento. Vale la pena soffermarsi sul risparmio che puoi ottenere usando i servizi gestiti che rimuovono o riducono i costi di licenza.

Solitamente, i servizi gestiti presentano attributi che puoi impostare per garantire la capacità necessaria. Devi impostare e monitorare questi attributi in modo che la tua capacità in eccesso sia mantenuta al minimo e le prestazioni siano massimizzate. Puoi modificare gli attributi di AWS Managed Services utilizzando la Console di gestione AWS o le API e gli SDK AWS per allineare le risorse necessarie con le variazioni della domanda. Ad esempio, puoi aumentare o diminuire il numero di nodi su un cluster Amazon EMR o Amazon RedShift per dimensionarlo.

Puoi anche unire più istanze in una risorsa AWS per ottenere un utilizzo di densità più elevata. Ad esempio, puoi predisporre più database di dimensioni ridotte su una singola istanza DB di Amazon Relational Database Service (Amazon RDS). Quando l'utilizzo si intensifica, puoi migrare uno dei database su un'istanza DB di RDS dedicata utilizzando uno snapshot e ripristinare il processo.

Quando predisponi carichi di lavoro su servizi gestiti, devi comprendere i requisiti inerenti alla modifica della capacità del servizio. Tali requisiti solitamente riguardano il tempo, l'impegno e qualunque impatto sul normale funzionamento del carico di lavoro. La risorsa predisposta deve offrire il tempo necessario per l'applicazione delle modifiche, pertanto procurati i mezzi necessari a tal fine. L'impegno costante richiesto per modificare i servizi può essere ridotto praticamente a zero grazie alle API e agli SDK integrati nel sistema, nonché grazie a strumenti di monitoraggio come Amazon CloudWatch.

Amazon Relational Database Service (RDS), Amazon Redshift e Amazon ElastiCache forniscono un servizio di database gestito. Amazon Athena, Amazon Elastic Map Reduce (EMR) e Amazon Elasticsearch offrono un servizio di analisi gestito.

AWS Managed Services (AMS) è un servizio che gestisce l'infrastruttura AWS per conto di clienti e partner aziendali. Fornisce un ambiente sicuro e conforme in cui è possibile distribuire i carichi di lavoro. AMS utilizza modelli operativi cloud aziendali dotati di automazione per consentirti di soddisfare i requisiti aziendali, di passare più rapidamente al cloud e di ridurre i costi di gestione correnti.

Servizi serverless o a livello di applicazione: puoi usare anche servizi serverless o a livello di applicazione, come AWS Lambda, Amazon Simple Queue Service (Amazon SQS), Amazon Simple Notification Service (Amazon SNS) e Amazon Simple Email Service (Amazon SES). Questi servizi eliminano la necessità di gestire una risorsa e forniscono funzioni di esecuzione del codice, servizi di accodamento e consegna dei messaggi. L'altro vantaggio è che le prestazioni e i costi vengono adattati in base all'utilizzo, garantendo l'allocazione e l'attribuzione dei costi in modo efficiente.

Per ulteriori informazioni su Serverless, consulta il whitepaper Well-Architected Serverless Application Lens.

Analizza il carico di lavoro per un utilizzo diverso nel corso del tempo: quando AWS rilascia nuovi servizi e funzionalità, è possibile che i servizi ottimali per il carico di lavoro cambino. Tale cambiamento comporta un impegno, che dovrebbe essere commensurato ai vantaggi potenziali. La frequenza di revisione del carico di lavoro dipende dai requisiti dell'organizzazione. Se si tratta di un carico di lavoro con costi importanti, una rapida implementazione dei nuovi servizi massimizzerà i risparmi sui costi, e in tal caso una revisione più frequente può risultare vantaggiosa. Un altro stimolo importante per la revisione è il cambiamento dei modelli di utilizzo. Se si verificassero notevoli cambiamenti nell'utilizzo, ciò potrebbe indicare un maggiore vantaggio dei servizi alternativi. Ad esempio, per velocità di trasferimento dei dati più elevate, un servizio di connessione diretta può risultare più economico di una VPN e garantire la connettività richiesta. Prevedi il potenziale impatto dei cambiamenti del servizio, così da monitorare le variazioni a livello di utilizzo e implementare con anticipo i servizi più convenienti.

Costi di licenza: utilizzando software open source è possibile eliminare il costo delle licenze software. Con l'aumentare delle dimensioni del carico di lavoro, l'impatto sui costi può essere significativo. Misura i vantaggi di usare software con licenza in rapporto ai costi totali per assicurarti di disporre di un carico di lavoro che sia il più ottimizzato possibile. Crea modelli per le eventuali modifiche alla licenza e il relativo impatto sui costi del carico di lavoro. Se un fornitore modifica il costo della licenza del database, valuta come questo incide sull'efficienza complessiva del carico di lavoro. Effettua un'analisi dello storico dei prezzi dei tuoi fornitori per scoprire le tendenze dei cambiamenti relativi alle licenze dei loro prodotti. I costi delle licenze possono anche essere adattati indipendentemente dal throughput o dall'utilizzo, come nel caso delle licenze che si adattano in base all'hardware (licenze legate alla CPU). È necessario evitare queste licenze poiché i costi possono aumentare rapidamente senza che vi siano vantaggi correlati.

Puoi utilizzare AWS License Manager per gestire le licenze software del tuo carico di lavoro. Puoi configurare regole per le licenze e applicare condizioni necessarie per prevenire le violazioni delle licenze e ridurre i costi legati all'eccedenza di licenze.

Seleziona il tipo, le dimensioni e il numero di risorse in modo corretto

Selezionando il tipo, le dimensioni e il numero di risorse nel modo migliore, puoi soddisfare i requisiti tecnici con le risorse dal costo più basso. Le attività di dimensionamento appropriato tengono conto di tutte le risorse di un carico di lavoro, di tutti gli attributi di ogni singola risorsa e dell'impegno necessario alle operazioni di dimensionamento appropriato. Il dimensionamento appropriato può corrispondere a un processo iterativo, attivato dalle modifiche nei modelli di utilizzo e da fattori esterni come la riduzione dei prezzi di AWS e nuovi tipi di risorse di AWS. La ricerca del dimensionamento appropriato può anche essere svolta una tantum se il suo costo supera il potenziale risparmio nell'arco di vita del carico di lavoro.

AWS offre una serie di strategie diverse:

- Esecuzione della modellizzazione dei costi
- Seleziona la dimensione in base a parametri o dati
- Seleziona automaticamente la dimensione (in base ai parametri)

Modellazione dei costi: esegui la modellazione dei costi per il tuo carico di lavoro e ciascuno dei suoi componenti per comprendere il bilanciamento tra le risorse e trovare la dimensione appropriata per ogni risorsa nel carico di lavoro, sulla base di un determinato livello di prestazioni. Esegui attività di analisi comparativa per il carico di lavoro in base ai diversi carichi previsti e confronti i costi. L'attività di modellazione deve riflettere i potenziali benefici; ad esempio, il tempo speso è proporzionale al costo dei componenti o al risparmio previsto. Per le best practice, consulta la sezione Revisione del whitepaper sul Principio dell'efficienza delle prestazioni del Canone di architettura AWS.

AWS Compute Optimizer è in grado di supportare la modellazione dei costi per l'esecuzione di carichi di lavoro. Fornisce consigli di dimensionamento appropriato per le risorse di calcolo in base a una valutazione cronologica dell'utilizzo. Questa è la fonte di dati ideale per le risorse di calcolo perché è un servizio gratuito e utilizza il machine learning per generare più raccomandazioni a seconda dei livelli di rischio. Inoltre, è possibile utilizzare Amazon CloudWatch e CloudWatch Logs con log personalizzati come fonti di dati per operazioni di dimensionamento appropriato per altri servizi e componenti del carico di lavoro.

Di seguito sono riportate le raccomandazioni per i parametri e i dati di modellazione dei costi:

- Il monitoraggio deve corrispondere in modo preciso all'esperienza degli utenti finali. Seleziona la granularità corretta per un dato periodo di tempo e scegli in modo ponderato il 99° percentile o quello massimo invece del valore medio.

- Seleziona la granularità corretta per il periodo di analisi richiesto per coprire tutti i cicli del carico di lavoro. Ad esempio, se esegui un'analisi di due settimane, potresti ignorare un ciclo mensile di utilizzo elevato, e questo potrebbe causare un provisioning insufficiente.

Selezione basata su dati o parametri: seleziona la dimensione o il tipo di risorsa in base al carico di lavoro e alle caratteristiche delle risorse; per esempio, elaborazione, memoria, throughput o scrittura intensiva. Questa selezione è tipicamente effettuata ricorrendo alla modellazione dei costi, utilizzando una versione precedente del carico di lavoro (ad esempio una versione in locale), utilizzando la documentazione o altre fonti di informazione sul carico di lavoro (come whitepaper e soluzioni pubblicate).

Selezione automatica basata su parametri: crea un ciclo di feedback all'interno del carico di lavoro che utilizza i parametri attivi del carico di lavoro in esecuzione per apportare modifiche a tale carico di lavoro. Puoi utilizzare un servizio gestito, ad esempio AWS Auto Scaling che configuri per eseguire le operazioni di dimensionamento appropriato più adatte a te. AWS offre anche API, SDK e funzionalità che consentono alle risorse di essere modificate con il minimo sforzo. È possibile programmare un carico di lavoro per arrestare e avviare un'istanza EC2 per consentire una modifica delle dimensioni dell'istanza o del tipo di istanza. Ciò offre i vantaggi del dimensionamento appropriato, eliminando al contempo quasi tutti i costi operativi necessari per apportare la modifica.

Alcuni servizi AWS hanno integrato la selezione automatica del tipo o della dimensione, ad esempio S3 Intelligent-Tiering. Basandosi sui modelli di utilizzo, S3 Intelligent-Tiering sposta automaticamente i dati tra due livelli di accesso: frequente e poco frequente.

Seleziona il modello di prezzo migliore

Esegui la modellazione dei costi del carico di lavoro: considera i requisiti dei componenti del carico di lavoro e comprendi i potenziali modelli di prezzo. Definisci il requisito di disponibilità del componente. Determina se ci sono più risorse indipendenti che eseguono la funzione nel carico di lavoro e quali sono i requisiti del carico di lavoro nel corso del tempo. Confronta il costo delle risorse utilizzando il modello di prezzo on demand predefinito e altri modelli applicabili. Tieni conto di qualsiasi potenziale cambiamento nelle risorse o nei componenti del carico di lavoro.

Esegui con regolarità analisi a livello di account: l'esecuzione della modellazione dei costi con regolarità garantisce l'implementazione di opportunità di ottimizzazione su più carichi di lavoro. Ad esempio, se più carichi di lavoro utilizzano on demand, a livello aggregato, il rischio di modifica è inferiore e l'implementazione di uno sconto a fronte di impegni otterrà un costo complessivo inferiore. Si consiglia di eseguire l'analisi seguendo cicli regolari, da due settimane a un mese. In questo modo è possibile effettuare acquisti in piccoli incrementi, così che la copertura dei modelli di prezzo evolva di pari passo con i carichi di lavoro e i relativi componenti.

Utilizza lo strumento per i suggerimenti AWS Cost Explorer per trovare opportunità di sconti a fronte di impegni.

Per trovare opportunità per i carichi di lavoro Spot, utilizza una visualizzazione oraria dell'utilizzo complessivo e cerca periodi regolari di variazione dell'utilizzo o di elasticità.

Modelli di prezzo: AWS offre diversi modelli di prezzo che consentono di pagare per le risorse nel modo più conveniente e adatto alle esigenze della tua organizzazione. La sezione seguente descrive ciascun modello di acquisto:

- On-demand
- Spot
- Sconti a fronte di impegni - Savings Plans
- Sconti a fronte di impegni - Capacità/istanze riservate
- Selezione geografica
- Accordi con terze parti e prezzi

On demand: questo è il modello di prezzo predefinito, per pagare in base al consumo. Quando usi risorse (ad esempio, come le istanze EC2 o servizi come DynamoDB on demand), paghi una tariffa oraria fissa e non hai impegni a lungo termine. Puoi incrementare o diminuire la capacità delle tue risorse o dei tuoi servizi in base alle domande relative alla tua applicazione. Le istanze on demand hanno una tariffa oraria, ma a seconda del servizio, possono essere fatturate in incrementi di 1 secondo (ad esempio AWS Lambda o istanze EC2 Linux). La soluzione on demand è consigliata per le applicazioni con carichi di lavoro a breve termine (come un progetto di quattro mesi) che raggiungono il picco periodicamente, oppure per carichi di lavoro non prevedibili e che non è possibile interrompere. L'on demand è inoltre idoneo per i carichi di lavoro, ad esempio ambienti di pre-produzione, che richiedono runtime ininterrotti, ma che non durano abbastanza per usufruire di sconti a fronte di impegni (Savings Plans o istanze riservate).

Spot: un'istanza Spot dispone di capacità di elaborazione EC2 inutilizzata disponibile con sconti fino al 90% sui prezzi on demand senza alcun impegno a lungo termine. Con le istanze Spot, puoi ridurre in modo significativo il costo di esecuzione delle applicazioni o adattare la capacità di calcolo dell'applicazione mantenendo lo stesso budget. A differenza delle istanze on demand, le istanze Spot possono essere interrotte con un avviso di 2 minuti se EC2 necessita di capacità o se il prezzo dell'istanza Spot supera il prezzo configurato. In media, le istanze Spot vengono interrotte meno del 5% del tempo.

Spot è ideale quando è presente una coda o un buffer, oppure quando ci sono più risorse che lavorano in modo indipendente per elaborare le richieste (ad esempio, l'elaborazione dei dati Hadoop). Generalmente questi carichi di lavoro sono tolleranti ai guasti, stateless e flessibili, come nel caso di elaborazione in batch, Big Data e strumenti di analisi, ambienti containerizzati e High Performance Computing (HPC). Anche i carichi di lavoro non critici come gli ambienti di test e sviluppo sono idonei per le istanze Spot.

Le istanze Spot sono inoltre integrate in diversi servizi AWS, ad esempio gruppi Auto Scaling EC2 (ASG), Elastic MapReduce (EMR), Elastic Container Service (ECS) e AWS Batch.

Quando un'istanza Spot deve essere recuperata, EC2 invia un avviso di due minuti tramite una notifica di interruzione dell'istanza Spot fornita attraverso CloudWatch Events, così come nei metadati dell'istanza. Durante il periodo di due minuti, l'applicazione può utilizzare questo tempo per salvare il proprio stato, esaurire i container in esecuzione, caricare i file di log finali o rimuoversi da un sistema di bilanciamento del carico. Al termine dei due minuti, è possibile ibernare, arrestare o terminare l'istanza Spot.

Quando utilizzi le istanze Spot nei tuoi carichi di lavoro, tieni a mente le seguenti best practice:

- **Imposta il prezzo massimo come tariffa on demand:** in questo modo pagherai la tariffa Spot corrente (il prezzo più basso disponibile) e non pagherai mai più della tariffa on demand. Le tariffe correnti e storiche sono disponibili tramite la console e l'API.
- **Cerca di essere flessibile per il maggior numero di tipi di istanze possibile:** sii flessibile per quanto concerne la famiglia e le dimensioni del tipo di istanza per migliorare la probabilità di soddisfare i requisiti di capacità prefissati, per ottenere il costo più basso possibile e per ridurre al minimo l'impatto delle interruzioni.
- **Cerca di essere flessibile relativamente a dove verrà eseguito il carico di lavoro:** la capacità disponibile può variare in base alla zona di disponibilità. Questo migliora la probabilità di raggiungere la capacità prefissata attingendo a più pool di capacità inutilizzati, oltre a offrire il costo più basso possibile.
- **Progetta avendo come obiettivo la continuità:** progetta carichi di lavoro stateless e tolleranti ai guasti, in modo che se una parte della capacità EC2 viene interrotta, non avrà alcun impatto sulla disponibilità o sulle prestazioni del carico di lavoro.
- Consigliamo di utilizzare istanze Spot in combinazione con piani on demand e Savings Plans/istanze riservate per massimizzare l'ottimizzazione dei costi del carico di lavoro insieme alle prestazioni.

Sconti a fronte di impegni - Savings Plans: AWS offre diversi modi per consentirti di ridurre i costi prenotando o impegnandoti a utilizzare una determinata quantità di risorse e ricevendo una tariffa scontata per le risorse. Savings Plan ti consente di prendere un impegno di spesa oraria per uno o tre anni e di ricevere prezzi scontati sulle risorse. Savings Plans ti offre sconti per servizi di elaborazione AWS quali EC2, Fargate e Lambda. Quando assumi l'impegno, paghi il relativo importo su base oraria, e l'utilizzo on demand alla tariffa scontata viene ridotto di conseguenza. Ad esempio, ti impegni per 50 USD all'ora e hai 150 USD all'ora di utilizzo on demand. Considerando i prezzi di Savings Plans, il tuo utilizzo specifico ha una percentuale di sconto pari al 50%. Pertanto, il tuo impegno di 50 USD copre 100 USD di utilizzo on demand. Pagherai 50 USD (di impegno) e 50 USD dell'utilizzo on demand rimanente.

I Compute Savings Plans sono i più flessibili e offrono uno sconto fino al 66%. Si applicano automaticamente a ogni zona di disponibilità, dimensione dell'istanza, famiglia di istanze, sistema operativo, tenancy, regione e servizio di calcolo.

Gli Instance Savings Plans hanno una minore flessibilità, ma offrono un tasso di sconto più elevato (fino al 72%). Si applicano automaticamente a ogni zona di disponibilità, dimensione dell'istanza, famiglia di istanze, sistema operativo e tenancy.

Sono disponibili tre opzioni di pagamento:

- **Nessun pagamento anticipato:** non è previsto alcun pagamento anticipato; pagherai quindi una tariffa oraria ridotta ogni mese per le ore totali di quel mese.
- **Pagamento anticipato parziale:** offre una percentuale di sconto più elevata rispetto a Nessun pagamento anticipato. Una parte dell'utilizzo viene pagata in anticipo; pagherai quindi una tariffa oraria ridotta ogni mese per le ore totali di quel mese.
- **Pagamento anticipato totale:** l'utilizzo per l'intero periodo viene pagato in anticipo e non verranno addebitati altri costi per il resto del termine di utilizzo che è coperto dall'impegno.

Per i tuoi carichi di lavoro puoi applicare qualsiasi combinazione di queste tre opzioni di acquisto.

I Savings Plans si applicano innanzitutto all'utilizzo nell'account in cui vengono acquistati, dalla percentuale di sconto più alta a quella più bassa, quindi si applicano all'utilizzo consolidato di tutti gli altri account, dalla percentuale di sconto più alta a quella più bassa.

Si consiglia di acquistare tutti i Savings Plans in un account senza utilizzo o risorse, ad esempio l'account master. In questo modo, il Savings Plan si applica alle tariffe di sconto più elevate per ogni tipo di utilizzo, consentendo di approfittare del massimo importo di sconto.

I carichi di lavoro e l'utilizzo solitamente variano nel corso del tempo. Pertanto, si consiglia di acquistare di volta in volta piccole quantità di Savings Plans. In questo modo puoi mantenere alti livelli di copertura per massimizzare gli sconti, e al contempo i piani soddisfano accuratamente e costantemente i requisiti del carico di lavoro e dell'organizzazione.

Non impostare una copertura prefissata nei tuoi account, per via della variabilità dello sconto che è possibile ottenere. Una bassa copertura non indica necessariamente un elevato risparmio potenziale. Potresti avere una copertura bassa nel tuo account, ma se il tuo utilizzo è costituito da istanze di piccole dimensioni, con un sistema operativo concesso in licenza, il potenziale risparmio potrebbe essere di pochi punti percentuali. Al contrario, valuta e monitora i risparmi potenziali disponibili nello strumento per i suggerimenti di Savings Plan. Valuta frequentemente i suggerimenti di Savings Plans in Cost Explorer (esegui analisi periodiche) e continua ad acquistare impegni finché i risparmi stimati non sono inferiori allo sconto richiesto per l'organizzazione. Ad esempio, valuta e monitora che i tuoi sconti potenziali rimangano al di sotto del 20%, e se superano tale soglia significa che è necessario effettuare un acquisto.

Monitora l'utilizzo e la copertura, ma solo per rilevare le modifiche. Non mirare a una specifica percentuale di utilizzo o di copertura, in quanto non è garantito che il risparmio sia proporzionale ad essa. Assicurati che un acquisto di Savings Plans comporti un aumento della copertura e, se ci sono riduzioni nella copertura o nell'utilizzo, assicurati che siano quantificate e note. Ad esempio, esegui la migrazione di una risorsa del carico di lavoro a un tipo di istanza più recente, riducendo l'utilizzo di un piano esistente, ma il vantaggio in termini di prestazioni supera la riduzione del risparmio.

Sconti a fronte di impegni - Istanze riservate/impegni: analogamente ai Savings Plans, le istanze riservate offrono sconti fino al 72% a fronte dell'impegno di eseguire una quantità minima di risorse. Le istanze riservate sono disponibili per RDS, Elasticsearch, ElastiCache, Amazon Redshift e DynamoDB. Amazon CloudFront e AWS Elemental MediaConvert offrono ulteriori sconti quando prendi impegni di utilizzo minimo. Le istanze riservate sono attualmente disponibili per EC2, tuttavia Savings Plans offre gli stessi livelli di sconto con maggiore flessibilità e senza spese di gestione.

Le istanze riservate offrono le stesse opzioni di prezzo (nessun pagamento anticipato, pagamento anticipato parziale e pagamento anticipato) e gli stessi termini di uno o tre anni.

Le istanze riservate possono essere acquistate in una regione o in una zona di disponibilità specifica. Quando vengono acquistate in una zona di disponibilità, forniscono una prenotazione di capacità.

EC2 dispone di istanze riservate modificabili, tuttavia, i Savings Plans devono essere utilizzati per tutte le istanze EC2 per via di una maggiore flessibilità e costi operativi ridotti.

Lo stesso processo e i parametri devono essere utilizzati per monitorare ed effettuare acquisti di istanze riservate. Si consiglia di non tenere traccia della copertura delle istanze riservate nei tuoi account. Inoltre, piuttosto che monitorare e valutare la percentuale di utilizzo, è consigliabile fare riferimento al report di utilizzo in Cost Explorer e utilizzare la colonna di risparmio netto nella tabella. Se il risparmio netto è un importo negativo significativamente elevato, è necessario intervenire per correggere l'istanza riservata inutilizzata.

Parco istanze EC2: Parco istanze EC2 ti consente di definire una capacità di elaborazione prefissata e poi specificare i tipi di istanza e l'equilibrio tra istanze on demand e Spot per il parco istanze. Parco istanze EC2 avvierà la combinazione di risorse più economica per soddisfare la capacità prevista.

Selezione geografica: quando progetti le tue soluzioni, una best practice da seguire è quella di cercare di posizionare le risorse di calcolo vicino agli utenti per offrire una latenza inferiore e una forte sovranità dei dati. Per i gruppi di pubblico globali dovresti usare più ubicazioni al fine di soddisfare queste esigenze. Dovresti selezionare la posizione geografica che ti consente di ridurre al minimo i costi.

L'infrastruttura 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 in una struttura separata.

Ciascuna regione AWS opera all'interno di condizioni di mercato locali, e la determinazione dei prezzi delle risorse è diversa in ciascuna regione. Scegli una regione specifica per gestire un componente o tutta la tua soluzione in modo da eseguirla al minor prezzo possibile a livello globale. Puoi utilizzare il Calcolatore di costo mensile AWS per stimare i costi del carico di lavoro in varie regioni.

Contratti di terze parti e prezzi: quando utilizzi soluzioni o servizi di terze parti nel cloud, è importante che le strutture dei prezzi siano allineate ai risultati dell'ottimizzazione dei costi. I prezzi devono essere adattati in base ai risultati e al valore che forniscono. Un esempio di questo è un software che contempla una percentuale del risparmio che fornisce, più risparmi (come risultato) e più ti addebita. Gli accordi che si adeguano in base alla fattura in genere non sono allineati all'ottimizzazione dei costi, a meno che non forniscano risultati per ogni parte della fattura specifica. Ad esempio, una soluzione che fornisce suggerimenti per EC2 e addebita una percentuale dell'intera fattura, aumenterà se utilizzi altri servizi per i quali non fornisce alcun vantaggio. Un altro esempio è un servizio gestito che viene addebitato a una percentuale del costo delle risorse che vengono gestite. Una dimensione di istanza più grande potrebbe non richiedere necessariamente un maggiore impegno di gestione, ma comporterà un addebito superiore. Assicurati che queste disposizioni tariffarie dei servizi includano un programma di ottimizzazione dei costi o funzionalità di servizio volte a migliorare l'efficienza.

Piano per il trasferimento dei dati

Un vantaggio del cloud è che si tratta di un servizio di rete gestito. Non è più necessario gestire e utilizzare una flotta di switch, router e altre apparecchiature di rete associate. Le risorse di rete nel cloud vengono utilizzate e pagate nello stesso modo in cui si paga per CPU e storage, pagando unicamente in base all'uso. Per ottimizzare i costi nel cloud, è necessario un uso efficiente delle risorse di rete.

Esegui la modellazione del trasferimento dei dati: comprendi dove avviene il trasferimento dei dati nel carico di lavoro, il costo del trasferimento e il relativo vantaggio. In questo modo puoi prendere una decisione consapevole quando si tratta di modificare o accettare una decisione relativa all'architettura. Ad esempio, potresti disporre di una configurazione con più zone di disponibilità dove replichi i dati tra le varie zone di disponibilità. Puoi modellare il costo della struttura e decidere che questo sia un costo accettabile (simile a quello del calcolo e dello storage in entrambe le zone di disponibilità) per ottenere l'affidabilità e la resilienza richieste.

Modella i costi in base a livelli differenti di utilizzo. L'utilizzo del carico di lavoro può cambiare nel corso del tempo e servizi differenti possono risultare più convenienti a livelli differenti.

Utilizza AWS Cost Explorer o il report CUR (Cost and Usage Report) per comprendere e modellare i costi di trasferimento dei dati. Configura un proof of concept (PoC) o testa

il carico di lavoro ed esegui un test con un carico simulato realistico. Puoi modellare i costi in base alle diverse esigenze di carico di lavoro.

Ottimizza il trasferimento dei dati: una progettazione basata sul trasferimento dei dati ti assicura la massima riduzione dei costi di trasferimento dei dati. Potrebbe implicare l'uso di reti di distribuzione di contenuti per posizionare i dati vicino agli utenti, oppure l'uso di collegamenti di rete dedicati dalle tue sedi ad AWS. Puoi anche utilizzare l'ottimizzazione WAN e l'ottimizzazione delle applicazioni per ridurre la quantità di dati trasferiti tra i componenti.

Seleziona servizi per ridurre il costo del trasferimento dei dati: Amazon CloudFront è una rete globale di distribuzione di contenuti che trasferisce i dati con una latenza ridotta e una velocità di trasferimento elevata. Cattura i dati nelle edge location di tutto il mondo, riducendo così il carico sulle tue risorse. Utilizzando CloudFront puoi ridurre l'impegno amministrativo legato alla distribuzione dei contenuti per numeri elevati di utenti a livello globale, con una latenza minima.

AWS Direct Connect ti consente di creare una connessione di rete dedicata ad AWS. In questo modo puoi ridurre i costi di rete, aumentare la larghezza di banda e offrire un'esperienza di rete più costante rispetto alle connessioni Internet.

AWS VPN consente di stabilire una connessione sicura e privata tra la rete privata e la rete globale AWS. È ideale per piccoli uffici o partner aziendali perché offre una connettività semplice e rapida, ed è un servizio completamente gestito ed elastico.

Gli endpoint VPC consentono la connettività tra i servizi AWS su reti private e possono essere utilizzati per ridurre i costi di trasferimento di dati pubblici e dei gateway NAT. Gli endpoint VPC del gateway non hanno tariffe orarie e supportano Amazon S3 e Amazon DynamoDB. Gliendpoint VPC dell'interfaccia sono forniti da AWS PrivateLink e prevedono una tariffa oraria e un costo di utilizzo per GB.

Risorse

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

- AWS Managed Services: video sul percorso di trasformazione aziendale
- Analisi dei costi con Cost Explorer
- Accesso alle raccomandazioni di istanza riservata
- Nozioni di base sulle raccomandazioni per il dimensionamento appropriato
- Best practice per le istanze Spot
- Parco istanze Spot

- Come funzionano le istanze riservate
- Infrastruttura globale AWS
- Consulente istanze Spot
- Well-Architected Labs - Risorse a costi contenuti

Gestione delle risorse di domanda e offerta

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

I vantaggi economici della fornitura just in time dovrebbero essere bilanciati rispetto alla richiesta di provisioning, per rispondere a risorse insufficienti, elevata disponibilità e tempi di provisioning. A seconda del fatto che la domanda sia fissa o variabile, prevedi la creazione di parametri e automazione, che faranno in modo che la gestione del tuo ambiente sia minima, anche quando ti espandi. Quando modifichi la domanda, dovresti conoscere il ritardo massimo e accettabile che il carico di lavoro può consentire.

In AWS, puoi utilizzare diversi approcci per gestire la domanda e fornire risorse. Le seguenti sezioni descrivono come utilizzare queste strategie:

- Analizza il carico di lavoro
- Gestisci la domanda
- Fornitura basata sulla domanda
- Fornitura basata sul tempo

Analizza il carico di lavoro: è importante per sapere i requisiti del carico di lavoro. I requisiti dell'organizzazione devono indicare i tempi di risposta del carico di lavoro per le richieste. Il tempo di risposta può essere utilizzato per determinare se la domanda è gestita o se l'offerta di risorse cambierà per soddisfare la domanda.

L'analisi deve includere la prevedibilità e la ripetibilità della domanda, la velocità di variazione della domanda e la quantità di variazione della domanda. Assicurati che l'analisi venga eseguita per un periodo sufficientemente lungo da incorporare qualsiasi variazione stagionale, ad esempio l'elaborazione di fine mese o i picchi legati alle festività.

Assicurati che le attività di analisi siano commensurate ai potenziali vantaggi dell'implementazione del dimensionamento. Osserva il costo totale previsto del componente, ed eventuali aumenti o riduzioni di utilizzo e costi durante il ciclo di vita del carico di lavoro.

Per eseguire un'analisi visiva della domanda dei carichi di lavoro puoi utilizzare AWS Cost Explorer o Amazon QuickSight con il CUR o i log delle applicazioni.

Gestisci la domanda

Gestisci la domanda - throttling: se l'origine della richiesta dispone di funzionalità di ripetizione dei tentativi, è possibile implementare il throttling. Il throttling indica alla sorgente che, se non è in grado di soddisfare la richiesta all'ora corrente, dovrebbe riprovare più tardi. La sorgente attenderà per un determinato periodo di tempo e quindi riproverà a effettuare la richiesta. L'implementazione del throttling ha il vantaggio di limitare la quantità massima di risorse e i costi del carico di lavoro. Per implementare il throttling in AWS, puoi utilizzare Amazon API Gateway. Consulta il whitepaper sul principio dell'affidabilità del canone di architettura per ulteriori dettagli sull'implementazione del throttling.

Gestisci la domanda - basata sul buffering: analogamente al throttling, il buffering rinvia l'elaborazione delle richieste, consentendo alle applicazioni eseguite a velocità diverse di comunicare in modo efficace. Un approccio basato sul buffering impiega una coda per l'accettazione dei messaggi (unità di lavoro) dai produttori. I messaggi vengono letti ed elaborati dai consumatori e ciò consente ai messaggi di essere eseguiti alla velocità che soddisfa i requisiti aziendali del consumatore stesso. Non devi preoccuparti del fatto che i produttori debbano gestire i problemi legati al throttling, come la durabilità e la contropressione dei dati (per cui i produttori rallentano per adeguarsi alla velocità dei consumatori).

Su AWS, puoi scegliere fra più servizi per l'implementazione di una strategia di buffering. Amazon SQS è un servizio gestito che offre code che consentono a un singolo consumatore di leggere singoli messaggi. Amazon Kinesis offre un flusso che consente a più consumatori di leggere gli stessi messaggi.

Durante la progettazione con un approccio basato sul buffering, assicurati di progettare il carico di lavoro per soddisfare la richiesta nel tempo richiesto e di essere in grado di gestire le richieste duplicate per il lavoro.

Fornitura dinamica

Fornitura basata sulla domanda: sfrutta l'elasticità del cloud per fornire risorse in grado di soddisfare le mutevoli esigenze. Sfrutta API o funzionalità dei servizi per modificare in modo programmatico e dinamico la quantità di risorse del cloud nella tua architettura. Ciò ti consente di dimensionare i componenti nella tua architettura e aumentare automaticamente il numero di risorse durante i picchi di domanda per mantenere le prestazioni, nonché diminuire la capacità quando la domanda cala in modo da ridurre i costi.

Auto Scaling ti aiuta a regolare la capacità per mantenere prestazioni stabili e prevedibili al minor costo possibile. Si tratta di un servizio completamente gestito e gratuito che si integra

con istanze Amazon EC2 e parchi istanze Spot, Amazon ECS, Amazon DynamoDB e Amazon Aurora.

Auto Scaling fornisce il rilevamento automatico delle risorse per aiutare a trovare risorse nel carico di lavoro che possono essere configurate, dispone di strategie di ridimensionamento integrate per ottimizzare le prestazioni, i costi o un equilibrio tra i due e fornisce il ridimensionamento predittivo per aiutare a risolvere i picchi ricorrenti con regolarità.

Auto Scaling può implementare il ridimensionamento manuale, programmato o basato sulla domanda, e puoi anche utilizzare parametri e allarmi di Amazon CloudWatch per attivare eventi di ridimensionamento per il tuo carico di lavoro. I parametri tipici possono essere parametri standard di Amazon EC2, ad esempio l'utilizzo della CPU, il throughput di rete e la latenza di richiesta/risposta osservata da ELB. Quando possibile, è consigliabile utilizzare un parametro indicativo dell'esperienza del cliente, in genere si tratta di un parametro personalizzato che potrebbe avere origine dal codice dell'applicazione all'interno del carico di lavoro.

Quando prevedi una strategia basata sulla domanda in un progetto, tieni presenti due considerazioni principali. In primo luogo, devi capire con quale velocità è necessario predisporre le nuove risorse. In secondo luogo, devi capire che la dimensione del margine tra domanda e risorse fornite cambierà. Devi prepararti ad affrontare le variazioni nella domanda, nonché le risorse insufficienti.

Elastic Load Balancing (ELB) consente di ricalibrare le risorse distribuendo la domanda su più risorse. Quando si implementano più risorse, è necessario aggiungerle al sistema di bilanciamento del carico per rispondere alla domanda. AWS ELB supporta istanze EC2, container, indirizzi IP e funzioni Lambda.

Fornitura basata sul tempo: una strategia basata sul tempo allinea la capacità delle risorse alla domanda, che è prevedibile o ben definita nel tempo. In genere questa strategia non dipende dai livelli di utilizzo delle risorse. Una strategia basata sul tempo assicura che le risorse siano disponibili nel momento esatto in cui vengono richieste e possano essere fornite senza ritardi dovuti alle procedure di avvio e ai controlli di sistema o di coerenza. Attraverso una strategia basata sul tempo puoi fornire risorse aggiuntive o incrementare la capacità nei periodi più intensi.

Puoi utilizzare Auto Scaling pianificato per implementare un approccio basato sul tempo. I carichi di lavoro possono essere programmati per eseguire il dimensionamento in determinati momenti (ad esempio, all'inizio dell'orario di lavoro), garantendo quindi la disponibilità delle risorse all'arrivo degli utenti on demand.

Puoi anche sfruttare API e SDK di AWS e AWS CloudFormation per predisporre e disattivare automaticamente interi ambienti quando ne hai bisogno. Questa strategia risulta particolarmente adatta per gli ambienti di sviluppo o di prova che operano solo in determinati orari di lavoro o periodi di tempo.

Puoi usare le API per dimensionare le risorse all'interno di un ambiente (dimensionamento verticale). Ad esempio, potresti dimensionare verticalmente un carico di lavoro di produzione modificando la dimensione o la classe dell'istanza. Ciò è possibile interrompendo e avviando l'istanza e selezionando una dimensione o classe diversa. Questa tecnica può essere applicata anche ad altre risorse, come gli Elastic Volumes EBS, che possono essere modificati per aumentarne le dimensioni, regolarne le prestazioni (IOPS) o cambiare il tipo di volume durante l'utilizzo.

Quando prevedi una strategia basata sul tempo in un progetto, tieni presenti due considerazioni principali. In primo luogo, che livello di coerenza presenta il modello di utilizzo? In secondo luogo, qual è l'impatto se il modello cambia? Puoi migliorare l'accuratezza delle previsioni monitorando i tuoi carichi di lavoro e utilizzando la business intelligence. Se noti cambiamenti significativi nel modello di utilizzo, puoi modificare i tempi per assicurarti che la copertura sia fornita.

Fornitura dinamica: puoi utilizzare AWS Auto Scaling oppure incorporare il dimensionamento nel codice utilizzando API o SDK AWS. Ciò riduce i costi complessivi del carico di lavoro rimuovendo i costi operativi dall'apportare manualmente modifiche al tuo ambiente e può essere eseguito molto più rapidamente. In questo modo è possibile garantire che le risorse del carico di lavoro soddisfino al meglio la domanda, in qualsiasi momento.

Risorse

Consulta le seguenti risorse per ulteriori informazioni sulle best practice di AWS per gestire la domanda e fornire le risorse.

- Throttling di API Gateway
- Nozioni di base di Amazon SQS
- Nozioni di base su Amazon EC2 Auto Scaling

Apporta ottimizzazioni nel corso del tempo

In AWS puoi apportare ottimizzazioni nel corso del tempo esaminando nuovi servizi e implementandoli nel carico di lavoro.

Valuta e implementa nuovi servizi

Poiché AWS rilascia nuovi servizi e funzionalità, è consigliabile rivedere le decisioni correnti sull'architettura per garantire che continuino a essere le più convenienti. Man mano che le tue esigenze cambiano, disattiva con fermezza risorse, componenti e carichi di lavoro di cui non hai più bisogno. Per apportare ottimizzazioni nel corso del tempo, tieni a mente quanto segue:

- Sviluppa un processo di revisione del carico di lavoro
- Valuta e implementazione i servizi

Sviluppa un processo di revisione del carico di lavoro: per assicurarti di avere sempre il carico di lavoro più conveniente, devi valutare regolarmente il carico di lavoro per sapere se ci sono opportunità di implementare nuovi servizi, funzionalità e componenti. Per garantire costi complessivi ridotti, il processo deve essere proporzionale al potenziale risparmio. Ad esempio, I carichi di lavoro che rappresentano il 50% della spesa complessiva devono essere esaminati con maggiore regolarità e più nel dettaglio rispetto ai carichi di lavoro che rappresentano il 5% della spesa complessiva. Prendi in considerazione qualsiasi fattore esterno o volatilità. Se il carico di lavoro serve una determinata area geografica o un segmento di mercato e viene previsto un cambiamento in tale area, revisioni più frequenti possono portare a risparmi sui costi. Un altro fattore in fase di revisione è rappresentato dall'impegno necessario per implementare le modifiche. Se i test e la convalida delle modifiche comportassero costi significativi, le revisioni dovrebbero essere meno frequenti.

Prendi in considerazione il costo nel lungo termine della manutenzione di componenti e risorse obsoleti e legacy, e dell'impossibilità di implementare in essi nuove funzionalità. L'attuale costo del test e della convalida potrebbe superare il vantaggio auspicato. Tuttavia, nel corso del tempo, il costo di apportare modifiche potrebbe crescere in modo significativo all'aumentare del divario tra il carico di lavoro e le tecnologie attuali, generando costi ancora maggiori. Ad esempio, il costo del passaggio a un nuovo linguaggio di programmazione potrebbe attualmente non risultare conveniente. Tuttavia, nel giro di cinque anni, il costo del personale qualificato per tale linguaggio potrebbe aumentare e, a causa dell'aumento del carico di lavoro, potresti dover trasferire un sistema ancora più grande al nuovo linguaggio, richiedendo sforzi ancora maggiori rispetto a prima.

Suddividi il carico di lavoro in componenti, assegna il costo del componente (una stima è sufficiente) e quindi elenca i fattori (ad esempio, impegno richiesto e mercati esterni) accanto a ciascun componente. Utilizza questi indicatori per determinare una frequenza di revisione per ogni carico di lavoro. Ad esempio, potresti avere i server web come un costo elevato, con un impegno di modifica ridotto e fattori esterni elevati, e da questo potrebbe derivare un'alta frequenza di revisione. Un database centrale può essere un costo medio, con un impegno di modifica elevato e un basso fattore esterno, e da questo potrebbe derivare una frequenza di revisione media.

Esamina il carico di lavoro e implementa i servizi: per ottenere i vantaggi offerti dai nuovi servizi e funzionalità AWS, devi eseguire il processo di revisione sui carichi di lavoro e implementare nuovi servizi e funzionalità in base alle esigenze. Ad esempio, è possibile esaminare i carichi di lavoro e sostituire il componente di messaggistica con Amazon Simple Email Service (SES). Ciò elimina il costo di gestione e manutenzione di un parco istanze, fornendo al contempo tutte le funzionalità a un costo ridotto.

Conclusioni

L'ottimizzazione dei costi e la gestione finanziaria nel cloud rappresentano un impegno continuo. È auspicabile collaborare con i team finanziari e tecnologici, rivedere l'approccio in termini di architettura e aggiornare la selezione dei componenti con regolarità.

AWS si impegna per aiutarti a ridurre al minimo i costi, consentendoti al tempo stesso di creare distribuzioni fortemente resilienti, reattive e adattive. Per ottimizzare davvero il costo della tua distribuzione, sfrutta gli strumenti, le tecniche e le best practice di cui abbiamo parlato in questo documento.

Collaboratori

Hanno contribuito a questo documento:

- Philip Fitzsimons, Sr Manager Well-Architected, Amazon Web Services
- Nathan Besh, Cost Lead Well-Architected, Amazon Web Services
- Levon Stepanian,, Amazon Web Services
- Keith Jarrett, Business Development Lead – Cost Optimization
- PT Ng, Commercial Architect, Amazon Web Services
- Arthur Basbaum, Business Developer Manager, Amazon Web Services
- Jarman Hauser, Commercial Architect, Amazon Web Services

Approfondimenti

Per ulteriori informazioni, consulta:

- Canone di architettura AWS

Revisioni del documento

Data	Descrizione
Aprile 2020	Versione aggiornata per integrare CFM, nuovi servizi e integrazione anche con Well-Architected.
Luglio 2018	Aggiornamento finalizzato a rispecchiare le modifiche apportate ad AWS e integrare le nozioni apprese grazie
Novembre 2017	Aggiornamento finalizzato a rispecchiare le modifiche apportate ad AWS e integrare le nozioni apprese grazie alle revisioni con i clienti.
Novembre 2016	Prima pubblicazione