Scalabilità dell'infrastruttura Amazon EKS per ottimizzare elaborazione, carichi di lavoro e prestazioni di rete

AWS Guida prescrittiva

AWS Guida prescrittiva: Scalabilità dell'infrastruttura Amazon EKS per ottimizzare elaborazione, carichi di lavoro e prestazioni di rete

Table of Contents

Scalabilità dell'infrastruttura Amazon EKS per ottimizzare elaborazione, carichi di lavoro e prestazioni di rete

Aniket Dekate, Aniket Kurzadkar e Ishwar Chauthaiwale, Amazon Web Services ()AWS

Novembre 2024 (storia del documento)

Amazon Elastic Kubernetes Service (Amazon EKS) è un servizio Kubernetes gestito. Con Amazon EKS, puoi eseguire i pod Kubernetes in un ambiente cloud containerizzato senza dover installare e utilizzare il tuo piano di controllo. Con la AWS gestione del piano di controllo, Amazon EKS riduce la gestione operativa dell'organizzazione. Altri vantaggi dell'utilizzo di Amazon EKS includono scalabilità, affidabilità e sicurezza nell'ambiente cloud.

Questa guida è progettata per aiutare le organizzazioni a ottimizzare la propria infrastruttura Amazon EKS nelle seguenti aree:

- La scalabilità di calcolo è un componente fondamentale per le prestazioni delle applicazioni in un ambiente Kubernetes dinamico:
  - Allocazione efficiente delle risorse: scopri le tecniche per allocare le risorse elaborate in modo dinamico per soddisfare le diverse esigenze.
  - Strumenti di automazione: ottieni una panoramica degli strumenti e dei servizi che automatizzano la scalabilità di elaborazione, riducendo la necessità di interventi manuali.
- La scalabilità del carico di lavoro aiuta a garantire che le applicazioni siano in grado di gestire carichi di lavoro diversi senza un peggioramento delle prestazioni:
  - HORIZONTAL POD AUTOSCALER: dai un'occhiata approfondita a come un HPA aiuta a scalare i carichi di lavoro in base a metriche in tempo reale.
  - Cluster Proportional Autoscaler: scopri come CPA ridimensiona automaticamente e mantiene una relazione proporzionale tra nodi e repliche, aumentando o diminuendo i carichi di lavoro al variare delle dimensioni del cluster.
  - Scalabilità basata sugli eventi: esamina le strategie per scalare le applicazioni in risposta a eventi o trigger specifici.
- La scalabilità della rete aiuta a mantenere una comunicazione senza interruzioni tra i servizi e un flusso di dati efficiente in ambienti dinamici:
  - Plugin Amazon VPC CNI: scopri come il plug-in VPC CNI consente una rete scalabile all'interno dei cluster Amazon EKS.

  - Rete personalizzata: verifica la gestione degli indirizzi IP e la segregazione del traffico di rete sui cluster Amazon EKS.
  - Delega dei prefissi: ottieni una panoramica sulla semplificazione della gestione degli IP in cluster Amazon EKS di grandi dimensioni e scalabili.
  - Amazon VPC Lattice: ottieni una panoramica di come VPC Lattice può gestire il cross-VPC e il networking per una scalabilità perfetta. service-to-service
- L'ottimizzazione dei costi aiuta le aziende a vedere dove vengono spese le proprie risorse e ad assegnare in modo appropriato le spese ai reparti o ai progetti:
  - Risorse di dimensioni adeguate: prendi in considerazione le tecniche per dimensionare le risorse cloud in modo appropriato per il carico di lavoro.
  - Monitoraggio e controllo dei costi: rivedi gli strumenti e le migliori pratiche per tracciare e ottimizzare le spese relative al cloud.

Ogni sezione si concentra su obiettivi particolari necessari per creare un ambiente cloud affidabile, efficace e conveniente.

Obiettivi

Questa guida può aiutare te e la tua organizzazione a raggiungere i seguenti obiettivi aziendali:

- Maggiore efficienza delle risorse: ottieni un utilizzo ottimale delle risorse scalando dinamicamente l'elaborazione, i carichi di lavoro e le risorse di rete in base alle richieste in tempo reale.

  Questo obiettivo sottolinea l'importanza di aumentare e ridurre le risorse in risposta ai modelli di utilizzo effettivi. Strumenti come le scale automatiche a pod orizzontali e il plug-in Amazon VPC CNI aiutano le organizzazioni a utilizzare solo le risorse di cui hanno bisogno, riducendo al minimo gli sprechi e massimizzando le prestazioni.

- Prestazioni delle applicazioni migliorate: mantieni alte le prestazioni e la reattività delle applicazioni, anche in presenza di carichi di lavoro e modelli di traffico fluttuanti.

  Questo obiettivo si concentra su strategie volte a garantire che le applicazioni siano in grado di gestire picchi di traffico e carichi di lavoro pesanti senza compromettere le prestazioni. Tecniche come la scalabilità del carico di lavoro basata sugli eventi, l'allocazione efficiente dell'elaborazione e le architetture di rete scalabili sono fondamentali per raggiungere questo obiettivo.

- Scalabilità senza interruzioni: consente una scalabilità fluida dei componenti dell'infrastruttura, consentendo una crescita e un adattamento senza sforzo alle mutevoli esigenze aziendali.

  La scalabilità perfetta è fondamentale per le organizzazioni che prevedono una crescita o registrano livelli di traffico variabili. Questo obiettivo risponde all'importanza dell'implementazione di soluzioni scalabili tra risorse di elaborazione, carico di lavoro e rete, in modo che la scalabilità possa essere automatica, efficiente e trasparente.

- Ottimizzazione dei costi: riduci al minimo i costi del cloud mantenendo o migliorando le prestazioni e la scalabilità.

  L'ottimizzazione dei costi può comprendere la riduzione delle spese, ad esempio il corretto dimensionamento delle risorse, l'utilizzo di soluzioni di scalabilità convenienti e il monitoraggio della spesa. L'obiettivo è bilanciare i risparmi sui costi con la necessità di prestazioni e scalabilità elevate.

Scalabilità del calcolo

La scalabilità del calcolo è un componente fondamentale per le prestazioni delle applicazioni in un ambiente Kubernetes dinamico. Kubernetes riduce gli sprechi attraverso la regolazione dinamica delle risorse di elaborazione (come CPU e memoria) in risposta alla domanda in tempo reale. Questa funzionalità aiuta a evitare un approvvigionamento eccessivo o insufficiente, il che può anche far risparmiare sui costi operativi. Kubernetes elimina efficacemente la necessità di interventi manuali consentendo all'infrastruttura di scalare automaticamente verso l'alto nelle ore di punta e verso il basso durante i periodi non di punta.

La scalabilità complessiva dell'elaborazione di Kubernetes automatizza il processo di scalabilità, il che aumenta la flessibilità e la scalabilità dell'applicazione e ne migliora il comportamento di tolleranza ai guasti. In definitiva, le funzionalità di Kubernetes migliorano l'eccellenza operativa e la produttività.

Questa sezione descrive i seguenti tipi di scalabilità del calcolo:

- Cluster Autoscaler
- Cluster Autoscaler con over-provisioning
- Karpenter

Grappolo AutoScaler

A seconda delle esigenze dei pod, lo strumento Cluster Autoscaler modifica automaticamente le dimensioni aggiungendo nodi quando necessario o rimuovendo nodi quando non sono necessari e sono sottoutilizzati.

Considerate lo strumento Cluster Autoscaler come una soluzione di scalabilità per carichi di lavoro in cui la domanda aumenta gradualmente e la latenza nella scalabilità non è un problema importante.

Lo strumento Cluster Autoscaler offre le seguenti funzionalità chiave:

- Scalabilità: aumenta e riduce i nodi in modo dinamico in risposta alle effettive richieste di risorse.
- Pianificazione dei pod: aiuta a garantire che ogni pod sia operativo e disponga delle risorse necessarie per funzionare, prevenendo la scarsità di risorse.
- Efficienza in termini di costi: elimina le spese inutili legate alla gestione dei nodi sottoutilizzati eliminandoli.

Cluster Autoscaler con over-provisioning

Cluster Autoscaler con funzionalità di over-provisioning analoghe a Cluster Autoscaler, in quanto distribuisce i nodi in modo efficiente e consente di risparmiare tempo eseguendo pod a bassa priorità sui nodi. Con questa tecnica, il traffico viene reindirizzato verso questi pod in risposta a picchi improvvisi della domanda, permettendo all'applicazione di continuare a funzionare senza interruzioni.

Cluster Autoscaler con over-provisioning offre le funzionalità dei dummy pod che possono essere utilizzati per implementare ed eseguire facilmente i nodi quando il carico di lavoro è molto elevato, la latenza non è necessaria e la scalabilità deve essere rapida.

Cluster Autoscaler con over-provisioning offre le seguenti funzionalità chiave:

- Migliore reattività: rendendo costantemente accessibile la capacità in eccesso, è necessario meno tempo per scalare il cluster in risposta ai picchi di domanda.
- Riservazione delle risorse: la gestione di picchi di traffico imprevisti aiuta efficacemente una corretta gestione con tempi di inattività ridotti.
- Scalabilità fluida: la riduzione al minimo dei ritardi nell'allocazione delle risorse facilita un processo di scalabilità più semplice.

Karpenter

Karpenter for Kubernetes supera il tradizionale strumento Cluster Autoscaler in termini di open source, prestazioni e personalizzazione. Con Karpenter, puoi avviare automaticamente solo le risorse di elaborazione necessarie per gestire le richieste del cluster in tempo reale. Karpenter è progettato per offrire una scalabilità più efficiente e reattiva.

Le applicazioni con carichi di lavoro estremamente variabili o complessi, in cui sono essenziali decisioni rapide sulla scalabilità, traggono grandi vantaggi dall'uso di Karpenter. Si integra con AWS, offrendo una migliore implementazione e ottimizzazione della selezione dei nodi.

Karpenter include le seguenti funzionalità chiave:

- Provisioning dinamico: Karpenter fornisce le istanze e le dimensioni giuste per lo scopo e fornisce nuovi nodi in modo dinamico in base ai requisiti particolari dei pod.
- Pianificazione avanzata: utilizzando un posizionamento intelligente dei pod, Karpenter organizza i nodi in modo tale che risorse come GPU, CPU, memoria e storage vengano utilizzate nel modo più efficace possibile.

- Scalabilità rapida: Karpenter è in grado di scalare rapidamente, reagendo spesso in pochi secondi. Questa reattività è utile in caso di traffico improvviso o quando il carico di lavoro richiede una scalabilità immediata
- Efficienza in termini di costi: scegliendo con cura l'istanza più efficace, è possibile ridurre i costi operativi e sfruttare le alternative aggiuntive a basso costo offerte da AWS, come le istanze On-Demand, le istanze Spot e le istanze riservate.

Scalabilità del carico di lavoro

La scalabilità del carico di lavoro in Kubernetes è essenziale per mantenere le prestazioni delle applicazioni e l'efficienza delle risorse in ambienti dinamici. La scalabilità aiuta a garantire che le applicazioni siano in grado di gestire carichi di lavoro diversi senza un peggioramento delle prestazioni. Kubernetes offre la possibilità di aumentare o ridurre automaticamente le risorse in base a metriche in tempo reale, consentendo alle organizzazioni di rispondere rapidamente ai cambiamenti del traffico. Questa elasticità non solo migliora l'esperienza utente, ma ottimizza anche l'utilizzo delle risorse, contribuendo a ridurre al minimo i costi associati a risorse sottoutilizzate o sovrafornite.

Inoltre, un'efficace scalabilità del carico di lavoro supporta l'elevata disponibilità, garantendo che le applicazioni rimangano reattive anche nei periodi di picco della domanda. La scalabilità del carico di lavoro in Kubernetes consente alle organizzazioni di utilizzare meglio le risorse cloud adattando dinamicamente la capacità per soddisfare le esigenze attuali.

Questa sezione illustra i seguenti tipi di scalabilità del carico di lavoro:

- Pod Autoscaler orizzontale
- Cluster Proporional Autoscaler
- Autoscaler basato su eventi basato su Kubernetes

Horizontal Pod Autoscaler

L'Horizontal Pod Autoscaler (HPA) è una funzionalità di Kubernetes che regola automaticamente il numero di repliche dei pod in una distribuzione, un controller di replica o un set stateful, in base all'utilizzo della CPU osservato o ad altre metriche selezionate. L'HPA assicura che le applicazioni possano gestire livelli di traffico e carico di lavoro fluttuanti senza la necessità di un intervento manuale. L'HPA offre un mezzo per preservare prestazioni ottimali facendo al contempo un uso efficace delle risorse disponibili.

In contesti in cui la domanda degli utenti potrebbe variare considerevolmente nel tempo, le app Web, i microservizi e APIs l'HPA sono particolarmente utili.

L'Horizontal Pod Autoscaler offre le seguenti funzionalità chiave:

- Scalabilità automatica: HPA aumenta o diminuisce automaticamente il numero di repliche dei pod in risposta a metriche in tempo reale, garantendo la scalabilità delle applicazioni per soddisfare la domanda degli utenti.
- Decisioni basate su metriche: per impostazione predefinita, la scalabilità HPA si basa sull'utilizzo della CPU. Tuttavia, può anche utilizzare metriche personalizzate, come l'utilizzo della memoria o metriche specifiche dell'applicazione, consentendo strategie di scalabilità più personalizzate.
- Parametri configurabili: è possibile scegliere il numero minimo e massimo di repliche e le percentuali di utilizzo desiderate, in modo da poter decidere la severità del ridimensionamento.
- Integrazione con Kubernetes: per monitorare e modificare le risorse, HPA collabora con altri elementi dell'ecosistema Kubernetes, tra cui Metrics Server, l'API Kubernetes e gli adattatori di metrica personalizzati.
- Migliore utilizzo delle risorse: HPA aiuta a garantire che le risorse vengano utilizzate in modo efficace, riducendo i costi e migliorando le prestazioni, modificando dinamicamente il numero di pod.

Cluster Proportional Autoscaler

Il Cluster Proportional Autoscaler (CPA) è un componente Kubernetes progettato per regolare automaticamente il numero di repliche dei pod in un cluster in base al numero di nodi disponibili. A differenza degli autoscaler tradizionali che scalano in base a metriche di utilizzo delle risorse (come CPU e memoria), CPA ridimensiona i carichi di lavoro in proporzione alla dimensione del cluster stesso.

Questo approccio è particolarmente utile per le applicazioni che devono mantenere un certo livello di ridondanza o disponibilità rispetto alle dimensioni del cluster, come CoredNS e altri servizi di infrastruttura. Alcuni dei principali casi d'uso del CPA includono:

- Sovra-approvvigionamento
- Scalabilità orizzontale dei servizi della piattaforma di base
- Scalabilità orizzontale dei carichi di lavoro perché CPA non richiede un server di metrica o un adattatore Prometheus

Automatizzando il processo di scalabilità, CPA aiuta le aziende a mantenere una distribuzione equilibrata del carico di lavoro, aumentare l'efficienza delle risorse e garantire che le applicazioni siano adeguatamente fornite per soddisfare la domanda degli utenti.

Cluster Proportional Autoscaler offre le seguenti funzionalità chiave:

- Scalabilità basata sui nodi: CPA ridimensiona le repliche in base al numero di nodi del cluster che è possibile pianificare, consentendo alle applicazioni di espandersi o contrarsi in proporzione alla dimensione del cluster.
- Adattamento proporzionato: per garantire che l'applicazione possa scalare in base alle variazioni delle dimensioni del cluster, l'autoscaler stabilisce una relazione proporzionata tra il numero di nodi e il numero di repliche. Questa relazione viene utilizzata per calcolare il numero desiderato di repliche per un carico di lavoro.
- Integrazione con i componenti Kubernetes: CPA funziona con componenti Kubernetes standard come Horizontal Pod Autoscaler (HPA), ma si concentra specificamente sul numero di nodi piuttosto che sulle metriche di utilizzo delle risorse. Questa integrazione consente una strategia di scalabilità più completa.
- Client API Golang: per monitorare il numero di nodi e i relativi core disponibili, CPA utilizza i client API Golang che vengono eseguiti all'interno dei pod e comunicano con il server API Kubernetes.
- Parametri configurabili: utilizzando a`ConfigMap`, gli utenti possono impostare soglie e parametri di scala utilizzati da CPA per modificarne il comportamento e assicurarsi che segua il piano di scalabilità previsto.

Event-Driven Autoscaler basato su Kubernetes

Event Driven Autoscaler (KEDA) basato su Kubernetes è un progetto open source che consente la scalabilità dei carichi di lavoro Kubernetes in base al numero di eventi da elaborare. KEDA migliora la scalabilità delle applicazioni consentendo loro di rispondere in modo dinamico a carichi di lavoro diversi, in particolare quelli basati sugli eventi.

Automatizzando il processo di scalabilità in base agli eventi, KEDA aiuta le organizzazioni a ottimizzare l'utilizzo delle risorse, migliorare le prestazioni delle applicazioni e ridurre i costi associati all'over-provisioning. Questo approccio è particolarmente utile per le applicazioni che presentano modelli di traffico diversi, come microservizi, funzioni serverless e sistemi di elaborazione dati in tempo reale.

KEDA offre le seguenti funzionalità chiave:

- Scalabilità basata sugli eventi: KEDA consente di definire regole di ridimensionamento basate su fonti di eventi esterne, come code di messaggi, richieste HTTP o metriche personalizzate. Questa

funzionalità aiuta a garantire la scalabilità delle applicazioni in risposta alla domanda in tempo reale.

- Componente leggero: KEDA è un componente leggero e monouso che non richiede molta configurazione o sovraccarico per essere facilmente integrato nei cluster Kubernetes esistenti.
- Integrazione con Kubernetes: KEDA estende le funzionalità dei componenti nativi di Kubernetes, come Horizontal Pod Autoscaler (HPA). KEDA aggiunge funzionalità di scalabilità basate sugli eventi a questi componenti, migliorandoli anziché sostituendoli.
- Supporto per più fonti di eventi: KEDA è compatibile con un'ampia gamma di sorgenti di eventi, tra cui piattaforme di messaggistica popolari come RabbitMQ, Apache Kafka e altre. Grazie a questa adattabilità, puoi personalizzare la scalabilità per adattarla alla tua architettura unica basata sugli eventi.
- Scaler personalizzati: utilizzando gli scaler personalizzati, è possibile designare metriche specifiche che KEDA può utilizzare per avviare azioni di scalabilità in risposta a logiche o requisiti aziendali specifici.
- Configurazione dichiarativa: in linea con i principi di Kubernetes, puoi utilizzare KEDA per descrivere il comportamento di scalabilità in modo dichiarativo utilizzando risorse personalizzate di Kubernetes per definire come dovrebbe avvenire la scalabilità.

Scalabilità della rete

La scalabilità della rete in Kubernetes è fondamentale per mantenere una comunicazione senza interruzioni tra i servizi e supportare un flusso di dati efficiente in ambienti dinamici. La scalabilità dell'infrastruttura di rete aiuta a garantire che il cluster sia in grado di gestire diversi livelli di traffico senza riscontrare strozzature o problemi di latenza. Kubernetes fornisce strumenti e meccanismi per scalare le risorse di rete, consentendo alle organizzazioni di mantenere prestazioni ottimali al variare dei modelli di traffico.

Questa elasticità nella scalabilità della rete migliora l'esperienza utente complessiva garantendo connessioni veloci e affidabili. La scalabilità della rete ottimizza anche l'uso delle risorse di rete, contribuendo a ridurre i costi associati ai componenti di rete sottoutilizzati o sovraccarichi.

Inoltre, un'efficace scalabilità della rete è fondamentale per supportare disponibilità e resilienza elevate. Regolando dinamicamente la capacità e il routing della rete, le organizzazioni possono garantire che i servizi rimangano accessibili e reattivi anche durante i periodi di picco della domanda o picchi di traffico imprevisti. Questo approccio consente un migliore utilizzo delle risorse di rete cloud, garantendo che l'infrastruttura sia sempre allineata ai requisiti attuali.

In questa sezione vengono descritti i seguenti tipi di scalabilità della rete:

- Plugin Amazon VPC CNI per Kubernetes
- Rete personalizzata
- Delega con prefisso
- Amazon VPC Lattice

Plug-in CNI di Amazon VPC per Kubernetes

Il plug-in Amazon VPC Container Network Interface (CNI) per Kubernetes è un componente fondamentale di Amazon EKS. Il plug-in VPC CNI offre funzionalità di rete avanzate integrando i pod Kubernetes con Amazon VPC. Con questo plug-in, a ciascun pod viene assegnato un indirizzo IP univoco dal cloud privato virtuale (VPC), migliorando così l'isolamento e le prestazioni della rete. Man mano che i cluster crescono e le richieste di rete variano, il plug-in Amazon VPC CNI svolge un ruolo chiave nel garantire operazioni di rete efficienti e scalabili.

Il plug-in gestisce automaticamente l'allocazione e il routing degli indirizzi IP all'interno del VPC, semplificando la gestione della rete e riducendo il rischio di conflitti IP. Supporta funzionalità come la delega dei prefissi, che consente una gestione IP più flessibile.

Il plug-in VPC CNI aiuta le organizzazioni a ottimizzare le prestazioni di rete, migliorare la sicurezza e ridurre il rischio di esaurimento dell'IP. Queste funzionalità sono particolarmente utili per ambienti dinamici su larga scala in cui le richieste di rete variano, come architetture di microservizi, carichi di lavoro ad alta densità e applicazioni multi-tenant.

Il plugin Amazon VPC CNI offre le seguenti funzionalità chiave:

- Rete avanzata: il plug-in VPC CNI consente a ciascun pod di ricevere il proprio indirizzo IP direttamente dal VPC, garantendo un forte isolamento e prestazioni di rete. Questo approccio è fondamentale per i carichi di lavoro che richiedono un throughput di rete elevato e una bassa latenza.
- Delega dei prefissi: per ovviare ai problemi di esaurimento degli indirizzi IP in cluster di grandi dimensioni, la delega dei prefissi alloca dinamicamente blocchi più grandi tra i nodi, che vengono poi suddivisi IPs per essere utilizzati dai pod. Questo approccio garantisce un utilizzo efficiente dell'IP e semplifica la scalabilità della rete.
- Rete personalizzata: gli utenti possono configurare interfacce di rete personalizzate (ENIs) per i pod, il che aiuta a distribuire il traffico dei pod su più interfacce, riducendo la congestione della rete e migliorando la scalabilità.
- Supporto per IPv6: abilitando IPv6 i cluster Amazon EKS, gli utenti possono espandere in modo significativo lo spazio di indirizzi IP disponibile, facilitando la scalabilità di applicazioni distribuite di grandi dimensioni senza limiti. IPv4
- Integrazione con Kubernetes: il plug-in VPC CNI funziona perfettamente con i componenti di rete Kubernetes, garantendo una gestione efficiente su pod, servizi ed endpoint esterni e supporta funzionalità avanzate come i gruppi di sicurezza per i pod. IPs

Rete personalizzata

Il networking personalizzato in Amazon EKS consente l'assegnazione di interfacce di rete specifiche ai pod, fornendo un controllo migliorato sulla gestione degli indirizzi IP e sul traffico di rete. Questo approccio è particolarmente utile negli scenari in cui l'esaurimento degli indirizzi IP è un problema o quando è necessario separare il traffico di rete per motivi di sicurezza, conformità o prestazioni.

La rete personalizzata aiuta le organizzazioni a gestire in modo efficiente lo spazio degli indirizzi IP, separare il traffico e garantire prestazioni di rete scalabili.

Con una rete personalizzata, gli amministratori possono gestire le risorse di rete in modo più efficiente. Gli amministratori possono utilizzare reti personalizzate per assicurarsi che i pod abbiano il necessario isolamento di rete e che il cluster possa scalare senza incontrare limitazioni di indirizzo IP.

La rete personalizzata offre le seguenti funzionalità chiave:

- Gestione IP avanzata: la rete personalizzata consente l'assegnazione di interfacce di rete specifiche (ENIs) ai pod, aiutando a gestire l'esaurimento degli indirizzi IP distribuendo il traffico dei pod su più pod. ENIs Questa funzionalità è particolarmente importante nei cluster con carichi di lavoro ad alta densità.
- Segregazione del traffico: con interfacce di rete personalizzate, è possibile separare il traffico dei pod in base a criteri specifici, come il tipo di applicazione o i requisiti di sicurezza. Questo approccio offre un maggiore controllo sul flusso di traffico all'interno e all'esterno del cluster.
- Supporto per IPv6: supporta anche la rete personalizzata in Amazon EKS IPv6, che offre una soluzione ai limiti degli IPv4 indirizzi. La rete può scalare in modo efficiente senza conflitti di indirizzi IP, anche in implementazioni su larga scala.
- Scalabilità e flessibilità: man mano che il cluster cresce, la rete personalizzata consente la gestione dinamica delle interfacce di rete. Ai nuovi pod vengono assegnate le risorse di rete appropriate senza intervento manuale. Questo approccio aiuta a mantenere un ambiente di rete flessibile e scalabile in grado di adattarsi ai mutevoli carichi di lavoro.

Delega prefisso

La delega dei prefissi in Kubernetes, in particolare all'interno di Amazon EKS, è progettata per semplificare e ottimizzare la gestione degli indirizzi IP man mano che i cluster si espandono. Allocando dinamicamente blocchi più grandi di indirizzi IP (prefissi) ai nodi, la delega dei prefissi riduce il rischio di esaurimento dell'IP e semplifica la gestione dello spazio IP.

Questo approccio migliora l'efficienza della rete, riduce al minimo la frammentazione e aiuta i cluster a scalare senza problemi senza regolazioni manuali dell'intervallo IP. La delega dei prefissi è particolarmente utile per implementazioni su larga scala, carichi di lavoro ad alta densità e ambienti in cui una gestione IP flessibile e dinamica è fondamentale per mantenere le prestazioni e la scalabilità della rete.

La delega con prefisso offre le seguenti funzionalità chiave:

- Gestione efficiente degli indirizzi IP: la delega dei prefissi consente l'allocazione dinamica degli intervalli IP, riducendo il rischio di esaurimento degli IP e garantendo un uso efficiente dello spazio IP disponibile.
- Gestione semplificata della rete: consentendo ai nodi di gestire le proprie allocazioni IP, la delega dei prefissi riduce al minimo la frammentazione della rete e semplifica il processo di routing, facilitando la scalabilità dei cluster in base alle esigenze.
- Supporto per implementazioni su larga scala: in cluster di grandi dimensioni con carichi di lavoro ad alta densità, la delega dei prefissi consente una scalabilità perfetta consentendo ai nuovi nodi di unirsi al cluster senza regolazioni manuali dell'intervallo IP.

Amazon VPC Lattice

Amazon VPC Lattice consente service-to-service comunicazioni efficienti e sicure all'interno e all'esterno VPCs, in particolare nelle architetture di microservizi. VPC Lattice utilizza misure di sicurezza come gruppi di sicurezza e liste di controllo degli accessi alla rete (rete ACLs) oltre all'integrazione AWS Identity and Access Management (IAM) per l'autenticazione granulare delle applicazioni. Un servizio proxy di livello 7 al centro di VPC Lattice offre connessione, bilanciamento del carico, autenticazione, autorizzazione, osservabilità, gestione del traffico e scoperta dei servizi.

Semplificando le configurazioni di rete e sicurezza, VPC Lattice aiuta le organizzazioni a ottimizzare la gestione del traffico, migliorare le prestazioni delle applicazioni e scalare senza problemi su più piattaforme. VPCs Regioni AWS Ciò è particolarmente utile per le applicazioni distribuite che richiedono una rete coerente e affidabile, come microservizi, implementazioni interregionali e ambienti complessi nativi del cloud.

Amazon VPC Lattice offre le seguenti funzionalità chiave:

- Service-to-service networking: VPC Lattice semplifica la configurazione di rete e sicurezza tra i servizi all'interno di un'architettura di microservizi. Fornisce una piattaforma unificata per la gestione delle comunicazioni, in modo che i servizi possano scalare indipendentemente mantenendo prestazioni e sicurezza elevate.
- Rete cross-VPC: VPC Lattice è fondamentale per la gestione del traffico su più regioni. VPCs Fornisce un framework di rete coerente che consente ai servizi di comunicare senza interruzioni, indipendentemente dalla loro posizione fisica. Questa funzionalità è particolarmente importante per le applicazioni su larga scala che si estendono su più regioni VPCs o aree geografiche.

- Gestione avanzata della sicurezza: integrando le politiche di sicurezza direttamente nel livello di rete, VPC Lattice service-to-service supporta comunicazioni sicure ed efficienti. Questa funzionalità riduce la complessità della gestione della sicurezza in un ambiente distribuito, consentendo una scalabilità più semplice e una riduzione del sovraccarico operativo.
- Gestione semplificata del traffico: VPC Lattice offre funzionalità avanzate di gestione del traffico, tra cui routing, bilanciamento del carico e meccanismi di failover. Con queste funzionalità, il traffico viene distribuito in modo efficiente tra i servizi, ottimizzando le prestazioni di rete e migliorando la scalabilità dell'applicazione.

Ottimizzazione dei costi

Per supportare un controllo efficace delle risorse, la riduzione al minimo dei costi di Kubernetes è fondamentale per le aziende che utilizzano questa tecnologia di orchestrazione dei container. È difficile tracciare correttamente la spesa nelle impostazioni di Kubernetes a causa della loro complessità, che include più componenti come pod e nodi. Attraverso l'applicazione di tecniche di ottimizzazione dei costi, le aziende possono vedere dove vengono spese le proprie risorse e assegnare in modo appropriato le spese ai reparti o ai progetti.

Sebbene la scalabilità dinamica offra dei vantaggi, se non gestita correttamente può comportare spese impreviste. Una gestione efficiente dei costi aiuta ad allocare le risorse solo quando sono realmente necessarie, evitando aumenti di spesa imprevisti.

Questa sezione illustra i seguenti approcci all'ottimizzazione dei costi:

- Kubecost
- Riccioli dorati
- AWS Fargate
- Spot Instances
- Istanze riservate
- AWS Istanze Graviton

Kubecost

Kubecost è una soluzione di gestione dei costi che aiuta le aziende a tracciare, controllare e massimizzare le spese per l'infrastruttura cloud. È stato creato appositamente per i cluster Kubernetes. Kubecost ti offre informazioni sull'utilizzo delle risorse e una conoscenza dei costi in tempo reale, consentendoti di comprendere meglio dove e quanto vengono utilizzate le tue risorse cloud. Con queste informazioni, puoi ottimizzare la spesa per l'infrastruttura, migliorare l'efficienza delle risorse e prendere decisioni più informate sui tuoi investimenti nel cloud.

Kubecost offre le seguenti funzionalità chiave:

- Allocazione dei costi: Kubecost offre un'allocazione completa dei costi per le risorse Kubernetes, inclusi carichi di lavoro, servizi, namespace ed etichette. Questa funzionalità aiuta i team a monitorare i costi per ambiente, progetto o team.

- Monitoraggio dei costi in tempo reale: offre il monitoraggio in tempo reale dei costi del cloud, offrendo alle organizzazioni informazioni immediate sui modelli di spesa e aiutando a prevenire sforamenti imprevisti dei costi.
- Consigli per l'ottimizzazione: Kubecost offre suggerimenti pratici per ridurre al minimo l'utilizzo delle risorse, tra cui la riduzione delle risorse inattive, il corretto dimensionamento dei carichi di lavoro e la massimizzazione delle spese di archiviazione.
- Budget e avvisi: gli utenti di Kubecost possono creare budget e ricevere promemoria quando una spesa si avvicina o supera i criteri predeterminati. Questa funzionalità aiuta i team a rispettare i vincoli finanziari.

Riccioli d'oro

Goldilocks è un'utilità Kubernetes progettata per aiutare gli utenti a ottimizzare le richieste di risorse e i limiti per i carichi di lavoro Kubernetes. Fornisce consigli su come configurare la CPU e le risorse di memoria per i contenitori in esecuzione in un cluster Kubernetes. Questi consigli ti aiutano a garantire che le applicazioni dispongano del giusto numero di risorse per funzionare in modo efficiente senza sprechi. Questa ottimizzazione può portare a risparmi sui costi, prestazioni migliorate e un uso più efficiente dei cluster Kubernetes.

Goldilocks offre le seguenti funzionalità chiave:

- Consigli sulle risorse: Goldilocks determina le impostazioni ideali per le richieste e le restrizioni delle risorse analizzando le statistiche precedenti sul consumo di CPU e memoria per i carichi di lavoro Kubernetes. In questo modo, diventa più facile evitare un approvvigionamento insufficiente o eccessivo, che può causare problemi di prestazioni e spreco di risorse.
- Integrazione VPA: Goldilocks sfrutta Kubernetes Vertical Pod Autoscaler (VPA) per raccogliere dati e fornire consigli. Funziona in una «modalità di raccomandazione», il che significa che non modifica effettivamente le impostazioni delle risorse ma offre indicazioni su quali dovrebbero essere tali impostazioni.
- Analisi basata sul namespace: Goldilocks ti dà la possibilità di regolare con precisione quali carichi di lavoro sono ottimizzati e monitorati, consentendoti di indirizzare determinati namespace per l'analisi.
- Dashboard visivo: la dashboard basata sul Web mostra visivamente le richieste e le restrizioni suggerite in materia di risorse, il che semplifica la comprensione e l'azione sui dati.

- Funzionamento non intrusivo: Goldilocks non altera la configurazione del cluster perché funziona in modalità di raccomandazione. Se lo desideri, puoi applicare manualmente le impostazioni delle risorse consigliate dopo aver esaminato i consigli.

AWS Fargate

Nel contesto di Amazon EKS, https://docs.aws.amazon.com/eks/latest/userguide/fargate.htmlAWS Fargate consente di eseguire i pod Kubernetes senza gestire le istanze Amazon sottostanti. EC2 È un motore di elaborazione serverless che ti consente di concentrarti sulla distribuzione e sulla scalabilità di applicazioni containerizzate senza preoccuparti dell'infrastruttura.

AWS Fargate offre le seguenti funzionalità chiave:

- Nessuna gestione dell'infrastruttura: Fargate elimina la necessità di fornire, gestire o scalare le EC2 istanze Amazon o i nodi Kubernetes. AWS gestisce tutta la gestione dell'infrastruttura, incluse l'applicazione di patch e la scalabilità.
- Isolamento a livello di pod: a differenza dei nodi di lavoro basati su Amazon, EC2 Fargate offre l'isolamento a livello di task o pod. Ogni pod funziona nel proprio ambiente di elaborazione isolato, che migliora la sicurezza e le prestazioni.
- Scalabilità automatica: Fargate ridimensiona automaticamente i pod Kubernetes in base alla richiesta. Non è necessario gestire le politiche di scalabilità o i pool di nodi.
- Fatturazione al secondo: paghi solo per le risorse di vCPU e memoria consumate da ciascun pod per l'esatta durata di funzionamento, un'opzione conveniente per determinati carichi di lavoro.
- Sovraccarico ridotto: eliminando la necessità di gestire EC2 le istanze, Fargate consente di concentrarsi sulla creazione e sulla gestione delle applicazioni anziché sulle operazioni dell'infrastruttura.

Spot Instances

Le istanze Spot offrono risparmi significativi rispetto ai prezzi delle istanze on demand e sono un'opzione conveniente per eseguire EC2 nodi di lavoro Amazon in un cluster Amazon EKS. Tuttavia, AWS possono interrompere le istanze Spot nel caso in cui sia necessaria la capacità dell'istanza on demand. AWS può recuperare le istanze Spot con un preavviso di 2 minuti quando è necessaria la capacità necessaria, rendendole meno affidabili per carichi di lavoro critici e con stato elevato.

Per i carichi di lavoro sensibili ai costi e in grado di resistere alle interruzioni, le istanze Spot in Amazon EKS sono una buona opzione. L'utilizzo di una combinazione di istanze Spot e istanze On-Demand in un cluster Kubernetes consente di risparmiare denaro senza sacrificare la disponibilità per carichi di lavoro vitali.

Le istanze Spot offrono le seguenti funzionalità chiave:

- Risparmio sui costi: le istanze Spot possono essere meno costose delle istanze on demand, il che le rende ideali per carichi di lavoro sensibili ai costi.
- Ideale per carichi di lavoro con tolleranza ai guasti: ideale per carichi di lavoro stateless e tolleranti ai guasti come l'elaborazione in batch, i lavori CI/CD, l'apprendimento automatico o l'elaborazione di dati su larga scala in cui le istanze possono essere sostituite senza gravi interruzioni.
- Integrazione di gruppo con scalabilità automatica: Amazon EKS integra le istanze Spot con Kubernetes Cluster Autoscaler, che può sostituire automaticamente i nodi di istanze Spot interrotti con altre istanze Spot o istanze On-Demand disponibili.

Istanze riservate

In Amazon EKS, Reserved Instances è un modello di prezzo per i EC2 nodi di lavoro Amazon che eseguono i carichi di lavoro Kubernetes. Utilizzando le istanze riservate, ti impegni a utilizzare tipi di istanze specifici per un periodo di 1 o 3 anni, in cambio di risparmi sui costi rispetto ai prezzi delle istanze on demand. La prenotazione di istanze in Amazon EKS è un modo conveniente per eseguire carichi di lavoro coerenti a lungo termine sui nodi di lavoro Amazon EC2 .

Le istanze riservate sono comunemente utilizzate per Amazon EC2. Tuttavia, anche i nodi di lavoro del cluster Amazon EKS (che sono EC2 istanze) possono trarre vantaggio da questo modello di risparmio sui costi, a condizione che il carico di lavoro richieda un utilizzo prevedibile a lungo termine.

I servizi di produzione, i database e altre applicazioni stateful che richiedono elevata disponibilità e prestazioni costanti sono esempi di carichi di lavoro stabili che ben si adattano alle istanze riservate.

Le istanze riservate offrono le seguenti funzionalità chiave:

- Risparmi sui costi: le istanze riservate offrono risparmi rispetto alle istanze on demand, a seconda della durata del periodo (1 o 3 anni) e del piano di pagamento (tutto anticipato, anticipo parziale o nessun anticipo).

- Impegno a lungo termine: ti impegni a rispettare un periodo di 1 o 3 anni per uno specifico tipo, dimensione e. Regione AWSÈ ideale per carichi di lavoro stabili e funzionanti ininterrottamente nel tempo.
- Prezzi prevedibili: poiché ti impegni a rispettare un termine specifico, le istanze riservate offrono costi mensili o iniziali prevedibili, semplificando la definizione del budget per i carichi di lavoro a lungo termine.
- Flessibilità delle istanze: con le istanze riservate convertibili, puoi modificare il tipo, la famiglia o la dimensione dell'istanza durante il periodo di prenotazione. Le istanze riservate convertibili offrono maggiore flessibilità rispetto alle istanze riservate standard, che non consentono modifiche.
- Capacità garantita: le istanze riservate assicurano che la capacità sia disponibile nella zona di disponibilità in cui viene effettuata la prenotazione, il che è fondamentale per i carichi di lavoro critici che richiedono una potenza di calcolo costante.
- Nessun rischio di interruzione: a differenza delle istanze Spot, le istanze riservate non sono soggette a interruzioni da parte di. AWS Ciò le rende ideali per l'esecuzione di carichi di lavoro cruciali che richiedono tempi di attività garantiti.

AWS Istanze Graviton

AWS Graviton è una famiglia di processori basati su ARM progettata per fornire prestazioni migliorate ed efficienza in termini di costi AWS per i carichi di lavoro cloud. Nel contesto di Amazon EKS, puoi utilizzare le istanze Graviton come nodi di lavoro per eseguire i carichi di lavoro Kubernetes, offrendo significativi miglioramenti delle prestazioni e risparmi sui costi.

Le istanze Graviton sono un'opzione eccellente per le applicazioni native del cloud e ad alta intensità di calcolo perché offrono un rapporto prezzo/prestazioni più elevato rispetto alle istanze x86. Tuttavia, quando prendi in considerazione l'adozione di istanze Graviton, tieni conto della compatibilità ARM.

AWS Le istanze Graviton offrono le seguenti funzionalità chiave:

- Architettura basata su ARM: i processori AWS Graviton sono basati sull'architettura ARM, che è diversa dalle tradizionali architetture x86 ma altamente efficiente per molti carichi di lavoro.
- Conveniente: EC2 le istanze Amazon basate su Graviton offrono in genere un miglior rapporto prezzo/prestazioni rispetto alle istanze basate su x86. EC2 Ciò li rende un'opzione interessante per i cluster Kubernetes che eseguono Amazon EKS.

- Prestazioni: i processori Graviton2, la seconda generazione di AWS Graviton, offrono miglioramenti significativi in termini di prestazioni di calcolo, throughput di memoria ed efficienza energetica. Sono ideali per carichi di lavoro che richiedono un uso intensivo della CPU e della memoria.
- Diversi tipi di istanze: le istanze Graviton sono disponibili in diverse famiglie, come t4g, m7g, c7g e r7g, e coprono una vasta gamma di casi d'uso, dai carichi di lavoro generici a quelli ottimizzati per il calcolo, la memoria e i carichi di lavoro espandibili.
- Gruppi di nodi Amazon EKS: puoi configurare gruppi di nodi gestiti da Amazon EKS o gruppi di nodi autogestiti per includere istanze basate su Graviton. Con questo approccio, puoi eseguire carichi di lavoro ottimizzati per l'architettura ARM sullo stesso cluster Kubernetes insieme a istanze basate su x86.

Passaggi successivi

Questa guida fornisce informazioni per aiutarti a ottimizzare Amazon EKS per quanto riguarda la scalabilità di calcolo, la scalabilità del carico di lavoro, la scalabilità della rete e l'ottimizzazione dei costi. Comprendendo e applicando questi concetti, le organizzazioni possono ottenere un ambiente cloud altamente efficiente, scalabile ed economico che soddisfi le loro esigenze dinamiche.

L'implementazione efficace della scalabilità di calcolo e carico di lavoro aiuta a garantire che le risorse vengano utilizzate in modo efficiente e che le applicazioni mantengano prestazioni elevate anche nelle ore di punta. L'adozione di tecniche di scalabilità della rete, come la rete personalizzata e la delega dei prefissi, supporta la gestione delle risorse di rete e una scalabilità senza interruzioni. L'enfasi sull'ottimizzazione dei costi aiuta le organizzazioni a bilanciare le prestazioni con l'efficienza finanziaria.

L'integrazione di queste linee guida nella tua strategia cloud può aiutarti a migliorare le prestazioni e la scalabilità dell'infrastruttura e a ridurre i costi. Questo approccio completo può consentirti di creare un solido ambiente cloud che supporti la crescita dell'organizzazione e si adatti alle esigenze aziendali in continua evoluzione.

Risorse

AWS blog

- Sviluppo per l'ottimizzazione dei costi e la resilienza per EKS con le istanze Spot
- Combinazione di AWS Graviton con x86 CPUs per ottimizzare costi e resilienza utilizzando Amazon EKS

AWS documentazione

- CNI di Amazon VPC
- Amazon Elastic Kubernetes Service AWS (white paper: panoramica delle opzioni di distribuzione su) AWS
- Guida alle best practice di Amazon EKS
- Carpenter
- Scopri di più su Kubecost
- Semplifica la gestione dell'elaborazione con AWS Fargate

Altre risorse

- Cluster Autoscaling (documentazione Kubernetes)
- Goldilocks: uno strumento open source per consigliare richieste di risorse (Fairwinds Blog)
- Scalabilità automatica dei pod orizzontali (documentazione Kubernetes)
- Kubecost (documentazione Kubecost)
- Scalabilità automatica basata sugli eventi di Kubernetes (documentazione KEDA)

Cronologia dei documenti

La tabella seguente descrive le modifiche significative a questa guida, Scalabilità dell'infrastruttura Amazon EKS per ottimizzare elaborazione, carichi di lavoro e prestazioni di rete. Per ricevere notifiche sugli aggiornamenti futuri, puoi abbonarti a un feed RSS.

Modifica	Descrizione	Data
Pubblicazione iniziale	—	11 novembre 2024

AWS Glossario delle linee guida prescrittive

I seguenti sono termini di uso comune nelle strategie, nelle guide e nei modelli forniti da AWS Prescriptive Guidance. Per suggerire voci, utilizza il link Fornisci feedback alla fine del glossario.

Numeri

7 R

Sette strategie di migrazione comuni per trasferire le applicazioni sul cloud. Queste strategie si basano sulle 5 R identificate da Gartner nel 2011 e sono le seguenti:

- Rifattorizzare/riprogettare: trasferisci un'applicazione e modifica la sua architettura sfruttando appieno le funzionalità native del cloud per migliorare l'agilità, le prestazioni e la scalabilità. Ciò comporta in genere la portabilità del sistema operativo e del database. Esempio: migra il tuo database Oracle locale all'edizione compatibile con Amazon Aurora PostgreSQL.
- Ridefinire la piattaforma (lift and reshape): trasferisci un'applicazione nel cloud e introduci un certo livello di ottimizzazione per sfruttare le funzionalità del cloud. Esempio: migra il tuo database Oracle locale ad Amazon Relational Database Service (Amazon RDS) per Oracle in. Cloud AWS
- Riacquistare (drop and shop): passa a un prodotto diverso, in genere effettuando la transizione da una licenza tradizionale a un modello SaaS. Esempio: migra il tuo sistema di gestione delle relazioni con i clienti (CRM) su Salesforce.com.
- Eseguire il rehosting (lift and shift): trasferisci un'applicazione sul cloud senza apportare modifiche per sfruttare le funzionalità del cloud. Esempio: migra il database Oracle locale su Oracle su un'istanza in. EC2 Cloud AWS
- Trasferire (eseguire il rehosting a livello hypervisor): trasferisci l'infrastruttura sul cloud senza acquistare nuovo hardware, riscrivere le applicazioni o modificare le operazioni esistenti. Si esegue la migrazione dei server da una piattaforma locale a un servizio cloud per la stessa piattaforma. Esempio: migra un'applicazione suMicrosoft Hyper-V. AWS
- Riesaminare (mantenere): mantieni le applicazioni nell'ambiente di origine. Queste potrebbero includere applicazioni che richiedono una rifattorizzazione significativa che desideri rimandare a un momento successivo e applicazioni legacy che desideri mantenere, perché non vi è alcuna giustificazione aziendale per effettuarne la migrazione.
- Ritirare: disattiva o rimuovi le applicazioni che non sono più necessarie nell'ambiente di origine.

A

ABAC

Vedi controllo degli accessi basato sugli attributi.

servizi astratti

Vedi servizi gestiti.

ACIDO

Vedi atomicità, consistenza, isolamento, durata.

migrazione attiva-attiva

Un metodo di migrazione del database in cui i database di origine e di destinazione vengono mantenuti sincronizzati (utilizzando uno strumento di replica bidirezionale o operazioni di doppia scrittura) ed entrambi i database gestiscono le transazioni provenienti dalle applicazioni di connessione durante la migrazione. Questo metodo supporta la migrazione in piccoli batch controllati anziché richiedere una conversione una tantum. È più flessibile ma richiede più lavoro rispetto alla migrazione attiva-passiva.

migrazione attiva-passiva

Un metodo di migrazione di database in cui i database di origine e di destinazione vengono mantenuti sincronizzati, ma solo il database di origine gestisce le transazioni provenienti dalle applicazioni di connessione mentre i dati vengono replicati nel database di destinazione. Il database di destinazione non accetta alcuna transazione durante la migrazione.

funzione aggregata

Una funzione SQL che opera su un gruppo di righe e calcola un singolo valore restituito per il gruppo. Esempi di funzioni aggregate includono `SUM` e. `MAX`

Intelligenza artificiale

Vedi intelligenza artificiale.

AIOps

Guarda le operazioni di intelligenza artificiale.

anonimizzazione

Il processo di eliminazione permanente delle informazioni personali in un set di dati. L'anonimizzazione può aiutare a proteggere la privacy personale. I dati anonimi non sono più considerati dati personali.

anti-modello

Una soluzione utilizzata di frequente per un problema ricorrente in cui la soluzione è controproducente, inefficace o meno efficace di un'alternativa.

controllo delle applicazioni

Un approccio alla sicurezza che consente l'uso solo di applicazioni approvate per proteggere un sistema dal malware.

portfolio di applicazioni

Una raccolta di informazioni dettagliate su ogni applicazione utilizzata da un'organizzazione, compresi i costi di creazione e manutenzione dell'applicazione e il relativo valore aziendale. Queste informazioni sono fondamentali per il processo di scoperta e analisi del portfolio e aiutano a identificare e ad assegnare la priorità alle applicazioni da migrare, modernizzare e ottimizzare.

intelligenza artificiale (IA)

Il campo dell'informatica dedicato all'uso delle tecnologie informatiche per svolgere funzioni cognitive tipicamente associate agli esseri umani, come l'apprendimento, la risoluzione di problemi e il riconoscimento di schemi. Per ulteriori informazioni, consulta la sezione Che cos'è l'intelligenza artificiale?

operazioni di intelligenza artificiale (AIOps)

Il processo di utilizzo delle tecniche di machine learning per risolvere problemi operativi, ridurre gli incidenti operativi e l'intervento umano e aumentare la qualità del servizio. Per ulteriori informazioni su come AIOps viene utilizzato nella strategia di AWS migrazione, consulta la guida all'integrazione delle operazioni.

crittografia asimmetrica

Un algoritmo di crittografia che utilizza una coppia di chiavi, una chiave pubblica per la crittografia e una chiave privata per la decrittografia. Puoi condividere la chiave pubblica perché non viene utilizzata per la decrittografia, ma l'accesso alla chiave privata deve essere altamente limitato.

atomicità, consistenza, isolamento, durabilità (ACID)

Un insieme di proprietà del software che garantiscono la validità dei dati e l'affidabilità operativa di un database, anche in caso di errori, interruzioni di corrente o altri problemi.

Controllo degli accessi basato su attributi (ABAC)

La pratica di creare autorizzazioni dettagliate basate su attributi utente, come reparto, ruolo professionale e nome del team. Per ulteriori informazioni, consulta ABAC AWS nella documentazione AWS Identity and Access Management (IAM).

fonte di dati autorevole

Una posizione in cui è archiviata la versione principale dei dati, considerata la fonte di informazioni più affidabile. È possibile copiare i dati dalla fonte di dati autorevole in altre posizioni allo scopo di elaborarli o modificarli, ad esempio anonimizzandoli, oscurandoli o pseudonimizzandoli.

Zona di disponibilità

Una posizione distinta all'interno di un edificio Regione AWS che è isolata dai guasti in altre zone di disponibilità e offre una connettività di rete economica e a bassa latenza verso altre zone di disponibilità nella stessa regione.

AWS Cloud Adoption Framework (CAF)AWS

Un framework di linee guida e best practice AWS per aiutare le organizzazioni a sviluppare un piano efficiente ed efficace per passare con successo al cloud. AWS CAF organizza le linee guida in sei aree di interesse chiamate prospettive: business, persone, governance, piattaforma, sicurezza e operazioni. Le prospettive relative ad azienda, persone e governance si concentrano sulle competenze e sui processi aziendali; le prospettive relative alla piattaforma, alla sicurezza e alle operazioni si concentrano sulle competenze e sui processi tecnici. Ad esempio, la prospettiva relativa alle persone si rivolge alle parti interessate che gestiscono le risorse umane (HR), le funzioni del personale e la gestione del personale. In questa prospettiva, AWS CAF fornisce linee guida per lo sviluppo delle persone, la formazione e le comunicazioni per aiutare a preparare l'organizzazione all'adozione del cloud di successo. Per ulteriori informazioni, consulta il sito web di AWS CAF e il white paper AWS CAF.

AWS Workload Qualification Framework (WQF)AWS

Uno strumento che valuta i carichi di lavoro di migrazione dei database, consiglia strategie di migrazione e fornisce stime del lavoro. AWS WQF è incluso in (). AWS Schema Conversion Tool AWS SCT Analizza gli schemi di database e gli oggetti di codice, il codice dell'applicazione, le dipendenze e le caratteristiche delle prestazioni e fornisce report di valutazione.

B

bot difettoso

Un bot che ha lo scopo di interrompere o causare danni a individui o organizzazioni.

BCP

Vedi la pianificazione della continuità operativa.

grafico comportamentale

Una vista unificata, interattiva dei comportamenti delle risorse e delle interazioni nel tempo. Puoi utilizzare un grafico comportamentale con Amazon Detective per esaminare tentativi di accesso non riusciti, chiamate API sospette e azioni simili. Per ulteriori informazioni, consulta Dati in un grafico comportamentale nella documentazione di Detective.

sistema big-endian

Un sistema che memorizza per primo il byte più importante. Vedi anche endianness.

Classificazione binaria

Un processo che prevede un risultato binario (una delle due classi possibili). Ad esempio, il modello di machine learning potrebbe dover prevedere problemi come "Questa e-mail è spam o non è spam?" o "Questo prodotto è un libro o un'auto?"

filtro Bloom

Una struttura di dati probabilistica ed efficiente in termini di memoria che viene utilizzata per verificare se un elemento fa parte di un set.

distribuzioni blu/verdi

Una strategia di implementazione in cui si creano due ambienti separati ma identici. La versione corrente dell'applicazione viene eseguita in un ambiente (blu) e la nuova versione dell'applicazione nell'altro ambiente (verde). Questa strategia consente di ripristinare rapidamente il sistema con un impatto minimo.

bot

Un'applicazione software che esegue attività automatizzate su Internet e simula l'attività o l'interazione umana. Alcuni bot sono utili o utili, come i web crawler che indicizzano le informazioni su Internet. Alcuni altri bot, noti come bot dannosi, hanno lo scopo di disturbare o causare danni a individui o organizzazioni.

botnet

Reti di bot infettate da malware e controllate da un'unica parte, nota come bot herder o bot operator. Le botnet sono il meccanismo più noto per scalare i bot e il loro impatto.

ramo

Un'area contenuta di un repository di codice. Il primo ramo creato in un repository è il ramo principale. È possibile creare un nuovo ramo a partire da un ramo esistente e quindi sviluppare funzionalità o correggere bug al suo interno. Un ramo creato per sviluppare una funzionalità viene comunemente detto ramo di funzionalità. Quando la funzionalità è pronta per il rilascio, il ramo di funzionalità viene ricongiunto al ramo principale. Per ulteriori informazioni, consulta Informazioni sulle filiali (documentazione). GitHub

accesso break-glass

In circostanze eccezionali e tramite una procedura approvata, un mezzo rapido per consentire a un utente di accedere a un sito a Account AWS cui in genere non dispone delle autorizzazioni necessarie. Per ulteriori informazioni, vedere l'indicatore Implementate break-glass procedures nella guida Well-Architected AWS .

strategia brownfield

L'infrastruttura esistente nell'ambiente. Quando si adotta una strategia brownfield per un'architettura di sistema, si progetta l'architettura in base ai vincoli dei sistemi e dell'infrastruttura attuali. Per l'espansione dell'infrastruttura esistente, è possibile combinare strategie brownfield e greenfield.

cache del buffer

L'area di memoria in cui sono archiviati i dati a cui si accede con maggiore frequenza.

capacità di business

Azioni intraprese da un'azienda per generare valore (ad esempio vendite, assistenza clienti o marketing). Le architetture dei microservizi e le decisioni di sviluppo possono essere guidate dalle capacità aziendali. Per ulteriori informazioni, consulta la sezione Organizzazione in base alle funzionalità aziendali del whitepaper Esecuzione di microservizi containerizzati su AWS.

pianificazione della continuità operativa (BCP)

Un piano che affronta il potenziale impatto di un evento che comporta l'interruzione dell'attività, come una migrazione su larga scala, sulle operazioni e consente a un'azienda di riprendere rapidamente le operazioni.

C

CAF

Vedi AWS Cloud Adoption Framework.

implementazione canaria

Il rilascio lento e incrementale di una versione agli utenti finali. Quando sei sicuro, distribuisci la nuova versione e sostituisci la versione corrente nella sua interezza.

CCoE

Vedi Cloud Center of Excellence.

CDC

Vedi Change Data Capture.

Change Data Capture (CDC)

Il processo di tracciamento delle modifiche a un'origine dati, ad esempio una tabella di database, e di registrazione dei metadati relativi alla modifica. È possibile utilizzare CDC per vari scopi, ad esempio il controllo o la replica delle modifiche in un sistema di destinazione per mantenere la sincronizzazione.

ingegneria del caos

Introduzione intenzionale di guasti o eventi dirompenti per testare la resilienza di un sistema. Puoi usare AWS Fault Injection Service (AWS FIS) per eseguire esperimenti che stressano i tuoi AWS carichi di lavoro e valutarne la risposta.

CI/CD

Vedi integrazione continua e distribuzione continua.

classificazione

Un processo di categorizzazione che aiuta a generare previsioni. I modelli di ML per problemi di classificazione prevedono un valore discreto. I valori discreti sono sempre distinti l'uno dall'altro. Ad esempio, un modello potrebbe dover valutare se in un'immagine è presente o meno un'auto.

crittografia lato client

Crittografia dei dati a livello locale, prima che il destinatario li Servizio AWS riceva.

Centro di eccellenza cloud (CCoE)

Un team multidisciplinare che guida le iniziative di adozione del cloud in tutta l'organizzazione, tra cui lo sviluppo di best practice per il cloud, la mobilitazione delle risorse, la definizione delle tempistiche di migrazione e la guida dell'organizzazione attraverso trasformazioni su larga scala. Per ulteriori informazioni, consulta gli CCoE post sull' Cloud AWS Enterprise Strategy Blog.

cloud computing

La tecnologia cloud generalmente utilizzata per l'archiviazione remota di dati e la gestione dei dispositivi IoT. Il cloud computing è generalmente collegato alla tecnologia di edge computing.

modello operativo cloud

In un'organizzazione IT, il modello operativo utilizzato per creare, maturare e ottimizzare uno o più ambienti cloud. Per ulteriori informazioni, consulta Building your Cloud Operating Model.

fasi di adozione del cloud

Le quattro fasi che le organizzazioni in genere attraversano quando migrano verso Cloud AWS:

- Progetto: esecuzione di alcuni progetti relativi al cloud per scopi di dimostrazione e apprendimento
- Fondamento: effettuare investimenti fondamentali per scalare l'adozione del cloud (ad esempio, creazione di una landing zone, definizione di una CCo E, definizione di un modello operativo)
- Migrazione: migrazione di singole applicazioni
- Reinvenzione: ottimizzazione di prodotti e servizi e innovazione nel cloud

Queste fasi sono state definite da Stephen Orban nel post sul blog The Journey Toward Cloud-First & the Stages of Adoption on the Enterprise Strategy. Cloud AWS Per informazioni su come si relazionano alla strategia di AWS migrazione, consulta la guida alla preparazione alla migrazione.

CMDB

Vedi database di gestione della configurazione.

repository di codice

Una posizione in cui il codice di origine e altri asset, come documentazione, esempi e script, vengono archiviati e aggiornati attraverso processi di controllo delle versioni. Gli archivi cloud più comuni includono GitHub oBitbucket Cloud. Ogni versione del codice è denominata ramo. In una

struttura a microservizi, ogni repository è dedicato a una singola funzionalità. Una singola pipeline CI/CD può utilizzare più repository.

cache fredda

Una cache del buffer vuota, non ben popolata o contenente dati obsoleti o irrilevanti. Ciò influisce sulle prestazioni perché l'istanza di database deve leggere dalla memoria o dal disco principale, il che richiede più tempo rispetto alla lettura dalla cache del buffer.

dati freddi

Dati a cui si accede raramente e che in genere sono storici. Quando si eseguono interrogazioni di questo tipo di dati, le interrogazioni lente sono in genere accettabili. Lo spostamento di questi dati su livelli o classi di storage meno costosi e con prestazioni inferiori può ridurre i costi.

visione artificiale (CV)

Un campo dell'intelligenza artificiale che utilizza l'apprendimento automatico per analizzare ed estrarre informazioni da formati visivi come immagini e video digitali. Ad esempio, Amazon SageMaker AI fornisce algoritmi di elaborazione delle immagini per CV.

deriva della configurazione

Per un carico di lavoro, una modifica della configurazione rispetto allo stato previsto. Potrebbe causare la non conformità del carico di lavoro e in genere è graduale e involontaria.

database di gestione della configurazione (CMDB)

Un repository che archivia e gestisce le informazioni su un database e il relativo ambiente IT, inclusi i componenti hardware e software e le relative configurazioni. In genere si utilizzano i dati di un CMDB nella fase di individuazione e analisi del portafoglio della migrazione.

Pacchetto di conformità

Una raccolta di AWS Config regole e azioni correttive che puoi assemblare per personalizzare i controlli di conformità e sicurezza. È possibile distribuire un pacchetto di conformità come singola entità in una regione Account AWS and o all'interno di un'organizzazione utilizzando un modello YAML. Per ulteriori informazioni, consulta i Conformance Pack nella documentazione. AWS Config

integrazione e distribuzione continua (continuous integration and continuous delivery, CI/CD)

Il processo di automazione delle fasi di origine, compilazione, test, gestione temporanea e produzione del processo di rilascio del software. CI/CD is commonly described as a pipeline. CI/

CDpuò aiutarvi ad automatizzare i processi, migliorare la produttività, migliorare la qualità del codice e velocizzare le consegne. Per ulteriori informazioni, consulta Vantaggi della distribuzione continua. CD può anche significare continuous deployment (implementazione continua). Per ulteriori informazioni, consulta Distribuzione continua e implementazione continua a confronto.

CV

Vedi visione artificiale.

D

dati a riposo

Dati stazionari nella rete, ad esempio i dati archiviati.

classificazione dei dati

Un processo per identificare e classificare i dati nella rete in base alla loro criticità e sensibilità. È un componente fondamentale di qualsiasi strategia di gestione dei rischi di sicurezza informatica perché consente di determinare i controlli di protezione e conservazione appropriati per i dati. La classificazione dei dati è un componente del pilastro della sicurezza nel AWS Well-Architected Framework. Per ulteriori informazioni, consulta Classificazione dei dati.

deriva dei dati

Una variazione significativa tra i dati di produzione e i dati utilizzati per addestrare un modello di machine learning o una modifica significativa dei dati di input nel tempo. La deriva dei dati può ridurre la qualità, l'accuratezza e l'equità complessive nelle previsioni dei modelli ML.

dati in transito

Dati che si spostano attivamente attraverso la rete, ad esempio tra le risorse di rete.

rete di dati

Un framework architettonico che fornisce la proprietà distribuita e decentralizzata dei dati con gestione e governance centralizzate.

riduzione al minimo dei dati

Il principio della raccolta e del trattamento dei soli dati strettamente necessari. Praticare la riduzione al minimo dei dati in the Cloud AWS può ridurre i rischi per la privacy, i costi e l'impronta di carbonio delle analisi.

perimetro dei dati

Una serie di barriere preventive nell' AWS ambiente che aiutano a garantire che solo le identità attendibili accedano alle risorse attendibili delle reti previste. Per ulteriori informazioni, consulta Building a data perimeter on. AWS

pre-elaborazione dei dati

Trasformare i dati grezzi in un formato che possa essere facilmente analizzato dal modello di ML. La pre-elaborazione dei dati può comportare la rimozione di determinate colonne o righe e l'eliminazione di valori mancanti, incoerenti o duplicati.

provenienza dei dati

Il processo di tracciamento dell'origine e della cronologia dei dati durante il loro ciclo di vita, ad esempio il modo in cui i dati sono stati generati, trasmessi e archiviati.

soggetto dei dati

Un individuo i cui dati vengono raccolti ed elaborati.

data warehouse

Un sistema di gestione dei dati che supporta la business intelligence, come l'analisi. I data warehouse contengono in genere grandi quantità di dati storici e vengono generalmente utilizzati per interrogazioni e analisi.

linguaggio di definizione del database (DDL)

Istruzioni o comandi per creare o modificare la struttura di tabelle e oggetti in un database.

linguaggio di manipolazione del database (DML)

Istruzioni o comandi per modificare (inserire, aggiornare ed eliminare) informazioni in un database.

DDL

Vedi linguaggio di definizione del database.

deep ensemble

Combinare più modelli di deep learning per la previsione. È possibile utilizzare i deep ensemble per ottenere una previsione più accurata o per stimare l'incertezza nelle previsioni.

deep learning

Un sottocampo del ML che utilizza più livelli di reti neurali artificiali per identificare la mappatura tra i dati di input e le variabili target di interesse.

defense-in-depth

Un approccio alla sicurezza delle informazioni in cui una serie di meccanismi e controlli di sicurezza sono accuratamente stratificati su una rete di computer per proteggere la riservatezza, l'integrità e la disponibilità della rete e dei dati al suo interno. Quando si adotta questa strategia AWS, si aggiungono più controlli a diversi livelli della AWS Organizations struttura per proteggere le risorse. Ad esempio, un defense-in-depth approccio potrebbe combinare l'autenticazione a più fattori, la segmentazione della rete e la crittografia.

amministratore delegato

In AWS Organizations, un servizio compatibile può registrare un account AWS membro per amministrare gli account dell'organizzazione e gestire le autorizzazioni per quel servizio. Questo account è denominato amministratore delegato per quel servizio specifico. Per ulteriori informazioni e un elenco di servizi compatibili, consulta Servizi che funzionano con AWS Organizations nella documentazione di AWS Organizations .

implementazione

Il processo di creazione di un'applicazione, di nuove funzionalità o di correzioni di codice disponibili nell'ambiente di destinazione. L'implementazione prevede l'applicazione di modifiche in una base di codice, seguita dalla creazione e dall'esecuzione di tale base di codice negli ambienti applicativi.

Ambiente di sviluppo

Vedi ambiente.

controllo di rilevamento

Un controllo di sicurezza progettato per rilevare, registrare e avvisare dopo che si è verificato un evento. Questi controlli rappresentano una seconda linea di difesa e avvisano l'utente in caso di eventi di sicurezza che aggirano i controlli preventivi in vigore. Per ulteriori informazioni, consulta Controlli di rilevamento in Implementazione dei controlli di sicurezza in AWS.

mappatura del flusso di valore dello sviluppo (DVSM)

Un processo utilizzato per identificare e dare priorità ai vincoli che influiscono negativamente sulla velocità e sulla qualità nel ciclo di vita dello sviluppo del software. DVSM estende il processo di

mappatura del flusso di valore originariamente progettato per pratiche di produzione snella. Si concentra sulle fasi e sui team necessari per creare e trasferire valore attraverso il processo di sviluppo del software.

gemello digitale

Una rappresentazione virtuale di un sistema reale, ad esempio un edificio, una fabbrica, un'attrezzatura industriale o una linea di produzione. I gemelli digitali supportano la manutenzione predittiva, il monitoraggio remoto e l'ottimizzazione della produzione.

tabella delle dimensioni

In uno schema a stella, una tabella più piccola che contiene gli attributi dei dati quantitativi in una tabella dei fatti. Gli attributi della tabella delle dimensioni sono in genere campi di testo o numeri discreti che si comportano come testo. Questi attributi vengono comunemente utilizzati per il vincolo delle query, il filtraggio e l'etichettatura dei set di risultati.

disastro

Un evento che impedisce a un carico di lavoro o a un sistema di raggiungere gli obiettivi aziendali nella sua sede principale di implementazione. Questi eventi possono essere disastri naturali, guasti tecnici o il risultato di azioni umane, come errori di configurazione involontari o attacchi di malware.

disaster recovery (DR)

La strategia e il processo utilizzati per ridurre al minimo i tempi di inattività e la perdita di dati causati da un disastro. Per ulteriori informazioni, consulta Disaster Recovery of Workloads su AWS: Recovery in the Cloud in the AWS Well-Architected Framework.

DML

Vedi linguaggio di manipolazione del database.

progettazione basata sul dominio

Un approccio allo sviluppo di un sistema software complesso collegandone i componenti a domini in evoluzione, o obiettivi aziendali principali, perseguiti da ciascun componente. Questo concetto è stato introdotto da Eric Evans nel suo libro,Domain-Driven Design: Tackling Complexity in the Heart of Software (Boston: Addison-Wesley Professional, 2003). Per informazioni su come utilizzare la progettazione basata sul dominio con il modello del fico strangolatore (Strangler Fig), consulta la sezione Modernizzazione incrementale dei servizi Web Microsoft ASP.NET (ASMX) legacy utilizzando container e il Gateway Amazon API.

DOTT.

Vedi disaster recovery.

rilevamento della deriva

Tracciamento delle deviazioni da una configurazione di base. Ad esempio, puoi utilizzarlo AWS CloudFormation per rilevare la deriva nelle risorse di sistema oppure puoi usarlo AWS Control Tower per rilevare cambiamenti nella tua landing zone che potrebbero influire sulla conformità ai requisiti di governance.

DVSM

Vedi la mappatura del flusso di valore dello sviluppo.

E

EDA

Vedi analisi esplorativa dei dati.

MODIFICA

Vedi scambio elettronico di dati.

edge computing

La tecnologia che aumenta la potenza di calcolo per i dispositivi intelligenti all'edge di una rete IoT. Rispetto al cloud computing, l'edge computing può ridurre la latenza di comunicazione e migliorare i tempi di risposta.

scambio elettronico di dati (EDI)

Lo scambio automatizzato di documenti aziendali tra organizzazioni. Per ulteriori informazioni, vedere Cos'è lo scambio elettronico di dati.

crittografia

Un processo di elaborazione che trasforma i dati in chiaro, leggibili dall'uomo, in testo cifrato.

chiave crittografica

Una stringa crittografica di bit randomizzati generata da un algoritmo di crittografia. Le chiavi possono variare di lunghezza e ogni chiave è progettata per essere imprevedibile e univoca.

endianità

L'ordine in cui i byte vengono archiviati nella memoria del computer. I sistemi big-endian memorizzano per primo il byte più importante. I sistemi little-endian memorizzano per primo il byte meno importante.

endpoint

Vedi service endpoint.

servizio endpoint

Un servizio che puoi ospitare in un cloud privato virtuale (VPC) da condividere con altri utenti. Puoi creare un servizio endpoint con AWS PrivateLink e concedere autorizzazioni ad altri Account AWS o a AWS Identity and Access Management (IAM) principali. Questi account o principali possono connettersi al servizio endpoint in privato creando endpoint VPC di interfaccia. Per ulteriori informazioni, consulta Creazione di un servizio endpoint nella documentazione di Amazon Virtual Private Cloud (Amazon VPC).

pianificazione delle risorse aziendali (ERP)

Un sistema che automatizza e gestisce i processi aziendali chiave (come contabilità, MES e gestione dei progetti) per un'azienda.

crittografia envelope

Il processo di crittografia di una chiave di crittografia con un'altra chiave di crittografia. Per ulteriori informazioni, vedete Envelope encryption nella documentazione AWS Key Management Service (AWS KMS).

ambiente

Un'istanza di un'applicazione in esecuzione. Di seguito sono riportati i tipi di ambiente più comuni nel cloud computing:

- ambiente di sviluppo: un'istanza di un'applicazione in esecuzione disponibile solo per il team principale responsabile della manutenzione dell'applicazione. Gli ambienti di sviluppo vengono utilizzati per testare le modifiche prima di promuoverle negli ambienti superiori. Questo tipo di ambiente viene talvolta definito ambiente di test.
- ambienti inferiori: tutti gli ambienti di sviluppo di un'applicazione, ad esempio quelli utilizzati per le build e i test iniziali.

- ambiente di produzione: un'istanza di un'applicazione in esecuzione a cui gli utenti finali possono accedere. In una pipeline CI/CD, l'ambiente di produzione è l'ultimo ambiente di implementazione.
- ambienti superiori: tutti gli ambienti a cui possono accedere utenti diversi dal team di sviluppo principale. Si può trattare di un ambiente di produzione, ambienti di preproduzione e ambienti per i test di accettazione da parte degli utenti.

epica

Nelle metodologie agili, categorie funzionali che aiutano a organizzare e dare priorità al lavoro. Le epiche forniscono una descrizione di alto livello dei requisiti e delle attività di implementazione. Ad esempio, le epopee della sicurezza AWS CAF includono la gestione delle identità e degli accessi, i controlli investigativi, la sicurezza dell'infrastruttura, la protezione dei dati e la risposta agli incidenti. Per ulteriori informazioni sulle epiche, consulta la strategia di migrazione AWS , consulta la guida all'implementazione del programma.

ERP

Vedi pianificazione delle risorse aziendali.

analisi esplorativa dei dati (EDA)

Il processo di analisi di un set di dati per comprenderne le caratteristiche principali. Si raccolgono o si aggregano dati e quindi si eseguono indagini iniziali per trovare modelli, rilevare anomalie e verificare ipotesi. L'EDA viene eseguita calcolando statistiche di riepilogo e creando visualizzazioni di dati.

F

tabella dei fatti

Il tavolo centrale con schema a stella. Memorizza dati quantitativi sulle operazioni aziendali. In genere, una tabella dei fatti contiene due tipi di colonne: quelle che contengono misure e quelle che contengono una chiave esterna per una tabella di dimensioni.

fallire velocemente

Una filosofia che utilizza test frequenti e incrementali per ridurre il ciclo di vita dello sviluppo. È una parte fondamentale di un approccio agile.

limite di isolamento dei guasti

Nel Cloud AWS, un limite come una zona di disponibilità Regione AWS, un piano di controllo o un piano dati che limita l'effetto di un errore e aiuta a migliorare la resilienza dei carichi di lavoro. Per ulteriori informazioni, consulta AWS Fault Isolation Boundaries.

ramo di funzionalità

Vedi filiale.

caratteristiche

I dati di input che usi per fare una previsione. Ad esempio, in un contesto di produzione, le caratteristiche potrebbero essere immagini acquisite periodicamente dalla linea di produzione.

importanza delle caratteristiche

Quanto è importante una caratteristica per le previsioni di un modello. Di solito viene espresso come punteggio numerico che può essere calcolato con varie tecniche, come Shapley Additive Explanations (SHAP) e gradienti integrati. Per ulteriori informazioni, consulta Interpretabilità del modello di machine learning con AWS.

trasformazione delle funzionalità

Per ottimizzare i dati per il processo di machine learning, incluso l'arricchimento dei dati con fonti aggiuntive, il dimensionamento dei valori o l'estrazione di più set di informazioni da un singolo campo di dati. Ciò consente al modello di ML di trarre vantaggio dai dati. Ad esempio, se suddividi la data "2021-05-27 00:15:37" in "2021", "maggio", "giovedì" e "15", puoi aiutare l'algoritmo di apprendimento ad apprendere modelli sfumati associati a diversi componenti dei dati.

prompt con pochi scatti

Fornire a un LLM un numero limitato di esempi che dimostrino l'attività e il risultato desiderato prima di chiedergli di eseguire un'attività simile. Questa tecnica è un'applicazione dell'apprendimento contestuale, in cui i modelli imparano da esempi (immagini) incorporati nei prompt. I prompt con pochi passaggi possono essere efficaci per attività che richiedono una formattazione, un ragionamento o una conoscenza del dominio specifici. Vedi anche zero-shot prompting.

FGAC

Vedi il controllo granulare degli accessi.

controllo granulare degli accessi (FGAC)

L'uso di più condizioni per consentire o rifiutare una richiesta di accesso.

migrazione flash-cut

Un metodo di migrazione del database che utilizza la replica continua dei dati tramite l'acquisizione dei dati delle modifiche per migrare i dati nel più breve tempo possibile, anziché utilizzare un approccio graduale. L'obiettivo è ridurre al minimo i tempi di inattività.

FM

Vedi il modello di base.

modello di fondazione (FM)

Una grande rete neurale di deep learning che si è addestrata su enormi set di dati generalizzati e non etichettati. FMs sono in grado di svolgere un'ampia varietà di attività generali, come comprendere il linguaggio, generare testo e immagini e conversare in linguaggio naturale. Per ulteriori informazioni, consulta Cosa sono i modelli Foundation.

G

AI generativa

Un sottoinsieme di modelli di intelligenza artificiale che sono stati addestrati su grandi quantità di dati e che possono utilizzare un semplice prompt di testo per creare nuovi contenuti e artefatti, come immagini, video, testo e audio. Per ulteriori informazioni, consulta Cos'è l'IA generativa.

blocco geografico

Vedi restrizioni geografiche.

limitazioni geografiche (blocco geografico)

In Amazon CloudFront, un'opzione per impedire agli utenti di determinati paesi di accedere alle distribuzioni di contenuti. Puoi utilizzare un elenco consentito o un elenco di blocco per specificare i paesi approvati e vietati. Per ulteriori informazioni, consulta Limitare la distribuzione geografica dei contenuti nella CloudFront documentazione.

Flusso di lavoro di GitFlow

Un approccio in cui gli ambienti inferiori e superiori utilizzano rami diversi in un repository di codice di origine. Il flusso di lavoro Gitflow è considerato obsoleto e il flusso di lavoro basato su trunk è l'approccio moderno e preferito.

immagine dorata

Un'istantanea di un sistema o di un software che viene utilizzata come modello per distribuire nuove istanze di quel sistema o software. Ad esempio, nella produzione, un'immagine dorata può essere utilizzata per fornire software su più dispositivi e contribuire a migliorare la velocità, la scalabilità e la produttività nelle operazioni di produzione dei dispositivi.

strategia greenfield

L'assenza di infrastrutture esistenti in un nuovo ambiente. Quando si adotta una strategia greenfield per un'architettura di sistema, è possibile selezionare tutte le nuove tecnologie senza il vincolo della compatibilità con l'infrastruttura esistente, nota anche come brownfield. Per l'espansione dell'infrastruttura esistente, è possibile combinare strategie brownfield e greenfield.

guardrail

Una regola di alto livello che aiuta a governare le risorse, le politiche e la conformità tra le unità organizzative (). OUs I guardrail preventivi applicano le policy per garantire l'allineamento agli standard di conformità. Vengono implementati utilizzando le policy di controllo dei servizi e i limiti delle autorizzazioni IAM. I guardrail di rilevamento rilevano le violazioni delle policy e i problemi di conformità e generano avvisi per porvi rimedio. Sono implementati utilizzando Amazon AWS Config AWS Security Hub GuardDuty AWS Trusted Advisor, Amazon Inspector e controlli personalizzati AWS Lambda .

H

AH

Vedi disponibilità elevata.

migrazione di database eterogenea

Migrazione del database di origine in un database di destinazione che utilizza un motore di database diverso (ad esempio, da Oracle ad Amazon Aurora). La migrazione eterogenea fa in

genere parte di uno sforzo di riprogettazione e la conversione dello schema può essere un'attività complessa. AWS offre AWS SCT che aiuta con le conversioni dello schema.

alta disponibilità (HA)

La capacità di un carico di lavoro di funzionare in modo continuo, senza intervento, in caso di sfide o disastri. I sistemi HA sono progettati per il failover automatico, fornire costantemente prestazioni di alta qualità e gestire carichi e guasti diversi con un impatto minimo sulle prestazioni.

modernizzazione storica

Un approccio utilizzato per modernizzare e aggiornare i sistemi di tecnologia operativa (OT) per soddisfare meglio le esigenze dell'industria manifatturiera. Uno storico è un tipo di database utilizzato per raccogliere e archiviare dati da varie fonti in una fabbrica.

dati di esclusione

Una parte di dati storici etichettati che viene trattenuta da un set di dati utilizzata per addestrare un modello di apprendimento automatico. È possibile utilizzare i dati di holdout per valutare le prestazioni del modello confrontando le previsioni del modello con i dati di holdout.

migrazione di database omogenea

Migrazione del database di origine in un database di destinazione che condivide lo stesso motore di database (ad esempio, da Microsoft SQL Server ad Amazon RDS per SQL Server). La migrazione omogenea fa in genere parte di un'operazione di rehosting o ridefinizione della piattaforma. Per migrare lo schema è possibile utilizzare le utilità native del database.

dati caldi

Dati a cui si accede frequentemente, come dati in tempo reale o dati di traduzione recenti. Questi dati richiedono in genere un livello o una classe di storage ad alte prestazioni per fornire risposte rapide alle query.

hotfix

Una soluzione urgente per un problema critico in un ambiente di produzione. A causa della sua urgenza, un hotfix viene in genere creato al di fuori del tipico DevOps flusso di lavoro di rilascio.

periodo di hypercare

Subito dopo la conversione, il periodo di tempo in cui un team di migrazione gestisce e monitora le applicazioni migrate nel cloud per risolvere eventuali problemi. In genere, questo periodo dura

da 1 a 4 giorni. Al termine del periodo di hypercare, il team addetto alla migrazione in genere trasferisce la responsabilità delle applicazioni al team addetto alle operazioni cloud.

I

IaC

Considera l'infrastruttura come codice.

Policy basata su identità

Una policy associata a uno o più principi IAM che definisce le relative autorizzazioni all'interno dell' Cloud AWS ambiente.

applicazione inattiva

Un'applicazione che prevede un uso di CPU e memoria medio compreso tra il 5% e il 20% in un periodo di 90 giorni. In un progetto di migrazione, è normale ritirare queste applicazioni o mantenerle on-premise.

IIoT

Vedi Industrial Internet of Things.

infrastruttura immutabile

Un modello che implementa una nuova infrastruttura per i carichi di lavoro di produzione anziché aggiornare, applicare patch o modificare l'infrastruttura esistente. Le infrastrutture immutabili sono intrinsecamente più coerenti, affidabili e prevedibili delle infrastrutture mutabili. Per ulteriori informazioni, consulta la best practice Deploy using immutable infrastructure in Well-Architected AWS Framework.

VPC in ingresso (ingress)

In un'architettura AWS multi-account, un VPC che accetta, ispeziona e indirizza le connessioni di rete dall'esterno di un'applicazione. La AWS Security Reference Architecture consiglia di configurare l'account di rete con funzionalità in entrata, in uscita e di ispezione VPCs per proteggere l'interfaccia bidirezionale tra l'applicazione e la rete Internet in generale.

migrazione incrementale

Una strategia di conversione in cui si esegue la migrazione dell'applicazione in piccole parti anziché eseguire una conversione singola e completa. Ad esempio, inizialmente potresti spostare

solo alcuni microservizi o utenti nel nuovo sistema. Dopo aver verificato che tutto funzioni correttamente, puoi spostare in modo incrementale microservizi o utenti aggiuntivi fino alla disattivazione del sistema legacy. Questa strategia riduce i rischi associati alle migrazioni di grandi dimensioni.

Industria 4.0

Un termine introdotto da Klaus Schwab nel 2016 per riferirsi alla modernizzazione dei processi di produzione attraverso progressi in termini di connettività, dati in tempo reale, automazione, analisi e AI/ML.

infrastruttura

Tutte le risorse e gli asset contenuti nell'ambiente di un'applicazione.

infrastruttura come codice (IaC)

Il processo di provisioning e gestione dell'infrastruttura di un'applicazione tramite un insieme di file di configurazione. Il processo IaC è progettato per aiutarti a centralizzare la gestione dell'infrastruttura, a standardizzare le risorse e a dimensionare rapidamente, in modo che i nuovi ambienti siano ripetibili, affidabili e coerenti.

IIoInternet delle cose industriale (T)

L'uso di sensori e dispositivi connessi a Internet nei settori industriali, come quello manifatturiero, energetico, automobilistico, sanitario, delle scienze della vita e dell'agricoltura. Per ulteriori informazioni, vedere Creazione di una strategia di trasformazione digitale per l'Internet of Things (IIoT) industriale.

VPC di ispezione

In un'architettura AWS multi-account, un VPC centralizzato che gestisce le ispezioni del traffico di rete tra VPCs (nello stesso o in modo diverso Regioni AWS), Internet e le reti locali. La AWS Security Reference Architecture consiglia di configurare l'account di rete con informazioni in entrata, in uscita e di ispezione VPCs per proteggere l'interfaccia bidirezionale tra l'applicazione e Internet in generale.

Internet of Things (IoT)

La rete di oggetti fisici connessi con sensori o processori incorporati che comunicano con altri dispositivi e sistemi tramite Internet o una rete di comunicazione locale. Per ulteriori informazioni, consulta Cos'è l'IoT?

interpretabilità

Una caratteristica di un modello di machine learning che descrive il grado in cui un essere umano è in grado di comprendere in che modo le previsioni del modello dipendono dai suoi input. Per ulteriori informazioni, vedere Interpretabilità del modello di machine learning con. AWS

IoT

Vedi Internet of Things.

libreria di informazioni IT (ITIL)

Una serie di best practice per offrire servizi IT e allinearli ai requisiti aziendali. ITIL fornisce le basi per ITSM.

gestione dei servizi IT (ITSM)

Attività associate alla progettazione, implementazione, gestione e supporto dei servizi IT per un'organizzazione. Per informazioni sull'integrazione delle operazioni cloud con gli strumenti ITSM, consulta la guida all'integrazione delle operazioni.

ITIL

Vedi la libreria di informazioni IT.

ITSM

Vedi Gestione dei servizi IT.

L

controllo degli accessi basato su etichette (LBAC)

Un'implementazione del controllo di accesso obbligatorio (MAC) in cui agli utenti e ai dati stessi viene assegnato esplicitamente un valore di etichetta di sicurezza. L'intersezione tra l'etichetta di sicurezza utente e l'etichetta di sicurezza dei dati determina quali righe e colonne possono essere visualizzate dall'utente.

zona di destinazione

Una landing zone è un AWS ambiente multi-account ben progettato, scalabile e sicuro. Questo è un punto di partenza dal quale le organizzazioni possono avviare e distribuire rapidamente carichi di lavoro e applicazioni con fiducia nel loro ambiente di sicurezza e infrastruttura. Per ulteriori

informazioni sulle zone di destinazione, consulta la sezione Configurazione di un ambiente AWS multi-account sicuro e scalabile.

modello linguistico di grandi dimensioni (LLM)

Un modello di intelligenza artificiale di deep learning preaddestrato su una grande quantità di dati. Un LLM può svolgere più attività, come rispondere a domande, riepilogare documenti, tradurre testo in altre lingue e completare frasi. Per ulteriori informazioni, consulta Cosa sono. LLMs

migrazione su larga scala

Una migrazione di 300 o più server.

BIANCO

Vedi controllo degli accessi basato su etichette.

Privilegio minimo

La best practice di sicurezza per la concessione delle autorizzazioni minime richieste per eseguire un'attività. Per ulteriori informazioni, consulta Applicazione delle autorizzazioni del privilegio minimo nella documentazione di IAM.

eseguire il rehosting (lift and shift)

Vedi 7 R.

sistema little-endian

Un sistema che memorizza per primo il byte meno importante. Vedi anche endianità.

LLM

Vedi modello linguistico di grandi dimensioni.

ambienti inferiori

Vedi ambiente.

M

machine learning (ML)

Un tipo di intelligenza artificiale che utilizza algoritmi e tecniche per il riconoscimento e l'apprendimento di schemi. Il machine learning analizza e apprende dai dati registrati, come i dati

dell'Internet delle cose (IoT), per generare un modello statistico basato su modelli. Per ulteriori informazioni, consulta la sezione Machine learning.

ramo principale

Vedi filiale.

malware

Software progettato per compromettere la sicurezza o la privacy del computer. Il malware potrebbe interrompere i sistemi informatici, divulgare informazioni sensibili o ottenere accessi non autorizzati. Esempi di malware includono virus, worm, ransomware, trojan horse, spyware e keylogger.

servizi gestiti

Servizi AWS per cui AWS gestisce il livello di infrastruttura, il sistema operativo e le piattaforme e si accede agli endpoint per archiviare e recuperare i dati. Amazon Simple Storage Service (Amazon S3) Simple Storage Service (Amazon S3) e Amazon DynamoDB sono esempi di servizi gestiti. Questi sono noti anche come servizi astratti.

sistema di esecuzione della produzione (MES)

Un sistema software per tracciare, monitorare, documentare e controllare i processi di produzione che convertono le materie prime in prodotti finiti in officina.

MAP

Vedi Migration Acceleration Program.

meccanismo

Un processo completo in cui si crea uno strumento, si promuove l'adozione dello strumento e quindi si esaminano i risultati per apportare le modifiche. Un meccanismo è un ciclo che si rafforza e si migliora man mano che funziona. Per ulteriori informazioni, consulta Creazione di meccanismi nel AWS Well-Architected Framework.

account membro

Tutti gli account Account AWS diversi dall'account di gestione che fanno parte di un'organizzazione in. AWS Organizations Un account può essere membro di una sola organizzazione alla volta.

MEH.

Vedi sistema di esecuzione della produzione.

Message Queuing Telemetry Transport (MQTT)

Un protocollo di comunicazione machine-to-machine (M2M) leggero, basato sul modello di pubblicazione/sottoscrizione, per dispositivi IoT con risorse limitate.

microservizio

Un servizio piccolo e indipendente che comunica tramite canali ben definiti ed è in genere di proprietà di piccoli team autonomi. APIs Ad esempio, un sistema assicurativo potrebbe includere microservizi che si riferiscono a funzionalità aziendali, come vendite o marketing, o sottodomini, come acquisti, reclami o analisi. I vantaggi dei microservizi includono agilità, dimensionamento flessibile, facilità di implementazione, codice riutilizzabile e resilienza. Per ulteriori informazioni, consulta Integrazione dei microservizi utilizzando servizi serverless. AWS

architettura di microservizi

Un approccio alla creazione di un'applicazione con componenti indipendenti che eseguono ogni processo applicativo come microservizio. Questi microservizi comunicano attraverso un'interfaccia ben definita utilizzando sistemi leggeri. APIs Ogni microservizio in questa architettura può essere aggiornato, distribuito e dimensionato per soddisfare la richiesta di funzioni specifiche di un'applicazione. Per ulteriori informazioni, vedere Implementazione dei microservizi su. AWS

Programma di accelerazione della migrazione (MAP)

Un AWS programma che fornisce consulenza, supporto, formazione e servizi per aiutare le organizzazioni a costruire una solida base operativa per il passaggio al cloud e per contribuire a compensare il costo iniziale delle migrazioni. MAP include una metodologia di migrazione per eseguire le migrazioni precedenti in modo metodico e un set di strumenti per automatizzare e accelerare gli scenari di migrazione comuni.

migrazione su larga scala

Il processo di trasferimento della maggior parte del portfolio di applicazioni sul cloud avviene a ondate, con più applicazioni trasferite a una velocità maggiore in ogni ondata. Questa fase utilizza le migliori pratiche e le lezioni apprese nelle fasi precedenti per implementare una fabbrica di migrazione di team, strumenti e processi per semplificare la migrazione dei carichi di lavoro attraverso l'automazione e la distribuzione agile. Questa è la terza fase della strategia di migrazione AWS.

fabbrica di migrazione

Team interfunzionali che semplificano la migrazione dei carichi di lavoro attraverso approcci automatizzati e agili. I team di Migration Factory in genere includono addetti alle operazioni,

analisti e proprietari aziendali, ingegneri addetti alla migrazione, sviluppatori e DevOps professionisti che lavorano nell'ambito degli sprint. Tra il 20% e il 50% di un portfolio di applicazioni aziendali è costituito da schemi ripetuti che possono essere ottimizzati con un approccio di fabbrica. Per ulteriori informazioni, consulta la discussione sulle fabbriche di migrazione e la Guida alla fabbrica di migrazione al cloud in questo set di contenuti.

metadati di migrazione

Le informazioni sull'applicazione e sul server necessarie per completare la migrazione. Ogni modello di migrazione richiede un set diverso di metadati di migrazione. Esempi di metadati di migrazione includono la sottorete, il gruppo di sicurezza e l'account di destinazione. AWS

modello di migrazione

Un'attività di migrazione ripetibile che descrive in dettaglio la strategia di migrazione, la destinazione della migrazione e l'applicazione o il servizio di migrazione utilizzati. Esempio: riorganizza la migrazione su Amazon EC2 con AWS Application Migration Service.

Valutazione del portfolio di migrazione (MPA)

Uno strumento online che fornisce informazioni per la convalida del business case per la migrazione a. Cloud AWS MPA offre una valutazione dettagliata del portfolio (dimensionamento corretto dei server, prezzi, confronto del TCO, analisi dei costi di migrazione) e pianificazione della migrazione (analisi e raccolta dei dati delle applicazioni, raggruppamento delle applicazioni, prioritizzazione delle migrazioni e pianificazione delle ondate). Lo strumento MPA (richiede l'accesso) è disponibile gratuitamente per tutti i AWS consulenti e i consulenti dei partner APN.

valutazione della preparazione alla migrazione (MRA)

Il processo di acquisizione di informazioni sullo stato di preparazione al cloud di un'organizzazione, l'identificazione dei punti di forza e di debolezza e la creazione di un piano d'azione per colmare le lacune identificate, utilizzando il CAF. AWS Per ulteriori informazioni, consulta la guida di preparazione alla migrazione. MRA è la prima fase della strategia di migrazione AWS.

strategia di migrazione

L'approccio utilizzato per migrare un carico di lavoro verso. Cloud AWS Per ulteriori informazioni, consulta la voce 7 R in questo glossario e consulta Mobilita la tua organizzazione per accelerare le migrazioni su larga scala.

ML

Vedi machine learning.

modernizzazione

Trasformazione di un'applicazione obsoleta (legacy o monolitica) e della relativa infrastruttura in un sistema agile, elastico e altamente disponibile nel cloud per ridurre i costi, aumentare l'efficienza e sfruttare le innovazioni. Per ulteriori informazioni, vedere Strategia per la modernizzazione delle applicazioni in. Cloud AWS

valutazione della preparazione alla modernizzazione

Una valutazione che aiuta a determinare la preparazione alla modernizzazione delle applicazioni di un'organizzazione, identifica vantaggi, rischi e dipendenze e determina in che misura l'organizzazione può supportare lo stato futuro di tali applicazioni. Il risultato della valutazione è uno schema dell'architettura di destinazione, una tabella di marcia che descrive in dettaglio le fasi di sviluppo e le tappe fondamentali del processo di modernizzazione e un piano d'azione per colmare le lacune identificate. Per ulteriori informazioni, vedere Valutazione della preparazione alla modernizzazione per le applicazioni in. Cloud AWS

applicazioni monolitiche (monoliti)

Applicazioni eseguite come un unico servizio con processi strettamente collegati. Le applicazioni monolitiche presentano diversi inconvenienti. Se una funzionalità dell'applicazione registra un picco di domanda, l'intera architettura deve essere dimensionata. L'aggiunta o il miglioramento delle funzionalità di un'applicazione monolitica diventa inoltre più complessa man mano che la base di codice cresce. Per risolvere questi problemi, puoi utilizzare un'architettura di microservizi. Per ulteriori informazioni, consulta la sezione Scomposizione dei monoliti in microservizi.

MAPPA

Vedi Migration Portfolio Assessment.

MQTT

Vedi Message Queuing Telemetry Transport.

classificazione multiclasse

Un processo che aiuta a generare previsioni per più classi (prevedendo uno o più di due risultati). Ad esempio, un modello di machine learning potrebbe chiedere "Questo prodotto è un libro, un'auto o un telefono?" oppure "Quale categoria di prodotti è più interessante per questo cliente?"

infrastruttura mutabile

Un modello che aggiorna e modifica l'infrastruttura esistente per i carichi di lavoro di produzione. Per migliorare la coerenza, l'affidabilità e la prevedibilità, il AWS Well-Architected Framework consiglia l'uso di un'infrastruttura immutabile come best practice.

O

OAC

Vedi Origin Access Control.

QUERCIA

Vedi Origin Access Identity.

OCM

Vedi gestione delle modifiche organizzative.

migrazione offline

Un metodo di migrazione in cui il carico di lavoro di origine viene eliminato durante il processo di migrazione. Questo metodo prevede tempi di inattività prolungati e viene in genere utilizzato per carichi di lavoro piccoli e non critici.

OI

Vedi l'integrazione delle operazioni.

OLA

Vedi accordo a livello operativo.

migrazione online

Un metodo di migrazione in cui il carico di lavoro di origine viene copiato sul sistema di destinazione senza essere messo offline. Le applicazioni connesse al carico di lavoro possono continuare a funzionare durante la migrazione. Questo metodo comporta tempi di inattività pari a zero o comunque minimi e viene in genere utilizzato per carichi di lavoro di produzione critici.

OPC-UA

Vedi Open Process Communications - Unified Architecture.

Comunicazioni a processo aperto - Architettura unificata (OPC-UA)

Un protocollo di comunicazione machine-to-machine (M2M) per l'automazione industriale. OPC-UA fornisce uno standard di interoperabilità con schemi di crittografia, autenticazione e autorizzazione dei dati.

accordo a livello operativo (OLA)

Un accordo che chiarisce quali sono gli impegni reciproci tra i gruppi IT funzionali, a supporto di un accordo sul livello di servizio (SLA).

revisione della prontezza operativa (ORR)

Un elenco di domande e best practice associate che aiutano a comprendere, valutare, prevenire o ridurre la portata degli incidenti e dei possibili guasti. Per ulteriori informazioni, vedere Operational Readiness Reviews (ORR) nel Well-Architected AWS Framework.

tecnologia operativa (OT)

Sistemi hardware e software che interagiscono con l'ambiente fisico per controllare le operazioni, le apparecchiature e le infrastrutture industriali. Nella produzione, l'integrazione di sistemi OT e di tecnologia dell'informazione (IT) è un obiettivo chiave per le trasformazioni dell'Industria 4.0.

integrazione delle operazioni (OI)

Il processo di modernizzazione delle operazioni nel cloud, che prevede la pianificazione, l'automazione e l'integrazione della disponibilità. Per ulteriori informazioni, consulta la guida all'integrazione delle operazioni.

trail organizzativo

Un percorso creato da noi AWS CloudTrail che registra tutti gli eventi di un'organizzazione per tutti Account AWS . AWS Organizations Questo percorso viene creato in ogni Account AWS che fa parte dell'organizzazione e tiene traccia dell'attività in ogni account. Per ulteriori informazioni, consulta Creazione di un percorso per un'organizzazione nella CloudTrail documentazione.

gestione del cambiamento organizzativo (OCM)

Un framework per la gestione di trasformazioni aziendali importanti e che comportano l'interruzione delle attività dal punto di vista delle persone, della cultura e della leadership. OCM aiuta le organizzazioni a prepararsi e passare a nuovi sistemi e strategie accelerando l'adozione del cambiamento, affrontando i problemi di transizione e promuovendo cambiamenti culturali e organizzativi. Nella strategia di AWS migrazione, questo framework si chiama accelerazione delle

persone, a causa della velocità di cambiamento richiesta nei progetti di adozione del cloud. Per ulteriori informazioni, consultare la Guida OCM.

controllo dell'accesso all'origine (OAC)

In CloudFront, un'opzione avanzata per limitare l'accesso per proteggere i contenuti di Amazon Simple Storage Service (Amazon S3). OAC supporta tutti i bucket S3 in generale Regioni AWS, la crittografia lato server con AWS KMS (SSE-KMS) e le richieste dinamiche e dirette al bucket S3. `PUT` `DELETE`

identità di accesso origine (OAI)

Nel CloudFront, un'opzione per limitare l'accesso per proteggere i tuoi contenuti Amazon S3. Quando usi OAI, CloudFront crea un principale con cui Amazon S3 può autenticarsi. I principali autenticati possono accedere ai contenuti in un bucket S3 solo tramite una distribuzione specifica. CloudFront Vedi anche OAC, che fornisce un controllo degli accessi più granulare e avanzato.

ORR

Vedi la revisione della prontezza operativa.

- NON

Vedi la tecnologia operativa.

VPC in uscita (egress)

In un'architettura AWS multi-account, un VPC che gestisce le connessioni di rete avviate dall'interno di un'applicazione. La AWS Security Reference Architecture consiglia di configurare l'account di rete con funzionalità in entrata, in uscita e di ispezione VPCs per proteggere l'interfaccia bidirezionale tra l'applicazione e Internet in generale.

P

limite delle autorizzazioni

Una policy di gestione IAM collegata ai principali IAM per impostare le autorizzazioni massime che l'utente o il ruolo possono avere. Per ulteriori informazioni, consulta Limiti delle autorizzazioni nella documentazione di IAM.

informazioni di identificazione personale (PII)

Informazioni che, se visualizzate direttamente o abbinate ad altri dati correlati, possono essere utilizzate per dedurre ragionevolmente l'identità di un individuo. Esempi di informazioni personali includono nomi, indirizzi e informazioni di contatto.

Informazioni che consentono l'identificazione personale degli utenti

Visualizza le informazioni di identificazione personale.

playbook

Una serie di passaggi predefiniti che raccolgono il lavoro associato alle migrazioni, come l'erogazione delle funzioni operative principali nel cloud. Un playbook può assumere la forma di script, runbook automatici o un riepilogo dei processi o dei passaggi necessari per gestire un ambiente modernizzato.

PLC

Vedi controllore logico programmabile.

PLM

Vedi la gestione del ciclo di vita del prodotto.

policy

Un oggetto in grado di definire le autorizzazioni (vedi politica basata sull'identità), specificare le condizioni di accesso (vedi politicabasata sulle risorse) o definire le autorizzazioni massime per tutti gli account di un'organizzazione in (vedi politica di controllo dei servizi). AWS Organizations

persistenza poliglotta

Scelta indipendente della tecnologia di archiviazione di dati di un microservizio in base ai modelli di accesso ai dati e ad altri requisiti. Se i microservizi utilizzano la stessa tecnologia di archiviazione di dati, possono incontrare problemi di implementazione o registrare prestazioni scadenti. I microservizi vengono implementati più facilmente e ottengono prestazioni e scalabilità migliori se utilizzano l'archivio dati più adatto alle loro esigenze. Per ulteriori informazioni, consulta la sezione Abilitazione della persistenza dei dati nei microservizi.

valutazione del portfolio

Un processo di scoperta, analisi e definizione delle priorità del portfolio di applicazioni per pianificare la migrazione. Per ulteriori informazioni, consulta la pagina Valutazione della preparazione alla migrazione.

predicate

Una condizione di interrogazione che restituisce o, in genere, si trova in una clausola`WHERE`. `true` `false`

predicato pushdown

Una tecnica di ottimizzazione delle query del database che filtra i dati della query prima del trasferimento. Ciò riduce la quantità di dati che devono essere recuperati ed elaborati dal database relazionale e migliora le prestazioni delle query.

controllo preventivo

Un controllo di sicurezza progettato per impedire il verificarsi di un evento. Questi controlli sono la prima linea di difesa per impedire accessi non autorizzati o modifiche indesiderate alla rete. Per ulteriori informazioni, consulta Controlli preventivi in Implementazione dei controlli di sicurezza in AWS.

principale

Un'entità in AWS grado di eseguire azioni e accedere alle risorse. Questa entità è in genere un utente root per un Account AWS ruolo IAM o un utente. Per ulteriori informazioni, consulta Principali in Termini e concetti dei ruoli nella documentazione di IAM.

privacy fin dalla progettazione

Un approccio di ingegneria dei sistemi che tiene conto della privacy durante l'intero processo di sviluppo.

zone ospitate private

Un contenitore che contiene informazioni su come desideri che Amazon Route 53 risponda alle query DNS per un dominio e i relativi sottodomini all'interno di uno o più. VPCs Per ulteriori informazioni, consulta Utilizzo delle zone ospitate private nella documentazione di Route 53.

controllo proattivo

Un controllo di sicurezza progettato per impedire l'implementazione di risorse non conformi. Questi controlli analizzano le risorse prima del loro provisioning. Se la risorsa non è conforme al controllo, non viene fornita. Per ulteriori informazioni, consulta la guida di riferimento sui controlli nella AWS Control Tower documentazione e consulta Controlli proattivi in Implementazione dei controlli di sicurezza su. AWS

gestione del ciclo di vita del prodotto (PLM)

La gestione dei dati e dei processi di un prodotto durante l'intero ciclo di vita, dalla progettazione, sviluppo e lancio, attraverso la crescita e la maturità, fino al declino e alla rimozione.

Ambiente di produzione

Vedi ambiente.

controllore logico programmabile (PLC)

Nella produzione, un computer altamente affidabile e adattabile che monitora le macchine e automatizza i processi di produzione.

concatenamento rapido

Utilizzo dell'output di un prompt LLM come input per il prompt successivo per generare risposte migliori. Questa tecnica viene utilizzata per suddividere un'attività complessa in sottoattività o per perfezionare o espandere iterativamente una risposta preliminare. Aiuta a migliorare l'accuratezza e la pertinenza delle risposte di un modello e consente risultati più granulari e personalizzati.

pseudonimizzazione

Il processo di sostituzione degli identificatori personali in un set di dati con valori segnaposto. La pseudonimizzazione può aiutare a proteggere la privacy personale. I dati pseudonimizzati sono ancora considerati dati personali.

publish/subscribe (pub/sub)

Un modello che consente comunicazioni asincrone tra microservizi per migliorare la scalabilità e la reattività. Ad esempio, in un MES basato su microservizi, un microservizio può pubblicare messaggi di eventi su un canale a cui altri microservizi possono abbonarsi. Il sistema può aggiungere nuovi microservizi senza modificare il servizio di pubblicazione.

Q

Piano di query

Una serie di passaggi, come le istruzioni, utilizzati per accedere ai dati in un sistema di database relazionale SQL.

regressione del piano di query

Quando un ottimizzatore del servizio di database sceglie un piano non ottimale rispetto a prima di una determinata modifica all'ambiente di database. Questo può essere causato da modifiche a statistiche, vincoli, impostazioni dell'ambiente, associazioni dei parametri di query e aggiornamenti al motore di database.

R

Matrice RACI

Vedi responsabile, responsabile, consultato, informato (RACI).

STRACCIO

Vedi Retrieval Augmented Generation.

ransomware

Un software dannoso progettato per bloccare l'accesso a un sistema informatico o ai dati fino a quando non viene effettuato un pagamento.

Matrice RASCI

Vedi responsabile, responsabile, consultato, informato (RACI).

RCAC

Vedi controllo dell'accesso a righe e colonne.

replica di lettura

Una copia di un database utilizzata per scopi di sola lettura. È possibile indirizzare le query alla replica di lettura per ridurre il carico sul database principale.

riprogettare

Vedi 7 Rs.

obiettivo del punto di ripristino (RPO)

Il periodo di tempo massimo accettabile dall'ultimo punto di ripristino dei dati. Questo determina ciò che si considera una perdita di dati accettabile tra l'ultimo punto di ripristino e l'interruzione del servizio.

obiettivo del tempo di ripristino (RTO)

Il ritardo massimo accettabile tra l'interruzione del servizio e il ripristino del servizio.

rifattorizzare

Vedi 7 R.

Regione

Una raccolta di AWS risorse in un'area geografica. Ciascuna Regione AWS è isolata e indipendente dalle altre per fornire tolleranza agli errori, stabilità e resilienza. Per ulteriori informazioni, consulta Specificare cosa può usare Regioni AWS il tuo account.

regressione

Una tecnica di ML che prevede un valore numerico. Ad esempio, per risolvere il problema "A che prezzo verrà venduta questa casa?" un modello di ML potrebbe utilizzare un modello di regressione lineare per prevedere il prezzo di vendita di una casa sulla base di dati noti sulla casa (ad esempio, la metratura).

riospitare

Vedi 7 R.

rilascio

In un processo di implementazione, l'atto di promuovere modifiche a un ambiente di produzione.

trasferisco

Vedi 7 Rs.

ripiattaforma

Vedi 7 Rs.

riacquisto

Vedi 7 Rs.

resilienza

La capacità di un'applicazione di resistere o ripristinare le interruzioni. L'elevata disponibilità e il disaster recovery sono considerazioni comuni quando si pianifica la resilienza in. Cloud AWSPer ulteriori informazioni, vedere Cloud AWS Resilience.

policy basata su risorse

Una policy associata a una risorsa, ad esempio un bucket Amazon S3, un endpoint o una chiave di crittografia. Questo tipo di policy specifica a quali principali è consentito l'accesso, le azioni supportate e qualsiasi altra condizione che deve essere soddisfatta.

matrice di assegnazione di responsabilità (RACI)

Una matrice che definisce i ruoli e le responsabilità di tutte le parti coinvolte nelle attività di migrazione e nelle operazioni cloud. Il nome della matrice deriva dai tipi di responsabilità definiti nella matrice: responsabile (R), responsabile (A), consultato (C) e informato (I). Il tipo di supporto (S) è facoltativo. Se includi il supporto, la matrice viene chiamata matrice RASCI e, se la escludi, viene chiamata matrice RACI.

controllo reattivo

Un controllo di sicurezza progettato per favorire la correzione di eventi avversi o deviazioni dalla baseline di sicurezza. Per ulteriori informazioni, consulta Controlli reattivi in Implementazione dei controlli di sicurezza in AWS.

retain

Vedi 7 R.

andare in pensione

Vedi 7 Rs.

Retrieval Augmented Generation (RAG)

Una tecnologia di intelligenza artificiale generativa in cui un LLM fa riferimento a una fonte di dati autorevole esterna alle sue fonti di dati di formazione prima di generare una risposta. Ad esempio, un modello RAG potrebbe eseguire una ricerca semantica nella knowledge base o nei dati personalizzati di un'organizzazione. Per ulteriori informazioni, consulta Cos'è il RAG.

rotazione

Processo di aggiornamento periodico di un segreto per rendere più difficile l'accesso alle credenziali da parte di un utente malintenzionato.

controllo dell'accesso a righe e colonne (RCAC)

L'uso di espressioni SQL di base e flessibili con regole di accesso definite. RCAC è costituito da autorizzazioni di riga e maschere di colonna.

RPO

Vedi l'obiettivo del punto di ripristino.

RTO

Vedi l'obiettivo del tempo di ripristino.

runbook

Un insieme di procedure manuali o automatizzate necessarie per eseguire un'attività specifica. In genere sono progettati per semplificare operazioni o procedure ripetitive con tassi di errore elevati.

S

SAML 2.0

Uno standard aperto utilizzato da molti provider di identità (IdPs). Questa funzionalità abilita il single sign-on (SSO) federato, in modo che gli utenti possano accedere AWS Management Console o chiamare le operazioni AWS API senza che tu debba creare un utente in IAM per tutti i membri dell'organizzazione. Per ulteriori informazioni sulla federazione basata su SAML 2.0, consulta Informazioni sulla federazione basata su SAML 2.0 nella documentazione di IAM.

SCADA

Vedi controllo di supervisione e acquisizione dati.

SCP

Vedi la politica di controllo del servizio.

Secret

In AWS Secrets Manager, informazioni riservate o riservate, come una password o le credenziali utente, archiviate in forma crittografata. È costituito dal valore segreto e dai relativi metadati. Il valore segreto può essere binario, una stringa singola o più stringhe. Per ulteriori informazioni, consulta Cosa c'è in un segreto di Secrets Manager? nella documentazione di Secrets Manager.

sicurezza fin dalla progettazione

Un approccio di ingegneria dei sistemi che tiene conto della sicurezza durante l'intero processo di sviluppo.

controllo di sicurezza

Un guardrail tecnico o amministrativo che impedisce, rileva o riduce la capacità di un autore di minacce di sfruttare una vulnerabilità di sicurezza. Esistono quattro tipi principali di controlli di sicurezza: preventivi, investigativi, reattivi e proattivi.

rafforzamento della sicurezza

Il processo di riduzione della superficie di attacco per renderla più resistente agli attacchi. Può includere azioni come la rimozione di risorse che non sono più necessarie, l'implementazione di best practice di sicurezza che prevedono la concessione del privilegio minimo o la disattivazione di funzionalità non necessarie nei file di configurazione.

sistema di gestione delle informazioni e degli eventi di sicurezza (SIEM)

Strumenti e servizi che combinano sistemi di gestione delle informazioni di sicurezza (SIM) e sistemi di gestione degli eventi di sicurezza (SEM). Un sistema SIEM raccoglie, monitora e analizza i dati da server, reti, dispositivi e altre fonti per rilevare minacce e violazioni della sicurezza e generare avvisi.

automazione della risposta alla sicurezza

Un'azione predefinita e programmata progettata per rispondere o porre rimedio automaticamente a un evento di sicurezza. Queste automazioni fungono da controlli di sicurezza investigativi o reattivi che aiutano a implementare le migliori pratiche di sicurezza. AWS Esempi di azioni di risposta automatizzate includono la modifica di un gruppo di sicurezza VPC, l'applicazione di patch a un'istanza EC2 Amazon o la rotazione delle credenziali.

Crittografia lato server

Crittografia dei dati a destinazione, da parte di chi li riceve. Servizio AWS

Policy di controllo dei servizi (SCP)

Una politica che fornisce il controllo centralizzato sulle autorizzazioni per tutti gli account di un'organizzazione in. AWS Organizations SCPs definire barriere o fissare limiti alle azioni che un amministratore può delegare a utenti o ruoli. È possibile utilizzarli SCPs come elenchi consentiti o elenchi di rifiuto, per specificare quali servizi o azioni sono consentiti o proibiti. Per ulteriori informazioni, consulta le politiche di controllo del servizio nella AWS Organizations documentazione.

endpoint del servizio

L'URL del punto di ingresso per un Servizio AWS. Puoi utilizzare l'endpoint per connetterti a livello di programmazione al servizio di destinazione. Per ulteriori informazioni, consulta Endpoint del Servizio AWS nei Riferimenti generali di AWS.

accordo sul livello di servizio (SLA)

Un accordo che chiarisce ciò che un team IT promette di offrire ai propri clienti, ad esempio l'operatività e le prestazioni del servizio.

indicatore del livello di servizio (SLI)

Misurazione di un aspetto prestazionale di un servizio, ad esempio il tasso di errore, la disponibilità o la velocità effettiva.

obiettivo a livello di servizio (SLO)

Una metrica target che rappresenta lo stato di un servizio, misurato da un indicatore del livello di servizio.

Modello di responsabilità condivisa

Un modello che descrive la responsabilità condivisa AWS per la sicurezza e la conformità del cloud. AWS è responsabile della sicurezza del cloud, mentre tu sei responsabile della sicurezza nel cloud. Per ulteriori informazioni, consulta Modello di responsabilità condivisa.

SIEM

Vedi il sistema di gestione delle informazioni e degli eventi sulla sicurezza.

punto di errore singolo (SPOF)

Un guasto in un singolo componente critico di un'applicazione che può disturbare il sistema.

SLAM

Vedi il contratto sul livello di servizio.

SLI

Vedi l'indicatore del livello di servizio.

LENTA

Vedi obiettivo del livello di servizio.

split-and-seed modello

Un modello per dimensionare e accelerare i progetti di modernizzazione. Man mano che vengono definite nuove funzionalità e versioni dei prodotti, il team principale si divide per creare nuovi team di prodotto. Questo aiuta a dimensionare le capacità e i servizi dell'organizzazione, migliora la produttività degli sviluppatori e supporta una rapida innovazione. Per ulteriori informazioni, vedere Approccio graduale alla modernizzazione delle applicazioni in. Cloud AWS

SPOF

Vedi punto di errore singolo.

schema a stella

Una struttura organizzativa di database che utilizza un'unica tabella dei fatti di grandi dimensioni per archiviare i dati transazionali o misurati e utilizza una o più tabelle dimensionali più piccole per memorizzare gli attributi dei dati. Questa struttura è progettata per l'uso in un data warehouse o per scopi di business intelligence.

modello del fico strangolatore

Un approccio alla modernizzazione dei sistemi monolitici mediante la riscrittura e la sostituzione incrementali delle funzionalità del sistema fino alla disattivazione del sistema legacy. Questo modello utilizza l'analogia di una pianta di fico che cresce fino a diventare un albero robusto e alla fine annienta e sostituisce il suo ospite. Il modello è stato introdotto da Martin Fowler come metodo per gestire il rischio durante la riscrittura di sistemi monolitici. Per un esempio di come applicare questo modello, consulta Modernizzazione incrementale dei servizi Web legacy di Microsoft ASP.NET (ASMX) mediante container e Gateway Amazon API.

sottorete

Un intervallo di indirizzi IP nel VPC. Una sottorete deve risiedere in una singola zona di disponibilità.

controllo di supervisione e acquisizione dati (SCADA)

Nella produzione, un sistema che utilizza hardware e software per monitorare gli asset fisici e le operazioni di produzione.

crittografia simmetrica

Un algoritmo di crittografia che utilizza la stessa chiave per crittografare e decrittografare i dati.

test sintetici

Test di un sistema in modo da simulare le interazioni degli utenti per rilevare potenziali problemi o monitorare le prestazioni. Puoi usare Amazon CloudWatch Synthetics per creare questi test.

prompt di sistema

Una tecnica per fornire contesto, istruzioni o linee guida a un LLM per indirizzarne il comportamento. I prompt di sistema aiutano a impostare il contesto e stabilire regole per le interazioni con gli utenti.

T

tags

Coppie chiave-valore che fungono da metadati per l'organizzazione delle risorse. AWS Con i tag è possibile a gestire, identificare, organizzare, cercare e filtrare le risorse. Per ulteriori informazioni, consulta Tagging delle risorse AWS.

variabile di destinazione

Il valore che stai cercando di prevedere nel machine learning supervisionato. Questo è indicato anche come variabile di risultato. Ad esempio, in un ambiente di produzione la variabile di destinazione potrebbe essere un difetto del prodotto.

elenco di attività

Uno strumento che viene utilizzato per tenere traccia dei progressi tramite un runbook. Un elenco di attività contiene una panoramica del runbook e un elenco di attività generali da completare. Per ogni attività generale, include la quantità stimata di tempo richiesta, il proprietario e lo stato di avanzamento.

Ambiente di test

Vedi ambiente.

training

Fornire dati da cui trarre ispirazione dal modello di machine learning. I dati di training devono contenere la risposta corretta. L'algoritmo di apprendimento trova nei dati di addestramento i pattern che mappano gli attributi dei dati di input al target (la risposta che si desidera prevedere). Produce un modello di ML che acquisisce questi modelli. Puoi quindi utilizzare il modello di ML per creare previsioni su nuovi dati di cui non si conosce il target.

Transit Gateway

Un hub di transito di rete che puoi utilizzare per interconnettere le tue reti VPCs e quelle locali. Per ulteriori informazioni, consulta Cos'è un gateway di transito nella AWS Transit Gateway documentazione.

flusso di lavoro basato su trunk

Un approccio in cui gli sviluppatori creano e testano le funzionalità localmente in un ramo di funzionalità e quindi uniscono tali modifiche al ramo principale. Il ramo principale viene quindi integrato negli ambienti di sviluppo, preproduzione e produzione, in sequenza.

Accesso attendibile

Concessione delle autorizzazioni a un servizio specificato dall'utente per eseguire attività all'interno dell'organizzazione AWS Organizations e nei suoi account per conto dell'utente. Il servizio attendibile crea un ruolo collegato al servizio in ogni account, quando tale ruolo è necessario, per eseguire attività di gestione per conto dell'utente. Per ulteriori informazioni, consulta Utilizzo AWS Organizations con altri AWS servizi nella AWS Organizations documentazione.

regolazione

Modificare alcuni aspetti del processo di training per migliorare la precisione del modello di ML. Ad esempio, puoi addestrare il modello di ML generando un set di etichette, aggiungendo etichette e quindi ripetendo questi passaggi più volte con impostazioni diverse per ottimizzare il modello.

team da due pizze

Una piccola DevOps squadra che puoi sfamare con due pizze. Un team composto da due persone garantisce la migliore opportunità possibile di collaborazione nello sviluppo del software.

U

incertezza

Un concetto che si riferisce a informazioni imprecise, incomplete o sconosciute che possono minare l'affidabilità dei modelli di machine learning predittivi. Esistono due tipi di incertezza: l'incertezza epistemica, che è causata da dati limitati e incompleti, mentre l'incertezza aleatoria è causata dal rumore e dalla casualità insiti nei dati. Per ulteriori informazioni, consulta la guida Quantificazione dell'incertezza nei sistemi di deep learning.

compiti indifferenziati

Conosciuto anche come sollevamento di carichi pesanti, è un lavoro necessario per creare e far funzionare un'applicazione, ma che non apporta valore diretto all'utente finale né offre vantaggi competitivi. Esempi di attività indifferenziate includono l'approvvigionamento, la manutenzione e la pianificazione della capacità.

ambienti superiori

Vedi ambiente.

V

vacuum

Un'operazione di manutenzione del database che prevede la pulizia dopo aggiornamenti incrementali per recuperare lo spazio di archiviazione e migliorare le prestazioni.

controllo delle versioni

Processi e strumenti che tengono traccia delle modifiche, ad esempio le modifiche al codice di origine in un repository.

Peering VPC

Una connessione tra due VPCs che consente di indirizzare il traffico utilizzando indirizzi IP privati. Per ulteriori informazioni, consulta Che cos'è il peering VPC? nella documentazione di Amazon VPC.

vulnerabilità

Un difetto software o hardware che compromette la sicurezza del sistema.

W

cache calda

Una cache del buffer che contiene dati correnti e pertinenti a cui si accede frequentemente. L'istanza di database può leggere dalla cache del buffer, il che richiede meno tempo rispetto alla lettura dalla memoria dal disco principale.

dati caldi

Dati a cui si accede raramente. Quando si eseguono interrogazioni di questo tipo di dati, in genere sono accettabili query moderatamente lente.

funzione finestra

Una funzione SQL che esegue un calcolo su un gruppo di righe che si riferiscono in qualche modo al record corrente. Le funzioni della finestra sono utili per l'elaborazione di attività, come il calcolo di una media mobile o l'accesso al valore delle righe in base alla posizione relativa della riga corrente.

Carico di lavoro

Una raccolta di risorse e codice che fornisce valore aziendale, ad esempio un'applicazione rivolta ai clienti o un processo back-end.

flusso di lavoro

Gruppi funzionali in un progetto di migrazione responsabili di una serie specifica di attività. Ogni flusso di lavoro è indipendente ma supporta gli altri flussi di lavoro del progetto. Ad esempio, il flusso di lavoro del portfolio è responsabile della definizione delle priorità delle applicazioni, della pianificazione delle ondate e della raccolta dei metadati di migrazione. Il flusso di lavoro del portfolio fornisce queste risorse al flusso di lavoro di migrazione, che quindi migra i server e le applicazioni.

VERME

Vedi scrivere una volta, leggere molti.

WQF

Vedi AWS Workload Qualification Framework.

scrivi una volta, leggi molte (WORM)

Un modello di storage che scrive i dati una sola volta e ne impedisce l'eliminazione o la modifica. Gli utenti autorizzati possono leggere i dati tutte le volte che è necessario, ma non possono modificarli. Questa infrastruttura di archiviazione dei dati è considerata immutabile.

Z

exploit zero-day

Un attacco, in genere malware, che sfrutta una vulnerabilità zero-day.

vulnerabilità zero-day

Un difetto o una vulnerabilità assoluta in un sistema di produzione. Gli autori delle minacce possono utilizzare questo tipo di vulnerabilità per attaccare il sistema. Gli sviluppatori vengono spesso a conoscenza della vulnerabilità causata dall'attacco.

prompt zero-shot

Fornire a un LLM le istruzioni per eseguire un'attività ma non esempi (immagini) che possano aiutarla. Il LLM deve utilizzare le sue conoscenze pre-addestrate per gestire l'attività. L'efficacia del prompt zero-shot dipende dalla complessità dell'attività e dalla qualità del prompt. Vedi anche few-shot prompting.

applicazione zombie

Un'applicazione che prevede un utilizzo CPU e memoria inferiore al 5%. In un progetto di migrazione, è normale ritirare queste applicazioni.

Le traduzioni sono generate tramite traduzione automatica. In caso di conflitto tra il contenuto di una traduzione e la versione originale in Inglese, quest'ultima prevarrà.