Whitepaper AWS

Ripristino di emergenza dei carichi di lavoro in AWS: ripristino nel cloud

Copyright © Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Ripristino di emergenza dei carichi di lavoro in AWS: ripristino nel cloud: Whitepaper AWS

Copyright © Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

I marchi e il trade dress di Amazon non possono essere utilizzati in relazione a prodotti o servizi che non siano di Amazon, in qualsiasi modo che possa causare confusione tra i clienti o in qualsiasi modo che denigri o discrediti Amazon. Tutti gli altri marchi non di proprietà di Amazon sono di proprietà dei rispettivi proprietari, che possono o meno essere affiliati, collegati o sponsorizzati da Amazon.

Table of Contents

Ripristino di emergenza dei carichi di lavoro in AWS: ripristino nel cloud

Data di pubblicazione: 12 febbraio 2021 (Revisioni del documento)

Riassunto

Il ripristino di emergenza è il processo di preparazione e ripristino da un'emergenza. Un evento che impedisce a un carico di lavoro o a un sistema di realizzare gli obiettivi aziendali nella posizione di distribuzione primaria è considerato un'emergenza. Questo documento delinea le best practice per la pianificazione e i test del ripristino di emergenza per qualsiasi carico di lavoro distribuito in AWS e offre diversi approcci per attenuare i rischi e soddisfare l'obiettivo del tempo di ripristino (RTO) e l'obiettivo del punto di ripristino (RPO) per il carico di lavoro.

Introduzione

Un carico di lavoro deve svolgere la funzione prevista in modo corretto e coerente. Per realizzare questo obiettivo, deve essere progettarlo per la resilienza. La resilienza è la capacità di un carico di lavoro di ripristinare se stesso in seguito a interruzioni dell'infrastruttura o del servizio, acquisire dinamicamente le risorse di calcolo per soddisfare la domanda e ridurre le interruzioni, dovute ad esempio a configurazioni errate o problemi di rete transitori.

Il ripristino di emergenza è una parte importante della strategia di resilienza e riguarda il modo in cui il carico di lavoro risponde quando si verifica un'emergenza, ovvero un evento che produce un impatto negativo grave sull'azienda. Questa risposta deve basarsi sugli obiettivi aziendali dell'organizzazione, che specificano la strategia del carico di lavoro per evitare la perdita di dati, nota come obiettivo del punto di ripristino (RPO), e quella per ridurre i tempi di inattività quando il carico di lavoro non è disponibile per l'uso, nota come obiettivo del tempo di ripristino (RTO). Dovrai quindi implementare la resilienza nella fase di progettazione dei carichi di lavoro nel cloud per realizzare gli obiettivi di ripristino (RPO e RTO) per un determinato evento di emergenza occasionale. Questo approccio permette alla tua organizzazione di mantenere la continuità aziendale nell'ambito della pianificazione della continuità aziendale.

Questo documento è incentrato su come pianificare, progettare e implementare in AWS architetture in grado di soddisfare gli obiettivi di ripristino di emergenza per la tua azienda. Le informazioni fornite qui sono destinate a chi riveste ruoli tecnologici, ad esempio a Chief Technology Officer (CTO), architetti, sviluppatori e membri dei team operativi.

Ripristino di emergenza e disponibilità

Il ripristino di emergenza può essere raffrontato con la disponibilità, che è un altro componente importante della strategia di resilienza. Mentre il ripristino di emergenza misura gli obiettivi per eventi occasionali, gli obiettivi di disponibilità misurano i valori medi in un periodo di tempo.

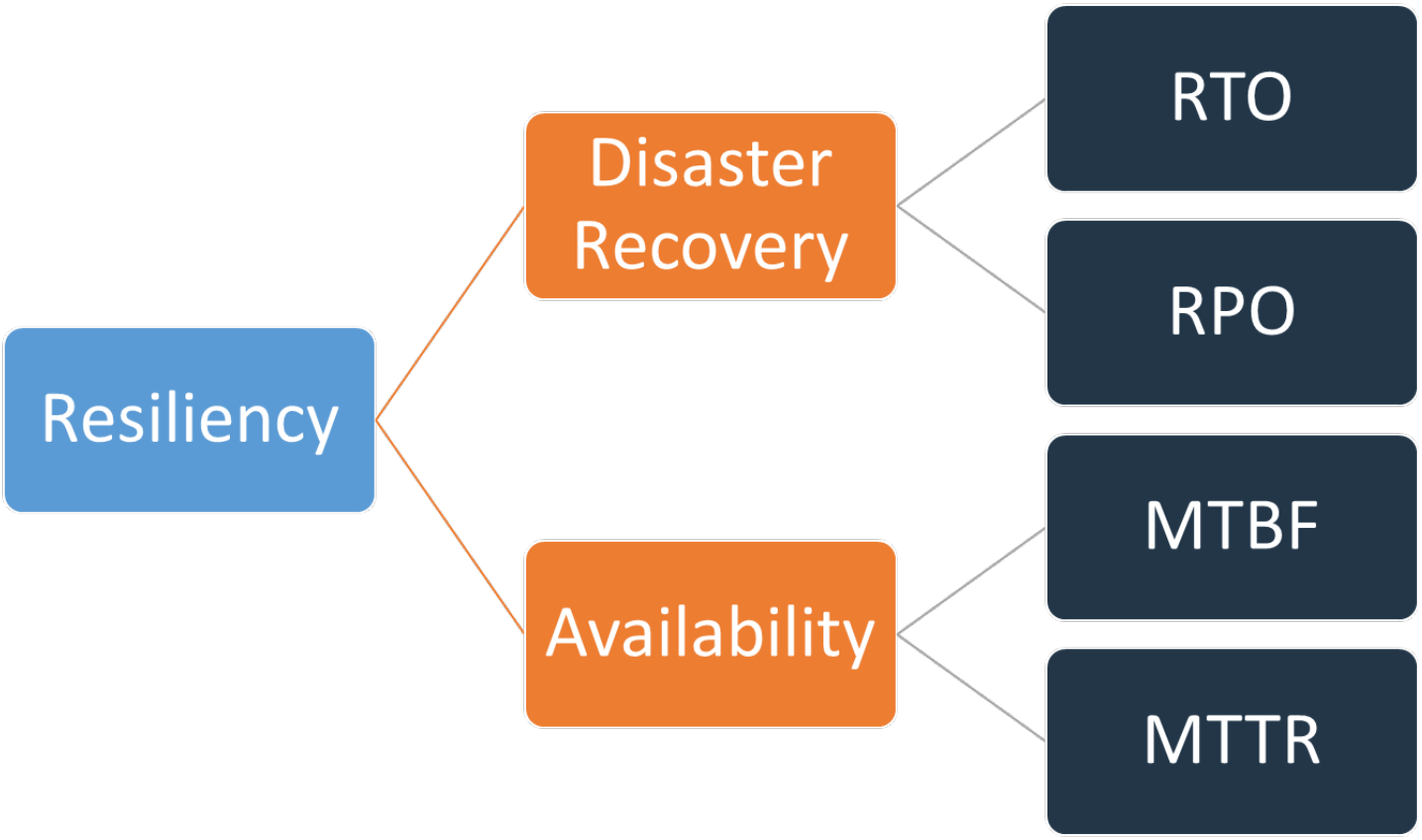

Figura 1 – Obiettivi di resilienza

La disponibilità viene calcolata usando il tempo medio tra gli errori (MTBF) e il tempo medio di ripristino (MTTR):

$$Availability = \frac{Available\ for\ Use\ Time}{Total\ Time} = \frac{MTBF}{MTBF + MTTR}$$

Questo approccio viene spesso definito usando un certo numero di "nove", per cui un obiettivo di disponibilità del 99,9% è noto come "tre nove".

Per il carico di lavoro, può essere più semplice conteggiare le richieste riuscite e non riuscite invece di usare un approccio basato sul tempo. In questo caso, è possibile usare il calcolo seguente:

$$Availability = \frac{Successful\ Responses}{Valid\ Requests}$$

Il ripristino di emergenza è incentrato sugli eventi di emergenza, mentre la disponibilità riguarda interruzioni più comuni su scala più ridotta, ad esempio errori dei componenti, problemi di rete e picchi di carico. L'obiettivo del ripristino di emergenza è la continuità aziendale, mentre la disponibilità riguarda la massimizzazione del tempo per cui un carico di lavoro è disponibile per l'esecuzione della funzionalità aziendale prevista. Entrambi devono fare parte di una strategia di resilienza.

Modello di responsabilità condivisa per la resilienza

La resilienza è una responsabilità condivisa tra AWS e il cliente. È importante comprendere come operano il ripristino di emergenza e la disponibilità, nell'ambito della resilienza, in questo modello condiviso.

Responsabilità di AWS "Resilienza del cloud"

AWS è responsabile della resilienza dell'infrastruttura in cui vengono eseguiti tutti i servizi offerti in AWS Cloud. Questa infrastruttura comprende l'hardware, il software, i componenti di rete e le strutture che eseguono servizi AWS Cloud. AWS intraprende iniziative commercialmente ragionevoli per rendere questi servizi AWS Cloud disponibili, garantendo che la disponibilità dei servizi soddisfi o superi gli accordi sul livello di servizio (SLA) AWS.

L'infrastruttura cloud globale di AWS è progettata per permettere ai clienti di creare architetture di carichi di lavoro altamente resilienti. Ogni regione AWS è completamente isolata ed è costituita da più zone di disponibilità, che sono partizioni fisicamente isolate dell'infrastruttura. Le zone di disponibilità isolano gli errori che possono influire sulla resilienza dei carichi di lavoro, impedendo che abbiano impatto sulle altre zone nella regione. Allo stesso tempo, tuttavia, tutte le zone di disponibilità in una regione AWS sono interconnesse tramite una rete a elevata larghezza di banda e a bassa latenza, su una fibra ottica metropolitana dedicata e completamente ridondante che fornisce connessioni di rete a velocità effettiva elevata e a bassa latenza. Tutto il traffico tra zone è crittografato. Le prestazioni di rete sono sufficienti per l'esecuzione di una replica sincrona tra zone. Le zone di disponibilità semplificano il processo di partizionamento delle applicazioni per offrire disponibilità elevata.

Responsabilità del cliente "Resilienza nel cloud"

La tua responsabilità sarà determinata dai servizi AWS Cloud selezionati. Questo determina la quantità di attività di configurazione che dovrai eseguire nell'ambito delle tue responsabilità in fatto di resilienza. Ad esempio, un servizio come Amazon Elastic Compute Cloud (Amazon EC2) richiede che il cliente esegua tutte le necessarie attività di configurazione e gestione della resilienza. I clienti che distribuiscono istanze Amazon EC2 sono responsabili della distribuzione delle istanze EC2 in più posizioni (come le zone di disponibilità AWS), dell'implementazione della riparazione automatica tramite servizi come AWS Auto Scaling, nonché dell'uso di best practice per l'architettura di carichi di lavoro resilienti per le applicazioni installate nelle istanze. Per i servizi gestiti, come Amazon S3 e Amazon DynamoDB, AWS gestisce il livello dell'infrastruttura, del sistema operativo e delle

piattaforme e i clienti accedono agli endpoint per archiviare e recuperare i dati. La responsabilità della gestione della resilienza dei dati, incluse le strategie di backup, controllo delle versioni e replica, è tua.

La distribuzione del carico di lavoro tra più zone di disponibilità in una regione AWS fa parte di una strategia di disponibilità elevata progettata per proteggere i carichi di lavoro isolando i problemi in una zona di disponibilità e usa la ridondanza delle altre zone di disponibilità per continuare a gestire le richieste. Anche un'architettura Multi-AZ fa parte di una strategia di ripristino di emergenza progettata per rendere i carichi di lavoro meglio isolati e protetti da problemi come interruzioni di corrente, fulmini, tornado, terremoti e altro. Le strategie di ripristino di emergenza possono usare anche più regioni AWS. Ad esempio, in una configurazione attiva/passiva, il servizio per il carico di lavoro eseguirà il failover dalla regione attiva alla regione di ripristino di emergenza se la regione attiva non è più in grado di gestire le richieste.

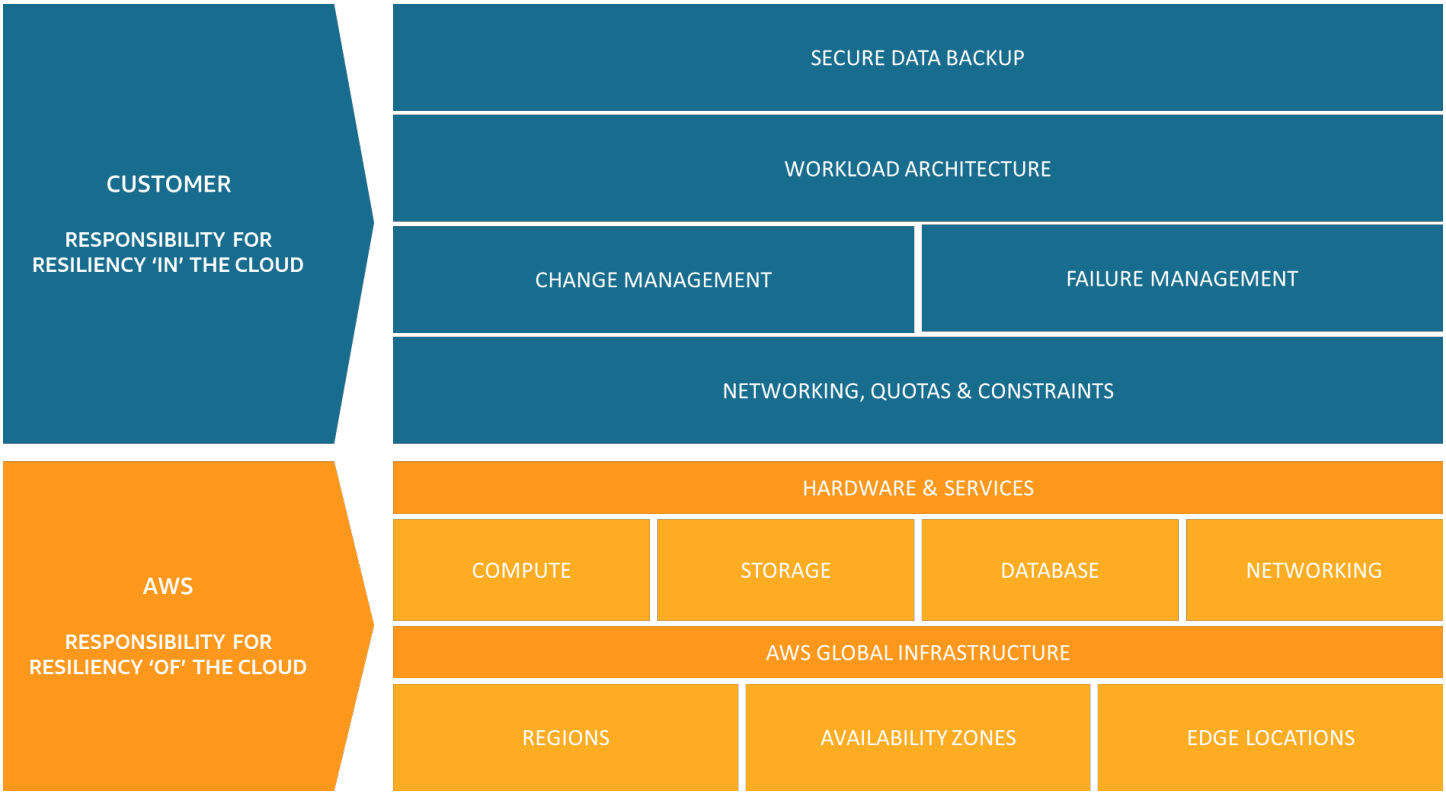

Figura 2 – La resilienza è una responsabilità condivisa tra AWS e il cliente

Che cos'è un'emergenza?

Nel pianificare la strategia di ripristino di emergenza, valuta il piano in base a queste tre categorie di emergenze:

- Calamità naturali, come terremoti o inondazioni
- Guasti tecnici, come mancanza di corrente o connettività di rete insufficiente
- Intervento umano, ad esempio una configurazione involontariamente errata o la modifica o l'accesso non autorizzato o di una parte esterna

Ognuna di queste potenziali emergenze avrà anche un impatto a livello geografico che può essere locale, regionale, nazionale, continentale o globale. La natura dell'emergenza e l'impatto geografico sono due fattori importanti nel considerare la strategia di ripristino di emergenza. Ad esempio, puoi attenuare la gravità di un'inondazione locale che causa l'interruzione dell'alimentazione in un data center adottando una strategia Multi-AZ, perché in questo caso il problema non interesserebbe più di una zona di disponibilità. Tuttavia, un attacco ai dati di produzione richiederebbe una strategia di ripristino di emergenza che includa il failover per eseguire il backup dei dati in un'altra regione AWS.

La disponibilità elevata e il ripristino di emergenza sono due cose diverse

La disponibilità e il ripristino di emergenza si basano su alcune delle stesse best practice, come il monitoraggio degli errori, la distribuzione in più posizioni e il failover automatico. Tuttavia, la disponibilità è incentrata sui componenti del carico di lavoro, mentre il ripristino di emergenza è basato su copie distinte dell'intero carico di lavoro. Il ripristino di emergenza ha obiettivi diversi dalla disponibilità, in quanto misura il tempo di ripristino dopo gli eventi su larga scala che si qualificano come emergenze. Devi innanzitutto assicurarti che il carico di lavoro realizzi gli obiettivi di disponibilità, in quanto un'architettura a disponibilità elevata ti permetterà di soddisfare le esigenze dei clienti in caso di eventi che hanno impatto sulla disponibilità. La strategia di ripristino di emergenza richiede approcci diversi rispetto a quelli per la disponibilità, concentrandosi sulla distribuzione di sistemi distinti in più posizioni per poter eseguire il failover dell'intero carico di lavoro, se necessario.

Devi tenere conto della disponibilità del carico di lavoro nella pianificazione del ripristino di emergenza, in quanto influirà sull'approccio adottato. Un carico di lavoro in esecuzione in una singola istanza Amazon EC2 in una zona di disponibilità non ha disponibilità elevata. Se un'inondazione locale colpisce la zona di disponibilità, questo scenario richiede il failover in un'altra zona di disponibilità per realizzare gli obiettivi di ripristino di emergenza. Confronta questo scenario con un carico di lavoro a disponibilità elevata distribuito come attivo/attivo multisito tra più regioni attive, in cui tutte le regioni gestiscono il traffico di produzione. In questo caso, anche nell'improbabile eventualità che un'emergenza su larga scala neutralizzi un'intera regione, la strategia di ripristino di emergenza viene attuata instradando tutto il traffico verso le regioni rimanenti.

Anche il modo in cui accostarsi ai dati è diverso tra disponibilità e ripristino di emergenza. Supponiamo una soluzione di archiviazione continuamente replicata in un altro sito per ottenere disponibilità elevata, ad esempio un carico di lavoro attivo/attivo multisito. Se uno o più file vengono eliminati o danneggiati nel dispositivo di archiviazione principale, le modifiche dannose possono essere replicate nel dispositivo di archiviazione secondario. In questo scenario, nonostante la disponibilità elevata, la capacità di failover in caso di eliminazione o corruzione dei dati risulterà compromessa. È invece necessario anche un backup point-in-time come parte di una strategia di ripristino di emergenza.

Piano di continuità aziendale

Il piano di ripristino di emergenza deve essere un sottoinsieme del piano di continuità aziendale dell'organizzazione e non un documento autonomo. Non ha senso mantenere obiettivi di ripristino di emergenza aggressivi per il ripristino di un carico di lavoro se gli obiettivi aziendali del carico di lavoro non possono essere realizzati a causa dell'impatto dell'emergenza su elementi dell'azienda diversi dal carico di lavoro stesso. Ad esempio, un terremoto potrebbe impedire il trasporto di prodotti acquistati sull'applicazione di e-commerce: anche se una strategia di ripristino di emergenza efficace mantiene in funzione il carico di lavoro, il piano di continuità aziendale deve soddisfare le esigenze in fatto di trasporti. La strategia di ripristino di emergenza deve essere basata su priorità, contesto e requisiti aziendali.

Analisi dell'impatto aziendale e valutazione del rischio

Un'analisi dell'impatto aziendale deve quantificare l'impatto aziendale di un'interruzione subita dai carichi di lavoro. Deve identificare l'impatto su clienti interni ed esterni dell'impossibilità di usare i carichi di lavoro e l'effetto che ha sulla tua azienda. L'analisi deve aiutare a determinare quanto rapidamente deve essere reso disponibile il carico di lavoro e l'entità della perdita di dati tollerata. Tuttavia, è importante notare che gli obiettivi di ripristino non devono essere definiti in modo isolato. La probabilità di un'interruzione e il costo del ripristino sono fattori chiave per definire il valore aziendale dell'implementazione di una strategia di ripristino di emergenza per un carico di lavoro.

L'impatto aziendale può dipendere dal momento in cui si verifica l'emergenza. Può essere utile prendere in considerazione questo aspetto nella pianificazione del ripristino di emergenza. Ad esempio, è probabile che un'interruzione al sistema di gestione delle retribuzioni abbia un impatto molto elevato sull'azienda se si verifica appena prima di pagare tutti i dipendenti, ma può avere un impatto ridotto se avviene subito dopo il pagamento.

Una valutazione del rischio del tipo di emergenza e dell'impatto geografico insieme a una panoramica dell'implementazione tecnica del carico di lavoro determinerà la probabilità di interruzione causata da ogni tipo di emergenza.

Per carichi di lavoro altamente critici, puoi prendere in considerazione la disponibilità elevata tra più regioni con backup continui per ridurre al minimo l'impatto aziendale. Per i carichi di lavoro meno critici, una strategia valida può essere quella di non definire alcun ripristino di emergenza. E per alcuni scenari di emergenza, è anche ragionevole non prevedere alcuna strategia di ripristino di emergenza come decisione informata basata su una ridotta probabilità del verificarsi dell'emergenza.

Ricorda che le zone di disponibilità all'interno di una regione AWS sono già progettate con una distanza significativa le une dalle altre e un'attenta pianificazione della posizione, in modo che le emergenze più comuni abbiano impatto solo su una zona e non sulle altre. Di conseguenza, un'architettura Multi-AZ all'interno di una regione AWS potrebbe già soddisfare le tue esigenze di attenuazione dei rischi.

È bene valutare il costo delle opzioni di ripristino di emergenza per garantire che la strategia di ripristino di emergenza fornisca il livello corretto di valore aziendale considerando l'impatto e il rischio per l'azienda.

Con tutte queste informazioni, potrai documentare la minaccia, il rischio, l'impatto e il costo di diversi scenari di emergenza e le opzioni di ripristino associate. Queste informazioni devono essere usate per determinare gli obiettivi di ripristino per ognuno dei carichi di lavoro.

Obiettivi di ripristino (RTO e RPO)

Nel creare una strategia di ripristino di emergenza, le organizzazioni basano in genere la propria pianificazione sull'obiettivo del tempo di ripristino (RTO) e sull'obiettivo del punto di ripristino (RPO).

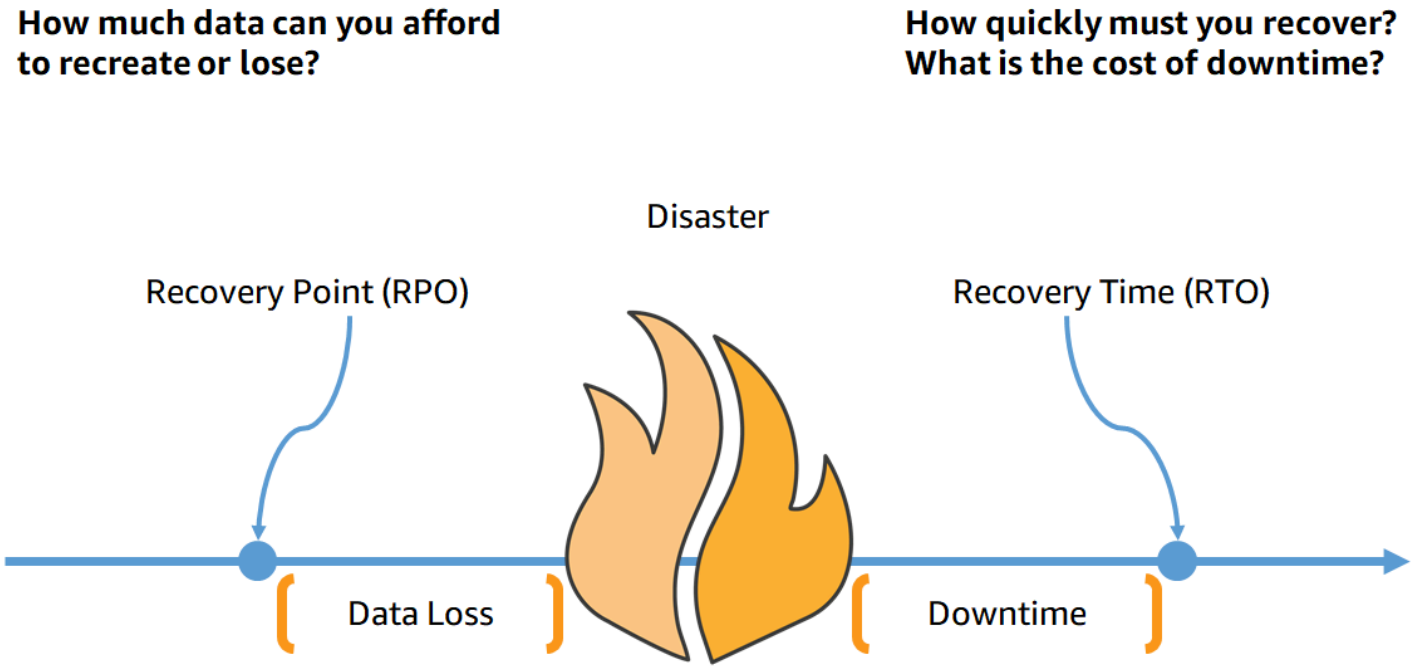

Figura 3 – Obiettivi di ripristino

L'obiettivo del tempo di ripristino (RTO) è il ritardo massimo accettabile tra l'interruzione del servizio e il suo ripristino. Questo obiettivo determina una finestra temporale accettabile per la non disponibilità del servizio, definita dall'organizzazione.

In questo documento vengono presentate quattro strategie di ripristino di emergenza: backup e ripristino, Pilot Light, standby a freddo e attivo/attivo multisito (consulta Opzioni di ripristino di emergenza nel cloud). Nel diagramma seguente l'azienda ha determinato il proprio RTO massimo accettabile e il limite di spesa per la propria strategia di ripristino del servizio. In base agli obiettivi aziendali, le strategie di ripristino di emergenza Pilot Light e standby a freddo soddisfano entrambe i criteri di RTO e costo.

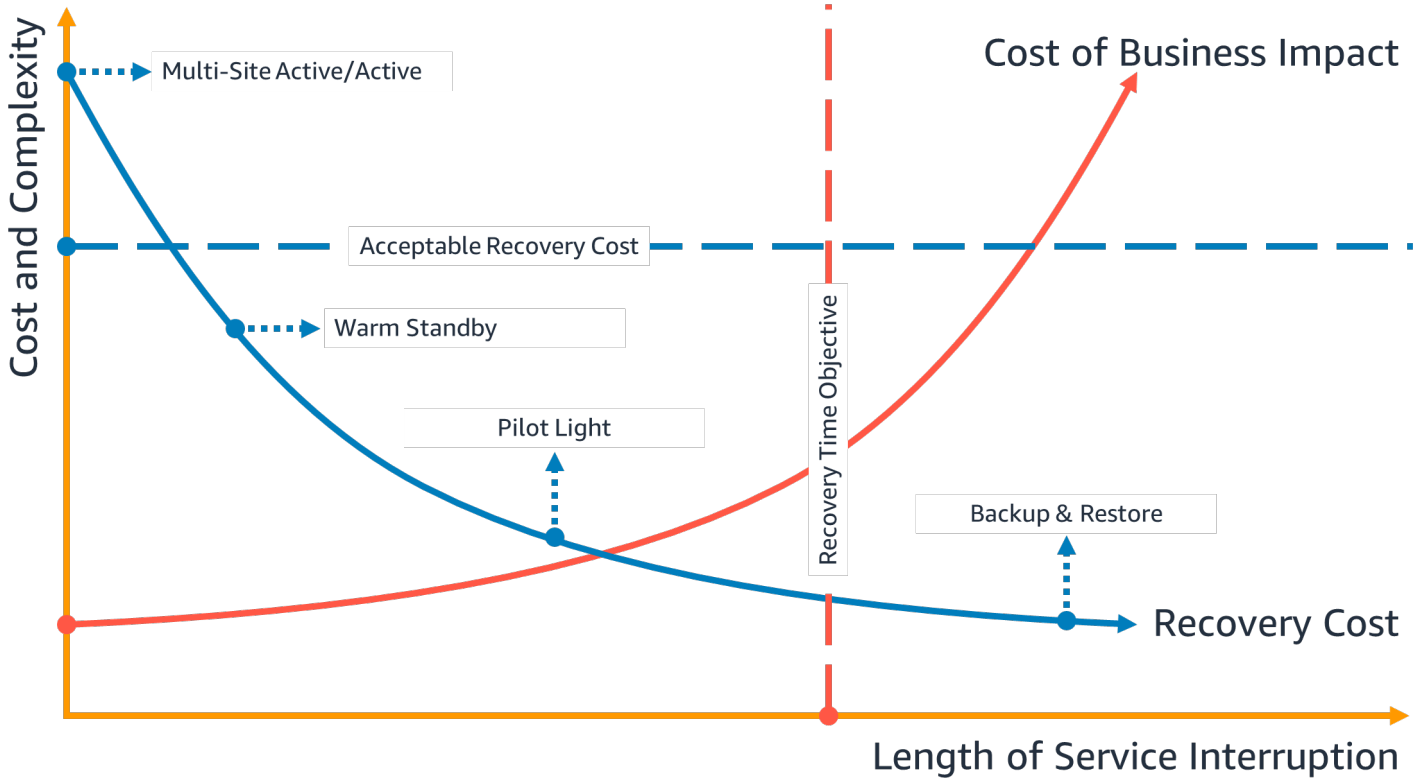

Figura 4 – Obiettivo del tempo di ripristino

L'obiettivo del punto di ripristino (RPO) è il periodo di tempo massimo accettabile dall'ultimo punto di ripristino dei dati. Questo obiettivo determina una perdita di dati accettabile tra l'ultimo punto di ripristino e l'interruzione del servizio, definita dall'organizzazione.

Nel diagramma seguente l'azienda ha determinato il proprio RPO massimo consentito e il limite di spesa per la propria strategia di ripristino dei dati. Delle quattro strategie di ripristino di emergenza, le strategie Pilot Light e standby a freddo soddisfano entrambe i criteri di RPO e costo.

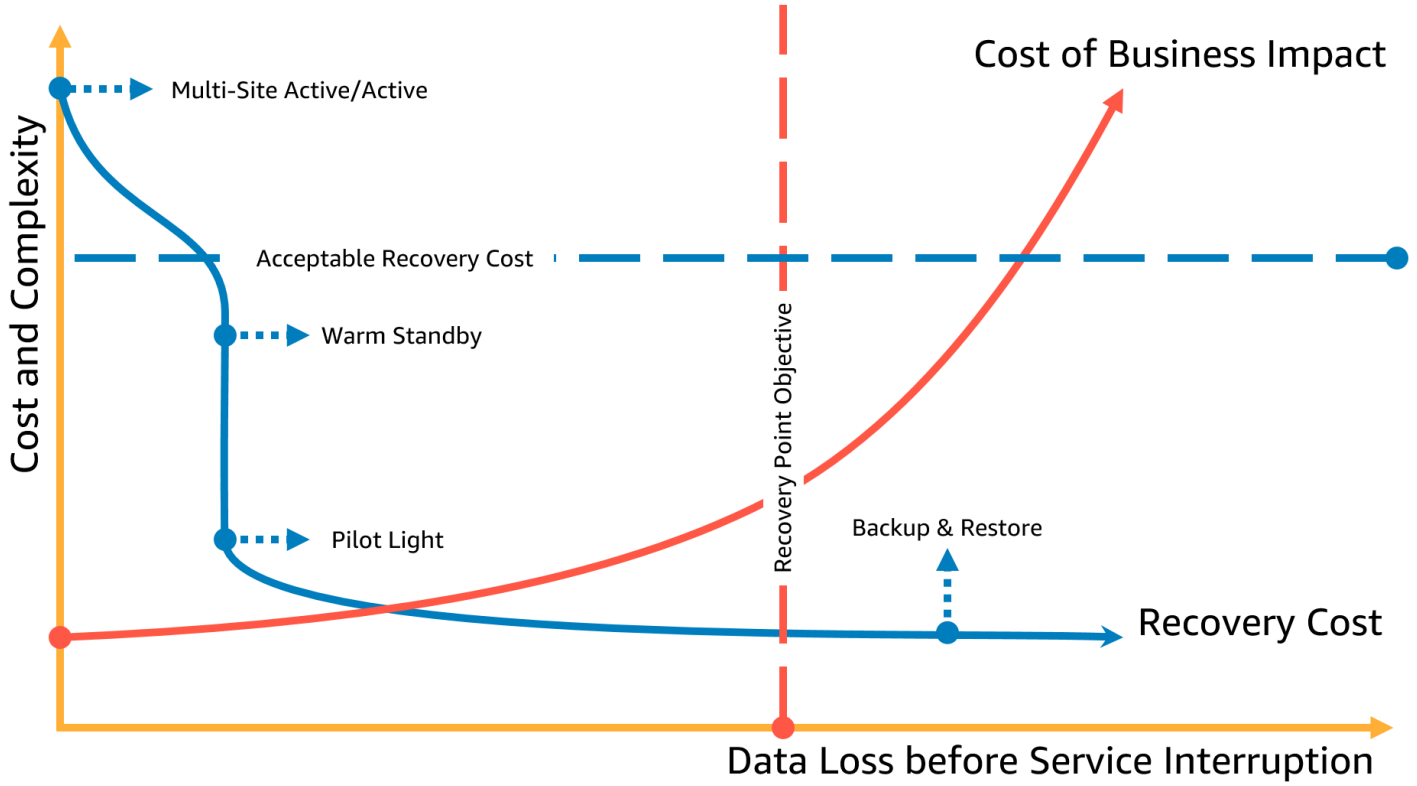

Figura 5 – Obiettivo del punto di ripristino

> **Note**
>
> Se il costo del ripristino è superiore al costo dell'errore o della perdita, l'opzione di ripristino non deve essere implementata a meno che non sia necessario tenere conto di un fattore secondario, ad esempio i requisiti normativi.

Il ripristino di emergenza è diverso nel cloud

Le strategie di ripristino di emergenza evolvono con l'innovazione tecnica. Un piano di ripristino di emergenza On-Premise può includere il trasporto fisico di nastri o la replica dei dati in un altro sito. La tua organizzazione deve rivalutare l'impatto aziendale, il rischio e il costo delle precedenti strategie di ripristino di emergenza per poter realizzare i propri obiettivi di ripristino di emergenza in AWS. Il ripristino di emergenza in AWS Cloud include i vantaggi seguenti rispetto agli ambienti tradizionali:

- Rapido ripristino da un'emergenza con complessità ridotta
- Test semplici e ripetibili che possono essere eseguiti più facilmente e più spesso
- Sovraccarico di gestione minore per ridurre il peso operativo
- Opportunità per automatizzare, ridurre le possibilità di errore e migliorare i tempi di ripristino

AWS ti permette di sostituire le spese in conto capitale fisse di un data center di backup fisico con i costi operativi variabili di un ambiente di dimensioni adeguate nel cloud, per ottenere risparmi significativi.

Per molte organizzazioni, il ripristino di emergenza On-Premise era basato sul rischio di interruzione per uno o più carichi di lavoro in un data center e sul ripristino dei dati sottoposti a backup o replicati in un data center secondario. Quando le organizzazioni distribuiscono carichi di lavoro in AWS, possono implementare un carico di lavoro Well-Architected e affidarsi alla progettazione dell'infrastruttura cloud globale di AWS per ridurre l'impatto di tali interruzioni. Consulta il[whitepaper Il principio dell'affidabilità – AWS Well-Architected Framework] per ulteriori informazioni sulle best practice sull'architettura per la progettazione e l'esecuzione di carichi di lavoro affidabili, sicuri, efficienti e convenienti nel cloud.

Se i tuoi carichi di lavoro sono in AWS, non devi preoccuparti di connettività del data center (ad eccezione della tua capacità di accedervi), alimentazione, aria condizionata, estinzione degli incendi e hardware. Tutti questi aspetti vengono gestiti al posto tuo e hai accesso a più zone di disponibilità con isolamento dagli errori, ognuna costituita da uno o più data center distinti.

Singola regione AWS

Per un evento di emergenza basato sull'interruzione o sulla perdita di un data center fisico, l'implementazione di un carico di lavoro a disponibilità elevata in più zone di disponibilità all'interno di una singola regione AWS contribuisce ad attenuare l'impatto di calamità naturali ed emergenze

tecniche e riduce il rischio di minacce legate all'intervento umano, come un errore o un'attività non autorizzata che può causare la perdita di dati. Ogni regione AWS è costituita da più zone di disponibilità, ognuna con isolamento dagli errori nella altre zone. Ogni zona di disponibilità è a sua volta costituita da più data center fisici. Per isolare meglio i problemi più seri e ottenere disponibilità elevata, puoi dividere i carichi di lavoro tra più zone nella stessa regione. Le zone di disponibilità sono progettate per la ridondanza fisica e forniscono resilienza in modo da assicurare prestazioni costanti, anche in caso di interruzioni dell'alimentazione, tempi di inattività di Internet, inondazioni e altre calamità naturali. Consulta Infrastruttura cloud globale di AWS per informazioni su questi aspetti.

Attraverso la distribuzione tra più zone di disponibilità in una singola regione AWS, il carico di lavoro è protetto meglio dagli errori di un singolo data center (o anche più di uno). Per una maggiore sicurezza della distribuzione in un'unica regione, puoi eseguire il backup dei dati e della configurazione (inclusa la definizione dell'infrastruttura) in un'altra regione. Questa strategia riduce l'ambito del piano di ripristino di emergenza includendo solo il backup e il ripristino dei dati. La resilienza multi-regione attraverso il backup in un'altra regione AWS è una scelta semplice e conveniente rispetto alle altre opzioni multi-regione descritte nella sezione seguente. Ad esempio, il backup in Amazon Simple Storage Service (Amazon S3) dà accesso al recupero immediato dei dati. Tuttavia, se la strategia di ripristino di emergenza per parti dei dati ha requisiti meno rigorosi in fatto di tempi di ripristino (da pochi minuti a più ore), l'uso di Amazon S3 Glacier o Amazon S3 Glacier Deep Archive ridurrà notevolmente i costi della strategia di backup e ripristino.

Alcuni carichi di lavoro possono essere soggetti a requisiti normativi sulla residenza dei dati. Se questo vale per il tuo carico di lavoro in una località in cui è attualmente disponibile una sola regione AWS, oltre a progettare carichi di lavoro Multi-AZ per la disponibilità elevata come spiegato in precedenza, puoi anche usare le zone di disponibilità all'interno della regione come posizioni distinte. Questa scelta può essere utile per soddisfare i requisiti di residenza dei dati previsti per il carico di lavoro all'interno della regione. Le strategie di ripristino di emergenza descritte nelle sezioni seguenti usano più regioni AWS, ma possono anche essere implementate tramite zone di disponibilità anziché regioni.

Più regioni AWS

Per un evento di emergenza che include il rischio di perdere più data center posti a una distanza significativa l'uno dall'altro, è bene valutare opzioni di ripristino di emergenza in grado di attenuare l'impatto di calamità naturali ed emergenze tecniche che possono colpire un'intera regione in AWS. Tutte le opzioni descritte nelle sezioni seguenti possono essere implementate come architetture multi-regione per la protezione da tali emergenze.

Opzioni di ripristino di emergenza nel cloud

Le strategie di ripristino di emergenza disponibili in AWS possono essere generalmente suddivise in quattro approcci, che vanno dall'esecuzione di backup a basso costo e con complessità ridotta a strategie più complesse che usano più regioni attive. È essenziale testare regolarmente la strategia di ripristino di emergenza in modo da poterla richiamare con fiducia, se necessario.

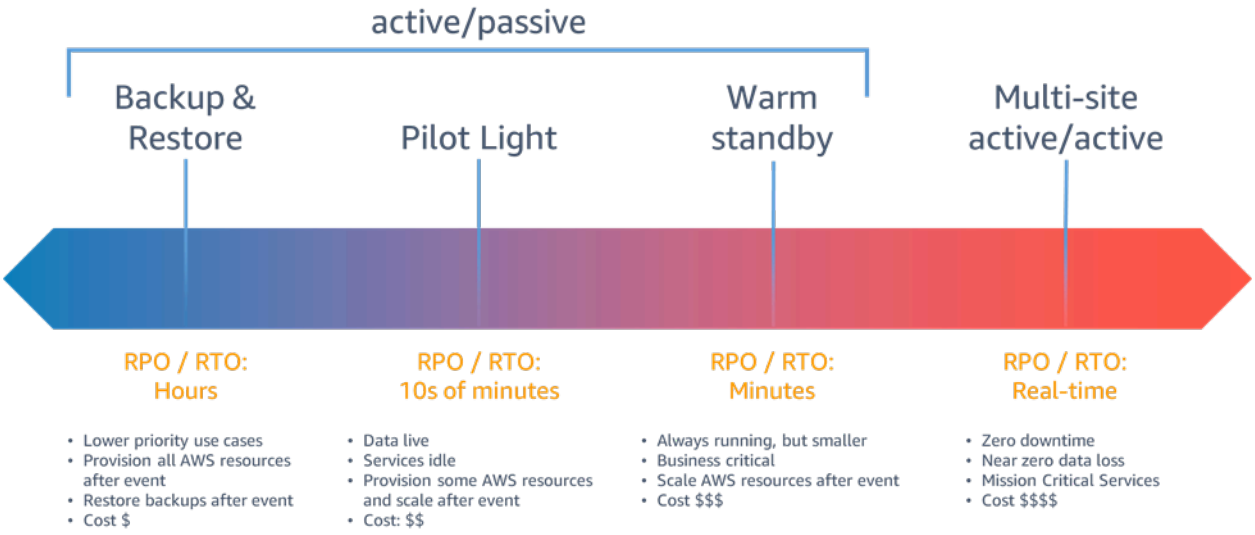

Figura 6 – Strategie di ripristino di emergenza

Per un evento di emergenza basato sull'interruzione o sulla perdita di un data center fisico per un carico di lavoro Well-Architected a disponibilità elevata, può essere necessario un approccio al ripristino di emergenza che preveda solo il backup e il ripristino. Se la tua definizione di emergenza si estende oltre l'interruzione o la perdita di un data center fisico fino a un'intera regione o se devi soddisfare requisiti normativi che lo richiedono, dovrai prendere in considerazione l'approccio Pilot Light, standby a freddo o attivo/attivo multisito.

Backup e ripristino

L'approccio di backup e ripristino è adatto per ridurre l'eventualità di perdita o corruzione dei dati. Questo approccio può essere usato anche per ridurre l'entità di un'emergenza regionale tramite la replica dei dati in altre regioni AWS o per attenuare la mancanza di ridondanza per i carichi di lavoro distribuiti in un'unica zona di disponibilità. Oltre ai dati, devi ridistribuire l'infrastruttura, la configurazione e il codice dell'applicazione nella regione di ripristino. Per permettere una ridistribuzione dell'infrastruttura rapida e senza errori, devi usare sempre il modello Infrastructure

as Code (IaC) tramite servizi comeAWS CloudFormation o il AWS Cloud Development Kit (AWS CDK). Senza IaC, il ripristino dei carichi di lavoro nella regione di ripristino può risultare complesso, causando tempi di ripristino più estesi e il probabile superamento dell'RTO. Oltre ai dati utente, assicurati di eseguire anche il backup del codice e della configurazione, incluse le Amazon Machine Image (AMI) usate per creare istanze Amazon EC2. Puoi usare AWS CodePipeline per automatizzare la ridistribuzione del codice e della configurazione dell'applicazione.

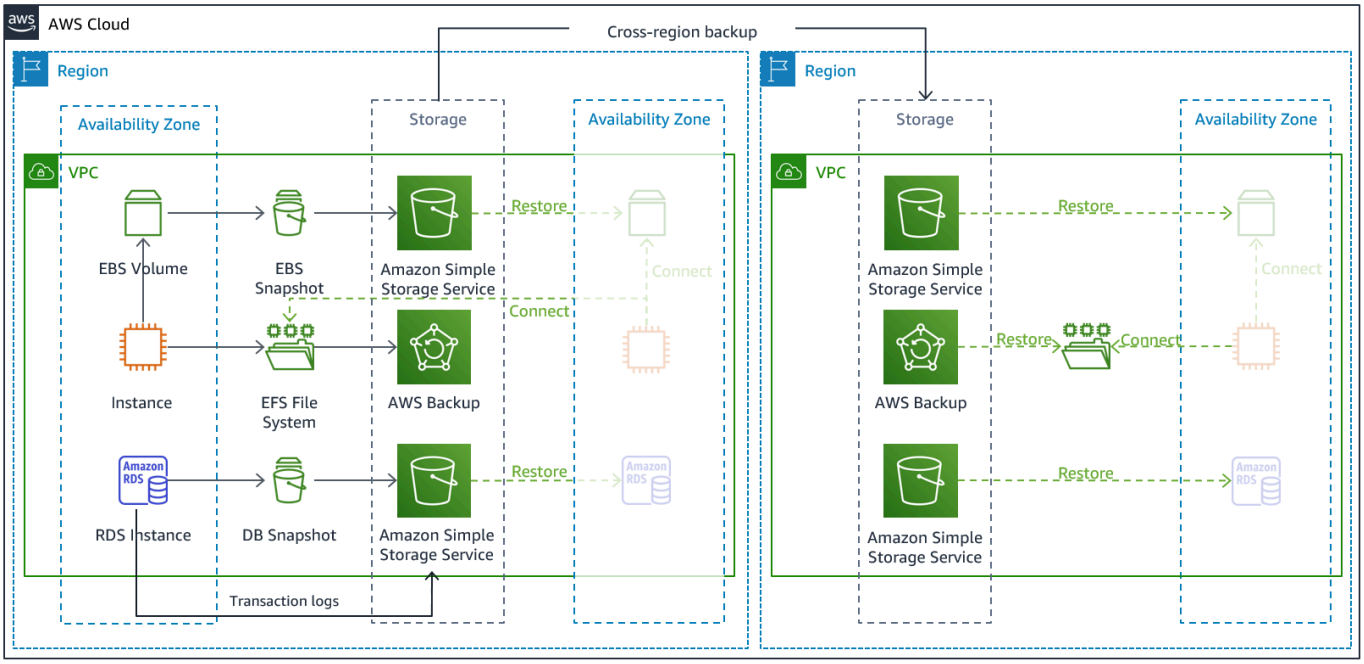

Figura 7 – Architettura di backup e ripristino

Servizi AWS

I dati dei carichi di lavoro richiedono una strategia di backup da eseguire periodicamente o continuamente. La frequenza con cui esegui il backup determinerà il punto di ripristino raggiungibile (che deve essere allineato all'RPO). Il backup deve anche offrire un modo per eseguire il ripristino in base al momento specifico in cui è stato eseguito. Il backup con ripristino point-in-time (PITR) è disponibile tramite i servizi e le risorse seguenti:

- Snapshot Amazon Elastic Block Store (Amazon EBS)
- Backup Amazon DynamoDB
- Snapshot Amazon RDS
- Snapshot DB Amazon Aurora

- Backup Amazon EFS (quando si usa AWS Backup)
- Snapshot Amazon Redshift
- Snapshot Amazon Neptune

Per Amazon Simple Storage Service (Amazon S3), puoi usare la replica tra regioni in Amazon S3 per copiare in modo asincrono e continuamente gli oggetti in un bucket S3 nella regione di ripristino di emergenza, fornendo il controllo delle versioni per gli oggetti archiviati in modo da poter scegliere il punto di ripristino. La replica continua dei dati ha il vantaggio di offrire il tempo più breve (vicino a zero) per il backup dei dati, ma può non proteggere da eventi di emergenza come la corruzione dei dati o attacchi dannosi, ad esempio un'eliminazione non autorizzata dei dati, come d'altronde i backup point-in-time. La replica continua viene trattata nella sezione Servizi AWS per Pilot Light.

AWS Backup fornisce una posizione centralizzata per configurare, pianificare e monitorare le funzionalità di backup di AWS per i servizi e le risorse seguenti:

- Volumi Amazon Elastic Block Store (Amazon EBS)
- Istanze Amazon EC2
- Database Amazon Relational Database Service (Amazon RDS), inclusi database Amazon Aurora
- Tabelle Amazon DynamoDB
- File system Amazon Elastic File System (Amazon EFS)
- Volumi AWS Storage Gateway
- Amazon FSx per Windows File Server e Amazon FSx for Lustre

AWS Backup supporta la copia di backup tra regioni, ad esempio in una regione di ripristino di emergenza.

Come strategia di ripristino di emergenza aggiuntiva per i dati in Amazon S3, abilita il controllo delle versioni degli oggetti S3. Il controllo delle versioni degli oggetti protegge i dati in S3 dalle conseguenze delle operazioni di eliminazione o modifica mantenendo la versione originale prima dell'operazione. Il controllo delle versioni degli oggetti può essere un'utile misura di attenuazione per emergenze dovute a errore umano. Se usi la replica di S3 per eseguire il backup dei dati nella regione di ripristino di emergenza, quando un oggetto viene eliminato nel bucket di origine, per default Amazon S3 aggiunge un contrassegno di eliminazione solo nel bucket di origine. Questo approccio protegge i dati nella regione di ripristino di emergenza da eliminazioni dannose nella regione di origine.

Oltre ai dati, devi eseguire anche il backup della configurazione e dell'infrastruttura necessarie per ridistribuire il carico di lavoro e soddisfare l'obiettivo del tempo di ripristino (RTO). AWS CloudFormation fornisce il modello Infrastructure as Code (IaC) e ti permette di definire tutte le risorse AWS nel carico di lavoro in modo da poter eseguire la distribuzione e la ridistribuzione in modo affidabile in più account AWS e regioni AWS. Puoi eseguire il backup delle istanze Amazon EC2 usate dal carico di lavoro come Amazon Machine Image (AMI). L'AMI viene creata da snapshot del volume radice dell'istanza e da qualsiasi altro volume EBS collegato all'istanza. Puoi usare questa AMI per avviare una versione ripristinata dell'istanza EC2. Un'AMI può essere copiata all'interno di una regione o tra regioni. In alternativa, puoi usare AWS Backup per copiare backup tra account e in altre regioni AWS. La funzionalità di backup tra account favorisce la protezione da eventi di emergenza che includono minacce interne o compromissione degli account. AWS Backup aggiunge anche altre funzionalità per il backup di istanze EC2: oltre ai singoli volumi EBS dell'istanza, AWS Backup archivia e monitora anche questi metadati: tipo di istanza, cloud privato virtuale (VPC) configurato, gruppo di sicurezza, ruolo IAM, configurazione di monitoraggio e tag. Tuttavia, questi metadati aggiuntivi vengono usati solo per il ripristino del backup di istanze EC2 nella stessa regione AWS.

Tutti i dati archiviati nella regione di ripristino di emergenza come backup devono essere ripristinati al momento del failover. AWS Backup offre la funzionalità di ripristino, ma attualmente non permette il ripristino pianificato o automatico. Puoi implementare il ripristino automatico nella regione di ripristino di emergenza usando l'SDK AWS per chiamare API per AWS Backup. Puoi configurarlo come normale processo ricorrente o attivare il ripristino ogni volta che viene completato un backup. La figura seguente mostra un esempio di ripristino automatico tramite Amazon Simple Notification Service (Amazon SNS) e AWS Lambda.

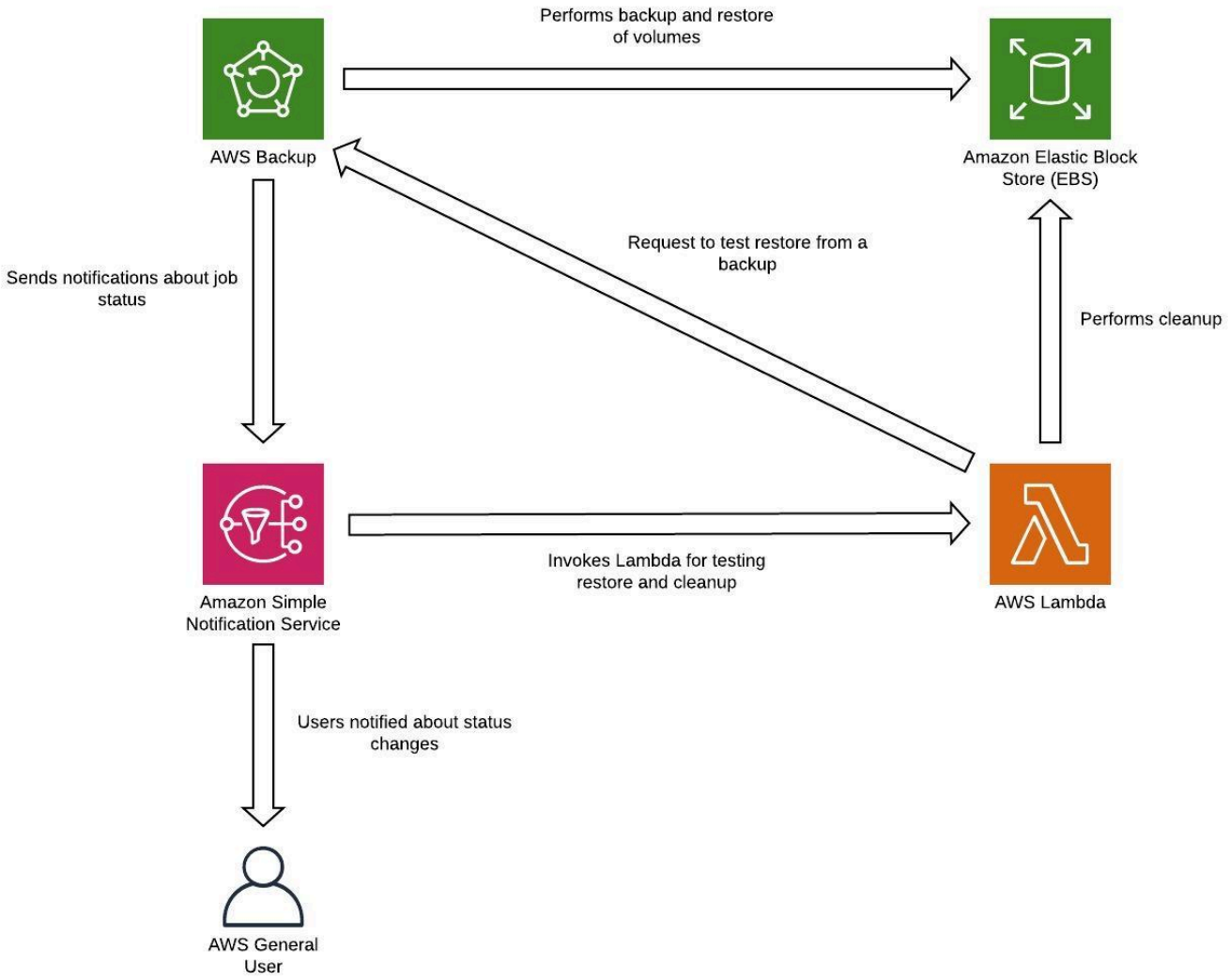

Figura 8 – Ripristino e test dei backup

> **Note**
>
> La strategia di backup deve includere il test dei backup. Per ulteriori informazioni, consulta la sezione Test del ripristino di emergenza. Per una dimostrazione pratica dell'implementazione, fai riferimento a AWS Well-Architected Lab: Testing Backup and Restore of Data.

Pilot Light

Con l'approccio Pilot Light puoi replicare i dati da una regione a un'altra ed effettuare il provisioning di una copia dell'infrastruttura principale del carico di lavoro. Le risorse necessarie per supportare

la replica e il backup dei dati, come database e archiviazione di oggetti, sono sempre attive. Altri elementi, come i server applicazioni, vengono caricati con il codice e le configurazioni dell'applicazione, ma vengono disattivati e usati solo durante i test o quando viene richiamato il failover del ripristino di emergenza. A differenza dell'approccio di backup e ripristino, l'infrastruttura principale è sempre disponibile e hai sempre la possibilità di effettuare rapidamente il provisioning di un ambiente di produzione su scala completa attivando e aumentando orizzontalmente i server applicazioni.

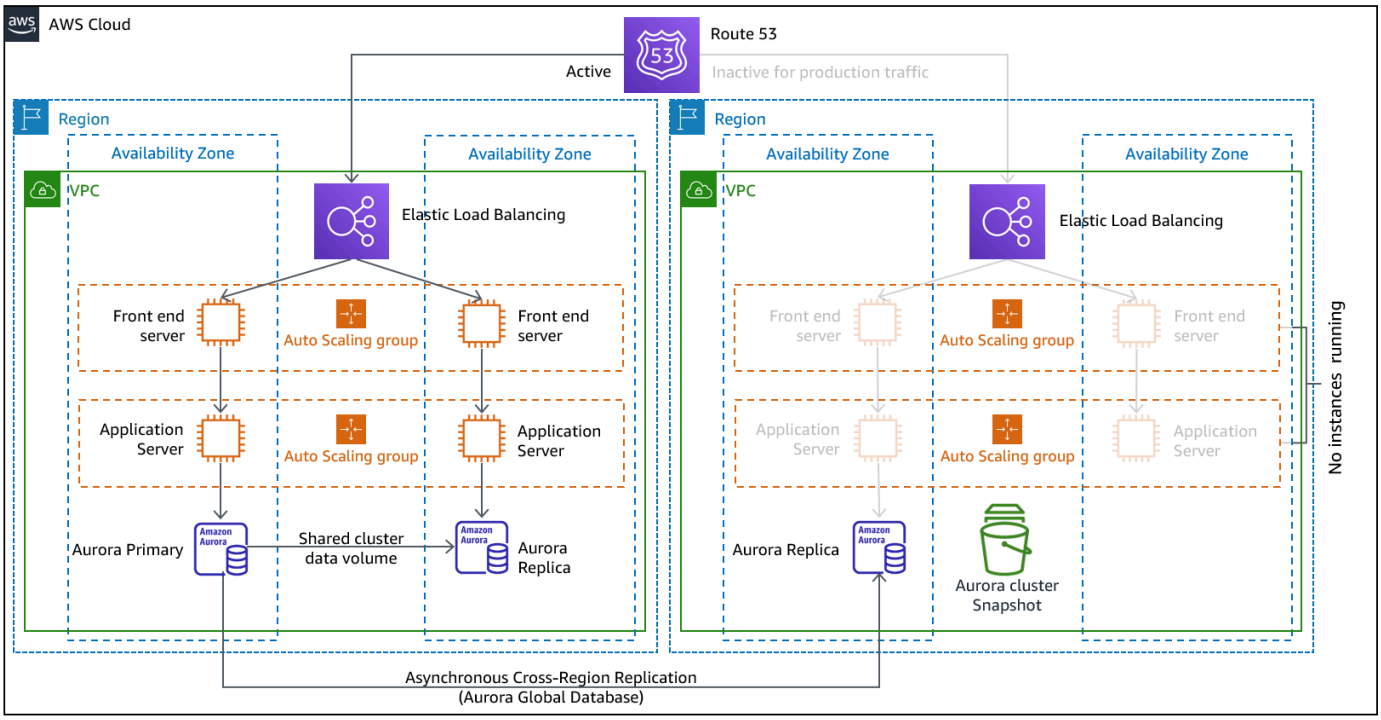

Figura 9 – Architettura per Pilot Light

Un approccio Pilot Light taglia i costi continui del ripristino di emergenza riducendo al minimo le risorse attive e semplifica il ripristino al momento di un'emergenza grazie alla possibilità di soddisfare tutti i requisiti dell'infrastruttura principale. Questa opzione di ripristino richiede la modifica dell'approccio di distribuzione. Devi apportare modifiche all'infrastruttura principale in ogni regione e distribuire le modifiche al carico di lavoro (configurazione e codice) contemporaneamente in ciascuna. Questo passaggio può essere semplificato automatizzando le distribuzioni e usando Infrastructure as Code (IaC) per distribuire l'infrastruttura tra più account e regioni (distribuzione dell'infrastruttura completa nella regione primaria e distribuzione dell'infrastruttura ridotta/disattivata nelle regioni di ripristino di emergenza). Ti consigliamo di usare un account diverso per ogni regione per fornire il massimo livello di isolamento delle risorse e della sicurezza (qualora la compromissione delle credenziali sia inclusa nei piani di ripristino di emergenza).

Con questo approccio, devi anche attenuare l'impatto di un'emergenza relativa ai dati. La replica continua dei dati ti protegge da alcuni tipi di emergenza, ma può non essere sufficiente in caso di corruzione ed eliminazione dei dati, a meno che la strategia non includa anche il controllo delle versioni dei dati archiviati o opzioni per il ripristino point-in-time (PITR). Puoi eseguire il backup dei dati replicati nella regione in cui si verifica l'emergenza per creare backup point-in-time nella stessa regione.

Servizi AWS

Oltre a usare i servizi AWS presentati nella sezione Backup e ripristino per creare backup point-in-time, tieni conto anche dei servizi seguenti per la tua strategia Pilot Light.

Per l'approccio Pilot Light, la replica continua dei dati in database e archivi dati attivi nella regione di ripristino di emergenza è l'approccio migliore per un RPO ridotto (se usata insieme ai backup point-in-time di cui abbiamo parlato in precedenza). AWS fornisce la replica continua e asincrona dei dati tra regioni tramite i servizi e le risorse seguenti:

- Replica di Amazon Simple Storage Service (Amazon S3)
- Repliche di lettura di Amazon RDS
- Amazon Aurora Global Database
- Tabelle globali Amazon DynamoDB

Con la replica continua, le versioni dei dati sono disponibili quasi immediatamente nella regione di ripristino di emergenza. Gli effettivi tempi di replica possono essere monitorati tramite caratteristiche dei servizi come S3 Replication Time Control (S3 RTC) per oggetti S3 e le caratteristiche di gestione di Amazon Aurora Global Database.

Durante il failover per eseguire il carico di lavoro di lettura/scrittura dalla regione di ripristino di emergenza, devi promuovere una replica di lettura RDS in modo da renderla l'istanza primaria. Per istanze database diverse da Aurora, il completamento del processo richiede alcuni minuti e il riavvio è parte del processo. Per la replica tra regioni e il failover con RDS, l'uso di Amazon Aurora Global Database offre diversi vantaggi. Il database globale usa un'infrastruttura dedicata che lascia i database completamente disponibili per la gestione dell'applicazione e può replicarsi nella regione secondaria con una latenza tipica inferiore a un secondo (e all'interno di una regione AWS molto inferiore a 100 millisecondi). Con Amazon Aurora Global Database, se la regione primaria subisce un calo delle prestazioni o un'interruzione, puoi promuovere una delle regioni secondarie in modo che si

occupi delle letture/scritture in meno di 1 minuto anche in caso di interruzione completa nella regione. La promozione può essere automatica e non è previsto alcun riavvio.

Una versione ridotta dell'infrastruttura principale del carico di lavoro con un numero inferiore di risorse deve essere distribuita nella regione di ripristino di emergenza. Usando AWS CloudFormation, puoi definire l'infrastruttura e distribuirla in modo coerente tra più account AWS e regioni AWS. AWS CloudFormation usa pseudoparametri predefiniti per identificare l'account AWS e la regione AWS in cui viene distribuito. Di conseguenza, puoi implementare logica condizionale nei modelli CloudFormation per distribuire solo la versione ridotta dell'infrastruttura nella regione di ripristino di emergenza. Per le distribuzioni di istanze EC2, un'Amazon Machine Image (AMI) fornisce informazioni come la configurazione hardware e il software installato. Puoi implementare una pipeline Image Builder che crea le AMI necessarie e le copia nella regione primaria e in quella di backup. In questo modo, puoi garantire che le versioni definitive delle AMI abbiano tutto il necessario per ridistribuire o aumentare orizzontalmente il carico di lavoro in una regione, nel caso di un evento di emergenza. Le istanze Amazon EC2 vengono distribuite in una configurazione ridotta (meno istanze rispetto alla regione primaria). Puoi usare l'ibernazione per porre le istanze EC2 in uno stato di arresto che evita il pagamento dei costi di EC2, ma solo per l'archiviazione usata. Per avviare istanze EC2, puoi creare script tramite AWS Command Line Interface (CLI) o l'SDK AWS. Per aumentare orizzontalmente l'infrastruttura in modo da supportare il traffico di produzione, consulta AWS Auto Scaling nella sezione Standby a freddo.

Per una configurazione attiva/in standby come nell'approccio Pilot Light, tutto il traffico va inizialmente alla regione primaria e quindi passa alla regione di ripristino di emergenza se la regione primaria non è più disponibile. Sono disponibili due opzioni di gestione del traffico da prendere in considerazione usando servizi AWS. La prima opzione consiste nell'usare Amazon Route 53. Con Amazon Route 53 puoi associare più endpoint IP in una o più regioni AWS a un nome di dominio Route 53. Puoi quindi instradare il traffico verso l'endpoint appropriato all'interno del nome di dominio. I controlli dell'integrità in Amazon Route 53 monitorano questi endpoint. Usando questi controlli dell'integrità, puoi configurare il failover DNS per garantire che il traffico venga inviato a endpoint integri.

La seconda opzione consiste nell'usare AWS Global Accelerator. Usando un indirizzo IP AnyCast, puoi associare più endpoint in una o più regioni AWS agli stessi indirizzi IP statici. AWS Global Accelerator instrada quindi il traffico verso l'endpoint appropriato associato all'indirizzo. I controlli dell'integrità in Global Accelerator monitorano gli endpoint. Usando questi controlli dell'integrità, AWS Global Accelerator controlla automaticamente lo stato delle applicazioni e instrada il traffico degli utenti solo verso un endpoint dell'applicazione integro. Global Accelerator offre latenze inferiori per il traffico all'endpoint dell'applicazione, in quanto usa l'estesa rete edge di AWS per spostare il traffico

sulla dorsale di rete (backbone) AWS il prima possibile. Global Accelerator evita anche i problemi di memorizzazione nella cache che possono verificarsi con i sistemi DNS (come Route 53).

CloudEndure Disaster Recovery

CloudEndure Disaster Recovery, disponibile da Marketplace AWS, replica continuamente le applicazioni e i database ospitati dal server da qualsiasi origine in AWS usando la replica a livello di blocco del server sottostante. CloudEndure Disaster Recovery ti permette di usare AWS Cloud come regione di ripristino di emergenza per un carico di lavoro On-Premise e il suo ambiente. Può essere usato anche per il ripristino di emergenza dei carichi di lavoro ospitati da AWS se sono costituiti solo da applicazioni e database ospitati in EC2 (ovvero non in RDS). CloudEndure Disaster Recovery usa la strategia Pilot Light, mantenendo una copia dei dati e delle risorse disattivate in un ambiente Amazon Virtual Private Cloud (Amazon VPC) usato come area di gestione temporanea. Quando viene attivato un evento di failover, le risorse con gestione temporanea vengono usate per creare automaticamente una distribuzione a capacità completa nell'ambiente Amazon VPC di destinazione usato come posizione di ripristino.

Figura 10 – Architettura di CloudEndure Disaster Recovery

Standby a freddo

L'approccio standby a freddo garantisce che vi sia una copia ridotta ma completamente funzionale dell'ambiente di produzione in un'altra regione. Questo approccio estende il concetto di Pilot Light e riduce il tempo di ripristino, in quanto il carico di lavoro è sempre attivo in un'altra regione. Questo approccio permette anche di eseguire test più facilmente o di implementare test continui per aumentare la fiducia nella possibilità del ripristino da un'emergenza.

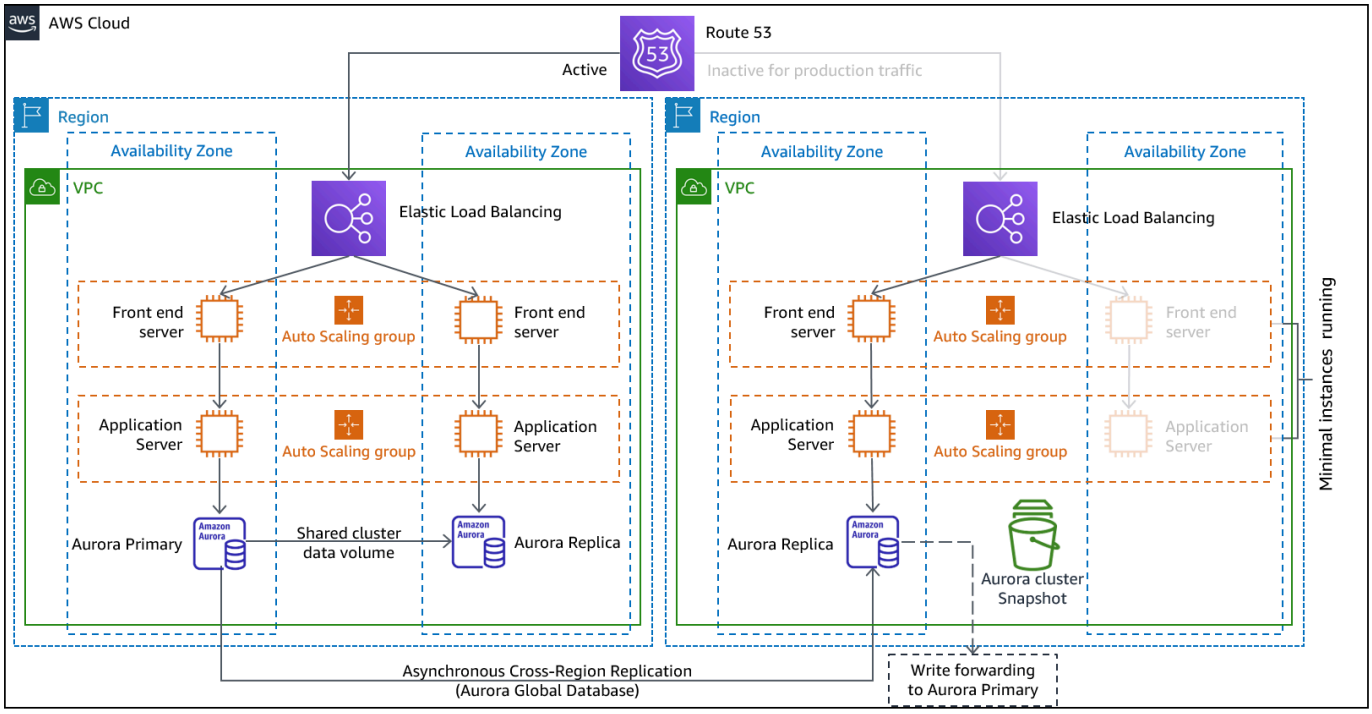

Figura 11 – Architettura dell'approccio standby a freddo

Nota: la differenza tra Pilot Light e standby a freddo può essere talvolta di difficile comprensione. Entrambi gli approcci includono un ambiente nella regione di ripristino di emergenza con copie delle risorse della regione primaria. La differenza è che l'approccio Pilot Light non è in grado di elaborare le richieste senza prima ulteriori operazioni, mentre l'approccio standby a freddo può gestire il traffico (a livelli di capacità ridotti) immediatamente. L'approccio Pilot Light richiede l'attivazione dei server, possibilmente la distribuzione di un'infrastruttura (non principale) aggiuntiva e l'aumento di risorse, mentre lo standby a freddo richiede solo l'aumento di risorse (tutto è già distribuito e in esecuzione). Usa le tue esigenze in fatto di RTO e RPO come base per scegliere tra questi approcci.

Servizi AWS

Tutti i servizi AWS presentati nelle sezioni Backup e ripristino e Pilot Light vengono usati anche nell'approccio standby a freddo per il backup dei dati, la replica dei dati, il routing del traffico attivo/in standby e la distribuzione dell'infrastruttura, incluse le istanze EC2.

Viene usato AWS Auto Scaling per dimensionare le risorse, incluse le istanze Amazon EC2, le attività Amazon ECS, la velocità effettiva di Amazon DynamoDB e le repliche di Amazon Aurora all'interno di una regione AWS. Amazon EC2 Auto Scaling dimensiona la distribuzione dell'istanza EC2 tra zone di disponibilità all'interno di una regione AWS, fornendo resilienza all'interno della regione.

Usa Auto Scaling per aumentare orizzontalmente le risorse della regione di ripristino di emergenza fino alla capacità completa di produzione, come parte di una strategia Pilot Light o di standby a freddo. Ad esempio, per EC2, aumenta l'impostazione della capacità desiderata nel gruppo Auto Scaling. Puoi modificare questa impostazione manualmente tramite la AWS Management Console, automaticamente tramite l'SDK AWS o ridistribuendo il modello AWS CloudFormation usando il nuovo valore di capacità desiderato. Puoi usare parametri AWS CloudFormation per semplificare la ridistribuzione del modello CloudFormation. Assicurati che le quote del servizio nella regione di ripristino di emergenza siano sufficientemente elevate da non limitare l'aumento fino alla capacità di produzione.

Attivo/attivo multisito

Puoi eseguire il carico di lavoro contemporaneamente in più regioni come parte di un approccio attivo/attivo multisito o attivo/passivo con standby a caldo. L'approccio attivo/attivo multisito rende disponibile il traffico da tutte le regioni in cui viene distribuito, mentre l'approccio standby a caldo rende disponibile il traffico da un'unica regione, usando le altre regioni solo per il ripristino di emergenza. Con un approccio attivo/attivo multisito, gli utenti possono accedere al carico di lavoro in una qualsiasi delle regioni in cui viene distribuito. Questo è l'approccio più complesso e costoso al ripristino di emergenza, ma può ridurre il tempo di ripristino fino quasi a zero per la maggior parte delle emergenze con l'implementazione e le scelte tecnologiche corrette. Tuttavia, può essere necessario usare i backup per la corruzione dei dati, il che in genere comporta un punto di ripristino diverso da zero. Lo standby a caldo usa una configurazione attiva/passiva in cui gli utenti vengono indirizzati a una sola regione, mentre le regioni di ripristino di emergenza non accettano traffico. La maggior parte dei clienti che intende creare un ambiente completo nella seconda regione, ritiene opportuno usare una configurazione attiva/attiva. In alternativa, se non vuoi usare entrambe le regioni per gestire il traffico degli utenti, lo standby a freddo offre un approccio più economico e meno complesso dal punto di vista operativo.

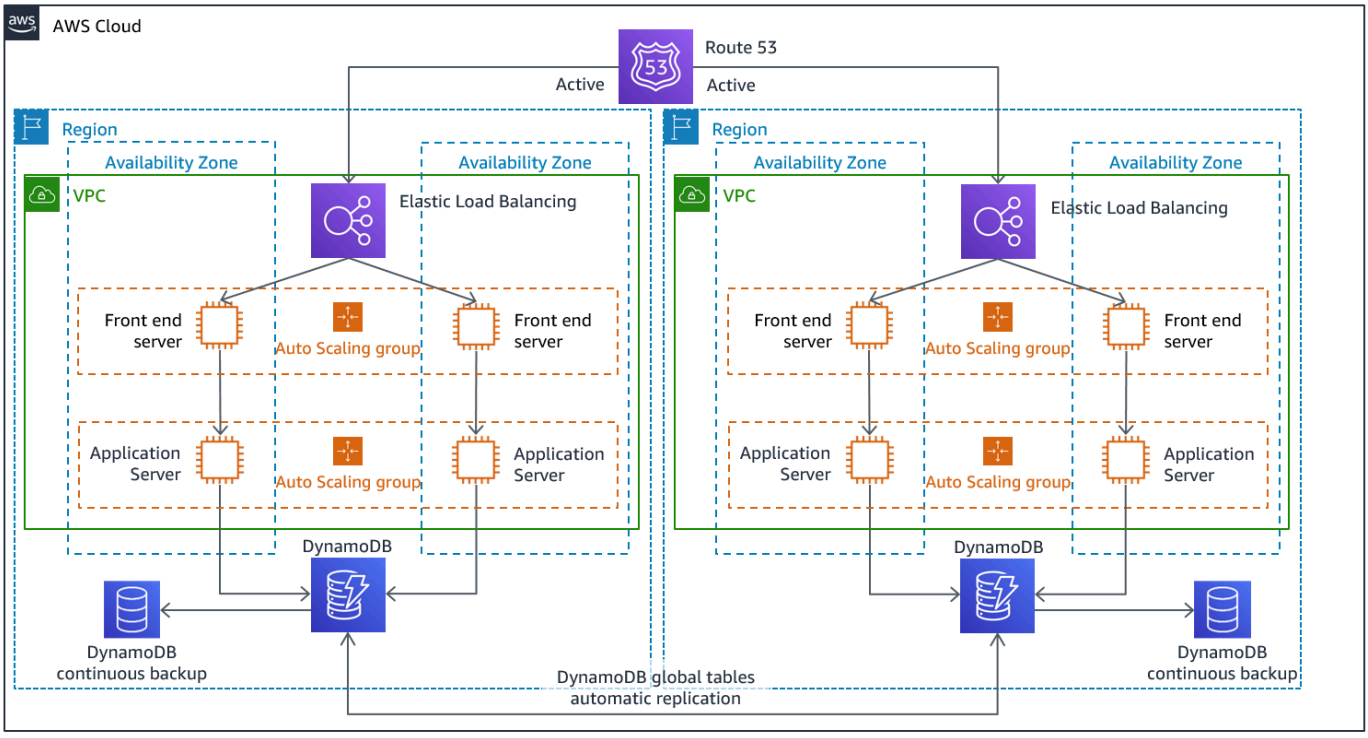

Figura 12 – Architettura dell'approccio attivo/attivo multisito (cambia un percorso attivo in inattivo per lo standby a caldo)

Poiché il carico di lavoro è in esecuzione in più di una regione, l'approccio attivo/attivo multisito è uno scenario senza failover. I test del ripristino di emergenza in questo caso sono incentrati sul modo in cui il carico di lavoro reagisce alla perdita di una regione: il traffico viene allontanato dalla regione con errori? Le altre regioni possono gestire tutto il traffico? Sono necessari test anche per un'emergenza relativa ai dati. Il backup e il ripristino sono comunque necessari e devono essere testati regolarmente. Va anche notato che i tempi di ripristino per un'emergenza relativa ai dati che comporta corruzione, eliminazione o offuscamento dei dati saranno sempre maggiori di zero e che il punto di ripristino corrisponderà sempre a un punto precedente all'individuazione dell'emergenza. Se la complessità e i costi aggiuntivi di un approccio multisito attivo/attivo (o standby a caldo) sono necessarie per mantenere tempi di ripristino prossimi a zero, dovranno essere intraprese altre iniziative per garantire la sicurezza e prevenire l'errore umano in modo da attenuare l'impatto di questo tipo di emergenze.

Servizi AWS

Tutti i servizi AWS presentati nelle sezioni Backup e ripristino,Pilot Light e Standby a freddo vengono usati anche qui per backup dei dati point-in-time, replica dei dati, routing del traffico attivo/attivo e distribuzione e dimensionamento dell'infrastruttura, incluse le istanze EC2.

Per gli scenari attivi/passivi discussi in precedenza (Pilot Light e standby a freddo), è possibile usare sia Amazon Route 53 sia AWS Global Accelerator per instradare il traffico verso la regione attiva. Per l'approccio attivo/attivo descritto qui, entrambi questi servizi permettono anche la definizione di policy che determinano quali utenti accedono a ogni endpoint regionale attivo. Con AWS Global Accelerator puoi impostare una chiamata al traffico per controllare la percentuale di traffico indirizzata a ogni endpoint dell'applicazione. Amazon Route 53 supporta questo approccio in percentuale e molte altre policy disponibili, tra cui quelle basate su geoprossimità e latenza. Global Accelerator utilizza automaticamente l'estesa rete di edge server AWS per inserire il traffico nella dorsale di rete (backbone) AWS il prima possibile, con la conseguente riduzione della latenza delle richieste.

La replica dei dati con questa strategia assicura un RPO vicino a zero. Servizi come Amazon Aurora Global Database usano un'infrastruttura dedicata che lascia il database completamente disponibile per la gestione dell'applicazione e può replicarsi in una regione secondaria con latenza tipica inferiore a un secondo. Con un approccio attivo/passivo, le scritture vengono eseguite solo nella regione primaria. La differenza con l'approccio attivo/attivo sta nella progettazione di quante scritture vengono gestite da ogni regione attiva. È prassi comune progettare la strategia in modo che le letture degli utenti vengano gestite dalla regione a esse più vicina, nota come lettura in locale. Con le scritture, sono disponibili diverse opzioni:

- Una strategia di scrittura globale instrada tutte le scritture verso un'unica regione. In caso di errore nella regione, un'altra regione viene promossa in modo da accettare le scritture. Aurora Global Database è adatto per la scrittura globale, in quanto supporta la sincronizzazione con repliche di lettura tra regioni e permette di promuovere una delle regioni secondarie in modo che si occupino delle letture/scritture in meno di un minuto.
- Una strategia di scrittura in locale instrada le scritture nella regione più vicina (proprio come per le letture). Le tabelle globali Amazon DynamoDB supportano questa strategia, permettendo le letture e le scritture da ogni regione in cui viene distribuita la tabella globale. Le tabelle globali Amazon DynamoDB usano una riconciliazione per cui l'ultima scrittura è quella vincente tra aggiornamenti simultanei.
- Una strategia di scrittura partizionata assegna le scritture a una regione specifica in base a una chiave di partizione (come l'ID utente) per evitare conflitti di scrittura. Per questo caso,

è possibile usare la replica di Amazon S3 configurata in modo bidirezionale, che supporta attualmente la replica tra due regioni. Nell'implementare questo approccio, assicurati di abilitare la sincronizzazione delle modifiche di replica nei due bucket A e B per replicare le modifiche ai metadati della replica, come liste di controllo accessi agli oggetti, tag degli oggetti o blocchi degli oggetti, negli oggetti replicati. Puoi anche configurare se replicare o meno i contrassegni di eliminazione tra bucket nelle regioni attive. Oltre alla replica, la strategia deve includere anche backup point-in-time per la protezione da eventi di corruzione o eliminazione dei dati.

AWS CloudFormation è un potente strumento per applicare un'infrastruttura distribuita in modo coerente tra account AWS in più regioni AWS. AWS CloudFormation StackSets estende questa funzionalità permettendoti di creare, aggiornare o eliminare stack CloudFormation tra più account e regioni con un'unica operazione. Benché AWS CloudFormation usi YAML o JSON per definire un modello Infrastructure as Code, il AWS Cloud Development Kit (AWS CDK) permette di usare i linguaggi di programmazione più familiari. Il codice viene convertito in CloudFormation e viene quindi usato per distribuire risorse in AWS.

Rilevamento

È importante determinare al più presto quando i carichi di lavoro non forniscono i risultati aziendali previsti. In questo modo, puoi dichiarare rapidamente un'emergenza e attivare il ripristino da un incidente. Per obiettivi di ripristino aggressivi, questo tempo di risposta, insieme a informazioni appropriate, è essenziale per realizzare gli obiettivi di ripristino. Se l'obiettivo del punto di ripristino è un'ora, devi rilevare l'incidente, informare il personale appropriato, attivare i processi di riassegnazione, valutare le informazioni (se disponibili) sul tempo previsto per il ripristino (senza eseguire il piano di ripristino di emergenza), dichiarare un'emergenza ed eseguire il ripristino entro un'ora.

> **Note**
>
> Se le parti interessate decidono di non richiamare il ripristino di emergenza anche se l'RTO è a rischio, rivaluta i piani e gli obiettivi di ripristino di emergenza. La decisione di non richiamare i piani di ripristino di emergenza può essere determinata dall'inadeguatezza dei piani stessi o da una mancanza di fiducia nella loro esecuzione.

Nella pianificazione e nelle finalità è essenziale tenere conto di rilevamento dell'incidente, notifica, riassegnazione, individuazione e dichiarazione, in modo da fornire obiettivi realizzabili realistici che assicurino valore aziendale.

AWS pubblica le informazioni più aggiornate relative alla disponibilità dei servizi in Service Health Dashboard. Consulta questa risorsa in qualsiasi momento per ottenere informazioni aggiornate sullo stato o iscriviti a un feed RSS per ricevere notifiche sulle interruzioni per ogni singolo servizio. Se riscontri un problema operativo in tempo reale relativo a uno dei nostri servizi che non viene visualizzato in Service Health Dashboard, puoi creare una richiesta di assistenza.

AWS Health Dashboard fornisce informazioni su eventi AWS Health che possono interessare il tuo account. Le informazioni vengono presentate in due modi: un pannello di controllo che mostra eventi recenti e imminenti organizzati per categoria e un registro eventi completo che mostra tutti gli eventi degli ultimi 90 giorni.

Per i requisiti più rigorosi in fatto di RTO, puoi implementare il failover automatico in base a controlli dell'integrità. Progetta controlli dell'integrità che siano rappresentativi dell'esperienza utente e basati su indicatori chiave delle prestazioni (KPI). I controlli dell'integrità approfonditi utilizzano le

funzionalità principali del carico di lavoro e vanno oltre semplici controlli heartbeat. Usa controlli dell'integrità approfonditi basati su più segnali. Usa questo approccio con cautela, in modo da non attivare falsi allarmi, perché il failover, se non è necessario, può di per sé introdurre rischi di disponibilità.

Test del ripristino di emergenza

Testa l'implementazione del ripristino di emergenza per convalidarla e testa regolarmente il failover nella regione di ripristino di emergenza del carico di lavoro per garantire che l'RTO e l'RPO vengano soddisfatti.

Un modello da evitare è lo sviluppo di percorsi di ripristino eseguiti raramente. Ad esempio, è possibile che si disponga di un archivio dati secondario utilizzato per query di sola lettura. Quando scrivi in un archivio dati e quello principale ha un guasto, puoi eseguire il failover verso l'archivio dati secondario. Se non testi frequentemente questo failover, è possibile che i presupposti relativi alle funzionalità dell'archivio dati secondario non siano corretti. La capacità dell'archivio secondario, che potrebbe essere stata sufficiente durante l'ultimo test, può non essere più in grado di tollerare il carico in questo scenario oppure le quote dei servizi nella regione secondaria possono non essere sufficienti.

La nostra esperienza ha dimostrato che l'unico ripristino da errore che funziona è il percorso testato più spesso. Per questo motivo, l'opzione migliore consiste nel definire un numero ridotto di percorsi di ripristino.

Puoi stabilire modelli di ripristino e testarli regolarmente. In presenza di un percorso di ripristino complesso o critico, devi comunque riprodurre regolarmente l'errore nell'ambiente di produzione per convalidare il funzionamento del percorso di ripristino.

Gestisci la deviazione della configurazione nella regione di ripristino di emergenza. Assicurati che l'infrastruttura, i dati e la configurazione siano quelli necessari nella regione di ripristino di emergenza. Ad esempio, controlla che le AMI e le quote dei servizi siano aggiornate.

Puoi utilizzare AWS Config per monitorare e registrare continuamente le configurazioni delle risorse AWS. AWS Config è in grado di rilevare la deviazione e attivare AWS Systems Manager Automation per correggerla e generare allarmi. AWS CloudFormation può anche rilevare la deviazione negli stack distribuiti.

Conclusione

I clienti sono responsabili della disponibilità delle proprie applicazioni nel cloud. È importante definire che cosa sia un'emergenza e predisporre un piano di ripristino di emergenza che rifletta questa definizione e l'impatto che può avere sui risultati aziendali. Crea un obiettivo del tempo di ripristino (RTO) e un obiettivo del punto di ripristino (RPO) in base all'analisi dell'impatto e alle valutazioni dei rischi, quindi scegli l'architettura appropriata per attenuare la portata delle emergenze. Verifica che il rilevamento delle emergenze sia possibile e tempestivo: è essenziale identificare i casi in cui gli obiettivi sono a rischio. Assicurati di avere un piano e convalidalo tramite test. I piani di ripristino di emergenza non convalidati rischiano di non essere implementati a causa di mancanza di fiducia o del non raggiungimento degli obiettivi di ripristino di emergenza.

Collaboratori

Hanno contribuito a questo documento:

- Alex Livingstone, Practice Lead Cloud Operations, AWS Enterprise Support
- Seth Eliot, Principal Reliability Solutions Architect, Amazon Web Services

Approfondimenti

Per ulteriori informazioni, consulta:

- Principio dell'affidabilità: AWS Well-Architected Framework
- Lista di controllo per i piani di ripristino di emergenza
- Implementazione dei controlli dell'integrità
- Implementazione di soluzioni AWS: Multi-Region Application Architecture
- AWS re:Invent 2018: Architecture Patterns for Multi-Region Active-Active Applications (ARC209-R2)

Cronologia dei documenti

Modifica	Descrizione	Data
Pubblicazione iniziale	Prima pubblicazione.	12 febbraio 2021

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

Avvisi

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

© 2021, Amazon Web Services, Inc., o sue affiliate. Tutti i diritti riservati.