aws

Guida per l'utente

Application Auto Scaling

Application Auto Scaling: Guida per l'utente

Table of Contents

Che cos'è Application Auto Scaling?

Application Auto Scaling è un servizio Web per sviluppatori e amministratori di sistema che necessitano di una soluzione per scalare automaticamente le proprie risorse scalabili per singoli servizi oltre AWS ad Amazon EC2. Con Application Auto Scaling, è possibile configurare il ridimensionamento automatico per le seguenti risorse: : AWS

- AppStream flotte 2.0
- Repliche Aurora
- Endpoint di classificazione dei documenti Amazon Comprehend e di riconoscimento delle identità
- Tabelle DynamoDB e indici secondari globali
- Servizi Amazon ECS
- ElastiCache per cluster Redis (gruppi di replica)
- Cluster Amazon EMR
- Tabelle di Amazon Keyspaces (per Apache Cassandra)
- Provisioning simultaneo della funzione Lambda
- Archiviazione broker Amazon Managed Streaming for Apache Kafka (MSK)
- Cluster Amazon Neptune
- SageMaker varianti degli endpoint
- SageMaker componenti di inferenza
- SageMaker Concorrenza fornita senza server
- Richieste di parchi istanze Spot
- Risorse personalizzate fornite dalle tue applicazioni o dai tuoi servizi. Per ulteriori informazioni, consulta il repository. GitHub

Per vedere la disponibilità regionale per uno qualsiasi dei AWS servizi sopra elencati, consulta la tabella delle regioni nella tabella delle

Per ulteriori informazioni sul dimensionamento del parco istanze Amazon EC2 utilizzando i gruppi Auto Scaling, consulta la Guida per l'utente di Amazon EC2 Auto Scaling.

Caratteristiche di Application Auto Scaling

Application Auto Scaling ti consente di dimensionare automaticamente le risorse scalabili in base alle condizioni da te definite.

- Ridimensionamento del tracciamento degli obiettivi: ridimensiona una risorsa in base a un valore target per una CloudWatch metrica specifica.
- Dimensionamento per fasi: esegue il dimensionamento di una risorsa in base a un set di adeguamenti del dimensionamento che variano in base alle dimensioni dell'utilizzo fuori limite segnalato dall'allarme.
- Dimensionamento pianificato: esegue il dimensionamento di una risorsa solamente una tantum o in base a una pianificazione ricorrente.

Funzionano con Application Auto Scaling

È possibile configurare il dimensionamento utilizzando le seguenti interfacce a seconda della risorsa che si sta scalando:

- AWS Management Console: fornisce un'interfaccia Web da utilizzare per configurare il dimensionamento. Se hai registrato un AWS account, accedi ad Application Auto Scaling accedendo a. AWS Management Console Apri quindi la console di servizio per una delle risorse elencate nell'introduzione. Assicurati di aprire la console nella Regione AWS stessa risorsa con cui desideri lavorare.

  > **Note**
  >
  > L'accesso alla console non è disponibile per tutte le risorse. Per ulteriori informazioni, consulta Servizi AWS che puoi usare con Application Auto Scaling.

- AWS Command Line Interface (AWS CLI) — Fornisce comandi per un ampio set di Servizi AWS ed è supportato su Windows, macOS e Linux. Per iniziare, consulta AWS Command Line Interface. Per un elenco di comandi, vedete application-autoscaling nel Command Reference.AWS CLI
- AWS Tools for Windows PowerShell— Fornisce comandi per un'ampia gamma di AWS prodotti per coloro che eseguono script nell'ambiente. PowerShell Per iniziare, consulta la Guida per l'utente di AWS Tools for Windows PowerShell. Per ulteriori informazioni, consulta la Documentazione di riferimento per Cmdlet AWS Tools for PowerShell.

- AWS SDK: forniscono operazioni API specifiche per la lingua e si occupano di molti dettagli di connessione, come il calcolo delle firme, la gestione dei tentativi di richiesta e la gestione degli errori. Per ulteriori informazioni, consulta Strumenti su cui basarsi. AWS
- HTTPS API: forniscono operazioni API di basso livello accessibili tramite richieste HTTPS. Per ulteriori informazioni, consulta Documentazione di riferimento sull'API Application Auto Scaling.
- AWS CloudFormation— Supporta la configurazione del ridimensionamento utilizzando un CloudFormation modello. Per ulteriori informazioni, consulta Configurare le risorse di Application Auto Scaling utilizzando AWS CloudFormation.

Per connettersi a livello di codice a un Servizio AWS, si utilizza un endpoint. l'utente della regione.

Concetti relativi all'Application Auto Scaling

In questo argomento vengono illustrati i concetti chiave di Application Auto Scaling che consentono di iniziare a utilizzarlo.

Obiettivo scalabile

Un'entità creata per specificare la risorsa che si desidera dimensionare. Ogni obiettivo scalabile è identificato in modo univoco da uno spazio dei nomi del servizio, un ID risorsa e una dimensione scalabile, che rappresenta una dimensione della capacità del servizio sottostante. Ad esempio, un servizio Amazon ECS supporta la scalabilità automatica del conteggio delle attività, una tabella DynamoDB supporta la scalabilità automatica della capacità di lettura e scrittura della tabella e dei relativi indici secondari globali e un cluster Aurora supporta il dimensionamento del conteggio delle repliche.

> **Tip**
>
> Ogni obiettivo scalabile ha inoltre una capacità minima e massima. Le policy di dimensionamento non saranno mai superiori o inferiori all'intervallo minimo-massimo. È possibile apportare out-of-band modifiche direttamente alle risorse sottostanti che non rientrano in questo intervallo, di cui Application Auto Scaling non è a conoscenza. Tuttavia, ogni volta che viene richiamata una policy di dimensionamento o l'API `RegisterScalableTarget`, Application Auto Scaling recupera la capacità corrente e la confronta con la capacità minima e massima. Se non rientra nell'intervallo minimo-

massimo, la capacità viene aggiornata in modo da rispettare il minimo e il massimo impostati.

Dimensionamento orizzontale (riduzione)

Quando Application Auto Scaling riduce automaticamente la capacità per un obiettivo scalabile, l'obiettivo scalabile si riduce orizzontalmente. Quando vengono impostate le policy di dimensionamento, non possono scalare nella destinazione scalabile una capacità inferiore alla capacità minima.

Aumento orizzontale

Quando Application Auto Scaling aumenta automaticamente la capacità per un obiettivo scalabile, l'obiettivo scalabile aumenta orizzontalmente. Quando vengono impostate le policy di dimensionamento, non possono scalare nella destinazione scalabile una capacità superiore alla capacità massima.

Policy di dimensionamento

Una politica di scalabilità indica ad Application Auto Scaling di tenere traccia di una metrica specifica. CloudWatch Quindi, determina l'operazione di dimensionamento da eseguire quando il parametro è superiore o inferiore a un determinato valore di soglia. Ad esempio, è possibile che desideri aumentare orizzontalmente se l'utilizzo della CPU nel cluster inizia ad aumentare, e ridurre orizzontalmente quando scende di nuovo.

Le metriche utilizzate per la scalabilità automatica vengono pubblicate dal servizio di destinazione, ma puoi anche pubblicare la tua metrica CloudWatch e quindi utilizzarla con una politica di scalabilità.

Un periodo di tempo di raffreddamento tra le attività di dimensionamento consente alla risorsa di stabilizzarsi prima che inizi un'altra attività di dimensionamento. Application Auto Scaling continua a valutare i parametri durante il tempo di raffreddamento. Al termine del tempo di raffreddamento, la policy di dimensionamento avvia un'altra attività di dimensionamento, se necessario. Mentre è attivo un tempo di raffreddamento, se è necessario un aumento orizzontale maggiore in base al valore del parametro corrente, la policy di dimensionamento aumenta orizzontalmente immediatamente.

Operazioni pianificate

Le operazioni pianificate dimensionano automaticamente le risorse in una data e un'ora specifiche. Funzionano modificando la capacità minima e massima per un obiettivo scalabile

e possono quindi essere utilizzate per ridurre orizzontalmente in base a una pianificazione impostando una capacità minima elevata o una capacità massima bassa. Ad esempio, è possibile utilizzare le operazioni pianificate per dimensionare un'applicazione che non consuma risorse nei fine settimana diminuendo la capacità il venerdì e aumentando la capacità il lunedì successivo.

È inoltre possibile utilizzare le operazioni pianificate per ottimizzare i valori minimi e massimi nel tempo per adattarsi a situazioni in cui è previsto un traffico superiore al normale, ad esempio campagne di marketing o fluttuazioni stagionali. In questo modo è possibile migliorare le prestazioni nei momenti in cui è necessario aumentare orizzontalmente le risorse per far fronte al maggiore utilizzo, e ridurre i costi quando si utilizzano meno risorse.

Ulteriori informazioni

Servizi AWS che puoi usare con Application Auto Scaling - Questa sezione illustra i servizi che è possibile dimensionare e consente di impostare la scalabilità automatica registrando un obiettivo scalabile. Vengono inoltre descritti tutti i ruoli collegati ai servizi IAM creati da Application Auto Scaling per accedere alle risorse nel servizio obiettivo.

Policy di dimensionamento con monitoraggio degli obiettivi per Application Auto Scaling - Una delle caratteristiche principali di Application Auto Scaling è la disponibilità di policy di dimensionamento con monitoraggio degli obiettivi. Scopri come le policy di monitoraggio degli obiettivi regolano automaticamente la capacità desiderata per mantenere l'utilizzo a un livello costante in base ai parametri e ai valori obiettivo configurati. Ad esempio, si può configurare il monitoraggio degli obiettivi per mantenere al 50% l'utilizzo della CPU di un Parco istanze Spot. Application Auto Scaling avvia o termina le istanze EC2 in base alle esigenze per mantenere l'utilizzo aggregato della CPU su tutti i server al 50%.

Servizi AWS che puoi usare con Application Auto Scaling

Application Auto Scaling si integra con altri AWS servizi in modo da poter aggiungere funzionalità di scalabilità per soddisfare la domanda dell'applicazione. La scalabilità automatica è una caratteristica facoltativa del servizio disabilitata per impostazione predefinita in quasi tutti i casi.

La tabella seguente elenca i AWS servizi che è possibile utilizzare con Application Auto Scaling, incluse informazioni sui metodi supportati per la configurazione dell'auto scaling. È inoltre possibile utilizzare Application Auto Scaling con risorse personalizzate.

- Accesso tramite console - È possibile configurare un servizio AWS compatibile per avviare la scalabilità automatica configurando una policy di dimensionamento nella console del servizio obiettivo.
- Accesso tramite CLI - È possibile configurare un servizio AWS compatibile per avviare la scalabilità automatica utilizzando la AWS CLI.
- Accesso SDK: puoi configurare un AWS servizio compatibile per avviare la scalabilità automatica utilizzando gli AWS SDK.
- CloudFormation accesso: è possibile configurare un AWS servizio compatibile per avviare la scalabilità automatica utilizzando un modello di AWS CloudFormation stack. Per ulteriori informazioni, consulta Configurare le risorse di Application Auto Scaling utilizzando AWS CloudFormation.

AWS servizio	Accesso alla console[1]	Accesso tramite CLI	Accesso tramite SDK	CloudFormation accesso
AppStream 2.0	 Sì	 Sì	 Sì	 Sì
Aurora	 Sì	 Sì	 Sì	 Sì

AWS servizio	Accesso alla console[1]	Accesso tramite CLI	Accesso tramite SDK	CloudFormation accesso
Amazon Comprehend	No	Sì	Sì	Sì
Amazon DynamoDB	Sì	Sì	Sì	Sì
Amazon ECS	Sì	Sì	Sì	Sì
Amazon ElastiCache	Sì	Sì	Sì	Sì
Amazon EMR	Sì	Sì	Sì	Sì
Amazon Keyspaces	Sì	Sì	Sì	Sì
Lambda	No	Sì	Sì	Sì
Amazon MSK	Sì	Sì	Sì	Sì

AWS servizio	Accesso alla console[1]	Accesso tramite CLI	Accesso tramite SDK	CloudFormation accesso
Amazon Neptune	No	Sì	Sì	Sì
SageMaker	Sì	Sì	Sì	Sì
Spot Fleet	Sì	Sì	Sì	Sì
Risorse personalizzate	No	Sì	Sì	Sì

[1] Accesso alla console per la configurazione delle politiche di scalabilità. La maggior parte dei servizi non supporta la configurazione del ridimensionamento pianificato dalla console. Attualmente, solo Amazon AppStream 2.0 e Spot Fleet forniscono l'accesso alla console per la scalabilità pianificata. ElastiCache

Amazon AppStream 2.0 e Application Auto Scaling

Puoi scalare le flotte AppStream 2.0 utilizzando le politiche di scalabilità di Target Tracking, le politiche di scalabilità per fasi e la scalabilità pianificata.

Utilizzate le seguenti informazioni per aiutarvi a integrare la AppStream versione 2.0 con Application Auto Scaling.

Ruolo collegato al servizio creato per la versione 2.0 AppStream

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione delle risorse AppStream 2.0 come destinazioni scalabili con Application Auto Scaling.

Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_AppStreamFleet`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `appstream.application-autoscaling.amazonaws.com`

Registrazione di flotte AppStream 2.0 come obiettivi scalabili con Application Auto Scaling

Application Auto Scaling richiede un target scalabile prima di poter creare policy di scalabilità o azioni pianificate per una flotta 2.0. AppStream Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Se configuri la scalabilità automatica utilizzando la console AppStream 2.0, la AppStream versione 2.0 registra automaticamente una destinazione scalabile per te.

Se desideri configurare la scalabilità automatica utilizzando la AWS CLI o uno degli SDK, puoi utilizzare AWS le seguenti opzioni:

- AWS CLI:

  Chiamate il register-scalable-targetcomando per una AppStream flotta 2.0. Nell'esempio seguente viene registrata la capacità desiderata di un parco istanze denominato `sample-fleet`, con una capacità minima di un'istanza e una capacità massima di cinque istanze del parco istanze.

```
aws application-autoscaling register-scalable-target \
   --service-namespace appstream \
   --scalable-dimension appstream:fleet:DesiredCapacity \
   --resource-id fleet/sample-fleet \
   --min-capacity 1 \
   --max-capacity 5
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità delle tue risorse AppStream 2.0 nella seguente documentazione:

Fleet Auto Scaling for AppStream 2.0 nella Guida all'amministrazione di Amazon AppStream 2.0

Amazon Aurora e Application Auto Scaling

È possibile dimensionare cluster database di Aurora utilizzando le policy di dimensionamento con monitoraggio degli obiettivi, le policy di dimensionamento per fasi e il dimensionamento pianificato.

Utilizza le informazioni riportate di seguito per semplificare l'integrazione di Aurora con Application Auto Scaling.

Ruolo collegato ai servizi creato per Aurora

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione delle risorse Aurora come destinazioni scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_RDSCluster`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `rds.application-autoscaling.amazonaws.com`

Registrazione di cluster database Aurora come obiettivi scalabili con Application Auto Scaling

Application Auto Scaling richiede un obiettivo scalabile prima di poter creare policy di dimensionamento o operazioni pianificate per un cluster Aurora. Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Se configuri la scalabilità automatica usando la console Aurora, Aurora registra automaticamente un obiettivo scalabile per tuo conto.

Se desideri configurare la scalabilità automatica utilizzando la AWS CLI o uno degli SDK, puoi utilizzare AWS le seguenti opzioni:

- AWS CLI:

  Chiama il comando register-scalable-target per un cluster Aurora. Nell'esempio seguente viene registrato il conteggio di repliche Aurora in un cluster denominato `my-db-cluster`, con una capacità minima di una replica Aurora e una capacità massima di otto repliche Aurora.

```
aws application-autoscaling register-scalable-target \
  --service-namespace rds \
  --scalable-dimension rds:cluster:ReadReplicaCount \
  --resource-id cluster:my-db-cluster \
  --min-capacity 1 \
  --max-capacity 8
```

  In caso di esito positivo, questo comando restituisce l'ARN dell'obiettivo scalabile.

```
{
```

```
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità delle risorse Aurora nella seguente documentazione:

Utilizzo della scalabilità automatica di Amazon Aurora con le repliche Aurora nella Guida per l'utente di Amazon RDS

Amazon Comprehend e Application Auto Scaling

Puoi dimensionare la classificazione dei documenti e gli endpoint di riconoscimento delle entità di Amazon Comprehend utilizzando le policy di dimensionamento con monitoraggio degli obiettivi e il dimensionamento pianificato.

Utilizza le informazioni riportate di seguito per semplificare l'integrazione di Amazon Comprehend con Application Auto Scaling.

Ruolo collegato ai servizi creato per Amazon Comprehend

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione delle risorse Amazon Comprehend come destinazioni scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_ComprehendEndpoint`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `comprehend.application-autoscaling.amazonaws.com`

Registrazione delle risorse Amazon Comprehend come obiettivi scalabili con Application Auto Scaling

Application Auto Scaling richiede un obiettivo scalabile prima di poter creare policy di dimensionamento o operazioni pianificate per un endpoint di classificazione di documenti o riconoscimento delle entità di Amazon Comprehend. Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Per configurare la scalabilità automatica utilizzando la AWS CLI o uno AWS degli SDK, puoi utilizzare le seguenti opzioni:

- AWS CLI:

  Chiama il comando register-scalable-target per un endpoint di classificazione dei documenti. Nell'esempio seguente viene registrato il numero desiderato di unità di inferenza che il modello deve utilizzare per un endpoint di classificazione di documenti utilizzando l'ARN dell'endpoint, con una capacità minima di un'unità di inferenza e una capacità massima di tre unità di inferenza.

```
aws application-autoscaling register-scalable-target \
  --service-namespace comprehend \
  --scalable-dimension comprehend:document-classifier-endpoint:DesiredInferenceUnits
 \
  --resource-id arn:aws:comprehend:us-west-2:123456789012:document-classifier-
endpoint/EXAMPLE \
  --min-capacity 1 \
  --max-capacity 3
```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Chiama il comando `register-scalable-target` per un endpoint di riconoscimento delle entità. Nell'esempio seguente viene registrato il numero desiderato di unità di inferenza che il modello deve utilizzare per un riconoscitore delle entità utilizzando l'ARN dell'endpoint, con una capacità minima di un'unità di inferenza e una capacità massima di tre unità di inferenza.

```
aws application-autoscaling register-scalable-target \
  --service-namespace comprehend \
  --scalable-dimension comprehend:entity-recognizer-endpoint:DesiredInferenceUnits \
  --resource-id arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-
endpoint/EXAMPLE \
  --min-capacity 1 \
  --max-capacity 3
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità delle tue risorse Amazon Comprehend nella seguente documentazione:

Scalabilità automatica con gli endpoint nella Guida per gli sviluppatori di Amazon Comprehend

Amazon DynamoDB e Application Auto Scaling

È possibile dimensionare le tabelle DynamoDB utilizzando le policy di dimensionamento con monitoraggio degli obiettivi e il dimensionamento pianificato.

Utilizza le informazioni riportate di seguito per semplificare l'integrazione di DynamoDB con Application Auto Scaling.

Ruolo collegato ai servizi creato per DynamoDB

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione delle risorse DynamoDB come destinazioni scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_DynamoDBTable`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `dynamodb.application-autoscaling.amazonaws.com`

Registrazione delle risorse DynamoDB come obiettivi scalabili con Application Auto Scaling

Application Auto Scaling richiede un obiettivo scalabile prima di poter creare policy di dimensionamento o operazioni pianificate per una tabella o un indice secondario globale DynamoDB. Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Se configuri la scalabilità automatica usando la console DynamoDB, DynamoDB registra automaticamente un obiettivo scalabile per tuo conto.

Se desideri configurare la scalabilità automatica utilizzando la AWS CLI o uno degli SDK, puoi utilizzare AWS le seguenti opzioni:

- AWS CLI:

  Chiamate il register-scalable-targetcomando per determinare la capacità di scrittura di una tabella. Nell'esempio seguente viene registrata la capacità in scrittura assegnata di una tabella denominata `my-table`, con una capacità minima di cinque unità di capacità in scrittura e una capacità massima di 10 unità di capacità in scrittura.

```
aws application-autoscaling register-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:WriteCapacityUnits \
  --resource-id table/my-table \
  --min-capacity 5 \
  --max-capacity 10
```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

  Chiama il register-scalable-targetcomando per determinare la capacità di lettura di una tabella. Nell'esempio seguente viene registrata la capacità in lettura assegnata di una tabella denominata `my-table`, con una capacità minima di cinque unità di capacità in lettura e una capacità massima di 10 unità di lettura.

```
aws application-autoscaling register-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits \
  --resource-id table/my-table \
  --min-capacity 5 \
  --max-capacity 10
```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
```

```
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Chiama il register-scalable-targetcomando per la capacità di scrittura di un indice secondario globale. Nell'esempio seguente viene registrata la capacità in scrittura assegnata di un indice secondario globale denominato `my-table-index`, con una capacità minima di cinque unità di capacità in scrittura e una capacità massima di 10 unità di capacità in scrittura.

```
aws application-autoscaling register-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:index:WriteCapacityUnits \
  --resource-id table/my-table/index/my-table-index \
  --min-capacity 5 \
  --max-capacity 10
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Chiama il register-scalable-targetcomando per la capacità di lettura di un indice secondario globale. Nell'esempio seguente viene registrata la capacità in lettura assegnata di un indice secondario globale `my-table-index`, con una capacità minima di cinque unità di capacità in lettura e una capacità massima di 10 unità di capacità in lettura.

```
aws application-autoscaling register-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:index:ReadCapacityUnits \
  --resource-id table/my-table/index/my-table-index \
  --min-capacity 5 \
  --max-capacity 10
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
```

```
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità delle tue risorse DynamoDB nella seguente documentazione:

- Gestione della capacità di velocità effettiva con DynamoDB Auto Scaling nella Guida per gli sviluppatori di Amazon DynamoDB
- Valuta le impostazioni di ridimensionamento automatico della tabella nella Amazon DynamoDB Developer Guide
- Come usare AWS CloudFormation per configurare la scalabilità automatica per le tabelle e gli indici DynamoDB sul blog AWS

Puoi anche trovare un tutorial per il ridimensionamento programmato in. Tutorial: Guida di base su dimensionamento pianificato utilizzando AWS CLI In questo tutorial, apprenderai i passaggi di base per configurare il dimensionamento in modo che la tabella DynamoDB si dimensioni negli orari pianificati.

Amazon ECS e Application Auto Scaling

È possibile dimensionare i servizi ECS utilizzando le policy di dimensionamento con monitoraggio degli obiettivi, le policy di dimensionamento per fasi e il dimensionamento pianificato.

Utilizza le informazioni riportate di seguito per semplificare l'integrazione di Amazon ECS con Application Auto Scaling.

Ruolo collegato ai servizi creato per Amazon ECS

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione delle risorse Amazon ECS come destinazioni scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_ECSService`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `ecs.application-autoscaling.amazonaws.com`

Registrazione dei servizi ECS come obiettivi scalabili con Application Auto Scaling

Application Auto Scaling richiede un obiettivo scalabile prima di poter creare policy di dimensionamento o operazioni pianificate per un servizio Amazon ECS. Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Se configuri la scalabilità automatica usando la console Amazon ECS, Amazon ECS registra automaticamente un obiettivo scalabile per tuo conto.

Se desideri configurare la scalabilità automatica utilizzando la AWS CLI o uno degli SDK, puoi utilizzare AWS le seguenti opzioni:

- AWS CLI:

  Chiama il comando register-scalable-target per un servizio Amazon ECS. Nell'esempio seguente viene registrato un obiettivo scalabile per un servizio denominato `sample-app-service`, in esecuzione sul cluster `default`, con un numero minimo di attività di un'attività e un numero massimo di 10 attività.

```
aws application-autoscaling register-scalable-target \
  --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/default/sample-app-service \
  --min-capacity 1 \
  --max-capacity 10
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità delle tue risorse Amazon ECS nella seguente documentazione:

- Scalabilità automatica dei servizi nella Amazon Elastic Container Service Developer Guide
- Configurazione della scalabilità automatica del servizio nella Guida alle best practice di Amazon Elastic Container Service

> **Note**
>
> Per istruzioni su come sospendere i processi di scalabilità orizzontale mentre le distribuzioni di Amazon ECS sono in corso, consulta la seguente documentazione: Scalabilità e implementazioni automatiche dei servizi nella Amazon Elastic Container Service Developer Guide

ElastiCache per Redis e Application Auto Scaling

È possibile scalare ElastiCache per i gruppi di replica Redis utilizzando le politiche di scalabilità di Target Tracking e il ridimensionamento pianificato.

Utilizzate le seguenti informazioni per facilitare l'integrazione ElastiCache con Application Auto Scaling.

Ruolo collegato ai servizi creato per ElastiCache

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione ElastiCache delle risorse come destinazioni scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_ElastiCacheRG`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `elasticache.application-autoscaling.amazonaws.com`

Registrazione ElastiCache per i gruppi di replica Redis come destinazioni scalabili con Application Auto Scaling

Application Auto Scaling richiede una destinazione scalabile prima di poter creare policy di scalabilità o azioni pianificate per un gruppo di replica. ElastiCache Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Se configuri la scalabilità automatica utilizzando la ElastiCache console, registra ElastiCache automaticamente una destinazione scalabile per te.

Se desideri configurare la scalabilità automatica utilizzando la AWS CLI o uno degli SDK, puoi utilizzare AWS le seguenti opzioni:

- AWS CLI:

  Richiamate il register-scalable-targetcomando per un ElastiCache gruppo di replica. Nell'esempio seguente viene registrato il numero desiderato di gruppi di nodi per un gruppo di replica denominato `mycluster`, con una capacità minima di uno e una capacità massima di cinque.

```
aws application-autoscaling register-scalable-target \
   --service-namespace elasticache \
   --scalable-dimension elasticache:replication-group:NodeGroups \
   --resource-id replication-group/mycluster \
   --min-capacity 1 \
   --max-capacity 5
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Nell'esempio seguente viene registrato il numero desiderato di repliche per gruppo di nodi per un gruppo di replica denominato `mycluster`, con una capacità minima di 1 e una capacità massima di 5.

```
aws application-autoscaling register-scalable-target \
   --service-namespace elasticache \
   --scalable-dimension elasticache:replication-group:Replicas \
   --resource-id replication-group/mycluster \
   --min-capacity 1 \
   --max-capacity 5
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità ElastiCache delle tue risorse nella seguente documentazione:

Auto Scaling ElastiCache per cluster Redis nella Guida per l'utente di Amazon ElastiCache for Redis

Amazon Keyspaces (per Apache Cassandra) e Application Auto Scaling

È possibile dimensionare le tabelle Amazon Keyspaces utilizzando le policy di dimensionamento con monitoraggio degli obiettivi e il dimensionamento pianificato.

Utilizza le informazioni riportate di seguito per semplificare l'integrazione di Amazon Keyspaces con Application Auto Scaling.

Ruolo collegato ai servizi creato per Amazon Keyspaces

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione delle risorse Amazon Keyspaces come destinazioni scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_CassandraTable`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `cassandra.application-autoscaling.amazonaws.com`

Registrazione delle tabelle Amazon Keyspaces come obiettivi scalabili con Application Auto Scaling

Application Auto Scaling richiede un obiettivo scalabile prima di poter creare policy di dimensionamento o operazioni pianificate per una tabella Amazon Keyspaces. Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Se configuri la scalabilità automatica usando la console Amazon Keyspaces, Amazon Keyspaces registra automaticamente un obiettivo scalabile per tuo conto.

Se desideri configurare la scalabilità automatica utilizzando la AWS CLI o uno degli SDK, puoi utilizzare AWS le seguenti opzioni:

- AWS CLI:

  Chiama il comando register-scalable-target per una tabella Amazon Keyspaces. Nell'esempio seguente viene registrata la capacità in scrittura assegnata di una tabella denominata `mytable`, con una capacità minima di cinque unità di capacità in scrittura e una capacità massima di 10 unità di capacità in scrittura.

  ```
  aws application-autoscaling register-scalable-target \
    --service-namespace cassandra \
    --scalable-dimension cassandra:table:WriteCapacityUnits \
    --resource-id keyspace/mykeyspace/table/mytable \
    --min-capacity 5 \
    --max-capacity 10
  ```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

  ```
  {
      "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
  }
  ```

  Nell'esempio seguente viene registrata la capacità in lettura assegnata di una tabella denominata `mytable`, con una capacità minima di cinque unità di capacità in lettura e una capacità massima di 10 unità di capacità in lettura.

```
aws application-autoscaling register-scalable-target \
  --service-namespace cassandra \
  --scalable-dimension cassandra:table:ReadCapacityUnits \
  --resource-id keyspace/mykeyspace/table/mytable \
  --min-capacity 5 \
  --max-capacity 10
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità delle tue risorse Amazon Keyspaces nella seguente documentazione:

Gestione della capacità di throughput con la scalabilità automatica di Amazon Keyspaces nella Amazon Keyspaces (per Apache Cassandra) Developer Guide

AWS Lambda e Application Auto Scaling

È possibile scalare la concorrenza AWS Lambda fornita utilizzando le politiche di scalabilità di Target Tracking e la scalabilità pianificata.

Utilizza le informazioni riportate di seguito per semplificare l'integrazione di Lambda con Application Auto Scaling.

Ruolo collegato ai servizi creato per Lambda

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione delle risorse Lambda come destinazioni scalabili con Application Auto Scaling.

Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_LambdaConcurrency`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `lambda.application-autoscaling.amazonaws.com`

Registrazione delle funzioni Lambda come obiettivi scalabili con Application Auto Scaling

Application Auto Scaling richiede un obiettivo scalabile prima di poter creare policy di dimensionamento o operazioni pianificate per una funzione Lambda. Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Per configurare la scalabilità automatica utilizzando la AWS CLI o uno AWS degli SDK, puoi utilizzare le seguenti opzioni:

- AWS CLI:

  Chiama il comando register-scalable-target per una funzione Lambda. Nell'esempio seguente viene registrato il provisioning simultaneo per un alias chiamato `BLUE` per una funzione chiamata `my-function` con una capacità minima di 0 e una capacità massima di 100.

```
aws application-autoscaling register-scalable-target \
  --service-namespace lambda \
  --scalable-dimension lambda:function:ProvisionedConcurrency \
  --resource-id function:my-function:BLUE \
  --min-capacity 0 \
  --max-capacity 100
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità delle funzioni Lambda nella seguente documentazione:

- Configurazione della concorrenza fornita nella Developer Guide AWS Lambda
- Pianificazione di Lambda Provisioned Concurrency per i picchi di utilizzo ricorrenti sul blog AWS

Amazon Managed Streaming for Apache Kafka (MSK) e Application Auto Scaling

Puoi aumentare orizzontalmente l'archiviazione cluster Amazon MSK utilizzando le policy di dimensionamento con monitoraggio degli obiettivi. La riduzione orizzontale in base alla policy del monitoraggio degli obiettivi è disabilitata.

Utilizza le informazioni riportate di seguito per semplificare l'integrazione di Amazon MSK con Application Auto Scaling.

Ruolo collegato ai servizi creato per Amazon MSK

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione delle risorse Amazon MSK come destinazioni scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_KafkaCluster`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `kafka.application-autoscaling.amazonaws.com`

Registrazione dell'archiviazione cluster Amazon MSK come obiettivi scalabili con Application Auto Scaling

Application Auto Scaling richiede un obiettivo scalabile prima di poter creare una policy di dimensionamento per la dimensione del volume di archiviazione per broker di un cluster Amazon MSK. Un obiettivo scalabile è una risorsa la cui dimensione può essere dimensionta tramite Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Se configuri la scalabilità automatica usando la console Amazon MSK, Amazon MSK registra automaticamente un obiettivo scalabile per tuo conto.

Se desideri configurare la scalabilità automatica utilizzando la AWS CLI o uno degli SDK, puoi utilizzare AWS le seguenti opzioni:

- AWS CLI:

  Chiama il comando register-scalable-target per un cluster Amazon MSK. Nell'esempio seguente viene registrata la dimensione del volume di archiviazione per broker di un cluster Amazon MSK, con una capacità minima di 100 GiB e una capacità massima di 800 GiB.

```
aws application-autoscaling register-scalable-target \
   --service-namespace kafka \
   --scalable-dimension kafka:broker-storage:VolumeSize \
   --resource-id arn:aws:kafka:us-east-1:123456789012:cluster/demo-
cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5 \
   --min-capacity 100 \
   --max-capacity 800
```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

> **Note**
>
> Quando un cluster Amazon MSK è l'obiettivo scalabile, la riduzione orizzontale è disabilitata e non può essere abilitata.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità delle tue risorse Amazon MSK nella seguente documentazione:

Scalabilità automatica nella Amazon Managed Streaming for Apache Kafka Developer Guide

Amazon Neptune e Application Auto Scaling

È possibile dimensionare i cluster Neptune utilizzando le policy di dimensionamento con monitoraggio degli obiettivi e il dimensionamento pianificato.

Utilizza le informazioni riportate di seguito per semplificare l'integrazione di Neptune con Application Auto Scaling.

Ruolo collegato ai servizi creato per Neptune

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione delle risorse di Neptune come destinazioni scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_NeptuneCluster`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `neptune.application-autoscaling.amazonaws.com`

Registrazione di cluster Neptune come target scalabili con Application Auto Scaling

Application Auto Scaling richiede un target scalabile, prima di poter creare policy di dimensionamento o operazioni pianificate per un cluster Neptune. Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Per configurare la scalabilità automatica utilizzando la AWS CLI o uno AWS degli SDK, puoi utilizzare le seguenti opzioni:

- AWS CLI:

  Chiama il register-scalable-targetcomando per un cluster Neptune. Nell'esempio seguente viene registrata la capacità desiderata di un cluster denominato `mycluster`, con una capacità minima di un'istanza e una capacità massima di otto.

```
aws application-autoscaling register-scalable-target \
  --service-namespace neptune \
  --scalable-dimension neptune:cluster:ReadReplicaCount \
  --resource-id cluster:mycluster \
  --min-capacity 1 \
  --max-capacity 8
```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità delle risorse di Neptune nella seguente documentazione:

Scalabilità automatica del numero di repliche in un cluster di database Amazon Neptune nella Guida per l'utente di Neptune

Amazon SageMaker e Application Auto Scaling

Puoi scalare le varianti SageMaker degli endpoint, il provisioning della concorrenza per gli endpoint serverless e i componenti di inferenza utilizzando le policy di scalabilità di Target Tracking, le policy di scalabilità in fasi e la scalabilità pianificata.

Utilizzate le seguenti informazioni per facilitare l'integrazione SageMaker con Application Auto Scaling.

Ruolo collegato ai servizi creato per SageMaker

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione SageMaker delle risorse come destinazioni scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_SageMakerEndpoint`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `sagemaker.application-autoscaling.amazonaws.com`

Registrazione delle varianti SageMaker degli endpoint come destinazioni scalabili con Application Auto Scaling

Application Auto Scaling richiede un target scalabile prima di poter creare politiche di scalabilità o azioni pianificate per un SageMaker modello (variante). Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Se configuri la scalabilità automatica utilizzando la SageMaker console, registra SageMaker automaticamente una destinazione scalabile per te.

Se desideri configurare la scalabilità automatica utilizzando la AWS CLI o uno degli SDK, puoi utilizzare AWS le seguenti opzioni:

- AWS CLI:

  Richiama il register-scalable-targetcomando per una variante di prodotto. Nell'esempio seguente viene registrato il conteggio delle istanze desiderato per una variante prodotto denominata `my-variant`, in esecuzione sull'endpoint `my-endpoint`, con una capacità minima di un'istanza e una capacità massima di otto istanze.

  ```
  aws application-autoscaling register-scalable-target \
    --service-namespace sagemaker \
    --scalable-dimension sagemaker:variant:DesiredInstanceCount \
    --resource-id endpoint/my-endpoint/variant/my-variant \
    --min-capacity 1 \
    --max-capacity 8
  ```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

  ```
  {
      "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
  }
  ```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Registrazione del provisioning simultaneo degli endpoint serverless come obiettivi dimensionabili con Application Auto Scaling

Inoltre, Application Auto Scaling richiede un obiettivo dimensionabile prima di poter creare policy di dimensionamento oppure operazioni pianificate per il provisioning simultaneo degli endpoint serverless.

Se configuri la scalabilità automatica utilizzando la SageMaker console, registra SageMaker automaticamente una destinazione scalabile per te.

Altrimenti, utilizza uno dei seguenti metodi per registrare l'obiettivo dimensionabile:

- AWS CLI:

  Richiama il register-scalable-targetcomando per una variante di prodotto. Nell'esempio seguente viene registrato il provisioning simultaneo per una variante di prodotto denominata `my-variant`, in esecuzione sull'endpoint `my-endpoint`, con una capacità minima di uno e una capacità massima di dieci.

  ```
  aws application-autoscaling register-scalable-target \
    --service-namespace sagemaker \
    --scalable-dimension sagemaker:variant:DesiredProvisionedConcurrency \
    --resource-id endpoint/my-endpoint/variant/my-variant \
    --min-capacity 1 \
    --max-capacity 10
  ```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

  ```
  {
      "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
  target/1234abcd56ab78cd901ef1234567890ab123"
  }
  ```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Registrazione di componenti di inferenza come target scalabili con Application Auto Scaling

Application Auto Scaling richiede un target scalabile, prima di poter creare policy di dimensionamento o operazioni pianificate per i componenti di inferenza.

- AWS CLI:

  Chiama il register-scalable-targetcomando per un componente di inferenza. Nell'esempio seguente viene registrato il numero desiderato di copie per un componente di inferenza denominato `my-inference-component`, con una capacità minima di zero copie e una capacità massima di tre copie.

  ```
  aws application-autoscaling register-scalable-target \
    --service-namespace sagemaker \
    --scalable-dimension sagemaker:inference-component:DesiredCopyCount \
    --resource-id inference-component/my-inference-component \
    --min-capacity 0 \
    --max-capacity 3
  ```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

  ```
  {
      "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
  }
  ```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità SageMaker delle tue risorse nella Amazon SageMaker Developer Guide:

- Ridimensiona automaticamente SageMaker i modelli Amazon
- Scala automaticamente Provisioned Concurrency per un endpoint serverless

- Imposta politiche di scalabilità automatica per implementazioni di endpoint multimodello
- Scalabilità automatica di un endpoint asincrono

> **Note**
>
> Nel 2023, SageMaker ha introdotto nuove funzionalità di inferenza basate su endpoint di inferenza in tempo reale. Si crea un SageMaker endpoint con una configurazione dell'endpoint che definisce il tipo di istanza e il numero iniziale di istanze per l'endpoint. Quindi, crea un componente di inferenza, che è un oggetto di SageMaker hosting che puoi utilizzare per distribuire un modello su un endpoint. Per informazioni sulla scalabilità dei componenti di inferenza, consulta Amazon SageMaker aggiunge nuove funzionalità di inferenza per aiutare a ridurre i costi e la latenza di implementazione del modello di base e riduce i costi di implementazione del modello del 50% in media utilizzando le funzionalità più recenti di Amazon SageMaker sul blog. AWS

Serie di istanze Spot Amazon EC2 e Application Auto Scaling

È possibile dimensionare le serie di istanze Spot utilizzando le policy di dimensionamento con monitoraggio degli obiettivi, le policy di dimensionamento per fasi e il dimensionamento pianificato.

Utilizza le informazioni riportate di seguito per semplificare l'integrazione di serie di istanze Spot con Application Auto Scaling.

Ruolo collegato ai servizi creato per la serie di istanze Spot

Il seguente ruolo collegato ai servizi viene creato automaticamente al Account AWS momento della registrazione delle risorse Spot Fleet come obiettivi scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `ec2.application-autoscaling.amazonaws.com`

Registrazione della serie di istanze Spot come obiettivi scalabili tramite Application Auto Scaling

Application Auto Scaling richiede un obiettivo scalabile prima di poter creare policy di dimensionamento o operazioni pianificate per una serie di istanze Spot. Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Se configuri la scalabilità automatica utilizzando la console serie di istanze Spot, la serie di istanze Spot registra automaticamente un obiettivo scalabile per tuo conto.

Se desideri configurare la scalabilità automatica utilizzando la AWS CLI o uno degli SDK, puoi utilizzare AWS le seguenti opzioni:

- AWS CLI:

  Chiama il comando register-scalable-target per una serie di istanze Spot. Nell'esempio seguente viene registrata la capacità obiettivo di una serie di istanze Spot utilizzando il relativo ID richiesta, con una capacità minima di due istanze e una capacità massima di 10 istanze.

```
aws application-autoscaling register-scalable-target \
  --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
  --min-capacity 2 \
  --max-capacity 10
```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
```

```
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità della tua flotta Spot nella seguente documentazione:

Scalabilità automatica per serie di istanze Spot nella Guida per l'utente di Amazon EC2

Risorse personalizzate e Application Auto Scaling

È possibile dimensionare risorse personalizzate utilizzando le policy di dimensionamento con monitoraggio degli obiettivi, le policy di dimensionamento per fasi e il dimensionamento pianificato.

Utilizza le informazioni riportate di seguito per semplificare l'integrazione delle risorse personalizzate con Application Auto Scaling.

Ruolo collegato ai servizi creato per le risorse personalizzate

Il seguente ruolo collegato ai servizi viene creato automaticamente Account AWS quando si registrano risorse personalizzate come destinazioni scalabili con Application Auto Scaling. Questo ruolo consente ad Application Auto Scaling di eseguire le operazioni supportate all'interno dell'account. Per ulteriori informazioni, consulta Ruoli collegati ai servizi per Application Auto Scaling.

- `AWSServiceRoleForApplicationAutoScaling_CustomResource`

Principale del servizio utilizzato dal ruolo collegato ai servizi

Il ruolo collegato ai servizi nella sezione precedente può essere assunto solo dal principale del servizio autorizzato dalle relazioni di attendibilità definite per il ruolo. Il ruolo collegato ai servizi utilizzato da Application Auto Scaling concede l'accesso al seguente principale del servizio:

- `custom-resource.application-autoscaling.amazonaws.com`

Registrazione delle risorse personalizzate come obiettivi scalabili con Application Auto Scaling

Application Auto Scaling richiede un obiettivo scalabile prima di poter creare policy di dimensionamento o operazioni pianificate per una risorsa personalizzata. Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling. Gli obiettivi scalabili sono identificati in modo univoco dalla combinazione di ID risorsa, dimensione scalabile e spazio dei nomi.

Per configurare la scalabilità automatica utilizzando la AWS CLI o uno AWS degli SDK, puoi utilizzare le seguenti opzioni:

- AWS CLI:

  Chiama il comando register-scalable-target per una risorsa personalizzata. Nell'esempio seguente viene registrata una risorsa personalizzata come obiettivo scalabile, con un conteggio minimo desiderato di un'unità di capacità e un conteggio massimo desiderato di 10 unità di capacità. Il file `custom-resource-id.txt` contiene una stringa che identifichi l'ID della risorsa, che rappresenta il percorso della risorsa personalizzata tramite l'endpoint Amazon API Gateway.

  ```
  aws application-autoscaling register-scalable-target \
    --service-namespace custom-resource \
    --scalable-dimension custom-resource:ResourceType:Property \
    --resource-id file://~/custom-resource-id.txt \
    --min-capacity 1 \
    --max-capacity 10
  ```

  Contenuto di `custom-resource-id.txt`.

  ```
  https://example.execute-api.us-west-2.amazonaws.com/prod/
  scalableTargetDimensions/1-23456789
  ```

  In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

  ```
  {
      "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
  target/1234abcd56ab78cd901ef1234567890ab123"
  ```

```
}
```

- AWS SDK:

  Chiama l'operazione RegisterScalableTarget e fornisci `ResourceId`, `ScalableDimension`, `ServiceNamespace`, `MinCapacity` e `MaxCapacity` come parametri.

Risorse correlate

Se hai appena iniziato a usare Application Auto Scaling, puoi trovare ulteriori informazioni utili sulla scalabilità delle tue risorse personalizzate nella seguente documentazione:

GitHubdeposito

Configurare le risorse di Application Auto Scaling utilizzando AWS CloudFormation

Application Auto Scaling è integrato con AWS CloudFormation un servizio che consente di modellare e configurare le AWS risorse in modo da dedicare meno tempo alla creazione e alla gestione delle risorse e dell'infrastruttura. Crei un modello che descrive tutte le AWS risorse che desideri e fornisce e AWS CloudFormation configura tali risorse per te.

Quando lo utilizzi AWS CloudFormation, puoi riutilizzare il modello per configurare le risorse Application Auto Scaling in modo coerente e ripetuto. Descrivi le tue risorse una sola volta, quindi fornisci le stesse risorse più e più volte in più regioni Account AWS .

Application Auto Scaling e modelli AWS CloudFormation

Per eseguire il provisioning e la configurazione delle risorse per Application Auto Scaling e i servizi correlati, devi conoscere i modelli AWS CloudFormation. I modelli sono file di testo formattati in JSON o YAML. Questi modelli descrivono le risorse che desideri inserire negli AWS CloudFormation stack. Se non conosci JSON o YAML, puoi usare AWS CloudFormation Designer per iniziare a usare i modelli. AWS CloudFormation Per ulteriori informazioni, consulta Che cos'è AWS CloudFormation Designer? nella Guida per l'utente di AWS CloudFormation .

Quando crei un modello di stack per le risorse Application Auto Scaling, è necessario fornire quanto segue:

- Uno spazio dei nomi per il servizio obiettivo (ad esempio, **`appstream`**). Consultate il AWS::ApplicationAutoScaling::ScalableTargetriferimento per ottenere i namespace dei servizi.
- Una dimensione scalabile associata alla risorsa obiettivo (ad esempio, **`appstream:fleet:DesiredCapacity`**). Vedi il AWS::ApplicationAutoScaling::ScalableTargetriferimento per ottenere dimensioni scalabili.
- Un ID risorsa per la risorsa obiettivo (ad esempio, **`fleet/sample-fleet`**). Vedi il AWS::ApplicationAutoScaling::ScalableTargetriferimento per informazioni sulla sintassi ed esempi di ID di risorse specifici.
- Un ruolo collegato ai servizi per la risorsa obiettivo (ad esempio **`arn:aws:iam::012345678910:role/aws-service-role/appstream.application-autoscaling.amazonaws.com/`**

`AWSServiceRoleForApplicationAutoScaling_AppStreamFleet`). Consulta la tabella Riferimento ARN del ruolo collegato ai servizi per ottenere gli ARN del ruolo.

Per ulteriori informazioni sulle risorse di Application Auto Scaling, consulta il riferimento Application Auto Scaling nella Guida per l'utente di AWS CloudFormation .

Frammenti di modello di esempio

Puoi trovare frammenti di esempio da includere nei AWS CloudFormation modelli nelle seguenti sezioni della Guida per l'AWS CloudFormation utente:

- Per esempi di policy di scalabilità e azioni pianificate, consulta Configurare le risorse di Application Auto Scaling con. AWS CloudFormation
- Per altri esempi di politiche di scalabilità, vedere. AWS::ApplicationAutoScaling::ScalingPolicy

Scopri di più su AWS CloudFormation

Per ulteriori informazioni AWS CloudFormation, consulta le seguenti risorse:

- AWS CloudFormation
- AWS CloudFormation Guida per l'utente
- AWS CloudFormation Documentazione di riferimento delle API
- Guida per l'utente dell'interfaccia a riga di comando di AWS CloudFormation

Dimensionamento pianificato per Application Auto Scaling

Con il dimensionamento pianificato, puoi impostare il dimensionamento automatico per la tua applicazione in base a variazioni di carico prevedibili, creando azioni pianificate che aumentano o diminuiscono la capacità in momenti specifici. Ciò consente di dimensionare l'applicazione in modo proattivo per far fronte alle variazioni di carico prevedibili.

Ad esempio, supponiamo che si verifichi uno schema di traffico settimanale regolare in cui il carico aumenta a metà settimana e diminuisce verso la fine. È possibile configurare una pianificazione di dimensionamento in Application Auto Scaling che si allinei a questo schema:

- Mercoledì mattina, un'azione pianificata aumenta la capacità aumentando la capacità minima precedentemente impostata del target scalabile.
- Venerdì sera, un'altra azione pianificata riduce la capacità diminuendo la capacità massima precedentemente impostata del target scalabile.

Queste azioni di dimensionamento pianificate consentono di ottimizzare costi e prestazioni. L'applicazione dispone di una capacità sufficiente per gestire il picco di traffico infrasettimanale ma in altri momenti non fornisce una capacità non necessaria eccedente.

È possibile utilizzare simultaneamente il dimensionamento pianificato e le policy di dimensionamento per ottenere i vantaggi degli approccia proattivi e reattivi al dimensionamento. Dopo l'esecuzione di un'operazione pianificata di dimensionamento, la policy di dimensionamento può continuare a prendere decisioni sull'opportunità di dimensionare ulteriormente la capacità. In questo modo è possibile garantire di disporre di capacità sufficiente per la gestione dei carichi dell'applicazione. Sebbene l'applicazione si dimensioni per soddisfare la domanda, la capacità corrente deve rientrare nei valori di capacità minima e massima impostati dall'operazione pianificata.

Indice

Come funziona la scalabilità pianificata per Application Auto Scaling

Questo argomento descrive come funziona il ridimensionamento programmato e introduce le considerazioni chiave da comprendere per utilizzarlo in modo efficace.

Indice

Come funziona

Per utilizzare il dimensionamento pianificato, è possibile creare operazioni pianificate che indicano ad Application Auto Scaling di eseguire attività di dimensionamento a orari specifici. Quando crei un'operazione pianificata, specifichi l'obiettivo scalabile, il momento in cui l'attività di dimensionamento dovrebbe verificarsi, la capacità minima e la capacità massima. È possibile creare operazioni pianificate sia una tantum che ricorrenti.

All'ora specificata, Application Auto Scaling dimensiona in base ai nuovi valori di capacità, confrontando la capacità attuale con la capacità minima e massima specificate.

- Se l'attuale capacità è inferiore alla capacità minima specificata, Application Auto Scaling aumenta orizzontalmente la capacità fino alla capacità minima specificata.
- Se l'attuale capacità è superiore alla capacità massima specificata, Application Auto Scaling riduce orizzontalmente la capacità fino alla capacità massima specificata.

Considerazioni

Quando crei un'operazione pianificata, tieni presente quanto segue:

- Un'operazione pianificata imposta `MinCapacity` e `MaxCapacity` in base a quanto specificato dall'operazione pianificata alla data e all'orario specificati. La richiesta può includere facoltativamente anche solo uno di questi dimensionamenti. Ad esempio, è possibile creare

un'operazione pianificata con solo la capacità minima specificata. In alcuni casi, tuttavia, è necessario includere entrambe le dimensioni, per garantire che la nuova capacità minima non sia superiore alla capacità massima, oppure che la nuova capacità massima non sia inferiore alla capacità minima.

- Per impostazione predefinita, le pianificazioni ricorrenti impostate sono in formato UTC. Puoi modificare l'orario affinché corrisponda al fuso orario locale o a un fuso orario di un'altra parte della rete. Quando specifichi un fuso orario che osserva l'ora legale, l'operazione si adegua automaticamente in funzione dell'ora legale. Per ulteriori informazioni, consulta Pianifica operazioni di dimensionamento ricorrenti utilizzando espressioni cron.
- Puoi temporaneamente disattivare il dimensionamento pianificato per un obiettivo scalabile. Ciò permette di impedire che le operazioni pianificate siano attive senza doverle eliminare. Sarà poi quindi possibile riprendere il dimensionamento programmato quando si vorrà utilizzarlo nuovamente. Per ulteriori informazioni, consulta Sospendi e riprendi il dimensionamento per Application Auto Scaling.
- L'ordine di esecuzione delle operazioni pianificate è garantito per lo stesso obiettivo scalabile, ma non per le operazioni pianificate tra obiettivi scalabili.
- Per completare correttamente un'operazione pianificata, la risorsa specificata deve trovarsi in uno stato scalabile nel servizio dell'obiettivo. In caso contrario, la richiesta fallisce e restituisce un messaggio di errore, ad esempio `Resource Id [ActualResourceId] is not scalable. Reason: The status of all DB instances must be 'available' or 'incompatible-parameters'`.
- A causa della natura distribuita di Application Auto Scaling e dei servizi obiettivo, l'intervallo tra il momento in cui l'operazione pianificata viene attivata e il momento in cui il servizio obiettivo esegue l'operazione di dimensionamento potrebbe ammontare ad alcuni secondi. Poiché le operazioni pianificate vengono eseguite nell'ordine in cui sono specificate, l'esecuzione delle operazioni pianificate con orari di inizio ravvicinati può richiedere più tempo.

Comandi comunemente utilizzati per la creazione, la gestione e l'eliminazione delle operazioni pianificate

I comandi comunemente utilizzati per il dimensionamento pianificato includono:

- register-scalable-target per registrare AWS o personalizzare le risorse come destinazioni scalabili (una risorsa scalabile da Application Auto Scaling) e per sospendere e riprendere il ridimensionamento.

- put-scheduled-action per aggiungere o modificare le operazioni pianificate per un obiettivo scalabile esistente.
- describe-scaling-activities per restituire informazioni sulle attività di scalabilità in una regione. AWS
- describe-scheduled-actions per restituire informazioni sulle azioni pianificate in una regione. AWS
- delete-scheduled-action per eliminare un'operazione pianificata.

Risorse correlate

Per un esempio dettagliato dell'utilizzo della scalabilità pianificata, consulta il post del blog Scheduling AWS Lambda Provisioned Concurrency for recurring Peak Usage on the Compute Blog. AWS

Per un tutorial dettagliato su come creare operazioni pianificate tramite risorse AWS di esempio, consulta Tutorial: Guida di base su dimensionamento pianificato utilizzando AWS CLI.

Per ulteriori informazioni su come creare azioni programmate per i gruppi con dimensionamento automatico, consulta Dimensionamento programmato per Dimensionamento automatico Amazon EC2 nella Guida per l'utente di Dimensionamento automatico Amazon EC2.

Limitazioni

Di seguito sono riportate le limitazioni quando si utilizza il dimensionamento pianificato:

- I nomi delle operazioni pianificate devono essere univoci per ciascun obiettivo scalabile.
- Application Auto Scaling non fornisce precisione di secondo livello nelle espressioni di pianificazione. La risoluzione più alta che utilizza un'espressione cron è un minuto.
- L'obiettivo scalabile non può essere un cluster Amazon MSK. Il dimensionamento pianificato non è supportato per Amazon MSK.
- L'accesso da console per visualizzare, aggiungere, aggiornare o rimuovere azioni pianificate su risorse scalabili dipende dalla risorsa utilizzata. Per ulteriori informazioni, consulta Servizi AWS che puoi usare con Application Auto Scaling.

Pianifica operazioni di dimensionamento ricorrenti utilizzando espressioni cron

> **Important**
>
> Per un supporto con le espressioni cron per il dimensionamento automatico Amazon EC2, consulta l'argomento Pianificazioni ricorrenti nella Guida per l'utente del Dimensionamento automatico Amazon EC2. Con il Dimensionamento automatico Amazon EC2, usi la sintassi cron tradizionale anziché la sintassi cron personalizzata utilizzata da Application Auto Scaling.

Puoi creare un'operazione pianificata con una pianificazione ricorrente utilizzando un'espressione cron.

Per creare una pianificazione ricorrente specifica un'espressione cron e un fuso orario per descrivere a quali intervalli l'operazione pianificata deve ripetersi. I valori del fuso orario supportati sono i nomi canonici dei fusi orari IANA supportati da Joda-Time (come ad esempio `Etc/GMT+9` o `Pacific/Tahiti`). È possibile specificare una data e un'ora per l'ora di avvio, per l'ora di fine o per entrambi i campi. Per un comando di esempio che utilizza il per creare un'azione pianificata, AWS CLI vedi. Creazione di un'operazione pianificata ricorrente che specifica un fuso orario

Il formato dell'espressione cron supportato è costituito da cinque campi separati da spazi: [Minutes] [Hours] [Day_of_Month] [Month] [Day_of_Week] [Year]. Ad esempio, l'espressione cron `30 6 ? * MON *`, configura un'operazione pianificata che ricorre ogni lunedì alle 6:30. L'asterisco viene utilizzato come carattere jolly per abbinare tutti i valori di un campo.

Per ulteriori informazioni sulla sintassi cron per le azioni pianificate di Application Auto Scaling, consulta il riferimento alle espressioni Cron nella Amazon User Guide. EventBridge

Quando crei una pianificazione ricorrente, scegli con attenzione l'ora di inizio e di fine. Ricorda quanto segue:

- Se si specifica un'ora di inizio, Application Auto Scaling esegue l'operazione a quell'ora, quindi esegue l'operazione in base alla pianificazione ricorrente.
- Se si specifica un'ora di fine, l'iterazione dell'operazione si interrompe dopo tale orario. Application Auto Scaling non tiene traccia dei valori precedenti e torna a tali valori precedenti dopo l'ora di fine.

- L'ora di inizio e l'ora di fine devono essere impostate in UTC quando utilizzi gli SDK AWS CLI o gli AWS SDK per creare o aggiornare un'azione pianificata.

Esempi

Puoi fare riferimento alla tabella seguente quando crei una pianificazione ricorrente per un obiettivo scalabile Application Auto Scaling. Gli esempi riportati di seguito sono la sintassi corretta per l'utilizzo di Application Auto Scaling per creare o aggiornare un'operazione pianificata.

Minuti	Ore	Giorno del mese	Mese	Giorno della settimana	Anno	Significato
0	10	*	*	?	*	Esegui ogni giorno alle 10:00 (UTC)
15	12	*	*	?	*	Esegui ogni giorno alle 12:15 (UTC)
0	18	?	*	LUN-VEN	*	Esegui dal lunedì al venerdì alle 18:00 (UTC)
0	8	1	*	?	*	Esegui ogni primo giorno del mese alle 8:00 (UTC)
0/15	*	*	*	?	*	Esegui ogni 15 minuti

Minuti	Ore	Giorno del mese	Mese	Giorno della settimana	Anno	Significato
0/10	*	?	*	LUN-VEN	*	Esegui dal lunedì al venerdì ogni 10 minuti
0/5	8-17	?	*	LUN-VEN	*	Esegui dal lunedì al venerdì dalle 8:00 alle 17:55 (UTC) ogni 5 minuti

Eccezione

È inoltre possibile creare un'espressione cron con un valore stringa che contiene sette campi. In questo caso, è possibile utilizzare i primi tre campi per specificare l'ora di esecuzione di un'operazione pianificata, inclusi i secondi. L'espressione cron completa ha i seguenti campi separati dallo spazio: [Seconds] [Minutes] [Hours] [Day_of_Month] [Month] [Day_of_Week] [Year]. Tuttavia, questo approccio non garantisce che l'operazione pianificata venga eseguita nel secondo preciso specificato. Inoltre, alcune console di servizio potrebbero non supportare il campo secondi in un'espressione cron.

Operazioni pianificate di esempio per Application Auto Scaling

Gli esempi seguenti mostrano come creare azioni pianificate con il comando AWS CLI put-scheduled-action. Quando si specifica la nuova capacità, è possibile indicare il valore minimo, quello massimo o entrambi.

Per brevità, gli esempi di questo argomento illustrano i comandi della CLI per alcuni dei servizi che si integrano con Application Auto Scaling. Per specificare un altro target scalabile, specificare il suo spazio dei nomi in `--service-namespace`, la sua dimensione scalabile `--scalable-`

`dimension` e l'ID di risorsa in `--resource-id`. Per maggiori informazioni ed esempi per ogni servizio, consultare gli argomenti in Servizi AWS che puoi usare con Application Auto Scaling.

Quando usi il AWS CLI, ricorda che i comandi vengono eseguiti nella Regione AWS configurazione per il tuo profilo. Per eseguire i comandi in un'altra regione, modificare la regione predefinita per il profilo oppure utilizzare il parametro `--region` con il comando.

Indice

Creazione di un'operazione pianificata che si verifica una sola volta

Per dimensionare automaticamente l'obiettivo scalabile una sola volta, a una data e un'ora specificate, utilizza l'opzione `--schedule "at(yyyy-mm-ddThh:mm:ss)"`.

Example Esempio: aumento orizzontale una tantum

Di seguito è riportato un esempio di creazione di un'operazione pianificata per aumentare orizzontalmente la capacità in una data e a un'ora specifiche.

Alla data e all'ora specificate per `--schedule` (22:00 UTC del 31 marzo 2021), se il valore indicato per `MinCapacity` supera la capacità corrente, Application Auto Scaling aumenta orizzontalmente fino a `MinCapacity`.

Linux, macOS o Unix

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource \
  --scalable-dimension custom-resource:ResourceType:Property \
  --resource-id file://~/custom-resource-id.txt \
  --scheduled-action-name scale-out \
  --schedule "at(2021-03-31T22:00:00)" \
  --scalable-target-action MinCapacity=3
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource --
scalable-dimension custom-resource:ResourceType:Property --resource-id file://~/custom-
resource-id.txt --scheduled-action-name scale-out --schedule "at(2021-03-31T22:00:00)"
 --scalable-target-action MinCapacity=3
```

> **Note**
>
> Quando viene eseguita questa operazione pianificata, se la capacità massima è inferiore al valore specificato per la capacità minima, è necessario specificare una nuova capacità minima e massima e non solo la capacità minima.

Example Esempio: riduzione orizzontale una tantum

Di seguito è riportato un esempio di creazione di un'operazione pianificata per ridurre orizzontalmente la capacità in una data e a un'ora specifiche.

Alla data e all'ora specificate per `--schedule` (22:30 UTC del 31 marzo 2021), se il valore indicato per `MaxCapacity` supera la capacità corrente, Application Auto Scaling riduce orizzontalmente fino a `MaxCapacity`.

Linux, macOS o Unix

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource \
  --scalable-dimension custom-resource:ResourceType:Property \
  --resource-id file://~/custom-resource-id.txt \
  --scheduled-action-name scale-in \
  --schedule "at(2021-03-31T22:30:00)" \
  --scalable-target-action MinCapacity=0,MaxCapacity=0
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource --
scalable-dimension custom-resource:ResourceType:Property --resource-id file://~/custom-
resource-id.txt --scheduled-action-name scale-in --schedule "at(2021-03-31T22:30:00)"
 --scalable-target-action MinCapacity=0,MaxCapacity=0
```

Crea un'operazione pianificata eseguita a intervalli ricorrenti

Per pianificare il dimensionamento a intervalli ricorrenti, utilizza l'opzione `--schedule "rate(value unit)"`. Il valore deve essere un numero intero positivo. L'unità può essere `minute`, `minutes`, `hour`, `hours`, `day` oppure `days`. Per ulteriori informazioni, consulta Rate expression nella Amazon EventBridge User Guide.

Di seguito è riportato un esempio di un'operazione pianificata che utilizza un'espressione rate.

Secondo la pianificazione specificata (ogni 5 ore a partire dal 30 gennaio 2021 alle 12:00 UTC fino al 31 gennaio 2021 alle 22:00 UTC), se il valore specificato per `MinCapacity` è superiore alla capacità attuale, Application Auto Scaling aumenta orizzontalmente fino a `MinCapacity`. Se il valore specificato per `MaxCapacity` è inferiore all'attuale capacità, Application Auto Scaling riduce orizzontalmente a `MaxCapacity`.

Linux, macOS o Unix

```
aws application-autoscaling put-scheduled-action --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --scheduled-action-name my-recurring-action \
  --schedule "rate(5 hours)" \
  --start-time 2021-01-30T12:00:00 \
  --end-time 2021-01-31T22:00:00 \
  --scalable-target-action MinCapacity=3,MaxCapacity=10
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace ecs --scalable-
dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
 --scheduled-action-name my-recurring-action --schedule "rate(5 hours)" --start-
time 2021-01-30T12:00:00 --end-time 2021-01-31T22:00:00 --scalable-target-action
 MinCapacity=3,MaxCapacity=10
```

Creazione di un'operazione pianificata eseguita in base a una pianificazione periodica

Per pianificare il dimensionamento in base a una pianificazione ricorrente, usa l'opzione `--schedule "cron(fields)"`. Per ulteriori informazioni, consulta Pianifica operazioni di dimensionamento ricorrenti utilizzando espressioni cron.

Di seguito è riportato un esempio di un'operazione pianificata che utilizza un'espressione cron.

Alla pianificazione specificata (ogni giorno alle 9:00 UTC), se il valore indicato per `MinCapacity` è superiore alla capacità attuale, Application Auto Scaling aumenta orizzontalmente fino a `MinCapacity`. Se il valore specificato per `MaxCapacity` è inferiore all'attuale capacità, Application Auto Scaling riduce orizzontalmente a `MaxCapacity`.

Linux, macOS o Unix

```
aws application-autoscaling put-scheduled-action --service-namespace appstream \
  --scalable-dimension appstream:fleet:DesiredCapacity \
  --resource-id fleet/sample-fleet \
  --scheduled-action-name my-recurring-action \
  --schedule "cron(0 9 * * ? *)" \
  --scalable-target-action MinCapacity=10,MaxCapacity=50
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace appstream --
scalable-dimension appstream:fleet:DesiredCapacity --resource-id fleet/sample-fleet --
scheduled-action-name my-recurring-action --schedule "cron(0 9 * * ? *)" --scalable-
target-action MinCapacity=10,MaxCapacity=50
```

Crea un'operazione pianificata occasionale che specifica un fuso orario

Le operazioni pianificate vengono impostate sul fuso orario UTC per impostazione predefinita. Per specificare un fuso orario diverso, includi l'opzione `--timezone` e specifica il nome canonico per il fuso orario (per esempio `America/New_York`). Per ulteriori informazioni, consulta https://www.joda.org/joda-time/timezones.html, che fornisce informazioni sui fusi orari IANA supportati quando si chiama put-scheduled-action.

Di seguito è riportato un esempio di utilizzo dell'opzione `--timezone` durante la creazione di un'operazione pianificata per dimensionare la capacità in una data e a un'ora specifiche.

Alla data e all'ora specificate per `--schedule` (17:00 ora locale del 31 gennaio 2021), se il valore indicato per `MinCapacity` supera la capacità corrente, Application Auto Scaling aumenta orizzontalmente fino a `MinCapacity`. Se il valore specificato per `MaxCapacity` è inferiore all'attuale capacità, Application Auto Scaling riduce orizzontalmente a `MaxCapacity`.

Linux, macOS o Unix

```
aws application-autoscaling put-scheduled-action --service-namespace comprehend \
  --scalable-dimension comprehend:document-classifier-endpoint:DesiredInferenceUnits \
  --resource-id arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/
EXAMPLE \
  --scheduled-action-name  my-one-time-action \
  --schedule "at(2021-01-31T17:00:00)" --timezone "America/New_York" \
  --scalable-target-action MinCapacity=1,MaxCapacity=3
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace comprehend --
scalable-dimension comprehend:document-classifier-endpoint:DesiredInferenceUnits
 --resource-id arn:aws:comprehend:us-west-2:123456789012:document-classifier-
endpoint/EXAMPLE --scheduled-action-name  my-one-time-action --schedule
 "at(2021-01-31T17:00:00)" --timezone "America/New_York" --scalable-target-action
 MinCapacity=1,MaxCapacity=3
```

Creazione di un'operazione pianificata ricorrente che specifica un fuso orario

Di seguito è riportato un esempio di utilizzo del `--timezone` quando viene creata un'operazione pianificata ricorrente per dimensionare la capacità. Per ulteriori informazioni, consulta Pianifica operazioni di dimensionamento ricorrenti utilizzando espressioni cron.

Alla pianificazione specificata (ogni giorno da lunedì a venerdì alle 18:00 ora locale), se il valore indicato per `MinCapacity` è superiore alla capacità attuale, Application Auto Scaling aumenta orizzontalmente fino a `MinCapacity`. Se il valore specificato per `MaxCapacity` è inferiore all'attuale capacità, Application Auto Scaling riduce orizzontalmente a `MaxCapacity`.

Linux, macOS o Unix

```
aws application-autoscaling put-scheduled-action --service-namespace lambda \
  --scalable-dimension lambda:function:ProvisionedConcurrency \
  --resource-id function:my-function:BLUE \
  --scheduled-action-name my-recurring-action \
  --schedule "cron(0 18 ? * MON-FRI *)" --timezone "Etc/GMT+9" \
  --scalable-target-action MinCapacity=10,MaxCapacity=50
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace lambda
 --scalable-dimension lambda:function:ProvisionedConcurrency --resource-
id function:my-function:BLUE --scheduled-action-name my-recurring-action --schedule
 "cron(0 18 ? * MON-FRI *)" --timezone "Etc/GMT+9" --scalable-target-action
 MinCapacity=10,MaxCapacity=50
```

Gestisci il dimensionamento pianificato per Application Auto Scaling

AWS CLI Include diversi altri comandi che ti aiutano a gestire le azioni pianificate.

Per brevità, gli esempi di questo argomento illustrano i comandi della CLI per alcuni dei servizi che si integrano con Application Auto Scaling. Per specificare un altro target scalabile, specificare il suo spazio dei nomi in `--service-namespace`, la sua dimensione scalabile `--scalable-dimension` e l'ID di risorsa in `--resource-id`. Per maggiori informazioni ed esempi per ogni servizio, consultare gli argomenti in Servizi AWS che puoi usare con Application Auto Scaling.

Quando usi il AWS CLI, ricorda che i comandi vengono eseguiti nella Regione AWS configurazione per il tuo profilo. Per eseguire i comandi in un'altra regione, modificare la regione predefinita per il profilo oppure utilizzare il parametro `--region` con il comando.

Indice

Visualizza le attività di dimensionamento per un servizio specificato

Per visualizzare le attività di dimensionamento per tutti gli obiettivi scalabili in uno spazio dei nomi del servizio specificato, utilizza il comando describe-scaling-activities.

Nell'esempio seguente vengono recuperate le attività di dimensionamento associate allo spazio dei nomi del servizio `dynamodb`.

Linux, macOS o Unix

```
aws application-autoscaling describe-scaling-activities --service-namespace dynamodb
```

Windows

```
aws application-autoscaling describe-scaling-activities --service-namespace dynamodb
```

Se il comando viene eseguito correttamente, verrà visualizzato un output simile al seguente.

```
{
    "ScalingActivities": [
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting write capacity units to 10.",
            "ResourceId": "table/my-table",
            "ActivityId": "4d1308c0-bbcf-4514-a673-b0220ae38547",
            "StartTime": 1561574415.086,
            "ServiceNamespace": "dynamodb",
            "EndTime": 1561574449.51,
            "Cause": "maximum capacity was set to 10",
            "StatusMessage": "Successfully set write capacity units to 10. Change
 successfully fulfilled by dynamodb.",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting min capacity to 5 and max capacity to 10",
            "ResourceId": "table/my-table",
            "ActivityId": "f2b7847b-721d-4e01-8ef0-0c8d3bacc1c7",
            "StartTime": 1561574414.644,
            "ServiceNamespace": "dynamodb",
            "Cause": "scheduled action name my-second-scheduled-action was triggered",
            "StatusMessage": "Successfully set min capacity to 5 and max capacity to
 10",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting write capacity units to 15.",
            "ResourceId": "table/my-table",
            "ActivityId": "d8ea4de6-9eaa-499f-b466-2cc5e681ba8b",
            "StartTime": 1561574108.904,
            "ServiceNamespace": "dynamodb",
```

```
            "EndTime": 1561574140.255,
            "Cause": "minimum capacity was set to 15",
            "StatusMessage": "Successfully set write capacity units to 15. Change
 successfully fulfilled by dynamodb.",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting min capacity to 15 and max capacity to 20",
            "ResourceId": "table/my-table",
            "ActivityId": "3250fd06-6940-4e8e-bb1f-d494db7554d2",
            "StartTime": 1561574108.512,
            "ServiceNamespace": "dynamodb",
            "Cause": "scheduled action name my-first-scheduled-action was triggered",
            "StatusMessage": "Successfully set min capacity to 15 and max capacity to
 20",
            "StatusCode": "Successful"
        }
    ]
}
```

Per modificare questo comando in modo che recuperi le attività di dimensionamento per una sola degli obiettivi scalabili, aggiungi l'opzione `--resource-id`.

Descrizione di tutte le operazioni pianificate per un servizio specificato

È possibile descrivere tutte le operazioni pianificate per gli obiettivi scalabili in uno spazio dei nomi del servizio specificato, utilizza il comando describe-scheduled-actions.

Nell'esempio seguente vengono recuperate le operazioni pianificate associate allo spazio dei nomi del servizio `ec2`.

Linux, macOS o Unix

```
aws application-autoscaling describe-scheduled-actions --service-namespace ec2
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-namespace ec2
```

Se il comando viene eseguito correttamente, verrà visualizzato un output simile al seguente.

```
{
    "ScheduledActions": [
        {
            "ScheduledActionName": "my-one-time-action",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:493a6261-fbb9-432d-855d-3c302c14bdb9:resource/ec2/
spot-fleet-request/sfr-107dc873-0802-4402-a901-37294EXAMPLE:scheduledActionName/my-one-
time-action",
            "ServiceNamespace": "ec2",
            "Schedule": "at(2021-01-31T17:00:00)",
            "Timezone": "America/New_York",
            "ResourceId": "spot-fleet-request/sfr-107dc873-0802-4402-
a901-37294EXAMPLE",
            "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
            "ScalableTargetAction": {
                "MaxCapacity": 1
            },
            "CreationTime": 1607454792.331
        },
        {
            "ScheduledActionName": "my-recurring-action",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:493a6261-fbb9-432d-855d-3c302c14bdb9:resource/ec2/
spot-fleet-request/sfr-107dc873-0802-4402-a901-37294EXAMPLE:scheduledActionName/my-
recurring-action",
            "ServiceNamespace": "ec2",
            "Schedule": "rate(5 minutes)",
            "ResourceId": "spot-fleet-request/sfr-107dc873-0802-4402-
a901-37294EXAMPLE",
            "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
            "StartTime": 1604059200.0,
            "EndTime": 1612130400.0,
            "ScalableTargetAction": {
                "MinCapacity": 3,
                "MaxCapacity": 10
            },
            "CreationTime": 1607454949.719
        },
        {
            "ScheduledActionName": "my-one-time-action",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:4bce34c7-bb81-4ecf-b776-5c726efb1567:resource/ec2/
```

```
spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE:scheduledActionName/my-one-
time-action",
            "ServiceNamespace": "ec2",
            "Schedule": "at(2020-12-08T9:36:00)",
            "Timezone": "America/New_York",
            "ResourceId": "spot-fleet-request/sfr-40edeb7b-9ae7-44be-
bef2-5c4c8EXAMPLE",
            "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
            "ScalableTargetAction": {
                "MinCapacity": 1,
                "MaxCapacity": 3
            },
            "CreationTime": 1607456031.391
        }
    ]
}
```

Descrivi una o più operazioni pianificate per un obiettivo scalabile

Per recuperare informazioni sulle operazioni pianificate per un obiettivo scalabile specificato, aggiungi l'opzione `--resource-id` quando descrivi le operazioni pianificate utilizzando il comando describe-scheduled-actions.

Se includi l'opzione `--scheduled-action-names` e specifichi il nome di un'operazione pianificata come valore, il comando restituisce solo l'operazione pianificata il cui nome corrisponde, come illustrato nell'esempio seguente.

Linux, macOS o Unix

```
aws application-autoscaling describe-scheduled-actions --service-namespace ec2 \
  --resource-id spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE \
  --scheduled-action-names my-one-time-action
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-namespace ec2 --
resource-id spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE --scheduled-
action-names my-one-time-action
```

Di seguito è riportato un output di esempio.

```
{
```

```
    "ScheduledActions": [
        {
            "ScheduledActionName": "my-one-time-action",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:4bce34c7-bb81-4ecf-b776-5c726efb1567:resource/ec2/
spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE:scheduledActionName/my-one-
time-action",
            "ServiceNamespace": "ec2",
            "Schedule": "at(2020-12-08T9:36:00)",
            "Timezone": "America/New_York",
            "ResourceId": "spot-fleet-request/sfr-40edeb7b-9ae7-44be-
bef2-5c4c8EXAMPLE",
            "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
            "ScalableTargetAction": {
                "MinCapacity": 1,
                "MaxCapacity": 3
            },
            "CreationTime": 1607456031.391
        }
    ]
}
```

Se vengono forniti più valori per il valore specificato per l'opzione `--scheduled-action-names`, tutte le operazioni pianificate i cui nomi corrispondono sono incluse nell'output.

Disattiva il dimensionamento pianificato per un obiettivo scalabile

È possibile disattivare temporaneamente il dimensionamento pianificato senza eliminare le operazioni pianificate. Per ulteriori informazioni, consulta Sospendi e riprendi il dimensionamento per Application Auto Scaling.

Puoi sospendere il dimensionamento pianificato su un obiettivo scalabile utilizzando il comando register-scalable-target con l'opzione `--suspended-state` e specificando `true` come valore dell'attributo `ScheduledScalingSuspended`, come mostrato nell'esempio seguente.

Linux, macOS o Unix

```
aws application-autoscaling register-scalable-target --service-namespace rds \
  --scalable-dimension rds:cluster:ReadReplicaCount --resource-id cluster:my-db-cluster
 \
  --suspended-state '{"ScheduledScalingSuspended": true}'
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace rds --
scalable-dimension rds:cluster:ReadReplicaCount --resource-id cluster:my-db-cluster --
suspended-state "{\"ScheduledScalingSuspended\": true}"
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Per riattivare il dimensionamento pianificato, esegui nuovamente questo comando, specificando `false` come valore dell'attributo `ScheduledScalingSuspended`.

Eliminazione di un'operazione pianificata

Una volta terminato con un'operazione pianificata, è possibile eliminarla utilizzando il comando delete-scheduled-action.

Linux, macOS o Unix

```
aws application-autoscaling delete-scheduled-action --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-37294EXAMPLE \
  --scheduled-action-name my-recurring-action
```

Windows

```
aws application-autoscaling delete-scheduled-action --service-namespace ec2 --scalable-
dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-request/
sfr-73fbd2ce-aa30-494c-8788-37294EXAMPLE --scheduled-action-name my-recurring-action
```

In caso di esito positivo, il comando torna al prompt.

Tutorial: Guida di base su dimensionamento pianificato utilizzando AWS CLI

Il seguente tutorial mostra come utilizzare il per iniziare con il AWS CLI ridimensionamento pianificato aiutandoti a creare azioni pianificate che ridimensionano una tabella DynamoDB di esempio chiamata. `TestTable` Se non disponi già di una tabella DynamoDB `TestTable` che utilizzi per testare, puoi crearla ora eseguendo il comando create-table mostrato nella Fase 1: creazione di una tabella DynamoDB nella Guida per gli sviluppatori di Amazon DynamoDB.

Quando usi la AWS CLI, ricorda che i comandi vengono eseguiti nella AWS regione configurata per il tuo profilo. Per eseguire i comandi in un'altra regione, modificare la regione predefinita per il profilo oppure utilizzare il parametro `--region` con il comando.

> **Note**
>
> Nell'ambito di questo tutorial potresti incorrere in AWS costi aggiuntivi. Monitora il tuo utilizzo del Piano gratuito e assicurati di comprendere i costi associati al numero di unità di capacità di lettura e scrittura utilizzate dal database DynamoDB.

Indice

Fase 1: registrazione dell'obiettivo scalabile

Inizia registrando la tabella DynamoDB come obiettivo scalabile tramite Application Auto Scaling.

Per registrare un obiettivo scalabile tramite Application Auto Scaling

1. Innanzitutto, utilizza il comando describe-scalable-targets per controllare se le risorse DynamoDB sono già registrate. Ciò consente di verificare che la tabella `TestTable` non sia registrata, nel caso in cui non si tratti di una nuova tabella.

Linux, macOS o Unix

```
aws application-autoscaling describe-scalable-targets \
  --service-namespace dynamodb
```

Windows

```
aws application-autoscaling describe-scalable-targets --service-namespace dynamodb
```

Se non esistono target scalabili, questa è la risposta.

```
{
    "ScalableTargets": []
}
```

2. Utilizza il comando register-scalable-target per registrare la capacità in scrittura della tua tabella DynamoDB denominata `TestTable`. Imposta un capacità desiderata minima di 5 unità di capacità in scrittura e una capacità desiderata massima di 10 unità di capacità in scrittura.

   Linux, macOS o Unix

   ```
   aws application-autoscaling register-scalable-target \
     --service-namespace dynamodb \
     --scalable-dimension dynamodb:table:WriteCapacityUnits \
     --resource-id table/TestTable \
     --min-capacity 5 --max-capacity 10
   ```

   Windows

   ```
   aws application-autoscaling register-scalable-target --service-namespace dynamodb
    --scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/
   TestTable --min-capacity 5 --max-capacity 10
   ```

   In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

   ```
   {
       "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-
   id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
   ```

```
}
```

Fase 2: creazione di due operazioni pianificate

Application Auto Scaling consente di pianificare il momento in cui si verifica un'operazione di dimensionamento. Specifica il target scalabile, la pianificazione e le capacità minima e massima. Al momento specificato, Application Auto Scaling aggiorna il valore minimo e massimo per l'obiettivo scalabile. Se la sua capacità attuale è esterna a questo intervallo, questo comporta un'attività di dimensionamento.

La pianificazione degli aggiornamenti alla capacità minima e massima è anche utile se decidi di creare una policy di dimensionamento. Una policy di dimensionamento consente di scalare dinamicamente le risorse in base all'utilizzo di risorse corrente. Un guardrail comune per una policy di dimensionamento è avere valori appropriati per le capacità minima e massima.

Per questo esercizio, creiamo due operazioni una tantum per diminuire e incrementare.

Per creare e visualizzare le operazioni pianificate

1. Per creare la prima operazione pianificata, utilizza il comando put-scheduled-action.

   Il comando at in `--schedule` pianifica l'operazione per l'esecuzione una tantum alla data e ora specificate in futuro. Le ore sono in formato 24 ore in UTC. Pianificare l'operazione da eseguire tra circa 5 minuti da ora.

   Alla data e all'ora specificate, Application Auto Scaling aggiorna i valori `MinCapacity` e `MaxCapacity`. Ipotizzando che la tabella disponga attualmente di 5 unità di capacità in scrittura, Application Auto Scaling esegue la riduzione orizzontale fino a `MinCapacity` per inserire la tabella all'interno del nuovo intervallo desiderato di 15-20 unità di capacità in scrittura.

   Linux, macOS o Unix

   ```
   aws application-autoscaling put-scheduled-action \
     --service-namespace dynamodb \
     --scalable-dimension dynamodb:table:WriteCapacityUnits \
     --resource-id table/TestTable \
     --scheduled-action-name my-first-scheduled-action \
     --schedule "at(2019-05-20T17:05:00)" \
     --scalable-target-action MinCapacity=15,MaxCapacity=20
   ```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace dynamodb
 --scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/
TestTable --scheduled-action-name my-first-scheduled-action --schedule
 "at(2019-05-20T17:05:00)" --scalable-target-action MinCapacity=15,MaxCapacity=20
```

Questo comando non restituisce alcun output se va a buon fine.

2. Per creare la seconda operazione pianificata utilizzata da Application Auto Scaling per la riduzione orizzontale, utilizza il comando put-scheduled-action.

   Pianificare l'operazione affinché venga eseguita tra circa 10 minuti da ora.

   Alla data e all'ora specificate, Application Auto Scaling aggiorna i valori `MinCapacity` e `MaxCapacity` della tabella ed esegue la riduzione orizzontale fino a `MaxCapacity` per ripristinare la tabella nell'intervallo desiderato originale di 5-10 unità di capacità in scrittura.

   Linux, macOS o Unix

   ```
   aws application-autoscaling put-scheduled-action \
     --service-namespace dynamodb \
     --scalable-dimension dynamodb:table:WriteCapacityUnits \
     --resource-id table/TestTable \
     --scheduled-action-name my-second-scheduled-action \
     --schedule "at(2019-05-20T17:10:00)" \
     --scalable-target-action MinCapacity=5,MaxCapacity=10
   ```

   Windows

   ```
   aws application-autoscaling put-scheduled-action --service-namespace dynamodb
    --scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/
   TestTable --scheduled-action-name my-second-scheduled-action --schedule
    "at(2019-05-20T17:10:00)" --scalable-target-action MinCapacity=5,MaxCapacity=10
   ```

3. (Facoltativo) Ottieni un elenco di operazioni pianificate per lo spazio dei nomi del servizio specificato utilizzando il comando describe-scheduled-actions.

   Linux, macOS o Unix

   ```
   aws application-autoscaling describe-scheduled-actions \
   ```

```
  --service-namespace dynamodb
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-namespace dynamodb
```

Di seguito è riportato un output di esempio.

```
{
    "ScheduledActions": [
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Schedule": "at(2019-05-20T18:35:00)",
            "ResourceId": "table/TestTable",
            "CreationTime": 1561571888.361,
            "ScheduledActionARN": "arn:aws:autoscaling:us-
east-1:123456789012:scheduledAction:2d36aa3b-cdf9-4565-b290-81db519b227d:resource/
dynamodb/table/TestTable:scheduledActionName/my-first-scheduled-action",
            "ScalableTargetAction": {
                "MinCapacity": 15,
                "MaxCapacity": 20
            },
            "ScheduledActionName": "my-first-scheduled-action",
            "ServiceNamespace": "dynamodb"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Schedule": "at(2019-05-20T18:40:00)",
            "ResourceId": "table/TestTable",
            "CreationTime": 1561571946.021,
            "ScheduledActionARN": "arn:aws:autoscaling:us-
east-1:123456789012:scheduledAction:2d36aa3b-cdf9-4565-b290-81db519b227d:resource/
dynamodb/table/TestTable:scheduledActionName/my-second-scheduled-action",
            "ScalableTargetAction": {
                "MinCapacity": 5,
                "MaxCapacity": 10
            },
            "ScheduledActionName": "my-second-scheduled-action",
            "ServiceNamespace": "dynamodb"
        }
    ]
}
```

Fase 3: visualizzazione delle attività di dimensionamento

In questa fase è possibile visualizzare le attività di dimensionamento attivate dalle operazioni pianificate e quindi verificare che DynamoDB abbia modificato la capacità di scrittura della tabella.

Per visualizzare le attività di dimensionamento

1. Attendi il tempo scelto e verifica che le operazioni pianificate siano in funzione utilizzando il comando describe-scaling-activities.

   Linux, macOS o Unix

   ```
   aws application-autoscaling describe-scaling-activities \
     --service-namespace dynamodb
   ```

   Windows

   ```
   aws application-autoscaling describe-scaling-activities --service-
   namespace dynamodb
   ```

   Di seguito è riportato l'output di esempio per la prima operazione pianificata mentre questa è in corso.

   Le attività di dimensionamento sono ordinate in base alla data di creazione, con le attività di dimensionamento più recenti restituite per prime.

   ```
   {
       "ScalingActivities": [
           {
               "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
               "Description": "Setting write capacity units to 15.",
               "ResourceId": "table/TestTable",
               "ActivityId": "d8ea4de6-9eaa-499f-b466-2cc5e681ba8b",
               "StartTime": 1561574108.904,
               "ServiceNamespace": "dynamodb",
               "Cause": "minimum capacity was set to 15",
               "StatusMessage": "Successfully set write capacity units to 15. Waiting
    for change to be fulfilled by dynamodb.",
               "StatusCode": "InProgress"
           },
           {
   ```

```
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting min capacity to 15 and max capacity to 20",
            "ResourceId": "table/TestTable",
            "ActivityId": "3250fd06-6940-4e8e-bb1f-d494db7554d2",
            "StartTime": 1561574108.512,
            "ServiceNamespace": "dynamodb",
            "Cause": "scheduled action name my-first-scheduled-action was
 triggered",
            "StatusMessage": "Successfully set min capacity to 15 and max capacity
 to 20",
            "StatusCode": "Successful"
        }
    ]
}
```

L'esempio seguente è l'output dopo l'esecuzione di entrambe le operazioni pianificate.

```
{
    "ScalingActivities": [
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting write capacity units to 10.",
            "ResourceId": "table/TestTable",
            "ActivityId": "4d1308c0-bbcf-4514-a673-b0220ae38547",
            "StartTime": 1561574415.086,
            "ServiceNamespace": "dynamodb",
            "EndTime": 1561574449.51,
            "Cause": "maximum capacity was set to 10",
            "StatusMessage": "Successfully set write capacity units to 10. Change
 successfully fulfilled by dynamodb.",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting min capacity to 5 and max capacity to 10",
            "ResourceId": "table/TestTable",
            "ActivityId": "f2b7847b-721d-4e01-8ef0-0c8d3bacc1c7",
            "StartTime": 1561574414.644,
            "ServiceNamespace": "dynamodb",
            "Cause": "scheduled action name my-second-scheduled-action was
 triggered",
            "StatusMessage": "Successfully set min capacity to 5 and max capacity
 to 10",
```

```
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting write capacity units to 15.",
            "ResourceId": "table/TestTable",
            "ActivityId": "d8ea4de6-9eaa-499f-b466-2cc5e681ba8b",
            "StartTime": 1561574108.904,
            "ServiceNamespace": "dynamodb",
            "EndTime": 1561574140.255,
            "Cause": "minimum capacity was set to 15",
            "StatusMessage": "Successfully set write capacity units to 15. Change
 successfully fulfilled by dynamodb.",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting min capacity to 15 and max capacity to 20",
            "ResourceId": "table/TestTable",
            "ActivityId": "3250fd06-6940-4e8e-bb1f-d494db7554d2",
            "StartTime": 1561574108.512,
            "ServiceNamespace": "dynamodb",
            "Cause": "scheduled action name my-first-scheduled-action was
 triggered",
            "StatusMessage": "Successfully set min capacity to 15 and max capacity
 to 20",
            "StatusCode": "Successful"
        }
    ]
}
```

2. Dopo avere eseguito le operazioni pianificate, apri la console DynamoDB e scegli la tabella che desideri utilizzare. Visualizzare Write capacity units (Unità di capacità in scrittura) nella scheda Capacity (Capacità). Dopo l'esecuzione della seconda operazione di dimensionamento, le unità di capacità in scrittura sono state scalate da 15 a 10.

   Puoi anche verificare la capacità in scrittura corrente della tabella utilizzando il seguente comando describe-table. Per filtrare l'output, includi l'opzione `--query`. Per ulteriori informazioni sulle funzionalità di filtraggio dell'output di AWS CLI, vedere Controlling command output from the AWS CLI nella Guida per l'utente.AWS Command Line Interface

   Linux, macOS o Unix

```
aws dynamodb describe-table --table-name TestTable \
  --query 'Table.[TableName,TableStatus,ProvisionedThroughput]'
```

Windows

```
aws dynamodb describe-table --table-name TestTable --query "Table.
[TableName,TableStatus,ProvisionedThroughput]"
```

Di seguito è riportato un output di esempio.

```
[
    "TestTable",
    "ACTIVE",
    {
        "NumberOfDecreasesToday": 1,
        "WriteCapacityUnits": 10,
        "LastIncreaseDateTime": 1561574133.264,
        "ReadCapacityUnits": 5,
        "LastDecreaseDateTime": 1561574435.607
    }
]
```

Fase 4: fasi successive

Se vuoi provare a scalare sia con il dimensionamento pianificato che con una policy di ridimensionamento, segui i passaggi descritti in Tutorial: configura il dimensionamento automatico per gestire un carico di lavoro pesante.

Fase 5: rimozione

Al termine dell'utilizzo degli esercizi sulle nozioni di base, puoi eliminare le risorse associate come segue.

Per eliminare le operazioni pianificate

Il comando delete-scheduled-action elimina una specifica operazione pianificata. Puoi ignorare questa fase se desideri mantenere l'operazione pianificata per l'utilizzo futuro.

Linux, macOS o Unix

```
aws application-autoscaling delete-scheduled-action \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:WriteCapacityUnits \
  --resource-id table/TestTable \
  --scheduled-action-name my-second-scheduled-action
```

Windows

```
aws application-autoscaling delete-scheduled-action --service-namespace dynamodb --
scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/TestTable --
scheduled-action-name my-second-scheduled-action
```

Per annullare la registrazione di un obiettivo scalabile

Utilizza il comando deregister-scalable-target per annullare la registrazione dell'obiettivo scalabile. Se hai a disposizione policy di dimensionamento create o eventuali operazioni pianificate che non sono state ancora eliminate, verranno eliminate tramite questo comando. Se desideri mantenere l'obiettivo scalabile registrato per utilizzarlo in futuro, puoi ignorare questa fase.

Linux, macOS o Unix

```
aws application-autoscaling deregister-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:WriteCapacityUnits \
  --resource-id table/TestTable
```

Windows

```
aws application-autoscaling deregister-scalable-target --service-namespace dynamodb --
scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/TestTable
```

Per eliminare la tabella DynamoDB

Utilizza il comando delete-table per eliminare la tabella utilizzata in questo tutorial. Puoi ignorare questa fase se desideri mantenere la tabella per un utilizzo futuro.

Linux, macOS o Unix

```
aws dynamodb delete-table --table-name TestTable
```

Windows

```
aws dynamodb delete-table --table-name TestTable
```

Policy di dimensionamento con monitoraggio degli obiettivi per Application Auto Scaling

Una policy di dimensionamento del monitoraggio degli obiettivi dimensiona automaticamente l'applicazione in base al valore di un parametro target. Ciò consente all'applicazione di mantenere prestazioni ottimali ed efficienza in termini di costi senza intervento manuale.

Con il monitoraggio degli obiettivi, devi selezionare un parametro e un valore target che rappresenti il livello di utilizzo o velocità di trasmissione effettiva ideale per la tua applicazione. Application Auto Scaling crea e gestisce gli CloudWatch allarmi che attivano gli eventi di scalabilità quando la metrica si discosta dall'obiettivo. È simile al modo in cui un termostato mantiene una temperatura target.

Ad esempio, immaginiamo di avere un'applicazione Web attualmente eseguita su una serie di istanze Spot e di volere che l'utilizzo della CPU del parco istanze con scalabilità automatica rimanga intorno al 50% quando il carico sull'applicazione varia. Questo ti offre la capacità aggiuntiva per gestire i picchi di traffico senza dover mantenere un numero eccessivo di risorse inattive.

È possibile soddisfare questa esigenza creando una policy di dimensionamento del monitoraggio degli obiettivi che si rivolge a un utilizzo medio della CPU del 50%. Quindi, Application Auto Scaling impiegherà la scalabilità orizzontale (aumento della capacità) quando la CPU supera il 50 percento per gestire un carico maggiore. Impiegherà la riduzione orizzontale (diminuzione della capacità) quando la CPU scende al di sotto del 50% per ottimizzare i costi nei periodi di basso utilizzo.

Le policy di tracciamento di Target eliminano la necessità di definire CloudWatch manualmente gli allarmi e le regolazioni di ridimensionamento. Application Auto Scaling lo gestisce automaticamente in base all'obiettivo impostato.

Puoi basare le policy di dimensionamento con monitoraggio degli obiettivi sia su parametri predefiniti che personalizzati.

- Parametri predefiniti: parametri forniti da Application Auto Scaling come l'utilizzo medio della CPU o il numero medio di richieste per destinazione.
- Metriche personalizzate: puoi utilizzare la matematica metrica per combinare metriche, sfruttare le metriche esistenti o utilizzare le tue metriche personalizzate pubblicate su. CloudWatch

Scegli un parametro che cambi in modo inversamente proporzionale alla variazione della capacità del tuo target scalabile. Quindi, se raddoppi la capacità, la metrica diminuisce del 50 percento. Ciò consente ai dati dei parametri di attivare con precisione gli eventi di scalabilità proporzionale.

Indice

Come funziona il target tracking scaling per Application Auto Scaling

Questo argomento descrive come funziona il ridimensionamento del tracciamento del target e introduce gli elementi chiave di una politica di ridimensionamento del tracciamento del target.

Indice

Come funziona

Per utilizzare il ridimensionamento del tracciamento del target, create una politica di ridimensionamento del tracciamento del target e specificate quanto segue:

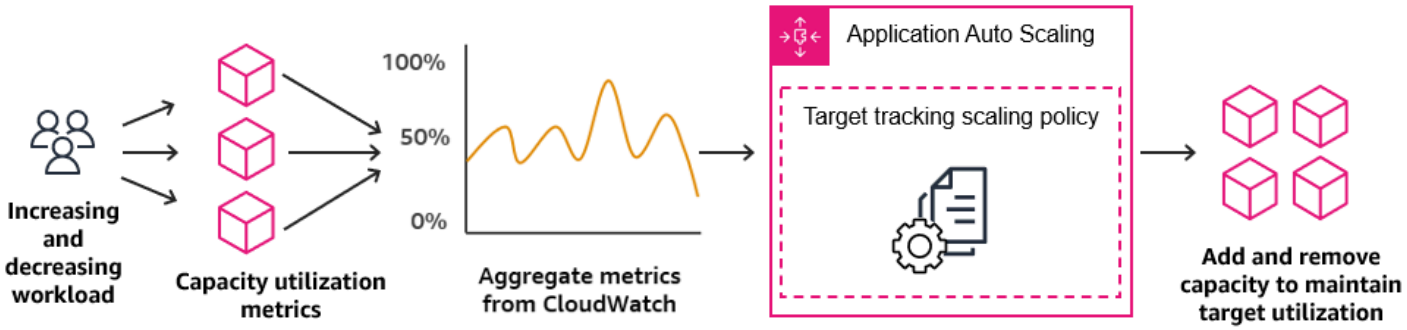
- Metrica: una CloudWatch metrica da monitorare, ad esempio l'utilizzo medio della CPU o il numero medio di richieste per destinazione.
- Valore target: il valore target per il parametro, ad esempio il 50% di utilizzo della CPU o 1000 richieste per target al minuto.

Application Auto Scaling crea e gestisce gli CloudWatch allarmi che richiamano la politica di scalabilità e calcola la regolazione della scalabilità in base alla metrica e al valore target. Aggiunge e rimuove la capacità in base alle necessità, per mantenere il parametro al valore di destinazione specificato o vicino a esso.

Quando il parametro è superiore al valore target, Application Auto Scaling impiega la scalabilità orizzontale aggiungendo capacità per ridurre la differenza tra il valore del parametro e il valore di destinazione. Quando il parametro è inferiore al valore target, Application Auto Scaling impiega il ridimensionamento rimuovendo la capacità.

Le attività di dimensionamento vengono eseguite con periodi di raffreddamento intermedi per evitare rapide fluttuazioni della capacità. Facoltativamente, puoi configurare i tempi di raffreddamento per la tua policy di dimensionamento.

Il diagramma seguente mostra una panoramica del funzionamento di una policy di dimensionamento del monitoraggio delle destinazioni una volta completata la configurazione.



Una policy di dimensionamento del monitoraggio della destinazione è più aggressiva nell'aggiunta di capacità quando l'utilizzo aumenta, rispetto alla rimozione della capacità quando l'utilizzo diminuisce. Ad esempio, se il parametro specificato della policy raggiunge il valore obiettivo, la policy presuppone che l'applicazione sia già sottoposta a un forte carico. Quindi risponde aggiungendo capacità proporzionale al valore del parametro il più velocemente possibile. Più alto è il parametro, maggiore è la capacità aggiunta.

Quando il parametro scende al di sotto del valore di destinazione, la policy prevede che l'utilizzo aumenterà di nuovo. In questo caso, rallenta il dimensionamento rimuovendo la capacità solo

quando l'utilizzo supera una soglia che è sufficientemente inferiore al valore obiettivo (in genere almeno del 10% inferiore) per considerare l'utilizzo rallentato. Lo scopo di questo comportamento più conservativo è garantire che la rimozione della capacità avvenga solo quando l'applicazione non riscontra più una domanda allo stesso livello elevato come in precedenza.

Selezionare i parametri.

Puoi creare policy di dimensionamento con monitoraggio degli obiettivi sia con parametri predefiniti che personalizzati.

Quando crei una policy di dimensionamento del monitoraggio degli obiettivi con un tipo parametro predefinito, scegli un parametro dalla lista di parametri predefiniti in Policy di dimensionamento del monitoraggio degli obiettivi con parametri predefiniti.

Quando scegli un parametro, tieni presente quanto segue:

- Non tutti i parametri personalizzati funzionano per il monitoraggio degli obiettivi. Il parametro deve essere un parametro di utilizzo valido e deve descrivere il livello di occupazione di un target scalabile. Il valore del parametro deve aumentare o diminuire in modo proporzionale alla capacità del target scalabile in modo che i dati del parametro possano essere utilizzati per eseguire il dimensionamento proporzionale del target scalabile.
- Per utilizzare il parametro `ALBRequestCountPerTarget`, è necessario specificare il parametro `ResourceLabel` per identificare il gruppo di destinazione associato al parametro.
- Quando una metrica emette valori 0 reali su CloudWatch (ad esempio`ALBRequestCountPerTarget`), Application Auto Scaling può scalare fino a 0 quando non c'è traffico verso l'applicazione per un periodo di tempo prolungato. Per fare in modo che l'obiettivo scalabile si riduca orizzontalmente a 0 istanze quando non riceve richieste instradate, la capacità minima dell'obiettivo scalabile deve essere impostata su 0.
- Invece di pubblicare nuove metriche da utilizzare nella politica di scalabilità, è possibile utilizzare la matematica delle metriche per combinare quelle esistenti. Per ulteriori informazioni, consulta Creazione di una policy di dimensionamento con monitoraggio degli obiettivi per l'Applicazione di Dimensionamento automatico tramite la matematica dei parametri.
- Per vedere se il servizio che stai utilizzando supporta la possibilità di specificare un parametro personalizzato nella console del servizio, consulta la documentazione per tale servizio.
- Ti consigliamo di utilizzare parametri disponibili a intervalli di un minuto per aiutarti a dimensionare più rapidamente quando viene modificato l'utilizzo. Il monitoraggio degli obiettivi valuta i parametri aggregati con una granularità di un minuto per tutti i parametri predefiniti e personalizzati, ma

il parametro sottostante potrebbe pubblicare i dati meno frequentemente. Ad esempio, tutti i parametri di Amazon EC2 vengono inviati a intervalli di cinque minuti per impostazione predefinita, ma sono configurabili fino a un minuto (noto come monitoraggio dettagliato). Questa scelta spetta ai singoli servizi. La maggior parte dei servizi cerca di utilizzare l'intervallo più breve possibile.

Definire il valore target

Quando si crea una policy di dimensionamento del monitoraggio degli obiettivi, è necessario specificare un valore target. Il valore di destinazione rappresenta l'utilizzo medio ideale o la velocità di trasmissione effettiva per l'applicazione. Per utilizzare le risorse in modo efficiente in termini di costi, impostare il valore target il più alto possibile con un buffer ragionevole per aumenti di traffico imprevisti. Quando l'applicazione viene aumentata orizzontalmente in modo ottimale per un normale flusso di traffico, il valore del parametro effettivo deve essere pari o appena inferiore al valore di destinazione.

Quando una policy di scalabilità si basa sulla velocità di trasmissione effettiva, ad esempio il conteggio delle richieste per destinazione per un Application Load Balancer, I/O di rete o altri parametri di conteggio, il valore di destinazione rappresenta la velocità di trasmissione effettiva media ottimale da una singola entità (ad esempio una singola destinazione del gruppo di destinazione Application Load Balancer), per un periodo di un minuto.

Definizione dei tempi di raffreddamento

Facoltativamente, puoi definire i tempi di raffreddamento nella policy di dimensionamento del monitoraggio degli obiettivi.

Il tempo di raffreddamento indica la quantità di tempo che la policy di dimensionamento attende prima di rendere effettiva una precedente attività di dimensionamento.

Esistono due tipi di tempi di raffreddamento:

- Con il periodo di attesa di incremento, l'intenzione è di dimensionare in modo continuo (ma non eccessivamente). Dopo che Application Auto Scaling ha impiegato correttamente la scalabilità orizzontale utilizzando una policy di dimensionamento, inizia a calcolare il tempo di raffreddamento. Una policy di dimensionamento non aumenta di nuovo la capacità desiderata, a meno che non venga impiegata una scalabilità orizzontale maggiore o che il tempo di raffreddamento finisca. Mentre è attivo il periodo di attesa di incremento, la capacità aggiunta dall'attività di incremento iniziale viene calcolata come parte della capacità desiderata per il successivo evento di incremento.

- Con il tempo di raffreddamento per il ridimensionamento, l'intenzione è di ridimensionare in modo conservativo per proteggere la disponibilità dell'applicazione, per cui le attività di ridimensionamento vengono bloccate fino alla scadenza del tempo di raffreddamento per il ridimensionamento. Tuttavia, se un altro allarme attiva un'attività di incremento durante il periodo di attesa, Application Auto Scaling incrementa immediatamente la destinazione. In questo caso, il tempo di raffreddamento per il ridimensionamento si interrompe e non viene completato.

Ogni periodo di attesa viene misurato in secondi e si applica solo alle attività di dimensionamento correlate alle policy. Durante un periodo di attesa, quando un'operazione pianificata inizia all'ora pianificata, può attivare immediatamente un'attività di dimensionamento senza attendere la scadenza del periodo di attesa.

È possibile iniziare con i valori predefiniti, che possono essere poi ottimizzati. Ad esempio, potrebbe essere necessario aumentare un periodo di attesa per evitare che la policy di dimensionamento del monitoraggio di destinazione sia troppo aggressiva rispetto alle modifiche che si verificano in brevi periodi di tempo.

Valori predefiniti

Application Auto Scaling fornisce un valore predefinito di 600 per i gruppi di ElastiCache replica e un valore predefinito di 300 per le seguenti destinazioni scalabili:

- AppStream flotte 2.0
- Cluster di database Aurora
- Servizi ECS
- Cluster di Neptune
- SageMaker varianti di endpoint
- SageMaker componenti di inferenza
- SageMaker Concorrenza fornita senza server
- Parco istanze Spot
- Risorse personalizzate

Per tutti gli altri obiettivi dimensionabili, il valore predefinito è 0 oppure null:

- Endpoint di classificazione dei documenti Amazon Comprehend e di riconoscimento delle identità

- Tabelle DynamoDB e indici secondari globali
- Tabelle di Amazon Keyspaces
- Concorrenza con provisioning di Lambda
- Storage di broker Amazon MSK

I valori null vengono considerati come valori zero quando Application Auto Scaling valuta il tempo di raffreddamento.

Puoi aggiornare qualsiasi valore predefinito, inclusi i valori null, per impostare i tempi di raffreddamento.

Considerazioni

Le seguenti considerazioni si applicano quando si usano le policy di dimensionamento con monitoraggio degli obiettivi:

- Non create, modificate o eliminate gli CloudWatch allarmi utilizzati con una politica di scalabilità di Target Tracking. Application Auto Scaling crea e gestisce gli CloudWatch allarmi associati alle politiche di scalabilità di tracciamento di Target e li elimina quando non sono più necessari.
- Se nella metrica mancano dei punti dati, lo stato di CloudWatch allarme passa a. `INSUFFICIENT_DATA` In questo caso, Application Auto Scaling non può dimensionare l'obiettivo scalabile finché non vengono trovati nuovi punti dati. Per ulteriori informazioni, consulta Configurazione del modo in cui gli CloudWatch allarmi trattano i dati mancanti nella Amazon CloudWatch User Guide.
- Se la metrica viene riportata scarsamente in base alla progettazione, la matematica metrica può essere utile. Ad esempio, per utilizzare i valori più recenti, utilizzate la `FILL(m1,REPEAT)` funzione dove `m1` è la metrica.
- Potrebbero esserci delle differenze tra il valore di destinazione e i punti di dati dei parametri reali. Ciò avviene perché Application Auto Scaling agisce sempre con prudenza, arrotondando per eccesso o per difetto quando determina la capacità da aggiungere o rimuovere. In questo modo si impedisce l'aggiunta di capacità insufficiente o la rimozione di capacità eccessiva. Tuttavia, per un obiettivo scalabile con piccole capacità, i punti di dati dei parametri reali potrebbero sembrare lontani dal valore di destinazione.

  Per un obiettivo scalabile con maggiori capacità, l'aggiunta o la rimozione di capacità fa sì che vi sia un intervallo minore tra il valore di destinazione e i punti di dati dei parametri reali.

- Una policy di dimensionamento di monitoraggio obiettivi presuppone che essa debba eseguire un dimensionamento orizzontale quando il parametro specificato supera il valore di destinazione. Non puoi utilizzare una policy di dimensionamento di monitoraggio obiettivi per il dimensionamento orizzontale quando il parametro specificato è inferiore al valore di destinazione.

Più policy di dimensionamento

È possibile avere più policy di dimensionamento del monitoraggio di target per un target scalabile, purché ciascuna di esse utilizzi un parametro diverso. Lo scopo di Application Auto Scaling è sempre quello di assegnare la priorità alla disponibilità, quindi il suo comportamento varia a seconda che le policy di monitoraggio degli obiettivi siano pronte o meno per l'aumento o la riduzione orizzontale. L'obiettivo scalabile viene aumentato se una qualsiasi delle policy di monitoraggio dei target è pronta per l'aumento e viene ridotto solo se tutte le policy di monitoraggio dei target (con la porzione di riduzione abilitata) sono pronte per essere ridotte.

Se più policy di dimensionamento impongono all'obiettivo scalabile una riduzione o un aumento orizzontale allo stesso tempo, Application Auto Scaling dimensiona in base alla policy che fornisce la capacità massima sia per la riduzione sia per l'aumento orizzontale. Ciò offre maggiore flessibilità per coprire scenari diversi e garantisce che vi sia sempre capacità sufficiente per elaborare i carichi di lavoro delle applicazioni.

Puoi disabilitare la porzione del ridimensionamento di una policy di dimensionamento del monitoraggio degli obiettivi per utilizzare un metodo di ridimensionamento diverso da quello usato per la scalabilità orizzontale. Ad esempio, è possibile utilizzare un altro tipo di policy di dimensionamento per il dimensionamento verticale e utilizzare una policy di dimensionamento del monitoraggio obiettivi per il dimensionamento orizzontale.

Suggeriamo, tuttavia, di prestare attenzione quando si utilizzano le policy di dimensionamento con monitoraggio degli obiettivi insieme alle policy di dimensionamento per fasi, per evitare che insorgano conflitti che possono causare comportamenti indesiderati. Ad esempio, se la policy di dimensionamento per fasi avvia un'attività di riduzione prima che la policy con monitoraggio degli obiettivi sia pronta per eseguirla, tale attività non verrà bloccata. Al termine dell'attività di scalabilità verticale, la policy di monitoraggio della destinazione potrebbe indicare alla destinazione scalabile di nuovo la scalabilità orizzontale.

Per i carichi di lavoro di natura ciclica, è inoltre possibile automatizzare le modifiche di capacità in una pianificazione utilizzando il dimensionamento pianificato. Per ogni operazione programmata, è possibile definire un nuovo valore di capacità minima e un nuovo valore di capacità massima.

Questi valori formano i limiti della policy di dimensionamento. La combinazione di dimensionamento pianificato e di monitoraggio della destinazione può contribuire a ridurre l'impatto di un forte aumento dei livelli di utilizzo, quando la capacità è immediatamente necessaria.

Comandi comunemente utilizzati per la creazione, la gestione e l'eliminazione delle policy di dimensionamento

I comandi comunemente utilizzati per le policy di dimensionamento includono:

- register-scalable-target per registrare AWS o personalizzare le risorse come destinazioni scalabili (una risorsa scalabile da Application Auto Scaling) e per sospendere e riprendere il ridimensionamento.
- put-scaling-policy per aggiungere o modificare policy di dimensionamento per un obiettivo scalabile esistente.
- describe-scaling-activities per restituire informazioni sulle attività di scalabilità in una regione. AWS
- describe-scaling-policies per restituire informazioni sulle policy di dimensionamento in una Regione AWS .
- delete-scaling-policy per eliminare una policy di dimensionamento.

Risorse correlate

Per ulteriori informazioni su come creare policy di dimensionamento con monitoraggio della destinazione per i gruppi con dimensionamento automatico, consulta Policy di dimensionamento semplici e con monitoraggio della destinazione per Dimensionamento automatico Amazon EC2 nella Guida per l'utente di Dimensionamento automatico Amazon EC2.

Limitazioni

Di seguito sono riportate le restrizioni quando si utilizzano le policy di dimensionamento con monitoraggio degli obiettivi:

- L'obiettivo scalabile non può essere un cluster Amazon EMR. Le policy di dimensionamento con monitoraggio degli obiettivi non sono supportate per Amazon EMR.
- Quando un cluster Amazon MSK è l'obiettivo scalabile, la riduzione orizzontale è disabilitata e non può essere abilitata.
- Non è possibile utilizzare le operazioni `RegisterScalableTarget` o l'`PutScalingPolicy`API per aggiornare un piano di scalabilità. AWS Auto Scaling

- L'accesso da console per visualizzare, aggiungere, aggiornare o rimuovere policy di dimensionamento con monitoraggio della destinazione su risorse scalabili dipende dalla risorsa utilizzata. Per ulteriori informazioni, consulta Servizi AWS che puoi usare con Application Auto Scaling.

Creare una politica di ridimensionamento del tracciamento degli obiettivi per Application Auto Scaling utilizzando il AWS CLI

È possibile creare una politica di scalabilità di tracciamento degli obiettivi per Application Auto Scaling utilizzando per AWS CLI le seguenti attività di configurazione.

1. Registrazione di una destinazione dimensionabile.
2. Aggiungi una policy di dimensionamento con monitoraggio della destinazione sulla destinazione dimensionabile.

Per brevità, gli esempi di questo argomento illustrano i comandi della CLI per una serie di istanze Spot di Amazon EC2. Per specificare un altro target scalabile, specificare il suo spazio dei nomi in `--service-namespace`, la sua dimensione scalabile `--scalable-dimension` e l'ID di risorsa in `--resource-id`. Per maggiori informazioni ed esempi per ogni servizio, consultare gli argomenti in Servizi AWS che puoi usare con Application Auto Scaling.

Quando usi il AWS CLI, ricorda che i tuoi comandi vengono eseguiti nella Regione AWS configurazione per il tuo profilo. Per eseguire i comandi in un'altra regione, modificare la regione predefinita per il profilo oppure utilizzare il parametro `--region` con il comando.

Indice

Registrazione di un target scalabile

Se non lo hai ancora fatto, registra l'obiettivo scalabile. Utilizza il comando register-scalable-target per registrare una risorsa specifica nel servizio obiettivo come obiettivo scalabile. Nell'esempio seguente

viene registrata una richiesta alla serie di istanze Spot con Application Auto Scaling. Application Auto Scaling può dimensionare il numero di istanze della serie di istanze Spot a un minimo di 2 e un massimo di 10 istanze. Sostituisci ciascun *placeholder input dell'utente* con le tue informazioni.

Linux, macOS o Unix

```
aws application-autoscaling register-scalable-target --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
  --min-capacity 2 --max-capacity 10
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace ec2 --
scalable-dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-
request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE --min-capacity 2 --max-capacity 10
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Creazione di una policy di dimensionamento con monitoraggio degli obiettivi

Per creare una politica di ridimensionamento del tracciamento degli obiettivi, puoi utilizzare i seguenti esempi per iniziare.

Creazione di una policy di dimensionamento con monitoraggio degli obiettivi

1. Utilizzate il `cat` comando seguente per memorizzare un valore target per la vostra politica di scalabilità e una specifica metrica predefinita in un file JSON denominato `config.json` nella vostra home directory. Di seguito è riportato un esempio di configurazione di tracciamento degli obiettivi che mantiene l'utilizzo medio della CPU al 50%.

   ```
   $ cat ~/config.json
   {
     "TargetValue": 50.0,
   ```

```
  "PredefinedMetricSpecification":
    {
      "PredefinedMetricType": "EC2SpotFleetRequestAverageCPUUtilization"
    }
}
```

Per ulteriori informazioni, vedere PredefinedMetricSpecificazione nell'Application Auto Scaling API Reference.

In alternativa, puoi utilizzare una metrica personalizzata per il ridimensionamento creando una specifica metrica personalizzata e aggiungendo valori per ogni parametro da. CloudWatch Di seguito è riportato un esempio di configurazione di tracciamento degli obiettivi che mantiene l'utilizzo medio della metrica specificata a 100.

```
$ cat ~/config.json
{
   "TargetValue": 100.0,
   "CustomizedMetricSpecification":{
      "MetricName": "MyUtilizationMetric",
      "Namespace": "MyNamespace",
      "Dimensions": [
         {
            "Name": "MyOptionalMetricDimensionName",
            "Value": "MyOptionalMetricDimensionValue"
         }
      ],
      "Statistic": "Average",
      "Unit": "Percent"
   }
}
```

Per ulteriori informazioni, vedere CustomizedMetricSpecificazione nell'Application Auto Scaling API Reference.

2. Utilizza il comando put-scaling-policy insieme al file `config.json` che hai creato per generare una policy di dimensionamento denominata `cpu50-target-tracking-scaling-policy`.

   Linux, macOS o Unix

```
aws application-autoscaling put-scaling-policy --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
```

```
  --policy-name cpu50-target-tracking-scaling-policy --policy-type
 TargetTrackingScaling \
  --target-tracking-scaling-policy-configuration file://config.json
```

Windows

```
aws application-autoscaling put-scaling-policy --service-namespace ec2 --scalable-
dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-request/
sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE --policy-name cpu50-target-tracking-
scaling-policy --policy-type TargetTrackingScaling --target-tracking-scaling-
policy-configuration file://config.json
```

In caso di successo, questo comando restituisce gli ARN e i nomi dei due CloudWatch allarmi creati per tuo conto.

```
{
    "PolicyARN": "arn:aws:autoscaling:region:account-
id:scalingPolicy:policy-id:resource/ec2/spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE:policyName/cpu50-target-tracking-scaling-policy",
    "Alarms": [
        {
            "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-
spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-
b46e-434a-a60f-3b36d653feca",
            "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-b46e-434a-a60f-3b36d653feca"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-
spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-
d19b-4a63-a812-6c67aaf2910d",
            "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-d19b-4a63-a812-6c67aaf2910d"
        }
    ]
}
```

Descrizione delle policy di dimensionamento con monitoraggio degli obiettivi

È possibile descrivere tutte le policy di dimensionamento per lo spazio dei nomi dei servizi specificato utilizzando il comando describe-scaling-policies.

```
aws application-autoscaling describe-scaling-policies --service-namespace ec2
```

È possibile filtrare i risultati solo per le policy di dimensionamento del monitoraggio di target utilizzando il parametro `--query`. Per ulteriori informazioni sulla sintassi per `query`, consulta Controllo dell'output del comando dalla AWS CLI nella Guida per l'utente di AWS Command Line Interface .

Linux, macOS o Unix

```
aws application-autoscaling describe-scaling-policies --service-namespace ec2 \
  --query 'ScalingPolicies[?PolicyType==`TargetTrackingScaling`]'
```

Windows

```
aws application-autoscaling describe-scaling-policies --service-namespace ec2 --query
 "ScalingPolicies[?PolicyType==`TargetTrackingScaling`]"
```

Di seguito è riportato un output di esempio.

```
[
    {
        "PolicyARN": "PolicyARN",
        "TargetTrackingScalingPolicyConfiguration": {
            "PredefinedMetricSpecification": {
                "PredefinedMetricType": "EC2SpotFleetRequestAverageCPUUtilization"
            },
            "TargetValue": 50.0
        },
        "PolicyName": "cpu50-target-tracking-scaling-policy",
        "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
        "ServiceNamespace": "ec2",
        "PolicyType": "TargetTrackingScaling",
        "ResourceId": "spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE",
        "Alarms": [
            {
```

```
                "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-
spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-
b46e-434a-a60f-3b36d653feca",
                "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-b46e-434a-a60f-3b36d653feca"
            },
            {
                "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-
spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-
d19b-4a63-a812-6c67aaf2910d",
                "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-d19b-4a63-a812-6c67aaf2910d"
            }
        ],
        "CreationTime": 1515021724.807
    }
]
```

Eliminazione di una policy di dimensionamento con monitoraggio degli obiettivi

Una volta terminato con una policy di dimensionamento del monitoraggio di target, è possibile eliminarla utilizzando il comando delete-scaling-policy.

Il seguente comando elimina la policy di dimensionamento con monitoraggio degli obiettivi specificata per la richiesta della serie di istanze Spot indicata. Elimina anche gli CloudWatch allarmi creati da Application Auto Scaling per tuo conto.

Linux, macOS o Unix

```
aws application-autoscaling delete-scaling-policy --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
  --policy-name cpu50-target-tracking-scaling-policy
```

Windows

```
aws application-autoscaling delete-scaling-policy --service-namespace ec2 --scalable-
dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-request/
sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE --policy-name cpu50-target-tracking-scaling-
policy
```

Creazione di una policy di dimensionamento con monitoraggio degli obiettivi per l'Applicazione di Dimensionamento automatico tramite la matematica dei parametri

Utilizzando la matematica metrica, puoi interrogare più CloudWatch metriche e utilizzare espressioni matematiche per creare nuove serie temporali basate su queste metriche. Puoi visualizzare le serie temporali risultanti nella CloudWatch console e aggiungerle ai dashboard. Per ulteriori informazioni sulla matematica dei parametri, consulta Using metric Math nella Amazon User Guide. CloudWatch

Alle espressioni matematiche dei parametri si applicano le seguenti considerazioni:

- Puoi interrogare qualsiasi metrica disponibile. CloudWatch Ogni parametro è una combinazione univoca di nome del parametro, spazio dei nomi e nessuna o più dimensioni.
- È possibile utilizzare qualsiasi operatore aritmetico (+ - */^), funzione statistica (come AVG o SUM) o altra funzione che supporti. CloudWatch
- È possibile utilizzare i parametri e i risultati di altre espressioni matematiche nelle formule dell'espressione matematica.
- Qualsiasi espressione utilizzata in una specifica dei parametri deve restituire una singola serie temporale.
- Puoi verificare che un'espressione matematica metrica sia valida utilizzando la console o l' CloudWatch API Data. CloudWatch GetMetric

Argomenti

- Esempio: backlog della coda di Amazon SQS per attività
- Limitazioni

Esempio: backlog della coda di Amazon SQS per attività

Per calcolare il backlog della coda di Amazon SQS per attività, prendi il numero approssimativo di messaggi disponibili per il recupero dalla coda e dividi tale numero per il numero di attività Amazon ECS attive nel servizio. Per ulteriori informazioni, consulta Amazon Elastic Container Service (ECS) Auto Scaling using custom metrics sul AWS Compute Blog.

La logica dell'espressione è questa:

```
sum of (number of messages in the queue)/(number of tasks that are
currently in the RUNNING state)
```

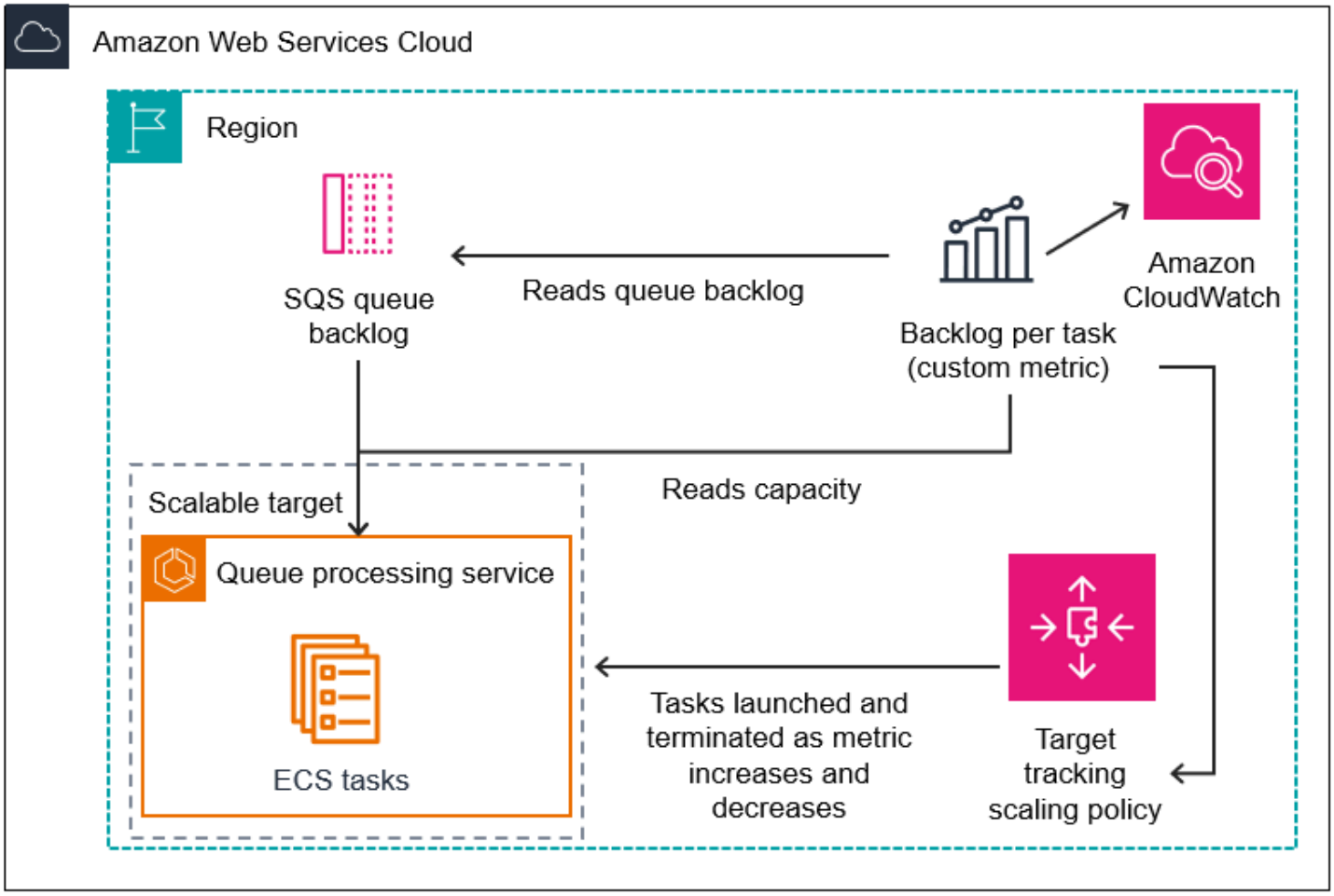
Quindi le informazioni sui CloudWatch parametri sono le seguenti.

ID	CloudWatch metrico	Statistic	Periodo
m1	ApproximateNumberOfMessagesVisibile	Somma	1 minuto
m2	RunningTaskConta	Media	1 minuto

L'ID dell'operazione matematica sui parametri e l'espressione sono i seguenti.

ID	Expression
e1	`(m1)/(m2)`

Il diagramma seguente illustra l'architettura di questa metrica:



Per utilizzare questa matematica dei parametri al fine di creare una policy di dimensionamento con monitoraggio degli obiettivi (AWS CLI)

1. Memorizza l'espressione matematica dei parametri come parte di un specifico parametro personalizzato in un file JSON denominato `config.json`.

   Utilizza la tabella seguente come guida. Sostituisci ciascun *`placeholder input dell'utente`* con le tue informazioni.

```
{
    "CustomizedMetricSpecification": {
        "Metrics": [
            {
                "Label": "Get the queue size (the number of messages waiting to be
 processed)",
                "Id": "m1",
                "MetricStat": {
```

```
                    "Metric": {
                        "MetricName": "ApproximateNumberOfMessagesVisible",
                        "Namespace": "AWS/SQS",
                        "Dimensions": [
                            {
                                "Name": "QueueName",
                                "Value": "my-queue"
                            }
                        ]
                    },
                    "Stat": "Sum"
                },
                "ReturnData": false
            },
            {
                "Label": "Get the ECS running task count (the number of currently
running tasks)",
                "Id": "m2",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "RunningTaskCount",
                        "Namespace": "ECS/ContainerInsights",
                        "Dimensions": [
                            {
                                "Name": "ClusterName",
                                "Value": "my-cluster"
                            },
                            {
                                "Name": "ServiceName",
                                "Value": "my-service"
                            }
                        ]
                    },
                    "Stat": "Average"
                },
                "ReturnData": false
            },
            {
                "Label": "Calculate the backlog per instance",
                "Id": "e1",
                "Expression": "m1 / m2",
                "ReturnData": true
            }
        ]
```

```
    },
    "TargetValue": 100
}
```

Per ulteriori informazioni, consulta TargetTrackingScalingPolicyConfigurazione nell'Application Auto Scaling API Reference.

> **Note**
>
> Di seguito sono riportate alcune risorse aggiuntive che possono aiutarti a trovare nomi di metriche, namespace, dimensioni e statistiche per le metriche: CloudWatch
>
> - Per informazioni sui parametri disponibili per AWS i servizi, consulta i AWS servizi che pubblicano CloudWatch metriche nella Amazon CloudWatch User Guide.
> - Per ottenere il nome esatto della metrica, lo spazio dei nomi e le dimensioni (se applicabili) di una CloudWatch metrica con, consulta list-metrics. AWS CLI

2. Per creare questa policy, esegui il comando put-scaling-policy utilizzando il file JSON come input, come mostrato nel seguente esempio.

```
aws application-autoscaling put-scaling-policy --policy-name sqs-backlog-target-tracking-scaling-policy \
  --service-namespace ecs --scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service \
  --policy-type TargetTrackingScaling --target-tracking-scaling-policy-configuration file://config.json
```

In caso di successo, questo comando restituisce l'Amazon Resource Name (ARN) della policy e gli ARN dei due CloudWatch allarmi creati per tuo conto.

```
{
    "PolicyARN": "arn:aws:autoscaling:us-west-2:012345678910:scalingPolicy:
 8784a896-b2ba-47a1-b08c-27301cc499a1:resource/ecs/service/my-cluster/my-
service:policyName/sqs-backlog-target-tracking-scaling-policy",
    "Alarms": [
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:012345678910:alarm:TargetTracking-service/my-cluster/my-service-
AlarmHigh-9bc77b56-0571-4276-ba0f-d4178882e0a0",
```

```
            "AlarmName": "TargetTracking-service/my-cluster/my-service-
AlarmHigh-9bc77b56-0571-4276-ba0f-d4178882e0a0"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:012345678910:alarm:TargetTracking-service/my-cluster/my-service-
AlarmLow-9b6ad934-6d37-438e-9e05-02836ddcbdc4",
            "AlarmName": "TargetTracking-service/my-cluster/my-service-
AlarmLow-9b6ad934-6d37-438e-9e05-02836ddcbdc4"
        }
    ]
}
```

> **Note**
>
> Se questo comando genera un errore, assicurati di averlo aggiornato AWS CLI localmente alla versione più recente.

Limitazioni

- La dimensione massima della richiesta è 50 KB. Questa è la dimensione totale del payload per la richiesta PutScalingPolicy API quando si utilizza la matematica metrica nella definizione della politica. Se si supera questo limite, l'Applicazione di Dimensionamento automatico rifiuta la richiesta.
- I seguenti servizi non sono supportati quando si usa la matematica dei parametri con le policy di dimensionamento con monitoraggio degli obiettivi:
  - Amazon Keyspaces (per Apache Cassandra)
  - DynamoDB
  - Amazon EMR
  - MSK Amazon
  - Amazon Neptune

Policy di dimensionamento per fasi per Application Auto Scaling.

Una politica di scalabilità graduale ridimensiona la capacità dell'applicazione in incrementi predefiniti in base agli allarmi. CloudWatch È possibile definire policy di dimensionamento separate per gestire il dimensionamento orizzontale (aumento della capacità) e il ridimensionamento (riduzione della capacità) in caso di superamento di una soglia di allarme.

Con le politiche di scalabilità graduale, puoi creare e gestire gli CloudWatch allarmi che richiamano il processo di scalabilità. Quando viene violato un allarme, Application Auto Scaling avvia la policy di dimensionamento associata a tale allarme.

La policy di dimensionamento graduale ridimensiona la capacità utilizzando una serie di regolazioni, note come regolazioni della fase. La portata della regolazione varia in base alla dimensione dell'utilizzo fuori limite segnalato dall'allarme.

- Se la violazione supera la prima soglia, Application Auto Scaling applicherà la prima modifica.
- Se la violazione supera la seconda soglia, Application Auto Scaling applicherà la seconda modifica.

Ciò consente alla policy di dimensionamento di rispondere in modo appropriato a modifiche minori e importanti nella metrica degli allarmi.

La policy continuerà a rispondere a ulteriori violazioni degli allarmi mentre è in corso l'attività di dimensionamento. Ciò significa che Application Auto Scaling valuterà tutte le violazioni degli allarmi man mano che si verificano. Viene utilizzato un tempo di raffreddamento per proteggere dal sovradimensionamento dovuto a molteplici violazioni degli allarmi che si verificano in rapida successione.

Come il monitoraggio degli obiettivi, il dimensionamento dei passaggi può aiutare a dimensionare automaticamente la capacità dell'applicazione in base alle variazioni del traffico. Tuttavia, le policy di monitoraggio degli obiettivi tendono ad essere più facili da implementare e gestire per esigenze di dimensionamento costanti.

È possibile utilizzare policy di dimensionamento graduale con i seguenti obiettivi scalabili:

- AppStream Flotte 2.0
- Cluster di database Aurora

- Servizi ECS
- Cluster EMR
- SageMaker varianti di endpoint
- SageMaker componenti di inferenza
- SageMaker Concorrenza fornita senza server
- Parco istanze Spot
- Risorse personalizzate

Indice

Come funziona la scalabilità a gradini per Application Auto Scaling

Questo argomento descrive come funziona la scalabilità dei gradini e introduce gli elementi chiave di una politica di scalabilità dei gradini.

Indice

Come funziona

Per utilizzare lo step scaling, crei un CloudWatch allarme che monitora una metrica per il tuo obiettivo scalabile. Definisci il parametro, il valore di soglia e il numero di periodi di valutazione che

determinano una violazione dell'allarme. È inoltre necessario creare una policy di dimensionamento graduale che definisca come dimensionare la capacità in caso di superamento della soglia di allarme e come associarla al target scalabile.

Aggiungi le modifiche ai passaggi nella policy. È possibile definire diverse regolazioni delle fasi in base alla dimensione dell'allarme di violazione. Per esempio:

- Se il parametro di allarme raggiunge il 60 percento, è possibile impiegare la scalabilità orizzontale di 10 unità di capacità
- Se il parametro di allarme raggiunge il 75 percento, è possibile impiegare la scalabilità orizzontale di 30 unità di capacità
- Se il parametro di allarme raggiunge l'85 percento, è possibile impiegare la scalabilità orizzontale di 40 unità di capacità

Quando la soglia di allarme viene superata per il numero specificato di periodi di valutazione, Application Auto Scaling applicherà le regolazioni della fase definite nella policy. Le regolazioni possono continuare in caso di ulteriori violazioni degli allarmi fino al ripristino dello stato di allarme su 0K.

Le attività di dimensionamento vengono eseguite con periodi di raffreddamento intermedi per evitare rapide fluttuazioni della capacità. Facoltativamente, puoi configurare i tempi di raffreddamento per la tua policy di dimensionamento.

Adeguamenti per fasi

Quando si crea una policy di dimensionamento a fasi, è possibile specificare una o più regolazioni per fasi che dimensionano automaticamente la capacità di istanze della destinazione in modo dinamico in base all'utilizzo fuori limite segnalato dall'allarme. Ogni regolazione per fasi specifica quanto segue:

- Un limite inferiore per il valore del parametro
- Un limite superiore per il valore del parametro
- La quantità da dimensionare, in base al tipo di regolazione del dimensionamento

CloudWatch aggrega i punti dati metrici in base alla statistica della metrica associata all'allarme. CloudWatch Quando la soglia dell'allarme viene superata, viene richiamata la policy di dimensionamento appropriata. Application Auto Scaling applica il tipo di aggregazione specificato ai punti dati metrici più recenti di CloudWatch (anziché ai dati metrici grezzi). Confronta il valore

dei parametri aggregati con i limiti superiore e inferiore definiti dagli adeguamenti delle fasi per determinare quali di queste eseguire.

Devi specificare i limiti superiore e inferiore relativi alla soglia dell'utilizzo fuori limite. Ad esempio, supponiamo che tu abbia creato un CloudWatch allarme e una politica di scalabilità orizzontale per quando la metrica è superiore al 50 percento. Hai quindi creato un secondo allarme e una policy di riduzione orizzontale per quando il parametro è inferiore al 50 percento. Hai apportato un insieme di regolazioni della fase con un tipo di regolazione di `PercentChangeInCapacity` per ogni policy:

Esempio: Regolazioni per fasi per la policy di aumento orizzontale

Limite inferiore	Limite superiore	Regolazione
0	10	0
10	20	10
20	null	30

Esempio: Regolazioni per fasi della policy di riduzione orizzontale

Limite inferiore	Limite superiore	Regolazione
-10	0	0
-20	-10	-10
null	-20	-30

Questo crea la seguente configurazione di dimensionamento.

```
Metric value

-infinity          30%    40%          60%     70%             infinity
-------------------------------------------------------------------------
          -30%      | -10% | Unchanged  | +10%  |       +30%
-------------------------------------------------------------------------
```

Ad esempio, supponiamo che tu abbia una configurazione scalabile con capacità di 10. I seguenti punti riassumono il comportamento della configurazione di dimensionamento in relazione alla capacità della destinazione dimensionabile:

- La capacità originale viene mantenuta, mentre il valore del parametro aggregato è maggiore di 40 e minore di 60.
- Se il valore del parametro raggiunge 60, il dimensionamento automatico dell'applicazione aumenta la capacità desiderata dell'obiettivo dimensionabile di 1, portandola a 11. Questo avviene in base al secondo adeguamento delle policy di scalabilità orizzontale (aggiungere 10% di 10). Una volta aggiunta la nuova capacità, Application Auto Scaling aumenta la capacità corrente a 11. Se il valore del parametro aumenta a 70 anche dopo questo incremento di capacità, Application Auto Scaling aumenta la capacità richiesta di 3, portandola a 14. Questo avviene in base al terzo adeguamento delle policy di scalabilità orizzontale (aggiungere 30% di 11, 3,3, arrotondato a 3).
- Se il valore del parametro raggiunge 40, il dimensionamento automatico dell'applicazione riduce la capacità della destinazione dimensionabile di 1, portandola a 13, in base al secondo adeguamento per fasi della policy di ridimensionamento (togliere il 10% di 14, ossia 1,4, arrotondato a 1). Se il valore del parametro scende a 30 anche dopo questa riduzione di capacità, Application Auto Scaling riduce la capacità dell'obiettivo di 3, portandola a 10, in base al terzo adeguamento per fasi della policy di riduzione (togliere il 30% di 13, ossia 3,9, arrotondato a 3).

Quando specifichi le regolazioni delle fasi per la policy di dimensionamento, tieni presente quanto segue:

- Gli intervalli delle regolazioni delle fasi non possono essere sovrapporsi o presentare scarti.
- Solo una regolazione delle fasi può avere un limite inferiore nullo (infinito negativo). Se una regolazione delle fasi presenta un limite inferiore negativo, allora deve esistere una regolazione con un limite inferiore nullo.
- Solo una regolazione delle fasi può avere un limite superiore nullo (infinito positivo). Se una regolazione delle fasi presenta un limite superiore positivo, deve esistere anche una regolazione con un limite superiore nullo.
- I limiti superiore e inferiore non possono essere nulli nella stessa regolazione delle fasi.
- Se il valore del parametro supera la soglia dell'utilizzo fuori limite, il limite inferiore è incluso e quello superiore è escluso. Se il valore del parametro è inferiore alla soglia dell'utilizzo fuori limite, il limite inferiore è escluso e il limite superiore è incluso.

Tipi di regolazioni per il dimensionamento

È possibile definire una policy di dimensionamento che esegue l'operazione di dimensionamento ottimale, in base al tipo di regolazione scelta. È possibile specificare il tipo di adeguamento come percentuale della capacità corrente di target scalabile o in numeri assoluti.

Application Auto Scaling supporta i seguenti tipi di adeguamenti per le policy di dimensionamento per fasi:

- ChangeInCapacità: aumenta o diminuisce la capacità corrente del target scalabile in base al valore specificato. Un valore positivo incrementa la capacità, mentre un valore negativo la decrementa. Esempio: se la capacità corrente è 3 e l'adeguamento è 5, Application Auto Scaling aggiunge 5 alla capacità, per un totale di 8.
- ExactCapacity—Modifica la capacità corrente del target scalabile portandolo al valore specificato. Con questo tipo di adeguamento, specifica un valore non negativo. Esempio: se la capacità corrente è 3 e l'adeguamento è 5, Application Auto Scaling modifica la capacità in 5.
- PercentChangeInCapacity—Aumenta o diminuisce la capacità corrente del target scalabile della percentuale specificata. Un valore positivo incrementa la capacità, mentre un valore negativo la decrementa. Esempio: se la capacità corrente è 10 e l'adeguamento è del 10%, Application Auto Scaling aggiunge 1 alla capacità, per un totale di 11.

> **Note**
>
> Se il valore risultante non è un numero intero, Application Auto Scaling lo arrotonda come segue:
>
> - I valori superiori a 1 vengono arrotondati per difetto. Ad esempio, `12.7` viene arrotondato in `12`.
> - I valori tra 0 e 1 vengono arrotondati a 1. Ad esempio, `.67` viene arrotondato in `1`.
> - I valori tra 0 e -1 vengono arrotondati a -1. Ad esempio, `-.58` viene arrotondato in `-1`.
> - I valori inferiori a -1 vengono arrotondati per eccesso. Ad esempio, `-6.67` viene arrotondato in `-6`.

Con PercentChangeInCapacity, è anche possibile specificare la quantità minima da scalare utilizzando il `MinAdjustmentMagnitude` parametro. Ad esempio, supponi di creare una policy che aggiunge 25 percento e di specificare una quantità minima di 2. Se il target scalabile ha una

capacità di 4 e viene eseguita la policy di dimensionamento, il 25 percento di 4 è 1. Tuttavia, poiché hai specificato un incremento minimo di 2, Application Auto Scaling aggiunge 2.

Periodo di attesa

Facoltativamente, puoi definire un tempo di raffreddamento nella tua policy di dimensionamento per fasi.

Il tempo di raffreddamento indica la quantità di tempo che la policy di dimensionamento attende prima di rendere effettiva una precedente attività di dimensionamento.

Esistono due modi per pianificare l'utilizzo dei tempi di raffreddamento per una configurazione di dimensionamento per fasi:

- Con il tempo di raffreddamento per le policy a scalabilità orizzontale, l'intenzione è di impiegare la scalabilità orizzontale in modo continuo (ma non eccessivo). Dopo che Application Auto Scaling ha impiegato correttamente la scalabilità orizzontale utilizzando una policy di dimensionamento, inizia a calcolare il tempo di raffreddamento. Una policy di dimensionamento non aumenta di nuovo la capacità desiderata, a meno che non venga impiegata una scalabilità orizzontale maggiore o che il tempo di raffreddamento finisca. Mentre è attivo il periodo di attesa di incremento, la capacità aggiunta dall'attività di incremento iniziale viene calcolata come parte della capacità desiderata per il successivo evento di incremento.
- Con il tempo di raffreddamento per le policy di ridimensionamento, l'intenzione è di ridimensionare in modo conservativo per proteggere la disponibilità dell'applicazione, per cui le attività di ridimensionamento vengono bloccate fino alla scadenza del tempo di raffreddamento per il ridimensionamento. Tuttavia, se un altro allarme attiva un'attività di incremento durante il periodo di attesa, Application Auto Scaling incrementa immediatamente la destinazione. In questo caso, il tempo di raffreddamento per il ridimensionamento si interrompe e non viene completato.

Ad esempio, quando si verifica un picco di traffico, viene attivato un allarme e Application Auto Scaling aggiunge automaticamente la capacità per aiutare a gestire l'aumento del carico. Se imposti un tempo di raffreddamento per la policy di scalabilità orizzontale, quando l'allarme attiva la policy per incrementare la capacità di 2, l'attività di dimensionamento viene completata e inizia il tempo di raffreddamento per la scalabilità orizzontale. Se un allarme si attiva nuovamente durante il tempo di raffreddamento, ma a un adeguamento di fasi maggiore, ad esempio di 3, l'aumento precedente di 2 è considerato parte della capacità attuale. Pertanto, solo 1 viene aggiunto alla capacità. Ciò

consente un dimensionamento più rapido rispetto all'attesa della scadenza del raffreddamento, ma senza aggiungere più capacità del necessario.

Il periodo di attesa viene misurato in secondi e si applica solo alle attività di dimensionamento correlate alle policy di dimensionamento. Durante un periodo di attesa, quando un'operazione pianificata inizia all'ora pianificata, può attivare immediatamente un'attività di dimensionamento senza attendere la scadenza del periodo di attesa.

Se non viene specificato alcun valore, quello predefinito è 300.

Comandi comunemente utilizzati per la creazione, la gestione e l'eliminazione delle policy di dimensionamento

I comandi comunemente utilizzati per le policy di dimensionamento includono:

- register-scalable-target per registrare AWS o personalizzare le risorse come destinazioni scalabili (una risorsa scalabile da Application Auto Scaling) e per sospendere e riprendere il ridimensionamento.
- put-scaling-policy per aggiungere o modificare policy di dimensionamento per un obiettivo scalabile esistente.
- describe-scaling-activities per restituire informazioni sulle attività di dimensionamento in una Regione AWS .
- describe-scaling-policies per restituire informazioni sulle policy di dimensionamento in una Regione AWS .
- delete-scaling-policy per eliminare una policy di dimensionamento.

Considerazioni

Le seguenti considerazioni si applicano quando si usano le policy di dimensionamento a fasi:

- Valuta se riesci a prevedere le regolazioni dei passaggi sull'applicazione in modo sufficientemente accurato da utilizzare il dimensionamento a fasi. Se il parametro di dimensionamento aumenta o diminuisce in proporzione alla capacità della destinazione scalabile, ti consigliamo di utilizzare una policy di dimensionamento di monitoraggio della destinazione. È comunque possibile utilizzare il dimensionamento per fasi come policy aggiuntiva per una configurazione più avanzata. Ad esempio, puoi configurare una risposta più aggressiva quando l'utilizzo raggiunge un determinato livello.

- Assicurati di scegliere un margine adeguato tra le soglie di scalabilità orizzontale e ridimensionamento per evitare che si verifichino sbalzi. Flapping è un ciclo infinito di riduzione e aumento orizzontale. Ciò significa che se viene eseguita un'azione di dimensionamento, il valore della metrica cambierebbe per avviare un'altra azione di dimensionamento nella direzione opposta.

Risorse correlate

Per ulteriori informazioni su come creare policy di dimensionamento per i gruppi con dimensionamento automatico, consulta Policy di dimensionamento semplici e a fasi per Dimensionamento automatico Amazon EC2 nella Guida per l'utente di Dimensionamento automatico Amazon EC2.

Limitazioni

- L'accesso da console per visualizzare, aggiungere, aggiornare o rimuovere policy di dimensionamento su risorse scalabili dipende dalla risorsa utilizzata. Per ulteriori informazioni, consulta Servizi AWS che puoi usare con Application Auto Scaling.

Creare una politica di scalabilità graduale per Application Auto Scaling utilizzando AWS CLI

È possibile creare una politica di scalabilità a fasi per Application Auto Scaling utilizzando per AWS CLI le seguenti attività di configurazione.

1. Registrazione di una destinazione dimensionabile.
2. Aggiungi una policy di dimensionamento a fasi sulla destinazione dimensionabile.
3. Crea un CloudWatch allarme per la politica.

Per brevità, gli esempi di questo argomento illustrano i comandi della CLI per un servizio Amazon ECS. Per specificare un altro target scalabile, specificare il suo spazio dei nomi in `--service-namespace`, la sua dimensione scalabile `--scalable-dimension` e l'ID di risorsa in `--resource-id`. Per maggiori informazioni ed esempi per ogni servizio, consultare gli argomenti in Servizi AWS che puoi usare con Application Auto Scaling.

Quando usi il AWS CLI, ricorda che i comandi vengono eseguiti nella Regione AWS configurazione per il tuo profilo. Per eseguire i comandi in un'altra regione, modificare la regione predefinita per il profilo oppure utilizzare il parametro `--region` con il comando.

Indice

Registrazione di un target scalabile

Se non lo hai ancora fatto, registra l'obiettivo scalabile. Utilizza il comando register-scalable-target per registrare una risorsa specifica nel servizio obiettivo come obiettivo scalabile. Nell'esempio seguente viene registrato un servizio Amazon ECS con Application Auto Scaling. Application Auto Scaling può dimensionare il numero di attività a un minimo di 2 e un massimo di 10 attività. Sostituisci ciascun *`placeholder input dell'utente`* con le tue informazioni.

Linux, macOS o Unix

```
aws application-autoscaling register-scalable-target --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --min-capacity 2 --max-capacity 10
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace ecs --
scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
 --min-capacity 2 --max-capacity 10
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Creazione di una policy di dimensionamento per fasi

Per creare una politica di scalabilità graduale per il tuo target scalabile, puoi utilizzare i seguenti esempi per iniziare.

Scale out

Per creare una politica di scalabilità graduale per la scalabilità orizzontale (aumento della capacità)

1. Utilizzate il `cat` comando seguente per memorizzare una configurazione della politica di scalabilità dei passaggi in un file JSON denominato `config.json` nella vostra home directory. Di seguito è riportato un esempio di configurazione con un tipo di regolazione `PercentChangeInCapacity` che aumenta la capacità del target scalabile in base alle seguenti regolazioni della fase (presupponendo una soglia di CloudWatch allarme di 70):

   - Aumenta la capacità del 10 percento quando il valore della metrica è maggiore o uguale a 70 ma inferiore a 85
   - Aumenta la capacità del 20 percento quando il valore della metrica è maggiore o uguale a 85 ma inferiore a 95
   - Aumenta la capacità del 30 percento quando il valore della metrica è maggiore o uguale a 95

```
$ cat ~/config.json
{
  "AdjustmentType": "PercentChangeInCapacity",
  "MetricAggregationType": "Average",
  "Cooldown": 60,
  "MinAdjustmentMagnitude": 1,
  "StepAdjustments": [
    {
      "MetricIntervalLowerBound": 0.0,
      "MetricIntervalUpperBound": 15.0,
      "ScalingAdjustment": 10
    },
    {
      "MetricIntervalLowerBound": 15.0,
      "MetricIntervalUpperBound": 25.0,
      "ScalingAdjustment": 20
    },
```

```
    {
      "MetricIntervalLowerBound": 25.0,
      "ScalingAdjustment": 30
    }
  ]
}
```

Per ulteriori informazioni, consulta StepScalingPolicyConfigurationl'Application Auto Scaling API Reference.

2. Utilizza il comando put-scaling-policy insieme al file `config.json` che hai creato per generare una policy di dimensionamento denominata `my-step-scaling-policy`.

   Linux, macOS o Unix

   ```
   aws application-autoscaling put-scaling-policy --service-namespace ecs \
     --scalable-dimension ecs:service:DesiredCount \
     --resource-id service/my-cluster/my-service \
     --policy-name my-step-scaling-policy --policy-type StepScaling \
     --step-scaling-policy-configuration file://config.json
   ```

   Windows

   ```
   aws application-autoscaling put-scaling-policy --service-namespace ecs --
   scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-
   service --policy-name my-step-scaling-policy --policy-type StepScaling --step-
   scaling-policy-configuration file://config.json
   ```

   L'output include l'ARN utilizzato come nome univoco per la policy. Ne hai bisogno per creare un CloudWatch allarme per la tua politica.

   ```
   {
       "PolicyARN":
    "arn:aws:autoscaling:region:123456789012:scalingPolicy:ac542982-cbeb-4294-891c-
   a5a941dfa787:resource/ecs/service/my-cluster/my-service:policyName/my-step-
   scaling-policy"
   }
   ```

Scale in

Per creare una politica di scalabilità graduale per la scalabilità in entrata (riduzione della capacità)

1. Utilizzate il `cat` comando seguente per memorizzare una configurazione della politica di scalabilità dei passaggi in un file JSON denominato `config.json` nella vostra home directory. Di seguito è riportato un esempio di configurazione con un tipo di regolazione `ChangeInCapacity` che riduce la capacità del target scalabile in base alle seguenti regolazioni della fase (presupponendo una soglia di CloudWatch allarme di 50):

   - Riduci la capacità di 1 quando il valore della metrica è inferiore o uguale a 50 ma maggiore di 40
   - Riduci la capacità del 2% quando il valore della metrica è inferiore o uguale a 40 ma maggiore di 30
   - Riduci la capacità del 3% quando il valore della metrica è inferiore o uguale a 30

```
$ cat ~/config.json
{
  "AdjustmentType": "ChangeInCapacity",
  "MetricAggregationType": "Average",
  "Cooldown": 60,
  "StepAdjustments": [
    {
      "MetricIntervalUpperBound": 0.0,
      "MetricIntervalLowerBound": -10.0,
      "ScalingAdjustment": -1
    },
    {
      "MetricIntervalUpperBound": -10.0,
      "MetricIntervalLowerBound": -20.0,
      "ScalingAdjustment": -2
    },
    {
      "MetricIntervalUpperBound": -20.0,
      "ScalingAdjustment": -3
    }
  ]
}
```

Per ulteriori informazioni, consulta StepScalingPolicyConfigurationl'Application Auto Scaling API Reference.

2. Utilizza il comando put-scaling-policy insieme al file `config.json` che hai creato per generare una policy di dimensionamento denominata `my-step-scaling-policy`.

   Linux, macOS o Unix

   ```
   aws application-autoscaling put-scaling-policy --service-namespace ecs \
     --scalable-dimension ecs:service:DesiredCount \
     --resource-id service/my-cluster/my-service \
     --policy-name my-step-scaling-policy --policy-type StepScaling \
     --step-scaling-policy-configuration file://config.json
   ```

   Windows

   ```
   aws application-autoscaling put-scaling-policy --service-namespace ecs --
   scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-
   service --policy-name my-step-scaling-policy --policy-type StepScaling --step-
   scaling-policy-configuration file://config.json
   ```

   L'output include l'ARN utilizzato come nome univoco per la policy. Ne hai bisogno per creare un CloudWatch allarme per la tua politica.

   ```
   {
       "PolicyARN":
    "arn:aws:autoscaling:region:123456789012:scalingPolicy:ac542982-cbeb-4294-891c-
   a5a941dfa787:resource/ecs/service/my-cluster/my-service:policyName/my-step-
   scaling-policy"
   }
   ```

Creazione di un allarme che richiami la policy di dimensionamento

Infine, usa il seguente comando CloudWatch put-metric-alarm per creare un allarme da utilizzare con la tua politica di scalabilità dei passaggi. In questo esempio, si dispone di un allarme basato sull'utilizzo medio della CPU. L'allarme viene configurato per essere in stato ALARM se raggiunge una soglia del 70% per almeno due periodi di valutazione consecutivi di 60 secondi. Per specificare

una metrica diversa o utilizzare una CloudWatch metrica personalizzata, specificane il nome in e il relativo spazio dei nomi in. `--metric-name` `--namespace`

Linux, macOS o Unix

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmHigh-ECS:service/my-
cluster/my-service \
  --metric-name CPUUtilization --namespace AWS/ECS --statistic Average \
  --period 60 --evaluation-periods 2 --threshold 70 \
  --comparison-operator GreaterThanOrEqualToThreshold \
  --dimensions Name=ClusterName,Value=default Name=ServiceName,Value=sample-app-service
 \
  --alarm-actions PolicyARN
```

Windows

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmHigh-ECS:service/
my-cluster/my-service --metric-name CPUUtilization --namespace AWS/ECS --statistic
 Average --period 60 --evaluation-periods 2 --threshold 70 --comparison-operator
 GreaterThanOrEqualToThreshold --dimensions Name=ClusterName,Value=default
 Name=ServiceName,Value=sample-app-service --alarm-actions PolicyARN
```

Descrizione delle policy di dimensionamento per fasi

È possibile descrivere tutte le policy di dimensionamento per lo spazio dei nomi dei servizi specificato utilizzando il comando describe-scaling-policies.

```
aws application-autoscaling describe-scaling-policies --service-namespace ecs
```

È possibile filtrare i risultati solo per le policy di dimensionamento per fasi utilizzando il parametro `--query`. Per ulteriori informazioni sulla sintassi per `query`, consulta Controllo dell'output del comando dalla AWS CLI nella Guida per l'utente di AWS Command Line Interface .

Linux, macOS o Unix

```
aws application-autoscaling describe-scaling-policies --service-namespace ecs \
  --query 'ScalingPolicies[?PolicyType==`StepScaling`]'
```

Windows

```
aws application-autoscaling describe-scaling-policies --service-namespace ecs --query
 "ScalingPolicies[?PolicyType==`StepScaling`]"
```

Di seguito è riportato un output di esempio.

```
[
    {
        "PolicyARN": "PolicyARN",
        "StepScalingPolicyConfiguration": {
            "MetricAggregationType": "Average",
            "Cooldown": 60,
            "StepAdjustments": [
                {
                    "MetricIntervalLowerBound": 0.0,
                    "MetricIntervalUpperBound": 15.0,
                    "ScalingAdjustment": 1
                },
                {
                    "MetricIntervalLowerBound": 15.0,
                    "MetricIntervalUpperBound": 25.0,
                    "ScalingAdjustment": 2
                },
                {
                    "MetricIntervalLowerBound": 25.0,
                    "ScalingAdjustment": 3
                }
            ],
            "AdjustmentType": "ChangeInCapacity"
        },
        "PolicyType": "StepScaling",
        "ResourceId": "service/my-cluster/my-service",
        "ServiceNamespace": "ecs",
        "Alarms": [
            {
                "AlarmName": "Step-Scaling-AlarmHigh-ECS:service/my-cluster/my-
service",
                "AlarmARN": "arn:aws:cloudwatch:region:012345678910:alarm:Step-Scaling-
AlarmHigh-ECS:service/my-cluster/my-service"
            }
        ],
        "PolicyName": "my-step-scaling-policy",
        "ScalableDimension": "ecs:service:DesiredCount",
        "CreationTime": 1515024099.901
```

```
    }
]
```

Eliminazione di una policy di dimensionamento per fasi

Quando una policy di dimensionamento per fasi non è più necessaria, puoi eliminarla. Per eliminare sia la politica di ridimensionamento che l' CloudWatch allarme, completa le seguenti attività.

Per eliminare la policy di dimensionamento

Utilizza il comando delete-scaling-policy.

Linux, macOS o Unix

```
aws application-autoscaling delete-scaling-policy --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --policy-name my-step-scaling-policy
```

Windows

```
aws application-autoscaling delete-scaling-policy --service-namespace ecs --scalable-
dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service --
policy-name my-step-scaling-policy
```

Per eliminare l'allarme CloudWatch

Utilizza il comando delete-alarms. Puoi eliminare uno o più allarmi alla volta. Ad esempio, utilizzare il seguente comando per eliminare gli allarmi `Step-Scaling-AlarmHigh-ECS:service/my-cluster/my-service` e `Step-Scaling-AlarmLow-ECS:service/my-cluster/my-service`.

```
aws cloudwatch delete-alarms --alarm-name Step-Scaling-AlarmHigh-ECS:service/my-
cluster/my-service Step-Scaling-AlarmLow-ECS:service/my-cluster/my-service
```

Tutorial: configura il dimensionamento automatico per gestire un carico di lavoro pesante

> **Important**
>
> Prima di esplorare questo tutorial, ti suggeriamo di esaminare il seguente tutorial introduttivo: Tutorial: Guida di base su dimensionamento pianificato utilizzando AWS CLI.

In questo tutorial imparerai ad aumentare o ridurre orizzontalmente in base a finestre temporali quando l'applicazione si troverà a gestire un carico di lavoro più pesante del normale. Ciò sarà di particolare utilità quando si lavora con un'applicazione che improvvisamente si trova a ospitare un gran numero di visitatori, o normalmente oppure su base stagionale.

Per gestire tale carico aggiuntivo, insieme al dimensionamento pianificato, si può utilizzare una policy di dimensionamento con monitoraggio degli obiettivi. Il dimensionamento pianificato avvia in automatico le modifiche ai valori `MinCapacity` e `MaxCapacity` per conto tuo sulla base della pianificazione specificata. Quando una policy di dimensionamento con monitoraggio degli obiettivi è attiva sulla risorsa, essa può dimensionare dinamicamente in base al corrente utilizzo della risorsa nel rispetto del nuovo intervallo che prevede un minimo e un massimo in termini di capacità.

Dopo aver completato il tutorial, saprai come:

- Utilizzare il dimensionamento pianificato per aggiungere l'ulteriore capacità necessaria a gestire un carico di lavoro pesante prima che si manifesti, e successivamente a ridurla quando non serve più.
- Utilizzare una policy di dimensionamento con monitoraggio degli obiettivi per dimensionare l'applicazione in base all'utilizzo corrente della risorsa.

Indice

Prerequisiti

Il tutorial presuppone che tu abbia già:

- Ha creato un Account AWS.
- Installato e configurato il AWS CLI.
- Sono state concesse le autorizzazioni necessarie per registrare e annullare la registrazione delle risorse come obiettivi scalabili con Application Auto Scaling. Inoltre, ha concesso le autorizzazioni necessarie per creare politiche di scalabilità e azioni pianificate. Per ulteriori informazioni, consulta Identity and Access Management per Application Auto Scaling.
- Ha creato una risorsa supportata in un ambiente non di produzione disponibile per questo tutorial. Se non disponi già dell'account, creane uno. Per informazioni sui servizi e risorse AWS che funzionano con Application Auto Scaling, consulta la sezione Servizi AWS che puoi usare con Application Auto Scaling.

> **Note**
>
> Quando avrai completato il tutorial, ci saranno due fasi in cui occorrerà impostare su 0 i valori massimi e minimi di capacità per ripristinare la capacità iniziale a 0. A seconda della risorsa che stai utilizzando con Application Auto Scaling, potrebbe non essere possibile reimpostare la capacità iniziale a 0 durante tali fasi. Per aiutarti a risolvere il problema, un messaggio nell'output indicherà che la capacità minima non può essere inferiore al valore specificato e fornirà il valore di capacità minima che la AWS risorsa può accettare.

Fase 1: registrazione dell'obiettivo scalabile

Inizia registrando la risorsa come obiettivo scalabile tramite Application Auto Scaling. Un obiettivo scalabile è una risorsa la cui dimensione può essere aumentata e ridotta orizzontalmente da Application Auto Scaling.

Per registrare un obiettivo scalabile tramite Application Auto Scaling

- Utilizza il comando register-scalable-target per registrare un nuovo obiettivo scalabile. Imposta i valori `--min-capacity` e `--max-capacity` su 0 per ripristinare la capacità corrente su 0.

Sostituisci il testo di esempio di `--service-namespace` con lo spazio dei nomi del servizio AWS che stai utilizzando con Application Auto Scaling, `--scalable-dimension` con la dimensione scalabile associata alla risorsa che stai registrando e `--resource-id` con un identificatore per la risorsa. Questi valori variano in base alla risorsa utilizzata e alla modalità di costruzione dell'ID della risorsa. Consulta gli argomenti nella sezione Servizi AWS che puoi usare con Application Auto Scaling per ulteriori informazioni. Questi argomenti includono comandi di esempio che mostrano come registrare obiettivi scalabili con Application Auto Scaling.

Linux, macOS o Unix

```
aws application-autoscaling register-scalable-target \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --min-capacity 0 --max-capacity 0
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace namespace
 --scalable-dimension dimension --resource-id identifier --min-capacity 0 --max-
capacity 0
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-
id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Fase 2: impostazione delle operazioni pianificate in base ai requisiti

Puoi utilizzare il comando put-scheduled-action per creare operazioni pianificate che soddisfino le tue esigenze di lavoro. In questo tutorial ci concentreremo su una configurazione che interrompe l'utilizzo di risorse fuori dall'orario di lavoro riducendo la capacità a 0.

Per creare un'operazione pianificata che aumenta orizzontalmente la capacità la mattina

1. Per aumentare orizzontalmente l'obiettivo scalabile, utilizza il comando put-scheduled-action. Includi il parametro `--schedule` prevedendo una pianificazione ricorrente espressa in UTC utilizzando un'espressione cron.

   Nella pianificazione specificata (ogni giorno alle 9:00 AM UTC), Application Auto Scaling aggiornerà i valori `MinCapacity` e `MaxCapacity` nell'intervallo desiderato di 1-5 unità di capacità.

   Linux, macOS o Unix

   ```
   aws application-autoscaling put-scheduled-action \
     --service-namespace namespace \
     --scalable-dimension dimension \
     --resource-id identifier \
     --scheduled-action-name my-first-scheduled-action \
     --schedule "cron(0 9 * * ? *)" \
     --scalable-target-action MinCapacity=1,MaxCapacity=5
   ```

   Windows

   ```
   aws application-autoscaling put-scheduled-action --service-namespace namespace --
   scalable-dimension dimension --resource-id identifier --scheduled-action-name my-
   first-scheduled-action --schedule "cron(0 9 * * ? *)" --scalable-target-action
    MinCapacity=1,MaxCapacity=5
   ```

   Questo comando non restituisce alcun output se va a buon fine.

2. Per confermare l'esistenza dell'operazione pianificata, utilizza il comando describe-scheduled-actions.

   Linux, macOS o Unix

   ```
   aws application-autoscaling describe-scheduled-actions \
     --service-namespace namespace \
     --query 'ScheduledActions[?ResourceId==`identifier`]'
   ```

   Windows

```
aws application-autoscaling describe-scheduled-actions --service-
namespace namespace --query "ScheduledActions[?ResourceId==`identifier`]"
```

Di seguito è riportato un output di esempio.

```
[
    {
        "ScheduledActionName": "my-first-scheduled-action",
        "ScheduledActionARN": "arn",
        "Schedule": "cron(0 9 * * ? *)",
        "ScalableTargetAction": {
            "MinCapacity": 1,
            "MaxCapacity": 5
        },
        ...
    }
]
```

Per creare un'operazione pianificata che riduce orizzontalmente la capacità durante la notte

1. Ripeti la procedura precedente per creare un'altra operazione pianificata che Application Auto Scaling utilizzerà per diminuire orizzontalmente la capacità al termine della giornata.

   Con la pianificazione specificata (ogni giorno alle 20:00 UTC), Application Auto Scaling aggiornerà i valori `MinCapacity` e `MaxCapacity` portandoli su 0, come richiesto dal comando put-scheduled-action.

   Linux, macOS o Unix

   ```
   aws application-autoscaling put-scheduled-action \
     --service-namespace namespace \
     --scalable-dimension dimension \
     --resource-id identifier \
     --scheduled-action-name my-second-scheduled-action \
     --schedule "cron(0 20 * * ? *)" \
     --scalable-target-action MinCapacity=0,MaxCapacity=0
   ```

   Windows

```
aws application-autoscaling put-scheduled-action --service-namespace namespace --
scalable-dimension dimension --resource-id identifier --scheduled-action-name my-
second-scheduled-action --schedule "cron(0 20 * * ? *)" --scalable-target-action
 MinCapacity=0,MaxCapacity=0
```

2. Per confermare l'esistenza dell'operazione pianificata, utilizza il comando describe-scheduled-actions.

   Linux, macOS o Unix

```
aws application-autoscaling describe-scheduled-actions \
  --service-namespace namespace \
  --query 'ScheduledActions[?ResourceId==`identifier`]'
```

   Windows

```
aws application-autoscaling describe-scheduled-actions --service-
namespace namespace --query "ScheduledActions[?ResourceId==`identifier`]"
```

   Di seguito è riportato un output di esempio.

```
[
    {
        "ScheduledActionName": "my-first-scheduled-action",
        "ScheduledActionARN": "arn",
        "Schedule": "cron(0 9 * * ? *)",
        "ScalableTargetAction": {
            "MinCapacity": 1,
            "MaxCapacity": 5
        },
        ...
    },
    {
        "ScheduledActionName": "my-second-scheduled-action",
        "ScheduledActionARN": "arn",
        "Schedule": "cron(0 20 * * ? *)",
        "ScalableTargetAction": {
            "MinCapacity": 0,
            "MaxCapacity": 0
        },
        ...
```

```
    }
]
```

Fase 3: creazione di una policy di dimensionamento con monitoraggio degli obiettivi

Ora che esiste una pianificazione di base, aggiungi una policy di dimensionamento con monitoraggio degli obiettivi da dimensionare in base al corrente utilizzo della risorsa.

Tramite il monitoraggio degli obiettivi, Application Auto Scaling confronta il valore dell'obiettivo nella policy con quello corrente del parametro specificato. Se non sono uguali per un determinato periodo di tempo, Application Auto Scaling aggiunge o rimuove capacità per mantenere le prestazioni costanti. Man mano che il carico di lavoro sull'applicazione e il valore del parametro aumentano, Application Auto Scaling aggiunge capacità il più velocemente possibile senza eccedere `MaxCapacity`. Quando invece Application Auto Scaling rimuove capacità perché il carico di lavoro è minimo, esegue l'operazione senza andare sotto `MinCapacity`. Adeguando la capacità in base all'utilizzo pagherai soltanto in base alle reali esigenze dell'applicazione.

Se il parametro dispone di dati insufficienti perché l'applicazione non ha alcun carico di lavoro, Application Auto Scaling non aggiunge né rimuove capacità. In altre parole, Application Auto Scaling dà priorità alla disponibilità in situazioni in cui non è presente un numero sufficiente di informazioni.

Potrai aggiungere più policy di dimensionamento, ma attenzione a non aggiungerne una che vada in conflitto con qualche fase, altrimenti potrebbero manifestarsi comportamenti indesiderati. Per esempio, se la policy di dimensionamento per fasi avvia un'attività di riduzione orizzontale prima che la policy di monitoraggio degli obiettivi sia pronta per eseguirla, l'attività di riduzione orizzontale non verrà bloccata. Al termine dell'attività di riduzione orizzontale, la policy di monitoraggio degli obiettivi potrebbe richiedere ad Application Auto Scaling di aumentare orizzontalmente di nuovo la capacità.

Per creare una policy di dimensionamento con monitoraggio degli obiettivi

1. Utilizza il comando put-scaling-policy per creare la policy.

   Le metriche utilizzate più frequentemente per il tracciamento degli obiettivi sono predefinite e possono essere utilizzate senza fornire le specifiche complete della metrica di. CloudWatch Per ulteriori informazioni sui parametri disponibili predefiniti consulta Policy di dimensionamento con monitoraggio degli obiettivi per Application Auto Scaling.

Prima di eseguire il comando, verifica che il parametro predefinito preveda un determinato valore obiettivo. Per esempio, per aumentare orizzontalmente la capacità quando la CPU raggiunge il 50% di utilizzo, indica un valore obiettivo pari a 50,0. Oppure, per aumentare orizzontalmente il provisioning simultaneo di Lambda quando l'utilizzo raggiunge il 70% di capacità, indica un valore obiettivo pari a 0,7. Per informazioni sui valori obiettivo con riferimento a una specifica risorsa, consulta la documentazione fornita dal servizio per configurare il monitoraggio degli obiettivi. Per ulteriori informazioni, consulta Servizi AWS che puoi usare con Application Auto Scaling.

Linux, macOS o Unix

```
aws application-autoscaling put-scaling-policy \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --policy-name my-scaling-policy --policy-type TargetTrackingScaling \
  --target-tracking-scaling-policy-configuration '{ "TargetValue": 50.0,
 "PredefinedMetricSpecification": { "PredefinedMetricType": "predefinedmetric" }}'
```

Windows

```
aws application-autoscaling put-scaling-policy --service-namespace namespace --
scalable-dimension dimension --resource-id identifier --policy-name my-scaling-
policy --policy-type TargetTrackingScaling --target-tracking-scaling-policy-
configuration "{ \"TargetValue\": 50.0, \"PredefinedMetricSpecification\":
 { \"PredefinedMetricType\": \"predefinedmetric\" }}"
```

In caso di successo, questo comando restituisce gli ARN e i nomi dei due CloudWatch allarmi che sono stati creati per tuo conto.

2. Per confermare l'esistenza dell'operazione pianificata, utilizza il comando describe-scaling-policies.

   Linux, macOS o Unix

   ```
   aws application-autoscaling describe-scaling-policies --service-namespace namespace
    \
     --query 'ScalingPolicies[?ResourceId==`identifier`]'
   ```

Windows

```
aws application-autoscaling describe-scaling-policies --service-namespace namespace
 --query "ScalingPolicies[?ResourceId==`identifier`]"
```

Di seguito è riportato un output di esempio.

```
[
    {
        "PolicyARN": "arn",
        "TargetTrackingScalingPolicyConfiguration": {
            "PredefinedMetricSpecification": {
                "PredefinedMetricType": "predefinedmetric"
            },
            "TargetValue": 50.0
        },
        "PolicyName": "my-scaling-policy",
        "PolicyType": "TargetTrackingScaling",
        "Alarms": [],
        ...
    }
]
```

Fase 4: fasi successive

Quando si verifica un'attività di dimensionamento, viene visualizzato un record nell'output delle attività di dimensionamento per il target scalabile, ad esempio:

```
Successfully set desired count to 1. Change successfully fulfilled by ecs.
```

Per monitorare le attività di dimensionamento tramite Application Auto Scaling, portai utilizzare il seguente comando describe-scaling-activities.

Linux, macOS o Unix

```
aws application-autoscaling describe-scaling-activities
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier
```

Windows

```
aws application-autoscaling describe-scaling-activities --service-namespace namespace
 --scalable-dimension dimension --resource-id identifier
```

Fase 5: Pulizia

Per evitare che il tuo account accumuli costi per le risorse create in fase di dimensionamento attivo, è possibile cancellare la relativa configurazione di dimensionamento procedendo come segue.

L'eliminazione della configurazione di ridimensionamento non elimina la risorsa sottostante. AWS Inoltre, non la fa ritornare alla sua capacità originale. È possibile utilizzare la console del servizio all'interno del quale la risorsa è stata creata per eliminarla o modificarne la capacità.

Per eliminare le operazioni pianificate

Il comando delete-scheduled-action elimina una specifica operazione pianificata. Se desideri mantenere l'operazione pianificata per utilizzarla in futuro, puoi ignorare questa fase.

Linux, macOS o Unix

```
aws application-autoscaling delete-scheduled-action \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --scheduled-action-name my-second-scheduled-action
```

Windows

```
aws application-autoscaling delete-scheduled-action --service-namespace namespace
 --scalable-dimension dimension --resource-id identifier --scheduled-action-name my-
second-scheduled-action
```

Per eliminare la policy di dimensionamento

Il comando delete-scaling-policy elimina una determinata policy di dimensionamento con monitoraggio degli obiettivi. Se desideri mantenere la policy di dimensionamento per utilizzarla in futuro, puoi ignorare questa fase.

Linux, macOS o Unix

```
aws application-autoscaling delete-scaling-policy \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --policy-name my-scaling-policy
```

Windows

```
aws application-autoscaling delete-scaling-policy --service-namespace namespace --
scalable-dimension dimension --resource-id identifier --policy-name my-scaling-policy
```

Per annullare la registrazione di un obiettivo scalabile

Utilizza il comando deregister-scalable-target per annullare la registrazione dell'obiettivo scalabile. Se hai a disposizione policy di dimensionamento create o eventuali operazioni pianificate che non sono state ancora eliminate, verranno eliminate tramite questo comando. Se desideri mantenere l'obiettivo scalabile registrato per utilizzarlo in futuro, puoi ignorare questa fase.

Linux, macOS o Unix

```
aws application-autoscaling deregister-scalable-target \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier
```

Windows

```
aws application-autoscaling deregister-scalable-target --service-namespace namespace --
scalable-dimension dimension --resource-id identifier
```

Sospendi e riprendi il dimensionamento per Application Auto Scaling

Questo argomento spiega come sospendere e riprendere una o più attività di dimensionamento per i target scalabili nell'applicazione. La caratteristica sospensione-ripresa è utilizzata per interrompere temporaneamente le attività di dimensionamento attivate da policy di dimensionamento e operazioni pianificate. Questa può essere utile, ad esempio, quando non desideri che la scalabilità automatica interferisca potenzialmente mentre apporti una modifica o esamini un problema di configurazione. Le policy di dimensionamento e le operazioni pianificate possono essere mantenute e quando sono pronte, le attività di dimensionamento possono essere riavviate.

Nei seguenti comandi della CLI di esempio, passerai i parametri in formato JSON in un file config.json. È inoltre possibile passare questi parametri sulla riga di comando utilizzando le virgolette per racchiudere la struttura dati JSON. Per ulteriori informazioni, consulta Utilizzo di virgolette con stringhe nella AWS CLI nella Guida per l'utente di AWS Command Line Interface .

Indice

> **Note**
>
> Per istruzioni su come sospendere i processi di scalabilità orizzontale mentre le distribuzioni di Amazon ECS sono in corso, consulta la seguente documentazione: Scalabilità e implementazioni automatiche dei servizi nella Amazon Elastic Container Service Developer Guide

Attività di dimensionamento

Application Auto Scaling supporta l'inserimento delle seguenti attività di dimensionamento in uno stato sospeso:

- Tutte le attività di scalabilità verticale attivate da una policy di dimensionamento.
- Tutte le attività di scalabilità orizzontale attivate da una policy di dimensionamento.

- Tutte le attività di dimensionamento che implicano operazioni pianificate.

Le seguenti descrizioni spiegano cosa succede quando le singole attività di dimensionamento vengono sospese. Ogni singola attività può essere sospesa e riavviata in modo indipendente. A seconda del motivo per cui si sospende un'attività di dimensionamento, potrebbe essere necessario sospendere più attività di dimensionamento.

`DynamicScalingInSuspended`

- Application Auto Scaling non rimuove la capacità quando viene attivata una policy di dimensionamento con monitoraggio degli obiettivi o una policy di dimensionamento per fasi. Ciò consente di disabilitare temporaneamente le attività di scalabilità associate alle politiche di scalabilità senza eliminare le politiche di scalabilità o gli allarmi associati. CloudWatch Quando riprendi il dimensionamento orizzontale (riduzione), Application Auto Scaling valuta le policy con le soglie di allarme attualmente in violazione.

`DynamicScalingOutSuspended`

- Application Auto Scaling non aggiunge la capacità quando viene attivata una policy di dimensionamento con monitoraggio degli obiettivi o una policy di dimensionamento per fasi. Ciò consente di disabilitare temporaneamente le attività di scalabilità orizzontale associate alle politiche di scalabilità senza eliminare le politiche di scalabilità o gli allarmi associati. CloudWatch Quando riprendi il dimensionamento orizzontale (aumento), Application Auto Scaling valuta le policy con le soglie di allarme attualmente in violazione.

`ScheduledScalingSuspended`

- Application Auto Scaling non avvia operazioni di dimensionamento programmate per l'esecuzione durante il periodo di sospensione. Quando riprendi il dimensionamento pianificato, Application Auto Scaling valuta solo le operazioni pianificate il cui orario di esecuzione non è ancora trascorso.

Sospendere e riprendere le attività di scalabilità

È possibile sospendere e riprendere le singole attività o tutte le attività di dimensionamento per l'obiettivo scalabile Application Auto Scaling.

> **Note**
>
> Per brevità, questi esempi illustrano come sospendere e riprendere il dimensionamento per una tabella DynamoDB. Per specificare un altro target scalabile, specificare il suo spazio dei nomi in `--service-namespace`, la sua dimensione scalabile `--scalable-dimension` e l'ID di risorsa in `--resource-id`. Per maggiori informazioni ed esempi per ogni servizio, consulta gli argomenti in Servizi AWS che puoi usare con Application Auto Scaling.

Per sospendere un'attività di dimensionamento

Apri una finestra a riga di comando e utilizza il comando register-scalable-target con l'opzione `--suspended-state` come segue.

Linux, macOS o Unix

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table \
  --suspended-state file://config.json
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb --
scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table --
suspended-state file://config.json
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Per sospendere solo le attività di scalabilità verticale attivate da una policy di dimensionamento, specificare quanto segue in config.json.

```
{
    "DynamicScalingInSuspended":true
}
```

Per sospendere solo le attività di scalabilità orizzontale attivate da una policy di dimensionamento, specificare quanto segue in config.json.

```
{
    "DynamicScalingOutSuspended":true
}
```

Per sospendere solo le attività di dimensionamento che implicano azioni pianificate, specificare quanto segue in config.json.

```
{
    "ScheduledScalingSuspended":true
}
```

Per sospendere tutte le attività di dimensionamento

Utilizza il comando register-scalable-target con l'opzione `--suspended-state` come segue.

Linux, macOS o Unix

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table \
  --suspended-state file://config.json
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb --
scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table --
suspended-state file://config.json
```

In questo esempio si presuppone che il file config.json contiene i seguenti parametri in formato JSON.

```
{
    "DynamicScalingInSuspended":true,
    "DynamicScalingOutSuspended":true,
    "ScheduledScalingSuspended":true
}
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Visualizzazione delle attività di dimensionamento sospese

Utilizza il comando describe-scalable-targets per determinare quali attività di dimensionamento sono in modalità sospesa per un obiettivo scalabile.

Linux, macOS o Unix

```
aws application-autoscaling describe-scalable-targets --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table
```

Windows

```
aws application-autoscaling describe-scalable-targets --service-namespace dynamodb --
scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table
```

Di seguito è riportato un output di esempio.

```
{
    "ScalableTargets": [
        {
            "ServiceNamespace": "dynamodb",
            "ScalableDimension": "dynamodb:table:ReadCapacityUnits",
            "ResourceId": "table/my-table",
            "MinCapacity": 1,
            "MaxCapacity": 20,
            "SuspendedState": {
                "DynamicScalingOutSuspended": true,
                "DynamicScalingInSuspended": true,
                "ScheduledScalingSuspended": true
            },
            "CreationTime": 1558125758.957,
            "RoleARN": "arn:aws:iam::123456789012:role/aws-
service-role/dynamodb.application-autoscaling.amazonaws.com/
AWSServiceRoleForApplicationAutoScaling_DynamoDBTable"
        }
    ]
```

```
}
```

Riprendere le attività di dimensionamento

Quando sei pronto, puoi riprendere le attività di dimensionamento utilizzando il comando register-scalable-target.

Il seguente comando di esempio riprende tutte le attività di dimensionamento per il target scalabile specificato.

Linux, macOS o Unix

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table \
  --suspended-state file://config.json
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb --
scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table --
suspended-state file://config.json
```

In questo esempio si presuppone che il file config.json contiene i seguenti parametri in formato JSON.

```
{
    "DynamicScalingInSuspended":false,
    "DynamicScalingOutSuspended":false,
    "ScheduledScalingSuspended":false
}
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
}
```

Attività di dimensionamento per Application Auto Scaling

Application Auto Scaling monitora le CloudWatch metriche della politica di scalabilità e avvia un'attività di scalabilità quando vengono superate le soglie. Inoltre, avvia le attività di dimensionamento quando modifichi la dimensione massima o minima dell'obiettivo scalabile, manualmente oppure seguendo una pianificazione.

Quando si verifica un'attività di dimensionamento, Application Auto Scaling esegue una delle seguenti operazioni:

- Aumenta la capacità dell'obiettivo scalabile (denominato aumento orizzontale)
- Riduce la capacità dell'obiettivo scalabile (denominato riduzione orizzontale)

Puoi cercare le attività di dimensionamento delle ultime sei settimane.

Cerca le attività di scalabilità per target scalabile

Per visualizzare le attività di dimensionamento di uno specifico obiettivo scalabile, utilizza il comando seguente describe-scaling-activities.

Linux, macOS o Unix

```
aws application-autoscaling describe-scaling-activities --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
```

Windows

```
aws application-autoscaling describe-scaling-activities --service-namespace ecs --
scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
```

Di seguito è riportato un esempio di risposta, dove `StatusCode` contiene lo stato corrente dell'attività e `StatusMessage` contiene informazioni sullo stato dell'attività di dimensionamento.

```
{
    "ScalingActivities": [
        {
            "ScalableDimension": "ecs:service:DesiredCount",
            "Description": "Setting desired count to 1.",
```

```
            "ResourceId": "service/my-cluster/my-service",
            "ActivityId": "e6c5f7d1-dbbb-4a3f-89b2-51f33e766399",
            "StartTime": 1462575838.171,
            "ServiceNamespace": "ecs",
            "EndTime": 1462575872.111,
            "Cause": "monitor alarm web-app-cpu-lt-25 in state ALARM triggered policy
 web-app-cpu-lt-25",
            "StatusMessage": "Successfully set desired count to 1. Change successfully
 fulfilled by ecs.",
            "StatusCode": "Successful"
        }
    ]
}
```

Per una descrizione dei campi della risposta, consulta ScalingActivityl'Application Auto Scaling API Reference.

I seguenti codici di stato indicano quando l'evento di dimensionamento che porta all'attività di dimensionamento raggiunge uno stato completato:

- `Successful`: il dimensionamento è stato completato con successo
- `Overridden`: la capacità desiderata è stata aggiornata mediante un nuovo evento di dimensionamento
- `Unfulfilled`: il dimensionamento è scaduto o il servizio di destinazione non può soddisfare la richiesta
- `Failed`: il dimensionamento non è riuscito con un'eccezione

> **Note**
>
> L'attività di dimensionamento potrebbe anche avere uno stato pari a `Pending` o `InProgress`. Tutte le attività di dimensionamento hanno uno stato `Pending` prima che il servizio di destinazione risponda. Dopo la risposta della destinazione, lo stato dell'attività di dimensionamento diventa `InProgress`.

Includi attività non ridimensionate

Per impostazione predefinita, le attività di dimensionamento non riflettono i momenti in cui Application Auto Scaling decide di non ridimensionare.

Ad esempio, si presuma che un servizio Amazon ECS superi la soglia massima di una determinata metrica, ma che il numero di attività sia già pari al numero massimo di attività consentite. In questo caso, Application Auto Scaling non aumenta orizzontalmente il numero di attività desiderato.

Per includere attività non ridimensionate (not scaled activities) nella risposta, aggiungi l'opzione `--include-not-scaled-activities` al comando describe-scaling-activities.

Linux, macOS o Unix

```
aws application-autoscaling describe-scaling-activities --include-not-scaled-activities
 \
  --service-namespace ecs --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service
```

Windows

```
aws application-autoscaling describe-scaling-activities --include-not-scaled-activities
 --service-namespace ecs --scalable-dimension ecs:service:DesiredCount --resource-
id service/my-cluster/my-service
```

> **Note**
>
> Se questo comando genera un errore, assicurati di averlo aggiornato AWS CLI localmente alla versione più recente.

Per confermare che la risposta include le attività non ridimensionate, l'elemento `NotScaledReasons` viene mostrato nell'output di alcune, se non tutte, attività di dimensionamento non riuscite.

```
{
    "ScalingActivities": [
        {
            "ScalableDimension": "ecs:service:DesiredCount",
            "Description": "Attempting to scale due to alarm triggered",
            "ResourceId": "service/my-cluster/my-service",
            "ActivityId": "4d759079-a31f-4d0c-8468-504c56e2eecf",
            "StartTime": 1664928867.915,
            "ServiceNamespace": "ecs",
            "Cause": "monitor alarm web-app-cpu-gt-75 in state ALARM triggered policy
 web-app-cpu-gt-75",
            "StatusCode": "Failed",
```

```
            "NotScaledReasons": [
                {
                    "Code": "AlreadyAtMaxCapacity",
                    "MaxCapacity": 4
                }
            ]
        }
    ]
}
```

Per una descrizione dei campi della risposta, consulta ScalingActivityl'Application Auto Scaling API Reference.

Se viene restituita un'attività non ridimensionata, in base al codice del motivo elencato nella risposta in `Code`, attributi come `CurrentCapacity`, `MaxCapacity` e `MinCapacity` potrebbero essere presenti nella risposta.

Per evitare grandi quantità di inserimenti duplicati, nella cronologia delle attività di ridimensionamento verrà registrata solo la prima attività non ridimensionata. Qualsiasi attività successiva non ridimensionata non genererà nuove voci, a meno che non venga modificato il motivo del mancato ridimensionamento.

Codici motivazionali

Di seguito sono riportati i codici motivo di un'attività non ridimensionata.

Codice di motivo	Definizione			
AutoScalingAnticipatedFlapping	L'algoritmo di dimensionamento automatico ha stabilito di non eseguire un'azione di dimension amento perché avrebbe comportato un flapping.			

Codice di motivo	Definizione			
	Flapping è un ciclo infinito di riduzione e aumento orizzontale. Ciò significa che se viene eseguita un'azione di dimensionamento, il valore della metrica cambierebbe per avviare un'altra azione di dimensionamento nella direzione opposta.			
TargetServicePutResourceAsUnscalable	Il servizio di destinazione ha temporaneamente messo la risorsa in uno stato non scalabile. Application Auto Scaling ritenterà, qualora venissero soddisfatte le condizioni di dimensionamento automatico configurate nella policy di dimensionamento.			

Codice di motivo	Definizione			
AlreadyAtMaxCapacity	Il dimensionamento è bloccato dalla capacità massima specificata. Se desideri che Application Auto Scaling aumenti orizzontalmente, è necessario aumentare la capacità massima.			
AlreadyAtMinCapacity	Il dimensionamento è bloccato dalla capacità minima specificata. Se desideri che Application Auto Scaling riduca orizzontalmente, è necessario ridurre la capacità minima.			
AlreadyAtDesiredCapacity	L'algoritmo di dimensionamento automatico ha calcolato la capacità aggiornata in modo che sia uguale alla capacità attuale.			

Applicazioni di monitoraggio Auto Scaling

Il monitoraggio è una parte importante per mantenere l'affidabilità, la disponibilità e le prestazioni di Application Auto Scaling e delle altre AWS soluzioni. È necessario raccogliere i dati di monitoraggio da tutte le parti della AWS soluzione in modo da poter eseguire più facilmente il debug di un errore multipunto, se si verifica uno. AWS fornisce strumenti di monitoraggio per monitorare Application Auto Scaling, segnalare quando qualcosa non va e intraprendere azioni automatiche quando necessario.

Puoi utilizzare le seguenti funzionalità per aiutarti a gestire le tue AWS risorse:

AWS CloudTrail

Con AWS CloudTrail, puoi tenere traccia delle chiamate effettuate all'API Application Auto Scaling da o per tuo conto. Account AWS CloudTrail archivia le informazioni nei file di registro nel bucket Amazon S3 specificato. Puoi identificare quali utenti e account hanno richiamato Application Auto Scaling, l'indirizzo IP di origine da cui sono state effettuate le chiamate e quando sono avvenute. Per ulteriori informazioni, consulta Registra le chiamate API Application Auto Scaling utilizzando AWS CloudTrail.

> **Note**
>
> Per informazioni su altri AWS servizi che possono aiutarti a registrare e raccogliere dati sui tuoi carichi di lavoro, consulta la guida alla registrazione e al monitoraggio per i proprietari delle applicazioni contenuta nella Prescriptive Guidance.AWS

Amazon CloudWatch

Amazon ti CloudWatch aiuta ad analizzare i log e, in tempo reale, a monitorare i parametri delle tue AWS risorse e delle applicazioni ospitate. Puoi raccogliere i parametri e tenerne traccia, creare pannelli di controllo personalizzati e impostare allarmi per inviare una notifica o intraprendere azioni quando un parametro specificato raggiunge una determinata soglia. Ad esempio, puoi tenere CloudWatch traccia dell'utilizzo delle risorse e ricevere notifiche quando l'utilizzo è molto elevato o quando l'allarme della metrica è entrato in funzione. `INSUFFICIENT_DATA` Per ulteriori informazioni, consulta Monitora l'utilizzo di risorse scalabili utilizzando CloudWatch.

CloudWatch tiene traccia anche delle metriche di utilizzo delle AWS API per Application Auto Scaling. Puoi utilizzare queste metriche per configurare allarmi che ti avvisano quando il volume

delle chiamate API viola una soglia da te definita. Per ulteriori informazioni, consulta i parametri di AWS utilizzo nella Amazon CloudWatch User Guide.

Amazon EventBridge

Amazon EventBridge è un servizio di bus eventi senza server che semplifica la connessione delle applicazioni con dati provenienti da una varietà di fonti. EventBridge fornisce un flusso di dati in tempo reale dalle tue applicazioni, applicazioni Software-as-a-Service (SaaS) AWS e servizi e indirizza tali dati verso destinazioni come Lambda. Questo ti consente di monitorare gli eventi che si verificano nei servizi e creare architetture basate su eventi. Per ulteriori informazioni, consulta Monitora gli eventi di Application Auto Scaling utilizzando Amazon EventBridge.

AWS Health Dashboard

Il AWS Health Dashboard (PHD) visualizza informazioni e fornisce anche notifiche richiamate dai cambiamenti nello stato delle risorse. AWS Le informazioni vengono presentate in due modi: su un pannello di controllo che mostra eventi recenti e prossimi organizzati per categoria e in un log completo che mostra tutti gli eventi degli ultimi 90 giorni. Per ulteriori informazioni, vedi Guida introduttiva a. AWS Health Dashboard

Monitora l'utilizzo di risorse scalabili utilizzando CloudWatch

Con Amazon CloudWatch, ottieni una visibilità quasi continua delle tue applicazioni su risorse scalabili. CloudWatch è un servizio di monitoraggio delle AWS risorse. È possibile CloudWatch utilizzarlo per raccogliere e tenere traccia delle metriche, impostare allarmi e reagire automaticamente ai cambiamenti nelle risorse. AWS Puoi anche creare pannelli di controllo per monitorare i parametri o i set di parametri specifici di cui hai bisogno.

Quando interagisci con i servizi che si integrano con Application Auto Scaling, questi inviano le metriche mostrate nella tabella seguente a. CloudWatch In CloudWatch, le metriche vengono raggruppate prima in base allo spazio dei nomi del servizio e poi in base alle varie combinazioni di dimensioni all'interno di ogni spazio dei nomi. Questi parametri possono aiutarti a monitorare l'utilizzo delle risorse e a pianificare la capacità delle tue applicazioni. Se il carico di lavoro dell'applicazione non è costante, è consigliabile utilizzare il dimensionamento automatico. Per descrizioni dettagliate di questi parametri, consulta la documentazione relativa al parametro di interesse.

Indice

CloudWatch metriche per il monitoraggio dell'utilizzo delle risorse

La tabella seguente elenca le CloudWatch metriche disponibili per supportare il monitoraggio dell'utilizzo delle risorse. L'elenco non è esaustivo ma fornisce un buon punto di partenza. Se non vedi queste metriche nella CloudWatch console, assicurati di aver completato la configurazione della risorsa. Per ulteriori informazioni, consulta la Amazon CloudWatch User Guide.

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
AppStream 2.0			
Parchi istanze	AWS/AppStream	Nome: AvailableCapacity Dimensione: parco istanze	AppStream Metriche 2.0
Parchi istanze	AWS/AppStream	Nome: CapacityUtilization Dimensione: parco istanze	AppStream Metriche 2.0
Aurora			
Repliche	AWS/RDS	Nome: CPUUtilization Dimensioni: DBClusterIdentifie	Parametri a livello di cluster di Amazon Aurora

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
		r, Role (READER)	
Repliche	AWS/RDS	Nome: DatabaseConnections Dimensioni: DBClusterIdentifier, Role (READER)	Parametri a livello di cluster di Amazon Aurora
Amazon Comprehend			
Endpoint di classificazione dei documenti	AWS/Comprehend	Nome: InferenceUtilization Dimensione: EndpointArn	Parametri relativi agli endpoint di Amazon Comprehend
Endpoint di riconoscimento delle entità	AWS/Comprehend	Nome: InferenceUtilization Dimensione: EndpointArn	Parametri relativi agli endpoint di Amazon Comprehend
DynamoDB			

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
Tabelle e indici secondari globali	AWS/DynamoDB	Nome: ProvisionedReadCapacityUnits Dimensioni: TableName, GlobalSecondaryIndexName	Parametri di DynamoDB
Tabelle e indici secondari globali	AWS/DynamoDB	Nome: ProvisionedWriteCapacityUnits Dimensioni: TableName, GlobalSecondaryIndexName	Parametri di DynamoDB

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
Tabelle e indici secondari globali	AWS/DynamoDB	Nome: ConsumedReadCapacityUnits Dimensioni: TableName, GlobalSecondaryIndexName	Parametri di DynamoDB
Tabelle e indici secondari globali	AWS/DynamoDB	Nome: ConsumedWriteCapacityUnits Dimensioni: TableName, GlobalSecondaryIndexName	Parametri di DynamoDB
Amazon ECS			

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
Servizi	AWS/ECS	Nome: CPUUtilization Dimensioni: ClusterName, ServiceName	Parametri di Amazon ECS
Servizi	AWS/ECS	Nome: MemoryUtilization Dimensioni: ClusterName, ServiceName	Parametri di Amazon ECS
Servizi	AWS/ApplicationELB	Nome: RequestCountPerTarget Dimensione: TargetGroup	Parametri di Application Load Balancer
ElastiCache			

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
Cluster (gruppi di replica)	AWS/ElastiCache	Nome: DatabaseMemoryUsageCountedForEvictPercentage Dimensione: ReplicationGroupId	ElastiCache per le metriche Redis
Cluster (gruppi di replica)	AWS/ElastiCache	Nome: DatabaseCapacityUsageCountedForEvictPercentage Dimensione: ReplicationGroupId	ElastiCache per le metriche Redis
Cluster (gruppi di replica)	AWS/ElastiCache	Nome: EngineCPUUtilization Dimensioni: ReplicationGroupId, Ruolo (primario)	ElastiCache per le metriche Redis

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
Cluster (gruppi di replica)	AWS/ElastiCache	Nome: EngineCPUUtilization Dimensioni: ReplicationGroupId, Role (Replica)	ElastiCache per le metriche Redis
Amazon EMR			
Cluster	AWS/ElasticMap Riduci	Nome: YARNPercentageMemoryAvailable Dimensione: ClusterId	Parametri di Amazon EMR
Amazon Keyspaces			
Tabelle	AWS/Cassandra	Nome: ProvisionedReadCapacityUnits Dimensioni: Keyspace, TableName	Parametri di Amazon Keyspaces

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
Tabelle	AWS/Cassandra	Nome: ProvisionedWriteCapacityUnits Dimensioni: Keyspace, TableName	Parametri di Amazon Keyspaces
Tabelle	AWS/Cassandra	Nome: ConsumedReadCapacityUnits Dimensioni: Keyspace, TableName	Parametri di Amazon Keyspaces
Tabelle	AWS/Cassandra	Nome: ConsumedWriteCapacityUnits Dimensioni: Keyspace, TableName	Parametri di Amazon Keyspaces
Lambda			

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
Simultaneità fornita	AWS/Lambda	Nome: ProvisionedConcurrencyUtilization Dimensioni: FunctionName, Risorsa	Parametri della funzione Lambda
Amazon MSK			
Archiviazione del broker	AWS/Kafka	Nome: KafkaDataLogsDiskUsed Dimensioni: Nome del cluster	Parametri Amazon ECR
Archiviazione del broker	AWS/Kafka	Nome: KafkaDataLogsDiskUsed Dimensioni: Nome del cluster, ID broker	Parametri Amazon ECR

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
Neptune			
Cluster	AWS/Neptune	Nome: CPUUtilization Dimensioni: DBClusterIdentifier, Role (READER)	Parametri di Neptune
SageMaker			
Varianti di endpoint	AWS/SageMaker	Nome: InvocationsPerInstance Dimensioni: EndpointName, VariantName	Parametri di invocazione

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
Componenti di inferenza	AWS/SageMaker	Nome: InvocationsPerCopy Dimensioni: InferenceComponentName	Parametri di invocazione
Provisioning simultaneo per un endpoint serverless	AWS/SageMaker	Nome: ServerlessProvisionedConcurrencyUtilization Dimensioni: EndpointName, VariantName	Parametri degli endpoint serverless
Serie di istanze Spot (Amazon EC2)			
Parco istanze Spot	AWS/EC2Spot	Nome: CPUUtilization Dimensione: FleetRequestId	Parametri della serie di istanze spot

Risorse scalabili	Spazio dei nomi	CloudWatch metrica	Collegamento alla documentazione
Parco istanze Spot	AWS/EC2Spot	Nome: NetworkIn Dimensione: FleetRequestId	Parametri della serie di istanze spot
Parco istanze Spot	AWS/EC2Spot	Nome: NetworkOut Dimensione: FleetRequestId	Parametri della serie di istanze spot
Parco istanze Spot	AWS/ApplicationELB	Nome: RequestCountPerTarget Dimensione: TargetGroup	Parametri di Application Load Balancer

Policy di dimensionamento del monitoraggio degli obiettivi con parametri predefiniti

La tabella seguente elenca i tipi di metrica predefiniti dall'Application Auto Scaling API Reference con il nome della metrica corrispondente. CloudWatch Ogni metrica predefinita rappresenta un'aggregazione dei valori della metrica sottostante. CloudWatch Il risultato è l'utilizzo medio delle risorse in un minuto, basato su una percentuale, se non diversamente specificato. I parametri

predefiniti vengono utilizzati solo nel contesto dell'impostazione delle policy di dimensionamento del monitoraggio degli obiettivi.

Puoi trovare ulteriori informazioni su questi parametri nella documentazione relativa al servizio, disponibile nella tabella in CloudWatch metriche per il monitoraggio dell'utilizzo delle risorse.

Tipo di parametro predefinito	CloudWatch nome della metrica
AppStream 2.0	
`AppStreamAverageCapacityUtilization`	CapacityUtilization
Aurora	
`RDSReaderAverageCPUUtilization`	CPUUtilization
`RDSReaderAverageDatabaseConnections`	DatabaseConnections[1]
Amazon Comprehend	
`ComprehendInferenceUtilization`	InferenceUtilization
DynamoDB	
`DynamoDBReadCapacityUtilization`	ProvisionedReadCapacityUnits, ConsumedReadCapacityUnits [2]
`DynamoDBWriteCapacityUtilization`	ProvisionedWriteCapacityUnits, ConsumedWriteCapacityUnits [2]
Amazon ECS	
`ECSServiceAverageCPUUtilization`	CPUUtilization
`ECSServiceAverageMemoryUtilization`	MemoryUtilization
`ALBRequestCountPerTarget`	RequestCountPerTarget[1]

Tipo di parametro predefinito	CloudWatch nome della metrica
ElastiCache	
`ElastiCacheDatabaseMemoryUsageCountedForEvictPercentage`	DatabaseMemoryUsageCountedForEvictPercentuale
`ElastiCacheDatabaseCapacityUsageCountedForEvictPercentage`	DatabaseCapacityUsageCountedForEvictPercentuale
`ElastiCachePrimaryEngineCPUUtilization`	EngineCPUUtilization
`ElastiCacheReplicaEngineCPUUtilization`	EngineCPUUtilization
Amazon Keyspaces	
`CassandraReadCapacityUtilization`	ProvisionedReadCapacityUnits, ConsumedReadCapacityUnits [2]
`CassandraWriteCapacityUtilization`	ProvisionedWriteCapacityUnits, ConsumedWriteCapacityUnits [2]
Lambda	
`LambdaProvisionedConcurrencyUtilization`	ProvisionedConcurrencyUtilizzo
Amazon MSK	
`KafkaBrokerStorageUtilization`	KafkaDataLogsDiskUsato
Neptune	
`NeptuneReaderAverageCPUUtilization`	CPUUtilization
SageMaker	

Tipo di parametro predefinito	CloudWatch nome della metrica
`SageMakerVariantInvocationsPerInstance`	InvocationsPerIstanza[1]
`SageMakerInferenceComponentInvocationsPerCopy`	InvocationsPerCopia[1]
`SageMakerVariantProvisionedConcurrencyUtilization`	ServerlessProvisionedConcurrencyUtilization
Spot Fleet	
`EC2SpotFleetRequestAverageCPUUtilization`	CPUUtilization[3]
`EC2SpotFleetRequestAverageNetworkIn`[3]	NetworkIn[1] [3]
`EC2SpotFleetRequestAverageNetworkOut`[3]	NetworkOut[1] [3]
`ALBRequestCountPerTarget`	RequestCountPerTarget[1]

[1] Il parametro si basa su un conteggio anziché su una percentuale.

[2] Per DynamoDB e Amazon Keyspaces, i parametri predefiniti sono un'aggregazione di due parametri per supportare la scalabilità basata sul consumo CloudWatch di throughput assegnato.

[3] Per prestazioni di dimensionamento ottimali, è necessario utilizzare il monitoraggio dettagliato di Amazon EC2.

Registra le chiamate API Application Auto Scaling utilizzando AWS CloudTrail

Application Auto Scaling è integrato con AWS CloudTrail un servizio che fornisce una registrazione delle azioni intraprese da un utente, un ruolo o un utente che Servizio AWS utilizza l'API Application Auto Scaling. CloudTrail acquisisce tutte le chiamate API per Application Auto Scaling come eventi.

Le chiamate acquisite includono chiamate provenienti da AWS Management Console e chiamate di codice all'API Application Auto Scaling. Se crei un trail, puoi abilitare la distribuzione continua di CloudTrail eventi a un bucket Amazon S3, inclusi gli eventi per Application Auto Scaling. Se non configuri un percorso, puoi comunque visualizzare gli eventi più recenti nella CloudTrail console nella cronologia degli eventi. Utilizzando le informazioni raccolte da CloudTrail, è possibile determinare la richiesta effettuata ad Application Auto Scaling, l'indirizzo IP da cui è stata effettuata la richiesta, chi ha effettuato la richiesta, quando è stata effettuata e dettagli aggiuntivi.

Per ulteriori informazioni CloudTrail, consulta la Guida per l'AWS CloudTrail utente.

Informazioni sull'Application Auto Scaling in CloudTrail

CloudTrail è abilitato sul tuo Account AWS quando crei l'account. Quando si verifica un'attività di Application Auto Scaling, tale attività viene registrata in un CloudTrail evento insieme ad altri eventi di AWS servizio nella cronologia degli eventi. Puoi visualizzare, cercare e scaricare eventi recenti in. Account AWS Per ulteriori informazioni, consulta Visualizzazione degli eventi con la cronologia degli CloudTrail eventi.

Per una registrazione continua degli eventi del tuo Account AWS, compresi gli eventi per Application Auto Scaling, crea un percorso. Un trail consente di CloudTrail inviare file di log a un bucket Amazon S3. Per impostazione predefinita, quando si crea un percorso nella console, questo sarà valido in tutte le Regioni AWS. Il trail registra gli eventi di tutte le regioni della AWS partizione e consegna i file di log al bucket Amazon S3 specificato. Inoltre, puoi configurare altri Amazon Web Services per analizzare ulteriormente e agire in base ai dati sugli eventi raccolti nei CloudTrail log. Per ulteriori informazioni, consulta gli argomenti seguenti:

- Panoramica della creazione di un percorso
- CloudTrail servizi e integrazioni supportati
- Configurazione delle notifiche Amazon SNS per CloudTrail
- Ricezione di file di CloudTrail registro da più regioni e ricezione di file di CloudTrail registro da più account

Tutte le azioni di Application Auto Scaling vengono registrate CloudTrail e documentate nell'Application Auto Scaling API Reference. Ad esempio, le chiamate alle `PutScalingPolicy` `DescribeScalingPolicies` azioni e generano voci nei file di registro. `DeleteScalingPolicy` CloudTrail

Ogni evento o voce di log contiene informazioni sull'utente che ha generato la richiesta. Le informazioni di identità consentono di determinare quanto segue:

- Se la richiesta è stata effettuata con credenziali utente root o AWS Identity and Access Management (IAM).
- Se la richiesta è stata effettuata con le credenziali di sicurezza temporanee per un ruolo o un utente federato.
- Se la richiesta è stata effettuata da un altro AWS servizio.

Per ulteriori informazioni, vedete l'elemento CloudTrail userIdentity.

Comprendere delle voci di file di log di Application Auto Scaling

Un trail è una configurazione che consente la distribuzione di eventi come file di log in un bucket Amazon S3 specificato dall'utente. CloudTrail i file di registro contengono una o più voci di registro. Un evento rappresenta una singola richiesta proveniente da qualsiasi fonte e include informazioni sull'azione richiesta, la data e l'ora dell'azione, i parametri della richiesta e così via. CloudTrail i file di registro non sono una traccia ordinata dello stack delle chiamate API pubbliche, quindi non vengono visualizzati in un ordine specifico.

L'esempio seguente mostra una voce di CloudTrail registro che illustra l'`DescribeScalableTargets`azione.

```
{
    "eventVersion": "1.05",
    "userIdentity": {
        "type": "Root",
        "principalId": "123456789012",
        "arn": "arn:aws:iam::123456789012:root",
        "accountId": "123456789012",
        "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
        "sessionContext": {
            "attributes": {
                "mfaAuthenticated": "false",
                "creationDate": "2018-08-21T17:05:42Z"
            }
        }
    },
    "eventTime": "2018-08-16T23:20:32Z",
    "eventSource": "autoscaling.amazonaws.com",
```

```
    "eventName": "DescribeScalableTargets",
    "awsRegion": "us-west-2",
    "sourceIPAddress": "72.21.196.68",
    "userAgent": "EC2 Spot Console",
    "requestParameters": {
        "serviceNamespace": "ec2",
        "scalableDimension": "ec2:spot-fleet-request:TargetCapacity",
        "resourceIds": [
            "spot-fleet-request/sfr-05ceaf79-3ba2-405d-e87b-612857f1357a"
        ]
    },
    "responseElements": null,
    "additionalEventData": {
        "service": "application-autoscaling"
    },
    "requestID": "0737e2ea-fb2d-11e3-bfd8-99133058e7bb",
    "eventID": "3fcfb182-98f8-4744-bd45-b38835ab61cb",
    "eventType": "AwsApiCall",
    "recipientAccountId": "123456789012"
}
```

Risorse correlate

Con CloudWatch Logs, è possibile monitorare e ricevere avvisi per eventi specifici acquisiti da. CloudTrail Gli eventi inviati a CloudWatch Logs sono quelli configurati per essere registrati dal tuo percorso, quindi assicurati di aver configurato il percorso o i percorsi per registrare i tipi di eventi che ti interessa monitorare. CloudWatch I log possono monitorare le informazioni nei file di registro e avvisare l'utente quando vengono raggiunte determinate soglie. Puoi inoltre archiviare i dati del log in storage estremamente durevole. Per ulteriori informazioni, consulta la Amazon CloudWatch Logs User Guide e l'argomento Monitoring CloudTrail log files with Amazon CloudWatch Logs nella AWS CloudTrail User Guide.

Monitora gli eventi di Application Auto Scaling utilizzando Amazon EventBridge

Amazon EventBridge, precedentemente chiamato CloudWatch Events, ti aiuta a monitorare eventi specifici di Application Auto Scaling e ad avviare azioni mirate che ne utilizzano altri. Servizi AWS Gli eventi di Servizi AWS vengono trasmessi quasi EventBridge in tempo reale.

In questo modo è possibile creare regole che corrispondano agli eventi in arrivo e indirizzarli verso le destinazioni per l'elaborazione. EventBridge

Per ulteriori informazioni, consulta la sezione Guida introduttiva ad Amazon EventBridge nella Amazon EventBridge User Guide.

Eventi Application Auto Scaling

Di seguito sono riportati eventi di esempio di Application Auto Scaling. Gli eventi vengono prodotti nel miglior modo possibile.

Attualmente per Application Auto Scaling sono disponibili solo gli eventi specifici per le CloudTrail chiamate scaled to max e API.

Event types (Tipi di evento)

- Evento per il cambiamento di stato: scalato al massimo
- Eventi per le chiamate API tramite CloudTrail

Evento per il cambiamento di stato: scalato al massimo

L'evento di esempio seguente mostra che Application Auto Scaling ha aumentato (orizzontalmente) la capacità della destinazione scalabile fino al limite massimo di dimensione. Se la domanda aumenta di nuovo, verrà impedito a Application Auto Scaling di aumentare la destinazione a una dimensione maggiore perché è già ridimensionata al massimo.

Nell'oggetto `detail`, i valori per gli attributi `resourceId`, `serviceNamespace` e `scalableDimension` identificano la destinazione scalabile. I valori per gli attributi `newDesiredCapacity` e `oldDesiredCapacity` si riferiscono alla nuova capacità dopo l'evento di aumento orizzontale e alla capacità originale precedente. `maxCapacity` è il limite massimo di dimensione della destinazione scalabile.


```
{
  "version": "0",
  "id": "11112222-3333-4444-5555-666677778888",
  "detail-type": "Application Auto Scaling Scaling Activity State Change",
  "source": "aws.application-autoscaling",
  "account": "123456789012",
  "time": "2019-06-12T10:23:40Z",
  "region": "us-west-2",
  "resources": [],
```

```
  "detail": {
    "startTime": "2022-06-12T10:20:43Z",
    "endTime": "2022-06-12T10:23:40Z",
    "newDesiredCapacity": 8,
    "oldDesiredCapacity": 5,
    "minCapacity": 2,
    "maxCapacity": 8,
    "resourceId": "table/my-table",
    "scalableDimension": "dynamodb:table:WriteCapacityUnits",
    "serviceNamespace": "dynamodb",
    "statusCode": "Successful",
    "scaledToMax": true,
    "direction": "scale-out"
}
```

Per creare una regola che acquisisca tutti gli `scaledToMax` eventi di modifica dello stato per tutte le destinazioni scalabili, utilizza il seguente modello di eventi di esempio.

```
{
  "source": [
    "aws.application-autoscaling"
  ],
  "detail-type": [
    "Application Auto Scaling Scaling Activity State Change"
  ],
  "detail": {
    "scaledToMax": [
      true
    ]
  }
}
```

Eventi per le chiamate API tramite CloudTrail

Un trail è una configurazione AWS CloudTrail utilizzata per fornire eventi come file di log a un bucket Amazon S3. CloudTrail i file di registro contengono voci di registro. Un evento rappresenta una voce di log e include informazioni sull'azione richiesta, la data e l'ora dell'operazione e i parametri della richiesta. Per informazioni su come iniziare CloudTrail, consulta Creazione di un percorso nella Guida per l'AWS CloudTrail utente.

Gli eventi che vengono erogati tramite CloudTrail hanno `AWS API Call via CloudTrail` come valore`detail-type`.

L'evento di esempio seguente rappresenta una voce del file di CloudTrail registro che mostra che un utente della console ha chiamato l'azione Application Auto Scaling RegisterScalableTarget.


```
{
  "version": "0",
  "id": "99998888-7777-6666-5555-444433332222",
  "detail-type": "AWS API Call via CloudTrail",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2022-07-13T16:50:15Z",
  "region": "us-west-2",
  "resources": [],
  "detail": {
    "eventVersion": "1.08",
    "userIdentity": {
      "type": "IAMUser",
      "principalId": "123456789012",
      "arn": "arn:aws:iam::123456789012:user/Bob",
      "accountId": "123456789012",
      "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
      "sessionContext": {
        "sessionIssuer": {
          "type": "Role",
          "principalId": "123456789012",
          "arn": "arn:aws:iam::123456789012:role/Admin",
          "accountId": "123456789012",
          "userName": "Admin"
        },
        "webIdFederationData": {},
        "attributes": {
          "creationDate": "2022-07-13T15:17:08Z",
          "mfaAuthenticated": "false"
        }
      }
    },
    "eventTime": "2022-07-13T16:50:15Z",
    "eventSource": "autoscaling.amazonaws.com",
    "eventName": "RegisterScalableTarget",
    "awsRegion": "us-west-2",
    "sourceIPAddress": "AWS Internal",
    "userAgent": "EC2 Spot Console",
    "requestParameters": {
      "resourceId": "spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE",
```

```
      "serviceNamespace": "ec2",
      "scalableDimension": "ec2:spot-fleet-request:TargetCapacity",
      "minCapacity": 2,
      "maxCapacity": 10
    },
    "responseElements": null,
    "additionalEventData": {
      "service": "application-autoscaling"
    },
    "requestID": "e9caf887-8d88-11e5-a331-3332aa445952",
    "eventID": "49d14f36-6450-44a5-a501-b0fdcdfaeb98",
    "readOnly": false,
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "123456789012",
    "eventCategory": "Management",
    "sessionCredentialFromConsole": "true"
  }
}
```

Per creare una regola basata su tutte le chiamate API DeleteScalingPolicy e DeregisterScalableTarget per tutte le destinazioni scalabili, utilizza il seguente modello di evento di esempio:

```
{
  "source": [
    "aws.autoscaling"
  ],
  "detail-type": [
    "AWS API Call via CloudTrail"
  ],
  "detail": {
    "eventSource": [
      "autoscaling.amazonaws.com"
    ],
    "eventName": [
       "DeleteScalingPolicy",
       "DeregisterScalableTarget"
    ],
    "additionalEventData": {
      "service": [
        "application-autoscaling"
      ]
```

```
    }
  }
}
```

Per ulteriori informazioni sull'utilizzo CloudTrail, consultaRegistra le chiamate API Application Auto Scaling utilizzando AWS CloudTrail.

Supporto al tagging per l'Applicazione di Dimensionamento automatico

Puoi utilizzare AWS CLI o un SDK per etichettare i target scalabili di Application Auto Scaling. Gli obiettivi scalabili sono le entità che rappresentano le AWS o le risorse personalizzate scalabili da Application Auto Scaling.

Ogni tag è un'etichetta composta da una chiave e da un valore definiti dall'utente utilizzando l'Applicazione di Dimensionamento automatico API. I tag possono aiutarti a configurare l'accesso granulare a obiettivi scalabili specifici in base alle esigenze dell'organizzazione. Per ulteriori informazioni, consulta ABAC con l'Applicazione di Dimensionamento automatico.

È possibile aggiungere tag a nuovi obiettivi scalabili al momento della loro registrazione oppure è possibile aggiungerli agli obiettivi scalabili esistenti.

I comandi comunemente utilizzati per la gestione dei tag includono:

- register-scalable-target per taggare nuovi obiettivi scalabili quando li registri.
- tag-resource per aggiungere tag a un obiettivo scalabile esistente.
- list-tags-for-resource per restituire i tag su un obiettivo scalabile.
- untag-resource per eliminare un tag.

Esempi di assegnazione di tag

Utilizza il comando register-scalable-target con l'opzione `--tags`. Questo esempio contrassegna un obiettivo scalabile con due tag: una chiave tag denominata **`environment`** con il valore tag **`production`** e una chiave tag denominata **`iscontainerbased`** con il valore tag **`true`**.

Sostituisci i valori di esempio per `--max-capacity` and `--min-capacity` e il testo di esempio per `--service-namespace` con lo spazio dei nomi del AWS servizio che stai utilizzando con Application Auto Scaling`--scalable-dimension`, con la dimensione scalabile associata alla risorsa che stai registrando `--resource-id` e con un identificatore per la risorsa. Per maggiori informazioni ed esempi per ogni servizio, consulta gli argomenti in Servizi AWS che puoi usare con Application Auto Scaling.

```
aws application-autoscaling register-scalable-target \
```

```
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --min-capacity 1 --max-capacity 10 \
  --tags environment=production,iscontainerbased=true
```

In caso di esito positivo, il comando restituisce l'ARN dell'obiettivo scalabile.

```
{
    "ScalableTargetARN": "arn:aws:application-autoscaling:region:account-id:scalable-target/1234abcd56ab78cd901ef1234567890ab123"
}
```

> **Note**
>
> Se questo comando genera un errore, assicurati di averlo aggiornato localmente alla versione più recente. AWS CLI

Tag di sicurezza

Utilizza i tag per verificare che il richiedente (ad esempio un utente IAM o un ruolo IAM) disponga delle autorizzazioni per eseguire determinate azioni. Fornire informazioni sui tag nell'elemento condizione di una policy IAM utilizzando una o più delle seguenti chiavi di condizione:

- Utilizza `aws:ResourceTag/tag-key: tag-value` per concedere (o negare) agli utenti operazioni su obiettivi scalabili con tag specifici.
- Utilizza `aws:RequestTag/tag-key: tag-value` per richiedere che un tag specifico sia presente (o non presente) in una richiesta.
- Utilizza `aws:TagKeys [tag-key, ...]` per richiedere che chiavi tag specifiche siano presenti (o non presenti) in una richiesta.

Per esempio, la seguente policy IAM concede all'utente le autorizzazioni per le operazioni seguenti: `DeregisterScalableTarget`, `DeleteScalingPolicy` e `DeleteScheduledAction`. Tuttavia, nega anche l'azione se l'obiettivo scalabile su cui si sta agendo ha il tag **`environment=production`**.

```
{
```

```
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "application-autoscaling:DeregisterScalableTarget",
                "application-autoscaling:DeleteScalingPolicy",
                "application-autoscaling:DeleteScheduledAction"
            ],
            "Resource": "*"
            }
        },
        {
            "Effect": "Deny",
            "Action": [
                "application-autoscaling:DeregisterScalableTarget",
                "application-autoscaling:DeleteScalingPolicy",
                "application-autoscaling:DeleteScheduledAction"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {"aws:ResourceTag/environment": "production"}
            }
        }
    ]
}
```

Controllo dell'accesso ai tag

Utilizza i tag per verificare che il richiedente (ad esempio un utente IAM o un ruolo IAM) disponga delle autorizzazioni necessarie per aggiungere, modificare o eliminare i tag per gli obiettivi scalabili.

Ad esempio, è possibile creare una policy IAM che consente di rimuovere dagli obiettivi scalabili solo il tag con la chiave **`temporary`**.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "application-autoscaling:UntagResource",
            "Resource": "*",
```

```
            "Condition": {
                "ForAllValues:StringEquals": { "aws:TagKeys": ["temporary"] }
            }
        }
    ]
}
```

Sicurezza in Application Auto Scaling

La sicurezza del cloud AWS è la massima priorità. In qualità di AWS cliente, puoi beneficiare di un data center e di un'architettura di rete progettati per soddisfare i requisiti delle organizzazioni più sensibili alla sicurezza.

La sicurezza è una responsabilità condivisa tra AWS te e te. Il modello di responsabilità condivisa descrive questo aspetto come sicurezza del cloud e sicurezza nel cloud:

- Sicurezza del cloud: AWS è responsabile della protezione dell'infrastruttura che gestisce AWS i servizi nel AWS cloud. AWS ti fornisce anche servizi che puoi utilizzare in modo sicuro. Per ulteriori informazioni sui programmi di conformità che si applicano all'Application Auto Scaling, consulta AWS i servizi rientranti nell'ambito del programma di conformità (AWS servizi compresi nell'ambito per programma).
- Sicurezza nel cloud: la tua responsabilità è determinata dal AWS servizio che utilizzi. Inoltre, sei responsabile anche di altri fattori, tra cui la riservatezza dei dati, i requisiti dell'azienda e le leggi e le normative applicabili.

Questa documentazione permette di comprendere come applicare il modello di responsabilità condivisa quando si usa Application Auto Scaling. Gli argomenti seguenti illustrano come configurare Application Auto Scaling in modo da soddisfare gli obiettivi di sicurezza e conformità. Scoprite anche come utilizzare altri AWS servizi che vi aiutano a monitorare e proteggere le risorse dell'Application Auto Scaling.

Indice

Protezione dei dati in Application Auto Scaling

Il modello di responsabilità AWS condivisa modello di di si applica alla protezione dei dati in Application Auto Scaling. Come descritto in questo modello, AWS è responsabile della protezione dell'infrastruttura globale che gestisce tutti i Cloud AWS. L'utente è responsabile del controllo dei contenuti ospitati su questa infrastruttura. L'utente è inoltre responsabile della configurazione della protezione e delle attività di gestione per i Servizi AWS utilizzati. Per ulteriori informazioni sulla privacy dei dati, vedi le Domande frequenti sulla privacy dei dati. Per informazioni sulla protezione dei dati in Europa, consulta il post del blog relativo al Modello di responsabilità condivisa AWS e GDPR nel Blog sulla sicurezza AWS .

Ai fini della protezione dei dati, consigliamo di proteggere Account AWS le credenziali e configurare i singoli utenti con AWS IAM Identity Center or AWS Identity and Access Management (IAM). In tal modo, a ogni utente verranno assegnate solo le autorizzazioni necessarie per svolgere i suoi compiti. Ti suggeriamo, inoltre, di proteggere i dati nei seguenti modi:

- Utilizza l'autenticazione a più fattori (MFA) con ogni account.
- Usa SSL/TLS per comunicare con le risorse. AWS È richiesto TLS 1.2 ed è consigliato TLS 1.3.
- Configura l'API e la registrazione delle attività degli utenti con. AWS CloudTrail
- Utilizza soluzioni di AWS crittografia, insieme a tutti i controlli di sicurezza predefiniti all'interno Servizi AWS.
- Utilizza i servizi di sicurezza gestiti avanzati, come Amazon Macie, che aiutano a individuare e proteggere i dati sensibili archiviati in Amazon S3.
- Se hai bisogno di moduli crittografici convalidati FIPS 140-2 per l'accesso AWS tramite un'interfaccia a riga di comando o un'API, utilizza un endpoint FIPS. Per ulteriori informazioni sugli endpoint FIPS disponibili, consulta il Federal Information Processing Standard (FIPS) 140-2.

Ti consigliamo vivamente di non inserire mai informazioni riservate o sensibili, ad esempio gli indirizzi e-mail dei clienti, nei tag o nei campi di testo in formato libero, ad esempio nel campo Nome. Ciò include quando lavori con Application Auto Scaling o altro Servizi AWS utilizzando la console, l'API o AWS gli AWS CLI SDK. I dati inseriti nei tag o nei campi di testo in formato libero utilizzati per i nomi possono essere utilizzati per i la fatturazione o i log di diagnostica. Quando fornisci un URL a un server esterno, ti suggeriamo vivamente di non includere informazioni sulle credenziali nell'URL per convalidare la tua richiesta al server.

Identity and Access Management per Application Auto Scaling

AWS Identity and Access Management (IAM) è un software Servizio AWS che aiuta un amministratore a controllare in modo sicuro l'accesso alle risorse. AWS Gli amministratori IAM controllano chi può essere autenticato (con accesso effettuato) e autorizzato (che dispone di autorizzazioni) a utilizzare risorse di Application Auto Scaling. IAM è un software Servizio AWS che puoi utilizzare senza costi aggiuntivi.

Per la documentazione IAM completa, consulta la Guida per l'utente IAM.

Controllo accessi

Per autenticare le richieste, è necessario disporre di credenziali valide, ma a meno che non si disponga delle autorizzazioni non è possibile creare o accedere alle risorse Application Auto Scaling. Ad esempio, è necessario disporre delle autorizzazioni per creare criteri di dimensionamento, configurare il dimensionamento pianificato e così via.

Le seguenti sezioni forniscono dettagli su come un amministratore IAM può utilizzare IAM per proteggere le AWS risorse, controllando chi può eseguire le azioni dell'API Application Auto Scaling.

Indice

Come funziona Application Auto Scaling con IAM

> **Note**
>
> Nel mese di dicembre 2017 è stato installato un aggiornamento per Application Auto Scaling, abilitando diversi ruoli collegati ai servizi per i servizi integrati Application Auto Scaling. Sono necessarie autorizzazioni IAM specifiche e un ruolo collegato al servizio Application Auto

Scaling (o un ruolo di servizio per la scalabilità automatica di Amazon EMR) in modo che gli utenti possano configurare il dimensionamento.

Prima di utilizzare IAM per gestire l'accesso ad Application Auto Scaling, è necessario imparare quali funzioni IAM sono disponibili per l'uso con Application Auto Scaling.

Funzionalità IAM che è possibile utilizzare con Application Auto Scaling

Funzionalità IAM	Supporto di Application Auto Scaling
Policy basate su identità	Sì
Azioni di policy	Sì
Risorse relative alle policy	Sì
Chiavi di condizione della policy (specifica del servizio)	Sì
Policy basate su risorse	No
Liste di controllo degli accessi (ACL)	No
ABAC (tag nelle policy)	Parziale
Credenziali temporanee	Sì
Ruoli di servizio	Sì
Ruoli collegati al servizio	Sì

Per avere una panoramica di alto livello su come Application Auto Scaling e Servizi AWS altre funzioni funzionano con la maggior parte delle funzionalità IAM, Servizi AWS consulta la sezione dedicata alla compatibilità con IAM nella IAM User Guide.

Policy basate su identità di Application Auto Scaling

Supporta le policy basate su identità	Sì

Le policy basate su identità sono documenti di policy di autorizzazione JSON che è possibile allegare a un'identità (utente, gruppo di utenti o ruolo IAM). Tali policy definiscono le azioni che utenti e ruoli possono eseguire, su quali risorse e in quali condizioni. Per informazioni su come creare una policy basata su identità, consulta Creazione di policy IAM nella Guida per l'utente IAM.

Con le policy basate su identità di IAM, è possibile specificare quali operazioni e risorse sono consentite o respinte, nonché le condizioni in base alle quali le operazioni sono consentite o respinte. Non è possibile specificare l'entità principale in una policy basata sull'identità perché si applica all'utente o al ruolo a cui è associato. Per informazioni su tutti gli elementi utilizzabili in una policy JSON, consulta Guida di riferimento agli elementi delle policy JSON IAM nella Guida per l'utente di IAM.

Esempi di policy basate su identità per Application Auto Scaling

Per visualizzare esempi di policy basate su identità Application Auto Scaling, consulta Esempi di policy basate su identità di Application Auto Scaling.

Azioni

Supporta le operazioni di policy	Sì

In una dichiarazione di policy IAM, è possibile specificare qualsiasi operazione API per qualsiasi servizio che supporta IAM. Per Application Auto Scaling, utilizzare il seguente prefisso con il nome dell'operazione API: `application-autoscaling:`. For example: `application-autoscaling:RegisterScalableTarget`, `application-autoscaling:PutScalingPolicy` e `application-autoscaling:DeregisterScalableTarget`.

Per specificare più operazioni in una singola istruzione, separale con virgole, come illustrato nell'esempio seguente.

```
"Action": [
      "application-autoscaling:DescribeScalingPolicies",
      "application-autoscaling:DescribeScalingActivities"
```

Puoi specificare più operazioni tramite caratteri jolly (*). Ad esempio, per specificare tutte le operazioni che iniziano con la parola `Describe`, includi la seguente operazione.

```
"Action": "application-autoscaling:Describe*"
```

Per un elenco delle azioni di Application Auto Scaling, vedere Azioni definite da AWS Application Auto Scaling nel Service Authorization Reference.

Risorse

Supporta le risorse di policy	Sì

In una dichiarazione di policy IAM, l'elemento `Resource` specifica l'oggetto o gli oggetti coperti dall'istruzione. Per l'Applicazione di Dimensionamento automatico, ogni dichiarazione di policy IAM si applica agli obiettivi scalabili specificati utilizzando i relativi nomi delle risorse Amazon (ARN).

Il formato della risorsa ARN per obiettivi scalabili:

```
arn:aws:application-autoscaling:region:account-id:scalable-target/unique-identifier
```

Ad esempio, nell'istruzione puoi indicare un obiettivo scalabile utilizzando il relativo ARN come segue: L'ID univoco (1234abcd56ab78cd901ef1234567890ab123) è un valore assegnato dall'Applicazione di Dimensionamento automatico all'obiettivo scalabile.

```
"Resource": "arn:aws:application-autoscaling:us-east-1:123456789012:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
```

Puoi specificare tutte le istanze appartenenti a un determinato account sostituendo l'identificatore univoco con il carattere jolly (*) come segue:

```
"Resource": "arn:aws:application-autoscaling:us-east-1:123456789012:scalable-target/*"
```

Per specificare tutte le risorse o se una determinata operazione API non supporta gli ARN, utilizza il carattere jolly (*) come l'elemento `Resource` nel modo seguente:

```
"Resource": "*"
```

Per ulteriori informazioni, vedere Tipi di risorse definiti da AWS Application Auto Scaling nel Service Authorization Reference.

Chiavi di condizione

Supporta le chiavi di condizione delle policy specifiche del servizio	Sì

È possibile specificare le condizioni nelle policy IAM che controllano l'accesso alle risorse dell'Applicazione di Dimensionamento automatico. L'istruzione di policy diventa effettiva solo quando le condizioni sono true.

L'Applicazione di Dimensionamento automatico supporta le seguenti chiavi di condizione definite dal servizio che puoi utilizzare nelle policy basate sull'identità per determinare chi può eseguire le azioni dell'API dell'Applicazione di Dimensionamento automatico.

- `application-autoscaling:scalable-dimension`
- `application-autoscaling:service-namespace`

Per sapere con quali azioni dell'API Application Auto Scaling è possibile utilizzare una chiave di condizione, vedere Azioni definite da AWS Application Auto Scaling nel Service Authorization Reference. Per ulteriori informazioni sull'utilizzo delle chiavi di condizione di Application Auto Scaling, vedere Chiavi di condizione per AWS Application Auto Scaling.

Per visualizzare le chiavi di condizione globali disponibili per tutti i servizi, consulta chiavi di condizione globali contestuali AWS nella Guida dell'utente IAM.

Policy basate su risorse

Supporta le policy basate su risorse	No

Altri AWS servizi, come Amazon Simple Storage Service, supportano politiche di autorizzazione basate sulle risorse. Ad esempio, è possibile allegare una policy di autorizzazione a un bucket S3 per gestire le autorizzazioni di accesso a quel bucket.

Application Auto Scaling non supporta le policy basate su risorse.

Liste di controllo degli accessi (ACL)

Supporta le ACL	No

Application Auto Scaling non supporta le liste di controllo accessi (ACL).

ABAC con l'Applicazione di Dimensionamento automatico

Supporta ABAC (tag nelle policy)	Parziale

Il controllo dell'accesso basato su attributi (ABAC) è una strategia di autorizzazione che definisce le autorizzazioni in base agli attributi. In AWS, questi attributi sono chiamati tag. Puoi allegare tag a entità IAM (utenti o ruoli) e a molte AWS risorse. L'assegnazione di tag alle entità e alle risorse è il primo passaggio di ABAC. In seguito, vengono progettate policy ABAC per consentire operazioni quando il tag dell'entità principale corrisponde al tag sulla risorsa a cui si sta provando ad accedere.

La strategia ABAC è utile in ambienti soggetti a una rapida crescita e aiuta in situazioni in cui la gestione delle policy diventa impegnativa.

Per controllare l'accesso basato su tag, fornisci informazioni sui tag nell'elemento condizione di una policy utilizzando le chiavi di condizione `aws:ResourceTag/key-name`, `aws:RequestTag/key-name` o `aws:TagKeys`.

ABAC è possibile per le risorse che supportano i tag, ma non tutto supporta i tag. Le azioni pianificate e le policy di dimensionamento non supportano i tag, ma gli obiettivi scalabili sì. Per ulteriori informazioni, consulta Supporto al tagging per l'Applicazione di Dimensionamento automatico.

Per ulteriori informazioni su ABAC, consulta Che cos'è ABAC? nella Guida per l'utente IAM. Per visualizzare un tutorial con i passaggi per l'impostazione di ABAC, consulta Utilizzo del controllo degli accessi basato su attributi (ABAC) nella Guida per l'utente di IAM.

Utilizzo di credenziali temporanee con Application Auto Scaling

Supporta le credenziali temporanee	Sì

Alcuni Servizi AWS non funzionano quando accedi utilizzando credenziali temporanee. Per ulteriori informazioni, incluse quelle che Servizi AWS funzionano con credenziali temporanee, consulta la sezione relativa alla Servizi AWS compatibilità con IAM nella IAM User Guide.

Stai utilizzando credenziali temporanee se accedi AWS Management Console utilizzando qualsiasi metodo tranne nome utente e password. Ad esempio, quando accedi AWS utilizzando il link Single Sign-On (SSO) della tua azienda, tale processo crea automaticamente credenziali temporanee. Le credenziali temporanee vengono create in automatico anche quando accedi alla console come utente e poi cambi ruolo. Per ulteriori informazioni sullo scambio dei ruoli, consulta Cambio di un ruolo (console) nella Guida per l'utente IAM.

È possibile creare manualmente credenziali temporanee utilizzando l'API or. AWS CLI AWS È quindi possibile utilizzare tali credenziali temporanee per accedere. AWS AWS consiglia di generare dinamicamente credenziali temporanee anziché utilizzare chiavi di accesso a lungo termine. Per ulteriori informazioni, consulta Credenziali di sicurezza provvisorie in IAM.

Ruoli di servizio

Supporta i ruoli di servizio	Sì

Se il cluster Amazon EMR utilizza la scalabilità automatica, questa funzionalità consente ad Application Auto Scaling di assumere un ruolo di servizio per tuo conto. Analogamente a un ruolo collegato ai servizi, un ruolo di servizio consente al servizio di accedere alle risorse di altri servizi per completare un'azione per conto dell'utente. I ruoli dei servizi sono visualizzati nell'account IAM e sono di proprietà dell'account. Ciò significa che un amministratore IAM può modificare le autorizzazioni per questo ruolo. Tuttavia, questo potrebbe pregiudicare la funzionalità del servizio.

Application Auto Scaling supporta i ruoli di servizio solo per Amazon EMR. Per la documentazione relativa al ruolo del servizio EMR, consulta Utilizzo della scalabilità automatica con una policy personalizzata per i gruppi di istanze nella Guida alla gestione di Amazon EMR.

> **Note**
>
> Con l'introduzione di ruoli collegati ai servizi, non sono più necessari diversi ruoli del servizio legacy, ad esempio per Amazon ECS e Parco istanze Spot.

Ruoli collegati ai servizi

Supporta i ruoli collegati ai servizi	Sì

Un ruolo collegato al servizio è un tipo di ruolo di servizio collegato a un. Servizio AWS Il servizio può assumere il ruolo per eseguire un'azione per tuo conto. I ruoli collegati al servizio vengono visualizzati nel tuo account Account AWS e sono di proprietà del servizio. Un amministratore IAM può visualizzare le autorizzazioni per i ruoli collegati ai servizi, ma non modificarle.

Per ulteriori informazioni sui ruoli collegati ai servizi di Application Auto Scaling, consultare Ruoli collegati ai servizi per Application Auto Scaling.

AWS politiche gestite per Application Auto Scaling

Una policy AWS gestita è una policy autonoma creata e amministrata da. AWS AWS le politiche gestite sono progettate per fornire autorizzazioni per molti casi d'uso comuni, in modo da poter iniziare ad assegnare autorizzazioni a utenti, gruppi e ruoli.

Tieni presente che le policy AWS gestite potrebbero non concedere le autorizzazioni con il privilegio minimo per i tuoi casi d'uso specifici, poiché sono disponibili per tutti i clienti. AWS Ti consigliamo pertanto di ridurre ulteriormente le autorizzazioni definendo policy gestite dal cliente specifiche per i tuoi casi d'uso.

Non è possibile modificare le autorizzazioni definite nelle politiche gestite. AWS Se AWS aggiorna le autorizzazioni definite in una politica AWS gestita, l'aggiornamento ha effetto su tutte le identità principali (utenti, gruppi e ruoli) a cui è associata la politica. AWS è più probabile che aggiorni una policy AWS gestita quando ne Servizio AWS viene lanciata una nuova o quando diventano disponibili nuove operazioni API per i servizi esistenti.

Per ulteriori informazioni, consultare Policy gestite da AWSnella Guida per l'utente di IAM.

AWS politica gestita: AppStream 2.0 e CloudWatch

Nome della policy: AWSApplicationAutoscalingAppStreamFleetPolicy

Questa policy è associata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_AppStreamFleetper consentire ad Application Auto Scaling di chiamare AppStream Amazon CloudWatch ed eseguire il ridimensionamento per tuo conto.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni su tutte le risorse correlate («Resource»: «*»):

- Operazione: `appstream:DescribeFleets`
- Operazione: `appstream:UpdateFleet`
- Operazione: `cloudwatch:DescribeAlarms`
- Operazione: `cloudwatch:PutMetricAlarm`
- Operazione: `cloudwatch:DeleteAlarms`

AWS politica gestita: Aurora e CloudWatch

Nome della policy: AWSApplicationAutoscalingRDSClusterPolicy

Questa policy è allegata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_RDSClusterper consentire ad Application Auto Scaling di chiamare Aurora CloudWatch ed eseguire il ridimensionamento per conto dell'utente.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni su tutte le risorse correlate («Resource»: «*»):

- Operazione: `rds:AddTagsToResource`
- Operazione: `rds:CreateDBInstance`
- Operazione: `rds:DeleteDBInstance`
- Operazione: `rds:DescribeDBClusters`
- Operazione: `rds:DescribeDBInstance`
- Operazione: `cloudwatch:DescribeAlarms`
- Operazione: `cloudwatch:PutMetricAlarm`
- Operazione: `cloudwatch:DeleteAlarms`

AWS politica gestita: Amazon Comprehend e CloudWatch

Nome della policy: AWSApplicationAutoscalingComprehendEndpointPolicy

Questa policy è allegata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_ComprehendEndpointper consentire ad Application Auto Scaling di chiamare Amazon Comprehend CloudWatch ed eseguire il ridimensionamento per tuo conto.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni su tutte le risorse correlate («Resource»: «*»):

- Operazione: `comprehend:UpdateEndpoint`
- Operazione: `comprehend:DescribeEndpoint`
- Operazione: `cloudwatch:DescribeAlarms`
- Operazione: `cloudwatch:PutMetricAlarm`
- Operazione: `cloudwatch:DeleteAlarms`

AWS policy gestita: DynamoDB e CloudWatch

Nome della policy: AWSApplicationAutoscalingDynamoDBTablePolicy

Questa policy è associata al ruolo collegato al servizio denominato per consentire AWSServiceRoleForApplicationAutoScaling_DynamoDBTablead Application Auto Scaling di chiamare CloudWatch DynamodBand ed eseguire il ridimensionamento per conto dell'utente.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni su tutte le risorse correlate («Resource»: «*»):

- Operazione: `dynamodb:DescribeTable`
- Operazione: `dynamodb:UpdateTable`
- Operazione: `cloudwatch:DescribeAlarms`
- Operazione: `cloudwatch:PutMetricAlarm`
- Operazione: `cloudwatch:DeleteAlarms`

AWS politica gestita: Amazon ECS e CloudWatch

Nome della policy: AWSApplicationAutoscalingECSServicePolicy

Questa policy è associata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_ECSServiceper consentire ad Application Auto Scaling di chiamare Amazon ECS CloudWatch ed eseguire il ridimensionamento per tuo conto.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni su tutte le risorse correlate («Resource»: «*»):

- Operazione: `ecs:DescribeServices`
- Operazione: `ecs:UpdateService`
- Operazione: `cloudwatch:DescribeAlarms`
- Operazione: `cloudwatch:PutMetricAlarm`
- Operazione: `cloudwatch:DeleteAlarms`

AWS politica gestita: e ElastiCache CloudWatch

Nome della policy: AWSApplicationAutoscalingElastiCacheRGPolicy

Questa policy è allegata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_ElastiCacheRGper consentire ad Application Auto Scaling di ElastiCache chiamare CloudWatch ed eseguire il ridimensionamento per conto dell'utente.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni sulle risorse specificate:

- Operazione: `elasticache:DescribeReplicationGroups` su tutte le risorse
- Operazione: `elasticache:ModifyReplicationGroupShardConfiguration` su tutte le risorse
- Operazione: `elasticache:IncreaseReplicaCount` su tutte le risorse
- Operazione: `elasticache:DecreaseReplicaCount` su tutte le risorse
- Operazione: `elasticache:DescribeCacheClusters` su tutte le risorse
- Operazione: `elasticache:DescribeCacheParameters` su tutte le risorse
- Operazione: `cloudwatch:DescribeAlarms` su tutte le risorse

- Operazione: `cloudwatch:PutMetricAlarm` sulla risorsa `arn:aws:cloudwatch:*:*:alarm:TargetTracking*`
- Operazione: `cloudwatch:DeleteAlarms` sulla risorsa `arn:aws:cloudwatch:*:*:alarm:TargetTracking*`
- Operazione: `cloudwatch:DeleteAlarms`

AWS politica gestita: Amazon Keyspaces e CloudWatch

Nome della policy: AWSApplicationAutoscalingCassandraTablePolicy

Questa policy è associata al ruolo collegato al servizio denominato per consentire AWSServiceRoleForApplicationAutoScaling_CassandraTablead Application Auto Scaling di chiamare Amazon Keyspaces CloudWatch ed eseguire il ridimensionamento per tuo conto.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni sulle risorse specificate:

- Azione: `cassandra:Select` sulle seguenti risorse:
  - `arn:*:cassandra:*:*:/keyspace/system/table/*`
  - `arn:*:cassandra:*:*:/keyspace/system_schema/table/*`
  - `arn:*:cassandra:*:*:/keyspace/system_schema_mcs/table/*`
- Operazione: `cassandra:Alter` su tutte le risorse
- Operazione: `cloudwatch:DescribeAlarms` su tutte le risorse
- Operazione: `cloudwatch:PutMetricAlarm` su tutte le risorse
- Operazione: `cloudwatch:DeleteAlarms` su tutte le risorse

AWS politica gestita: Lambda e CloudWatch

Nome della policy: AWSApplicationAutoscalingLambdaConcurrencyPolicy

Questa policy è allegata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_LambdaConcurrencyper consentire ad Application Auto Scaling di chiamare Lambda CloudWatch ed eseguire il ridimensionamento per tuo conto.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni su tutte le risorse correlate («Resource»: «*»):

- Operazione: `lambda:PutProvisionedConcurrencyConfig`
- Operazione: `lambda:GetProvisionedConcurrencyConfig`
- Operazione: `lambda:DeleteProvisionedConcurrencyConfig`
- Operazione: `cloudwatch:DescribeAlarms`
- Operazione: `cloudwatch:PutMetricAlarm`
- Operazione: `cloudwatch:DeleteAlarms`

AWS politica gestita: Amazon MSK e CloudWatch

Nome della policy: AWSApplicationAutoscalingKafkaClusterPolicy

Questa policy è associata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_KafkaClusterper consentire ad Application Auto Scaling di chiamare Amazon MSK CloudWatch ed eseguire il ridimensionamento per tuo conto.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni su tutte le risorse correlate («Resource»: «*»):

- Operazione: `kafka:DescribeCluster`
- Operazione: `kafka:DescribeClusterOperation`
- Operazione: `kafka:UpdateBrokerStorage`
- Operazione: `cloudwatch:DescribeAlarms`
- Operazione: `cloudwatch:PutMetricAlarm`
- Operazione: `cloudwatch:DeleteAlarms`

AWS politica gestita: Neptune e CloudWatch

Nome della policy: AWSApplicationAutoscalingNeptuneClusterPolicy

Questa policy è allegata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_NeptuneClusterper consentire ad Application Auto Scaling di chiamare Neptune CloudWatch ed eseguire il ridimensionamento per conto dell'utente.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni sulle risorse specificate:

- Operazione: `rds:ListTagsForResource` su tutte le risorse
- Operazione: `rds:DescribeDBInstances` su tutte le risorse
- Operazione: `rds:DescribeDBClusters` su tutte le risorse
- Operazione: `rds:DescribeDBClusterParameters` su tutte le risorse
- Operazione: `cloudwatch:DescribeAlarms` su tutte le risorse
- Operazione: `rds:AddTagsToResource` sulle risorse con il prefisso autoscaled-reader nel motore del database Amazon Neptune (`"Condition":{"StringEquals": {"rds:DatabaseEngine":"neptune"}`)
- Operazione: `rds:CreateDBInstance` sulle risorse con il prefisso autoscaled-reader in tutti i cluster DB (`"Resource":"arn:*:rds:*:*:db:autoscaled-reader*", "arn:aws:rds:*:*:cluster:*"`) nel motore del database Amazon Neptune (`"Condition": {"StringEquals":{"rds:DatabaseEngine":"neptune"}`)
- Operazione: `rds:DeleteDBInstance` sulla risorsa `arn:aws:rds:*:*:db:autoscaled-reader*`
- Operazione: `cloudwatch:PutMetricAlarm` sulla risorsa `arn:aws:cloudwatch:*:*:alarm:TargetTracking*`
- Operazione: `cloudwatch:DeleteAlarms` sulla risorsa `arn:aws:cloudwatch:*:*:alarm:TargetTracking*`

AWS politica gestita: e SageMaker CloudWatch

Nome della policy: AWSApplicationAutoscalingSageMakerEndpointPolicy

Questa policy è allegata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_SageMakerEndpointper consentire ad Application Auto Scaling di SageMaker chiamare CloudWatch ed eseguire il ridimensionamento per conto dell'utente.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni sulle risorse specificate:

- Operazione: `sagemaker:DescribeEndpoint` su tutte le risorse
- Operazione: `sagemaker:DescribeEndpointConfig` su tutte le risorse
- Operazione: `sagemaker:DescribeInferenceComponent` su tutte le risorse
- Operazione: `sagemaker:UpdateEndpointWeightsAndCapacities` su tutte le risorse
- Operazione: `sagemaker:UpdateInferenceComponentRuntimeConfig` su tutte le risorse
- Operazione: `cloudwatch:DescribeAlarms` su tutte le risorse
- Operazione: `cloudwatch:GetMetricData` su tutte le risorse
- Operazione: `cloudwatch:PutMetricAlarm` sulla risorsa `arn:aws:cloudwatch:*:*:alarm:TargetTracking*`
- Operazione: `cloudwatch:DeleteAlarms` sulla risorsa `arn:aws:cloudwatch:*:*:alarm:TargetTracking*`

AWS politica gestita: EC2 Spot Fleet e CloudWatch

Nome della policy: AWSApplicationAutoscalingEC2SpotFleetRequestPolicy

Questa policy è allegata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_EC2 SpotFleet Request per consentire ad Application Auto Scaling di chiamare Amazon EC2 CloudWatch ed eseguire il ridimensionamento per tuo conto.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni su tutte le risorse correlate («Resource»: «*»):

- Operazione: `ec2:DescribeSpotFleetRequests`
- Operazione: `ec2:ModifySpotFleetRequest`
- Operazione: `cloudwatch:DescribeAlarms`
- Operazione: `cloudwatch:PutMetricAlarm`
- Operazione: `cloudwatch:DeleteAlarms`

AWS politica gestita: risorse personalizzate e CloudWatch

Nome della policy: AWSApplicationAutoScalingCustomResourcePolicy

Questa policy è associata al ruolo collegato al servizio denominato AWSServiceRoleForApplicationAutoScaling_CustomResourceper consentire ad Application Auto Scaling di richiamare le risorse personalizzate disponibili tramite API Gateway CloudWatch ed eseguire il ridimensionamento per conto dell'utente.

Dettagli dell'autorizzazione

La politica di autorizzazione consente ad Application Auto Scaling di completare le seguenti azioni su tutte le risorse correlate («Resource»: «*»):

- Operazione: `execute-api:Invoke`
- Operazione: `cloudwatch:DescribeAlarms`
- Operazione: `cloudwatch:PutMetricAlarm`
- Operazione: `cloudwatch:DeleteAlarms`

Application Auto Scaling aggiorna le policy gestite AWS

Visualizza i dettagli sugli aggiornamenti delle policy AWS gestite per Application Auto Scaling da quando questo servizio ha iniziato a tenere traccia di queste modifiche. Per gli avvisi automatici sulle modifiche apportate a questa pagina, sottoscrivi il feed RSS nella pagina della cronologia dei documenti di Application Auto Scaling.

Modifica	Descrizione	Data
Application Auto Scaling aggiunge le autorizzazioni al suo ruolo collegato al servizio SageMaker	Questa politica ora concede al servizio le autorizzazioni per richiamare le azioni SageMaker `DescribeInferenceComponent` e le `UpdateInferenceComponentRuntimeConfig` API per supportare la compatibilità per il ridimensionamento automatico delle	13 novembre 2023

Modifica	Descrizione	Data
	SageMaker risorse per un'integrazione imminente . La policy ora limita inoltre le azioni CloudWatch `PutMetricAlarm` e le `DeleteAlarms` API agli CloudWatch allarmi utilizzati con le politiche di scalabilità di Target Tracking.	
Application Auto Scaling aggiunge la policy di Neptune	Application Auto Scaling ha aggiunto una nuova policy gestita per Neptune. Questa policy è associata a un ruolo collegato al servizio che consente ad Application Auto Scaling di chiamare Neptune CloudWatch ed eseguire il ridimensionamento per conto dell'utente.	6 ottobre 2021
Application Auto Scaling aggiunge la policy ElastiCache Redis	Application Auto Scaling ha aggiunto una nuova policy gestita per. ElastiCache Questa policy è associata a un ruolo collegato al servizio che consente ad Applicati on Auto Scaling di ElastiCac he chiamare CloudWatch ed eseguire il ridimensionamento per conto dell'utente.	19 agosto 2021

Modifica	Descrizione	Data
Application Auto Scaling ha iniziato a monitorare le modifiche	Application Auto Scaling ha iniziato a tenere traccia delle modifiche per le sue politiche AWS gestite.	19 agosto 2021

Ruoli collegati ai servizi per Application Auto Scaling

Application Auto Scaling utilizza ruoli collegati ai servizi per le autorizzazioni necessarie per chiamare altri AWS servizi per conto dell'utente. Un ruolo collegato al servizio è un tipo unico di ruolo AWS Identity and Access Management (IAM) collegato direttamente a un servizio. AWS I ruoli collegati ai servizi forniscono un modo sicuro per delegare le autorizzazioni ai AWS servizi perché solo il servizio collegato può assumere un ruolo collegato al servizio.

Per i servizi che si integrano con la Application Auto Scaling, Application Auto Scaling crea ruoli collegati ai servizi per tuo conto. Esiste un ruolo collegato ai servizi per ciascun servizio. Ogni ruolo collegato ai servizi considera attendibile il principale del servizio specificato ai fini dell'assunzione del ruolo. Per ulteriori informazioni, consulta Riferimento ARN del ruolo collegato ai servizi.

Application Auto Scaling include tutte le autorizzazioni necessarie per ciascun ruolo collegato ai servizi. Queste autorizzazioni gestite vengono create e gestite da Application Auto Scaling e definiscono le operazioni consentite per ogni tipo di risorsa. Per informazioni dettagliate sulle autorizzazioni concesse da ciascun ruolo, consulta AWS politiche gestite per Application Auto Scaling.

Indice

Autorizzazioni necessarie per creare un ruolo collegato al servizio

Application Auto Scaling richiede le autorizzazioni per creare un ruolo collegato al servizio la prima volta che un utente Account AWS chiama `RegisterScalableTarget` per un determinato servizio. Application Auto Scaling crea un ruolo collegato ai servizi per il servizio obiettivo nel tuo account, se non esiste ancora. Il ruolo collegato ai servizi concede le autorizzazioni ad Application Auto Scaling in modo che possa eseguire chiamate al servizio obiettivo per tuo conto.

Affinché la creazione automatica di un ruolo riesca, gli utenti devono disporre dell'autorizzazione per l'operazione `iam:CreateServiceLinkedRole`.

```
"Action": "iam:CreateServiceLinkedRole"
```

Di seguito viene illustrato un esempio di policy basata su identità che concede l'autorizzazione di creare un ruolo collegato ai servizi per il Parco istanze Spot. È possibile specificare il ruolo collegato ai servizi nel campo `Resource` della policy e come ARN, e il principale del servizio per il ruolo collegato ai servizi come condizione, come mostrato. Per l'ARN per ciascun servizio, consulta Riferimento ARN del ruolo collegato ai servizi.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "iam:CreateServiceLinkedRole",
            "Resource": "arn:aws:iam::*:role/aws-service-role/ec2.application-
autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest",
            "Condition": {
                "StringLike": {
                    "iam:AWSServiceName":"ec2.application-autoscaling.amazonaws.com"
                }
            }
        }
    ]
}
```

> **Note**
>
> La chiave di condizione IAM `iam:AWSServiceName` specifica il principale del servizio a cui è associato il ruolo, indicato in questa policy di esempio come *ec2.application-*

autoscaling.amazonaws.com. Non procedere per tentativi con il principale del servizio. Per visualizzare il principale del servizio, consulta Servizi AWS che puoi usare con Application Auto Scaling.

Creazione di ruoli collegati ai servizi (procedura automatica)

Non hai bisogno di creare manualmente un ruolo collegato ai servizi. Application Auto Scaling crea il ruolo collegato ai servizi appropriato per tuo conto quando chiami `RegisterScalableTarget`. Ad esempio, se imposti la funzione di scalabilità automatica per un servizio Amazon ECS, Application Auto Scaling crea il ruolo `AWSServiceRoleForApplicationAutoScaling_ECSService`.

Creazione di ruoli collegati ai servizi (procedura manuale)

Per creare il ruolo collegato al servizio, puoi utilizzare la console IAM o l'API IAM. AWS CLI Per ulteriori informazioni, consulta Creazione di un ruolo collegato ai servizi nella Guida per l'utente di IAM.

Come creare un ruolo collegato ai servizi (AWS CLI)

Usa il comando della CLI create-service-linked-role per creare il ruolo collegato ai servizi per Application Auto Scaling. Nella richiesta, specifica il "prefisso" del nome del servizio.

Per trovare il prefisso del nome del servizio, fai riferimento alle informazioni sul principale servizio per il ruolo collegato ai servizi per ciascun servizio nella sezione Servizi AWS che puoi usare con Application Auto Scaling. Il nome del servizio e il principale del servizio hanno lo stesso prefisso. Ad esempio, per creare il ruolo AWS Lambda collegato al servizio, usa. `lambda.application-autoscaling.amazonaws.com`

```
aws iam create-service-linked-role --aws-service-name prefix.application-
autoscaling.amazonaws.com
```

Modifica di ruoli collegati ai servizi

Dei ruoli collegati ai servizi creati da Application Auto Scaling puoi modificare solo le descrizioni. Per ulteriori informazioni, consulta Modifica di un ruolo collegato ai servizi nella Guida per l'utente di IAM.

Eliminazione di ruoli collegati ai servizi

Se non utilizzi più Application Auto Scaling con un tipo di servizio supportato, ti consigliamo di rimuovere il ruolo collegato ai servizi corrispondente.

È possibile eliminare un ruolo collegato ai servizi solo dopo avere eliminato le risorse AWS correlate. Questo ti protegge dalla possibilità di revocare inavvertitamente le autorizzazioni Application Auto Scaling per le risorse. Per ulteriori informazioni, consulta la documentazione relativa alla risorsa scalabile. Ad esempio, per eliminare un servizio Amazon ECS, consulta Eliminazione di un servizio nella Guida per gli sviluppatori di Amazon Elastic Container Service.

Per eliminare un ruolo collegato ai servizi, puoi utilizzare IAM. Per ulteriori informazioni, consulta Eliminazione del ruolo collegato ai servizi nella Guida per l'utente di IAM.

Un ruolo collegato ai servizi che è stato eliminato viene creato nuovamente da Application Auto Scaling quando chiami `RegisterScalableTarget`.

Regioni supportate per i ruoli collegati ai servizi Application Auto Scaling

Application Auto Scaling supporta l'utilizzo di ruoli collegati ai servizi in tutte le AWS regioni in cui il servizio è disponibile.

Riferimento ARN del ruolo collegato ai servizi

La tabella seguente elenca l'Amazon Resource Name (ARN) del ruolo collegato al servizio per ogni ruolo Servizio AWS che funziona con Application Auto Scaling.

Servizio	ARN
AppStream 2.0	`arn:aws:iam:: 012345678910 :role/aws-service-role/appstream.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_AppStreamFleet`
Aurora	`arn:aws:iam:: 012345678910 :role/aws-service-role/rds.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_RDSCluster`
Comprehend	`arn:aws:iam:: 012345678910 :role/aws-service-role/comprehend.application-autoscaling.amazonaws.com/`

Servizio	ARN
	AWSServiceRoleForApplicationAutoScaling_Comprehendendpoint
DynamoDB	arn:aws:iam:: *012345678910* :role/aws-service-role/dynamodb.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_DynamoDBTable
ECS	arn:aws:iam:: *012345678910* :role/aws-service-role/ecs.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_ECSService
ElastiCache	arn:aws:iam:: *012345678910* :role/aws-service-role/elasticache.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_ElastiCacheRG
Keyspaces	arn:aws:iam:: *012345678910* :role/aws-service-role/cassandra.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_CassandraTable
Lambda	arn:aws:iam:: *012345678910* :role/aws-service-role/lambda.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_LambdaConcurrency
MSK	arn:aws:iam:: *012345678910* :role/aws-service-role/kafka.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_KafkaCluster
Neptune	arn:aws:iam:: *012345678910* :role/aws-service-role/neptune.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_NeptuneCluster

Servizio	ARN
SageMaker	`arn:aws:iam:: 012345678910 :role/aws-service-role/sagemaker.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_SageMakerEndpoint`
Parco istanze Spot	`arn:aws:iam:: 012345678910 :role/aws-service-role/ec2.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest`
Risorse personalizzate	`arn:aws:iam:: 012345678910 :role/aws-service-role/custom-resource.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_CustomResource`

> **Note**
>
> Puoi specificare l'ARN di un ruolo collegato al servizio per la `RoleARN` proprietà di una AWS::ApplicationAutoScaling::ScalableTargetrisorsa nei tuoi modelli di AWS CloudFormation stack, anche se il ruolo collegato al servizio specificato non esiste ancora. Application Auto Scaling crea automaticamente il ruolo per tuo conto.

Esempi di policy basate su identità di Application Auto Scaling

Per impostazione predefinita, un nuovo utente non Account AWS ha le autorizzazioni per fare nulla. Un amministratore IAM deve creare e assegnare policy IAM che diano un'autorizzazione di identità IAM (ad esempio un utente o un ruolo) per eseguire operazioni API di Application Auto Scaling.

Per informazioni su come creare una policy IAM utilizzando i seguenti documenti di policy JSON di esempio, consulta Creazione di policy nella scheda JSON nella Guida per l'utente di IAM.

Indice

- Autorizzazioni per lavorare in AWS Management Console

Autorizzazioni necessarie per le operazioni API Application Auto Scaling

Le seguenti policy concedono le autorizzazioni per i casi d'uso comune durante la chiamata all'API Application Auto Scaling. Considera questa sezione quando scrivi policy basate su identità. Ogni policy concede autorizzazioni per accedere ad alcune o a tutte le operazioni API Application Auto Scaling. È inoltre necessario assicurarsi che gli utenti finali dispongano delle autorizzazioni per il servizio di destinazione e CloudWatch (consulta la sezione successiva per i dettagli).

La seguente policy basata su identità concede autorizzazioni ad alcune o a tutte le operazioni API Application Auto Scaling.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "application-autoscaling:*"
            ],
            "Resource": "*"
        }
    ]
}
```

Le seguenti policy basate su identità consentono autorizzazioni a tutte le operazioni API Application Auto Scaling necessarie per configurare le policy di dimensionamento e le operazioni non pianificate.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "application-autoscaling:RegisterScalableTarget",
              "application-autoscaling:DescribeScalableTargets",
              "application-autoscaling:DeregisterScalableTarget",
              "application-autoscaling:PutScalingPolicy",
              "application-autoscaling:DescribeScalingPolicies",
              "application-autoscaling:DescribeScalingActivities",
```

```
              "application-autoscaling:DeleteScalingPolicy"
            ],
            "Resource": "*"
        }
    ]
}
```

Le seguenti policy basate su identità consentono autorizzazioni a tutte le operazioni API Application Auto Scaling necessarie per configurare le operazioni pianificate e le policy non correlate al dimensionamento.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "application-autoscaling:RegisterScalableTarget",
              "application-autoscaling:DescribeScalableTargets",
              "application-autoscaling:DeregisterScalableTarget",
              "application-autoscaling:PutScheduledAction",
              "application-autoscaling:DescribeScheduledActions",
              "application-autoscaling:DescribeScalingActivities",
              "application-autoscaling:DeleteScheduledAction"
            ],
            "Resource": "*"
        }
    ]
}
```

Autorizzazioni necessarie per le azioni API sui servizi di destinazione e CloudWatch

Per configurare e utilizzare correttamente Application Auto Scaling con il servizio di destinazione, agli utenti finali devono essere concesse le autorizzazioni per Amazon CloudWatch e per ogni servizio di destinazione per il quale configureranno la scalabilità. Utilizza le seguenti politiche per concedere le autorizzazioni minime necessarie per lavorare con i servizi di destinazione e. CloudWatch

Indice

- Tabelle DynamoDB e indici secondari globali
- Servizi ECS
- ElastiCache gruppi di replica
- Cluster Amazon EMR
- Tabelle di Amazon Keyspaces
- Funzioni Lambda
- Archiviazione broker Amazon Managed Streaming for Apache Kafka (MSK)
- Cluster di Neptune
- SageMaker endpoint
- Parco istanze Spot (Amazon EC2)
- Risorse personalizzate

AppStream 2.0 flotte

La seguente politica basata sull'identità concede le autorizzazioni per tutte le azioni AppStream 2.0 e CloudWatch API richieste.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "appstream:DescribeFleets",
              "appstream:UpdateFleet",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Repliche Aurora

La seguente politica basata sull'identità concede le autorizzazioni a tutte le azioni CloudWatch Aurora e API necessarie.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "rds:AddTagsToResource",
              "rds:CreateDBInstance",
              "rds:DeleteDBInstance",
              "rds:DescribeDBClusters",
              "rds:DescribeDBInstances",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Endpoint di classificazione dei documenti Amazon Comprehend e di riconoscimento delle identità

La seguente politica basata sull'identità concede le autorizzazioni a tutte le azioni Amazon Comprehend e API richieste. CloudWatch

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "comprehend:UpdateEndpoint",
              "comprehend:DescribeEndpoint",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Tabelle DynamoDB e indici secondari globali

La seguente politica basata sull'identità concede le autorizzazioni a tutte le azioni DynamoDB e API necessarie. CloudWatch

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "dynamodb:DescribeTable",
              "dynamodb:UpdateTable",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Servizi ECS

La seguente politica basata sull'identità concede le autorizzazioni a tutte le azioni ECS e API richieste. CloudWatch

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "ecs:DescribeServices",
              "ecs:UpdateService",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
```

```
}
```

ElastiCache gruppi di replica

La seguente politica basata sull'identità concede le autorizzazioni a tutte ElastiCache le azioni API richieste. CloudWatch

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "elasticache:ModifyReplicationGroupShardConfiguration",
              "elasticache:IncreaseReplicaCount",
              "elasticache:DecreaseReplicaCount",
              "elasticache:DescribeReplicationGroups",
              "elasticache:DescribeCacheClusters",
              "elasticache:DescribeCacheParameters",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Cluster Amazon EMR

La seguente policy basata sull'identità concede le autorizzazioni per tutte le azioni Amazon EMR CloudWatch e API richieste.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "elasticmapreduce:ModifyInstanceGroups",
              "elasticmapreduce:ListInstanceGroups",
              "cloudwatch:DescribeAlarms",
```

```
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Tabelle di Amazon Keyspaces

La seguente politica basata sull'identità concede le autorizzazioni per tutte le azioni Amazon Keyspaces CloudWatch e API richieste.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "cassandra:Select",
              "cassandra:Alter",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Funzioni Lambda

La seguente politica basata sull'identità concede le autorizzazioni a tutte le azioni Lambda CloudWatch e API richieste.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "lambda:PutProvisionedConcurrencyConfig",
```

```
                "lambda:GetProvisionedConcurrencyConfig",
                "lambda:DeleteProvisionedConcurrencyConfig",
                "cloudwatch:DescribeAlarms",
                "cloudwatch:PutMetricAlarm",
                "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Archiviazione broker Amazon Managed Streaming for Apache Kafka (MSK)

La seguente politica basata sull'identità concede le autorizzazioni per tutte le azioni Amazon MSK CloudWatch e API richieste.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "kafka:DescribeCluster",
                "kafka:DescribeClusterOperation",
                "kafka:UpdateBrokerStorage",
                "cloudwatch:DescribeAlarms",
                "cloudwatch:PutMetricAlarm",
                "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Cluster di Neptune

La seguente politica basata sull'identità concede le autorizzazioni a tutte le azioni di CloudWatch Neptune e API necessarie.

```
{
    "Version": "2012-10-17",
    "Statement": [
```

```
        {
            "Effect": "Allow",
            "Action": [
              "rds:AddTagsToResource",
              "rds:CreateDBInstance",
              "rds:DescribeDBInstances",
              "rds:DescribeDBClusters",
              "rds:DescribeDBClusterParameters",
              "rds:DeleteDBInstance",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

SageMaker endpoint

La seguente politica basata sull'identità concede le autorizzazioni a tutte SageMaker le azioni API richieste. CloudWatch

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "sagemaker:DescribeEndpoint",
              "sagemaker:DescribeEndpointConfig",
              "sagemaker:DescribeInferenceComponent",
              "sagemaker:UpdateEndpointWeightsAndCapacities",
              "sagemaker:UpdateInferenceComponentRuntimeConfig",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Parco istanze Spot (Amazon EC2)

La seguente politica basata sull'identità concede le autorizzazioni a tutte le azioni Spot Fleet e API richieste. CloudWatch

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "ec2:DescribeSpotFleetRequests",
              "ec2:ModifySpotFleetRequest",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Risorse personalizzate

La seguente policy basata su identità concede autorizzazioni per l'operazione di esecuzione API di API Gateway. Questa politica concede inoltre le autorizzazioni per tutte le azioni richieste. CloudWatch

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "execute-api:Invoke",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
```

```
}
```

Autorizzazioni per lavorare in AWS Management Console

Non esiste una console autonoma per Application Auto Scaling. La maggior parte dei servizi che si integrano con Application Auto Scaling hanno funzionalità dedicate alla configurazione del dimensionamento nella rispettiva console.

Nella maggior parte dei casi, ogni servizio fornisce policy IAM AWS gestite (predefinite) che definiscono l'accesso alla propria console, che include le autorizzazioni per le azioni dell'API Application Auto Scaling. Per ulteriori informazioni, fai riferimento alla documentazione relativa al servizio di cui desideri utilizzare la console.

Puoi inoltre creare policy IAM personalizzate per concedere agli utenti le autorizzazioni granulari al fine di visualizzare e lavorare con operazioni API Application Auto Scaling specifiche nella AWS Management Console. Puoi utilizzare le policy di esempio nelle sezioni precedenti; tuttavia, sono progettate per le richieste effettuate con AWS CLI o con un SDK. La console utilizza operazioni API aggiuntive per le relative caratteristiche. Pertanto, queste policy potrebbero non funzionare come previsto. Ad esempio, per configurare il ridimensionamento dei passaggi, gli utenti potrebbero richiedere autorizzazioni aggiuntive per creare e gestire gli allarmi. CloudWatch

> **Tip**
>
> Per individuare le operazioni API necessarie per eseguire le attività nella console, puoi utilizzare un servizio, ad esempio AWS CloudTrail. Per ulteriori informazioni, consulta la Guida per l'utente AWS CloudTrail.

La seguente policy basata su identità concede autorizzazioni per la configurazione di policy di dimensionamento del Parco istanze Spot. In aggiunta alle autorizzazioni IAM per il Parco istanze Spot l'utente della console che accede alle impostazioni di dimensionamento del parco istanze dalla console Amazon EC2 deve disporre delle autorizzazioni appropriate per i servizi che supportano il dimensionamento dinamico.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
```

```
                "application-autoscaling:*",
                "ec2:DescribeSpotFleetRequests",
                "ec2:ModifySpotFleetRequest",
                "cloudwatch:DeleteAlarms",
                "cloudwatch:DescribeAlarmHistory",
                "cloudwatch:DescribeAlarms",
                "cloudwatch:DescribeAlarmsForMetric",
                "cloudwatch:GetMetricStatistics",
                "cloudwatch:ListMetrics",
                "cloudwatch:PutMetricAlarm",
                "cloudwatch:DisableAlarmActions",
                "cloudwatch:EnableAlarmActions",
                "sns:CreateTopic",
                "sns:Subscribe",
                "sns:Get*",
                "sns:List*"
            ],
            "Resource": "*"
        },
        {
            "Effect": "Allow",
            "Action": "iam:CreateServiceLinkedRole",
            "Resource": "arn:aws:iam::*:role/aws-service-role/ec2.application-
autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest",
            "Condition": {
                "StringLike": {
                    "iam:AWSServiceName":"ec2.application-autoscaling.amazonaws.com"
                }
            }
        }
    ]
}
```

Questa policy consente agli utenti della console di visualizzare e modificare le politiche di scalabilità nella console Amazon EC2 e di creare e CloudWatch gestire allarmi nella console. CloudWatch

È possibile regolare le operazioni API per limitare l'accesso degli utenti. Ad esempio, la sostituzione di `application-autoscaling:*` con `application-autoscaling:Describe*` significa che l'utente ha un accesso di sola lettura.

Puoi anche modificare le CloudWatch autorizzazioni necessarie per limitare l'accesso degli utenti alle funzionalità. CloudWatch Per ulteriori informazioni, consulta la sezione Autorizzazioni necessarie per la CloudWatch console nella Amazon CloudWatch User Guide.

Risoluzione dei problemi di accesso ad Application Auto Scaling

Se riscontri `AccessDeniedException` o problemi simili durante l'utilizzo di Application Auto Scaling, consulta le informazioni in questa sezione.

Non sono autorizzato a eseguire un'operazione in Application Auto Scaling

Se ricevi un messaggio `AccessDeniedException` quando chiami un'operazione AWS API, significa che le credenziali AWS Identity and Access Management (IAM) che stai utilizzando non dispongono delle autorizzazioni necessarie per effettuare quella chiamata.

L'errore di esempio seguente si verifica quando l'utente `mateojackson` tenta di visualizzare i dettagli relativi a un obiettivo scalabile, ma non dispone dell'autorizzazione `application-autoscaling:DescribeScalableTargets`.

```
An error occurred (AccessDeniedException) when calling the DescribeScalableTargets
 operation: User: arn:aws:iam::123456789012:user/mateojackson is not authorized to
 perform: application-autoscaling:DescribeScalableTargets
```

Se ricevi questo errore o altri simili, devi contattare il tuo amministratore per ricevere assistenza.

Un amministratore del tuo account dovrà assicurarsi che tu disponga delle autorizzazioni per accedere a tutte le azioni API utilizzate da Application Auto Scaling per accedere alle risorse nel servizio di destinazione e. CloudWatch Sono necessarie autorizzazioni diverse a seconda delle risorse con cui si sta lavorando. Application Auto Scaling richiede inoltre l'autorizzazione per creare un ruolo collegato ai servizi la prima volta che un utente configura il dimensionamento per una determinata risorsa.

Sono un amministratore e la mia policy IAM ha restituito un errore o non funziona come previsto

Oltre alle azioni Application Auto Scaling, le policy IAM devono concedere le autorizzazioni per chiamare il servizio di destinazione e. CloudWatch Se un utente o un'applicazione non dispone delle autorizzazioni IAM appropriate, il suo accesso potrebbe essere negato in modo imprevisto. Per scrivere policy IAM per utenti e applicazioni nei propri account, consultare le informazioni in Esempi di policy basate su identità di Application Auto Scaling.

Per informazioni su come viene eseguita la convalida, consulta Convalida delle autorizzazioni per le chiamate API Application Auto Scaling sulle risorse di destinazione.

Nota che alcuni problemi di autorizzazione possono anche essere dovuti a un problema con la creazione dei ruoli collegati ai servizi utilizzati da Application Auto Scaling. Per ulteriori informazioni sulla creazione di questi ruoli collegati ai servizi, consulta Ruoli collegati ai servizi per Application Auto Scaling.

Convalida delle autorizzazioni per le chiamate API Application Auto Scaling sulle risorse di destinazione

Per effettuare richieste autorizzate alle azioni dell'API Application Auto Scaling è necessario che il chiamante dell'API disponga delle autorizzazioni per accedere alle AWS risorse nel servizio di destinazione e in. CloudWatch Application Auto Scaling convalida le autorizzazioni per le richieste associate sia al servizio di destinazione che CloudWatch prima di procedere con la richiesta. A tale scopo, viene eseguita una serie di chiamate per convalidare le autorizzazioni IAM sulle risorse obiettivo. Quando viene restituita una risposta, essa viene letta da Application Auto Scaling. Se le autorizzazioni IAM non consentono una determinata operazione, Application Auto Scaling invalida la richiesta e restituisce un errore all'utente contenente informazioni sull'autorizzazione mancante. Ciò garantisce che la configurazione di dimensionamento che l'utente desidera implementare funzioni come previsto e che venga restituito un errore utile se la richiesta non riesce.

Come esempio di come funziona, le seguenti informazioni forniscono dettagli su come Application Auto Scaling esegue le convalide delle autorizzazioni con Aurora e. CloudWatch

Quando un utente chiama l'API `RegisterScalableTarget` su un cluster database Aurora, Application Auto Scaling esegue tutti i seguenti controlli per verificare che l'utente IAM disponga delle autorizzazioni richieste (in grassetto).

- rds:CreateDBInstance: per determinare se l'utente dispone di questa autorizzazione, inviamo una richiesta all'operazione API `CreateDBInstance`, tentando di creare un'istanza database con parametri non validi (ID istanza vuoto) nel cluster database di Aurora specificato dall'utente. Per un utente autorizzato, l'API restituisce una risposta con codice di errore `InvalidParameterValue` dopo avere verificato la richiesta. Tuttavia, per un utente non autorizzato, si ottiene un errore `AccessDenied` e la richiesta di Application Auto Scaling non riesce e invia un errore `ValidationException` all'utente che elenca le autorizzazioni mancanti.
- rds:DeleteDBInstance: inviamo un ID istanza vuoto all'operazione API `DeleteDBInstance`. Per un utente autorizzato, questa richiesta genera un errore `InvalidParameterValue`. Per un utente non autorizzato, si traduce in `AccessDenied` e invia un'eccezione di convalida all'utente (stesso trattamento descritto nel primo punto dell'elenco).

- rds: AddTagsToResource: Poiché l'operazione `AddTagsToResource` API richiede un Amazon Resource Name (ARN), è necessario specificare una risorsa «fittizia» utilizzando un ID account (12345) e un ID di istanza fittizio (non-existing-db) non validi per costruire l'ARN (). `arn:aws:rds:us-east-1:12345:db:non-existing-db` Per un utente autorizzato, questa richiesta genera un errore `InvalidParameterValue`. Per un utente non autorizzato, si traduce in `AccessDenied` e invia un'eccezione di convalida all'utente.
- rds:DescribeDBCluster - Viene descritto il nome del cluster per la risorsa registrata per la scalabilità automatica. Per un utente autorizzato, otteniamo un risultato di descrizione valido. Per un utente non autorizzato, si traduce in `AccessDenied` e invia un'eccezione di convalida all'utente.
- rds:DescribeDBInstance. Chiamiamo l'API `DescribeDBInstance` con un filtro `db-cluster-id` che filtra il nome del cluster fornito dall'utente per registrare l'obiettivo scalabile. Per un utente autorizzato, siamo autorizzati a descrivere tutte le istanze database nel cluster di database. Per un utente non autorizzato, questa chiamata si traduce in `AccessDenied` e invia un'eccezione di convalida all'utente.
- cloudwatch: Alarm: chiamiamo l'API senza alcun parametro. PutMetric `PutMetricAlarm` Poiché il nome dell'allarme è assente, la richiesta risulta in un `ValidationError` per un utente autorizzato. Per un utente non autorizzato, si traduce in `AccessDenied` e invia un'eccezione di convalida all'utente.
- cloudwatch:DescribeAlarms: Chiamiamo l'`DescribeAlarms`API con il valore del numero massimo di record impostato su 1. Per un utente autorizzato, ci aspettiamo informazioni su un allarme nella risposta. Per un utente non autorizzato, questa chiamata si traduce in `AccessDenied` e invia un'eccezione di convalida all'utente.
- cloudwatch:DeleteAlarms: Analogamente a `PutMetricAlarm` quanto sopra, non forniamo parametri da richiedere. `DeleteAlarms` Poiché nella richiesta non è presente il nome dell'allarme, questa chiamata non riesce con `ValidationError` per un utente autorizzato. Per un utente non autorizzato, si traduce in `AccessDenied` e invia un'eccezione di convalida all'utente.

Ogni volta che si verifica una di queste eccezioni di convalida, essa viene registrata. Puoi adottare misure per identificare manualmente quali chiamate non sono riuscite a convalidare utilizzando. AWS CloudTrail Per ulteriori informazioni, consulta la Guida per l'utente AWS CloudTrail.

> **Note**
>
> Se ricevi avvisi per l' CloudTrailutilizzo di eventi Application Auto Scaling, questi avvisi includeranno le chiamate Application Auto Scaling per convalidare le autorizzazioni degli utenti per impostazione predefinita. Per filtrare questi avvisi, utilizza il campo `invokedBy`,

che conterrà `application-autoscaling.amazonaws.com` per questi controlli di convalida.

Accedi all'Application Auto Scaling utilizzando gli endpoint VPC dell'interfaccia

Puoi usarlo AWS PrivateLink per creare una connessione privata tra il tuo VPC e Application Auto Scaling. Puoi accedere ad Application Auto Scaling come se fosse nel tuo VPC, senza l'uso di un gateway Internet, un dispositivo NAT, una connessione VPN o una connessione. AWS Direct Connect Le istanze nel tuo VPC non necessitano di indirizzi IP pubblici per accedere all'Application Auto Scaling.

Stabilisci questa connessione privata creando un endpoint di interfaccia attivato da AWS PrivateLink. In ciascuna sottorete viene creata un'interfaccia di rete endpoint da abilitare per l'endpoint di interfaccia. Si tratta di interfacce di rete gestite dai richiedenti che fungono da punto di ingresso per il traffico destinato all'Application Auto Scaling.

Per ulteriori informazioni, consulta Access Servizi AWS through nella Guida. AWS PrivateLinkAWS PrivateLink

Indice

Creazione di un endpoint VPC dell'interfaccia

Crea un endpoint per Application Auto Scaling utilizzando il seguente nome di servizio:

```
com.amazonaws.region.application-autoscaling
```

Per ulteriori informazioni, consulta Accedere a un AWS servizio utilizzando un endpoint VPC di interfaccia nella Guida.AWS PrivateLink

Non è necessario modificare nessun'altra impostazione. Application Auto Scaling chiama altri AWS servizi utilizzando endpoint di servizio o endpoint VPC con interfaccia privata, a seconda di quale dei due siano in uso.

Creazione di una policy di endpoint VPC

Puoi collegare una policy all'endpoint VPC per controllare l'accesso all'API Application Auto Scaling. La policy specifica:

- Il principale che può eseguire operazioni.
- Le operazioni che possono essere eseguite.
- La risorsa su cui è possibile eseguire le operazioni.

Nell'esempio seguente viene illustrata una policy di endpoint VPC che nega a chiunque l'autorizzazione per eliminare una policy di dimensionamento tramite l'endpoint. Inoltre, la policy di esempio concede a chiunque l'autorizzazione per eseguire tutte le altre operazioni.

```
{
   "Statement": [
        {
            "Action": "*",
            "Effect": "Allow",
            "Resource": "*",
            "Principal": "*"
        },
        {
            "Action": "application-autoscaling:DeleteScalingPolicy",
            "Effect": "Deny",
            "Resource": "*",
            "Principal": "*"
        }
    ]
}
```

Per ulteriori informazioni, consulta Policy dell'endpoint VPC nella Guida di AWS PrivateLink .

Resilienza in Application Auto Scaling

L'infrastruttura AWS globale è costruita attorno a AWS regioni e zone di disponibilità.

AWS Le regioni offrono più zone di disponibilità fisicamente separate e isolate, collegate con reti a bassa latenza, ad alto throughput e altamente ridondanti.

Con le zone di disponibilità, puoi progettare e gestire applicazioni e database che eseguono automaticamente il failover tra zone di disponibilità senza interruzioni. Le zone di disponibilità sono più disponibili, tolleranti ai guasti e scalabili rispetto alle infrastrutture a data center singolo o multiplo tradizionali.

Per ulteriori informazioni su AWS regioni e zone di disponibilità, consulta infrastruttura globale.AWS

Sicurezza dell'infrastruttura in Application Auto Scaling

In quanto servizio gestito, Application Auto Scaling è protetto dalla sicurezza di rete AWS globale. Per informazioni sui servizi di AWS sicurezza e su come AWS protegge l'infrastruttura, consulta AWS Cloud Security. Per progettare il tuo AWS ambiente utilizzando le migliori pratiche per la sicurezza dell'infrastruttura, vedi Infrastructure Protection in Security Pillar AWS Well-Architected Framework.

Utilizzate chiamate API AWS pubblicate per accedere all'Application Auto Scaling attraverso la rete. I client devono supportare quanto segue:

- Transport Layer Security (TLS). È richiesto TLS 1.2 ed è consigliato TLS 1.3.
- Suite di cifratura con Perfect Forward Secrecy (PFS), ad esempio Ephemeral Diffie-Hellman (DHE) o Elliptic Curve Ephemeral Diffie-Hellman (ECDHE). La maggior parte dei sistemi moderni, come Java 7 e versioni successive, supporta tali modalità.

Inoltre, le richieste devono essere firmate utilizzando un ID chiave di accesso e una chiave di accesso segreta associata a un principale IAM. O puoi utilizzare AWS Security Token Service (AWS STS) per generare credenziali di sicurezza temporanee per sottoscrivere le richieste.

Convalida della conformità per Application Auto Scaling

Per sapere se un Servizio AWS programma rientra nell'ambito di specifici programmi di conformità, consulta Servizi AWS la sezione Scope by Compliance Program Servizi AWS e scegli il programma di conformità che ti interessa. Per informazioni generali, consulta Programmi di AWS conformità Programmi di di .

È possibile scaricare report di audit di terze parti utilizzando AWS Artifact. Per ulteriori informazioni, consulta Scaricamento dei report in AWS Artifact .

La vostra responsabilità di conformità durante l'utilizzo Servizi AWS è determinata dalla sensibilità dei dati, dagli obiettivi di conformità dell'azienda e dalle leggi e dai regolamenti applicabili. AWS fornisce le seguenti risorse per contribuire alla conformità:

- Guide introduttive su sicurezza e conformità: queste guide all'implementazione illustrano considerazioni sull'architettura e forniscono i passaggi per l'implementazione di ambienti di base incentrati sulla AWS sicurezza e la conformità.
- Progettazione per la sicurezza e la conformità HIPAA su Amazon Web Services: questo white paper descrive in che modo le aziende possono utilizzare AWS per creare applicazioni idonee all'HIPAA.

> **Note**
>
> Non Servizi AWS tutte sono idonee all'HIPAA. Per ulteriori informazioni, consulta la sezione Riferimenti sui servizi conformi ai requisiti HIPAA.

- AWS Risorse per la per la conformità: questa raccolta di cartelle di lavoro e guide potrebbe essere valida per il tuo settore e la tua località.
- AWS Guide alla conformità dei clienti: comprendi il modello di responsabilità condivisa attraverso la lente della conformità. Le guide riassumono le migliori pratiche per la protezione Servizi AWS e mappano le linee guida per i controlli di sicurezza su più framework (tra cui il National Institute of Standards and Technology (NIST), il Payment Card Industry Security Standards Council (PCI) e l'International Organization for Standardization (ISO)).
- Valutazione delle risorse con regole nella Guida per gli AWS Config sviluppatori: il AWS Config servizio valuta la conformità delle configurazioni delle risorse alle pratiche interne, alle linee guida e alle normative del settore.
- AWS Security Hub— Ciò Servizio AWS fornisce una visione completa dello stato di sicurezza interno. AWS La Centrale di sicurezza utilizza i controlli di sicurezza per valutare le risorse AWS e verificare la conformità agli standard e alle best practice del settore della sicurezza. Per un elenco dei servizi e dei controlli supportati, consulta la pagina Documentazione di riferimento sui controlli della Centrale di sicurezza.
- Amazon GuardDuty: Servizio AWS rileva potenziali minacce ai tuoi carichi di lavoro Account AWS, ai contenitori e ai dati monitorando l'ambiente alla ricerca di attività sospette e dannose. GuardDuty può aiutarti a soddisfare vari requisiti di conformità, come lo standard PCI DSS, soddisfacendo i requisiti di rilevamento delle intrusioni imposti da determinati framework di conformità.

- AWS Audit Manager— Ciò Servizio AWS consente di verificare continuamente l' AWS utilizzo per semplificare la gestione del rischio e la conformità alle normative e agli standard di settore.

Quote per Application Auto Scaling

Hai Account AWS delle quote predefinite, precedentemente denominate limiti, per ciascuna di esse. Servizio AWS Salvo diversa indicazione, ogni quota si applica a una regione specifica. Se per alcune quote è possibile richiedere aumenti, altre quote non possono essere modificate.

Per visualizzare le quote per Application Auto Scaling, apri la Console di Service Quotas. Nel pannello di navigazione, scegli servizi AWS e seleziona Application Auto Scaling.

Per richiedere un aumento delle quote, consultare Richiesta di aumento delle quote nella Guida per l'utente di Service Quotas.

Hai Account AWS le seguenti quote relative all'Application Auto Scaling.

Nome	Predefinita	Adattabile
Obiettivi scalabili per tipo di risorsa	Amazon DynamoDB: 5.000 | Amazon ECS: 3.000 | Amazon Keyspaces: 1.500 | Altri tipi di risorse: 500	Sì
Politiche di scalabilità per target scalabile (sia politiche di scalabilità a fasi che politiche di tracciamento degli obiettivi)	50	No
Operazioni pianificate per obiettivo scalabile	200	No
Regolazioni di fase per policy di dimensionamento a fasi	20	No

Tieni a mente le quote di servizio mentre dimensioni i carichi di lavoro. Ad esempio, quando raggiungi il numero massimo di unità di capacità consentite da un servizio, la scalabilità orizzontale si interromperà. Se la domanda scende e la capacità attuale diminuisce, Application Auto Scaling può di nuovo aumentare orizzontalmente. Per evitare di raggiungere nuovamente questa capacità, puoi richiedere un aumento. Ogni servizio ha le proprie quote predefinite per la capacità massima della risorsa. Per informazioni sulle quote predefinite per altri servizi Amazon Web Services, consulta Endpoint e quote di servizio nella Riferimenti generali di Amazon Web Services.

Cronologia dei documenti per Application Auto Scaling

Nella seguente tabella sono descritte le aggiunte significative apportate alla documentazione di Application Auto Scaling a partire da gennaio 2018. Per ricevere notifiche sugli aggiornamenti della documentazione, puoi sottoscrivere il feed RSS.

Modifica	Descrizione	Data
Modifiche alla Guida	Aggiornato il Numero massimo di obiettivi scalabili per tipo di risorsa inseriscilo nella documentazione delle quote. Consulta Quote per Application Auto Scaling.	16 gennaio 2024
Support per componenti di SageMaker inferenza	Utilizza Application Auto Scaling per dimensionare le copie di un componente di inferenza.	29 novembre 2023
Aggiornamento alle autorizzazioni del ruolo collegato ai servizi di IAM	Il tipo di policy `AWSApplicationAutoscalingSageMakerEndpointPolicy` di Application Auto Scaling. Per ulteriori informazioni, consulta Aggiornamenti di Application Auto Scaling per le policy gestite di AWS.	13 novembre 2023
Support per la concorrenza con provisioning SageMaker senza server	Utilizza Application Auto Scaling per dimensionare il provisioning simultaneo di un endpoint serverless.	9 maggio 2023
Classifica i tuoi obiettivi scalabili utilizzando i tag	Puoi assegnare i metadati agli obiettivi scalabili dell'Applicazione di Dimensionamento	20 marzo 2023

	automatico sotto forma di tag. Consulta Supporto al tagging per l'Applicazione di Dimensionamento automatico.	
Support per la CloudWatch matematica metrica	Ora puoi utilizzare la matematica dei parametri durante la creazione di policy di dimensionamento con monitoraggio degli obiettivi. Con la matematica metrica, puoi interrogare più CloudWatch metriche e utilizzare espressioni matematiche per creare nuove serie temporali basate su queste metriche. Consulta Creazione di una policy di dimensionamento con monitoraggio degli obiettivi per l'Applicazione di Dimensionamento automatico tramite la matematica dei parametri.	14 marzo 2023
Modifiche alla Guida	Il nuovo argomento nella Guida per l'utente di Application Auto Scaling aiuta a iniziare a utilizzare AWS CloudShell con Application Auto Scaling. Vedi Utilizzare AWS CloudShell per lavorare con Application Auto Scaling dalla riga di comando.	17 febbraio 2023

Motivi di non dimensionamento	Grazie all'API Application Auto Scaling, ora puoi recuperare i motivi leggibili dal computer per cui Application Auto Scaling non ridimensiona le tue risorse. Vedi Attività di dimensionamento per Application Auto Scaling.	4 gennaio 2023
Modifiche alla Guida	Aggiornato il Numero massimo di obiettivi scalabili per tipo di risorsa inseriscilo nella documentazione delle quote. Consulta Quote per Application Auto Scaling.	6 maggio 2022
Aggiunto il supporto per i cluster Amazon Neptune	Utilizza Application Auto Scaling per dimensionare il numero di repliche in un cluster DB Amazon Neptune. Per ulteriori informazioni, consulta Amazon Neptune e Application Auto Scaling. L'argomento Aggiornamenti Application Auto Scaling a policy gestite da AWS è stato aggiornato e ora include una nuova policy gestita per l'integrazione con Neptune.	6 ottobre 2021

Application Auto Scaling ora riporta le modifiche alle sue AWS politiche gestite	A partire dal 19 agosto 2021, le modifiche alle policy gestite sono riportate nell'argomento Application Auto Scaling updates to AWS managed policy. La prima modifica elencata è l'aggiunta delle autorizzazioni necessarie ElastiCache per Redis.	19 agosto 2021
Aggiungi il supporto ElastiCache per i gruppi di replica Redis	Utilizzate Application Auto Scaling per scalare il numero di gruppi di nodi e il numero di repliche per gruppo di nodi per un gruppo di replica ElastiCache per Redis (cluster). Per ulteriori informazioni, consulta ElastiCache Redis e Application Auto Scaling.	19 agosto 2021

Modifiche alla Guida	I nuovi argomenti IAM nella Guida per l'utente di Applicati on Auto Scaling ti aiutano a risolvere i problemi di accesso ad Application Auto Scaling. Per ulteriori informazioni, consulta Identity and Access Management per Applicati on Auto Scaling. Sono stati inoltre aggiunti nuovi esempi di politiche di autorizzazione IAM per le azioni sui servizi di destinazione e Amazon CloudWatch. Per ulteriori informazioni, consulta Politiche di esempio per lavorare con AWS CLI o un SDK.	23 febbraio 2021
Aggiunto il supporto per i fusi orari locali	Ora puoi creare operazion i pianificate nel fuso orario locale. Se il fuso orario osserva l'ora legale, esso si adegua automaticamente in funzione dell'ora legale. Per ulteriori informazioni, consulta Dimensionamento pianificato.	2 febbraio 2021

Modifiche alla Guida	Un nuovo tutorial nella Guida per l'utente di Application Auto Scaling consente di comprendere come utilizzare le policy di dimensionamento con monitoraggio degli obiettivi e il dimensionamento pianificato per aumentare la disponibilità dell'applicazione quando si utilizza la Application Auto Scaling. Inoltre, un nuovo argomento spiega come attivare una notifica quando vengono CloudWatch rilevati problemi che potrebbero richiedere la tua attenzione.	15 ottobre 2020
Aggiunto il supporto per l'archiviazione cluster Amazon Managed Streaming for Apache Kafka	Utilizza una policy di dimensionamento con monitoraggio degli obiettivi per il dimensionamento orizzontale dell'archiviazione broker associata a un cluster Amazon MSK.	30 settembre 2020
Aggiunto il supporto per gli endpoint di riconoscimento delle entità di Amazon Comprehend	Utilizza Application Auto Scaling per dimensionare il numero di unità di inferenza fornite per gli endpoint di riconoscimento delle entità di Amazon Comprehend.	28 settembre 2020

Aggiunta del supporto per le tabelle di Amazon Keyspaces (per Apache Cassandra)	Utilizza Application Auto Scaling per dimensionare la velocità effettiva assegnata (capacità di lettura e scrittura) di una tabella di Amazon Keyspaces.	23 aprile 2020
Nuovo capitolo sulla sicurezza	Un nuovo capitolo sulla sicurezza nella Guida per l'utente di Application Auto Scaling consente di comprendere come applicare il modello di responsabilità condivisa quando si utilizza Application Auto Scaling. Come parte di questo aggiornamento, il capitolo della guida per l'utente "Autenticazione e controllo degli accessi" è stato sostituito da una nuova sezione più utile, Identity and Access Management per Application Auto Scaling.	16 gennaio 2020
Aggiornamenti minori	Vari miglioramenti e correzioni.	15 gennaio 2020
Aggiunta della funzionalità di notifica	Application Auto Scaling ora invia eventi ad Amazon EventBridge e notifiche al tuo utente AWS Health Dashboard quando si verificano determinate azioni. Per ulteriori informazioni, consulta Monitoraggio di Application Auto Scaling.	20 dicembre 2019

Aggiungi il supporto per le funzioni AWS Lambda	Utilizza Application Auto Scaling per dimensionare il provisioning simultaneo di una funzione Lambda.	3 dicembre 2019
Aggiunto il supporto per gli endpoint di classificazione dei documenti di Amazon Comprehend	Utilizza Application Auto Scaling per dimensionare la capacità di velocità effettiva di un endpoint di classificazione dei documenti di Amazon Comprehend.	25 novembre 2019
Aggiungi il supporto AppStream 2.0 per le politiche di scalabilità del tracciamento degli obiettivi	Utilizza le politiche di scalabilità di Target Tracking per scalare le dimensioni di una flotta AppStream 2.0.	25 novembre 2019
Supporto per gli endpoint Amazon VPC	Ora puoi stabilire una connessione privata tra il VPC e Application Auto Scaling. Per informazioni e istruzioni sulla migrazione, consulta Application Auto Scaling ed endpoint VPC di interfaccia.	22 novembre 2019
Sospensione e ripresa del dimensionamento	Aggiunto il supporto per la sospensione e la ripresa del dimensionamento. Per ulteriori informazioni, consulta Sospensione e ripresa del dimensionamento per Application Auto Scaling.	29 agosto 2019

Nuova sezione	La sezione Impostazione è stata aggiunta alla documentazione di Application Auto Scaling. Miglioramenti e correzioni meno importanti sono stati apportati in tutta la guida per l'utente.	28 giugno 2019
Modifiche alla Guida	Documentazione di Application Auto Scaling migliorata nelle sezioni Dimensionamento pianificato, Policy di dimensionamento per fasi e Policy di dimensionamento con monitoraggio degli obiettivi.	11 marzo 2019
Aggiunta del supporto per le risorse personalizzate	Utilizza Application Auto Scaling per dimensionare le risorse personalizzate fornite dalle tue applicazioni o dai servizi. Per ulteriori informazioni, consulta il nostro GitHubrepository.	9 luglio 2018
Aggiungi il supporto per le varianti SageMaker degli endpoint	Utilizza Application Auto Scaling per dimensionare il numero di istanze dell'endpoint fornite per una variante.	28 febbraio 2018

Nella seguente tabella sono descritte le modifiche significative apportate alla documentazione di Application Auto Scaling prima del gennaio 2018.

Modifica	Descrizione	Data
Aggiunta di supporto per le repliche Aurora	Utilizza Application Auto Scaling per dimensionare	17 Novembre 2017

Modifica	Descrizione	Data
	il numero desiderato. Per ulteriori informazioni, consulta Utilizzo della scalabilità automatica di Amazon Aurora con le repliche Aurora nella Guida per l'utente di Amazon RDS.	
Aggiunta del supporto per il dimensionamento pianificato	Utilizza il dimensionamento pianificato per dimension are le risorse a orari specifici o intervalli predefiniti. Per ulteriori informazioni, consulta Dimensionamento pianificato per Application Auto Scaling.	8 Novembre 2017
Aggiunta del supporto per le policy di dimensionamento del monitoraggio di target	Utilizza le policy di dimension amento del monitoraggio dei target per impostare il dimensionamento dinamico per la tua applicazione in poche e semplici fasi. Per ulteriori informazioni, consulta Policy di dimensionamento con monitoraggio degli obiettivi per Application Auto Scaling.	12 luglio 2017

Modifica	Descrizione	Data
Aggiunto il supporto per la capacità di lettura e scrittura assegnata per le tabelle e gli indici secondari globali DynamoDB	Utilizza Application Auto Scaling per dimensionare la velocità effettiva assegnata (capacità di lettura e scrittura). Per ulteriori informazioni, consulta Gestione della capacità di velocità effettiva con la scalabilità automatica di DynamoDB nella Guida per gli sviluppatori di Amazon DynamoDB.	14 giugno 2017
Aggiungi il supporto per le flotte AppStream 2.0	Utilizza Application Auto Scaling per dimensionare la dimensione del parco istanze. Per ulteriori informazioni, consulta Fleet Auto Scaling for AppStream 2.0 nella Amazon AppStream 2.0 Administration Guide.	23 marzo 2017
Aggiunto il supporto per i cluster Amazon EMR	Utilizza Application Auto Scaling per dimensionare i nodi principali e attività. Per ulteriori informazioni, consulta Utilizzo della scalabilità automatica in Amazon EMR nella Guida alla gestione di Amazon EMR.	18 novembre 2016

Modifica	Descrizione	Data
Aggiunto il supporto per i parchi istanze Spot	Utilizza Application Auto Scaling per dimensionare la capacità obiettivo. Per ulteriori informazioni, consulta la sezione Scalabilità automatica per la flotta Spot nella Guida per l'utente di Amazon EC2.	1 settembre 2016
Aggiunto il supporto per i servizi Amazon ECS	Utilizza Application Auto Scaling per dimensionare il numero desiderato. Per ulteriori informazioni, consulta Scalabilità automatica del servizio nella Guida per gli sviluppatori di Amazon Elastic Container Service.	9 agosto 2016

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.