Benutzerhandbuch für Skalierungspläne

# AWS Auto Scaling

# AWS Auto Scaling: Benutzerhandbuch für Skalierungspläne

# Table of Contents

# Was ist ein Skalierungsplan?

Verwenden Sie einen Skalierplan, um Auto Scaling für verwandte oder zugehörige skalierbare Ressourcen innerhalb weniger Minuten zu konfigurieren. Sie können beispielsweise Tags verwenden, um Ressourcen in Kategorien wie Produktion, Test oder Entwicklung zu gruppieren. Anschließend können Sie nach Skalierungsplänen für skalierbare Ressourcen suchen und diese einrichten, die zu jeder Kategorie gehören. Oder, falls Ihre Cloud-Infrastruktur dies umfasst AWS CloudFormation, können Sie Stack-Vorlagen definieren, die zum Erstellen von Ressourcensammlungen verwendet werden. Erstellen Sie dann einen Skalierungsplan für die skalierbaren Ressourcen, die zu jedem Stack gehören.

## Unterstützte Ressourcen

AWS Auto Scaling unterstützt die Verwendung von Skalierungsplänen für die folgenden Dienste und Ressourcen:

- Amazon Aurora – Erhöhen oder verringern Sie die Anzahl der Aurora-Lesereplikate, die für einen Aurora-DB-Cluster bereitgestellt werden.
- Amazon EC2 Auto Scaling – Starten oder beenden Sie EC2-Instances, indem Sie die gewünschte Kapazität einer Auto-Scaling-Gruppe erhöhen oder senken.
- Amazon Elastic Container Service – Die gewünschte Aufgabenzahl in Amazon ECS erhöhen oder senken.
- Amazon DynamoDB – Erhöhen oder verringern Sie die bereitgestellte Lese- und Schreibkapazität einer DynamoDB-Tabelle oder eines globalen sekundären Index.
- Spot-Flotte – Starten oder beenden Sie EC2-Instances, indem Sie die Zielkapazität einer Spot-Flotte erhöhen oder verringern.

## Funktionen und Vorteile des Skalierungsplans

Skalierungspläne bieten die folgenden Funktionen und Vorteile:

- Ressourcenerkennung — AWS Auto Scaling ermöglicht die automatische Ressourcenerkennung, um skalierbare Ressourcen in Ihrer Anwendung zu finden.
- Dynamische Skalierung – Skalierungspläne verwenden die Amazon-EC2-Auto-Scaling- und Application-Auto-Scaling-Services, um die Kapazität skalierbarer Ressourcen anzupassen,

um Änderungen des Datenverkehrs oder der Workload zu bewältigen. Dynamische Skalierungsmetriken können Standardauslastung oder Durchsatzmetriken oder benutzerdefinierte Metriken sein.

- Integrierte Skalierempfehlungen – AWS Auto Scaling bietet Skalierungsstrategien mit Empfehlungen, die Sie verwenden können, um Leistung, Kosten oder ein Gleichgewicht zwischen den beiden zu optimieren.
- Prädiktive Skalierung – Skalierungspläne unterstützen auch die Prognose-Skalierung für Auto-Scaling-Gruppen. Dies hilft, Ihre Amazon-EC2-Kapazität schneller zu skalieren, wenn regelmäßig Spitzen auftreten.

> **Important**
>
> Wenn Sie Skalierungspläne nur für vorausschauende Skalierung verwenden, empfehlen wir dringend, stattdessen Richtlinien für vorausschauende Skalierung direkt auf Ihren Auto Scaling Scaling-Ressourcen festzulegen. Diese Option bietet mehr Funktionen, z. B. die Verwendung von Metrikaggregationen, um neue benutzerdefinierte Metriken zu erstellen oder historische Metrikdaten für mehrere Bereitstellungen beizubehalten. blue/green Weitere Informationen zu Amazon EC2 Auto Scaling finden Sie unter Predictive Scaling for Amazon EC2 Auto Scaling im Amazon EC2 Auto Scaling Scaling-Benutzerhandbuch. Weitere Informationen zu Application Auto Scaling finden Sie unter Predictive Scaling for Application Auto Scaling im Application Auto Scaling Scaling-Benutzerhandbuch.
> Eine Anleitung zur Migration von Skalierungsplänen zu vorausschauenden Skalierungsrichtlinien von Amazon EC2 Auto Scaling finden Sie unter. Migrieren Sie Ihren Skalierungsplan

# Erste Schritte

Verwenden Sie die folgenden Ressourcen, um einen Skalierungsplan zu erstellen und zu verwenden:

- Funktionsweise von Skalierungsplänen
- Bewährte Methoden für Skalierungspläne
- Erste Schritte mit Skalierungsplänen

# Arbeiten Sie mit Skalierungsplänen

Sie können die folgenden Schnittstellen verwenden, um Ihre Skalierungspläne zu erstellen, auf sie zuzugreifen und sie zu verwalten:

- AWS-Managementkonsole – Bietet eine Webschnittstelle für den Zugriff auf Ihre Skalierungspläne. Wenn Sie sich für einen registriert haben AWS-Konto, können Sie auf Ihre Skalierungspläne zugreifen, indem Sie sich bei dem anmelden AWS-Managementkonsole, das Suchfeld in der Navigationsleiste verwenden, um danach zu suchen AWS Auto Scaling, und dann auswählen. AWS Auto Scaling
- AWS Command Line Interface (AWS CLI) — Stellt Befehle für eine Vielzahl von AWS-Services Befehlen bereit und wird unter Windows, MacOS und Linux unterstützt. Informationen zu den ersten Schritten finden Sie im AWS Command Line Interface -Benutzerhandbuch. Weitere Informationen finden Sie unter autoscaling-plans in der AWS CLI -Befehlsreferenz.
- AWS Tools for Windows PowerShell— Stellt Befehle für eine breite Palette von AWS Produkten für Benutzer bereit, die in der PowerShell Umgebung Skripts erstellen. Informationen zu den ersten Schritten finden Sie im AWS -Tools für PowerShell -Benutzerhandbuch. Weitere Informationen finden Sie in der AWS -Tools für PowerShell Cmdlet-Referenz.
- AWS SDKs— Stellt sprachspezifische API-Operationen bereit und kümmert sich um viele Verbindungsdetails, wie z. B. die Berechnung von Signaturen, die Behandlung von Wiederholungsversuchen von Anfragen und die Behandlung von Fehlern. Weitere Informationen finden Sie unter AWS SDKs.
- HTTPS-API – Bietet API-Aktionen auf niedriger Ebene, die Sie mithilfe von HTTPS-Anforderungen aufrufen. Weitere Informationen finden Sie in der AWS Auto Scaling -API-Referenz.
- CloudFormation— Unterstützt die Erstellung von Skalierungsplänen mithilfe von Vorlagen. CloudFormation Weitere Informationen finden Sie in der AWS::AutoScalingPlans::ScalingPlanReferenz im CloudFormation Benutzerhandbuch.

# Regionale Verfügbarkeit

Die AWS Auto Scaling API ist in mehreren Versionen verfügbar AWS-Regionen und bietet einen Endpunkt für jede dieser Regionen. Eine Liste aller Regionen und Endpunkte, in denen die API derzeit verfügbar ist, finden Sie unter AWS Auto Scaling Endpunkte und Kontingente in den Allgemeine AWS-Referenz

# Preisgestaltung

Alle Funktionen für Skalierungspläne sind für Ihre Verwendung aktiviert. Die Funktionen werden ohne zusätzliche Kosten bereitgestellt, die über die Servicegebühren für CloudWatch und die anderen AWS Cloud Ressourcen, die Sie verwenden, hinausgehen.

> **Note**
>
> Die Funktion zur vorausschauenden Skalierung basiert auf der CloudWatch GetMetricDataErfassung historischer Metrikdaten für Kapazitätsprognosen, was mit Kosten verbunden ist. Wenn Sie jedoch Predictive Scaling mit einer Amazon EC2 Auto Scaling-Skalierungsrichtlinie anstelle eines Skalierungsplans aktivieren, fallen keine Gebühren für Anrufe an an. `GetMetricData`

# Funktionsweise von Skalierungsplänen

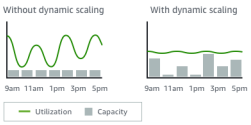
AWS Auto Scaling ermöglicht es Ihnen, Skalierungspläne zu verwenden, um eine Reihe von Anweisungen für die Skalierung Ihrer Ressourcen zu konfigurieren. Wenn Sie mit skalierbaren Ressourcen arbeiten CloudFormation oder Tags hinzufügen, können Sie Skalierungspläne für verschiedene Ressourcengruppen pro Anwendung einrichten. Die AWS Auto Scaling Konsole bietet Empfehlungen für Skalierungsstrategien, die auf jede Ressource zugeschnitten sind. Nachdem Sie Ihren Skalierungsplan erstellt haben, kombiniert es dynamische Skalierungsmethoden und prädiktive Skalierungsmethoden, um Ihre Skalierungsstrategie zu unterstützen.

Was ist eine Skalierungsstrategie?

In der Skalierungsstrategie erfahren Sie, AWS Auto Scaling wie Sie die Nutzung der Ressourcen in Ihrem Skalierungsplan optimieren können. Sie können die Optimierung auf Verfügbarkeit, auf Kosten oder ein ausgewogenes Verhältnis aus beidem festlegen. Alternativ können Sie auch Ihre eigene benutzerdefinierte Strategie mit Ihren definierten Metriken und Schwellenwerten erstellen. Sie können separate Strategien für die einzelnen Ressourcen oder Ressourcentypen festlegen.


Was ist dynamische Skalierung?

Mit der dynamischen Skalierung werden für die Ressourcen in Ihrem Skalierungsplan Skalierungsrichtlinien für die Zielnachverfolgung erstellt. Mithilfe dieser Skalierungsrichtlinien wird die Ressourcenkapazität als Reaktion auf Live-Änderungen in der Ressourcennutzung angepasst. Auf diese Weise soll ausreichend Kapazität bereitgestellt werden, um die Auslastung auf dem von der Skalierungsstrategie festgelegten Zielwert zu halten. Dies ähnelt der Art und Weise, wie ein Thermostat die Temperatur in Ihrem Zuhause konstant hält. Sie können eine Temperatur auswählen und der Thermostat erledigt den Rest.

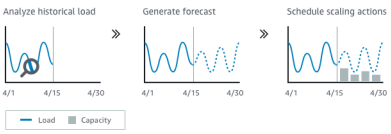



Sie können beispielsweise Ihren Skalierungsplan so konfigurieren, dass die Anzahl der Aufgaben, die Ihr Amazon Elastic Container Service (Amazon ECS)-Service ausführt, bei 75 Prozent der CPU-Auslastung bleibt. Wenn die CPU-Auslastung Ihres Services 75 Prozent überschreitet (was bedeutet, dass mehr als 75 Prozent der für den Service reservierten CPU verwendet wird), fügt Ihre Skalierungsrichtlinie Ihrem Service eine weitere Aufgabe hinzu, um bei der erhöhten Last zu helfen.

## Was ist prädiktive Skalierung?

Die prädiktive Skalierung verwendet Machine Learning, um den historischen Workload jeder Ressource zu analysieren und prognostiziert regelmäßig die zukünftige Belastung. Dies ähnelt der Funktionsweise von Wettervorhersagen. Mit der Prognose erstellt die prädiktive Skalierung geplante Skalierungsaktionen, um sicherzustellen, dass die Ressourcenkapazität verfügbar ist, bevor Ihre Anwendung sie benötigt. Wie bei der dynamischen Skalierung wird bei der prädiktiven Skalierung die Auslastung auf dem von der Skalierungsstrategie vorgegebenen Zielwert gehalten.

> **Important**
>
> Wenn Sie Skalierungspläne nur für vorausschauende Skalierung verwenden, empfehlen wir dringend, stattdessen Richtlinien für vorausschauende Skalierung direkt auf Ihren Auto Scaling Scaling-Ressourcen festzulegen. Diese Option bietet mehr Funktionen, z. B. die Verwendung von Metrikaggregationen, um neue benutzerdefinierte Metriken zu erstellen oder historische Metrikdaten für mehrere Bereitstellungen beizubehalten. blue/green Weitere Informationen zu Amazon EC2 Auto Scaling finden Sie unter Predictive Scaling for Amazon EC2 Auto Scaling im Amazon EC2 Auto Scaling Scaling-Benutzerhandbuch. Weitere Informationen zu Application Auto Scaling finden Sie unter Predictive Scaling for Application Auto Scaling im Application Auto Scaling Scaling-Benutzerhandbuch.
> Eine Anleitung zur Migration von Skalierungsplänen zu vorausschauenden Skalierungsrichtlinien von Amazon EC2 Auto Scaling finden Sie unter. Migrieren Sie Ihren Skalierungsplan




Sie können beispielsweise die prädiktive Skalierung aktivieren und Ihre Skalierungsstrategie so konfigurieren, dass die durchschnittliche CPU-Auslastung Ihrer Auto-Scaling-Gruppe bei 50 Prozent gehalten wird. Ihre Prognose sieht vor, dass täglich um 8:00 Uhr Verkehrsspitzen auftreten. Ihr Skalierungsplan erstellt die zukünftigen geplanten Skalierungsaktionen, um sicherzustellen, dass Ihre Auto-Scaling-Gruppe rechtzeitig bereit ist, den Datenverkehr zu bewältigen. Auf diese Weise kann die Anwendungsleistung konstant gehalten werden, mit dem Ziel, stets die erforderliche Kapazität bereitzustellen, um die Ressourcenauslastung jederzeit so nahe wie möglich bei 50 Prozent zu halten.

Die wichtigsten Komponenten zum Verständnis für die prädikative Skalierung sind folgende:

- Lastprognose: AWS Auto Scaling Analysiert den Verlauf von bis zu 14 Tagen für eine bestimmte Lastkennzahl und prognostiziert den future Bedarf für die nächsten zwei Tage. Diese Daten werden in 1-Stunden-Intervallen zur Verfügung gestellt und täglich aktualisiert.
- Geplante Skalierungsaktionen: AWS Auto Scaling plant die Skalierungsaktionen, mit denen die Kapazität proaktiv erhöht und verringert wird, um der Lastprognose zu entsprechen. AWS Auto Scaling Aktualisiert die Mindestkapazität zum geplanten Zeitpunkt mit dem Wert, der durch die geplante Skalierungsaktion angegeben wurde. Auf diese Weise soll die Ressourcenauslastung auf dem von der Skalierungsstrategie festgelegten Zielwert gehalten werden. Wenn Ihre Anwendung mehr Kapazität benötigt als vorhergesagt, lassen sich mit der dynamischen Skalierung zusätzliche Kapazitäten hinzufügen.
- Verhalten bei max. Kapazität: Die minimalen und maximalen Kapazitätsgrenzen für die automatische Skalierung gelten für jede Ressource. Sie können jedoch steuern, ob Ihre Anwendung die Kapazität über die maximale Kapazität hinaus erhöhen kann, wenn die prognostizierte Kapazität die maximale Kapazität überschreitet.

# Bewährte Methoden für Skalierungspläne

Die folgenden bewährten Methoden können Sie dabei unterstützen, Skalierungspläne optimal zu nutzen:

- Wenn Sie eine Startvorlage oder eine Startkonfiguration erstellen, aktivieren Sie die detaillierte Überwachung, um CloudWatch Metrikdaten für EC2-Instances im Abstand von einer Minute abzurufen, da so eine schnellere Reaktion auf Laständerungen gewährleistet ist. Erfolgt die Skalierung nach Metriken mit einem Intervall von 5 Minuten, kann dies zu einer verringerten Reaktionszeit und zu einer Skalierung nach veralteten Metrikdaten führen. Standardmäßig sind EC2-Instances für die grundlegende Überwachung aktiviert. Dies bedeutet, dass Metrikdaten für Instances mit einem Intervall von 5 Minuten verfügbar sind. Gegen eine zusätzliche Gebühr können Sie die detaillierte Überwachung aktivieren, um Metrikdaten für Instances mit einer Minute Frequenz zu erhalten. Weitere Informationen finden Sie unter Konfigurieren der Überwachung für Auto Scaling-Instances im Amazon EC2 Auto Scaling-Benutzerhandbuch.
- Wir empfehlen außerdem, die Auto-Scaling-Gruppen-Metriken zu aktivieren. Andernfalls wird die tatsächliche Kapazität nicht in den Kapazitätsprognose-Diagrammen angezeigt, die nach Abschluss des Assistenten für das Erstellen eines Skalierungsplans verfügbar sind. Weitere Informationen finden Sie unter CloudWatch Monitoring-Metriken für Ihre Auto Scaling Scaling-Gruppen und -Instances im Amazon EC2 Auto Scaling Scaling-Benutzerhandbuch.
- Prüfen Sie, welchen Instance-Typ Ihre Auto-Scaling-Gruppe verwendet, und achten Sie darauf, einen Typ der Instance mit Spitzenlastleistung zu verwenden. Amazon-EC2-Instances mit Spitzenlastleistung, wie T3- und T2-Instances, wurden entwickelt, um eine CPU-Basisleistung mit der Fähigkeit, die Leistung je nach Erfordernis Ihres Workloads zu steigern, bereitzustellen. Je nach der vom Skalierungsplan vorgegebenen Zielauslastung besteht das Risiko, dass Sie die Basisleistung überschreiten und kein CPU-Guthaben mehr haben, was die Leistung einschränkt. Weitere Informationen finden sie unter CPU-Guthaben und Basisleistung für Instances mit Spitzenlastleistung. Informationen zur Konfiguration dieser Instances finden Sie unter Using a Auto Scaling group to launch a Burstable Performance Instance as Unlimited im Amazon EC2 EC2-Benutzerhandbuch. `unlimited`

# Weitere Überlegungen

> **Important**
>
> Wenn Sie Skalierungspläne nur für vorausschauende Skalierung verwenden, empfehlen wir dringend, stattdessen Richtlinien für vorausschauende Skalierung direkt auf Ihren Auto Scaling Scaling-Ressourcen festzulegen. Diese Option bietet mehr Funktionen, z. B. die Verwendung von Metrikaggregationen, um neue benutzerdefinierte Metriken zu erstellen oder historische Metrikdaten für mehrere Bereitstellungen beizubehalten. blue/green Weitere Informationen zu Amazon EC2 Auto Scaling finden Sie unter Predictive Scaling for Amazon EC2 Auto Scaling im Amazon EC2 Auto Scaling Scaling-Benutzerhandbuch. Weitere Informationen zu Application Auto Scaling finden Sie unter Predictive Scaling for Application Auto Scaling im Application Auto Scaling Scaling-Benutzerhandbuch.
> Eine Anleitung zur Migration von Skalierungsplänen zu vorausschauenden Skalierungsrichtlinien von Amazon EC2 Auto Scaling finden Sie unter. Migrieren Sie Ihren Skalierungsplan

Berücksichtigen Sie die folgenden zusätzlichen Aspekte:

- Die prädiktive Skalierung nutzt Lastenprognosen, um die zukünftige Kapazität zu planen. Die Qualität der Prognosen variiert je nach Zyklizität der Last und der Anwendbarkeit des geschulten Prognosemodells. Die prädiktive Skalierung kann im Nur-Prognose-Modus ausgeführt werden, um die Qualität der Prognosen und der von den Prognosen erstellten Skalierungsaktionen zu beurteilen. Während dem Erstellen des Skalierungsplans können Sie den prädiktiven Skalierungsmodus als Forecast only (Nur Prognose) festlegen und ihn anschließend, nachdem Sie die Qualität der Prognose beurteilt haben, zu Forecast and scale (Prognose und Skalierung) ändern. Weitere Informationen erhalten Sie unter Prädiktive Skalierungseinstellungen und Überwachen und Auswerten von Prognosen.
- Wenn Sie andere Metriken für die prädiktive Skalierung festlegen möchten, müssen Sie sicherstellen, dass die Skalierungsmetrik und die Lastmetrik stark korrelieren. Der Wert der Metrik muss sich proportional zur Anzahl der Instances in der Auto-Scaling-Gruppe erhöhen oder verringern. Dadurch wird sichergestellt, dass die Metrikdaten verwendet werden können, um die Anzahl der Instances proportional zu skalieren. Beispiel: Die Lastmetrik entspricht der Gesamtanzahl der Anfragen und die Skalierungsmetrik der durchschnittlichen CPU-Auslastung. Wenn sich die Gesamtzahl an Anfragen um 50 Prozent erhöht, sollte sich die durchschnittliche

CPU-Auslastung ebenfalls um 50 Prozent erhöhen, vorausgesetzt, dass die Kapazität unverändert bleibt.

- Bevor Sie Ihren Skalierungsplan erstellen, sollten Sie alle zuvor geplanten Skalierungsaktionen löschen, die Sie nicht mehr benötigen, indem Sie auf die Konsolen zugreifen, auf denen sie erstellt wurden. AWS Auto Scaling erstellt keine vorausschauende Skalierungsaktion, die sich mit einer vorhandenen geplanten Skalierungsaktion überschneidet.
- Ihre benutzerdefinierten Einstellungen für die minimale und maximale Kapazität werden zusammen mit anderen Einstellungen, die für die dynamische Skalierung genutzt werden, in anderen Konsolen angezeigt. Wir empfehlen jedoch, dass Sie, nachdem Sie einen Skalierungsplan erstellt haben, diese Einstellungen nicht über andere Konsolen ändern, da Ihr Skalierungsplan die Änderungen von anderen Konsolen nicht erhält.
- Ihr Skalierungsplan kann Ressourcen aus mehreren Services enthalten, jede Ressource kann sich aber nur in jeweils einem Skalierungsplan befinden.

# Den Fehler vermeiden ActiveWithProblems

Ein Fehler ActiveWithProblems "" kann auftreten, wenn ein Skalierungsplan erstellt oder Ressourcen zu einem Skalierungsplan hinzugefügt werden. Der Fehler tritt auf, wenn der Skalierungsplan aktiv ist, aber die Skalierungskonfiguration für mindestens eine Ressource nicht angewendet werden konnte.

Dies geschieht normalerweise, weil eine Ressource bereits über eine Skalierungsrichtlinie verfügt oder eine Auto-Scaling-Gruppe die Mindestanforderungen für die prädiktive Skalierung nicht erfüllt.

Wenn eine Ihrer Ressourcen bereits über Skalierungsrichtlinien von verschiedenen Servicekonsolen verfügt, überschreibt AWS Auto Scaling diese anderen Skalierungsrichtlinien nicht und erstellt auch nicht standardmäßig neue Richtlinien. Sie können optional die vorhandenen Skalierungsrichtlinien löschen und sie durch Skalierungsrichtlinien für die Zielverfolgung ersetzen, die über die AWS Auto Scaling Konsole erstellt wurden. Dazu aktivieren Sie die Einstellung Replace external scaling policies (Externe Skalierungsrichtlinien ersetzen) für jede Ressource, für die Skalierungsrichtlinien überschrieben werden sollen.

Bei der prädiktiven Skalierung empfehlen wir, nach dem Erstellen einer neuen Auto-Scaling-Gruppe mit dem Konfigurieren der prädiktiven Skalierung 24 Stunden zu warten. Zum Generieren der anfänglichen Prognose müssen mindestens 24 Stunden an historischen Daten vorhanden sein. Wenn für die Gruppe weniger als 24 Stunden an historischen Daten vorhanden sind und die prädiktive Skalierung aktiviert ist, hat dies zur Folge, dass der Skalierungsplan erst in der nächsten Prognoseperiode, nachdem die erforderliche Datenmenge für die Gruppe erfasst wurde, eine

Prognose generieren kann. Sie können den Skalierungsplan jedoch auch bearbeiten und speichern, um den Prognoseprozess neu zu starten, sobald die 24 Stunden an Daten verfügbar sind.

# Erste Schritte mit Skalierungsplänen

Bevor Sie einen Skalierungsplan für die Verwendung mit Ihrer Anwendung erstellen, prüfen Sie die Anwendung gründlich, während sie in der AWS Cloud ausgeführt wird. Beachten Sie die folgenden Punkte:

- Ob über andere Konsolen erstellte Skalierungsrichtlinien vorhanden sind. Sie können die vorhandenen Skalierungsrichtlinien ersetzen oder beibehalten (ohne Änderungen an ihren Werten vornehmen zu können), wenn Sie Ihren Skalierungsplan erstellen.
- Die Zielauslastung, die für jede skalierbare Ressource in Ihrer Anwendung basierend auf der gesamten Ressource sinnvoll ist. Beispielsweise die CPU-Größe, die die EC2-Instances in einer Auto-Scaling-Gruppe voraussichtlich benötigen werden, im Vergleich zum für sie verfügbaren CPU. Oder für einen Service wie DynamoDB, der ein Modell mit bereitgestelltem Durchsatz verwendet, die Menge an Lese- und Schreibaktivitäten, die eine Tabelle oder ein Index voraussichtlich benötigen wird, im Vergleich zum verfügbaren Durchsatz. Anders gesagt: das Verhältnis von genutzter zu bereitgestellter Kapazität. Sie können die Zielauslastung jederzeit ändern, nachdem Sie Ihren Skalierungsplan erstellt haben.
- Wie lange dauert es, einen Server zu starten und zu konfigurieren? Wenn Sie dies wissen, können Sie für jede EC2-Instance ein Fenster zum Aufwärmen nach dem Starten konfigurieren, um sicherzustellen, dass kein neuer Server gestartet wird, während der vorherige Server noch gestartet wird.
- Ob der Metrikverlauf ausreichend lang ist, um mit der prädiktiven Skalierung verwendet zu werden (wenn neu erstellte Auto-Scaling-Gruppen genutzt werden). Im Allgemeinen sind Verlaufsdaten von 14 Tagen für genauere Vorhersagen erforderlich. Der Mindestwert beträgt 24 Stunden.

Je besser Sie Ihre Anwendung verstehen, desto effektiver können Sie den Skalierungsplan gestalten.

Die folgenden Aufgaben helfen Ihnen, sich mit Skalierungsplänen vertraut zu machen. Sie erstellen einen Skalierungsplan für eine einzelne Auto-Scaling-Gruppe und aktivieren die prädiktive Skalierung und die dynamische Skalierung.

Aufgaben

- Schritt 1: Ermitteln Ihrer skalierbaren Ressourcen
- Schritt 2: Festlegen der Skalierungsstrategie
- Schritt 3: Konfigurieren erweiterter Einstellungen (optional)

# Schritt 1: Ermitteln Ihrer skalierbaren Ressourcen

In diesem Abschnitt erhalten Sie eine praktische Einführung in die Erstellung von Skalierungsplänen in der AWS Auto Scaling -Konsole. Wenn dies Ihr erster Skalierungsplan ist, empfehlen wir Ihnen, zunächst einen Beispielskalierungsplan mit einer Amazon-EC2-Auto-Scaling-Gruppe zu erstellen.

## Voraussetzungen

Erstellen Sie zur Verwendung eines Skalierungsplans eine Auto-Scaling-Gruppe. Starten Sie mindestens eine Amazon-EC2-Instance in der Auto-Scaling-Gruppe. Weitere Informationen finden Sie unter Erste Schritte mit Amazon EC2 Auto Scaling im Benutzerhandbuch für Amazon EC2 Auto Scaling.

Verwenden Sie eine Auto Scaling Scaling-Gruppe mit aktivierten CloudWatch Metriken, damit Kapazitätsdaten in den Diagrammen angezeigt werden, die verfügbar sind, wenn Sie den Assistenten zum Erstellen von Skalierungsplänen ausführen. Weitere Informationen finden Sie unter Überwachen von CloudWatch Metriken für Ihre Auto Scaling Scaling-Gruppen und -Instances im Amazon EC2 Auto Scaling Scaling-Benutzerhandbuch.

Generieren Sie eine gewisse Auslastung für einige Tage oder länger, um CloudWatch Metrikdaten für die Funktion zur vorausschauenden Skalierung verfügbar zu haben, sofern dies möglich ist.

Vergewissern Sie sich, dass Sie über die erforderlichen Berechtigungen zum Arbeiten mit Skalierungsplänen verfügen. Weitere Informationen finden Sie unter Identitäts- und Zugriffsmanagement für Skalierungspläne.

## Hinzufügen Ihrer Auto-Scaling-Gruppe zu ihrem neuen Skalierungsplan

Wenn Sie einen Skalierungsplan über die Konsole erstellen, hilft es Ihnen als ersten Schritt, Ihre skalierbaren Ressourcen zu finden. Bevor Sie starten, sollten Sie überprüfen, ob die folgenden Anforderungen erfüllt sind:

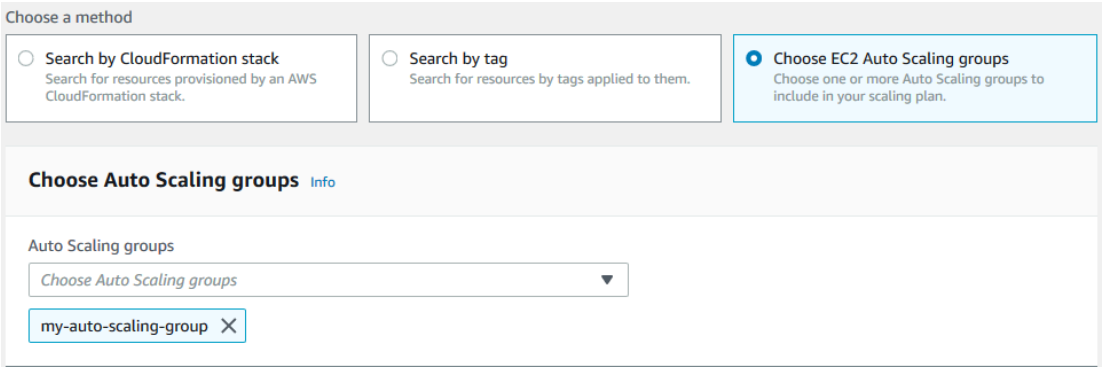
- Sie haben eine Auto-Scaling-Gruppe erstellt und mindestens eine EC2-Instance gestartet, wie im vorherigen Abschnitt beschrieben.

- Die Auto-Scaling-Gruppe, die Sie erstellt haben, existiert seit mindestens 24 Stunden.

So beginnen Sie mit der Erstellung eines Skalierungsplans

1. Öffnen Sie die AWS Auto Scaling Konsole unter. https://console.aws.amazon.com/awsautoscaling/
2. Wählen Sie in der Navigationsleiste oben dieselbe Region aus, die Sie beim Erstellen Ihrer Auto-Scaling-Gruppe verwendet haben.
3. Wählen Sie auf der Willkommensseite die Option Get started (Erste Schritte) aus.
4. Auf der Seite Skalierbare Ressourcen finden führen Sie einen der folgenden Schritte aus:
   - Wählen Sie Nach CloudFormation Stapel suchen und wählen Sie dann den zu CloudFormation verwendenden Stapel aus.
   - Wählen Sie Search by tag (Suche nach Tag). Wählen Sie dann für jedes Tag einen Tag-Schlüssel aus Schlüssel und Tag-Werte aus Wert aus. Um Tags hinzuzufügen, wählen Sie Add another row (Eine weitere Zeile hinzufügen). Um Tags zu entfernen, wählen Sie Remove (Entfernen).
   - Wählen Sie EC2-Auto Scaling-Gruppen auswählen und dann eine oder mehrere Auto-Scaling-Gruppen aus.

> **Note**
>
> Für ein einführendes Tutorial wählen Sie Wählen Sie EC2-Auto-Scaling-Gruppen und wählen Sie dann die Auto-Scaling-Gruppe, die Sie erstellt haben, aus.




5. Klicken Sie auf Weiter um mit dem Erstellungsprozess des Skalierungsplans fortzufahren.

# Weitere Informationen zum Ermitteln Ihrer skalierbaren Ressourcen

Wenn Sie bereits ein Beispiel für einen Skalierungsplan erstellt haben und weitere erstellen möchten, sehen Sie sich die folgenden Szenarien zur Verwendung eines CloudFormation Stacks oder einer Reihe von Tags genauer an. In diesem Abschnitt können Sie entscheiden, ob Sie die Option „Nach CloudFormation Stack suchen“ oder „Nach Tag suchen“ wählen möchten, um Ihre skalierbaren Ressourcen zu ermitteln, wenn Sie die Konsole zur Erstellung Ihres Skalierungsplans verwenden.

Wenn Sie in Schritt 1 des Assistenten „Skalierungsplan erstellen“ die Option „Nach CloudFormation Stapel suchen“ oder „Nach Tag suchen“ auswählen, werden die skalierbaren Ressourcen, die dem Stapel oder der Gruppe von Tags zugeordnet sind, für den Skalierungsplan verfügbar. Beim Definieren Ihres Skalierungsplans können Sie dann auswählen, welche dieser Ressourcen ein- oder ausgeschlossen werden sollen.

Ermitteln skalierbarer Ressourcen mithilfe eines CloudFormation Stacks

Wenn Sie verwenden CloudFormation, arbeiten Sie mit Stacks, um Ressourcen bereitzustellen. Alle Ressourcen in einem Stack werden durch die Vorlage des Stacks definiert. Der Skalierungsplan fügt einen Orchestrierungs-Layer über dem Stack hinzu, der die Konfiguration der Skalierung für mehrere Ressourcen vereinfacht. Ohne einen Skalierungsplan müssen Sie die Skalierung für jede skalierbare Ressource einzeln einrichten. Dies bedeutet, die Reihenfolge für die Bereitstellung von Ressourcen und Skalierungsrichtlinien herauszufinden und die Feinheiten der Funktionsweise dieser Abhängigkeiten zu verstehen.

In der AWS Auto Scaling Konsole können Sie einen vorhandenen Stack auswählen, um ihn nach Ressourcen zu durchsuchen, die für die automatische Skalierung konfiguriert werden können. AWS Auto Scaling findet nur Ressourcen, die im ausgewählten Stack definiert sind. Verschachtelte Stacks werden nicht durchlaufen.

Damit Ihre ECS-Services in einem CloudFormation Stack auffindbar sind, muss die AWS Auto Scaling Konsole wissen, auf welchem ECS-Cluster der Dienst ausgeführt wird. Dies erfordert, dass sich Ihre ECS-Services im selben CloudFormation Stack befinden wie der ECS-Cluster, auf dem der Dienst ausgeführt wird. Andernfalls müssen sie Teil des Standardclusters sein. Um korrekt identifiziert zu werden, muss der Name des ECS-Service auch in jedem dieser ECS-Cluster eindeutig sein.

Weitere Informationen zu CloudFormation finden Sie unter Was ist CloudFormation? im AWS CloudFormation Benutzerhandbuch.

Erkennen skalierbarer Ressourcen mithilfe von Tags

Tags stellen Metadaten bereit, anhand derer mithilfe von Tagfiltern verwandte skalierbare Ressourcen in der AWS Auto Scaling Konsole gefunden werden können.

Verwenden Sie Tags, um die folgenden Ressourcen zu finden:

- Aurora-DB-Cluster
- Auto-Scaling-Gruppen
- DynamoDB-Tabellen und globale sekundäre Indizes

Wenn Sie anhand von mehreren Tags suchen, muss jede Ressource alle aufgelisteten Tags aufweisen, um erkannt zu werden.

Weitere Informationen zur Markierung finden Sie in der folgenden Dokumentation.

- Erfahren Sie im Benutzerhandbuch für Amazon Aurora, wie Sie Aurora-Cluster markieren.
- Informationen zum Markieren von Auto-Scaling-Gruppen finden Sie im Benutzerhandbuch zu Amazon EC2 Auto Scaling.
- Informationen zum Markieren von DynamoDB-Ressourcen finden Sie im Entwicklerhandbuch für Amazon DynamoDB.

## Schritt 2: Festlegen der Skalierungsstrategie

Gehen Sie wie folgt vor, um Skalierungsstrategien für die Ressourcen anzugeben, die im vorherigen Schritt gefunden wurden.

AWS Auto Scaling Wählt für jeden Ressourcentyp die Metrik aus, die am häufigsten verwendet wird, um zu bestimmen, wie viel von der Ressource zu einem bestimmten Zeitpunkt genutzt wird. Sie wählen die am besten geeignete Skalierungsstrategie zur Optimierung der Leistung Ihrer Anwendung basierend auf dieser Metrik aus. Wenn Sie die Funktion der dynamischen Skalierung und die Funktion der prädiktiven Skalierung aktivieren, gilt die Skalierungsstrategie für beide Funktionen. Weitere Informationen finden Sie unter Funktionsweise von Skalierungsplänen.

Die folgenden Skalierungsstrategien stehen zur Verfügung:

- Im Hinblick auf Verfügbarkeit optimieren — Die Ressource wird automatisch nach oben und unten AWS Auto Scaling skaliert, um die Ressourcenauslastung bei 40 Prozent zu halten. Diese Option ist für Anwendungen mit dringendem und manchmal unvorhersehbarem Skalierungsbedarf hilfreich.

- Ausgewogenes Verhältnis zwischen Verfügbarkeit und Kosten — Die Ressource wird automatisch nach oben und unten AWS Auto Scaling skaliert, um die Ressourcenauslastung bei 50 Prozent zu halten. Diese Option unterstützt Sie bei der Aufrechterhaltung der hoher Verfügbarkeit bei gleichzeitiger Kostensenkung.
- Kostenoptimierung —AWS Auto Scaling Skaliert die Ressource automatisch nach oben und unten, um die Ressourcenauslastung bei 70 Prozent zu halten. Diese Option ist zur Senkung von Kosten nützlich, sofern Ihre Anwendung bei unerwarteten Bedarfsänderungen eine reduzierte Pufferkapazität tolerieren kann.

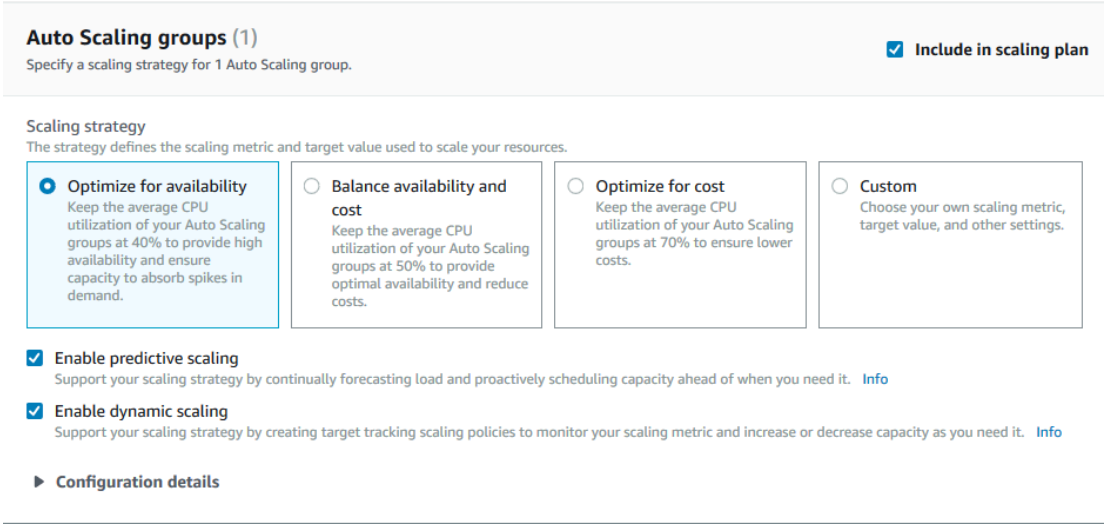
Beispiel: Der Skalierungsplan konfiguriert Ihre Auto-Scaling-Gruppe so, dass Amazon-EC2-Instances basierend auf der durchschnittlichen CPU-Auslastung aller Instances in der Gruppe hinzugefügt oder entfernt werden. Sie können bestimmen, ob die Auslastung für Verfügbarkeit, Kosten oder eine Kombination aus beidem optimiert werden soll, indem Sie die Skalierungsstrategie ändern.

Alternativ können Sie eine benutzerdefinierte Strategie konfigurieren, wenn eine vorhandene Strategie nicht Ihren Anforderungen entspricht. Mit einer benutzerdefinierten Strategie können Sie den Zielauslastungswert ändern, eine andere Metrik auswählen oder beides.

> **Important**
>
> Führen Sie für das Einführungstutorial nur den ersten Schritt des folgenden Verfahrens aus und wählen Sie dann Weiter, um fortzufahren.

So legen Sie eine Skalierungsstrategie fest

1. Geben Sie auf der Seite Specify scaling strategy (Festlegen der Skalierungsstrategie) unter Scaling plan details (Details des Skalierungsplans) für Name einen Namen für Ihren Skalierungsplan ein. Der Name Ihres Skalierungsplans muss innerhalb Ihrer Gruppe von Skalierungsplänen für die Region eindeutig sein. Es darf maximal 128 Zeichen lang sein und darf keine Pipes „|“, Schrägstriche „/“ oder Doppelpunkte „:“ enthalten.
2. Alle enthaltenen Ressourcen werden nach Ressourcentyp aufgelistet. Für Auto-Scaling-Gruppen gehen Sie wie folgt vor:




a. Überspringen Sie diesen Schritt, um die Standardskalierungsstrategie und -metriken zu verwenden. Um stattdessen eine andere Skalierungsstrategie oder Metriken zu verwenden, fahren Sie mit den folgenden Schritten fort:

   i. Wählen Sie für die Skalierungsstrategie die gewünschte Skalierungsstrategie aus.

      Stellen Sie für das Einführungs-Tutorial sicher, dass Sie Auf Verfügbarkeit optimieren auswählen. Dies gibt an, dass die durchschnittliche CPU-Auslastung Ihrer Auto-Scaling-Gruppe bei 40 Prozent gehalten wird.

   ii. Wenn Sie Benutzerdefiniert gewählt haben, erweitern Sie die Konfigurationsdetails, um die gewünschten Metriken und den Zielwert auszuwählen.

      - Wählen Sie für Scaling Metrik (Skalierungsmetrik) die gewünschte Skalierungsmetrik aus.
      - Wählen Sie für Zielwert den gewünschten Zielwert aus, z. B. die Zielauslastung oder den Zieldurchsatz während eines beliebigen einminütigen Intervalls.
      - Wählen Sie für Lastmetrik [nur Auto-Scaling-Gruppen] die gewünschte Lastmetrik aus, die für die prädiktive Skalierung verwendet werden soll.
      - Wählen Sie Externe Skalierungsrichtlinien ersetzen aus, um anzugeben, dass zuvor außerhalb des Skalierungsplans (z. B. von anderen Konsolen) erstellte Skalierungsrichtlinien gelöscht und durch neue Skalierungsrichtlinien für die Zielverfolgung ersetzt werden AWS Auto Scaling können, die vom Skalierungsplan erstellt wurden.

b. (Optional) Standardmäßig ist die prädiktive Skalierung für Auto-Scaling-Gruppen aktiviert. Um die vorausschauende Skalierung für die Auto Scaling-Gruppen zu deaktivieren, deaktivieren Sie Enable predictive scaling (Prädiktive Skalierung aktivieren).

c. (Optional) Standardmäßig ist die dynamische Skalierung für jeden Ressourcentyp aktiviert. Um die dynamische Skalierung für den Ressourcentyp zu deaktivieren, deaktivieren Sie Enable dynamic scaling (Dynamische Skalierung aktivieren).

d. (Optional) Wenn Sie eine Anwendungsquelle angeben, in der mehrere skalierbare Ressourcen erkannt werden, werden alle Ressourcentypen automatisch in Ihren Skalierungsplan eingeschlossen. Um einen Ressourcentyp aus Ihrem Skalierungsplan wegzulassen, deaktivieren Sie Include in scaling plan (In Skalierungsplan aufnehmen).

3. (Optional) Wiederholen Sie die vorhergehenden Schritte, um eine Skalierungsstrategie für einen anderen Ressourcentyp anzugeben.

4. Wenn Sie fertig sind, wählen Sie Weiter, um mit der Skalierungsplanerstellung fortzufahren.

## Schritt 3: Konfigurieren erweiterter Einstellungen (optional)

Nachdem Sie nun die Skalierungsstrategie festgelegt haben, die für jeden Ressourcentyp verwendet werden soll, können Sie mit dem Schritt Configure advanced settings (Konfigurieren von erweiterten Einstellungen) bei Bedarf alle Standardeinstellungen für jede Ressource anpassen. Für jeden Ressourcentyp gibt es mehrere Gruppen von Einstellungen, die Sie anpassen können. In den meisten Fällen sollten die Standardeinstellungen jedoch effizienter sein, mit Ausnahme der Werte für minimale Kapazität und maximale Kapazität, die sorgfältig angepasst werden sollten.

Überspringen Sie diese Schritte, falls Sie die Standardeinstellungen behalten möchten. Sie können diese Einstellungen jederzeit ändern, indem Sie den Skalierungsplan bearbeiten.

> **Important**
>
> Lassen Sie uns für das Einsteigertutorial einige Änderungen vornehmen, um die maximale Kapazität Ihrer Auto-Scaling-Gruppe zu aktualisieren und die prädiktive Skalierung im Nur-Prognose-Modus zu aktivieren. Obwohl Sie nicht alle Einstellungen für das Tutorial anpassen müssen, wird kurz auf die Einstellungen in jedem Abschnitt eingegangen.

# Allgemeine Einstellungen

Gehen Sie wie folgt vor, um die von Ihnen im vorherigen Schritt angegebenen Einstellungen für jede Ressource anzuzeigen und anzupassen. Sie können auch die minimale Kapazität und die maximale Kapazität einer jeden Ressource anpassen.

So verfahren Sie zum Anzeigen und Anpassen der allgemeinen Einstellungen

1. Wählen Sie auf der Seite Configure advanced settings (Konfigurieren von erweiterten Einstellungen) den Pfeil links neben beliebigen Abschnittsüberschrift aus, um den Abschnitt zu erweitern. Erweitern Sie für das Tutorial den Bereich Auto Scaling groups (Auto-Scaling-Gruppen).
2. Wählen Sie in der angezeigten Tabelle die Auto-Scaling-Gruppe aus, die Sie in diesem Tutorial verwenden.
3. Lassen Sie die Option Include in scaling plan (Aufnahme in den Skalierungsplan) ausgewählt. Wenn diese Option nicht ausgewählt ist, wird die Ressource aus dem Skalierungsplan ausgelassen. Wenn Sie nicht mindestens eine Ressource einschließen, kann der Skalierungsplan nicht erstellt werden.
4. Um die Ansicht zu erweitern und die Details des Abschnitts General Settings (Allgemeine Einstellungen) anzuzeigen, wählen Sie den Pfeil links neben der Abschnittsüberschrift aus.
5. Sie können für beliebige der folgenden Elemente eine Auswahl treffen. In diesem Tutorial suchen Sie die Einstellung Maximum capacity (Maximale Kapazität) und geben anstelle des aktuellen Wertes den Wert `3` ein.

   - Scaling strategy (Skalierungsstrategie) – Mit dieser Option können Sie die Optimierung auf Verfügbarkeit, Kosten oder ein ausgewogenes Verhältnis aus beidem festlegen oder eine benutzerdefinierte Strategie angeben.
   - Enable dynamic scaling (Dynamische Skalierung aktivieren) – Wenn diese Einstellung deaktiviert wird, kann die ausgewählte Ressource nicht mithilfe einer Zielverfolgungs-Skalierungskonfiguration skaliert werden.
   - Enable predictive scaling (Prädiktive Skalierung aktivieren) – [Nur Auto-Scaling-Gruppen] Wenn Sie diese Einstellung deaktivieren, kann die ausgewählte Gruppe mithilfe prädiktiver Skalierung nicht skaliert werden.
   - Scaling metric (Skalierungsmetrik) – Gibt die zu verwendende Skalierungsmetrik an. Wenn Sie Custom (Benutzerdefiniert) auswählen, können Sie eine benutzerdefinierte Metrik angeben,

die anstelle der in der Konsole verfügbaren vordefinierten Metriken verwendet werden soll. Weitere Informationen finden Sie im nächsten Thema in diesem Abschnitt.

- Target value (Zielwert) – Gibt den zu verwendenden Wert für die Zielauslastung an.
- Load metric (Lastmetrik) – [Nur Auto-Scaling-Gruppen] Legt die zu verwendende Lastmetrik fest. Wenn Sie Custom (Benutzerdefiniert) auswählen, können Sie eine benutzerdefinierte Metrik angeben, die anstelle der in der Konsole verfügbaren vordefinierten Metriken verwendet werden soll. Weitere Informationen finden Sie im nächsten Thema in diesem Abschnitt.
- Mindestkapazität — Gibt die Mindestkapazität für die Ressource an. AWS Auto Scaling stellt sicher, dass Ihre Ressource diese Größe niemals unterschreitet.
- Maximale Kapazität — Gibt die maximale Kapazität für die Ressource an. AWS Auto Scaling stellt sicher, dass Ihre Ressource diese Größe niemals überschreitet.

> **Note**
>
> Wenn Sie die prädiktive Skalierung verwenden, können Sie optional ein anderes Verhalten für die maximale Kapazität wählen, das basierend auf der prognostizierten Kapazität anzuwenden ist. Diese Einstellung befindet sich im Abschnitt Predictive scaling settings (Prädiktive Skalierungseinstellungen).

## Benutzerdefinierte Metriken

AWS Auto Scaling bietet die am häufigsten verwendeten Metriken für die automatische Skalierung. Je nach Ihren Anforderungen möchten Sie möglicherweise Daten von anderen Metriken als denen in der Konsole erhalten. Amazon CloudWatch hat viele verschiedene Metriken zur Auswahl. CloudWatch ermöglicht es Ihnen auch, Ihre eigenen Metriken zu veröffentlichen.

Sie verwenden JSON, um eine CloudWatch benutzerdefinierte Metrik anzugeben. Bevor Sie diese Anweisungen befolgen, empfehlen wir Ihnen, sich mit dem CloudWatch Amazon-Benutzerhandbuch vertraut zu machen.

Um eine benutzerdefinierte Metrik anzugeben, müssen Sie eine JSON-formatierte Nutzlast mit einer Reihe von erforderlichen Parametern aus einer Vorlage erstellen. Sie fügen die Werte für jeden Parameter von hinzu CloudWatch. Wir stellen die Vorlage als Teil der benutzerdefinierten Optionen für Scaling metric (Skalierungsmetrik) und Load metric (Lastmetrik) in den erweiterten Einstellungen Ihres Skalierungsplans bereit.

JSON stellt Daten auf zwei Arten dar:

- Ein Objekt, bei dem es sich um eine ungeordnete Sammlung von Name-Wert-Paaren handelt. Ein Objekt wird innerhalb von zwei Klammern ({ und }) definiert. Jedes Name-Wert-Paar beginnt mit dem Namen, gefolgt von einem Doppelpunkt und dem Wert. Name-Wert-Paare sind durch Kommas voneinander getrennt.
- Ein Array, bei dem es sich um eine geordnete Sammlung von Werten handelt. Ein Objekt wird innerhalb von zwei Klammern ([ und ]) definiert. Elemente im Array werden durch Kommas voneinander getrennt.

Hier ist ein Beispiel der JSON-Vorlage mit Beispielwerten für jeden Parameter:

```
{
  "MetricName": "MyBackendCPU",
  "Namespace": "MyNamespace",
  "Dimensions": [
    {
      "Name": "MyOptionalMetricDimensionName",
      "Value": "MyOptionalMetricDimensionValue"
    }
  ],
  "Statistic": "Sum"
}
```

Weitere Informationen finden Sie unter Angepasste Skalierungsmetrik-Spezifikation und Angepasste Lastmetrik-Spezifikation in der AWS Auto Scaling -API-Referenz.

## Dynamische Skalierungseinstellungen

Gehen Sie wie folgt vor, um die Einstellungen für die Ziel-Tracking-Skalierungsrichtlinie, die AWS Auto Scaling erstellt wird, anzuzeigen und anzupassen.

So verfahren Sie zum Anzeigen und Anpassen der Einstellungen für die dynamische Skalierung

1. Um die Ansicht zu erweitern und die Details des Abschnitts Dynamic scaling settings (Dynmaische Skalierungseinstellungen) anzuzeigen, wählen Sie den Pfeil links neben der Abschnittsüberschrift aus.
2. Sie können für die folgenden Elemente eine Auswahl treffen. Die Standardeinstellungen sind für dieses Tutorial aber völlig ausreichend.

- Replace external scaling policies (Ersetzen von externen Skalierungsrichtlinien) – Wenn diese Einstellung deaktiviert wird, werden vorhandene Skalierungsrichtlinien, die außerhalb des Skalierungsplans erstellt wurden, beibehalten und keine neuen erstellt.
- Disable scale-in (Abskalierung deaktivieren) – Wenn diese Einstellung deaktiviert wird, ist das automatische Abskalieren zur Verringerung der aktuellen Kapazität der Ressource zulässig, wenn die angegebene Metrik unter dem Zielwert liegt.
- Cooldown (Ruhephase) – Erstellt Ruhephasen für die Auf- und Abskalierung. Die Ruhephase ist die Zeitspanne, die die Skalierungsrichtlinie warten muss, bis eine vorherige Skalierungsaktivität wirksam ist. Weitere Informationen finden Sie unter Ruhephase im Benutzerhandbuch für Application Auto Scaling. (Diese Einstellung wird nicht angezeigt, wenn die Ressource eine Auto-Scaling-Gruppe ist.)
- Instance-Warmup — [Nur Auto Scaling Scaling-Gruppen] Steuert die Zeit, die vergeht, bis eine neu gestartete Instance anfängt, zu den Metriken beizutragen. CloudWatch Weitere Informationen finden Sie unter Instance-Aufwärmphase im Benutzerhandbuch für Amazon EC2 Auto Scaling.

## Prädiktive Skalierungseinstellungen

Wenn es sich bei Ihrer Ressource um eine Auto Scaling Scaling-Gruppe handelt, verwenden Sie dieses Verfahren, um die für die prädiktive Skalierung AWS Auto Scaling verwendeten Einstellungen anzuzeigen und anzupassen.

So verfahren Sie zum Anzeigen und Anpassen der Einstellungen für die prädiktive Skalierung

1. Um die Ansicht zu erweitern und die Details des Abschnitt Predictive scaling settings (Prädiktive Skalierungseinstellungen) anzuzeigen, wählen Sie den Pfeil links neben der Abschnittsüberschrift aus.
2. Sie können für die folgenden Elemente eine Auswahl treffen. In diesem Tutorial ändern Sie den Wert für Predictive scaling mode (Modus der prädiktiven Skalierung) in Forecast only (Nur Prognose).

   - Predictive scaling mode (Modus der prädiktiven Skalierung) – Legt die Skalierungsmethode fest. Der Standardwert ist Forecast and scale (Prognose und Skalierung). Wenn Sie diese Einstellung in Forecast only (Nur Prognose) ändern, prognostiziert der Skalierungsplan die zukünftige Kapazität, wendet die Skalierungsaktionen aber nicht an.

- Pre-launch instances (Vorabstarten von Instances) – Sorgt dafür, dass die Skalierungsaktionen bei der Aufskalierung früher ausgeführt werden. Beispiel: Die Prognose gibt vor, die Kapazität um 10:00 Uhr hinzuzufügen, und die Pufferzeit beträgt 5 Minuten (300 Sekunden). Die Laufzeit der entsprechenden Skalierungsaktion ist dann 9:55 Uhr. Dies ist hilfreich für Auto-Scaling-Gruppen, in denen es einige Minuten vom Start einer Instance bis zur Inbetriebnahme dauern kann. Die tatsächlich benötigte Zeit kann davon abweichen, da sie natürlich von mehreren Faktoren abhängt, z. B. der Größe der Instance und ob Startskripts ausgeführt werden müssen. Standardmäßig ist ein Zeitraum von 300 Sekunden festgelegt.
- Max Capacity Behavior (Verhalten bei max. Kapazität) – Steuert, ob die ausgewählte Ressource über die maximale Kapazität hochskaliert werden kann, wenn die prognostizierte Kapazität nahe an oder über der aktuell angegebenen maximalen Kapazität liegt.
  Die Standardeinstellung lautet Enforce the maximum capacity setting (Maximale Kapazitätseinstellung erzwingen).
  - Erzwingen Sie die Einstellung für die maximale Kapazität — die Ressourcenkapazität AWS Auto Scaling kann nicht höher als die maximale Kapazität skaliert werden. Die maximale Kapazität wird als ein hartes Limit durchgesetzt.
  - Stellen Sie die maximale Kapazität so ein, dass sie der prognostizierten Kapazität entspricht — die Ressourcenkapazität AWS Auto Scaling kann höher als die maximale Kapazität skaliert werden, um die prognostizierte Kapazität zu erreichen, aber nicht zu überschreiten.
  - Erhöhen Sie die maximale Kapazität über die prognostizierte Kapazität — die Ressourcenkapazität AWS Auto Scaling kann um einen bestimmten Pufferwert höher als die maximale Kapazität skaliert werden. Der Zweck ist, der Skalierungsrichtlinie für das Ziel-Tracking zusätzliche Kapazität zu verschaffen, wenn es zu unerwartetem Datenverkehr kommt.
- Max capacity behavior buffer (Puffer für Verhalten bei maximaler Kapazität) – Wenn Sie Increase maximum capacity above forecast capacity (Maximale Kapazität über die prognostizierte Kapazität hinaus erhöhen) gewählt haben, wählen Sie die Größe des Kapazitätpuffers aus, der verwendet werden soll, wenn die prognostizierte Kapazität nahe der maximalen Kapazität ist oder diese überschreitet. Der Wert wird als Prozentsatz relativ zur prognostizierten Kapazität angegeben. Wenn die prognostizierte Kapazität bei einem Puffer von 10 Prozent 50 ist und die maximale Kapazität 40 ist, dann ist die effektive maximale Kapazität 55.

3. Wenn Sie die Anpassung der Einstellungen abgeschlossen haben, klicken Sie auf Next (Weiter).

> **Note**
>
> Um eine Änderung rückgängig zu machen, markieren Sie die Ressourcen und wählen Sie Revert to original (Zurücksetzen auf Original) aus. Dadurch werden die ausgewählten Ressourcen in ihren letzten bekannten Zustand innerhalb des Skalierungsplans zurückgesetzt.

## Schritt 4: Erstellen des Skalierungsplans

Überprüfen Sie auf der Seite Review and create (Überprüfen und erstellen) die Details zu Ihrem Skalierungsplan und wählen Sie Create scaling plan (Skalierungsplan erstellen) aus. Sie werden zu einer Seite weitergeleitet, auf der der Status Ihres Skalierungsplans angezeigt wird. Es kann eine Weile dauern, bis der Skalierungsplan fertig erstellt ist, während Ihre Ressourcen aktualisiert werden.

Bei der prädiktiven Skalierung wird der Verlauf der angegebenen Lastmetrik in den letzten 14 Tagen AWS Auto Scaling analysiert (Daten von mindestens 24 Stunden sind erforderlich), um eine Prognose für die nächsten zwei Tage zu erstellen. Anschließend werden Skalierungsaktionen geplant, um die Ressourcenkapazität an die Prognosen für die einzelnen Stunden im Prognosezeitraum anzupassen.

Nachdem der Skalierungsplan fertig erstellt wurde, überprüfen Sie die Details des Skalierungsplans, indem Sie seinen Namen im Bildschirm Scaling plans (Skalierungspläne) auswählen.

### (Optional) Anzeigen von Skalierungsinformationen für eine Ressource

Verwenden Sie dieses Verfahrens, um die für eine Ressource erstellten Skalierungsinformationen anzuzeigen.

Die Daten werden folgendermaßen präsentiert:

- Grafiken mit aktuellen Metrikverläufen von CloudWatch.
- Diagramme zur prädiktiven Skalierung mit Last- und Kapazitätsprognosen auf der Grundlage von Daten von AWS Auto Scaling.
- Eine Tabelle, die alle prädiktiven Skalierungsaktionen auflistet, die für die Ressource geplant sind.

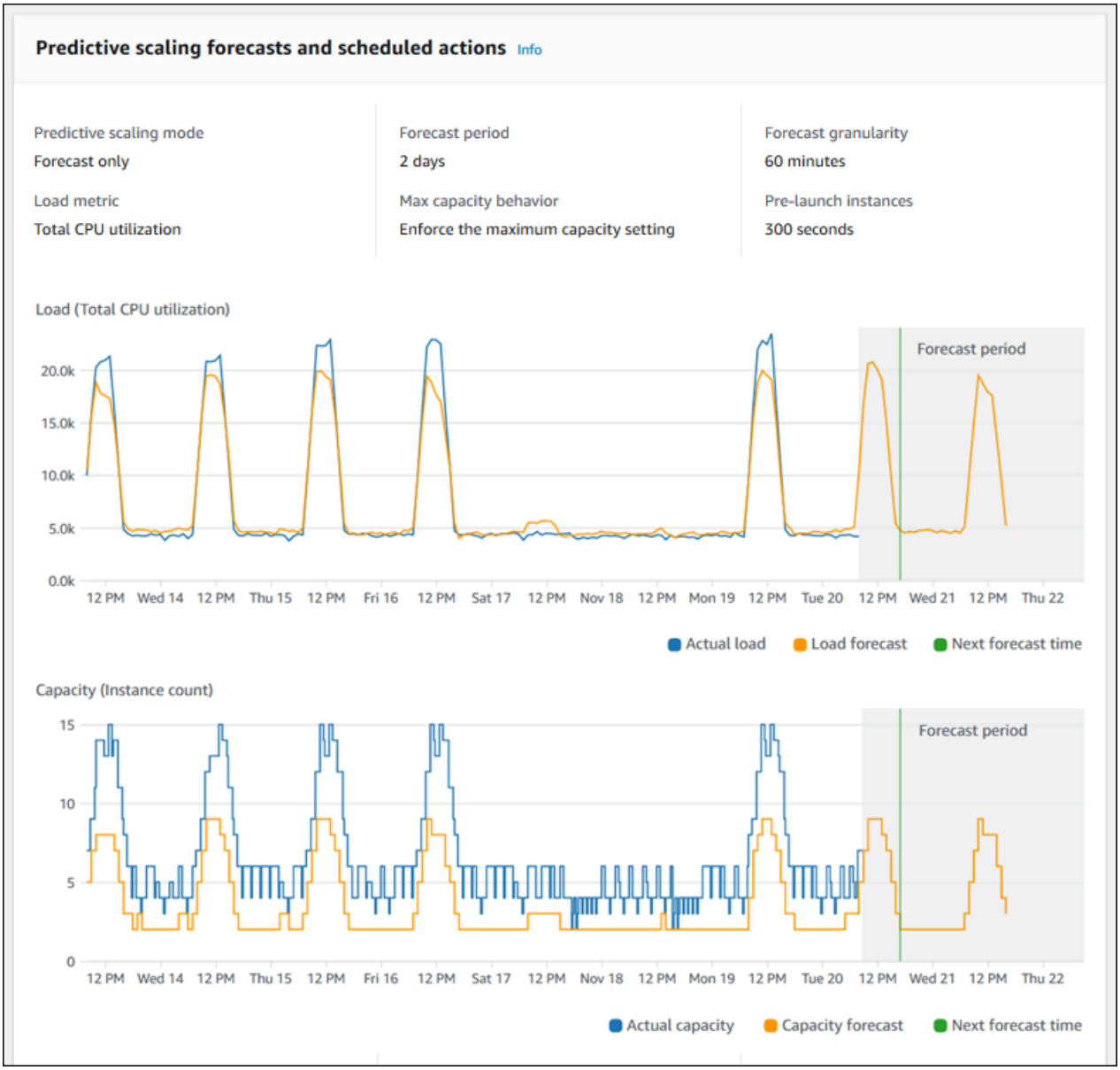
So zeigen Sie Skalierungsinformationen für eine Ressource an

1. Öffnen Sie die AWS Auto Scaling Konsole unter. https://console.aws.amazon.com/awsautoscaling/
2. Wählen Sie auf der Seite Scaling plans (Skalierungspläne) den Skalierungsplan aus.
3. Wählen Sie auf der Seite Scaling plan details (Details des Skalierungsplans) die anzuzeigende Ressource aus.

## Überwachen und Auswerten von Prognosen

Wenn Ihr Skalierungsplan eingerichtet ist und ausgeführt wird, können Sie die Lastprognose, die Kapazitätsprognose und die Skalierungsaktionen überwachen, um die Leistung der prädiktiven Skalierung zu überprüfen. All diese Daten sind in der AWS Auto Scaling Konsole für alle Auto Scaling Scaling-Gruppen verfügbar, die für prädiktive Skalierung aktiviert sind. Denken Sie daran, dass für Ihren Skalierungsplan mindestens 24 Stunden an historischen Daten benötigt werden, um die anfängliche Prognose zu fällen.

Im folgenden Beispiel zeigt die linke Seite der einzelnen Diagramme ein historisches Muster. Die rechte Seite des Diagramms zeigt die Prognose, die vom Skalierungsplan für den Prognosezeitraum erstellt wurde. Es werden sowohl tatsächliche als auch prognostizierte Werte (blau und orange) dargestellt.




AWS Auto Scaling lernt automatisch aus Ihren Daten. Zuerst wird eine Lastprognose erstellt. Anschließend bestimmt eine Berechnung der Kapazitätsprognose die Mindestanzahl der Instances, die erforderlich ist, um die Anwendung zu unterstützen. Basierend auf der Kapazitätsprognose, plant AWS Auto Scaling Skalierungsaktionen, die die Auto-Scaling-Gruppe entsprechend der vorausgesagten Laständerungen skalieren. Wenn die dynamische Skalierung aktiviert ist (empfohlen), kann die Auto-Scaling-Gruppe zusätzliche Kapazitäten skalieren (oder entfernen), je nachdem wie ausgelastet die Instances-Gruppe derzeit ist.

Wenn Sie die Leistung der prädiktiven Skalierung bewerten, überwachen Sie, wie stark die tatsächlichen und prognostizierten Werte im Laufe der Zeit übereinstimmen. Wenn Sie einen Skalierungsplan erstellen, AWS Auto Scaling stellt es Grafiken bereit, die auf den neuesten

tatsächlichen Daten basieren. Außerdem wird die anfängliche Prognose für die nächsten 48 Stunden zur Verfügung gestellt. Nachdem der Skalierungsplan erstellt wurde, ist jedoch nur eine sehr geringe Menge an Prognosedaten für einen Vergleich mit den tatsächlichen Daten vorhanden. Warten Sie, bis der Skalierungsplan Prognosewerte für einige Zeiträume abgerufen hat, bevor Sie die historischen Prognosewerte mit den tatsächlichen Werte vergleichen. Nach einigen Tagen täglicher Prognosen verfügen Sie über einen größere Stichprobe von Prognosewerten für den Vergleich mit tatsächlichen Werten.

Für Mustern, die täglich auftreten, kann das Zeitintervall zwischen der Erstellung Ihres Skalierungsplans und der Bewertung der Wirksamkeit der Prognose nur ein paar Tage betragen. Eine solche Dauer ist für die Bewertung der Prognose basierend auf einer kürzlichen Musteränderung allerdings nicht ausreichend. Angenommen, Sie überprüfen die Prognose für eine Auto-Scaling-Gruppe, die in der letzten Woche eine neue Marketingkampagne gestartet hat. Die Kampagne erhöht den Webdatenverkehr für die gleichen zwei Wochentage deutlich. In solchen Situationen empfehlen wir, mit der Auswertung der Wirksamkeit der Prognose zu warten, bis die Gruppe eine oder zwei vollständige Wochen an neuen Daten erfasst hat. Diese Empfehlung gilt auch für eine völlig neue Auto-Scaling-Gruppe, die gerade mit dem Sammeln von metrischen Daten begonnen hat.

Wenn die über einen geeigneten Zeitraum hinweg überwachten tatsächlichen und prognostizierten Werte nicht übereinstimmen, sollten Sie auch die von Ihnen ausgewählte Lastmetrik in Betracht ziehen. Um die Effizienz zu gewährleisten, muss die Lastmetrik eine zuverlässige und genaue Messung der Gesamtlast für alle Instances in der Auto-Scaling-Gruppe repräsentieren. Die Lastmetrik steht im Mittelpunkt der prädiktiven Skalierung. Wenn Sie eine suboptimale Lastmetrik wählen, können dadurch genaue Last- und Kapazitätsprognosen sowie die Planung der richtigen Kapazitätsberichtigungen für Ihre Auto-Scaling-Gruppe durch die prädiktive Skalierung verhindert werden.

# Schritt 5: Bereinigen

Nachdem Sie das Tutorial „Erste Schritte“ abgeschlossen haben, können Sie Ihren Skalierungsplan auf Wunsch beibehalten. Wenn Sie Ihren Skalierungsplan jedoch nicht aktiv verwenden, sollten Sie ihn löschen, damit Ihrem Konto keine unnötigen Gebühren berechnet werden.

Beim Löschen eines Skalierungsplans werden die Skalierungsrichtlinien für die Zielverfolgung, die zugehörigen CloudWatch Alarme und die prädiktiven Skalierungsaktionen, die in Ihrem Namen AWS Auto Scaling erstellt wurden, gelöscht.

Durch das Löschen eines Skalierungsplans werden Ihr CloudFormation Stack, Ihre Auto Scaling Scaling-Gruppe oder andere skalierbare Ressourcen nicht gelöscht.

So löschen Sie einen Skalierungsplan

1. Öffnen Sie die AWS Auto Scaling Konsole unter https://console.aws.amazon.com/awsautoscaling/.
2. Wählen Sie auf der Seite Scaling plans (Skalierungspläne) den Skalierungsplan aus, den Sie für dieses Tutorial erstellt haben, und klicken Sie auf Delete (Löschen).
3. Wenn Sie zur Bestätigung aufgefordert werden, wählen Sie Delete (Löschen).

Wenn Sie Ihren Skalierungsplan löschen, kehren Ihre Ressourcen nicht wieder zu ihrer ursprünglichen Kapazität zurück. Beispiel: Ist Ihre Auto-Scaling-Gruppe beim Löschen des Skalierungsplans auf 10 Instances skaliert, ist sie nach dem Löschen des Skalierungsplans weiterhin auf 10 Instances skaliert. Sie können die Kapazität bestimmter Ressourcen aktualisieren, indem Sie auf die Konsole für jeden einzelnen Service zugreifen.

### Löschen der Auto-Scaling-Gruppe

Um zu verhindern, dass für Ihr Konto Amazon-EC2-Gebühren entstehen, sollten Sie auch die Auto-Scaling-Gruppe löschen, die Sie für dieses Tutorial erstellt haben.

step-by-stepAnweisungen finden Sie unter Löschen Ihrer Auto Scaling Scaling-Gruppe im Amazon EC2 Auto Scaling Scaling-Benutzerhandbuch.

## Schritt 6: Nächste Schritte

Nachdem Sie sich nun mit Skalierungsplänen und einigen ihrer Funktionen vertraut gemacht haben, können Sie versuchen, Ihre eigene Skalierungsplanvorlage mit CloudFormation zu erstellen.

Eine CloudFormation Vorlage ist eine Textdatei im JSON- oder YAML-Format, die die Amazon Web Services Services-Infrastruktur beschreibt, die für die Ausführung einer Anwendung oder eines Dienstes erforderlich ist, sowie alle Verbindungen zwischen Infrastrukturkomponenten. Mit CloudFormation stellen Sie eine zugehörige Sammlung von Ressourcen als Stapel bereit und verwalten sie. CloudFormation ist ohne zusätzliche Kosten erhältlich, und Sie zahlen nur für die AWS Ressourcen, die Sie für die Ausführung Ihrer Anwendungen benötigen. Ressourcen können aus jeder AWS Ressource bestehen, die Sie in der Vorlage definieren. Weitere Informationen finden Sie im AWS CloudFormation Benutzerhandbuch unter So CloudFormation funktioniert es.

Im AWS CloudFormation -Benutzerhandbuch stellen wir Ihnen zum Einstieg eine einfache Vorlage zur Verfügung. Die Beispielvorlage ist als Beispiel im AWS::AutoScalingPlans::ScalingPlanAbschnitt

der Referenzdokumentation zur CloudFormation Vorlage verfügbar. Mit der Beispielvorlage werden einen Skalierungsplan für eine einzelne Auto-Scaling-Gruppe erstellt und sowohl eine prädiktive als auch dynamische Skalierung ermöglicht.

Weitere Informationen finden Sie unter Erste Schritte in CloudFormation im AWS CloudFormation - Benutzerhandbuch.

# Migrieren Sie Ihren Skalierungsplan

Sie können von einem Skalierungsplan zu Amazon EC2 Auto Scaling- und Application Auto Scaling Scaling-Richtlinien migrieren.

Migrationsprozess



> **Important**
>
> Um einen Skalierungsplan zu migrieren, müssen Sie mehrere Schritte in exakter Reihenfolge ausführen. Während Sie Ihren Skalierungsplan migrieren, sollten Sie ihn nicht aktualisieren, da dies die Reihenfolge der Vorgänge beeinträchtigt und zu unerwünschtem Verhalten führen kann.

## Schritt 1: Überprüfen Sie Ihr vorhandenes Setup

Verwenden Sie den describe-scaling-plansBefehl, um zu ermitteln, welche Skalierungseinstellungen Sie ändern müssen.

```
aws autoscaling-plans describe-scaling-plans \
  --scaling-plan-names my-scaling-plan
```

Notieren Sie sich die Elemente, die Sie aus dem vorhandenen Skalierungsplan beibehalten möchten. Dies kann Folgendes beinhalten:

- **`MinCapacity`**— Die Mindestkapazität der skalierbaren Ressource.
- **`MaxCapacity`**— Die maximale Kapazität der skalierbaren Ressource.
- **`PredefinedLoadMetricType`**— Eine Lastmetrik für vorausschauende Skalierung.
- **`PredefinedScalingMetricType`**— Eine Skalierungsmetrik für die (dynamische) Skalierung der Zielverfolgung und die prädiktive Skalierung.
- **`TargetValue`**— Der Zielwert für die Skalierungsmetrik.

### Unterschiede zwischen Skalierungsplänen und Skalierungsrichtlinien

Es gibt einige wichtige Unterschiede zwischen Skalierungsplänen und Skalierungsrichtlinien:

- Eine Skalierungsrichtlinie kann nur eine Art der Skalierung ermöglichen: entweder zielgerichtete Skalierung oder prädiktive Skalierung. Um beide Skalierungsmethoden zu verwenden, müssen Sie separate Richtlinien erstellen.
- Ebenso müssen Sie die Skalierungsmetrik für die prädiktive Skalierung und die Skalierungsmetrik für die Skalierung der Zielverfolgung innerhalb der jeweiligen Richtlinien getrennt definieren.

## Schritt 2: Erstellen Sie Richtlinien für die prädiktive Skalierung

Wenn Sie Predictive Scaling nicht verwenden, fahren Sie fort mit. Schritt 4: Bereiten Sie das Löschen des Skalierungsplans vor

Um Zeit für die Bewertung der Prognose zu haben, empfehlen wir, dass Sie Richtlinien für die vorausschauende Skalierung erstellen, bevor Sie andere Skalierungsrichtlinien verwenden.

Gehen Sie für alle Auto Scaling-Gruppen mit einer vorhandenen Lastmetrikspezifikation wie folgt vor, um sie in eine auf Amazon EC2 Auto Scaling basierende Predictive Scaling-Richtlinie umzuwandeln.

Um Richtlinien für vorausschauende Skalierung zu erstellen

1. Definieren Sie in einer JSON-Datei eine `MetricSpecifications` Struktur, wie im folgenden Beispiel gezeigt:

```
{
```

```
  "MetricSpecifications":[
    {
      ...
    }
  ]
}
```

2. Erstellen Sie in der `MetricSpecifications` Struktur für jede Lastmetrik in Ihrem Skalierungsplan eine `PredefinedLoadMetricSpecification` oder `CustomizedLoadMetricSpecification` mit den entsprechenden Einstellungen aus dem Skalierungsplan.

   Im Folgenden finden Sie Beispiele für die Struktur des Abschnitts „Lastmetrik“.

   With predefined metrics

   ```
   {
     "MetricSpecifications":[
       {
         "PredefinedLoadMetricSpecification":{
           "PredefinedMetricType":"ASGTotalCPUUtilization"
         },
         ...
       }
     ]
   }
   ```

   Weitere Informationen finden Sie PredictiveScalingPredefinedLoadMetricin der Amazon EC2 Auto Scaling API-Referenz.

   With custom metrics

   ```
   {
     "MetricSpecifications":[
       {
         "CustomizedLoadMetricSpecification":{
           "MetricDataQueries":[
             {
               "Id":"load_metric",
               "MetricStat":{
                 "Metric":{
                   "MetricName":"MyLoadMetric",
                   "Namespace":"MyNameSpace",
   ```

```
                "Dimensions":[
                  {
                    "Name":"MyOptionalMetricDimensionName",
                    "Value":"MyOptionalMetricDimensionValue"
                  }
                ]
              },
              "Stat":"Sum"
            }
          }
        ]
      },
      ...
    }
  ]
}
```

Weitere Informationen finden Sie PredictiveScalingCustomizedLoadMetricin der Amazon EC2 Auto Scaling API-Referenz.

3. Fügen Sie die Skalierungsmetrikspezifikation zur hinzu `MetricSpecifications` und definieren Sie einen Zielwert.

   Im Folgenden finden Sie Beispiele für die Struktur der Abschnitte Skalierungsmetrik und Zielwert.

   With predefined metrics

```
{
  "MetricSpecifications":[
    {
      "PredefinedLoadMetricSpecification":{
        "PredefinedMetricType":"ASGTotalCPUUtilization"
      },
      "PredefinedScalingMetricSpecification":{
        "PredefinedMetricType":"ASGCPUUtilization"
      },
      "TargetValue":50
    }
  ],
  ...
}
```

Weitere Informationen finden Sie PredictiveScalingPredefinedScalingMetricin der Amazon EC2 Auto Scaling API-Referenz.

With custom metrics

```
{
  "MetricSpecifications":[
    {
      "CustomizedLoadMetricSpecification":{
        "MetricDataQueries":[
          {
            "Id":"load_metric",
            "MetricStat":{
              "Metric":{
                "MetricName":"MyLoadMetric",
                "Namespace":"MyNameSpace",
                "Dimensions":[
                  {
                    "Name":"MyOptionalMetricDimensionName",
                    "Value":"MyOptionalMetricDimensionValue"
                  }
                ]
              },
              "Stat":"Sum"
            }
          }
        ]
      },
      "CustomizedScalingMetricSpecification":{
        "MetricDataQueries":[
          {
            "Id":"scaling_metric",
            "MetricStat":{
              "Metric":{
                "MetricName":"MyUtilizationMetric",
                "Namespace":"MyNameSpace",
                "Dimensions":[
                  {
                    "Name":"MyOptionalMetricDimensionName",
                    "Value":"MyOptionalMetricDimensionValue"
                  }
                ]
              },
```

```
                "Stat":"Average"
              }
            }
          ]
        },
        "TargetValue":50
      }
    ],
    ...
}
```

Weitere Informationen finden Sie PredictiveScalingCustomizedScalingMetricin der Amazon EC2 Auto Scaling API-Referenz.

4. Um nur Prognosen zu erstellen, fügen Sie die Eigenschaft `Mode` mit einem Wert von `ForecastOnly` hinzu. Nachdem Sie die Migration der prädiktiven Skalierung abgeschlossen und sichergestellt haben, dass die Prognose korrekt und zuverlässig ist, können Sie den Modus ändern, sodass eine Skalierung möglich ist. Weitere Informationen finden Sie unter Schritt 7: Reaktivieren Sie die prädiktive Skalierung.

```
{
  "MetricSpecifications":[
  ...
  ],
  "Mode":"ForecastOnly",
  ...
}
```

Weitere Informationen finden Sie PredictiveScalingConfigurationin der Amazon EC2 Auto Scaling API-Referenz.

5. Wenn die **`ScheduledActionBufferTime`** Eigenschaft in Ihrem Skalierungsplan enthalten ist, kopieren Sie ihren Wert in die `SchedulingBufferTime` Eigenschaft in Ihrer Richtlinie zur vorausschauenden Skalierung.

```
{
  "MetricSpecifications":[
  ...
  ],
  "Mode":"ForecastOnly",
  "SchedulingBufferTime":300,
  ...
```

```
}
```

Weitere Informationen finden Sie PredictiveScalingConfigurationin der Amazon EC2 Auto Scaling API-Referenz.

6. Wenn die **`PredictiveScalingMaxCapacityBuffer`** Eigenschaften **`PredictiveScalingMaxCapacityBehavior`** und in Ihrem Skalierungsplan vorhanden sind, können Sie die `MaxCapacityBuffer` Eigenschaften `MaxCapacityBreachBehavior` und in Ihrer Richtlinie für vorausschauende Skalierung konfigurieren. Diese Eigenschaften definieren, was passieren soll, wenn sich die prognostizierte Kapazität der für die Auto Scaling Scaling-Gruppe angegebenen maximalen Kapazität nähert oder diese überschreitet.

> **Warning**
>
> Wenn Sie die `MaxCapacityBreachBehavior` Eigenschaft auf setzen`IncreaseMaxCapacity`, könnten mehr Instances als vorgesehen gestartet werden, sofern Sie die erhöhte maximale Kapazität nicht überwachen und verwalten. Die erhöhte maximale Kapazität wird zur neuen normalen maximalen Kapazität für die Auto Scaling Scaling-Gruppe, bis Sie sie manuell aktualisieren. Die maximale Kapazität wird nicht automatisch wieder auf das ursprüngliche Maximum reduziert.

```
{
  "MetricSpecifications":[
  ...
  ],
  "Mode":"ForecastOnly",
  "SchedulingBufferTime":300,
  "MaxCapacityBreachBehavior": "IncreaseMaxCapacity",
  "MaxCapacityBuffer": 10
}
```

Weitere Informationen finden Sie PredictiveScalingConfigurationin der Amazon EC2 Auto Scaling API-Referenz.

7. Speichern Sie die JSON-Datei mit einem eindeutigen Namen. Notieren Sie sich den Dateinamen. Sie benötigen ihn im nächsten Schritt und erneut am Ende des Migrationsvorgangs, wenn Sie Ihre Predictive Scaling-Richtlinien reaktivieren. Weitere Informationen finden Sie unter Schritt 7: Reaktivieren Sie die prädiktive Skalierung.

8. Nachdem Sie Ihre JSON-Datei gespeichert haben, führen Sie den put-scaling-policyBefehl aus. Ersetzen Sie im folgenden Beispiel jede *user input placeholder* durch Ihre eigenen Informationen.

```
aws autoscaling put-scaling-policy --policy-name my-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://my-predictive-scaling-config.json
```

   Wenn der Befehl erfolgreich ausgeführt wurde, gibt er den Amazon-Ressourcennamen (ARN) der Richtlinie zurück.

```
{
  "PolicyARN": "arn:aws:autoscaling:region:account-id:scalingPolicy:2f4f5048-
d8a8-4d14-b13a-d1905620f345:autoScalingGroupName/my-asg:policyName/my-predictive-
scaling-policy",
  "Alarms": []
}
```

9. Wiederholen Sie diese Schritte für jede Lastmetrikspezifikation, die Sie zu einer auf Amazon EC2 Auto Scaling basierenden Predictive Scaling-Richtlinie migrieren.

## Schritt 3: Überprüfen Sie die Prognosen, die die Richtlinien für vorausschauende Skalierung generieren

Wenn Sie Predictive Scaling nicht verwenden, überspringen Sie das folgende Verfahren.

Kurz nachdem Sie eine Richtlinie für vorausschauende Skalierung erstellt haben, ist eine Prognose verfügbar. Nachdem Amazon EC2 Auto Scaling die Prognose generiert hat, können Sie die Prognose für die Richtlinie über die Amazon EC2 Auto Scaling Scaling-Konsole überprüfen und bei Bedarf anpassen.

Um die Prognose für eine prädiktive Skalierungsrichtlinie zu überprüfen

1. Öffnen Sie die Amazon-EC2-Konsole unter https://console.aws.amazon.com/ec2/.
2. Wählen Sie im Navigationsbereich Auto Scaling Scaling-Gruppen und dann den Namen Ihrer Auto Scaling Scaling-Gruppe aus der Liste aus.
3. Wählen Sie auf der Registerkarte Automatische Skalierung unter Richtlinien für vorausschauende Skalierung Ihre Richtlinie aus.

4. Im Abschnitt Überwachung können Sie die vergangenen und zukünftigen Prognosen Ihrer Richtlinie für Last und Kapazität im Vergleich zu tatsächlichen Werten anzeigen.

   Weitere Informationen finden Sie unter Übersichtsdiagramme zur vorausschauenden Skalierung im Amazon EC2 Auto Scaling Scaling-Benutzerhandbuch.
5. Wiederholen Sie diese Schritte für jede Predictive Scaling-Richtlinie, die Sie erstellt haben.

## Schritt 4: Bereiten Sie das Löschen des Skalierungsplans vor

Gehen Sie für alle Ressourcen mit einer vorhandenen Skalierungskonfiguration für die Zielverfolgung wie folgt vor, um alle zusätzlichen Informationen zu sammeln, die Sie aus dem Skalierungsplan benötigen, bevor Sie ihn löschen.

Verwenden Sie den describe-scaling-plan-resourcesBefehl, um die Informationen zur Skalierungsrichtlinie aus dem Skalierungsplan zu beschreiben. Ersetzen Sie den Befehl im folgenden Beispiel *my-scaling-plan* durch Ihre eigenen Informationen.

```
aws autoscaling-plans describe-scaling-plan-resources \
  --scaling-plan-name my-scaling-plan \
  --scaling-plan-version 1
```

Überprüfen Sie die Ausgabe und bestätigen Sie, dass Sie die beschriebenen Skalierungsrichtlinien migrieren möchten. Verwenden Sie diese Informationen, um neue Amazon EC2 Auto Scaling- und Application Auto Scaling-basierte Skalierungsrichtlinien für die Zielverfolgung in zu erstellen. Schritt 6: Reaktivieren Sie die dynamische Skalierung

## Schritt 5: Löschen Sie den Skalierungsplan

Bevor Sie neue Skalierungsrichtlinien für die Zielverfolgung erstellen, müssen Sie den Skalierungsplan löschen, um die von ihm erstellten Skalierungsrichtlinien zu löschen.

Verwenden Sie den delete-scaling-planBefehl, um Ihren Skalierungsplan zu löschen. Ersetzen Sie den Befehl im folgenden Beispiel *my-scaling-plan* durch Ihre eigenen Informationen.

```
aws autoscaling-plans delete-scaling-plan \
  --scaling-plan-name my-scaling-plan \
  --scaling-plan-version 1
```

Nachdem Sie den Skalierungsplan gelöscht haben, ist die dynamische Skalierung deaktiviert. Wenn es also zu einem plötzlichen Anstieg des Datenverkehrs oder der Arbeitslast kommt, erhöht sich die für jede skalierbare Ressource verfügbare Kapazität nicht von alleine. Als Vorsichtsmaßnahme sollten Sie die Kapazität Ihrer skalierbaren Ressourcen kurzfristig manuell erhöhen.

Um die Kapazität einer Auto Scaling Scaling-Gruppe zu erhöhen

1. Öffnen Sie die Amazon-EC2-Konsole unter https://console.aws.amazon.com/ec2/.
2. Wählen Sie im Navigationsbereich Auto Scaling Scaling-Gruppen und dann den Namen Ihrer Auto Scaling Scaling-Gruppe aus der Liste aus.
3. Wählen Sie auf der Registerkarte Details die Option Gruppendetails, Bearbeiten.
4. Erhöhen Sie für Gewünschte Kapazität die gewünschte Kapazität.
5. Wenn Sie fertig sind, wählen Sie Aktualisieren.

So fügen Sie eine Aurora Replica zu einem DB-Cluster hinzu:

1. Öffnen Sie die Amazon RDS-Konsole unter https://console.aws.amazon.com/rds/.
2. Wählen Sie im Navigationsbereich Datenbanken und dann Ihren DB-Cluster aus.
3. Stellen Sie sicher, dass sowohl der Cluster als auch die primäre Instance den Status Verfügbar aufweisen.
4. Wählen Sie Aktionen, Leser hinzufügen.
5. Geben Sie auf der Seite Leser hinzufügen Optionen für Ihr neues Aurora-Replikat an.
6. Wählen Sie „Leser hinzufügen“.

Um die bereitgestellte Lese- und Schreibkapazität einer DynamoDB-Tabelle oder eines globalen sekundären Indexes zu erhöhen

1. Öffnen Sie die DynamoDB-Konsole unter. https://console.aws.amazon.com/dynamodb/
2. Wählen Sie im Navigationsbereich Tabellen und dann den Namen Ihrer Tabelle aus der Liste aus.
3. Wählen Sie auf der Registerkarte Zusätzliche Einstellungen die Option Lese-/Schreibkapazität und Bearbeiten aus.
4. Erhöhen Sie auf der Seite  read/write Kapazität bearbeiten für Lesekapazität und Bereitgestellte Kapazitätseinheiten die bereitgestellte Lesekapazität der Tabelle.

5. (Optional) Wenn Sie möchten, dass Ihre globalen sekundären Indizes dieselben Lesekapazitätseinstellungen wie die Basistabelle verwenden, aktivieren Sie das Kontrollkästchen Dieselben Lesekapazitätseinstellungen für alle globalen sekundären Indizes verwenden.
6. Erhöhen Sie für Schreibkapazität unter Bereitgestellte Kapazitätseinheiten die bereitgestellte Schreibkapazität der Tabelle.
7. (Optional) Wenn Sie möchten, dass Ihre globalen sekundären Indizes dieselben Schreibkapazitätseinstellungen wie die Basistabelle verwenden, aktivieren Sie das Kontrollkästchen Dieselben Schreibkapazitätseinstellungen für alle globalen sekundären Indizes verwenden.
8. Wenn Sie die Kontrollkästchen in den Schritten 5 oder 7 nicht aktiviert haben, blättern Sie auf der Seite nach unten, um die Lese- und Schreibkapazität aller globalen sekundären Indizes zu aktualisieren.
9. Wählen Sie Änderungen speichern, um fortzufahren.

Um die Anzahl der laufenden Aufgaben für Ihren Amazon ECS-Service zu erhöhen

1. Öffnen Sie die Konsole auf https://console.aws.amazon.com/ecs/Version 2.
2. Wählen Sie im Navigationsbereich Clusters und dann den Namen Ihres Clusters aus der Liste aus.
3. Aktivieren Sie im Abschnitt Dienste das Kontrollkästchen neben dem Dienst und wählen Sie dann Aktualisieren aus.
4. Geben Sie unter Gewünschte Aufgaben die Anzahl der Aufgaben ein, die Sie für den Dienst ausführen möchten.
5. Wählen Sie Aktualisieren aus.

Um die Kapazität einer Spot-Flotte zu erhöhen

1. Öffnen Sie die Amazon-EC2-Konsole unter https://console.aws.amazon.com/ec2/.
2. Wählen Sie im Navigationsbereich Spot-Anfragen und dann Ihre Spot-Flotte-Anfrage aus.
3. Wählen Sie Actions (Aktionen)und dann Modify target capacity (Zielkapazität bearbeiten) aus.
4. Geben Sie unter Zielkapazität ändern die neue Zielkapazität und den Teil der On-Demand-Instance ein.
5. Wählen Sie Absenden aus.

# Schritt 6: Reaktivieren Sie die dynamische Skalierung

Reaktivieren Sie die dynamische Skalierung, indem Sie Skalierungsrichtlinien für die Zielverfolgung erstellen.

Wenn Sie eine Skalierungsrichtlinie für die Zielverfolgung für eine Auto Scaling Scaling-Gruppe erstellen, fügen Sie sie direkt der Gruppe hinzu. Wenn Sie eine Skalierungsrichtlinie für die Zielverfolgung für andere skalierbare Ressourcen erstellen, registrieren Sie die Ressource zunächst als skalierbares Ziel und fügen dann dem skalierbaren Ziel eine Skalierungsrichtlinie für die Zielverfolgung hinzu.

Themen

- Skalierungsrichtlinien für die Zielverfolgung für Auto Scaling Scaling-Gruppen erstellen
- Erstellen Sie Skalierungsrichtlinien für die Zielverfolgung für andere skalierbare Ressourcen

## Skalierungsrichtlinien für die Zielverfolgung für Auto Scaling Scaling-Gruppen erstellen

So erstellen Sie Skalierungsrichtlinien für die Zielverfolgung für Auto Scaling Scaling-Gruppen

1. Erstellen Sie in einer JSON-Datei eine `PredefinedMetricSpecification` oder `CustomizedMetricSpecification` verwenden Sie die entsprechenden Einstellungen aus dem Skalierungsplan.

   Im Folgenden finden Sie Beispiele für eine Ziel-Tracking-Konfiguration. Ersetzen Sie in diesen Beispielen jedes Beispiel *`user input placeholder`* durch Ihre eigenen Informationen.

   With predefined metrics

```
{
  "TargetValue": 50.0,
  "PredefinedMetricSpecification":
    {
      "PredefinedMetricType": "ASGAverageCPUUtilization"
    }
}
```

Weitere Informationen finden Sie PredefinedMetricSpecificationin der Amazon EC2 Auto Scaling API-Referenz.

With custom metrics

```
{
  "TargetValue": 100.0,
  "CustomizedMetricSpecification": {
    "MetricName": "MyBacklogPerInstance",
    "Namespace": "MyNamespace",
    "Dimensions": [{
      "Name": "MyOptionalMetricDimensionName",
      "Value": "MyOptionalMetricDimensionValue"
    }],
    "Statistic": "Average",
    "Unit": "None"
  }
}
```

Weitere Informationen finden Sie CustomizedMetricSpecificationin der Amazon EC2 Auto Scaling API-Referenz.

2. Um Ihre Skalierungsrichtlinie zu erstellen, verwenden Sie den put-scaling-policyBefehl zusammen mit der JSON-Datei, die Sie im vorherigen Schritt erstellt haben. Ersetzen Sie im folgenden Beispiel jede *user input placeholder* durch Ihre eigenen Informationen.

```
aws autoscaling put-scaling-policy --policy-name my-target-tracking-scaling-policy
 \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
```

3. Wiederholen Sie diesen Vorgang für jede Skalierungsplan-basierte Skalierungsrichtlinie, die Sie zu einer auf Amazon EC2 Auto Scaling basierenden Skalierungsrichtlinie für die Zielverfolgung migrieren.

## Erstellen Sie Skalierungsrichtlinien für die Zielverfolgung für andere skalierbare Ressourcen

Erstellen Sie als Nächstes Skalierungsrichtlinien für die Zielverfolgung für andere skalierbare Ressourcen, indem Sie die folgenden Konfigurationsaufgaben ausführen.

- Registrieren Sie ein skalierbares Ziel für Auto Scaling beim Application Auto Scaling Scaling-Dienst.
- Fügen Sie dem skalierbaren Ziel eine Skalierungsrichtlinie für die Ziel-Nachverfolgung hinzu.

Um Skalierungsrichtlinien für die Zielverfolgung für andere skalierbare Ressourcen zu erstellen

1. Verwenden Sie den register-scalable-targetBefehl, um die Ressource als skalierbares Ziel zu registrieren und die Skalierungsgrenzen für die Skalierungsrichtlinie zu definieren.

   Ersetzen Sie im folgenden Beispiel jede Information *user input placeholder* durch Ihre eigenen Informationen. Geben Sie für die Befehlsoptionen die folgenden Informationen an:

   - `--service-namespace`— Ein Namespace für den Zieldienst (zum Beispiel**ecs**). Informationen zum Abrufen von Dienst-Namespaces finden Sie in der Referenz. RegisterScalableTarget
   - `--scalable-dimension`— Eine skalierbare Dimension, die der Zielressource zugeordnet ist (z. B.). **ecs:service:DesiredCount** Informationen zu skalierbaren Dimensionen finden Sie in der RegisterScalableTargetReferenz.
   - `--resource-id`— Eine Ressourcen-ID für die Zielressource (zum Beispiel**service/*my-cluster*/*my-service***). Informationen zur Syntax und Beispiele für bestimmte Ressourcen IDs finden Sie in der RegisterScalableTargetReferenz.

   ```
   aws application-autoscaling register-scalable-target --service-namespace namespace
    \
     --scalable-dimension dimension \
     --resource-id identifier \
     --min-capacity 1 --max-capacity 10
   ```

   Bei Erfolg gibt dieser Befehl den ARN des skalierbaren Ziels zurück.

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

2. Erstellen Sie in einer JSON-Datei eine `PredefinedMetricSpecification` oder `CustomizedMetricSpecification` mit den entsprechenden Einstellungen aus dem Skalierungsplan.

   Im Folgenden finden Sie Beispiele für eine Ziel-Tracking-Konfiguration.

   With predefined metrics

   ```
   {
     "TargetValue": 70.0,
     "PredefinedMetricSpecification":
       {
         "PredefinedMetricType": "ECSServiceAverageCPUUtilization"
       }
   }
   ```

   Weitere Informationen finden Sie PredefinedMetricSpecificationin der API-Referenz für Application Auto Scaling.

   With custom metrics

   ```
   {
     "TargetValue": 70.0,
     "CustomizedMetricSpecification": {
       "MetricName": "MyUtilizationMetric",
       "Namespace": "MyNamespace",
       "Dimensions": [{
         "Name": "MyOptionalMetricDimensionName",
         "Value": "MyOptionalMetricDimensionValue"
       }],
       "Statistic": "Average",
       "Unit": "Percent"
     }
   }
   ```

   Weitere Informationen finden Sie CustomizedMetricSpecificationin der API-Referenz für Application Auto Scaling.

3. Verwenden Sie den put-scaling-policyBefehl zusammen mit der JSON-Datei, die Sie im vorherigen Schritt erstellt haben, um Ihre Skalierungsrichtlinie zu erstellen.

   ```
   aws application-autoscaling put-scaling-policy --service-namespace namespace \
   ```

```
  --scalable-dimension dimension \
  --resource-id identifier \
  --policy-name my-target-tracking-scaling-policy --policy-
type TargetTrackingScaling \
  --target-tracking-scaling-policy-configuration file://config.json
```

4. Wiederholen Sie diesen Vorgang für jede Skalierungsplan-basierte Skalierungsrichtlinie, die Sie zu einer auf Application Auto Scaling basierenden Skalierungsrichtlinie für die Zielverfolgung migrieren.

## Schritt 7: Reaktivieren Sie die prädiktive Skalierung

Wenn Sie Predictive Scaling nicht verwenden, überspringen Sie diesen Schritt.

Reaktivieren Sie die prädiktive Skalierung, indem Sie die prädiktive Skalierung auf Prognose und Skalierung umstellen.

Um diese Änderung vorzunehmen, aktualisieren Sie die JSON-Dateien, die Sie in erstellt haben, Schritt 2: Erstellen Sie Richtlinien für die prädiktive Skalierung und ändern Sie den Wert der `Mode` Option `ForecastAndScale` wie im folgenden Beispiel auf:

```
"Mode":"ForecastAndScale"
```

Aktualisieren Sie anschließend jede Richtlinie für prädiktive Skalierung mit dem put-scaling-policyBefehl. Ersetzen Sie in diesem Beispiel jede *`user input placeholder`* durch Ihre eigenen Informationen.

```
aws autoscaling put-scaling-policy --policy-name my-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://my-predictive-scaling-config.json
```

Alternativ können Sie diese Änderung von der Amazon EC2 Auto Scaling-Konsole aus vornehmen, indem Sie die Einstellung Skalierung auf Prognosebasis aktivieren. Weitere Informationen finden Sie unter Prädiktive Skalierung von Cooldowns für Amazon EC2 Auto Scaling im Amazon EC2 Auto Scaling-Benutzerhandbuch.

# Amazon EC2 Auto Scaling Scaling-Referenz für die Migration von Skalierungsrichtlinien für die Zielverfolgung

Zu Referenzzwecken sind in der folgenden Tabelle alle Eigenschaften der Ziel-Tracking-Konfiguration im Skalierungsplan mit ihren entsprechenden Eigenschaften im Amazon EC2 Auto Scaling `PutScalingPolicy` Scaling-API-Vorgang aufgeführt.

| Quelleigenschaft des Skalierungsplans | Zieleigenschaft von Amazon EC2 Auto Scaling |
|---|---|
| `PolicyName` | `PolicyName` |
| `PolicyType` | `PolicyType` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Name` | `TargetTrackingConfiguration.CustomizedMetricSpecification.Dimensions.Name` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Value` | `TargetTrackingConfiguration.CustomizedMetricSpecification.Dimensions.Value` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.MetricName` | `TargetTrackingConfiguration.CustomizedMetricSpecification.MetricName` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Namespace` | `TargetTrackingConfiguration.CustomizedMetricSpecification.Namespace` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Statistic` | `TargetTrackingConfiguration.CustomizedMetricSpecification.Statistic` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Unit` | `TargetTrackingConfiguration.CustomizedMetricSpecification.Unit` |

| Quelleigenschaft des Skalierungsplans | Zieleigenschaft von Amazon EC2 Auto Scaling |
| --- | --- |
| `TargetTrackingConfiguration.DisableScaleIn` | `TargetTrackingConfiguration.DisableScaleIn` |
| `TargetTrackingConfiguration.EstimatedInstanceWarmup` | `TargetTrackingConfiguration.EstimatedInstanceWarmup` [1] |
| `TargetTrackingConfiguration.PredefinedScalingMetricSpecification.PredefinedScalingMetricType` | `TargetTrackingConfiguration.PredefinedMetricSpecification.PredefinedMetricType` |
| `TargetTrackingConfiguration.PredefinedScalingMetricSpecification.ResourceLabel` | `TargetTrackingConfiguration.PredefinedMetricSpecification.ResourceLabel` |
| `TargetTrackingConfiguration.ScaleInCooldown` | `Not available` |
| `TargetTrackingConfiguration.ScaleOutCooldown` | `Not available` |
| `TargetTrackingConfiguration.TargetValue` | `TargetTrackingConfiguration.TargetValue` |

[1] Instance Warmup ist eine Funktion für Auto Scaling Scaling-Gruppen, die sicherstellt, dass neu gestartete Instances bereit sind, Traffic zu empfangen, bevor sie ihre Nutzungsdaten zur Skalierungsmetrik beitragen. Während sich die Instances noch in der Aufwärmphase befinden, verlangsamt Amazon EC2 Auto Scaling den Prozess des Hinzufügens oder Entfernens von Instances zur Gruppe. Anstatt eine Aufwärmzeit für eine Skalierungsrichtlinie anzugeben, empfehlen wir, die Standard-Instance-Aufwärmeinstellung Ihrer Auto Scaling Scaling-Gruppe zu verwenden, um sicherzustellen, dass alle Instance-Starts dieselbe Instance-Aufwärmzeit verwenden. Weitere Informationen finden Sie unter Festlegen des Standard-Instance-Warmup für eine Auto-Scaling-Gruppe im Amazon EC2 Auto Scaling-Benutzerhandbuch.

# Referenz für Application Auto Scaling zur Migration von Skalierungsrichtlinien für die Zielverfolgung

Zu Referenzzwecken sind in der folgenden Tabelle alle Konfigurationseigenschaften für die Zielverfolgung im Skalierungsplan mit ihren entsprechenden Eigenschaften im Application Auto Scaling `PutScalingPolicy` Scaling-API-Vorgang aufgeführt.

| Quelleigenschaft des Skalierungsplans | Zieleigenschaft von Application Auto Scaling |
|---|---|
| `PolicyName` | `PolicyName` |
| `PolicyType` | `PolicyType` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Name` | `TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Dimensions.Name` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Value` | `TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Dimensions.Value` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.MetricName` | `TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.MetricName` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Namespace` | `TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Namespace` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Statistic` | `TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Statistic` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Unit` | `TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Unit` |

| Quelleigenschaft des Skalierungsplans | Zieleigenschaft von Application Auto Scaling |
|---|---|
| `TargetTrackingConfiguration.DisableScaleIn` | `TargetTrackingScalingPolicyConfiguration.DisableScaleIn` |
| `TargetTrackingConfiguration.EstimatedInstanceWarmup` | `Not available` |
| `TargetTrackingConfiguration.PredefinedScalingMetricSpecification.PredefinedScalingMetricType` | `TargetTrackingScalingPolicyConfiguration.PredefinedMetricSpecification.PredefinedMetricType` |
| `TargetTrackingConfiguration.PredefinedScalingMetricSpecification.ResourceLabel` | `TargetTrackingScalingPolicyConfiguration.PredefinedMetricSpecification.ResourceLabel` |
| `TargetTrackingConfiguration.ScaleInCooldown` [1] | `TargetTrackingScalingPolicyConfiguration.ScaleInCooldown` |
| `TargetTrackingConfiguration.ScaleOutCooldown` [1] | `TargetTrackingScalingPolicyConfiguration.ScaleOutCooldown` |
| `TargetTrackingConfiguration.TargetValue` | `TargetTrackingScalingPolicyConfiguration.TargetValue` |

[1] Application Auto Scaling verwendet Abklingzeiten, um die Skalierung zu verlangsamen, wenn Ihre skalierbare Ressource horizontal skaliert (Kapazität erhöht) und hochskaliert (Kapazität reduziert). Weitere Informationen finden Sie unter Definieren von Ruhephasen im Benutzerhandbuch zum Auto Scaling von Anwendungen.

## Zusätzliche Informationen

Informationen zum Erstellen neuer Richtlinien für die vorausschauende Skalierung von der Konsole aus finden Sie im folgenden Thema:

- Amazon EC2 Auto Scaling — Erstellen Sie eine Richtlinie zur vorausschauenden Skalierung im Amazon EC2 Auto Scaling Scaling-Benutzerhandbuch.

In den folgenden Themen erfahren Sie, wie Sie mithilfe der Konsole neue Skalierungsrichtlinien für die Zielverfolgung erstellen:

- Amazon Aurora — Verwendung von Amazon Aurora Auto Scaling mit Aurora Replicas im Amazon RDS-Benutzerhandbuch.
- DynamoDB — Verwenden der auto Skalierung AWS-Managementkonsole mit DynamoDB im Amazon DynamoDB DynamoDB-Entwicklerhandbuch.
- Amazon EC2 Auto Scaling — Erstellen Sie eine Skalierungsrichtlinie für die Zielverfolgung im Amazon EC2 Auto Scaling Scaling-Benutzerhandbuch.
- Amazon ECS — Aktualisierung eines Service mithilfe der Konsole im Amazon Elastic Container Service Developer Guide.
- Spot-Flotte — Skalieren Sie Spot-Flotte mithilfe einer Zielverfolgungsrichtlinie im Amazon EC2 EC2-Benutzerhandbuch.

# Skalierungsplansicherheit

Cloud-Sicherheit AWS hat höchste Priorität. Als AWS Kunde profitieren Sie von einer Rechenzentrums- und Netzwerkarchitektur, die darauf ausgelegt sind, die Anforderungen der sicherheitssensibelsten Unternehmen zu erfüllen.

Sicherheit ist eine gemeinsame Verantwortung von Ihnen AWS und Ihnen. Das Modell der geteilten Verantwortung beschreibt dies als Sicherheit der Cloud und Sicherheit in der Cloud:

- Sicherheit der Cloud — AWS ist verantwortlich für den Schutz der Infrastruktur, die AWS Dienste in der AWS Cloud ausführt. AWS bietet Ihnen auch Dienste, die Sie sicher nutzen können. Externe Prüfer testen und verifizieren regelmäßig die Wirksamkeit unserer Sicherheitsmaßnahmen im Rahmen der AWS und . Weitere Informationen zu den Compliance-Programmen, die für gelten AWS Auto Scaling, finden Sie unter AWS Leistungen im Umfang nach Compliance-Programmen AWS .
- Sicherheit in der Cloud — Ihre Verantwortung richtet sich nach dem AWS Dienst, den Sie nutzen. Sie sind auch für andere Faktoren verantwortlich, einschließlich der Vertraulichkeit Ihrer Daten, für die Anforderungen Ihres Unternehmens und für die geltenden Gesetze und Vorschriften.

Diese Dokumentation hilft Ihnen, zu verstehen, wie das Modell der geteilten Verantwortung bei der Verwendung von Skalierungsplänen zum Tragen kommt, und hilft Ihnen, zu verstehen, wie der Zugriff auf Skalierungspläne verwaltet wird.

Topics



# Zugriff auf Skalierungspläne mithilfe von Schnittstellen-VPC-Endpunkten

Sie können verwenden AWS PrivateLink , um eine private Verbindung zwischen Ihrer VPC und AWS Auto Scaling herzustellen. Sie können darauf zugreifen, AWS Auto Scaling als ob es in Ihrer VPC

wäre, ohne ein Internet-Gateway, ein NAT-Gerät, eine VPN-Verbindung oder Direct Connect eine Verbindung zu verwenden. Instances in Ihrer VPC benötigen für den Zugriff AWS Auto Scaling keine öffentlichen IP-Adressen.

Sie stellen diese private Verbindung her, indem Sie einen Schnittstellen-Endpunkt erstellen, der von AWS PrivateLink unterstützt wird. Wir erstellen eine Endpunkt-Netzwerkschnittstelle in jedem Subnetz, das Sie für den Schnittstellen-Endpunkt aktivieren. Hierbei handelt es sich um vom Anforderer verwaltete Netzwerkschnittstellen, die als Eingangspunkt für den Datenverkehr dienen, der für AWS Auto Scaling bestimmt ist.

Weitere Informationen finden Sie AWS PrivateLink im AWS PrivateLink Leitfaden unter Zugriff AWS-Services durch.

Themen

- Erstellen eines Schnittstellen-VPC-Endpunkts für Skalierungspläne
- Erstellen einer VPC-Endpunktrichtlinie für Skalierungspläne
- Endpunkt-Migration

## Erstellen eines Schnittstellen-VPC-Endpunkts für Skalierungspläne

Erstellen Sie einen Endpunkt für AWS Auto Scaling Skalierungspläne mit dem folgenden Dienstnamen:

```
com.amazonaws.region.autoscaling-plans
```

Weitere Informationen finden Sie im AWS PrivateLink Handbuch unter Zugreifen auf einen AWS Dienst über einen Schnittstellen-VPC-Endpunkt.

Sie müssen keine anderen Einstellungen ändern. AWS Auto Scaling API-Aufrufe, die entweder Dienstendpunkte oder VPC-Endpunkte mit privater Schnittstelle AWS-Services verwenden, je nachdem, welche verwendet werden.

## Erstellen einer VPC-Endpunktrichtlinie für Skalierungspläne

Sie können Ihrem VPC-Endpunkt eine Richtlinie hinzufügen, um den Zugriff auf die AWS Auto Scaling API zu kontrollieren. Die Richtlinie legt Folgendes fest:

- Prinzipal, der die Aktionen ausführen kann.

- Die Aktionen, die ausgeführt werden können.
- Die Ressource, auf der die Aktionen ausgeführt werden können.

Das folgende Beispiel zeigt eine VPC-Endpunktrichtlinie, die jedem die Berechtigung zum Löschen eines Skalierungsplans über den Endpunkt verweigert. Die Beispielrichtlinie gewährt auch jedem die Berechtigung, alle anderen Aktionen auszuführen.

```
{
   "Statement": [
        {
            "Action": "*",
            "Effect": "Allow",
            "Resource": "*",
            "Principal": "*"
        },
        {
            "Action": "autoscaling-plans:DeleteScalingPlan",
            "Effect": "Deny",
            "Resource": "*",
            "Principal": "*"
        }
    ]
}
```

Weitere Informationen finden Sie unter VPC-Endpunkt-Richtlinien im AWS PrivateLink -Leitfaden.

## Endpunkt-Migration

Am 22. November 2019 haben wir den neuen Standard-DNS-Hostnamen und -Endpunkt für API-Aufrufe eingeführt`autoscaling-plans.region.amazonaws.com`. AWS Auto Scaling Der neue Endpunkt ist mit der neuesten Version von AWS CLI und SDKs kompatibel. Falls Sie dies noch nicht getan haben, installieren Sie die neueste Version AWS CLI und verwenden SDKs Sie den neuen Endpunkt. Informationen zur AWS CLI Aktualisierung von finden Sie unter Installation oder Aktualisierung von AWS CLI im AWS Command Line Interface Benutzerhandbuch. Informationen dazu finden Sie AWS SDKs unter Tools für Amazon Web Services.

> **Important**
>
> Aus Gründen der Abwärtskompatibilität wird der bestehende `autoscaling.region.amazonaws.com` Endpunkt weiterhin für AWS Auto Scaling

API-Aufrufe unterstützt. Um den `autoscaling.region.amazonaws.com`-Endpunkt als privaten Interface-VPC-Endpunkt einzurichten, siehe Amazon EC2 Auto Scaling und Interface-VPC-Endpunkte im Amazon EC2 Auto Scaling User Guide.

Aufzurufender Endpunkt bei Verwendung der CLI oder der AWS Auto Scaling API

In der aktuellen Version von AWS Auto Scaling gehen Ihre AWS Auto Scaling API-Aufrufe automatisch an den `autoscaling-plans.region.amazonaws.com` Endpunkt statt `anautoscaling.region.amazonaws.com`.

Sie können den neuen Endpunkt in der CLI aufrufen, indem Sie mit jedem Befehl den folgenden Parameter verwenden, um den Endpunkt anzugeben: `--endpoint-url https://autoscaling-plans.region.amazonaws.com`.

Obwohl es nicht empfohlen wird, können Sie den alten Endpunkt in der CLI auch aufrufen, indem Sie mit jedem Befehl den folgenden Parameter verwenden, um den Endpunkt anzugeben: `--endpoint-url https://autoscaling.region.amazonaws.com`.

Informationen zu den verschiedenen, die zum Aufrufen von SDKs verwendet werden APIs, finden Sie in der Dokumentation zum entsprechenden SDK. Dort erfahren Sie, wie Sie die Anfragen an einen bestimmten Endpunkt weiterleiten. Weitere Informationen finden Sie unter Tools für Amazon Web Services.

## Datenschutz für Skalierungspläne

Das Modell der AWS gemeinsamen Verantwortung und geteilter Verantwortung gilt für den Datenschutz in AWS Auto Scaling. Wie in diesem Modell beschrieben, AWS ist verantwortlich für den Schutz der globalen Infrastruktur, auf der alle Systeme laufen AWS Cloud. Sie sind dafür verantwortlich, die Kontrolle über Ihre in dieser Infrastruktur gehosteten Inhalte zu behalten. Sie sind auch für die Sicherheitskonfiguration und die Verwaltungsaufgaben für die von Ihnen verwendeten AWS-Services verantwortlich. Weitere Informationen zum Datenschutz finden Sie unter Häufig gestellte Fragen zum Datenschutz. Informationen zum Datenschutz in Europa finden Sie im Blog-Beitrag AWS -Modell der geteilten Verantwortung und in der DSGVO im AWS -Sicherheitsblog.

Aus Datenschutzgründen empfehlen wir, dass Sie AWS-Konto Anmeldeinformationen schützen und einzelne Benutzer mit AWS IAM Identity Center oder AWS Identity and Access Management (IAM) einrichten. So erhält jeder Benutzer nur die Berechtigungen, die zum Durchführen seiner Aufgaben erforderlich sind. Außerdem empfehlen wir, die Daten mit folgenden Methoden schützen:

- Verwenden Sie für jedes Konto die Multi-Faktor-Authentifizierung (MFA).
- Wird verwendet SSL/TLS , um mit AWS Ressourcen zu kommunizieren. Wir benötigen TLS 1.2 und empfehlen TLS 1.3.
- Richten Sie die API und die Protokollierung von Benutzeraktivitäten mit ein AWS CloudTrail. Informationen zur Verwendung von CloudTrail Pfaden zur Erfassung von AWS Aktivitäten finden Sie unter Arbeiten mit CloudTrail Pfaden im AWS CloudTrail Benutzerhandbuch.
- Verwenden Sie AWS Verschlüsselungslösungen zusammen mit allen darin enthaltenen Standardsicherheitskontrollen AWS-Services.
- Verwenden Sie erweiterte verwaltete Sicherheitsservices wie Amazon Macie, die dabei helfen, in Amazon S3 gespeicherte persönliche Daten zu erkennen und zu schützen.
- Wenn Sie für den Zugriff AWS über eine Befehlszeilenschnittstelle oder eine API FIPS 140-3-validierte kryptografische Module benötigen, verwenden Sie einen FIPS-Endpunkt. Weitere Informationen über verfügbare FIPS-Endpunkte finden Sie unter Federal Information Processing Standard (FIPS) 140-3.

Wir empfehlen dringend, in Freitextfeldern, z. B. im Feld Name, keine vertraulichen oder sensiblen Informationen wie die E-Mail-Adressen Ihrer Kunden einzugeben. Dies gilt auch, wenn Sie mit der Konsole, der AWS Auto Scaling API oder auf andere AWS-Services Weise arbeiten oder diese verwenden. AWS CLI AWS SDKs Alle Daten, die Sie in Tags oder Freitextfelder eingeben, die für Namen verwendet werden, können für Abrechnungs- oder Diagnoseprotokolle verwendet werden. Wenn Sie eine URL für einen externen Server bereitstellen, empfehlen wir dringend, keine Anmeldeinformationen zur Validierung Ihrer Anforderung an den betreffenden Server in die URL einzuschließen.

# Identitäts- und Zugriffsmanagement für Skalierungspläne

AWS Identity and Access Management (IAM) hilft einem Administrator AWS-Service , den Zugriff auf AWS Ressourcen sicher zu kontrollieren. IAM-Administratoren kontrollieren, wer authentifiziert (angemeldet) und autorisiert werden kann (über Berechtigungen verfügt), um Ressourcen zu verwenden. AWS Auto Scaling IAM ist ein Programm AWS-Service , das Sie ohne zusätzliche Kosten nutzen können.

Eine umfassende IAM-Dokumentation finden Sie im IAM User Guide.

# Zugriffskontrolle

Auch wenn Sie über gültige Anmeldeinformationen zur Authentifizierung Ihrer Anfragen verfügen, können Sie die Skalierungspläne nur mit entsprechenden Berechtigungen erstellen oder darauf zugreifen. Beispielsweise müssen Sie über Berechtigungen zum Erstellen von Skalierungsplänen, zum Konfigurieren der prädiktiven Skalierung usw. verfügen.

Dieses Thema enthält Informationen dazu, wie ein IAM-Administrator Ihre Skalierungspläne mithilfe von IAM sichern kann, indem er steuert, wer Aktionen durchführen darf.

Themen



## Funktionsweise von Skalierungsplänen mit IAM

Bevor Sie IAM verwenden, um zu verwalten, wer AWS Auto Scaling Skalierungspläne erstellen, darauf zugreifen und sie verwalten kann, sollten Sie sich darüber im Klaren sein, welche IAM-Funktionen für Skalierungspläne verfügbar sind.

Themen



### Identitätsbasierte Richtlinien

Mit identitätsbasierten IAM-Richtlinien können Sie angeben, welche Aktionen und Ressourcen erteilt oder abgelehnt werden. Darüber hinaus können Sie die Bedingungen festlegen, unter denen Aktionen zugelassen oder abgelehnt werden. Skalierungspläne unterstützten bestimmte Aktionen, Ressourcen und Bedingungsschlüssel. Informationen zu sämtlichen Elementen, die Sie in einer JSON-Richtlinie verwenden, finden Sie in der IAM-Referenz für JSON-Richtlinienelemente im IAM-Benutzerhandbuch.

Aktionen

Administratoren können mithilfe von AWS JSON-Richtlinien angeben, wer Zugriff auf was hat. Das heißt, welcher Prinzipal Aktionen für welche Ressourcen und unter welchen Bedingungen ausführen kann.

Das Element `Action` einer JSON-Richtlinie beschreibt die Aktionen, mit denen Sie den Zugriff in einer Richtlinie zulassen oder verweigern können. Nehmen Sie Aktionen in eine Richtlinie auf, um Berechtigungen zur Ausführung des zugehörigen Vorgangs zu erteilen.

Skalierungsplanaktionen in IAM-Richtlinienanweisungen verwenden das folgende Präfix vor der Aktion: `autoscaling-plans:`. Richtlinienanweisungen müssen entweder ein – `Action`oder ein `NotAction`-Element enthalten. Skalierungspläne verfügen über eigene Gruppen von Aktionen, die Aufgaben beschreiben, die Sie mit diesem Service durchführen können.

Um mehrere Aktionen in einer einzelnen Anweisung anzugeben, trennen Sie sie durch Beistriche, wie im folgenden Beispiel gezeigt.

```
"Action": [
      "autoscaling-plans:DescribeScalingPlans",
      "autoscaling-plans:DescribeScalingPlanResources"
```

Sie können auch Platzhalter (*) verwenden, um mehrere Aktionen anzugeben. Beispielsweise können Sie alle Aktionen festlegen, die mit dem Wort `Describe` beginnen, einschließlich der folgenden Aktion:

```
"Action": "autoscaling-plans:Describe*"
```

Eine vollständige Liste mit den Skalierungsplan-Aktionen, die in Richtlinienerklärungen verwendet werden können, finden Sie unter Aktionen, Ressourcen und Bedingungsschlüssel für AWS Auto Scaling in der Service-Autorisierungs-Referenz.

Ressourcen

Das Element `Resource` gibt die Objekte an, auf die die Aktion angewendet wird.

Skalierungspläne besitzten keine servicedefinierten Ressourcen, die als Element `Resource` einer IAM-Richtlinienanweisung verwendet werden können. Daher gibt es keine Amazon-Ressourcennamen (ARNs), die Sie in einer IAM-Richtlinie verwenden könnten. Um den Zugriff auf Skalierungsplan-Aktionen zu steuern, müssen Sie als Ressource immer ein „*“ (Sternchen) verwenden, wenn Sie eine IAM-Richtlinie schreiben.

Bedingungsschlüssel

Mithilfe des Elements `Condition` (oder des Blocks `Condition`) können Sie die Bedingungen angeben, unter denen eine Anweisung wirksam ist. Beispielsweise kann festgelegt werden, dass eine Richtlinie erst ab einem bestimmten Datum gilt. Bedingungen werden mithilfe vordefinierter Bedingungsschlüssel formuliert.

Skalierungspläne stellen keine servicespezifischen Bedingungsschlüssel bereit, unterstützen aber die Verwendung einiger globaler Bedingungsschlüssel. Eine Übersicht aller AWS globalen Bedingungsschlüssel finden Sie unter Kontextschlüssel für AWS globale Bedingungen im IAM-Benutzerhandbuch.

Das Element `Condition` ist optional.

Beispiele

Beispiele für identitätsbasierte Richtlinien für Skalierungspläne finden Sie unter Beispiele für identitätsbasierte Richtlinien für Skalierungspläne.

## Ressourcenbasierte Richtlinien

Andere Amazon Web Services, z. B. Amazon Simple Storage Service, unterstützen auch ressourcenbasierte Berechtigungsrichtlinien. Beispielsweise können Sie einem S3-Bucket eine Berechtigungsrichtlinie zuweisen, um die Zugriffsberechtigungen für diesen Bucket zu verwalten.

Skalierungspläne unterstützen keine ressourcenbasierten Richtlinien.

## Zugriffskontrolllisten () ACLs

Skalierungspläne unterstützen keine Zugriffskontrolllisten (ACLs).

## Autorisierung auf der Basis von Markierungen

Skalierungspläne können nicht markiert werden. Sie verfügen über keine servicedefinierten Ressourcen, die markiert werden können. Daher unterstützen sie nicht die Zugriffskontrolle basierend auf Tags auf einer Ressource.

Skalierungspläne können markierbare Ressourcen wie Auto-Scaling-Gruppen enthalten, die die Steuerung des Zugriffs basierend auf Tags unterstützen. Weitere Informationen finden Sie in der Dokumentation für diesen AWS-Service.

## IAM-Rollen

Eine IAM-Rolle ist eine Entität in Ihrem AWS-Konto mit spezifischen Berechtigungen.

### Verwenden temporärer Anmeldeinformationen

Sie können temporäre Anmeldeinformationen verwenden, um sich über einen Verbund anzumelden, eine IAM-Rolle anzunehmen oder eine kontenübergreifende Rolle anzunehmen. Sie erhalten temporäre Sicherheitsanmeldedaten, indem Sie AWS STS API-Operationen wie AssumeRoleoder aufrufen GetFederationToken.

Skalierungspläne unterstützten die Verwendung von temporären Anmeldeinformationen.

### Serviceverknüpfte Rollen für Skalierungspläne

AWS Auto Scaling verwendet dienstbezogene Rollen für die Berechtigungen, die erforderlich sind, um andere AWS Dienste in Ihrem Namen aufzurufen. Dienstgebundene Rollen erleichtern das Einrichten von Skalierungsplänen, da Sie die erforderlichen Berechtigungen nicht manuell hinzufügen müssen. Weitere Informationen finden Sie unter Verwenden von serviceverknüpften Rollen im -IAM-Benutzerhandbuch.

AWS Auto Scaling verwendet einige Typen von dienstbezogenen Rollen, um in Ihrem Namen andere AWS-Services Rollen aufzurufen, wenn Sie mit einem Skalierungsplan arbeiten:

- Serviceverknüpfte Rolle für vorausschauende Skalierung — Ermöglicht AWS Auto Scaling den Zugriff auf historische Metrikdaten von. CloudWatch Ermöglicht auch die Erstellung geplanter Aktionen für Auto-Scaling-Gruppen basierend auf einer Lastprognose und einer Kapazitätsprognose. Weitere Informationen finden Sie unter Serviceverknüpfte Rolle für vorausschauende Skalierung.
- Servicebezogene Rolle mit Amazon EC2 Auto Scaling — Ermöglicht den Zugriff auf und die Verwaltung von Skalierungsrichtlinien AWS Auto Scaling zur Zielverfolgung für Auto Scaling Scaling-Gruppen. Weitere Informationen finden Sie unter Servicebezogene Rollen für Amazon EC2 Auto Scaling im Benutzerhandbuch zum Amazon EC2 Auto Scaling.
- Service-verknüpfte Rolle für Application Auto Scaling — Ermöglicht AWS Auto Scaling den Zugriff auf und die Verwaltung von Skalierungsrichtlinien für die Zielverfolgung anderer skalierbarer Ressourcen. Für jeden Dienst gibt es eine serviceverknüpfte Rolle. Weitere Informationen finden Sie unter Serviceverknüpfte Rollen für Application Auto Scaling im Benutzerhandbuch zu Application Auto Scaling.

Mit dem folgenden Verfahren können Sie feststellen, ob Ihr Konto bereits über eine serviceverknüpfte Rolle verfügt.

So ermitteln Sie, ob bereits eine serviceverknüpfte Rolle vorhanden ist

1. Öffnen Sie unter https://console.aws.amazon.com/iam/ die IAM-Konsole.
2. Wählen Sie im Navigationsbereich Rollen.
3. Suchen Sie in der Liste nach `AWSServiceRole`, um die serviceverknüpften Rollen zu finden, die in Ihrem Konto vorhanden sind. Suchen Sie nach dem Namen der serviceverknüpften Rolle, die Sie prüfen möchten.

Servicerollen

AWS Auto Scaling hat keine Servicerollen für Skalierungspläne.

## Serviceverknüpfte Rolle für vorausschauende Skalierung

AWS Auto Scaling verwendet dienstbezogene Rollen für die Berechtigungen, die erforderlich sind, um andere in AWS Ihrem Namen anzurufen, wenn Sie mit einem Skalierungsplan arbeiten. Weitere Informationen finden Sie unter Serviceverknüpfte Rollen für Skalierungspläne.

In den folgenden Abschnitten wird beschrieben, wie Sie die serviceverknüpfte Rolle für die vorausschauende Skalierung erstellen und verwalten. Beginnen Sie mit dem Konfigurieren von Berechtigungen, damit eine IAM-Entität (z. B. ein Benutzer, eine Gruppe oder eine Rolle) eine serviceverknüpfte Rolle erstellen, bearbeiten oder löschen kann.

### Von der serviceverknüpften Rolle erteilte Berechtigungen

AWS Auto Scaling verwendet die angegebene dienstverknüpfte RolleAWSServiceRoleForAutoScalingPlans_EC2AutoScaling, um in Ihrem Namen andere AWS Dienste aufzurufen, wenn Sie die vorausschauende Skalierung aktivieren.

AWSServiceRoleForAutoScalingPlans_EC2AutoScalingvertraut darauf, dass der `autoscaling-plans.amazonaws.com` Dienst die Rolle übernimmt.

Diese serviceverknüpfte Rolle verwendet die verwaltete Richtlinie AWSAutoScalingPlansEC2AutoScalingPolicy. Informationen zu den Berechtigungen für diese Richtlinie finden Sie unter AWSAutoScalingPlansEC2AutoScalingPolicy in der Referenz zu von AWS verwalteten Richtlinien.

## Erstellen der serviceverknüpften Rolle (automatisch)

Sie müssen die AWSServiceRoleForAutoScalingPlans_ECbeiden AutoScaling Rollen nicht manuell erstellen. AWS erstellt diese Rolle für Sie, wenn Sie in Ihrem Konto einen Skalierungsplan erstellen und die vorausschauende Skalierung aktivieren.

AWS Um in Ihrem Namen eine serviceverknüpfte Rolle zu erstellen, müssen Sie über die erforderlichen Berechtigungen verfügen. Weitere Informationen finden Sie unter serviceverknüpfte Rollenberechtigung im IAM-Benutzerhandbuch.

## Erstellen der serviceverknüpften Rolle (manuell)

Sie können die serviceverknüpfte Rolle mithilfe der IAM-Konsole, IAM CLI oder IAM API manuell erstellen. Weitere Informationen finden Sie im IAM-Benutzerhandbuch unter Erstellen einer serviceverknüpften Rolle.

So erstellen Sie eine serviceverknüpfte Rolle (AWS CLI)

Verwenden Sie den folgenden create-service-linked-roleBefehl, um die serviceverknüpfte Rolle zu erstellen.

```
aws iam create-service-linked-role --aws-service-name autoscaling-plans.amazonaws.com
```

## Bearbeiten der serviceverknüpften Rolle

Sie können die Beschreibung von AWSServiceRoleForAutoScalingPlans_EC2 AutoScaling mithilfe von IAM bearbeiten. Weitere Informationen finden Sie unter Bearbeiten einer serviceverknüpften Rollenbeschreibung im IAM-Benutzerhandbuch.

## Löschen der serviceverknüpften Rolle

Wenn Sie Skalierungspläne nicht mehr verwenden müssen, empfehlen wir Ihnen, AWSServiceRoleForAutoScalingPlans_EC2 AutoScaling zu löschen.

Sie können eine serviceverknüpfte Rolle erst löschen, nachdem Sie alle Skalierungspläne in Ihrem AWS-Konto gelöscht haben, für die die vorausschauende Skalierung aktiviert ist. Auf diese Weise wird sichergestellt, dass Sie nicht versehentlich die Berechtigungen für den Zugriff auf Ihre Skalierungspläne entfernen.

Sie können die IAM-Konsole, die IAM-CLI oder die IAM-API verwenden, um serviceverknüpfte Rolle zu löschen. Weitere Informationen finden Sie unter Löschen einer serviceverknüpften Rolle im IAM-Benutzerhandbuch.

AWS Auto Scaling Erstellt die Rolle nach dem Löschen der AWSServiceRoleForAutoScalingPlans_EC2AutoScaling serviceverknüpften Rolle erneut, wenn Sie einen Skalierungsplan mit aktivierter vorausschauender Skalierung erstellen.

### Unterstützte Regionen

AWS Auto Scaling unterstützt die Verwendung von serviceverknüpften Rollen in allen verfügbaren AWS-Regionen Skalierungsplänen. Informationen zur regionalen Verfügbarkeit von Skalierungsplänen finden Sie unter AWS Auto Scaling -Endpunkte und -Kontingente in der Allgemeine AWS-Referenz.

## Beispiele für identitätsbasierte Richtlinien für Skalierungspläne

Standardmäßig besitzt ein völlig neuer IAM-Benutzer überhaupt keine Berechtigungen. Ein IAM-Administrator muss IAM-Richtlinien erstellen und zuweisen, die einer IAM-Identität (etwa einem Benutzer oder einer Rolle) die Berechtigung geben, mit Skalierungsplänen zu arbeiten.

Informationen dazu, wie Sie unter Verwendung dieser Beispiel-JSON-Richtliniendokumente eine IAM-Richtlinie erstellen, finden Sie unter Erstellen von Richtlinien auf der JSON-Registerkarte im IAM-Benutzerhandbuch.

Themen



### Best Practices für Richtlinien

Identitätsbasierte Richtlinien legen fest, ob jemand AWS Auto Scaling Ressourcen in Ihrem Konto erstellen, darauf zugreifen oder sie löschen kann. Dies kann zusätzliche Kosten für Ihr verursachen AWS-Konto. Beachten Sie beim Erstellen oder Bearbeiten identitätsbasierter Richtlinien die folgenden Richtlinien und Empfehlungen:

- Erste Schritte mit AWS verwalteten Richtlinien und Umstellung auf Berechtigungen mit den geringsten Rechten — Verwenden Sie die AWS verwalteten Richtlinien, die Berechtigungen für viele gängige Anwendungsfälle gewähren, um damit zu beginnen, Ihren Benutzern und Workloads Berechtigungen zu gewähren. Sie sind in Ihrem verfügbar. AWS-Konto Wir empfehlen Ihnen, die Berechtigungen weiter zu reduzieren, indem Sie vom AWS Kunden verwaltete Richtlinien definieren, die speziell auf Ihre Anwendungsfälle zugeschnitten sind. Weitere Informationen finden Sie unter Von AWS verwaltete Richtlinien oder Von AWS verwaltete Richtlinien für Auftragsfunktionen im IAM-Benutzerhandbuch.
- Anwendung von Berechtigungen mit den geringsten Rechten – Wenn Sie mit IAM-Richtlinien Berechtigungen festlegen, gewähren Sie nur die Berechtigungen, die für die Durchführung einer Aufgabe erforderlich sind. Sie tun dies, indem Sie die Aktionen definieren, die für bestimmte Ressourcen unter bestimmten Bedingungen durchgeführt werden können, auch bekannt als die geringsten Berechtigungen. Weitere Informationen zur Verwendung von IAM zum Anwenden von Berechtigungen finden Sie unter Richtlinien und Berechtigungen in IAM im IAM-Benutzerhandbuch.
- Verwenden von Bedingungen in IAM-Richtlinien zur weiteren Einschränkung des Zugriffs – Sie können Ihren Richtlinien eine Bedingung hinzufügen, um den Zugriff auf Aktionen und Ressourcen zu beschränken. Sie können beispielsweise eine Richtlinienbedingung schreiben, um festzulegen, dass alle Anforderungen mithilfe von SSL gesendet werden müssen. Sie können auch Bedingungen verwenden, um Zugriff auf Serviceaktionen zu gewähren, wenn diese für einen bestimmten Zweck verwendet werden AWS-Service, z. CloudFormation B. Weitere Informationen finden Sie unter IAM-JSON-Richtlinienelemente: Bedingung im IAM-Benutzerhandbuch.
- Verwenden von IAM Access Analyzer zur Validierung Ihrer IAM-Richtlinien, um sichere und funktionale Berechtigungen zu gewährleisten – IAM Access Analyzer validiert neue und vorhandene Richtlinien, damit die Richtlinien der IAM-Richtliniensprache (JSON) und den bewährten IAM-Methoden entsprechen. IAM Access Analyzer stellt mehr als 100 Richtlinienprüfungen und umsetzbare Empfehlungen zur Verfügung, damit Sie sichere und funktionale Richtlinien erstellen können. Weitere Informationen finden Sie unter Richtlinienvalidierung mit IAM Access Analyzer im IAM-Benutzerhandbuch.
- Multi-Faktor-Authentifizierung (MFA) erforderlich — Wenn Sie ein Szenario haben, das IAM-Benutzer oder einen Root-Benutzer in Ihrem System erfordert AWS-Konto, aktivieren Sie MFA für zusätzliche Sicherheit. Um MFA beim Aufrufen von API-Vorgängen anzufordern, fügen Sie Ihren Richtlinien MFA-Bedingungen hinzu. Weitere Informationen finden Sie unter Sicherer API-Zugriff mit MFA im IAM-Benutzerhandbuch.

Weitere Informationen zu bewährten Methoden in IAM finden Sie unter Best Practices für die Sicherheit in IAM im IAM-Benutzerhandbuch.

## Benutzern das Erstellen von Skalierungsplänen erlauben

Im Folgenden wird ein Beispiel für eine identitätsbasierte Richtlinie gezeigt, die Berechtigungen zum Erstellen von Skalierungsplänen gewährt.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "autoscaling-plans:*",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms",
              "cloudwatch:DescribeAlarms",
              "cloudformation:ListStackResources"
            ],
            "Resource": "*"
        }
    ]
}
```

Um mit einem Skalierungsplan arbeiten zu können, müssen Endnutzer über zusätzliche Berechtigungen verfügen, die es ihnen erlauben, mit bestimmten Ressourcen in ihren Konten zu arbeiten. Diese Berechtigungen sind in Zusätzliche erforderliche Berechtigungen aufgeführt.

Jeder Konsolenbenutzer benötigt außerdem Berechtigungen, die es ihm ermöglichen, die skalierbaren Ressourcen in seinem Konto zu ermitteln und Diagramme mit CloudWatch Metrikdaten von der Konsole aus anzuzeigen. AWS Auto Scaling Die zusätzlichen Berechtigungen, die für die Arbeit mit der AWS Auto Scaling Konsole erforderlich sind, sind unten aufgeführt:

- `cloudformation:ListStacks`: Auflisten von Stacks.
- `tag:GetTagKeys`: Für das Auffinden von skalierbaren Ressourcen, die bestimmte Tag-Schlüssel enthalten.

- `tag:GetTagValues`: Für das Auffinden von Ressourcen, die bestimmte Tag-Werte enthalten.
- `autoscaling:DescribeTags`: Zum Finden von Auto-Scaling-Gruppen, die bestimmte Tags enthalten.
- `cloudwatch:GetMetricData`: Zum Anzeigen von Daten in Metrikdiagrammen.

## Benutzern das Aktivieren der prädiktiven Skalierung erlauben

Im Folgenden wird ein Beispiel für eine identitätsbasierte Richtlinie gezeigt, die Berechtigungen zur Aktivierung der vorausschauenden Skalierung gewährt. Mit diesen Berechtigungen werden die Funktionen von Skalierungsplänen erweitert, die zum Skalieren von Auto-Scaling-Gruppen eingerichtet werden.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "cloudwatch:GetMetricData",
              "autoscaling:DescribeAutoScalingGroups",
              "autoscaling:DescribeScheduledActions",
              "autoscaling:BatchPutScheduledUpdateGroupAction",
              "autoscaling:BatchDeleteScheduledAction"
            ],
            "Resource": "*"
        }
    ]
}
```

## Zusätzliche erforderliche Berechtigungen

Um Skalierungspläne erfolgreich zu konfigurieren und zu verwenden, müssen die Benutzer über Berechtigungen für jeden Ziel-Service verfügen, für den sie die Skalierung konfigurieren. Um die für die Arbeit mit Zieldiensten erforderlichen Mindestberechtigungen zu gewähren, lesen Sie die Informationen in diesem Abschnitt und geben Sie die entsprechenden Aktionen im `Action` Element einer IAM-Richtlinienerklärung an.

### Auto-Scaling-Gruppen

Um Auto-Scaling-Gruppen zu einem Skalierungsplan hinzuzufügen, müssen Benutzer über die folgenden Berechtigungen für Amazon EC2 Auto Scaling verfügen:

- `autoscaling:UpdateAutoScalingGroup`
- `autoscaling:DescribeAutoScalingGroups`
- `autoscaling:PutScalingPolicy`
- `autoscaling:DescribePolicies`
- `autoscaling:DeletePolicy`

### ECS-Services

Um ECS-Services zu einem Skalierungsplan hinzuzufügen, müssen Benutzer über die folgenden Berechtigungen für Amazon ECS und Application Auto Scaling verfügen:

- `ecs:DescribeServices`
- `ecs:UpdateService`
- `application-autoscaling:RegisterScalableTarget`
- `application-autoscaling:DescribeScalableTargets`
- `application-autoscaling:DeregisterScalableTarget`
- `application-autoscaling:PutScalingPolicy`
- `application-autoscaling:DescribeScalingPolicies`
- `application-autoscaling:DeleteScalingPolicy`

### Spot-Flotte

Um Spot-Flotten zu einem Skalierungsplan hinzuzufügen, müssen Benutzer über die folgenden Berechtigungen für Amazon EC2 und Application Auto Scaling verfügen:

- `ec2:DescribeSpotFleetRequests`
- `ec2:ModifySpotFleetRequest`
- `application-autoscaling:RegisterScalableTarget`
- `application-autoscaling:DescribeScalableTargets`

- application-autoscaling:DeregisterScalableTarget
- application-autoscaling:PutScalingPolicy
- application-autoscaling:DescribeScalingPolicies
- application-autoscaling:DeleteScalingPolicy

DynamoDB-Tabellen oder globale Indizes

Um DynamoDB-Tabellen oder globale Indizes zu einem Skalierungsplan hinzuzufügen, müssen Benutzer über die folgenden Berechtigungen für DynamoDB und Application Auto Scaling verfügen:

- dynamodb:DescribeTable
- dynamodb:UpdateTable
- application-autoscaling:RegisterScalableTarget
- application-autoscaling:DescribeScalableTargets
- application-autoscaling:DeregisterScalableTarget
- application-autoscaling:PutScalingPolicy
- application-autoscaling:DescribeScalingPolicies
- application-autoscaling:DeleteScalingPolicy

Aurora-DB-Cluster

Um Aurora DB-Cluster zu einem Skalierungsplan hinzuzufügen, müssen Benutzer über die folgenden Berechtigungen für Amazon Aurora und Application Auto Scaling verfügen:

- rds:AddTagsToResource
- rds:CreateDBInstance
- rds:DeleteDBInstance
- rds:DescribeDBClusters
- rds:DescribeDBInstances
- application-autoscaling:RegisterScalableTarget
- application-autoscaling:DescribeScalableTargets
- application-autoscaling:DeregisterScalableTarget

- `application-autoscaling:PutScalingPolicy`
- `application-autoscaling:DescribeScalingPolicies`
- `application-autoscaling:DeleteScalingPolicy`

## Erforderliche Berechtigungen zum Erstellen einer serviceverknüpften Rolle

AWS Auto Scaling erfordert Berechtigungen zum Erstellen einer serviceverknüpften Rolle, wenn ein Benutzer in Ihrem Unternehmen zum ersten Mal einen Skalierungsplan mit aktivierter vorausschauender Skalierung AWS-Konto erstellt. Wenn die serviceverknüpfte Rolle noch nicht existiert, AWS Auto Scaling wird sie in Ihrem Konto erstellt. Die dienstverknüpfte Rolle gewährt Berechtigungen, AWS Auto Scaling sodass sie in Ihrem Namen andere Dienste aufrufen kann.

Damit diese automatische Rollenerstellung möglich ist, müssen Benutzer über Berechtigungen für die Aktion `iam:CreateServiceLinkedRole` verfügen.

```
"Action": "iam:CreateServiceLinkedRole"
```

Im Folgenden wird ein Beispiel für eine identitätsbasierte Richtlinie gezeigt, die Berechtigungen zum Erstellen einer dienstverknüpften Rolle gewährt.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "iam:CreateServiceLinkedRole",
            "Resource": "arn:aws:iam::*:role/aws-service-role/autoscaling-
plans.amazonaws.com/AWSServiceRoleForAutoScalingPlans_EC2AutoScaling",
            "Condition": {
                "StringLike": {
                    "iam:AWSServiceName":"autoscaling-plans.amazonaws.com"
                }
            }
        }
    ]
}
```

Weitere Informationen finden Sie unter Serviceverknüpfte Rolle für vorausschauende Skalierung.

## Überprüfung der Einhaltung der Vorschriften für Skalierungspläne

Informationen darüber, ob AWS-Service ein AWS-Services in den Geltungsbereich bestimmter Compliance-Programme fällt, finden Sie unter Umfang nach Compliance-Programm AWS-Services unter . Wählen Sie dort das Compliance-Programm aus, an dem Sie interessiert sind. Allgemeine Informationen finden Sie unter AWS Compliance-Programme AWS .

Sie können Prüfberichte von Drittanbietern unter herunterladen AWS Artifact. Weitere Informationen finden Sie unter Berichte herunterladen unter .

Ihre Verantwortung für die Einhaltung der Vorschriften bei der Nutzung AWS-Services hängt von der Vertraulichkeit Ihrer Daten, den Compliance-Zielen Ihres Unternehmens und den geltenden Gesetzen und Vorschriften ab. Weitere Informationen zu Ihrer Verantwortung für die Einhaltung der Vorschriften bei der Nutzung AWS-Services finden Sie in der AWS Sicherheitsdokumentation.

## Infrastruktursicherheit für Skalierungspläne

Als verwalteter Dienst AWS Auto Scaling ist er durch AWS globale Netzwerksicherheit geschützt. Informationen zu AWS Sicherheitsdiensten und zum AWS Schutz der Infrastruktur finden Sie unter AWS Cloud-Sicherheit. Informationen zum Entwerfen Ihrer AWS Umgebung unter Verwendung der bewährten Methoden für die Infrastruktursicherheit finden Sie unter Infrastructure Protection in Security Pillar AWS Well-Architected Framework.

Sie verwenden AWS veröffentlichte API-Aufrufe für den Zugriff AWS Auto Scaling über das Netzwerk. Kunden müssen Folgendes unterstützen:

- Transport Layer Security (TLS). Wir benötigen TLS 1.2 und empfehlen TLS 1.3.
- Verschlüsselungs-Suiten mit Perfect Forward Secrecy (PFS) wie DHE (Ephemeral Diffie-Hellman) oder ECDHE (Elliptic Curve Ephemeral Diffie-Hellman). Die meisten modernen Systeme wie Java 7 und höher unterstützen diese Modi.

# Kontingente für Ihre Skalierungspläne

Ihr AWS-Konto hat die Standardkontingente (früher als Limits bezeichnet) für Skalierungspläne. Wenn nicht anders angegeben, gilt jedes Kontingent spezifisch für eine Region. Sie können Erhöhungen für einige Kontingente beantragen und andere Kontingente können nicht erhöht werden.

Um die Kontingente für Application Auto Scaling anzuzeigen, öffnen Sie die Service-Quotas-Konsole. Wählen AWS-Servicesund wählen Sie im Navigationsbereich AWS Auto Scaling Plans aus.

Informationen zur Erhöhung eines Kontingents finden Sie unter Anfordern einer Kontingenterhöhung im Service-Quotas-Benutzerhandbuch.

Ihr AWS-Konto hat die folgenden Kontingente in Bezug auf Skalierungspläne.

| Name | Standard | Anpassbar |
|---|---|---|
| Skalierbare Ressourcen pro Ressourcentyp | Amazon DynamoDB: 3.000 \| Amazon EC2 Auto Scaling Scaling-Gruppen: 200 \| Alle anderen Ressourcentypen: 500 | Ja |
| Skalierungspläne | 100 | Ja |
| Skalierungsanweisungen pro Skalierungsplan | 500 | Nein |
| Konfiguration der Zielnachverfolgung pro Skalierungsanweisung | 10 | Nein |

Denken Sie an die Servicekontingente, wenn Sie Ihre Workloads skalieren. Wenn Sie beispielsweise die maximal zulässige Anzahl von Kapazitätseinheiten eines Services erreichen, wird die Skalierung beendet. Wenn die Nachfrage sinkt und die aktuelle Kapazität sinkt, AWS Auto Scaling kann wieder skaliert werden. Um das Service-Kontingent nicht erneut auszuschöpfen, können Sie eine Erhöhung beantragen. Jeder Service verfügt über eigene Standardkontingente für die maximale Kapazität der Ressource. Informationen zu den Standardkontingenten für andere Amazon Web Services finden Sie unter Service-Endpunkte und -Kontingente im Allgemeine Amazon Web Services-Referenz.

# Dokumentverlauf für Skalierpläne

In der folgenden Tabelle werden wichtige Ergänzungen der AWS Auto Scaling Dokumentation beschrieben. Um Benachrichtigungen über Aktualisierungen dieser Dokumentation zu erhalten, können Sie den RSS-Feed abonnieren.

| Änderung | Beschreibung | Datum |
| --- | --- | --- |
| Neue Inhalte für die Migration von AWS Auto Scaling zu alternativen Optionen | Sie können jetzt von AWS Auto Scaling Amazon EC2 Auto Scaling Predictive Scaling migrieren, das mehr Funktionen bietet. Weitere Informationen finden Sie unter Migrieren Sie Ihren Skalierungsplan. | 5. April 2024 |
| Neue Sicherheitsinhalte | Wir haben ein aktualisiertes Sicherheitskapitel veröffentlicht. Im Rahmen dieses Updates haben wir „Authentifizierung und Zugriffskontrolle“ durch das Identitäts- und Zugriffsmanagement für ersetzt AWS Auto Scaling. | 12. März 2020 |
| Unterstützung für Amazon VPC-Endpunkte | Sie können jetzt eine private Verbindung zwischen Ihrer VPC und AWS Auto Scaling herstellen. Überlegungen und Anleitungen zur Migration finden Sie unter Skalierungspläne und Schnittstellen-VPC-Endpunkte. | 22. November 2019 |
| Support für die Erhöhung der maximalen Kapazität über | Fügt Unterstützung hinzu, um für den Skalierun | 9. März 2019 |

| | | |
|---|---|---|
| die prognostizierte Kapazität hinaus | gsplan eine Erhöhung der maximale Kapazität um einen angegebenen Pufferwert über die prognostizierte Kapazität hinaus über die Konsole zuzulassen. Weitere Informationen finden Sie unter Einstellungen für prädiktive Skalierung. | |
| Prädiktive Skalierung und Verbesserungen | Sie können jetzt mit prädiktiver Skalierung Ihre Amazon-EC2-Auto-Scaling-Gruppen proaktiv skalieren. Diese Version unterstützt außerdem den Ersatz von Skalierungsrichtlinien, die außerhalb des Skalierungsplans erstellt wurden (z. B. von anderen Konsolen), sowie die Steuerung, ob Sie die Funktion zur dynamischen Skalierung Ihres Plans aktivieren. | 20. November 2018 |
| Unterstützung für benutzerdefinierte Ressourceneinstellungen | Unterstützung der Anpassung verschiedener Einstellungen für einzelne Ressourcen oder mehrere Ressourcen gleichzeitig hinzugefügt. | 9. Oktober 2018 |
| Tags als Anwendungsquelle | Diese Version bietet Unterstützung für die Angabe mehrerer Tags als Anwendungsquelle. | 23. April 2018 |
| Neuer Service | Erste Veröffentlichung von. AWS Auto Scaling | 16. Januar 2018 |

Die vorliegende Übersetzung wurde maschinell erstellt. Im Falle eines Konflikts oder eines Widerspruchs zwischen dieser übersetzten Fassung und der englischen Fassung (einschließlich infolge von Verzögerungen bei der Übersetzung) ist die englische Fassung maßgeblich.