aws

Panduan Pengguna Paket Penskalaan

# AWS Auto Scaling

# AWS Auto Scaling: Panduan Pengguna Paket Penskalaan

# Table of Contents

# Apa itu rencana penskalaan?

Gunakan rencana penskalaan untuk mengonfigurasi penskalaan otomatis untuk sumber daya terukur terkait atau terkait dalam hitungan menit. Misalnya, Anda dapat menggunakan tag untuk mengelompokkan sumber daya dalam kategori seperti produksi, pengujian, atau pengembangan. Kemudian, Anda dapat mencari dan menyiapkan rencana penskalaan untuk sumber daya yang dapat diskalakan yang termasuk dalam setiap kategori. Atau, jika infrastruktur cloud Anda termasuk AWS CloudFormation, Anda dapat menentukan template tumpukan yang akan digunakan untuk membuat koleksi sumber daya. Kemudian, buat rencana penskalaan untuk sumber daya yang dapat diskalakan milik setiap tumpukan.

## Sumber daya yang didukung

AWS Auto Scaling mendukung penggunaan rencana penskalaan untuk layanan dan sumber daya berikut:

- Amazon Aurora — Menambah atau mengurangi jumlah replika baca Aurora yang disediakan untuk klaster Aurora DB.
- Amazon EC2 Auto Scaling — Luncurkan atau hentikan instans EC2 dengan menambah atau mengurangi kapasitas grup Auto Scaling yang diinginkan.
- Amazon Elastic Container Service — Meningkatkan atau mengurangi jumlah tugas yang diinginkan di Amazon ECS.
- Amazon DynamoDB — Meningkatkan atau mengurangi kapasitas baca dan tulis yang disediakan dari tabel DynamoDB atau indeks sekunder global.
- Armada Spot — Meluncurkan atau menghentikan instans EC2 dengan menambah atau mengurangi kapasitas target Armada Spot.

## Fitur dan manfaat rencana penskalaan

Rencana penskalaan menyediakan fitur dan manfaat berikut:

- Penemuan sumber daya — AWS Auto Scaling menyediakan penemuan sumber daya otomatis untuk membantu menemukan sumber daya dalam aplikasi Anda yang dapat diskalakan.
- Penskalaan dinamis — Paket penskalaan menggunakan layanan Amazon EC2 Auto Scaling dan Application Auto Scaling untuk menyesuaikan kapasitas sumber daya yang dapat diskalakan guna

menangani perubahan lalu lintas atau beban kerja. Metrik penskalaan dinamis dapat berupa metrik pemanfaatan standar atau throughput, atau metrik khusus.

- Rekomendasi penskalaan bawaan — AWS Auto Scaling menyediakan strategi penskalaan dengan rekomendasi yang dapat Anda gunakan untuk mengoptimalkan kinerja, biaya, atau keseimbangan di antara keduanya.
- Penskalaan prediktif — Rencana penskalaan juga mendukung penskalaan prediktif untuk grup Auto Scaling. Ini membantu menskalakan kapasitas Amazon EC2 Anda lebih cepat ketika ada lonjakan yang terjadi secara teratur.

> **Important**
>
> Jika Anda menggunakan paket penskalaan hanya untuk penskalaan prediktif, kami sangat menyarankan agar Anda menetapkan kebijakan penskalaan prediktif secara langsung pada sumber daya Auto Scaling Anda. Opsi ini menawarkan lebih banyak fitur, seperti menggunakan agregasi metrik untuk membuat metrik kustom baru atau menyimpan data metrik historis di seluruh penerapan. blue/green Untuk informasi selengkapnya tentang Amazon EC2 Auto Scaling, lihat Penskalaan prediktif untuk Amazon EC2 Auto Scaling di Panduan Pengguna Amazon EC2 Auto Scaling. Untuk informasi selengkapnya tentang Application Auto Scaling, lihat Penskalaan Prediktif untuk Application Auto Scaling di Panduan Pengguna Application Auto Scaling.
> Untuk panduan migrasi dari paket penskalaan ke kebijakan penskalaan prediktif Amazon EC2 Auto Scaling, lihat. Migrasikan rencana penskalaan Anda

# Cara memulai

Gunakan sumber daya berikut untuk membantu Anda membuat dan menggunakan rencana penskalaan:

- Cara kerja rencana penyekalaan
- Praktik terbaik untuk rencana penskalaan
- Memulai dengan rencana penskalaan

# Bekerja dengan rencana penskalaan

Anda dapat membuat, mengakses, dan mengelola rencana penskalaan menggunakan salah satu antarmuka berikut:

- Konsol Manajemen AWS— Menyediakan antarmuka web yang dapat Anda gunakan untuk mengakses rencana penskalaan Anda. Jika Anda telah mendaftar Akun AWS, Anda dapat mengakses rencana penskalaan Anda dengan masuk ke Konsol Manajemen AWS, menggunakan kotak pencarian di bilah navigasi untuk mencari AWS Auto Scaling, dan kemudian memilih AWS Auto Scaling.
- AWS Command Line Interface (AWS CLI) — Menyediakan perintah untuk serangkaian luas Layanan AWS, dan didukung pada Windows, macOS, dan Linux. Untuk mulai, lihat Panduan Pengguna AWS Command Line Interface. Untuk informasi selengkapnya, lihat rencana penskalaan otomatis di Referensi Perintah.AWS CLI
- AWS Tools for Windows PowerShell— Menyediakan perintah untuk serangkaian AWS produk yang luas bagi mereka yang membuat skrip di PowerShell lingkungan. Untuk memulai, lihat Alat AWS untuk PowerShell Panduan Pengguna. Untuk informasi lebih lanjut, lihat Alat AWS untuk PowerShell Referensi Cmdlet.
- AWS SDKsMenyediakan operasi API khusus bahasa dan menangani banyak detail koneksi, seperti menghitung tanda tangan, menangani percobaan ulang permintaan, dan menangani kesalahan. Untuk informasi selengkapnya, lihat AWS SDKs.
- HTTPS API - Menyediakan tindakan API tingkat rendah yang Anda panggil menggunakan permintaan HTTPS. Untuk informasi lebih lanjut, lihat Referensi API AWS Auto Scaling.
- CloudFormation— Mendukung pembuatan rencana penskalaan menggunakan CloudFormation templat. Untuk informasi selengkapnya, lihat AWS::AutoScalingPlans::ScalingPlanreferensi di Panduan CloudFormation Pengguna.

# Ketersediaan wilayah

AWS Auto Scaling API tersedia di beberapa Region AWS dan menyediakan titik akhir untuk masing-masing Wilayah ini. .

# Harga

Semua fitur rencana penskalaan diaktifkan untuk Anda gunakan. Fitur disediakan tanpa biaya tambahan di luar biaya layanan untuk CloudWatch dan AWS Cloud sumber daya lain yang Anda gunakan.

> **Note**
>
> Fitur penskalaan prediktif bergantung pada CloudWatch GetMetricDataoperasi untuk mengumpulkan data metrik historis untuk peramalan kapasitas, yang menimbulkan biaya. Namun, jika Anda mengaktifkan penskalaan prediktif dengan kebijakan penskalaan Amazon EC2 Auto Scaling alih-alih paket penskalaan, tidak ada biaya untuk panggilan. `GetMetricData`

# Cara kerja rencana penyekalaan

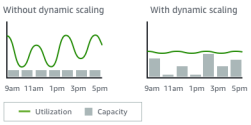
AWS Auto Scaling memungkinkan Anda menggunakan rencana penskalaan untuk mengonfigurasi serangkaian instruksi untuk menskalakan sumber daya Anda. Jika Anda bekerja dengan CloudFormation atau menambahkan tag ke sumber daya yang dapat diskalakan, Anda dapat menyiapkan rencana penskalaan untuk kumpulan sumber daya yang berbeda, per aplikasi. AWS Auto Scaling Konsol memberikan rekomendasi untuk strategi penskalaan yang disesuaikan untuk setiap sumber daya. Setelah Anda membuat rencana penskalaan Anda, ini menggabungkan penskalaan dinamis dan metode penskalaan prediktif bersama-sama untuk mendukung strategi penskalaan Anda.

Apa itu strategi penyekalaan?

Strategi penskalaan memberi tahu AWS Auto Scaling cara mengoptimalkan pemanfaatan sumber daya dalam rencana penskalaan Anda. Anda dapat mengoptimalkan ketersediaan, untuk biaya, atau keseimbangan keduanya. Atau, Anda juga dapat membuat strategi kustom sendiri, sesuai metrik dan ambang batas yang Anda tentukan. Anda dapat menetapkan strategi terpisah untuk setiap sumber daya atau jenis sumber daya.


Apa itu penyekalaan dinamis?

Skala dinamis menciptakan kebijakan penyekalaan pelacakan target untuk sumber daya dalam rencana penyekalaan Anda. Kebijakan penyekalaan ini menyesuaikan kapasitas sumber daya dalam menanggapi perubahan langsung dalam pemanfaatan sumber daya. Tujuannya adalah untuk memberikan kapasitas yang cukup untuk mempertahankan pemanfaatan pada nilai target yang ditentukan oleh strategi penyekalaan. Hal ini serupa dengan cara termostat Anda menjaga suhu di rumah Anda. Anda memilih suhu dan termostat akan melakukan sisanya.

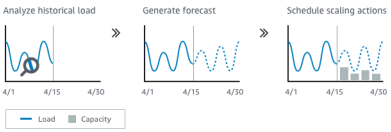



Misalnya, Anda dapat mengonfigurasi paket penskalaan untuk menjaga jumlah tugas yang dijalankan oleh layanan Amazon Elastic Container Service (Amazon ECS) pada 75 persen CPU. Ketika pemanfaatan CPU layanan Anda melebihi 75 persen (artinya lebih dari 75 persen CPU yang dicadangkan untuk layanan sedang digunakan), maka kebijakan penskalaan Anda menambahkan tugas lain ke layanan Anda untuk membantu peningkatan beban.

Apa itu penyekalaan prediktif?

Penskalaan prediktif menggunakan pembelajaran mesin untuk menganalisis beban kerja historis setiap sumber daya dan secara teratur memperkirakan beban masa depan. Hal ini serupa dengan cara kerja prakiraan cuaca. Menggunakan perkiraan, skala prediktif menghasilkan tindakan penyekalaan terjadwal untuk memastikan kapasitas sumber daya tersedia sebelum aplikasi Anda membutuhkannya. Seperti penyekalaan dinamis, penyekalaan prediktif bekerja untuk mempertahankan pemanfaatan pada nilai target yang ditentukan oleh strategi penyekalaan.

> **Important**
>
> Jika Anda menggunakan paket penskalaan hanya untuk penskalaan prediktif, kami sangat menyarankan agar Anda menetapkan kebijakan penskalaan prediktif secara langsung pada sumber daya Auto Scaling Anda. Opsi ini menawarkan lebih banyak fitur, seperti menggunakan agregasi metrik untuk membuat metrik kustom baru atau menyimpan data metrik historis di seluruh penerapan. blue/green Untuk informasi selengkapnya tentang Amazon EC2 Auto Scaling, lihat Penskalaan prediktif untuk Amazon EC2 Auto Scaling di Panduan Pengguna Amazon EC2 Auto Scaling. Untuk informasi selengkapnya tentang Application Auto Scaling, lihat Penskalaan Prediktif untuk Application Auto Scaling di Panduan Pengguna Application Auto Scaling.
> Untuk panduan migrasi dari paket penskalaan ke kebijakan penskalaan prediktif Amazon EC2 Auto Scaling, lihat. Migrasikan rencana penskalaan Anda




Misalnya, Anda dapat mengaktifkan penyekalaan prediktif dan mengkonfigurasi strategi penyekalaan Anda untuk menjaga pemanfaatan CPU rata-rata dari grup Auto Scaling Anda sebesar 50 persen. Perkiraan Anda meminta lonjakan lalu lintas terjadi setiap hari pada pukul 8:00. Rencana penyekalaan Anda menciptakan tindakan penyekalaan terjadwal di masa depan untuk memastikan bahwa kelompok Auto Scaling Anda siap menangani lalu lintas tersebut di masa mendatang. Hal ini membantu menjaga kinerja aplikasi tetap konstan, dengan tujuan untuk selalu memiliki kapasitas yang diperlukan untuk mempertahankan pemanfaatan sumber daya sedekat 50 persen mungkin setiap saat.

Berikut ini adalah konsep kunci untuk memahami penskalaan prediktif:

- Peramalan beban: AWS Auto Scaling menganalisis hingga 14 hari sejarah untuk metrik beban tertentu dan memperkirakan permintaan masa depan untuk dua hari ke depan. Data ini tersedia dalam interval satu jam dan diperbarui setiap hari.
- Tindakan penskalaan AWS Auto Scaling terjadwal: menjadwalkan tindakan penskalaan yang secara proaktif meningkatkan dan mengurangi kapasitas agar sesuai dengan perkiraan beban. Pada waktu yang dijadwalkan, AWS Auto Scaling perbarui kapasitas minimum dengan nilai yang ditentukan oleh tindakan penskalaan terjadwal. Tujuannya adalah untuk mempertahankan pemanfaatan sumber daya pada nilai target yang ditentukan oleh strategi penyekalaan. Jika aplikasi Anda memerlukan kapasitas yang lebih besar daripada prakiraan, skala dinamis tersedia untuk menambahkan kapasitas tambahan.
- Perilaku kapasitas maksimum: Batas kapasitas minimum dan maksimum untuk penskalaan otomatis berlaku untuk setiap sumber daya. Namun, Anda dapat mengontrol apakah aplikasi Anda dapat meningkatkan kapasitas melebihi kapasitas maksimum ketika kapasitas perkiraan lebih tinggi dari kapasitas maksimum.

# Praktik terbaik untuk rencana penskalaan

Praktik terbaik berikut dapat membantu Anda memaksimalkan rencana penyekalaan:

- Saat Anda membuat templat peluncuran atau konfigurasi peluncuran, aktifkan pemantauan terperinci untuk mendapatkan data CloudWatch metrik untuk instans EC2 pada frekuensi satu menit karena hal itu memastikan respons yang lebih cepat terhadap perubahan pemuatan. Penskalaan pada metrik dengan frekuensi lima menit dapat menghasilkan waktu respons yang lebih lambat dan penskalaan pada data metrik basi. Secara default, instans EC2 diaktifkan untuk pemantauan dasar, yang berarti data metrik untuk instans tersedia pada interval lima menit. Dengan biaya tambahan, Anda dapat mengaktifkan pemantauan terperinci untuk mendapatkan data metrik untuk instans pada frekuensi satu menit. Untuk informasi lebih lanjut, lihat Konfigurasi pemantauan untuk instans Auto Scaling dalam Panduan Pengguna Amazon EC2 Auto Scaling.
- Kami juga merekomendasikan Anda untuk mengaktifkan metrik grup Auto Scaling. Jika tidak, data kapasitas aktual tidak ditampilkan dalam grafik prakiraan kapasitas yang tersedia pada saat penyelesaian wizard Pembuatan Rencana Skala. Untuk informasi selengkapnya, lihat CloudWatch Metrik pemantauan untuk grup dan instans Auto Scaling di Panduan Pengguna Amazon EC2 Auto Scaling.
- Periksa jenis instans mana yang digunakan grup Auto Scaling Anda dan waspadai penggunaan jenis instans performa yang dapat dibobol. Instans Amazon EC2 dengan kinerja burstable, seperti instans T3 dan T2, dirancang untuk memberikan tingkat kinerja CPU dasar dengan kemampuan untuk meledak ke tingkat yang lebih tinggi bila diperlukan oleh beban kerja Anda. Bergantung pada pemanfaatan target yang ditentukan oleh rencana penyekalaan, Anda dapat menjalankan risiko melebihi data awal dan kemudian kehabisan kredit CPU, yang membatasi kinerja. Untuk informasi lebih lanjut, lihat Kredit CPU dan kinerja dasar untuk instans kinerja yang dapat pecah. Untuk mengonfigurasi instance ini sebagai`unlimited`, lihat Menggunakan grup Auto Scaling untuk meluncurkan instans performa burstable sebagai Unlimited di Panduan Pengguna Amazon EC2.

# Pertimbangan lainnya

> **Important**
>
> Jika Anda menggunakan paket penskalaan hanya untuk penskalaan prediktif, kami sangat menyarankan agar Anda menetapkan kebijakan penskalaan prediktif secara langsung

pada sumber daya Auto Scaling Anda. Opsi ini menawarkan lebih banyak fitur, seperti menggunakan agregasi metrik untuk membuat metrik kustom baru atau menyimpan data metrik historis di seluruh penerapan. blue/green Untuk informasi selengkapnya tentang Amazon EC2 Auto Scaling, lihat Penskalaan prediktif untuk Amazon EC2 Auto Scaling di Panduan Pengguna Amazon EC2 Auto Scaling. Untuk informasi selengkapnya tentang Application Auto Scaling, lihat Penskalaan Prediktif untuk Application Auto Scaling di Panduan Pengguna Application Auto Scaling.
Untuk panduan migrasi dari paket penskalaan ke kebijakan penskalaan prediktif Amazon EC2 Auto Scaling, lihat. Migrasikan rencana penskalaan Anda

Ingatlah pertimbangan tambahan berikut:

- Penskalaan prediktif menggunakan prakiraan beban untuk menjadwalkan kapasitas di masa depan. Kualitas prakiraan bervariasi berdasarkan seberapa siklus beban dan penerapan model peramalan terlatih. Skala prediktif dapat dijalankan dalam mode hanya prakiraan untuk menilai kualitas prakiraan dan tindakan penyekalaan yang dibuat oleh prakiraan. Anda dapat mengatur mode penyekalaan prediktif ke Hanya prakiraan saat Anda membuat rencana penyekalaan dan kemudian mengubahnya menjadi Prakiraan dan skala setelah selesai menilai kualitas prakiraan. Untuk informasi selengkapnya, lihat Pengaturan penyekalaan prediktif dan Memantau dan mengevaluasi prakiraan.
- Jika Anda memilih untuk menentukan metrik yang berbeda untuk penyekalaan prediktif, Anda harus memastikan bahwa metrik pennyekalaan dan metrik beban sangat terkorelasi. Nilai metrik harus meningkat dan menurun secara proporsional terhadap jumlah kasus dalam kelompok Auto Scaling. Ini memastikan bahwa data metrik dapat digunakan untuk menyekalakan atau dalam jumlah kejadian secara proporsional. Misalnya, metrik beban adalah jumlah permintaan total dan metrik penyekalaan adalah pemanfaatan CPU rata-rata. Jika jumlah permintaan total meningkat sebesar 50 persen, utilisasi CPU rata-rata juga harus meningkat sebesar 50 persen, asalkan kapasitas tidak berubah.
- Sebelum membuat rencana penskalaan Anda, Anda harus menghapus tindakan penskalaan yang dijadwalkan sebelumnya yang tidak lagi Anda perlukan dengan mengakses konsol tempat mereka dibuat. AWS Auto Scaling tidak membuat tindakan penskalaan prediktif yang tumpang tindih dengan tindakan penskalaan terjadwal yang ada.
- Pengaturan yang disesuaikan untuk kapasitas minimum dan maksimum, serta pengaturan lain yang digunakan untuk penyekalaan dinamis, muncul di konsol lainnya. Namun, kami menyarankan

agar setelah Anda membuat rencana penyekalaan, Anda tidak mengubah pengaturan ini dari konsol lain karena rencana penyekalaan Anda tidak menerima pembaruan dari konsol lain.

- Rencana penyekalaan Anda dapat berisi sumber daya dari berbagai layanan, tetapi setiap sumber daya hanya dapat berada dalam satu rencana penyekalaan pada satu waktu.

# Menghindari ActiveWithProblems kesalahan

Kesalahan "ActiveWithProblems" dapat terjadi ketika rencana penskalaan dibuat, atau sumber daya ditambahkan ke rencana penskalaan. Kesalahan terjadi saat rencana penyekalaan aktif, tetapi konfigurasi penyekalaan untuk satu atau beberapa sumber daya tidak dapat diterapkan.

Biasanya, hal ini terjadi karena sumber daya sudah memiliki kebijakan penyekalaan atau kelompok Auto Scaling tidak memenuhi persyaratan minimum untuk penyekalaan prediktif.

Jika salah satu sumber daya Anda sudah memiliki kebijakan penskalaan dari berbagai konsol layanan, AWS Auto Scaling jangan menimpa kebijakan penskalaan lain ini atau membuat yang baru secara default. Anda dapat menghapus kebijakan penskalaan yang ada secara opsional dan menggantinya dengan kebijakan penskalaan pelacakan target yang dibuat dari konsol. AWS Auto Scaling Anda melakukan ini dengan mengaktifkan pengaturan Ganti kebijakan penyekalaan eksternal untuk setiap sumber daya yang memiliki kebijakan penyekalaan untuk ditimpa.

Dengan skala prediktif, kami merekomendasikan menunggu 24 jam setelah membuat grup Auto Scaling baru untuk mengkonfigurasi skala prediktif. Minimal, harus ada data historis selama 24 jam untuk membuat prakiraan awal. Jika grup memiliki data historis kurang dari 24 jam dan penskalaan prediktif diaktifkan, maka rencana penskalaan tidak dapat menghasilkan perkiraan hingga periode perkiraan berikutnya, setelah grup mengumpulkan jumlah data yang diperlukan. Namun, Anda juga dapat mengedit dan menyimpan rencana penyekalaan untuk memulai ulang proses prakiraan segera setelah data 24 jam tersedia.

# Memulai dengan rencana penskalaan

Sebelum Anda membuat rencana penskalaan untuk digunakan dengan aplikasi Anda, tinjau aplikasi Anda secara menyeluruh saat berjalan di AWS Cloud. Perhatikan hal-hal berikut ini:

- Apakah Anda sudah memiliki kebijakan penyekalaan yang dibuat dari konsol lain. Anda dapat mengganti kebijakan penyekalaan yang ada, atau Anda dapat menyimpannya (tanpa diizinkan untuk mengubah nilai-nilai tersebut) saat Anda membuat rencana penyekalaan.
- Pemanfaatan target yang masuk akal untuk setiap sumber daya yang dapat diskalakan dalam aplikasi Anda berdasarkan sumber daya secara keseluruhan. Sebagai contoh, jumlah CPU yang ada di EC2 pada grup Auto Scaling diharapkan akan digunakan dibandingkan dengan CPU yang tersedia. Atau untuk layanan seperti DynamoDB yang menggunakan model throughput yang disediakan, jumlah aktivitas baca dan tulis yang diharapkan digunakan tabel atau indeks dibandingkan dengan throughput yang tersedia. Dengan kata lain, rasio konsumsi terhadap kapasitas yang tersedia. Anda dapat mengubah pemanfaatan target kapan saja setelah Anda membuat rencana penskalaan Anda.
- Berapa lama waktu yang diperlukan untuk meluncurkan dan mengonfigurasi server. Mengetahui hal ini membantu Anda mengonfigurasi jendela untuk setiap instans EC2 untuk pemanasan setelah diluncurkan untuk memastikan bahwa server baru tidak diluncurkan saat yang sebelumnya masih diluncurkan.
- Apakah riwayat metrik cukup panjang untuk digunakan dengan skala prediktif (jika menggunakan grup Auto Scaling yang baru dibuat). Secara umum, memiliki 14 hari penuh data historis diterjemahkan menjadi prakiraan yang lebih akurat. Minimal adalah 24 jam.

Semakin baik Anda memahami aplikasi Anda, semakin efektif Anda dapat membuat rencana penyekalaan.

Tugas-tugas berikut membantu Anda menjadi terbiasa dengan rencana penskalaan. Anda akan membuat rencana penskalaan untuk satu grup Auto Scaling dan mengaktifkan penskalaan prediktif dan penskalaan dinamis.

Tugas

- Langkah 1: Temukan sumber daya Anda yang dapat diskalakan
- Langkah 2: Tentukan strategi penskalaan
- Langkah 3: Mengonfigurasi pengaturan lanjutan (opsional)

# Langkah 1: Temukan sumber daya Anda yang dapat diskalakan

Bagian ini mencakup pengantar langsung untuk membuat rencana penskalaan di konsol. AWS Auto Scaling Jika ini adalah rencana penskalaan pertama Anda, sebaiknya Anda mulai dengan membuat paket penskalaan sampel menggunakan grup Penskalaan Otomatis Amazon EC2.

## Prasyarat

Untuk berlatih menggunakan rencana penskalaan, buat grup Auto Scaling. Luncurkan setidaknya satu instans Amazon EC2 di grup Auto Scaling. Untuk informasi selengkapnya, lihat Memulai Penskalaan Otomatis Amazon EC2 di Panduan Pengguna Penskalaan Otomatis Amazon EC2.

Gunakan grup Auto Scaling dengan CloudWatch metrik yang diaktifkan untuk memiliki data kapasitas pada grafik yang tersedia saat Anda menyelesaikan wizard Buat Rencana Penskalaan. Untuk informasi selengkapnya, lihat Monitor CloudWatch metrik untuk grup dan instans Auto Scaling di Panduan Pengguna Amazon EC2 Auto Scaling.

Hasilkan beberapa beban selama beberapa hari atau lebih agar data CloudWatch metrik tersedia untuk fitur penskalaan prediktif, jika memungkinkan.

Verifikasi bahwa Anda memiliki izin yang diperlukan untuk bekerja dengan rencana penskalaan. Untuk informasi selengkapnya, lihat Manajemen identitas dan akses untuk rencana penskalaan.

## Tambahkan grup Auto Scaling Anda ke paket penskalaan baru

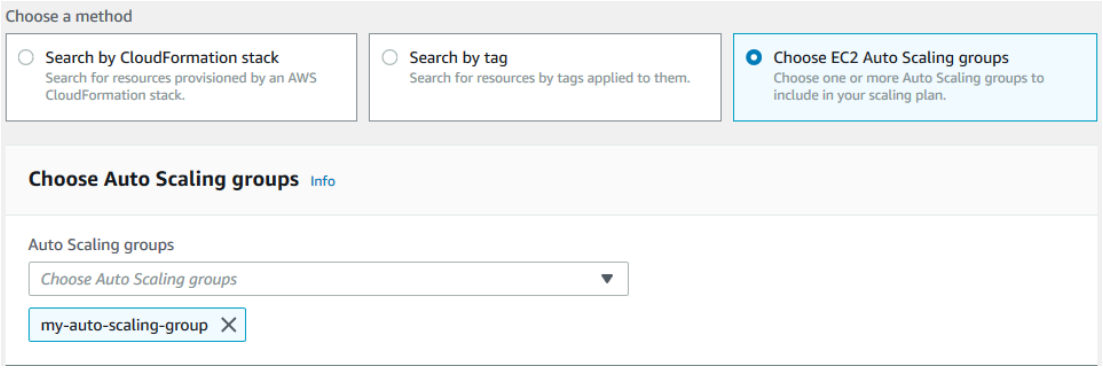
Saat Anda membuat rencana penskalaan dari konsol, ini membantu Anda menemukan sumber daya yang dapat diskalakan sebagai langkah pertama. Sebelum Anda mulai, konfirmasikan bahwa Anda memenuhi persyaratan berikut:

- Anda membuat grup Auto Scaling dan meluncurkan setidaknya satu instans EC2, seperti yang dijelaskan di bagian sebelumnya.
- Grup Auto Scaling yang Anda buat telah ada setidaknya selama 24 jam.

Untuk mulai membuat rencana penskalaan

1. Buka AWS Auto Scaling konsol di https://console.aws.amazon.com/awsautoscaling/.
2. Pada bilah navigasi di bagian atas layar, pilih Wilayah yang sama dengan yang Anda gunakan saat membuat grup Auto Scaling.
3. Dari halaman selamat datang, pilih Memulai.
4. Pada halaman Temukan sumber daya yang dapat diskalakan, lakukan salah satu hal berikut:
   - Pilih Cari berdasarkan CloudFormation tumpukan, lalu pilih CloudFormation tumpukan yang akan digunakan.
   - Pilih Cari berdasarkan tag. Kemudian, untuk setiap tag, pilih kunci tag dari Key dan nilai tag dari Value. Untuk menambahkan tanda, pilih Tambahkan baris lain. Untuk menghapus tanda, pilih Hapus.
   - Pilih Pilih grup Auto Scaling EC2, lalu pilih satu atau beberapa grup Auto Scaling.

> **Note**
>
> Untuk tutorial pengantar, pilih Pilih grup Auto Scaling EC2, lalu pilih grup Auto Scaling yang Anda buat.




5. Pilih Berikutnya untuk melanjutkan proses pembuatan rencana penskalaan.

# Pelajari lebih lanjut tentang menemukan sumber daya yang dapat diskalakan

Jika Anda telah membuat contoh rencana penskalaan dan ingin membuat lebih banyak lagi, lihat skenario berikut untuk menggunakan CloudFormation tumpukan atau sekumpulan tag secara lebih rinci. Anda dapat menggunakan bagian ini untuk memutuskan apakah akan memilih opsi Cari berdasarkan CloudFormation tumpukan atau Cari berdasarkan tag untuk menemukan sumber daya yang dapat diskalakan saat menggunakan konsol untuk membuat rencana penskalaan Anda.

Saat Anda memilih opsi Cari berdasarkan CloudFormation tumpukan atau Cari berdasarkan tag di langkah 1 panduan Buat Rencana Penskalaan, ini membuat sumber daya terukur yang terkait dengan tumpukan atau kumpulan tag tersedia untuk rencana penskalaan. Saat Anda menentukan rencana penyekalaan, Anda kemudian dapat memilih sumber daya mana yang akan disertakan atau dikecualikan.

Menemukan sumber daya yang dapat diskalakan menggunakan tumpukan CloudFormation

Saat Anda menggunakan CloudFormation, Anda bekerja dengan tumpukan untuk menyediakan sumber daya. Semua sumber daya dalam susunan ditentukan oleh templat tumpukan. Rencana penskalaan Anda menambahkan lapisan orkestrasi di atas tumpukan yang memudahkan konfigurasi penskalaan untuk berbagai sumber daya. Tanpa rencana penyekalaan, Anda perlu mengatur penyekalaan untuk setiap sumber daya yang dapat diskalakan secara individu. Ini berarti mencari tahu urutan penyediaan sumber daya dan menyekalakan kebijakan, serta memahami samar-samar tentang cara kerja dependensi ini.

Di AWS Auto Scaling konsol, Anda dapat memilih tumpukan yang ada untuk memindai sumber daya yang dapat dikonfigurasi untuk penskalaan otomatis. AWS Auto Scaling hanya menemukan sumber daya yang ditentukan dalam tumpukan yang dipilih. Tumpukan yang bertingkat tidak dijelajahi.

Agar layanan ECS Anda dapat ditemukan dalam CloudFormation tumpukan, AWS Auto Scaling konsol harus mengetahui cluster ECS mana yang menjalankan layanan. Ini mengharuskan layanan ECS Anda berada dalam CloudFormation tumpukan yang sama dengan cluster ECS yang menjalankan layanan. Jika tidak, mereka harus menjadi bagian dari klaster default. Untuk diidentifikasi dengan benar, nama layanan ECS juga harus unik di setiap klaster ECS ini.

Untuk informasi lebih lanjut tentang CloudFormation, lihat Apa itu CloudFormation? dalam AWS CloudFormation User Guide.

Menemukan sumber daya yang dapat diskalakan menggunakan tanda

Tag menyediakan metadata yang dapat digunakan untuk menemukan sumber daya terukur terkait di AWS Auto Scaling konsol, menggunakan filter tag.

Gunakan tanda untuk menemukan sumber daya berikut:

- Klaster Aurora DB
- Grup Auto Scaling
- Tabel DynamoDB dan indeks sekunder global

Saat Anda mencari lebih dari satu tanda, setiap sumber daya harus memiliki semua tanda yang tercantum untuk ditemukan.

Untuk informasi lebih lanjut tentang penandaan, baca dokumentasi berikut.

- Pelajari cara menandai klaster Aurora di Panduan Pengguna Amazon Aurora.
- Pelajari cara menandai grup Auto Scaling di Panduan Pengguna Amazon EC2 Auto Scaling.
- Pelajari cara menandai sumber daya DynamoDB di Panduan Pengembang Amazon DynamoDB.

## Langkah 2: Tentukan strategi penskalaan

Gunakan prosedur berikut untuk menentukan strategi penyekalaan sumber daya yang ditemukan di langkah sebelumnya.

Untuk setiap jenis sumber daya, AWS Auto Scaling pilih metrik yang paling umum digunakan untuk menentukan berapa banyak sumber daya yang digunakan pada waktu tertentu. Anda memilih strategi penyekalaan yang paling sesuai untuk mengoptimalkan kinerja aplikasi Anda berdasarkan metrik ini. Saat Anda mengaktifkan fitur penyekalaan dinamis dan fitur penyekalaan prediktif, strategi penyekalaan digunakan bersama di antaranya. Untuk informasi selengkapnya, lihat Cara kerja rencana penyekalaan.

Tersedia strategi penyekalaan berikut:

- Optimalkan ketersediaan -AWS Auto Scaling skala sumber daya keluar dan masuk secara otomatis untuk mempertahankan pemanfaatan sumber daya sebesar 40 persen. Opsi ini berguna ketika aplikasi Anda memiliki kebutuhan penyekalaan yang mendesak dan terkadang tidak dapat diprediksi.

- Menyeimbangkan ketersediaan dan biaya -AWS Auto Scaling skala sumber daya keluar dan masuk secara otomatis untuk mempertahankan pemanfaatan sumber daya sebesar 50 persen. Opsi ini membantu Anda menjaga ketersediaan yang tinggi sekaligus mengurangi biaya.
- Optimalkan biaya —AWS Auto Scaling skala sumber daya keluar dan masuk secara otomatis untuk mempertahankan pemanfaatan sumber daya sebesar 70 persen. Opsi ini berguna untuk menurunkan biaya jika aplikasi Anda dapat menangani pengurangan kapasitas buffer ketika ada perubahan permintaan yang tidak terduga.

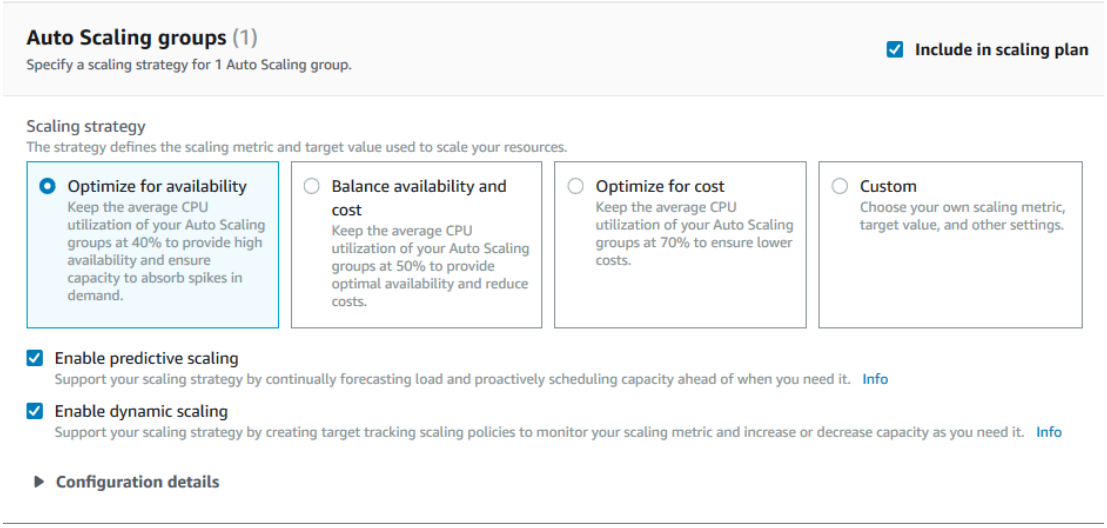
Misalnya, rencana penyekalaan mengkonfigurasi grup Auto Scaling Anda untuk menambahkan atau menghapus instans Amazon EC2 berdasarkan berapa banyak CPU digunakan secara rata-rata untuk semua instans dalam grup. Anda memilih untuk mengoptimalkan penggunaan untuk ketersediaan, biaya, atau kombinasi keduanya dengan mengubah strategi penyekalaan.

Atau, Anda dapat mengonfigurasi strategi khusus jika strategi yang ada tidak memenuhi kebutuhan Anda. Dengan strategi khusus, Anda dapat mengubah nilai pemanfaatan target, memilih metrik yang berbeda, atau keduanya.

> **Important**
>
> Untuk tutorial pengantar, selesaikan hanya langkah pertama dari prosedur berikut dan kemudian pilih Berikutnya untuk melanjutkan.

Untuk menentukan strategi penskalaan

1. Di halaman Menentukan strategi penyekalaan, untuk Perincian rencana penyekalaan, Nama, masukkan nama untuk rencana penyekalaan Anda. Nama rencana penskalaan Anda harus unik dalam rangkaian rencana penskalaan Anda untuk Wilayah. Ini dapat memiliki maksimum 128 karakter, dan tidak boleh berisi pipa “|”, garis miring “/”, atau titik dua “:”.
2. Semua sumber daya yang disertakan terdaftar berdasarkan jenis sumber daya. Untuk grup Auto Scaling, lakukan hal berikut:




a. Lewati langkah ini untuk menggunakan strategi dan metrik penskalaan default. Untuk menggunakan strategi atau metrik penskalaan yang berbeda, lanjutkan dengan langkah-langkah berikut:

   i. Untuk strategi Scaling, pilih strategi penskalaan yang diinginkan.

      Untuk tutorial pengantar, pastikan untuk memilih Optimalkan untuk ketersediaan. Ini menentukan bahwa penggunaan CPU rata-rata grup Auto Scaling Anda akan dipertahankan pada 40 persen.

   ii. Jika Anda memilih Kustom, perluas detail Konfigurasi untuk memilih metrik dan nilai target yang diinginkan.

      - Untuk Metrik penyekalaan, pilih metrik penyekalaan yang diinginkan.
      - Untuk nilai Target, pilih nilai target yang diinginkan, seperti pemanfaatan target atau throughput target selama interval satu menit.
      - Untuk metrik Muat [Khusus grup Auto Scaling], pilih metrik beban yang diinginkan untuk digunakan untuk penskalaan prediktif.
      - Pilih Ganti kebijakan penskalaan eksternal untuk menentukan kebijakan penskalaan yang AWS Auto Scaling dapat menghapus kebijakan penskalaan yang sebelumnya dibuat dari luar rencana penskalaan (seperti dari konsol lain) dan ganti dengan kebijakan penskalaan pelacakan target baru yang dibuat oleh rencana penskalaan.

b. (Opsional) Secara default, penskalaan prediktif diaktifkan untuk grup Auto Scaling. Untuk menonaktifkan penskalaan prediktif untuk grup Auto Scaling, hapus Aktifkan penskalaan prediktif.

c. (Opsional) Secara default, penskalaan dinamis diaktifkan untuk setiap jenis sumber daya. Untuk menonaktifkan penskalaan dinamis untuk jenis sumber daya, hapus Aktifkan penskalaan dinamis.

d. (Opsional) Secara default, ketika Anda menentukan sumber aplikasi dari mana sumber daya yang dapat diskalakan ditemukan, semua jenis sumber daya secara otomatis disertakan dalam rencana penyekalaan Anda. Untuk menghilangkan jenis sumber daya dari rencana penyekalaan Anda, hapus Termasuk dalam rencana penyekalaan.

3. (Opsional) Untuk menentukan strategi penskalaan untuk jenis sumber daya lain, ulangi langkah sebelumnya.
4. Setelah selesai, pilih Berikutnya untuk melanjutkan proses pembuatan rencana penskalaan.

## Langkah 3: Mengonfigurasi pengaturan lanjutan (opsional)

Setelah Anda menentukan strategi penyekalaan yang akan digunakan untuk setiap jenis sumber daya, Anda dapat memilih untuk menyesuaikan pengaturan default mana pun berdasarkan setiap sumber daya dengan menggunakan Mengkonfigurasi pengaturan lanjutan langkah selanjutnya. Untuk setiap jenis sumber daya, ada beberapa kelompok pengaturan yang dapat Anda sesuaikan. Namun, dalam kebanyakan kasus, pengaturan default harus lebih efisien, dengan kemungkinan pengecualian nilai untuk kapasitas minimum dan kapasitas maksimum, yang harus disesuaikan dengan hati-hati.

Lewati prosedur ini jika Anda ingin mempertahankan pengaturan default. Anda dapat mengubah pengaturan ini kapan saja dengan mengedit rencana penyekalaan.

> **Important**
>
> Untuk tutorial pengantar, mari buat beberapa perubahan untuk memperbarui kapasitas maksimum grup Auto Scaling Anda dan mengaktifkan penskalaan prediktif dalam mode perkiraan saja. Meskipun Anda tidak perlu menyesuaikan semua pengaturan untuk tutorial, mari kita periksa pengaturan tersebut secara singkat di setiap bagian.

### Pengaturan umum

Gunakan prosedur ini untuk melihat dan menyesuaikan pengaturan yang Anda tentukan di langkah sebelumnya, berdasarkan setiap sumber daya. Anda juga dapat menyesuaikan kapasitas minimum dan maksimum untuk setiap sumber daya.

Untuk melihat dan menyesuaikan pengaturan umum

1. Di Mengonfigurasi pengaturan lanjutan , pilih panah di sebelah kiri judul bagian mana pun untuk memperluas bagian tersebut. Untuk tutorial, perluas Kelompok Auto Scaling bagian.
2. Dari tabel yang ditampilkan, pilih kelompok Auto Scaling yang Anda gunakan dalam tutorial ini.
3. Tinggalkan Termasuk dalam rencana penyekalaan yang dipilih. Jika opsi ini tidak dipilih, sumber daya dihilangkan dari rencana penyekalaan. Jika Anda tidak menyertakan setidaknya satu sumber daya, rencana penyekalaan tidak dapat dibuat.
4. Untuk memperluas tampilan dan melihat detail Pengaturan Umum , pilih panah di sebelah kiri judul bagian.
5. Anda dapat membuat pilihan untuk item berikut. Untuk tutorial ini, temukan Kapasitas maksimum mengatur dan memasukkan nilai 3 menggantikan nilai saat ini.

   - Strategi penyekalaan—Memungkinkan Anda mengoptimalkan ketersediaan, biaya, atau keseimbangan keduanya, atau menentukan strategi kustom.
   - Aktifkan penyekalaan dinamis—Jika pengaturan ini dihapus, sumber daya yang dipilih tidak dapat diskalakan menggunakan konfigurasi penyekalaan pelacakan target.
   - Aktifkan penyekalaan prediktif—[Grup Auto Scaling saja] Jika pengaturan ini dihapus, grup yang dipilih tidak dapat diskalakan menggunakan penyekalaan prediktif.
   - Metrik penyekalaan—Tentukan metrik penyekalaan yang akan digunakan. Jika Anda memilih Kustom, Anda dapat menentukan metrik khusus untuk digunakan, bukan metrik yang telah ditentukan yang tersedia di konsol. Untuk informasi selengkapnya, lihat topik berikutnya dalam bagian ini.
   - Nilai target—Tentukan nilai pemanfaatan target yang akan digunakan.
   - Metrik muatan—[Hanya grup Auto Scaling Menentukan metrik beban yang akan digunakan. Jika Anda memilih Kustom, Anda dapat menentukan metrik khusus untuk digunakan, bukan metrik yang telah ditentukan yang tersedia di konsol. Untuk informasi selengkapnya, lihat topik berikutnya dalam bagian ini.
   - Kapasitas minimum —Menentukan kapasitas minimum untuk sumber daya. AWS Auto Scaling memastikan bahwa sumber daya Anda tidak pernah berada di bawah ukuran ini.
   - Kapasitas maksimum —Menentukan kapasitas maksimum untuk sumber daya. AWS Auto Scaling memastikan bahwa sumber daya Anda tidak pernah melebihi ukuran ini.

> **Note**
>
> Saat Anda menggunakan penyekalaan prediktif, Anda dapat memilih secara opsional perilaku kapasitas maksimum yang berbeda untuk digunakan berdasarkan kapasitas prakiraan. Pengaturan ini berada di bagian Pengaturan penyekalaan prediktif.

### Metrik khusus

AWS Auto Scaling menyediakan metrik yang paling umum digunakan untuk penskalaan otomatis. Namun, tergantung pada kebutuhan Anda, Anda mungkin lebih memilih untuk mendapatkan data dari metrik yang berbeda daripada metrik di konsol. Amazon CloudWatch memiliki banyak metrik berbeda untuk dipilih. CloudWatch juga memungkinkan Anda mempublikasikan metrik Anda sendiri.

Anda menggunakan JSON untuk menentukan metrik CloudWatch kustom. Sebelum Anda mengikuti petunjuk ini, kami sarankan agar Anda terbiasa dengan Panduan CloudWatch Pengguna Amazon.

Untuk menentukan metrik khusus, Anda membuat kapasitas muatan berformat JSON menggunakan serangkaian parameter yang diperlukan dari templat. Anda menambahkan nilai untuk setiap parameter dari CloudWatch. Kami menyediakan template sebagai bagian dari opsi khusus untuk metrik Penskalaan dan metrik Muat di pengaturan lanjutan rencana penskalaan Anda.

JSON mewakili data dalam dua cara:

- Objek, yang merupakan kumpulan pasangan nilai nama yang tidak dipesan. Objek ditentukan di dalam kurung kiri ({) dan kanan (}). Setiap pasangan nilai dimulai dengan nama, diikuti dengan titik dua, diikuti dengan nilai. Pasangan nilai-nama dipisahkan dengan koma.
- Sebuah rangkaian, yang merupakan kumpulan nilai yang dipesan. Sebuah rangkaian didefinisikan di dalam bracket kiri ([) dan kanan (]). Item dalam rangkaian dipisahkan dengan koma.

Berikut adalah contoh templat JSON dengan nilai sampel untuk setiap parameter:

```
{
  "MetricName": "MyBackendCPU",
  "Namespace": "MyNamespace",
  "Dimensions": [
    {
      "Name": "MyOptionalMetricDimensionName",
```

```
        "Value": "MyOptionalMetricDimensionValue"
      }
    ],
    "Statistic": "Sum"
  }
```

Untuk informasi selengkapnya, lihat Spesifikasi metrik penskalaan yang disesuaikan dan Spesifikasi metrik beban yang disesuaikan di Referensi AWS Auto Scaling API.

## Pengaturan penyekalaan dinamis

Gunakan prosedur ini untuk melihat dan menyesuaikan pengaturan untuk kebijakan penskalaan pelacakan target yang AWS Auto Scaling dibuat.

Untuk melihat dan menyesuaikan pengaturan untuk penyekalaan dinamis

1. Untuk memperluas tampilan dan melihat detail Pengaturan penyekalaan dinamis, pilih panah di sebelah kiri judul bagian.
2. Anda dapat membuat pilihan untuk item berikut. Namun, pengaturan default dapat digunakan untuk tutorial ini.

   - Ganti kebijakan penyekalaan eksternal—Jika pengaturan ini dihapus, kebijakan penskalaan yang ada tetap dibuat dari luar rencana penyekalaan ini, dan tidak membuat yang baru.
   - Nonaktifkan scale-in—Jika pengaturan ini dibersihkan, skala otomatis untuk mengurangi kapasitas sumber daya saat ini diperbolehkan ketika metrik yang ditentukan berada di bawah nilai target.
   - Jeda pakai—Menciptakan periode jeda pakai skala masuk dan skala keluar. Periode jeda pakai adalah jumlah waktu saat kebijakan penyekalaan menunggu aktivitas penskalaan sebelumnya diterapkan. Untuk informasi lebih lanjut, lihat Periode jeda pakai dalam Panduan Pengguna Auto Scaling Aplikasi. (Pengaturan ini tidak ditampilkan jika sumber daya adalah grup Auto Scaling.)
   - Pemanasan instans — [Hanya grup Auto Scaling] Mengontrol jumlah waktu yang berlalu sebelum instance yang baru diluncurkan mulai berkontribusi pada metrik. CloudWatch Untuk informasi lebih lanjut, lihat Pemanasan instans dalam Panduan Pengguna Amazon EC2 Auto Scaling.

# Pengaturan penyekalaan prediktif

Jika sumber daya Anda adalah grup Auto Scaling, gunakan prosedur ini untuk melihat dan menyesuaikan pengaturan yang AWS Auto Scaling digunakan untuk penskalaan prediktif.

Untuk melihat dan menyesuaikan pengaturan untuk penskalaan prediktif

1. Untuk memperluas tampilan dan melihat detail Pengaturan penyekalaan prediktif, pilih panah di sebelah kiri judul bagian.
2. Anda dapat membuat pilihan untuk item berikut. Untuk tutorial ini, ubah mode penskalaan Prediktif menjadi Forecast saja.

   - Mode penyekalaan prediktif—Tentukan mode penyekalaan. Standarnya adalah Prakiraan dan skala. Jika Anda mengubahnya menjadi Hanya prakiraan, rencana penyekalaan memprediksi kapasitas masa depan tetapi tidak menerapkan tindakan penyekalaan.
   - Contoh-contoh sebelum peluncuran—Menyesuaikan tindakan penyekalaan untuk dijalankan lebih awal saat menyekalakan keluar. Misalnya, prakiraan mengatakan untuk menambah kapasitas pada pukul 10.00 pagi, dan waktu penyangga adalah 5 menit (300 detik). Waktu pengoperasian dari tindakan penyekalaan yang terkait adalah pukul 9.55 pagi. Hal ini berguna untuk grup Auto Scaling, yang dapat memakan waktu beberapa menit dari waktu peluncuran instans hingga masuk layanan. Waktu yang sebenarnya dapat bervariasi karena bergantung pada beberapa faktor, seperti ukuran contoh dan apakah ada naskah startup yang harus diselesaikan. Nilai default adalah 300 detik.
   - Perilaku kapasitas maksimal—Mengendalikan apakah sumber daya yang dipilih dapat meningkat di atas kapasitas maksimum ketika kapasitas prakiraan mendekati atau melebihi kapasitas maksimum yang ditentukan saat ini. Standarnya adalah Menegakkan pengaturan kapasitas maksimum.
     - Menerapkan pengaturan kapasitas maksimum —AWS Auto Scaling tidak dapat menskalakan kapasitas sumber daya lebih tinggi dari kapasitas maksimum. Kapasitas maksimum diterapkan sebagai batas keras.
     - Mengatur kapasitas maksimum ke kapasitas perkiraan yang sama —AWS Auto Scaling dapat menskalakan kapasitas sumber daya lebih tinggi dari kapasitas maksimum untuk sama tetapi tidak melebihi kapasitas perkiraan.
     - Meningkatkan kapasitas maksimum di atas kapasitas perkiraan —AWS Auto Scaling dapat menskalakan kapasitas sumber daya lebih tinggi dari kapasitas maksimum dengan nilai

buffer yang ditentukan. Tujuannya adalah untuk memberikan kapasitas tambahan kebijakan penyekalaan target jika lalu lintas tidak terduga terjadi.

- Penyangga perilaku kapasitas maks—Jika Anda memilih Meningkatkan kapasitas maksimum di atas kapasitas prakiraan, pilih ukuran buffer kapasitas untuk digunakan saat kapasitas prakiraan hampir atau melebihi kapasitas maksimum. Nilai ditentukan sebagai persentase terkait dengan kapasitas prakiraan. Misalnya, dengan penyangga 10 persen, jika kapasitas prakiraan adalah 50, dan kapasitas maksimum adalah 40, maka kapasitas maksimum yang efektif adalah 55.

3. Setelah selesai menyesuaikan pengaturan, pilih Selanjutnya.

> **Note**
>
> Untuk mengembalikan perubahan, pilih sumber daya dan pilih Kembali ke versi asli. Ini akan mereset sumber daya yang dipilih ke status yang terakhir diketahui di dalam rencana penyekalaan.

# Langkah 4: Buat rencana penskalaan Anda

Di Tinjau dan buat , tinjau detail rencana penyekalaan Anda dan pilih Buat rencana penyekalaan. Anda diarahkan ke halaman yang menunjukkan status rencana penyekalaan Anda. Rencana penyekalaan dapat memakan waktu hingga selesai dibuat sementara sumber daya Anda diperbarui.

Dengan penskalaan prediktif, AWS Auto Scaling menganalisis riwayat metrik beban yang ditentukan dari 14 hari terakhir (minimal 24 jam data diperlukan) untuk menghasilkan perkiraan selama dua hari ke depan. Kemudian, jadwalkan tindakan penyekalaan untuk menyesuaikan kapasitas sumber daya agar sesuai dengan prakiraan untuk setiap jam dalam periode prakiraan.

Setelah pembuatan rencana penyekalaan selesai, lihat perincian rencana penyekalaan dengan memilih namanya dari layar Rencana penyekalaan.

## (Opsional) Lihat informasi penyekalaan untuk sumber daya

Gunakan prosedur ini untuk melihat informasi penyekalaan yang dibuat untuk sumber daya.

Data disajikan dengan cara berikut:

- Grafik yang menunjukkan data riwayat metrik terbaru dari CloudWatch.

- Grafik penskalaan prediktif yang menunjukkan prakiraan beban dan perkiraan kapasitas berdasarkan data dari. AWS Auto Scaling
- Tabel yang mencantumkan semua tindakan penyekalaan prediktif yang dijadwalkan untuk sumber daya.

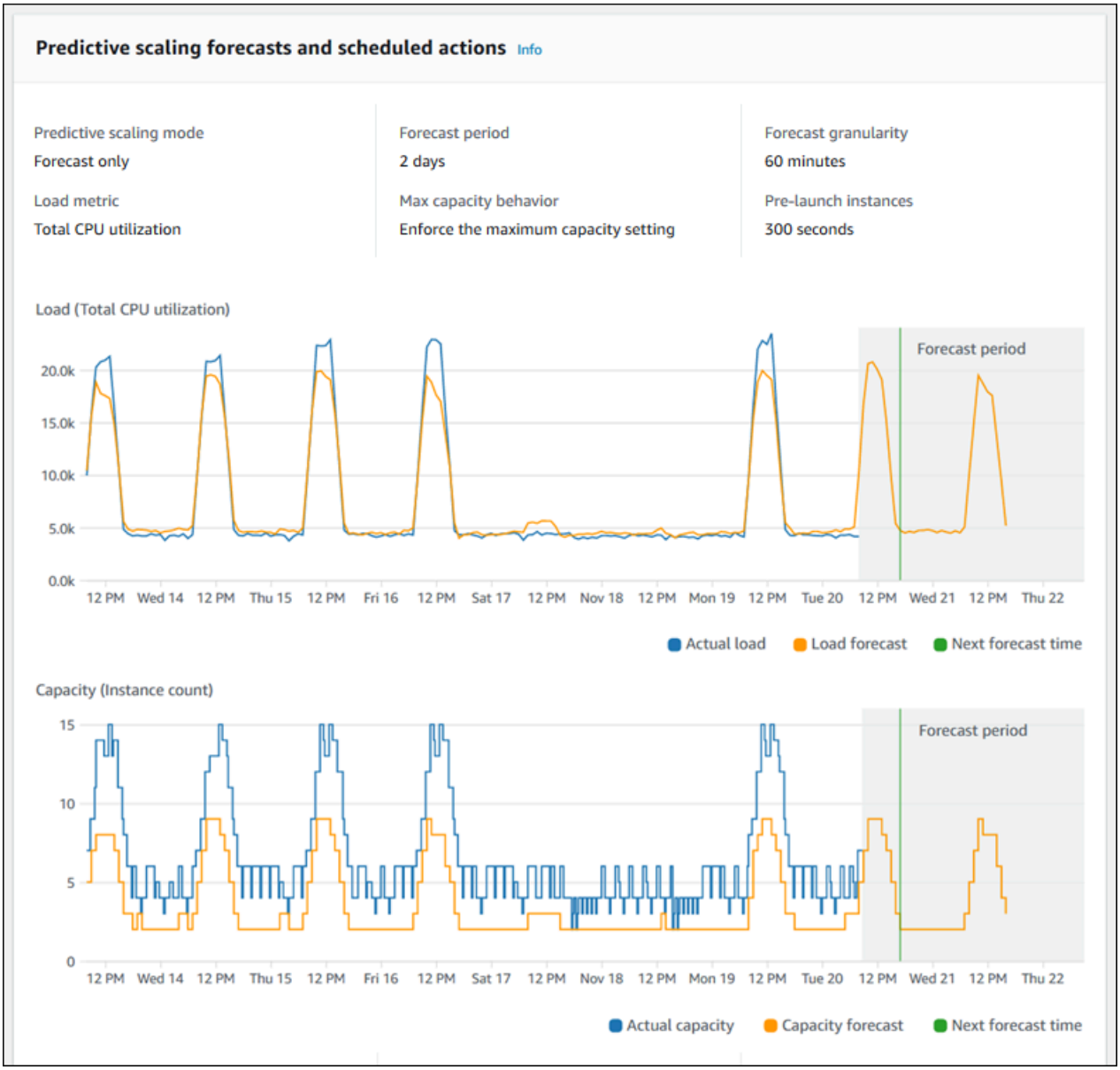
Untuk melihat informasi penyekalaan untuk sumber daya

1. Buka AWS Auto Scaling konsol di https://console.aws.amazon.com/awsautoscaling/.
2. Di Rencana penyekalaan, pilih rencana penyekalaan.
3. Di Perincian rencana penyekalaan, pilih sumber daya untuk dilihat.

## Memantau dan mengevaluasi prakiraan

Ketika rencana penyekalaan Anda sedang berjalan, Anda dapat memantau prakiraan beban, prakiraan kapasitas, dan tindakan penyekalaan untuk memeriksa kinerja penyekalaan prediktif. Semua data ini tersedia di AWS Auto Scaling konsol untuk semua grup Auto Scaling yang diaktifkan untuk penskalaan prediktif. Harap diingat bahwa rencana penyekalaan Anda memerlukan setidaknya 24 jam data beban historis untuk membuat prakiraan awal.

Dalam contoh berikut, sisi kiri setiap grafik menunjukkan pola historis. Sisi kanan menunjukkan prakiraan yang dihasilkan oleh rencana penyekalaan untuk periode prakiraan. Nilai aktual dan prakiraan (biru dan oranye) diplot.




AWS Auto Scaling belajar dari data Anda secara otomatis. Pertama, ini membuat prakiraan beban. Kemudian, perhitungan prakiraan kapasitas menentukan jumlah minimum instans yang diperlukan untuk mendukung aplikasi. Berdasarkan perkiraan kapasitas, AWS Auto Scaling menjadwalkan tindakan penskalaan yang menskalakan grup Auto Scaling sebelum perubahan beban yang diprediksi. Jika penyekalaan dinamis diaktifkan (disarankan), kelompok Auto Scaling dapat menyekalakan kapasitas tambahan (atau menghapus kapasitas) berdasarkan pemanfaatan kelompok contoh saat ini.

Saat mengevaluasi seberapa baik kinerja penyekalaan prediktif, pantau seberapa sesuai nilai aktual dan prakiraan dari waktu ke waktu. Saat Anda membuat rencana penskalaan, AWS Auto Scaling berikan grafik berdasarkan data aktual terbaru. Ini juga memberikan prakiraan awal untuk

48 jam berikutnya. Namun, ketika rencana penyekalaan dibuat, sangat sedikit data prakiraan untuk membandingkan data sebenarnya. Tunggu sampai rencana penyekalaan telah memperoleh nilai perkiraan untuk beberapa periode sebelum membandingkan nilai prakiraan historis dengan nilai aktual. Setelah beberapa hari prakiraan harian, Anda akan memiliki sampel nilai prakiraan yang lebih besar untuk dibandingkan dengan nilai aktual.

Untuk pola yang terjadi setiap hari, interval waktu antara membuat rencana penyekalaan Anda dan mengevaluasi efektivitas prakiraan dapat sesingkat beberapa hari. Namun, lama waktu ini tidak cukup untuk mengevaluasi prakiraan berdasarkan perubahan pola terbaru. Misalnya, katakanlah Anda sedang melihat prakiraan untuk sebuah kelompok Auto Scaling yang memulai kampanye pemasaran baru minggu lalu. Kampanye ini secara signifikan meningkatkan lalu lintas web Anda selama dua hari yang sama setiap minggunya. Dalam situasi seperti ini, kami merekomendasikan menunggu grup mengumpulkan satu atau dua minggu penuh dari data baru sebelum mengevaluasi efektivitas prakiraan. Rekomendasi yang sama berlaku untuk grup Auto Scaling baru yang baru mulai mengumpulkan data metrik.

Jika nilai aktual dan prakiraan tidak cocok setelah pemantauan terhadap durasi yang sesuai, Anda juga harus mempertimbangkan pilihan metrik beban. Agar efektif, metrik muatan harus menunjukkan pengukuran yang andal dan akurat dari total beban pada semua kasus dalam kelompok Auto Scaling. Metrik beban adalah inti untuk skala prediktif. Jika Anda memilih metrik beban non-optimal, itu dapat mencegah skala prediktif dari membuat prakiraan beban dan kapasitas yang akurat dan menjadwalkan penyesuaian kapasitas yang tepat untuk grup Auto Scaling Anda.

# Langkah 5: Bersihkan

Setelah menyelesaikan tutorial memulai, Anda dapat memilih untuk menyimpan rencana penyekalaan Anda. Namun, jika Anda tidak aktif menggunakan rencana penyekalaan, Anda harus mempertimbangkan untuk menghapusnya sehingga akun Anda tidak menimbulkan biaya yang tidak perlu.

Menghapus rencana penskalaan akan menghapus kebijakan penskalaan pelacakan target, CloudWatch alarm terkait, dan tindakan penskalaan prediktif yang dibuat atas nama Anda. AWS Auto Scaling

Menghapus rencana penskalaan tidak akan menghapus CloudFormation tumpukan, grup Auto Scaling, atau sumber daya lain yang dapat diskalakan.

Untuk menghapus rencana penyekalaan

1. Buka AWS Auto Scaling konsol di https://console.aws.amazon.com/awsautoscaling/.
2. Di Rencana penyekalaan, pilih rencana penyekalaan yang Anda buat untuk tutorial ini dan pilih Hapus.
3. Saat diminta konfirmasi, pilih Hapus.

Setelah Anda menghapus rencana penyekalaan, sumber daya Anda tidak kembali ke kapasitas semula. Misalnya, jika kelompok Auto Scaling Anda diskalakan menjadi 10 kasus ketika Anda menghapus rencana skala, kelompok Anda masih diskalakan ke 10 kasus setelah rencana penyekalaan dihapus. Anda dapat memperbarui kapasitas sumber daya tertentu dengan mengakses konsol untuk setiap layanan individu.

## Hapus grup Auto Scaling Anda

Untuk mencegah akun Anda memperoleh biaya Amazon EC2, Anda juga harus menghapus grup Auto Scaling yang Anda buat untuk tutorial ini.

Untuk step-by-step petunjuknya, lihat Menghapus grup Auto Scaling di Panduan Pengguna Amazon EC2 Auto Scaling.

# Langkah 6: Langkah selanjutnya

Sekarang setelah Anda membiasakan diri dengan rencana penskalaan dan beberapa fitur-fiturnya, Anda mungkin ingin mencoba membuat templat rencana penskalaan Anda sendiri. CloudFormation

CloudFormation Template adalah file teks berformat JSON atau YAML yang menjelaskan infrastruktur Amazon Web Services yang diperlukan untuk menjalankan aplikasi atau layanan bersama dengan interkoneksi antar komponen infrastruktur. Dengan CloudFormation, Anda menerapkan dan mengelola kumpulan sumber daya terkait sebagai tumpukan. CloudFormation tersedia tanpa biaya tambahan, dan Anda hanya membayar untuk AWS sumber daya yang dibutuhkan untuk menjalankan aplikasi Anda. Sumber daya dapat terdiri dari AWS sumber daya apa pun yang Anda tentukan dalam template. Untuk informasi selengkapnya, lihat Cara CloudFormation kerja di Panduan AWS CloudFormation Pengguna.

Dalam Panduan AWS CloudFormation Pengguna, kami menyediakan template sederhana untuk membantu Anda memulai. Template sampel tersedia sebagai contoh di AWS::AutoScalingPlans::ScalingPlanbagian dokumentasi referensi CloudFormation template.

Templat sampel membuat rencana penyekalaan untuk satu kelompok Auto Scaling dan memungkinkan penyekalaan prediktif dan penyekalaan dinamis.

Untuk informasi selengkapnya, lihat Memulai dengan CloudFormation dalam Panduan Pengguna AWS CloudFormation .

# Migrasikan rencana penskalaan Anda

Anda dapat bermigrasi dari paket penskalaan ke Amazon EC2 Auto Scaling dan kebijakan penskalaan Application Auto Scaling.

Proses migrasi

- Langkah 1: Tinjau pengaturan yang ada
- Langkah 2: Buat kebijakan penskalaan prediktif
- Langkah 3: Tinjau prakiraan yang dihasilkan oleh kebijakan penskalaan prediktif
- Langkah 4: Bersiaplah untuk menghapus rencana penskalaan
- Langkah 5: Hapus rencana penskalaan
- Langkah 6: Aktifkan kembali penskalaan dinamis
- Langkah 7: Aktifkan kembali penskalaan prediktif
- Referensi Penskalaan Otomatis Amazon EC2 untuk memigrasi kebijakan penskalaan pelacakan target
- Referensi Application Auto Scaling untuk memigrasi kebijakan penskalaan pelacakan target
- Informasi tambahan

> **Important**
>
> Untuk memigrasikan rencana penskalaan, Anda harus menyelesaikan beberapa langkah dalam urutan yang tepat. Saat Anda memigrasikan rencana penskalaan Anda, jangan perbarui, karena itu merusak urutan operasi dan dapat menyebabkan perilaku yang tidak diinginkan.

## Langkah 1: Tinjau pengaturan yang ada

Untuk menentukan pengaturan penskalaan mana yang harus Anda pindahkan, gunakan describe-scaling-plansperintah.

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

Catat item yang ingin Anda pertahankan dari rencana penskalaan yang ada, yang dapat mencakup hal-hal berikut:

- **`MinCapacity`**— Kapasitas minimum sumber daya yang dapat diskalakan.
- **`MaxCapacity`**— Kapasitas maksimum sumber daya yang dapat diskalakan.
- **`PredefinedLoadMetricType`**— Metrik beban untuk penskalaan prediktif.
- **`PredefinedScalingMetricType`**— Metrik penskalaan untuk penskalaan pelacakan target (dinamis) dan penskalaan prediktif.
- **`TargetValue`**— Nilai target untuk metrik penskalaan.

### Perbedaan antara rencana penskalaan dan kebijakan penskalaan

Ada beberapa perbedaan penting antara rencana penskalaan dan kebijakan penskalaan:

- Kebijakan penskalaan hanya dapat mengaktifkan satu jenis penskalaan: penskalaan pelacakan target atau penskalaan prediktif. Untuk menggunakan kedua metode penskalaan, Anda harus membuat kebijakan terpisah.
- Demikian juga, Anda harus menentukan metrik penskalaan untuk penskalaan prediktif dan metrik penskalaan untuk penskalaan pelacakan target secara terpisah dalam kebijakan masing-masing.

## Langkah 2: Buat kebijakan penskalaan prediktif

Jika Anda tidak menggunakan penskalaan prediktif, lewati ke depan. Langkah 4: Bersiaplah untuk menghapus rencana penskalaan

Untuk menyediakan waktu untuk mengevaluasi perkiraan, sebaiknya Anda membuat kebijakan penskalaan prediktif sebelum kebijakan penskalaan lainnya.

Untuk setiap grup Auto Scaling dengan spesifikasi metrik beban yang ada, lakukan hal berikut untuk mengubahnya menjadi kebijakan penskalaan prediktif berbasis Amazon EC2 Auto Scaling.

Untuk membuat kebijakan penskalaan prediktif

1. Dalam file JSON, tentukan `MetricSpecifications` struktur seperti yang ditunjukkan pada contoh berikut:

```
{
```

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

2. Dalam `MetricSpecifications` struktur, untuk setiap metrik pemuatan dalam rencana penskalaan Anda, buat `PredefinedLoadMetricSpecification` atau `CustomizedLoadMetricSpecification` gunakan pengaturan setara dari rencana penskalaan.

   Berikut ini adalah contoh struktur bagian metrik beban.

   With predefined metrics

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

   Untuk informasi lebih lanjut, lihat PredictiveScalingPredefinedLoadMetric dalam Referensi API Amazon EC2 Auto Scaling.

   With custom metrics

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

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

Untuk informasi lebih lanjut, lihat PredictiveScalingCustomizedLoadMetric dalam Referensi API Amazon EC2 Auto Scaling.

3. Tambahkan spesifikasi metrik penskalaan ke `MetricSpecifications` dan tentukan nilai target.

   Berikut ini adalah contoh struktur metrik penskalaan dan bagian nilai target.

   With predefined metrics

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

Untuk informasi lebih lanjut, lihat PredictiveScalingPredefinedScalingMetric dalam Referensi API Amazon EC2 Auto Scaling.

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
      }
    ],
    ...
}
```

Untuk informasi lebih lanjut, lihat PredictiveScalingCustomizedScalingMetric dalam Referensi API Amazon EC2 Auto Scaling.

4. Untuk memperkirakan saja, tambahkan properti `Mode` dengan nilai`ForecastOnly`. Setelah Anda selesai memigrasikan penskalaan prediktif dan memastikan bahwa prakiraan akurat dan andal, Anda dapat mengubah mode untuk memungkinkan penskalaan. Untuk informasi selengkapnya, lihat Langkah 7: Aktifkan kembali penskalaan prediktif.

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

Untuk informasi lebih lanjut, lihat PredictiveScalingConfiguration dalam Referensi API Amazon EC2 Auto Scaling.

5. Jika **`ScheduledActionBufferTime`** properti ada dalam rencana penskalaan Anda, salin nilainya ke `SchedulingBufferTime` properti dalam kebijakan penskalaan prediktif Anda.

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

Untuk informasi lebih lanjut, lihat PredictiveScalingConfiguration dalam Referensi API Amazon EC2 Auto Scaling.

6. Jika **`PredictiveScalingMaxCapacityBuffer`** properti **`PredictiveScalingMaxCapacityBehavior`** dan properti ada dalam rencana penskalaan Anda, maka Anda dapat mengonfigurasi `MaxCapacityBreachBehavior` dan `MaxCapacityBuffer` properti dalam kebijakan penskalaan prediktif Anda. Properti ini menentukan apa yang harus terjadi jika kapasitas perkiraan mendekati atau melebihi kapasitas maksimum yang ditentukan untuk grup Auto Scaling.

   > **Warning**
   >
   > Jika Anda menyetel `MaxCapacityBreachBehavior` properti ke`IncreaseMaxCapacity`, maka lebih banyak instance dapat diluncurkan daripada yang dimaksudkan kecuali Anda memantau dan mengelola peningkatan kapasitas maksimum. Kapasitas maksimum yang meningkat menjadi kapasitas maksimum normal baru untuk grup Auto Scaling hingga Anda memperbaruinya secara manual. Kapasitas maksimum tidak secara otomatis berkurang kembali ke maksimum semula.

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

   Untuk informasi lebih lanjut, lihat PredictiveScalingConfiguration dalam Referensi API Amazon EC2 Auto Scaling.

7. Simpan file JSON dengan nama unik. Catat nama file. Anda memerlukannya di langkah berikutnya dan lagi di akhir prosedur migrasi saat Anda mengaktifkan kembali kebijakan penskalaan prediktif Anda. Untuk informasi selengkapnya, lihat Langkah 7: Aktifkan kembali penskalaan prediktif.

8. Setelah Anda menyimpan file JSON Anda, jalankan put-scaling-policyperintah. Dalam contoh berikut, ganti masing-masing *user input placeholder* dengan informasi Anda sendiri.

```
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
```

   Jika berhasil, perintah ini mengembalikan Amazon Resource Name (ARN) kebijakan.

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

9. Ulangi langkah-langkah ini untuk setiap spesifikasi metrik pemuatan yang Anda migrasi ke kebijakan penskalaan prediktif berbasis Amazon EC2 Auto Scaling.

## Langkah 3: Tinjau prakiraan yang dihasilkan oleh kebijakan penskalaan prediktif

Jika Anda tidak menggunakan penskalaan prediktif, lewati prosedur berikut.

Prakiraan tersedia segera setelah Anda membuat kebijakan penskalaan prediktif. Setelah Amazon EC2 Auto Scaling menghasilkan perkiraan, Anda dapat meninjau perkiraan kebijakan tersebut melalui konsol Amazon EC2 Auto Scaling dan menyesuaikannya seperlunya.

Untuk meninjau perkiraan untuk kebijakan penskalaan prediktif

1. Buka konsol Amazon EC2 di. https://console.aws.amazon.com/ec2/
2. Di panel navigasi, pilih Grup Auto Scaling, lalu pilih nama grup Auto Scaling Anda dari daftar.
3. Pada tab Penskalaan otomatis, di Kebijakan penskalaan prediktif, pilih kebijakan Anda.
4. Di bagian Pemantauan, Anda dapat melihat perkiraan masa lalu dan masa depan kebijakan Anda untuk beban dan kapasitas terhadap nilai aktual.

   Untuk informasi selengkapnya, lihat Meninjau grafik pemantauan penskalaan prediktif di Panduan Pengguna Amazon EC2 Auto Scaling.

5. Ulangi langkah-langkah ini untuk setiap kebijakan penskalaan prediktif yang Anda buat.

## Langkah 4: Bersiaplah untuk menghapus rencana penskalaan

Untuk sumber daya apa pun dengan konfigurasi penskalaan pelacakan target yang ada, lakukan hal berikut untuk mengumpulkan informasi tambahan apa pun yang Anda perlukan dari rencana penskalaan sebelum menghapusnya.

Untuk menjelaskan informasi kebijakan penskalaan dari rencana penskalaan, gunakan perintah. describe-scaling-plan-resources Dalam contoh perintah berikut, ganti *my-scaling-plan* dengan informasi Anda sendiri.

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

Tinjau output dan konfirmasikan bahwa Anda ingin memigrasikan kebijakan penskalaan yang dijelaskan. Gunakan informasi ini untuk membuat kebijakan penskalaan pelacakan target berbasis Amazon EC2 Auto Scaling dan Application Auto Scaling baru. Langkah 6: Aktifkan kembali penskalaan dinamis

## Langkah 5: Hapus rencana penskalaan

Sebelum membuat kebijakan penskalaan pelacakan target baru, Anda harus menghapus rencana penskalaan untuk menghapus kebijakan penskalaan yang dibuatnya.

Untuk menghapus rencana penskalaan Anda, gunakan delete-scaling-planperintah. Dalam contoh perintah berikut, ganti *my-scaling-plan* dengan informasi Anda sendiri.

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

Setelah Anda menghapus rencana penskalaan, penskalaan dinamis dinonaktifkan. Jadi, jika ada lonjakan lalu lintas atau beban kerja yang tiba-tiba, kapasitas yang tersedia untuk setiap sumber daya yang dapat diskalakan tidak akan meningkat dengan sendirinya. Sebagai tindakan pencegahan, Anda mungkin ingin meningkatkan kapasitas sumber daya Anda yang dapat diskalakan secara manual dalam jangka pendek.

Untuk meningkatkan kapasitas grup Auto Scaling

1. Buka konsol Amazon EC2 di. https://console.aws.amazon.com/ec2/
2. Di panel navigasi, pilih Grup Auto Scaling, lalu pilih nama grup Auto Scaling Anda dari daftar.
3. Pada tab Detail, pilih Detail grup, Edit.
4. Untuk kapasitas yang diinginkan, tingkatkan kapasitas yang diinginkan.
5. Setelah selesai, pilih Perbarui.

Untuk menambahkan Replika Aurora ke klaster DB

1. Buka konsol Amazon RDS di https://console.aws.amazon.com/rds/.
2. Di panel navigasi, pilih Databases, lalu pilih cluster DB Anda.
3. Pastikan klaster dan instans primer berada dalam status Tersedia.
4. Pilih Tindakan, Tambahkan pembaca.
5. Pada halaman Tambah pembaca, tentukan opsi untuk replika Aurora baru Anda.
6. Pilih Tambahkan pembaca.

Untuk meningkatkan kapasitas baca dan tulis yang disediakan dari tabel DynamoDB atau indeks sekunder global

1. Buka konsol DynamoDB di. https://console.aws.amazon.com/dynamodb/
2. Di panel navigasi, pilih Tabel, lalu pilih nama tabel Anda dari daftar.
3. Pada tab Pengaturan tambahan, pilih Kapasitas baca/tulis, Edit.
4. Pada halaman Edit read/write kapasitas, untuk kapasitas Baca, Unit kapasitas yang disediakan, tingkatkan kapasitas baca tabel yang disediakan.
5. (Opsional) Jika Anda ingin indeks sekunder global Anda menggunakan pengaturan kapasitas baca yang sama dengan tabel dasar, lalu pilih kotak centang Gunakan pengaturan kapasitas baca yang sama untuk semua indeks sekunder global.
6. Untuk kapasitas Tulis, Unit kapasitas yang disediakan, tingkatkan kapasitas tulis tabel yang disediakan.
7. (Opsional) Jika Anda ingin indeks sekunder global Anda menggunakan pengaturan kapasitas tulis yang sama dengan tabel dasar, lalu pilih Gunakan pengaturan kapasitas tulis yang sama untuk semua indeks sekunder global kotak centang.

8. Jika Anda tidak memilih kotak centang pada langkah 5 atau 7, gulir ke bawah halaman untuk memperbarui kapasitas baca dan tulis dari indeks sekunder global apa pun.
9. Pilih Simpan perubahan untuk melanjutkan.

Untuk meningkatkan jumlah tugas yang berjalan untuk layanan Amazon ECS Anda

1. Buka konsol di https://console.aws.amazon.com/ecs/v2.
2. Di panel navigasi, pilih Cluster, lalu pilih nama cluster Anda dari daftar.
3. Di bagian Layanan, pilih kotak centang di sebelah layanan, lalu pilih Perbarui.
4. Untuk tugas yang diinginkan, masukkan jumlah tugas yang ingin Anda jalankan untuk layanan.
5. Pilih Perbarui.

Untuk meningkatkan kapasitas Armada Spot

1. Buka konsol Amazon EC2 di. https://console.aws.amazon.com/ec2/
2. Di panel navigasi, pilih Permintaan Spot, lalu pilih permintaan Armada Spot Anda.
3. Pilih Tindakan, Modifikasi kapasitas target.
4. Di Modify target capacity, masukkan kapasitas target baru dan porsi On-Demand Instance.
5. Pilih Kirim.

# Langkah 6: Aktifkan kembali penskalaan dinamis

Aktifkan kembali penskalaan dinamis dengan membuat kebijakan penskalaan pelacakan target.

Saat membuat kebijakan penskalaan pelacakan target untuk grup Auto Scaling, Anda menambahkannya langsung ke grup. Saat membuat kebijakan penskalaan pelacakan target untuk sumber daya terukur lainnya, pertama-tama Anda mendaftarkan sumber daya sebagai target yang dapat diskalakan, lalu menambahkan kebijakan penskalaan pelacakan target ke target yang dapat diskalakan.

Topik

- Membuat kebijakan penskalaan pelacakan target untuk grup Auto Scaling
- Buat kebijakan penskalaan pelacakan target untuk sumber daya lain yang dapat diskalakan

# Membuat kebijakan penskalaan pelacakan target untuk grup Auto Scaling

Untuk membuat kebijakan penskalaan pelacakan target untuk grup Auto Scaling

1. Dalam file JSON, buat `PredefinedMetricSpecification` atau `CustomizedMetricSpecification` gunakan pengaturan yang setara dari rencana penskalaan.

   Berikut ini adalah contoh konfigurasi pelacakan target. Dalam contoh ini, ganti masing-masing *`user input placeholder`* dengan informasi Anda sendiri.

   With predefined metrics

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

   Untuk informasi lebih lanjut, lihat PredefinedMetricSpecification dalam Referensi API Amazon EC2 Auto Scaling.

   With custom metrics

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

Untuk informasi lebih lanjut, lihat CustomizedMetricSpecification dalam Referensi API Amazon EC2 Auto Scaling.

2. Untuk membuat kebijakan penskalaan Anda, gunakan put-scaling-policyperintah, bersama dengan file JSON yang Anda buat pada langkah sebelumnya. Dalam contoh berikut, ganti masing-masing *user input placeholder* dengan informasi Anda sendiri.

```
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. Ulangi proses ini untuk setiap kebijakan penskalaan berbasis rencana penskalaan yang Anda migrasi ke kebijakan penskalaan pelacakan target berbasis Amazon EC2 Auto Scaling.

## Buat kebijakan penskalaan pelacakan target untuk sumber daya lain yang dapat diskalakan

Selanjutnya, buat kebijakan penskalaan pelacakan target untuk sumber daya terukur lainnya dengan melakukan tugas konfigurasi berikut.

- Daftarkan target yang dapat diskalakan untuk penskalaan otomatis dengan layanan Application Auto Scaling.
- Tambahkan kebijakan penskalaan pelacakan target pada target yang dapat diskalakan.

Untuk membuat kebijakan penskalaan pelacakan target untuk sumber daya lain yang dapat diskalakan

1. Gunakan register-scalable-targetperintah untuk mendaftarkan sumber daya sebagai target yang dapat diskalakan dan tentukan batas penskalaan untuk kebijakan penskalaan.

   Dalam contoh berikut, ganti masing-masing *user input placeholder* dengan informasi Anda sendiri. Untuk opsi perintah, berikan informasi berikut:

   - `--service-namespace`— Namespace untuk layanan target (misalnya,**`ecs`**). Untuk mendapatkan ruang nama layanan, lihat referensi. RegisterScalableTarget

- `--scalable-dimension`— Dimensi skalabel yang terkait dengan sumber daya target (misalnya,**`ecs:service:DesiredCount`**). Untuk mendapatkan dimensi yang dapat diskalakan, lihat RegisterScalableTargetreferensi.
- `--resource-id`— ID sumber daya untuk sumber daya target (misalnya,**`service/my-cluster/my-service`**). Untuk informasi tentang sintaks dan contoh sumber daya tertentu IDs, lihat RegisterScalableTargetreferensi.

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

Jika berhasil, perintah ini mengembalikan ARN dari target yang dapat diskalakan.

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

2. Dalam file JSON, buat `PredefinedMetricSpecification` atau `CustomizedMetricSpecification` gunakan pengaturan yang setara dari rencana penskalaan.

   Berikut ini adalah contoh konfigurasi pelacakan target.

   With predefined metrics

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

   Untuk informasi lebih lanjut, lihat PredefinedMetricSpecification dalam Referensi API Application Auto Scaling.

With custom metrics

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

Untuk informasi lebih lanjut, lihat CustomizedMetricSpecification dalam Referensi API Application Auto Scaling.

3. Untuk membuat kebijakan penskalaan Anda, gunakan put-scaling-policyperintah, bersama dengan file JSON yang Anda buat pada langkah sebelumnya.

```
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. Ulangi proses ini untuk setiap kebijakan penskalaan berbasis rencana penskalaan yang Anda migrasi ke kebijakan penskalaan pelacakan target berbasis Penskalaan Otomatis Aplikasi.

## Langkah 7: Aktifkan kembali penskalaan prediktif

Jika Anda tidak menggunakan penskalaan prediktif, lewati langkah ini.

Aktifkan kembali penskalaan prediktif dengan mengalihkan penskalaan prediktif ke perkiraan dan skala.

Untuk membuat perubahan ini, perbarui file JSON yang Anda buat Langkah 2: Buat kebijakan penskalaan prediktif dan ubah nilai `Mode` opsi menjadi `ForecastAndScale` seperti pada contoh berikut:

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

Kemudian, perbarui setiap kebijakan penskalaan prediktif dengan perintah. put-scaling-policy Dalam contoh ini, ganti masing-masing *user input placeholder* dengan informasi Anda sendiri.

```
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
```

Atau, Anda dapat membuat perubahan ini dari konsol Amazon EC2 Auto Scaling dengan mengaktifkan pengaturan Skala berdasarkan perkiraan. Untuk informasi selengkapnya, lihat Penskalaan prediktif untuk Penskalaan Otomatis Amazon EC2 di Panduan Pengguna Penskalaan Otomatis Amazon EC2.

## Referensi Penskalaan Otomatis Amazon EC2 untuk memigrasi kebijakan penskalaan pelacakan target

Untuk tujuan referensi, tabel berikut mencantumkan semua properti konfigurasi pelacakan target dalam rencana penskalaan dengan properti yang sesuai dalam operasi Amazon EC2 Auto `PutScalingPolicy` Scaling API.

| Properti sumber rencana penskalaan | Properti target Amazon EC2 Auto Scaling |
|---|---|
| `PolicyName` | `PolicyName` |
| `PolicyType` | `PolicyType` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Name` | `TargetTrackingConfiguration.CustomizedMetricSpecification.Dimensions.Name` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Value` | `TargetTrackingConfiguration.CustomizedMetricSpecification.Dimensions.Value` |

| Properti sumber rencana penskalaan | Properti target Amazon EC2 Auto Scaling |
| --- | --- |
| TargetTrackingConfiguration .CustomizedScalingMetricSpecification.MetricName | TargetTrackingConfiguration .CustomizedMetricSpecification.MetricName |
| TargetTrackingConfiguration .CustomizedScalingMetricSpecification.Namespace | TargetTrackingConfiguration .CustomizedMetricSpecification.Namespace |
| TargetTrackingConfiguration .CustomizedScalingMetricSpecification.Statistic | TargetTrackingConfiguration .CustomizedMetricSpecification.Statistic |
| TargetTrackingConfiguration .CustomizedScalingMetricSpecification.Unit | TargetTrackingConfiguration .CustomizedMetricSpecification.Unit |
| 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 |

| Properti sumber rencana penskalaan | Properti target Amazon EC2 Auto Scaling |
| --- | --- |
| `TargetTrackingConfiguration.TargetValue` | `TargetTrackingConfiguration.TargetValue` |

[1] Instance warmup adalah fitur untuk grup Auto Scaling yang membantu memastikan bahwa instans yang baru diluncurkan siap menerima lalu lintas sebelum menyumbangkan data penggunaannya ke metrik penskalaan. Saat instans masih memanas, Amazon EC2 Auto Scaling memperlambat proses penambahan atau penghapusan instans ke grup. Alih-alih menentukan waktu pemanasan untuk kebijakan penskalaan, sebaiknya gunakan setelan pemanasan instans default grup Auto Scaling untuk memastikan bahwa semua peluncuran instance menggunakan waktu pemanasan instance yang sama. Untuk informasi selengkapnya, lihat Menyetel pemanasan instans default untuk grup Auto Scaling di Panduan Pengguna Amazon EC2 Auto Scaling.

## Referensi Application Auto Scaling untuk memigrasi kebijakan penskalaan pelacakan target

Untuk tujuan referensi, tabel berikut mencantumkan semua properti konfigurasi pelacakan target dalam rencana penskalaan dengan properti yang sesuai dalam operasi Application Auto `PutScalingPolicy` Scaling API.

| Properti sumber rencana penskalaan | Properti target Application Auto Scaling |
| --- | --- |
| `PolicyName` | `PolicyName` |
| `PolicyType` | `PolicyType` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Name` | `TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Dimensions.Name` |
| `TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Value` | `TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Dimensions.Value` |

| Properti sumber rencana penskalaan | Properti target Application Auto Scaling |
| --- | --- |
| TargetTrackingConfiguration.CustomizedScalingMetricSpecification.MetricName | TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.MetricName |
| TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Namespace | TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Namespace |
| TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Statistic | TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Statistic |
| TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Unit | TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Unit |
| 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 |

| Properti sumber rencana penskalaan | Properti target Application Auto Scaling |
| --- | --- |
| `TargetTrackingConfiguration.TargetValue` | `TargetTrackingScalingPolicyConfiguration.TargetValue` |

[1] Application Auto Scaling menggunakan periode cooldown untuk memperlambat penskalaan saat sumber daya Anda yang dapat diskalakan keluar (meningkatkan kapasitas) dan penskalaan (mengurangi kapasitas). Untuk informasi selengkapnya, lihat Menentukan periode cooldown di Panduan Pengguna Application Auto Scaling.

## Informasi tambahan

Untuk mempelajari cara membuat kebijakan penskalaan prediktif baru dari konsol, lihat topik berikut:

- Penskalaan Otomatis Amazon EC2 — Buat kebijakan penskalaan prediktif di Panduan Pengguna Penskalaan Otomatis Amazon EC2.

Untuk mempelajari cara membuat kebijakan penskalaan pelacakan target baru menggunakan konsol, lihat topik berikut:

- Amazon Aurora — Menggunakan Auto Scaling Amazon Aurora dengan Replika Aurora di Panduan Pengguna Amazon RDS.
- DynamoDB - Menggunakan penskalaan otomatis with Konsol Manajemen AWS DynamoDB di Panduan Pengembang Amazon DynamoDB.
- Penskalaan Otomatis Amazon EC2 — Buat kebijakan penskalaan pelacakan target di Panduan Pengguna Penskalaan Otomatis Amazon EC2.
- Amazon ECS — Memperbarui layanan menggunakan konsol di Panduan Pengembang Layanan Amazon Elastic Container.
- Armada Spot — Skala Armada Spot menggunakan kebijakan pelacakan target di Panduan Pengguna Amazon EC2.

# Pencatatan Panggilan AWS Auto Scaling API dengan AWS CloudTrail

AWS Auto Scaling terintegrasi dengan AWS CloudTrail, layanan yang menyediakan catatan tindakan yang diambil oleh pengguna, peran, atau AWS layanan di AWS Auto Scaling. CloudTrail menangkap semua panggilan API untuk AWS Auto Scaling sebagai peristiwa. Panggilan yang diambil termasuk panggilan dari AWS Auto Scaling konsol dan panggilan kode ke AWS Auto Scaling API. Jika Anda membuat jejak, Anda dapat mengaktifkan pengiriman CloudTrail acara secara berkelanjutan ke bucket Amazon S3, termasuk acara untuk. AWS Auto Scaling Jika Anda tidak mengonfigurasi jejak, Anda masih dapat melihat peristiwa terbaru di CloudTrail konsol dalam Riwayat acara. Dengan menggunakan informasi yang dikumpulkan oleh CloudTrail, Anda dapat menentukan permintaan yang dibuat AWS Auto Scaling, alamat IP dari mana permintaan dibuat, siapa yang membuat permintaan, kapan dibuat, dan detail tambahan.

Untuk mempelajari selengkapnya CloudTrail, lihat Panduan AWS CloudTrail Pengguna.

## AWS Auto Scaling Informasi di CloudTrail

CloudTrail diaktifkan di AWS akun Anda saat Anda membuat akun. Ketika AWS Auto Scaling aktivitas terjadi, aktivitas tersebut dicatat dalam suatu CloudTrail peristiwa bersama dengan peristiwa AWS layanan lainnya dalam riwayat Acara. Anda dapat melihat, mencari, dan mengunduh acara terbaru di AWS akun Anda. Untuk informasi selengkapnya, lihat Melihat Acara dengan Riwayat CloudTrail Acara.

Untuk catatan peristiwa yang sedang berlangsung di AWS akun Anda, termasuk acara untuk AWS Auto Scaling, buat jejak. Jejak memungkinkan CloudTrail untuk mengirimkan file log ke bucket Amazon S3. Secara default, saat Anda membuat jejak di konsol, jejak tersebut berlaku untuk semua Region AWS. Jejak mencatat peristiwa dari semua Wilayah di AWS partisi dan mengirimkan file log ke bucket Amazon S3 yang Anda tentukan. Selain itu, Anda dapat mengonfigurasi Amazon Web Services lainnya untuk menganalisis lebih lanjut dan menindaklanjuti data peristiwa yang dikumpulkan dalam CloudTrail log. Untuk informasi selengkapnya, lihat berikut:

- Gambaran Umum untuk Membuat Jejak
- CloudTrail Layanan dan Integrasi yang Didukung
- Mengkonfigurasi Notifikasi Amazon SNS untuk CloudTrail

- Menerima File CloudTrail Log dari Beberapa Wilayah dan Menerima File CloudTrail Log dari Beberapa Akun

Semua AWS Auto Scaling tindakan dicatat oleh CloudTrail dan didokumentasikan dalam Referensi AWS Auto Scaling API. Misalnya, panggilan ke`CreateScalingPlan`,`DeleteScalingPlan`, dan `DescribeScalingPlans` tindakan menghasilkan entri dalam file CloudTrail log.

Setiap entri peristiwa atau log berisi informasi tentang entitas yang membuat permintaan tersebut. Informasi identitas membantu Anda menentukan hal berikut ini:

- Apakah permintaan itu dibuat dengan kredenal pengguna root atau AWS Identity and Access Management (IAM).
- Apakah permintaan tersebut dibuat dengan kredensial keamanan sementara untuk satu peran atau pengguna gabungan.
- Apakah permintaan itu dibuat oleh AWS layanan lain.

Untuk informasi lain, lihat Elemen userIdentity CloudTrail .

## Memahami Entri File AWS Auto Scaling Log

Trail adalah konfigurasi yang memungkinkan pengiriman peristiwa sebagai file log ke bucket Amazon S3 yang Anda tentukan. CloudTrail file log berisi satu atau lebih entri log. Peristiwa mewakili permintaan tunggal dari sumber manapun dan mencakup informasi tentang tindakan yang diminta, tanggal dan waktu tindakan, parameter permintaan, dan sebagainya. CloudTrail file log bukanlah jejak tumpukan yang diurutkan dari panggilan API publik, jadi file tersebut tidak muncul dalam urutan tertentu.

Contoh berikut menunjukkan entri CloudTrail log yang menunjukkan `CreateScalingPlan` tindakan.

```
{
    "eventVersion": "1.05",
    "userIdentity": {
        "type": "Root",
        "principalId": "123456789012",
        "arn": "arn:aws:iam::123456789012:root",
        "accountId": "123456789012",
        "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
        "sessionContext": {
            "attributes": {
```

```
                "mfaAuthenticated": "false",
                "creationDate": "2018-08-21T17:05:42Z"
            }
        }
    },
    "eventTime": "2018-08-01T23:17:19Z",
    "eventSource": "autoscaling.amazonaws.com",
    "eventName": "CreateScalingPlan",
    "awsRegion": "us-west-2",
    "sourceIPAddress": "72.21.196.68",
    "userAgent": "aws-internal/3",
    "requestParameters": {
        "applicationSource": {
            "tagFilters": [
                {
                    "key": "TagText",
                    "values": [
                        "MyApplication"
                    ]
                }
            ]
        },
        "scalingInstructions": [
            {
                "resourceId": "autoScalingGroup/MyAutoScalingGroup",
                "targetTrackingConfigurations": [
                    {
                        "predefinedScalingMetricSpecification": {
                            "predefinedScalingMetricType": "ASGAverageCPUUtilization"
                        },
                        "targetValue": 40
                    }
                ],
                "maxCapacity": 10,
                "serviceNamespace": "autoscaling",
                "scalableDimension": "autoscaling:autoScalingGroup:DesiredCapacity",
                "minCapacity": 1
            }
        ],
        "scalingPlanName": "MyScalingPlan"
    },
    "responseElements": {
        "scalingPlanVersion": 1
    },
```

```
    "additionalEventData": {
        "service": "autoscaling-plans"
    },
    "requestID": "0737e2ea-fb2d-11e3-bfd8-99133058e7bb",
    "eventID": "3fcfb182-98f8-4744-bd45-b38835ab61cb",
    "eventType": "AwsApiCall",
    "recipientAccountId": "123456789012"
}
```

# Keamanan rencana penskalaan

Keamanan cloud di AWS adalah prioritas tertinggi. Sebagai AWS pelanggan, Anda mendapat manfaat dari pusat data dan arsitektur jaringan yang dibangun untuk memenuhi persyaratan organisasi yang paling sensitif terhadap keamanan.

Keamanan adalah tanggung jawab bersama antara Anda AWS dan Anda. Model tanggung jawab bersama menjelaskan hal ini sebagai keamanan dari cloud dan keamanan dalam cloud:

- Keamanan cloud — AWS bertanggung jawab untuk melindungi infrastruktur yang menjalankan AWS layanan di AWS Cloud. AWS juga memberi Anda layanan yang dapat Anda gunakan dengan aman. Third-party Auditor secara teratur menguji dan memverifikasi efektivitas keamanan kami sebagai bagian dari program AWS kepatuhan program AWS . Untuk mempelajari tentang program kepatuhan yang berlaku AWS Auto Scaling, lihat AWS layanan dalam cakupan berdasarkan AWS layanan program kepatuhan .
- Keamanan di cloud — Tanggung jawab Anda ditentukan oleh AWS layanan yang Anda gunakan. Anda juga bertanggung jawab atas faktor lain, termasuk sensitivitas data Anda, persyaratan perusahaan Anda, serta undang-undang dan peraturan yang berlaku.

Dokumentasi ini membantu Anda memahami cara menerapkan model tanggung jawab bersama saat menggunakan rencana penskalaan, dan juga membantu Anda memahami cara mengelola akses ke rencana penskalaan.

Topik

- Akses rencana penskalaan menggunakan titik akhir VPC antarmuka
- Perlindungan data untuk rencana penskalaan
- Manajemen identitas dan akses untuk rencana penskalaan
- Validasi kepatuhan untuk rencana penskalaan
- Keamanan infrastruktur untuk rencana penskalaan

## Akses rencana penskalaan menggunakan titik akhir VPC antarmuka

Anda dapat menggunakan AWS PrivateLink untuk membuat koneksi pribadi antara VPC Anda dan. AWS Auto Scaling Anda dapat mengakses AWS Auto Scaling seolah-olah itu ada di VPC Anda,

tanpa menggunakan gateway internet, perangkat NAT, koneksi VPN, atau koneksi. Direct Connect Instans di VPC Anda tidak memerlukan alamat IP publik untuk mengakses. AWS Auto Scaling

Anda membuat koneksi pribadi ini dengan membuat titik akhir antarmuka, didukung oleh AWS PrivateLink. Kami membuat antarmuka jaringan endpoint di setiap subnet yang Anda aktifkan untuk titik akhir antarmuka. Ini adalah antarmuka jaringan yang dikelola pemohon yang berfungsi sebagai titik masuk untuk lalu lintas yang ditakdirkan. AWS Auto Scaling

Untuk informasi selengkapnya, lihat Akses Layanan AWS melalui AWS PrivateLink di AWS PrivateLink Panduan.

Topik

- Buat titik akhir VPC antarmuka untuk rencana penskalaan
- Membuat kebijakan titik akhir VPC untuk rencana penskalaan
- Migrasi titik akhir

## Buat titik akhir VPC antarmuka untuk rencana penskalaan

Buat titik akhir untuk rencana AWS Auto Scaling penskalaan menggunakan nama layanan berikut:

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

Untuk informasi selengkapnya, lihat Mengakses AWS layanan menggunakan titik akhir VPC antarmuka di Panduan.AWS PrivateLink

Anda tidak perlu mengubah pengaturan lainnya. AWS Auto Scaling API memanggil lainnya Layanan AWS menggunakan titik akhir layanan atau titik akhir VPC antarmuka pribadi, mana pun yang digunakan.

## Membuat kebijakan titik akhir VPC untuk rencana penskalaan

Anda dapat melampirkan kebijakan ke titik akhir VPC untuk mengontrol akses ke API. AWS Auto Scaling Kebijakan menentukan:

- Prinsipal yang dapat melakukan tindakan.
- Tindakan yang dapat dilakukan.
- Sumber daya tempat tindakan dapat dilakukan.

Contoh berikut menunjukkan kebijakan titik akhir VPC yang menolak izin setiap orang untuk menghapus rencana penskalaan melalui titik akhir. Kebijakan contoh juga memberikan izin kepada semua orang untuk melakukan semua tindakan lainnya.

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

Untuk informasi selengkapnya, lihat kebijakan titik akhir VPC di Panduan.AWS PrivateLink

## Migrasi titik akhir

Pada 22 November 2019, kami memperkenalkan `autoscaling-plans.region.amazonaws.com` nama host DNS default baru dan titik akhir untuk panggilan ke API. AWS Auto Scaling Endpoint baru ini kompatibel dengan rilis terbaru AWS CLI dan SDK. Jika Anda belum melakukannya, instal yang terbaru AWS CLI dan SDK untuk menggunakan titik akhir baru. Untuk memperbarui AWS CLI, lihat Menginstal atau memperbarui AWS CLI di Panduan AWS Command Line Interface Pengguna. Untuk informasi tentang AWS SDK, lihat Alat untuk Amazon Web Services.

> **Important**
>
> Untuk kompatibilitas mundur, `autoscaling.region.amazonaws.com` titik akhir yang ada akan terus didukung untuk panggilan ke API. AWS Auto Scaling Untuk mengatur `autoscaling.region.amazonaws.com` titik akhir sebagai titik akhir VPC antarmuka pribadi, lihat Amazon EC2 Auto Scaling dan antarmuka titik akhir VPC dalam Panduan Pengguna Amazon EC2 Auto Scaling.

Titik Akhir untuk Panggilan Saat Menggunakan CLI atau API AWS Auto Scaling

Untuk rilis saat ini AWS Auto Scaling, panggilan Anda ke AWS Auto Scaling API secara otomatis menuju ke `autoscaling-plans.region.amazonaws.com` titik akhir, bukan. `autoscaling.region.amazonaws.com`

Anda dapat menyebut titik akhir baru di CLI dengan menggunakan parameter berikut dengan setiap perintah untuk menentukan titik akhir: `--endpoint-url https://autoscaling-plans.region.amazonaws.com`.

Meskipun tidak disarankan, Anda juga dapat memanggil titik akhir lama di CLI dengan menggunakan parameter berikut dengan setiap perintah untuk menentukan titik akhir: `--endpoint-url https://autoscaling.region.amazonaws.com`.

Untuk berbagai SDK yang digunakan untuk memanggil API, lihat dokumentasi untuk SDK yang diinginkan guna mempelajari cara mengarahkan permintaan ke titik akhir tertentu. Untuk informasi lebih lanjut, lihat Alat untuk Amazon Web Services.

# Perlindungan data untuk rencana penskalaan

Model tanggung jawab AWS bersama model berlaku untuk perlindungan data di AWS Auto Scaling. Seperti yang dijelaskan dalam model AWS ini, bertanggung jawab untuk melindungi infrastruktur global yang menjalankan semua AWS Cloud. Anda bertanggung jawab untuk mempertahankan kendali atas konten yang di-host pada infrastruktur ini. Anda juga bertanggung jawab atas tugas-tugas konfigurasi dan manajemen keamanan untuk Layanan AWS yang Anda gunakan. Untuk informasi selengkapnya tentang privasi data, lihat FAQ Privasi Data AWS . Untuk informasi tentang perlindungan data di Eropa, lihat Pusat Peraturan Perlindungan Data Umum (GDPR).

Untuk tujuan perlindungan data, kami menyarankan Anda melindungi Akun AWS kredensil dan mengatur pengguna individu dengan AWS IAM Identity Center atau AWS Identity and Access Management (IAM). Dengan cara itu, setiap pengguna hanya diberi izin yang diperlukan untuk memenuhi tanggung jawab tugasnya. Kami juga menyarankan supaya Anda mengamankan data dengan cara-cara berikut:

- Gunakan autentikasi multi-faktor (MFA) pada setiap akun.
- Gunakan SSL/TLS untuk berkomunikasi dengan AWS sumber daya. Kami mensyaratkan TLS 1.2 dan menganjurkan TLS 1.3.

- Siapkan API dan logging aktivitas pengguna dengan AWS CloudTrail. Untuk informasi tentang penggunaan CloudTrail jejak untuk menangkap AWS aktivitas, lihat Bekerja dengan CloudTrail jejak di AWS CloudTrail Panduan Pengguna.
- Gunakan solusi AWS enkripsi, bersama dengan semua kontrol keamanan default di dalamnya Layanan AWS.
- Gunakan layanan keamanan terkelola tingkat lanjut seperti Amazon Macie, yang membantu menemukan dan mengamankan data sensitif yang disimpan di Amazon S3.
- Jika Anda memerlukan modul kriptografi tervalidasi FIPS 140-3 saat mengakses AWS melalui antarmuka baris perintah atau API, gunakan titik akhir FIPS. Lihat informasi selengkapnya tentang titik akhir FIPS yang tersedia di Standar Pemrosesan Informasi Federal (FIPS) 140-3.

Kami sangat merekomendasikan agar Anda tidak pernah memasukkan informasi identifikasi yang sensitif, seperti nomor rekening pelanggan Anda, ke dalam tanda atau bidang isian bebas seperti bidang Nama. Ini termasuk saat Anda bekerja dengan AWS Auto Scaling atau lainnya Layanan AWS menggunakan konsol, API AWS CLI, atau AWS SDK. Data apa pun yang Anda masukkan ke dalam tanda atau bidang isian bebas yang digunakan untuk nama dapat digunakan untuk log penagihan atau log diagnostik. Saat Anda memberikan URL ke server eksternal, kami sangat menganjurkan supaya Anda tidak menyertakan informasi kredensial di dalam URL untuk memvalidasi permintaan Anda ke server itu.

# Manajemen identitas dan akses untuk rencana penskalaan

AWS Identity and Access Management (IAM) adalah Layanan AWS yang membantu administrator mengontrol akses ke AWS sumber daya dengan aman. Administrator IAM mengontrol siapa yang dapat diautentikasi (masuk) dan diberi wewenang (memiliki izin) untuk menggunakan sumber daya. AWS Auto Scaling IAM adalah Layanan AWS yang dapat Anda gunakan tanpa biaya tambahan.

Untuk dokumentasi lengkap IAM, lihat Panduan Pengguna IAM.

## Kontrol akses

Anda dapat memiliki kredensi yang valid untuk mengautentikasi permintaan Anda, tetapi kecuali Anda memiliki izin, Anda tidak dapat membuat atau mengakses rencana penskalaan. Misalnya, Anda harus memiliki izin untuk membuat rencana penyekalaan, mengkonfigurasi penyekalaan prediktif, dan sebagainya.

Bagian berikut memberikan rincian tentang bagaimana administrator IAM dapat menggunakan IAM untuk membantu mengamankan rencana penskalaan Anda, dengan mengontrol siapa yang dapat bekerja dengan rencana penskalaan.

Topik



## Bagaimana rencana penskalaan bekerja dengan IAM

Sebelum Anda menggunakan IAM untuk mengelola siapa yang dapat membuat, mengakses, dan mengelola rencana AWS Auto Scaling penskalaan, Anda harus memahami fitur IAM apa yang tersedia untuk digunakan dengan rencana penskalaan.

Topik



### Identity-based kebijakan

Dengan kebijakan berbasis identitas IAM, Anda dapat menentukan tindakan dan sumber daya yang diizinkan atau ditolak, dan kondisi di mana tindakan tersebut diperbolehkan atau ditolak. Rencana penskalaan mendukung tindakan, sumber daya, dan kunci kondisi tertentu. Untuk mempelajari semua elemen yang Anda gunakan dalam kebijakan JSON, lihat Referensi elemen kebijakan IAM JSON dalam Panduan Pengguna IAM.

Tindakan

Administrator dapat menggunakan kebijakan AWS JSON untuk menentukan siapa yang memiliki akses ke apa. Yaitu, di mana utama dapat melakukan tindakan pada sumber daya, dan dalam kondisi apa.

Elemen `Action` dari kebijakan JSON menjelaskan tindakan yang dapat Anda gunakan untuk mengizinkan atau menolak akses dalam sebuah kebijakan. Sertakan tindakan dalam kebijakan untuk memberikan izin untuk melakukan operasi terkait.

Tindakan rencana penskalaan dalam pernyataan kebijakan IAM menggunakan awalan berikut sebelum tindakan:. `autoscaling-plans:` Pernyataan kebijakan harus memuat elemen `Action` atau `NotAction`. Rencana penskalaan memiliki serangkaian tindakan sendiri yang menggambarkan tugas yang dapat Anda lakukan dengan layanan ini.

Untuk menetapkan beberapa tindakan dalam satu pernyataan, pisahkan dengan koma seperti yang ditunjukkan dalam contoh berikut.

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

Anda dapat menentukan beberapa tindakan menggunakan wildcard (*). Misalnya, untuk menentukan semua tindakan yang dimulai dengan kata `Describe`, sertakan tindakan berikut.

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

Untuk melihat daftar lengkap tindakan rencana penskalaan yang dapat digunakan dalam pernyataan kebijakan, lihat Kunci tindakan, sumber daya, dan kondisi AWS Auto Scaling di Referensi Otorisasi Layanan.

Sumber daya

Elemen `Resource` menentukan objek di mana tindakan berlaku.

Rencana penskalaan tidak memiliki sumber daya yang ditentukan layanan yang dapat digunakan sebagai `Resource` elemen pernyataan kebijakan IAM. Oleh karena itu, tidak ada Nama Sumber Daya Amazon (ARN) untuk Anda gunakan dalam kebijakan IAM. Untuk mengontrol akses ke tindakan rencana penskalaan, selalu gunakan * (tanda bintang) sebagai sumber daya saat menulis kebijakan IAM.

Kunci syarat

Elemen `Condition` (atau blok `Condition`) memungkinkan Anda menentukan ketentuan yang mengizinkan Anda untuk menerapkan pernyataan. Misalnya, Anda mungkin ingin kebijakan diterapkan hanya setelah tanggal tertentu. Untuk menyatakan kondisi, gunakan kunci kondisi yang telah ditentukan sebelumnya.

Paket penskalaan tidak menyediakan kunci kondisi khusus layanan apa pun, tetapi mereka mendukung penggunaan beberapa kunci kondisi global. Untuk melihat semua kunci kondisi AWS global, lihat kunci konteks kondisi AWS global di Panduan Pengguna IAM.

Elemen `Condition` bersifat opsional.

Contoh

Untuk melihat contoh kebijakan berbasis identitas untuk rencana penskalaan, lihat. Identity-based contoh kebijakan untuk rencana penskalaan

## Resource-based kebijakan

Amazon Web Services lainnya, seperti Amazon Simple Storage Service, mendukung kebijakan izin berbasis sumber daya. Misalnya, Anda dapat melampirkan kebijakan izin ke bucket S3 untuk mengelola izin akses ke bucket tersebut.

Rencana penskalaan tidak mendukung kebijakan berbasis sumber daya.

## Daftar Kontrol Akses (ACL)

Paket penskalaan tidak mendukung Daftar Kontrol Akses (ACL).

## Otorisasi berdasarkan tanda

Rencana penskalaan tidak dapat ditandai. Mereka juga tidak memiliki sumber daya yang ditentukan layanan yang dapat ditandai. Oleh karena itu, mereka tidak mendukung pengendalian akses berdasarkan tag pada sumber daya.

Paket penskalaan mungkin berisi sumber daya yang dapat diberi tag, seperti grup Auto Scaling, yang mendukung pengendalian akses berdasarkan tag. Untuk informasi lebih lanjut, lihat dokumentasi untuk itu Layanan AWS.

## Peran IAM

IAM role adalah entitas dalam Akun AWS Anda yang memiliki izin khusus.

Menggunakan kredensial sementara

Anda dapat menggunakan kredensial sementara untuk masuk dengan federasi, untuk memainkan peran IAM, atau untuk mengambil peran lintas akun. Anda memperoleh kredensil keamanan sementara dengan memanggil operasi AWS STS API seperti AssumeRoleatau. GetFederationToken

Dukungan rencana penskalaan menggunakan kredensyal sementara.

Service-linked peran untuk rencana penskalaan

AWS Auto Scaling menggunakan peran terkait layanan untuk izin yang diperlukan untuk memanggil AWS layanan lain atas nama Anda. Service-linked peran membuat pengaturan rencana penskalaan lebih mudah karena Anda tidak perlu menambahkan izin yang diperlukan secara manual. Untuk informasi selengkapnya, lihat Menggunakan peran terkait layanan dalam Panduan Pengguna IAM.

AWS Auto Scaling menggunakan beberapa jenis peran terkait layanan untuk memanggil orang lain Layanan AWS atas nama Anda saat Anda bekerja dengan rencana penskalaan:

- Peran terkait layanan penskalaan prediktif — Memungkinkan AWS Auto Scaling untuk mengakses data metrik historis dari. CloudWatch Juga memungkinkan pembuatan tindakan terjadwal untuk grup Auto Scaling berdasarkan perkiraan beban dan prediksi kapasitas. Untuk informasi selengkapnya, lihat Peran terkait layanan penskalaan prediktif.
- Peran terkait layanan Amazon EC2 Auto Scaling AWS Auto Scaling — Memungkinkan untuk mengakses dan mengelola kebijakan penskalaan pelacakan target untuk grup Auto Scaling. Untuk informasi selengkapnya, lihat Service-linked peran untuk Amazon EC2 Auto Scaling di Panduan Pengguna Amazon EC2 Auto Scaling.
- Peran terkait layanan Application Auto Scaling — Memungkinkan AWS Auto Scaling untuk mengakses dan mengelola kebijakan penskalaan pelacakan target untuk sumber daya lain yang dapat diskalakan. Ada satu peran terkait layanan untuk setiap layanan. Untuk informasi selengkapnya, lihat Service-linked peran untuk Application Auto Scaling di Panduan Pengguna Application Auto Scaling.

Anda dapat menggunakan prosedur berikut untuk menentukan apakah akun Anda sudah memiliki peran terkait layanan.

Untuk menentukan apakah peran terkait layanan sudah ada

1. Buka konsol IAM di https://console.aws.amazon.com/iam/.
2. Di panel navigasi, pilih Peran.
3. Cari daftar `AWSServiceRole` untuk menemukan peran terkait layanan yang ada di akun Anda. Cari nama peran terkait layanan yang ingin Anda periksa.

Peran layanan

AWS Auto Scaling tidak memiliki peran layanan untuk rencana penskalaan.

## Peran terkait layanan penskalaan prediktif

AWS Auto Scaling menggunakan peran terkait layanan untuk izin yang diperlukan untuk memanggil orang lain AWS atas nama Anda saat Anda bekerja dengan rencana penskalaan. Untuk informasi selengkapnya, lihat Service-linked peran untuk rencana penskalaan.

Bagian berikut menjelaskan cara membuat dan mengelola peran terkait layanan untuk penskalaan prediktif. Mulai dengan mengonfigurasikan izin untuk memungkinkan entitas IAM (seperti pengguna, grup, atau peran) membuat, mengedit, atau menghapus peran yang ditautkan dengan layanan.

### Izin yang diberikan oleh peran tertaut layanan

AWS Auto Scaling menggunakan peran terkait layanan yang diberi nama AWSServiceRoleForAutoScalingPlans_EC2AutoScaling untuk memanggil AWS layanan lain atas nama Anda saat Anda mengaktifkan penskalaan prediktif.

AWSServiceRoleForAutoScalingPlans_EC2AutoScalingmempercayai `autoscaling-plans.amazonaws.com` layanan untuk mengambil peran.

Peran terkait layanan ini menggunakan kebijakan terkelola. AWSAutoScalingPlansEC2AutoScalingPolicy Untuk melihat izin kebijakan ini, lihat AWSAutoScalingPlansEC2AutoScalingPolicydi Referensi Kebijakan AWS Terkelola.

### Buat peran yang berkaitan dengan layanan (otomatis)

Anda tidak perlu membuat AWSServiceRoleForAutoScalingPlans_EC2AutoScalingperan secara manual. AWS membuat peran ini untuk Anda saat Anda membuat rencana penskalaan di akun Anda dan mengaktifkan penskalaan prediktif.

AWS Untuk membuat peran terkait layanan atas nama Anda, Anda harus memiliki izin yang diperlukan. Untuk informasi selengkapnya, lihat izin Service-linked peran di Panduan Pengguna IAM.

### Buat peran yang berkaitan dengan layanan (manual)

Untuk membuat peran terkait layanan secara manual, Anda dapat menggunakan konsol IAM, IAM CLI, atau IAM API. Untuk informasi selengkapnya, lihat Membuat peran terkait layanan di Panduan Pengguna IAM.

Untuk membuat peran terkait layanan (AWS CLI)

Gunakan perintah create-service-linked-role berikut untuk membuat peran terkait layanan.

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

### Mengedit peran terkait layanan

Anda dapat mengedit deskripsi AWSServiceRoleForAutoScalingPlans_EC2AutoScalingmenggunakan IAM. Untuk informasi selengkapnya, lihat Mengedit deskripsi peran terkait layanan di Panduan Pengguna IAM.

### Menghapus peran terkait layanan

Jika Anda tidak perlu lagi menggunakan rencana penskalaan, kami sarankan Anda menghapusnya AWSServiceRoleForAutoScalingPlans_EC2AutoScaling.

Anda dapat menghapus peran yang ditautkan layanan hanya setelah menghapus semua rencana penskalaan Akun AWS yang mengaktifkan penskalaan prediktif. Ini memastikan bahwa Anda tidak dapat secara tidak sengaja menghapus izin untuk mengakses paket penskalaan Anda.

Anda dapat menggunakan konsol IAM, IAM CLI, atau IAM API untuk menghapus peran terkait layanan. Untuk informasi selengkapnya, lihat Menghapus peran terkait layanan di Panduan Pengguna IAM.

Setelah menghapus peran yang AWSServiceRoleForAutoScalingPlans_EC2AutoScaling ditautkan layanan, AWS Auto Scaling buat kembali peran tersebut jika Anda membuat rencana penskalaan dengan penskalaan prediktif diaktifkan.

### Wilayah yang Didukung

AWS Auto Scaling mendukung penggunaan peran terkait layanan di semua tempat rencana Region AWS penskalaan tersedia. Untuk informasi tentang ketersediaan Regional dari rencana penskalaan, lihat AWS Auto Scaling titik akhir dan kuota di. Referensi Umum AWS

## Identity-based contoh kebijakan untuk rencana penskalaan

Secara default, pengguna IAM baru tidak memiliki izin untuk melakukan apa pun. Administrator IAM harus membuat dan menetapkan kebijakan IAM yang memberikan izin identitas IAM (seperti pengguna atau peran) untuk bekerja dengan rencana penskalaan.

Untuk mempelajari cara membuat kebijakan IAM dengan menggunakan contoh dokumen kebijakan JSON ini, lihat Membuat kebijakan di tab JSON dalam Panduan Pengguna IAM.

Topik

- Praktik terbaik kebijakan
- Perbolehkan pengguna membuat rencana penyekalaan
- Memungkinkan pengguna untuk mengaktifkan penyekalaan prediktif
- Izin tambahan yang diperlukan
- Izin yang diperlukan untuk membuat peran yang berkaitan dengan layanan

## Praktik terbaik kebijakan

Identity-based kebijakan menentukan apakah seseorang dapat membuat, mengakses, atau menghapus AWS Auto Scaling sumber daya di akun Anda. Tindakan ini membuat Akun AWS Anda dikenai biaya. Ketika Anda membuat atau mengedit kebijakan berbasis identitas, ikuti panduan dan rekomendasi ini:

- Mulailah dengan kebijakan AWS terkelola dan beralih ke izin hak istimewa paling sedikit — Untuk mulai memberikan izin kepada pengguna dan beban kerja Anda, gunakan kebijakan AWS terkelola yang memberikan izin untuk banyak kasus penggunaan umum. Mereka tersedia di Anda Akun AWS. Kami menyarankan Anda mengurangi izin lebih lanjut dengan menentukan kebijakan yang dikelola AWS pelanggan yang khusus untuk kasus penggunaan Anda. Untuk informasi selengkapnya, lihat Kebijakan yang dikelola AWS atau Kebijakan yang dikelola AWS untuk fungsi tugas dalam Panduan Pengguna IAM.
- Menerapkan izin dengan hak akses paling rendah – Ketika Anda menetapkan izin dengan kebijakan IAM, hanya berikan izin yang diperlukan untuk melakukan tugas. Anda melakukannya dengan mendefinisikan tindakan yang dapat diambil pada sumber daya tertentu dalam kondisi tertentu, yang juga dikenal sebagai izin dengan hak akses paling rendah. Untuk informasi selengkapnya tentang cara menggunakan IAM untuk mengajukan izin, lihat Kebijakan dan izin dalam IAM dalam Panduan Pengguna IAM.
- Gunakan kondisi dalam kebijakan IAM untuk membatasi akses lebih lanjut – Anda dapat menambahkan suatu kondisi ke kebijakan Anda untuk membatasi akses ke tindakan dan sumber daya. Sebagai contoh, Anda dapat menulis kondisi kebijakan untuk menentukan bahwa semua permintaan harus dikirim menggunakan SSL. Anda juga dapat menggunakan ketentuan untuk memberikan akses ke tindakan layanan jika digunakan melalui yang spesifik Layanan AWS, seperti

CloudFormation. Untuk informasi selengkapnya, lihat Elemen kebijakan JSON IAM: Kondisi dalam Panduan Pengguna IAM.

- Gunakan IAM Access Analyzer untuk memvalidasi kebijakan IAM Anda untuk memastikan izin yang aman dan fungsional – IAM Access Analyzer memvalidasi kebijakan baru dan yang sudah ada sehingga kebijakan tersebut mematuhi bahasa kebijakan IAM (JSON) dan praktik terbaik IAM. IAM Access Analyzer menyediakan lebih dari 100 pemeriksaan kebijakan dan rekomendasi yang dapat ditindaklanjuti untuk membantu Anda membuat kebijakan yang aman dan fungsional. Untuk informasi selengkapnya, lihat Validasi kebijakan dengan IAM Access Analyzer dalam Panduan Pengguna IAM.
- Memerlukan otentikasi multi-faktor (MFA) - Jika Anda memiliki skenario yang mengharuskan pengguna IAM atau pengguna root di Anda, Akun AWS aktifkan MFA untuk keamanan tambahan. Untuk meminta MFA ketika operasi API dipanggil, tambahkan kondisi MFA pada kebijakan Anda. Untuk informasi selengkapnya, lihat Amankan akses API dengan MFA dalam Panduan Pengguna IAM.

Untuk informasi selengkapnya tentang praktik terbaik dalam IAM, lihat Praktik terbaik keamanan di IAM dalam Panduan Pengguna IAM.

## Perbolehkan pengguna membuat rencana penyekalaan

Berikut ini menunjukkan contoh kebijakan berbasis identitas yang memberikan izin untuk membuat rencana penskalaan.

JSON

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

```
    ]
}
```

Untuk bekerja dengan rencana penskalaan, pengguna akhir harus memiliki izin tambahan yang memungkinkan mereka bekerja dengan sumber daya tertentu di akun mereka. Izin ini tercantum dalam Izin tambahan yang diperlukan.

Setiap pengguna konsol juga memerlukan izin yang memungkinkan mereka menemukan sumber daya yang dapat diskalakan di akun mereka dan untuk melihat grafik data CloudWatch metrik dari konsol. AWS Auto Scaling Kumpulan izin tambahan yang diperlukan untuk bekerja dengan AWS Auto Scaling konsol tercantum di bawah ini:

- `cloudformation:ListStacks`: Untuk daftar tumpukan.
- `tag:GetTagKeys`: Untuk menemukan sumber daya yang dapat diskalakan yang berisi kunci tag tertentu.
- `tag:GetTagValues`: Untuk menemukan sumber daya yang berisi nilai tag tertentu.
- `autoscaling:DescribeTags`: Untuk menemukan grup Auto Scaling yang berisi tag tertentu.
- `cloudwatch:GetMetricData`: Untuk melihat data dalam grafik metrik.

## Memungkinkan pengguna untuk mengaktifkan penyekalaan prediktif

Berikut ini menunjukkan contoh kebijakan berbasis identitas yang memberikan izin untuk mengaktifkan penskalaan prediktif. Izin ini memperluas fitur rencana penskalaan yang diatur untuk menskalakan grup Auto Scaling.

JSON

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

```
                "autoscaling:BatchDeleteScheduledAction"
            ],
            "Resource": "*"
        }
    ]
}
```

## Izin tambahan yang diperlukan

Agar berhasil mengonfigurasi rencana penskalaan, pengguna akhir harus diberikan izin untuk setiap layanan target yang akan mereka konfigurasi penskalaannya. Untuk memberikan izin minimum yang diperlukan untuk bekerja dengan layanan target, baca informasi di bagian ini dan tentukan tindakan yang relevan dalam `Action` elemen pernyataan kebijakan IAM.

### Grup Auto Scaling

Untuk menambahkan grup Auto Scaling ke rencana penskalaan, pengguna harus memiliki izin berikut dari Amazon EC2 Auto Scaling:

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

### Layanan ECS

Untuk menambahkan layanan ECS ke rencana penyekalaan, pengguna harus memiliki izin berikut dari Amazon ECS dan Auto Scaling Aplikasi:

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

- `application-autoscaling:DeleteScalingPolicy`

Temukan Armada

Untuk menambahkan Armada Spot ke rencana penyekalaan, pengguna harus memiliki izin berikut dari Amazon EC2 dan Auto Scaling Aplikasi:

- `ec2:DescribeSpotFleetRequests`
- `ec2:ModifySpotFleetRequest`
- `application-autoscaling:RegisterScalableTarget`
- `application-autoscaling:DescribeScalableTargets`
- `application-autoscaling:DeregisterScalableTarget`
- `application-autoscaling:PutScalingPolicy`
- `application-autoscaling:DescribeScalingPolicies`
- `application-autoscaling:DeleteScalingPolicy`

Tabel DynamoDB atau indeks global

Untuk menambahkan tabel DynamoDB atau indeks global ke rencana penskalaan, pengguna harus memiliki izin berikut dari DynamoDB dan Application Auto Scaling:

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

klaster Aurora DB

Untuk menambahkan klaster Aurora DB ke rencana penyekalaan, pengguna harus memiliki izin berikut dari Amazon Aurora dan Auto Scaling Aplikasi:

- `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`

### Izin yang diperlukan untuk membuat peran yang berkaitan dengan layanan

AWS Auto Scaling memerlukan izin untuk membuat peran terkait layanan saat pertama kali pengguna Akun AWS membuat rencana penskalaan dengan penskalaan prediktif diaktifkan. Jika peran terkait layanan belum ada, AWS Auto Scaling buat di akun Anda. Peran terkait layanan memberikan izin AWS Auto Scaling agar dapat memanggil layanan lain atas nama Anda.

Agar pembuatan peran otomatis berhasil, pengguna harus memiliki izin untuk tindakan `iam:CreateServiceLinkedRole` nyata.

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

Berikut ini menunjukkan contoh kebijakan berbasis identitas yang memberikan izin untuk membuat peran terkait layanan.

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"
                }
            }
        }
    ]
}
```

Untuk informasi selengkapnya, lihat Peran terkait layanan penskalaan prediktif.

## Validasi kepatuhan untuk rencana penskalaan

Untuk mempelajari apakah an Layanan AWS berada dalam lingkup program kepatuhan tertentu, lihat Layanan AWS di Lingkup oleh Program Kepatuhan Layanan AWS dan pilih program kepatuhan yang Anda minati. Untuk informasi umum, lihat Program AWS Kepatuhan Program AWS .

Anda dapat mengunduh laporan audit pihak ketiga menggunakan AWS Artifact. Untuk informasi selengkapnya, lihat Mengunduh Laporan di AWS Artifact .

Tanggung jawab kepatuhan Anda saat menggunakan Layanan AWS ditentukan oleh sensitivitas data Anda, tujuan kepatuhan perusahaan Anda, dan hukum dan peraturan yang berlaku. Untuk informasi selengkapnya tentang tanggung jawab kepatuhan Anda saat menggunakan Layanan AWS, lihat Dokumentasi AWS Keamanan.

## Keamanan infrastruktur untuk rencana penskalaan

Sebagai layanan terkelola, AWS Auto Scaling dilindungi oleh keamanan jaringan AWS global. Untuk informasi tentang layanan AWS keamanan dan cara AWS melindungi infrastruktur, lihat Keamanan AWS Cloud. Untuk mendesain AWS lingkungan Anda menggunakan praktik terbaik untuk keamanan infrastruktur, lihat Perlindungan Infrastruktur dalam Kerangka Kerja yang AWS Diarsiteksikan dengan Baik Pilar Keamanan.

Anda menggunakan panggilan API yang AWS dipublikasikan untuk mengakses AWS Auto Scaling melalui jaringan. Klien harus mendukung hal-hal berikut:

- Keamanan Lapisan Pengangkutan (TLS). Kami mensyaratkan TLS 1.2 dan menganjurkan TLS 1.3.

- Cipher suite dengan perfect forward secrecy (PFS) seperti DHE (Ephemeral) atau ECDHE (Elliptic Curve Ephemeral Diffie-Hellman). Diffie-Hellman Sebagian besar sistem modern seperti Java 7 dan versi lebih baru mendukung mode-mode ini.

# Kuota untuk rencana penskalaan Anda

Anda Akun AWS memiliki kuota default (sebelumnya disebut sebagai batas) yang terkait dengan rencana penskalaan. Kecuali dinyatakan lain, setiap kuota bersifat khusus per Wilayah. Anda dapat meminta peningkatan untuk beberapa kuota dan kuota lainnya tidak dapat ditingkatkan.

Untuk melihat kuota Application Auto Scaling, buka konsol Service Quotas. Di panel navigasi, pilih Layanan AWSdan pilih AWS Auto Scaling Plans.

Untuk meminta penambahan kuota, lihat Meminta penambahan kuota di Panduan Pengguna Service Quotas.

Anda Akun AWS memiliki kuota berikut yang terkait dengan rencana penskalaan.

| Nama | Default | Dapat disesuaikan |
|---|---|---|
| Sumber daya yang dapat diskalakan per jenis sumber daya | Amazon DynamoDB: 3.000 \| Grup Auto EC2 Scaling Amazon: 200 \| Semua jenis sumber daya lainnya: 500 | Ya |
| Rencana penskalaan | 100 | Ya |
| Instruksi penskalaan per rencana penskalaan | 500 | Tidak |
| Konfigurasi pelacakan target per instruksi penskalaan | 10 | Tidak |

Selalu ingat kuota layanan saat Anda meningkatkan beban kerja Anda. Misalnya, ketika Anda mencapai jumlah maksimum unit kapasitas yang diizinkan oleh layanan, penskalaan akan berhenti. Jika permintaan turun dan kapasitas saat ini menurun, AWS Auto Scaling dapat skala lagi. Untuk menghindari mencapai batas kuota layanan ini lagi, Anda dapat meminta peningkatan. Setiap layanan memiliki kuota default sendiri untuk kapasitas maksimum sumber daya. Untuk informasi tentang kuota default untuk Amazon Web Services lainnya, lihat Titik akhir dan kuota layanan di. Referensi Umum Amazon Web Services

# Riwayat dokumen untuk rencana penskalaan

Tabel berikut menjelaskan penambahan penting pada AWS Auto Scaling dokumentasi. Untuk notifikasi tentang pembaruan-pembaruan dokumentasi ini, Anda dapat berlangganan ke sebuah umpan RSS.

| Perubahan | Deskripsi | Tanggal |
| --- | --- | --- |
| Konten baru untuk migrasi dari AWS Auto Scaling ke opsi alternatif | Anda sekarang dapat bermigrasi dari penskalaan prediktif Amazon EC2 Auto Scaling AWS Auto Scaling ke Amazon EC2, yang menawarkan lebih banyak fungsionalitas. Untuk informasi selengkapnya, lihat Memigrasi paket penskalaan Anda. | April 5, 2024 |
| Konten keamanan baru | Kami merilis chapter Keamanan yang diperbarui. Sebagai bagian dari pembaruan ini, kami mengganti “Otentikasi dan Kontrol Akses” dengan Identitas dan manajemen akses untuk AWS Auto Scaling. | 12 Maret 2020 |
| Dukungan untuk titik akhir VPC Amazon VPC | Anda sekarang dapat membuat koneksi pribadi antara VPC Anda dan. AWS Auto Scaling Untuk pertimbangan dan instruksi migrasi, lihat Penskalaan paket dan titik akhir VPC antarmuka. | 22 November 2019 |

| | | |
|---|---|---|
| Support untuk meningkatkan kapasitas maksimum di atas kapasitas perkiraan | Menambahkan dukungan konsol untuk memungkin kan rencana penyekalaan untuk meningkatkan kapasitas maksimum di atas kapasitas prakiraan sebesar nilai penyangga yang ditentukan. Untuk informasi selengkapnya, lihat Pengaturan penskalaan prediktif. | 9 Maret 2019 |
| Penskalaan dan penyempur naan prediktif | Sekarang Anda dapat menggunakan skala prediktif untuk secara proaktif menyekalakan grup Amazon EC2 Auto Scaling. Rilis ini juga menambahkan dukungan untuk mengganti kebijakan penyekalaan yang dibuat di luar rencana penyekalaan (seperti dari konsol lain) dan mengontrol apakah Anda mengaktifkan fitur penyekala an dinamis dari program Anda. | 20 November 2018 |
| Support untuk pengaturan sumber daya kustom | Menambahkan dukungan untuk menyesuaikan berbagai pengaturan untuk setiap sumber daya individu atau beberapa sumber daya pada waktu yang sama. | 9 Oktober 2018 |
| Tag sebagai sumber aplikasi | Rilis ini menambahkan dukungan untuk menentukan satu set tanda sebagai sumber aplikasi. | 23 April 2018 |

| | | |
|---|---|---|
| Layanan baru | Rilis awal AWS Auto Scaling. | 16 Januari 2018 |

Terjemahan disediakan oleh mesin penerjemah. Jika konten terjemahan yang diberikan bertentangan dengan versi bahasa Inggris aslinya, utamakan versi bahasa Inggris.