AWS Auto Scaling

User Guide

AWS Auto Scaling: User Guide

Table of Contents

What Is AWS Auto Scaling?

AWS Auto Scaling enables you to quickly discover the scalable AWS resources that are part of your application and configure dynamic scaling in a matter of minutes. The AWS Auto Scaling console provides a single user interface to use the automatic scaling features of multiple AWS services. It also offers recommendations to configure scaling for the scalable resources in your application.

For more information about the benefits of this service, see the AWS Auto Scaling FAQs.

Scalable Resources

Use AWS Auto Scaling to automatically scale the following resources that support your application:

- Amazon EC2 Auto Scaling groups
- Aurora DB clusters
- DynamoDB global secondary indexes
- DynamoDB tables
- ECS services
- Spot Fleet requests

How AWS Auto Scaling Works

With AWS Auto Scaling, you create a scaling plan with a set of instructions used to configure dynamic scaling for the scalable resources in your application. AWS Auto Scaling creates target tracking policies for the scalable resources. Target tracking policies add and remove capacity for each resource as required to maintain resource utilization at the specified target value. AWS Auto Scaling offers recommendations for target tracking scaling policies based on the most popular scaling metrics and thresholds used for automatic scaling.

You create one scaling plan per application source (an AWS CloudFormation stack or a set of tags) and choose a scaling strategy for each type of scalable resource in your application. You can choose to prioritize application availability, cost optimization, or a combination of the two.

After you have selected an appropriate strategy for each resource type, you are directed to a screen where you can customize the scaling plan according to your needs. The default settings can vary depending on the selected resource and resource type. The default settings should be optimal in most cases.

How to Get Started

To get started, create a scaling plan. For more information, see Getting Started with AWS Auto Scaling (p. 3).

To see the regional availability for AWS Auto Scaling, see the AWS Region Table.

Related Services

To learn more about AWS CloudFormation, see the AWS CloudFormation User Guide.

For more information on scaling your fleet of Amazon EC2 instances, see the Amazon EC2 Auto Scaling User Guide.

For more information on automatic scaling for resources beyond EC2, see the Application Auto Scaling User Guide.

Getting Started with AWS Auto Scaling

This tutorial provides a hands-on introduction to AWS Auto Scaling through the AWS Management Console, a web-based interface. To create your first scaling plan, complete the following steps.

Tasks

Prerequisites

Scalable resources must be created outside of AWS Auto Scaling through the AWS Management Console, an API, or via AWS CloudFormation. To learn more about AWS CloudFormation, see the AWS CloudFormation User Guide.

For your scalable resources to be discoverable in the AWS Auto Scaling console, you need the name of your CloudFormation stack or a set of tags. Tags can be assigned in a number of different ways, such as through the console of individual services by using the **Tags** tab on the relevant resource screen, or from the Tag Editor. Please note that currently, ECS services and Spot Fleet requests cannot be discovered using tags.

To ensure that your ECS services are discoverable, AWS Auto Scaling needs to know which ECS cluster is running the service. For AWS Auto Scaling to know this, your ECS services must be in the same CloudFormation stack as the ECS cluster that is running the service. Otherwise, they must be part of the default cluster. To be identified correctly, the service name must also be unique across each of these ECS clusters.

Considerations

Keep the following considerations in mind:

- You can create one scaling plan per application source and add each scalable resource to one scaling plan.
- The scaling plan does not apply new target tracking policies to any resources with scaling policies that were created from outside of the plan. The external polices are kept instead. To apply the scaling plan to all scalable resources, delete any existing scaling policies.

Step 1: Search for Your Scalable Resources

Use one of the following procedures to specify the application source for your scalable resources.

To specify a CloudFormation stack as the application source

1. Open the AWS Auto Scaling console at https://console.aws.amazon.com/awsautoscaling/. From the welcome page, choose **Get started**.
2. Select **Search by CloudFormation stack**.
3. Select your AWS CloudFormation stack and choose **Next**.

To specify a set of tags as the application source

1. Open the AWS Auto Scaling console at https://console.aws.amazon.com/awsautoscaling/. From the welcome page, choose **Get started**.
2. Select **Search by tag**.
3. For each tag, select a tag key from **Key** and tag values from **Value**. To add tags, choose **Add another row**. To remove tags, choose **Remove**.
4. When you are finished specifying tags, choose **Next**.

Step 2: Configure Your Scaling Plan

On the **Configure scaling plan** page, for **Scaling plan details**, **Name**, type a name for your scaling plan. For each type of resource, choose a strategy. **Optimize for availability** is chosen by default. To omit a type of resource from your scaling plan, clear **Include in scaling plan**. When you are finished, choose **Next**.

Step 3: Specify Custom Settings (Optional)

Use the following procedure to set custom scaling settings for one or more scalable resources. In most cases, however, the default settings should be optimal, with the possible exception of the values for minimum capacity and maximum capacity which should be carefully adjusted.

To specify custom settings

1. On the **Specify custom settings** page, expand the section for the resource type you want to see, and then select any number of resources from the list.
2. Under **General settings**, you can customize the following settings:

   - **Include in scaling plan** - If this setting is disabled, it omits the selected resources from your scaling plan.
   - **Scaling strategy** - Specifies a target value for the default utilization metric for that resource type. Choose one of the following options:
     - Optimize for availability
     - Balance availability and cost
     - Optimize for cost
     - Custom
   - **Scaling metric** - Changes the utilization metric for the scaling strategy to the specified metric.
   - **Target value** - Changes the target value to a specified value of between 1 and 100 percent.
3. Under **Dynamic scaling settings**, you can customize the following settings:

   - **Minimum capacity** - Specifies the minimum value to scale to in response to a change in demand. When AWS Auto Scaling scales in, it can't decrease the capacity of the selected resources below the minimum capacity. The default value depends on the selected resources.

- **Maximum capacity** - Specifies the maximum value to scale to in response to a change in demand. When AWS Auto Scaling scales out, it can't increase the capacity of the selected resources above the maximum capacity. The default value depends on the selected resources.
- **Cooldown** - Specifies the amount of time, in seconds, to wait for the previous scaling activity to take effect before starting another scaling action. The default value is 300 seconds. This setting is not used if the resource is an Auto Scaling group.

4. When you are finished specifying custom settings, choose **Next**.

   Note
   To revert any of your changes, select the resources and choose **Revert to original**. This resets the selected resources to their last known state within the scaling plan.

Step 4: Create Your Scaling Plan

On the **Review and create** page, review the details of your scaling plan and choose **Create scaling plan**.

Delete Your Scaling Plan

Deleting a scaling plan deletes the target tracking policies that AWS Auto Scaling created on your behalf. Deleting a scaling plan does not delete your AWS CloudFormation stack or the scalable resources.

To delete a scaling plan

1. Open the AWS Auto Scaling console at https://console.aws.amazon.com/awsautoscaling/.
2. On the **Scaling plans** page, select the scaling plan and choose **Delete**.
3. When prompted for confirmation, choose **Delete**.

After you delete your scaling plan, your resources do not revert to their original capacity. Your resources are left in the state they were in when the scaling plan was deleted. For example, if your Auto Scaling group is scaled to 10 instances when you delete the scaling plan, your group will still be scaled to 10 instances after the scaling plan is deleted.

Authentication and Access Control for AWS Auto Scaling

Access to AWS Auto Scaling requires credentials that AWS can use to authenticate your requests. Those credentials must have permissions to perform AWS Auto Scaling actions, such as creating scaling plans.

This topic provides details on how you can use AWS Identity and Access Management (IAM) to help secure your resources by controlling who can perform AWS Auto Scaling actions.

By default, a brand new IAM user has no permissions to do anything. To grant permissions to call AWS Auto Scaling actions, you attach an IAM policy to the IAM users or groups that require the permissions it grants.

Specifying Actions in a Policy

You can specify any and all AWS Auto Scaling actions in an IAM policy. For more information, see Actions in the *AWS Auto Scaling API Reference.*

To specify a single policy, you can use the following prefix with the name of the action: `autoscaling-plans:`. For example:

```
"Action": "autoscaling-plans:DescribeScalingPlans"
```

Wildcards are supported. For example, you can use `autoscaling-plans:*` to specify all AWS Auto Scaling actions.

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

You can also use `Describe*` to specify all actions whose names start with `Describe`.

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

Specifying the Resource

AWS Auto Scaling has no service-defined resources that can be used as the `Resource` element of an IAM policy statement. Therefore, there are no Amazon Resource Names (ARNs) for you to use in an IAM policy. To control access to AWS Auto Scaling actions, always use an * (asterisk) as the resource when writing an IAM policy.

Specifying Conditions in a Policy

When you grant permissions, you can use IAM policy language to specify the conditions when a policy should take effect. For example, you might want a policy to be applied only after a specific date. To express conditions, use predefined condition keys.

For a list of context keys supported by each AWS service and a list of AWS-wide policy keys, see Actions, Resources, and Condition Keys for AWS Services and AWS Global Condition Context Keys in the *IAM User Guide*.

AWS Auto Scaling does not provide additional condition keys.

Example Policies

To create a scaling plan, users must have permission to use the actions in the following example policy.

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

Additional IAM Permissions

Users must have the following IAM additional permissions for each type of scalable resource they must add to a scaling plan. You can specify the following actions in the `Action` element of an IAM policy statement.

Auto Scaling groups

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

Resource types other than Auto Scaling groups

- `application-autoscaling:RegisterScalableTarget`
- `application-autoscaling:DescribeScalableTargets`
- `application-autoscaling:DeregisterScalableTarget`
- `application-autoscaling:PutScalingPolicy`
- `application-autoscaling:DescribeScalingPolicies`
- `application-autoscaling:DeleteScalingPolicy`

ECS services

- `ecs:DescribeServices`
- `ecs:UpdateServices`

Spot Fleet requests

- `ec2:DescribeSpotFleetRequests`
- `ec2:ModifySpotFleetRequest`

DynamoDB tables or global indexes

- `dynamodb:DescribeTable`
- `dynamodb:UpdateTable`

Aurora DB clusters

- `rds:AddTagsToResource`
- `rds:CreateDBInstance`
- `rds:DeleteDBInstance`
- `rds:DescribeDBClusters`
- `rds:DescribeDBInstances`

AWS Auto Scaling Limits

Your AWS account has the following limits related to AWS Auto Scaling. To request a limit increase, use the Auto Scaling Limits form.

- Scaling plans: 100
- Target tracking configurations per instruction: 10
- Target tracking configurations per scaling plan: 500

AWS Auto Scaling Resources

The following related resources can help you as you work with this service.

- **AWS Auto Scaling** – The primary web page for information about AWS Auto Scaling.
- **AWS Auto Scaling FAQ** – The answers to questions customers ask about AWS Auto Scaling.
- **AWS Auto Scaling Discussion Forum** – Get help from the community.
- **Target Tracking Scaling Policies** for Amazon EC2 Auto Scaling – Get information about target tracking scaling policies for Amazon EC2 Auto Scaling groups.
- **Target Tracking Scaling Policies** for all other resources – Get information about target tracking scaling policies for resources beyond EC2, such as DynamoDB indexes and tables and ECS services.
- **AWS Auto Scaling API and CLI Reference Guides** – Documentation for the API calls and the AWS CLI commands that you can use to create, modify, and delete Auto Scaling plans.
- **Logging API Calls with CloudTrail** – Get information about monitoring calls made to the API for your account, including calls made by the AWS Management Console, command line tools, and other services.

The following additional resources are available to help you learn more about AWS.

- **Classes & Workshops** – Links to role-based and specialty courses as well as self-paced labs to help sharpen your AWS skills and gain practical experience.
- **AWS Developer Tools** – Links to developer tools, SDKs, IDE toolkits, and command line tools for developing and managing AWS applications.
- **AWS Whitepapers** – Links to a comprehensive list of technical AWS whitepapers, covering topics such as architecture, security, and economics and authored by AWS Solutions Architects or other technical experts.
- **AWS Support Center** – The hub for creating and managing your AWS Support cases. Also includes links to other helpful resources, such as forums, technical FAQs, service health status, and AWS Trusted Advisor.
- **AWS Support** – The primary web page for information about AWS Support, a one-on-one, fast-response support channel to help you build and run applications in the cloud.
- **Contact Us** – A central contact point for inquiries concerning AWS billing, account, events, abuse, and other issues.
- **AWS Site Terms** – Detailed information about our copyright and trademark; your account, license, and site access; and other topics.

Document History

The following table describes important additions to the AWS Auto Scaling documentation. For notification about updates to this documentation, you can subscribe to the RSS feed.

update-history-change	update-history-description	update-history-date
Support for custom resource settings (p. 11)	Added support for customizing various settings for each individual resource or multiple resources at the same time. For more information, see Getting Started with AWS Auto Scaling.	October 9, 2018
Tags as an application source (p. 11)	This release adds support for specifying a set of tags as an application source.	April 23, 2018
New service (p. 11)	Initial release of AWS Auto Scaling.	January 16, 2018