AWS Auto Scaling

API Reference

API Version 2018-01-06

AWS Auto Scaling: API Reference

Table of Contents

Welcome 1
Actions 2
CreateScalingPlan 3
Request Syntax 3
Request Parameters 4
Response Syntax 4
Response Elements 4
Errors 5
See Also 5
DeleteScalingPlan 6
Request Syntax 6
Request Parameters 6
Response Elements 6
Errors 6
See Also 7
DescribeScalingPlanResources 8
Request Syntax 8
Request Parameters 8
Response Syntax 9
Response Elements 9
Errors 10
See Also 10
DescribeScalingPlans 11
Request Syntax 11
Request Parameters 11
Response Syntax 12
Response Elements 13
Errors 13
See Also 14
GetScalingPlanResourceForecastData 15
Request Syntax 15
Request Parameters 15
Response Syntax 16
Response Elements 17
Errors 17
See Also 17
UpdateScalingPlan 18
Request Syntax 18
Request Parameters 19
Response Elements 19
Errors 20
See Also 20
Data Types 21
ApplicationSource 22
Contents 22
See Also 22
CustomizedLoadMetricSpecification 23
Contents 23
See Also 24
CustomizedScalingMetricSpecification 25
Contents 25
See Also 26
Datapoint 27
Contents 27

See Also 27
MetricDimension 28
Contents 28
See Also 28
PredefinedLoadMetricSpecification 29
Contents 29
See Also 29
PredefinedScalingMetricSpecification 30
Contents 30
See Also 30
ScalingInstruction 32
Contents 32
See Also 35
ScalingPlan 37
Contents 37
See Also 38
ScalingPlanResource 39
Contents 39
See Also 41
ScalingPolicy 42
Contents 42
See Also 42
TagFilter 43
Contents 43
See Also 43
TargetTrackingConfiguration 44
Contents 44
See Also 45
Common Parameters 46
Common Errors 48
Logging API Calls with CloudTrail 50
AWS Auto Scaling Information in CloudTrail 50
Understanding AWS Auto Scaling Log File Entries 51
..... 51

Welcome

This is the *AWS Auto Scaling API Reference*. Use AWS Auto Scaling to create scaling plans for your applications to automatically scale your scalable AWS resources.

API Summary

You can use the AWS Auto Scaling service API to accomplish the following tasks:

- Create and manage scaling plans
- Define target tracking scaling policies to dynamically scale your resources based on utilization
- Scale Amazon EC2 Auto Scaling groups using predictive scaling and dynamic scaling to scale your Amazon EC2 capacity faster
- Set minimum and maximum capacity limits
- Retrieve information on existing scaling plans
- Access current forecast data and historical forecast data for up to 56 days previous

The documentation for each action shows the Query API request syntax, the request parameters, and the response elements and provides links to language-specific SDK reference topics. For more information, see AWS SDKs.

To learn more about AWS Auto Scaling, including information about granting IAM users required permissions for AWS Auto Scaling actions, see the AWS Auto Scaling User Guide.

This document was last published on October 6, 2021.

Actions

The following actions are supported:

- CreateScalingPlan (p. 3)
- DeleteScalingPlan (p. 6)
- DescribeScalingPlanResources (p. 8)
- DescribeScalingPlans (p. 11)
- GetScalingPlanResourceForecastData (p. 15)
- UpdateScalingPlan (p. 18)

CreateScalingPlan

Creates a scaling plan.

Request Syntax

```
{
   "ApplicationSource": {
      "CloudFormationStackARN": "string",
      "TagFilters": [
         {
            "Key": "string",
            "Values": [ "string" ]
         }
      ]
   },
   "ScalingInstructions": [
      {
         "CustomizedLoadMetricSpecification": {
            "Dimensions": [
               {
                  "Name": "string",
                  "Value": "string"
               }
            ],
            "MetricName": "string",
            "Namespace": "string",
            "Statistic": "string",
            "Unit": "string"
         },
         "DisableDynamicScaling": boolean,
         "MaxCapacity": number,
         "MinCapacity": number,
         "PredefinedLoadMetricSpecification": {
            "PredefinedLoadMetricType": "string",
            "ResourceLabel": "string"
         },
         "PredictiveScalingMaxCapacityBehavior": "string",
         "PredictiveScalingMaxCapacityBuffer": number,
         "PredictiveScalingMode": "string",
         "ResourceId": "string",
         "ScalableDimension": "string",
         "ScalingPolicyUpdateBehavior": "string",
         "ScheduledActionBufferTime": number,
         "ServiceNamespace": "string",
         "TargetTrackingConfigurations": [
            {
               "CustomizedScalingMetricSpecification": {
                  "Dimensions": [
                     {
                        "Name": "string",
                        "Value": "string"
                     }
                  ],
                  "MetricName": "string",
                  "Namespace": "string",
                  "Statistic": "string",
                  "Unit": "string"
               },
               "DisableScaleIn": boolean,
               "EstimatedInstanceWarmup": number,
               "PredefinedScalingMetricSpecification": {
                  "PredefinedScalingMetricType": "string",
```

```
                "ResourceLabel": "string"
            },
            "ScaleInCooldown": number,
            "ScaleOutCooldown": number,
            "TargetValue": number
        }
    ]
  }
 ],
 "ScalingPlanName": "string"
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters (p. 46).

The request accepts the following data in JSON format.

ApplicationSource (p. 3)

A CloudFormation stack or set of tags. You can create one scaling plan per application source.

Type: ApplicationSource (p. 22) object

Required: Yes

ScalingInstructions (p. 3)

The scaling instructions.

Type: Array of ScalingInstruction (p. 32) objects

Required: Yes

ScalingPlanName (p. 3)

The name of the scaling plan. Names cannot contain vertical bars, colons, or forward slashes.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\p{Print}&&[^|:/]]+`

Required: Yes

Response Syntax

```
{
   "ScalingPlanVersion": number
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

ScalingPlanVersion (p. 4)

The version number of the scaling plan. This value is always `1`. Currently, you cannot have multiple scaling plan versions.

Type: Long

Errors

For information about the errors that are common to all actions, see Common Errors (p. 48).

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to a scaling plan that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

LimitExceededException

Your account exceeded a limit. This exception is thrown when a per-account resource limit is exceeded.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the parameters provided.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DeleteScalingPlan

Deletes the specified scaling plan.

Deleting a scaling plan deletes the underlying ScalingInstruction (p. 32) for all of the scalable resources that are covered by the plan.

If the plan has launched resources or has scaling activities in progress, you must delete those resources separately.

Request Syntax

```
{
   "ScalingPlanName": "string",
   "ScalingPlanVersion": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters (p. 46).

The request accepts the following data in JSON format.

ScalingPlanName (p. 6)

The name of the scaling plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\p{Print}&&[^|:/]]+`

Required: Yes

ScalingPlanVersion (p. 6)

The version number of the scaling plan. Currently, the only valid value is `1`.

Type: Long

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 48).

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to a scaling plan that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

ObjectNotFoundException

The specified object could not be found.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the parameters provided.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeScalingPlanResources

Describes the scalable resources in the specified scaling plan.

Request Syntax

```
{
   "MaxResults": number,
   "NextToken": "string",
   "ScalingPlanName": "string",
   "ScalingPlanVersion": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters (p. 46).

The request accepts the following data in JSON format.

MaxResults (p. 8)

The maximum number of scalable resources to return. The value must be between 1 and 50. The default value is 50.

Type: Integer

Required: No

NextToken (p. 8)

The token for the next set of results.

Type: String

Required: No

ScalingPlanName (p. 8)

The name of the scaling plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\p{Print}&&[^|:/]]+`

Required: Yes

ScalingPlanVersion (p. 8)

The version number of the scaling plan. Currently, the only valid value is `1`.

Type: Long

Required: Yes

Response Syntax

```
{
   "NextToken": "string",
   "ScalingPlanResources": [
      {
         "ResourceId": "string",
         "ScalableDimension": "string",
         "ScalingPlanName": "string",
         "ScalingPlanVersion": number,
         "ScalingPolicies": [
            {
               "PolicyName": "string",
               "PolicyType": "string",
               "TargetTrackingConfiguration": {
                  "CustomizedScalingMetricSpecification": {
                     "Dimensions": [
                        {
                           "Name": "string",
                           "Value": "string"
                        }
                     ],
                     "MetricName": "string",
                     "Namespace": "string",
                     "Statistic": "string",
                     "Unit": "string"
                  },
                  "DisableScaleIn": boolean,
                  "EstimatedInstanceWarmup": number,
                  "PredefinedScalingMetricSpecification": {
                     "PredefinedScalingMetricType": "string",
                     "ResourceLabel": "string"
                  },
                  "ScaleInCooldown": number,
                  "ScaleOutCooldown": number,
                  "TargetValue": number
               }
            }
         ],
         "ScalingStatusCode": "string",
         "ScalingStatusMessage": "string",
         "ServiceNamespace": "string"
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken (p. 9)

The token required to get the next set of results. This value is `null` if there are no more results to return.

Type: String

ScalingPlanResources (p. 9)

Information about the scalable resources.

Type: Array of ScalingPlanResource (p. 39) objects

Errors

For information about the errors that are common to all actions, see Common Errors (p. 48).

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to a scaling plan that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

InvalidNextTokenException

The token provided is not valid.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the parameters provided.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

DescribeScalingPlans

Describes one or more of your scaling plans.

Request Syntax

```
{
   "ApplicationSources": [
      {
         "CloudFormationStackARN": "string",
         "TagFilters": [
            {
               "Key": "string",
               "Values": [ "string" ]
            }
         ]
      }
   ],
   "MaxResults": number,
   "NextToken": "string",
   "ScalingPlanNames": [ "string" ],
   "ScalingPlanVersion": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters (p. 46).

The request accepts the following data in JSON format.

ApplicationSources (p. 11)

The sources for the applications (up to 10). If you specify scaling plan names, you cannot specify application sources.

Type: Array of ApplicationSource (p. 22) objects

Required: No

MaxResults (p. 11)

The maximum number of scalable resources to return. This value can be between 1 and 50. The default value is 50.

Type: Integer

Required: No

NextToken (p. 11)

The token for the next set of results.

Type: String

Required: No

ScalingPlanNames (p. 11)

The names of the scaling plans (up to 10). If you specify application sources, you cannot specify scaling plan names.

Type: Array of strings

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\p{Print}&&[^|:/]]+`

Required: No

ScalingPlanVersion (p. 11)

The version number of the scaling plan. Currently, the only valid value is `1`.

Note
If you specify a scaling plan version, you must also specify a scaling plan name.

Type: Long

Required: No

Response Syntax

```
{
   "NextToken": "string",
   "ScalingPlans": [
      {
         "ApplicationSource": {
            "CloudFormationStackARN": "string",
            "TagFilters": [
               {
                  "Key": "string",
                  "Values": [ "string" ]
               }
            ]
         },
         "CreationTime": number,
         "ScalingInstructions": [
            {
               "CustomizedLoadMetricSpecification": {
                  "Dimensions": [
                     {
                        "Name": "string",
                        "Value": "string"
                     }
                  ],
                  "MetricName": "string",
                  "Namespace": "string",
                  "Statistic": "string",
                  "Unit": "string"
               },
               "DisableDynamicScaling": boolean,
               "MaxCapacity": number,
               "MinCapacity": number,
               "PredefinedLoadMetricSpecification": {
                  "PredefinedLoadMetricType": "string",
                  "ResourceLabel": "string"
               },
               "PredictiveScalingMaxCapacityBehavior": "string",
               "PredictiveScalingMaxCapacityBuffer": number,
               "PredictiveScalingMode": "string",
               "ResourceId": "string",
               "ScalableDimension": "string",
               "ScalingPolicyUpdateBehavior": "string",
               "ScheduledActionBufferTime": number,
               "ServiceNamespace": "string",
               "TargetTrackingConfigurations": [
```

```
                {
                   "CustomizedScalingMetricSpecification": {
                      "Dimensions": [
                         {
                            "Name": "string",
                            "Value": "string"
                         }
                      ],
                      "MetricName": "string",
                      "Namespace": "string",
                      "Statistic": "string",
                      "Unit": "string"
                   },
                   "DisableScaleIn": boolean,
                   "EstimatedInstanceWarmup": number,
                   "PredefinedScalingMetricSpecification": {
                      "PredefinedScalingMetricType": "string",
                      "ResourceLabel": "string"
                   },
                   "ScaleInCooldown": number,
                   "ScaleOutCooldown": number,
                   "TargetValue": number
                }
             ]
          }
       ],
       "ScalingPlanName": "string",
       "ScalingPlanVersion": number,
       "StatusCode": "string",
       "StatusMessage": "string",
       "StatusStartTime": number
    }
 ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

NextToken (p. 12)

The token required to get the next set of results. This value is `null` if there are no more results to return.

Type: String

ScalingPlans (p. 12)

Information about the scaling plans.

Type: Array of ScalingPlan (p. 37) objects

Errors

For information about the errors that are common to all actions, see Common Errors (p. 48).

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to a scaling plan that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

InvalidNextTokenException

The token provided is not valid.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the parameters provided.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

GetScalingPlanResourceForecastData

Retrieves the forecast data for a scalable resource.

Capacity forecasts are represented as predicted values, or data points, that are calculated using historical data points from a specified CloudWatch load metric. Data points are available for up to 56 days.

Request Syntax

```
{
   "EndTime": number,
   "ForecastDataType": "string",
   "ResourceId": "string",
   "ScalableDimension": "string",
   "ScalingPlanName": "string",
   "ScalingPlanVersion": number,
   "ServiceNamespace": "string",
   "StartTime": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters (p. 46).

The request accepts the following data in JSON format.

EndTime (p. 15)

The exclusive end time of the time range for the forecast data to get. The maximum time duration between the start and end time is seven days.

Although this parameter can accept a date and time that is more than two days in the future, the availability of forecast data has limits. AWS Auto Scaling only issues forecasts for periods of two days in advance.

Type: Timestamp

Required: Yes

ForecastDataType (p. 15)

The type of forecast data to get.

- `LoadForecast`: The load metric forecast.
- `CapacityForecast`: The capacity forecast.
- `ScheduledActionMinCapacity`: The minimum capacity for each scheduled scaling action. This data is calculated as the larger of two values: the capacity forecast or the minimum capacity in the scaling instruction.
- `ScheduledActionMaxCapacity`: The maximum capacity for each scheduled scaling action. The calculation used is determined by the predictive scaling maximum capacity behavior setting in the scaling instruction.

Type: String

Valid Values: `CapacityForecast | LoadForecast | ScheduledActionMinCapacity | ScheduledActionMaxCapacity`

Required: Yes

ResourceId (p. 15)

The ID of the resource. This string consists of a prefix (`autoScalingGroup`) followed by the name of a specified Auto Scaling group (`my-asg`). Example: `autoScalingGroup/my-asg`.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension (p. 15)

The scalable dimension for the resource. The only valid value is `autoscaling:autoScalingGroup:DesiredCapacity`.

Type: String

Valid Values: `autoscaling:autoScalingGroup:DesiredCapacity`

Required: Yes

ScalingPlanName (p. 15)

The name of the scaling plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\p{Print}&&[^|:/]]+`

Required: Yes

ScalingPlanVersion (p. 15)

The version number of the scaling plan. Currently, the only valid value is `1`.

Type: Long

Required: Yes

ServiceNamespace (p. 15)

The namespace of the AWS service. The only valid value is `autoscaling`.

Type: String

Valid Values: `autoscaling`

Required: Yes

StartTime (p. 15)

The inclusive start time of the time range for the forecast data to get. The date and time can be at most 56 days before the current date and time.

Type: Timestamp

Required: Yes

Response Syntax

```
{
```

```
   "Datapoints": [
      {
         "Timestamp": number,
         "Value": number
      }
   ]
}
```

Response Elements

If the action is successful, the service sends back an HTTP 200 response.

The following data is returned in JSON format by the service.

Datapoints (p. 16)

The data points to return.

Type: Array of Datapoint (p. 27) objects

Errors

For information about the errors that are common to all actions, see Common Errors (p. 48).

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the parameters provided.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

UpdateScalingPlan

Updates the specified scaling plan.

You cannot update a scaling plan if it is in the process of being created, updated, or deleted.

Request Syntax

```
{
   "ApplicationSource": {
      "CloudFormationStackARN": "string",
      "TagFilters": [
         {
            "Key": "string",
            "Values": [ "string" ]
         }
      ]
   },
   "ScalingInstructions": [
      {
         "CustomizedLoadMetricSpecification": {
            "Dimensions": [
               {
                  "Name": "string",
                  "Value": "string"
               }
            ],
            "MetricName": "string",
            "Namespace": "string",
            "Statistic": "string",
            "Unit": "string"
         },
         "DisableDynamicScaling": boolean,
         "MaxCapacity": number,
         "MinCapacity": number,
         "PredefinedLoadMetricSpecification": {
            "PredefinedLoadMetricType": "string",
            "ResourceLabel": "string"
         },
         "PredictiveScalingMaxCapacityBehavior": "string",
         "PredictiveScalingMaxCapacityBuffer": number,
         "PredictiveScalingMode": "string",
         "ResourceId": "string",
         "ScalableDimension": "string",
         "ScalingPolicyUpdateBehavior": "string",
         "ScheduledActionBufferTime": number,
         "ServiceNamespace": "string",
         "TargetTrackingConfigurations": [
            {
               "CustomizedScalingMetricSpecification": {
                  "Dimensions": [
                     {
                        "Name": "string",
                        "Value": "string"
                     }
                  ],
                  "MetricName": "string",
                  "Namespace": "string",
                  "Statistic": "string",
                  "Unit": "string"
               },
               "DisableScaleIn": boolean,
```

```
                "EstimatedInstanceWarmup": number,
                "PredefinedScalingMetricSpecification": {
                   "PredefinedScalingMetricType": "string",
                   "ResourceLabel": "string"
                },
                "ScaleInCooldown": number,
                "ScaleOutCooldown": number,
                "TargetValue": number
             }
          ]
       }
   ],
   "ScalingPlanName": "string",
   "ScalingPlanVersion": number
}
```

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters (p. 46).

The request accepts the following data in JSON format.

ApplicationSource (p. 18)

A CloudFormation stack or set of tags.

Type: ApplicationSource (p. 22) object

Required: No

ScalingInstructions (p. 18)

The scaling instructions.

Type: Array of ScalingInstruction (p. 32) objects

Required: No

ScalingPlanName (p. 18)

The name of the scaling plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\p{Print}&&[^|:/]]+`

Required: Yes

ScalingPlanVersion (p. 18)

The version number of the scaling plan. The only valid value is `1`. Currently, you cannot have multiple scaling plan versions.

Type: Long

Required: Yes

Response Elements

If the action is successful, the service sends back an HTTP 200 response with an empty HTTP body.

Errors

For information about the errors that are common to all actions, see Common Errors (p. 48).

ConcurrentUpdateException

Concurrent updates caused an exception, for example, if you request an update to a scaling plan that already has a pending update.

HTTP Status Code: 400

InternalServiceException

The service encountered an internal error.

HTTP Status Code: 400

ObjectNotFoundException

The specified object could not be found.

HTTP Status Code: 400

ValidationException

An exception was thrown for a validation issue. Review the parameters provided.

HTTP Status Code: 400

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS Command Line Interface
- AWS SDK for .NET
- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V3

Data Types

The AWS Auto Scaling API contains several data types that various actions use. This section describes each data type in detail.

> **Note**
> The order of each element in a data type structure is not guaranteed. Applications should not assume a particular order.

The following data types are supported:

- ApplicationSource (p. 22)
- CustomizedLoadMetricSpecification (p. 23)
- CustomizedScalingMetricSpecification (p. 25)
- Datapoint (p. 27)
- MetricDimension (p. 28)
- PredefinedLoadMetricSpecification (p. 29)
- PredefinedScalingMetricSpecification (p. 30)
- ScalingInstruction (p. 32)
- ScalingPlan (p. 37)
- ScalingPlanResource (p. 39)
- ScalingPolicy (p. 42)
- TagFilter (p. 43)
- TargetTrackingConfiguration (p. 44)

ApplicationSource

Represents an application source.

Contents

CloudFormationStackARN

The Amazon Resource Name (ARN) of a CloudFormation stack.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

TagFilters

A set of tags (up to 50).

Type: Array of TagFilter (p. 43) objects

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomizedLoadMetricSpecification

Represents a CloudWatch metric of your choosing that can be used for predictive scaling.

For predictive scaling to work with a customized load metric specification, AWS Auto Scaling needs access to the `Sum` and `Average` statistics that CloudWatch computes from metric data.

When you choose a load metric, make sure that the required `Sum` and `Average` statistics for your metric are available in CloudWatch and that they provide relevant data for predictive scaling. The `Sum` statistic must represent the total load on the resource, and the `Average` statistic must represent the average load per capacity unit of the resource. For example, there is a metric that counts the number of requests processed by your Auto Scaling group. If the `Sum` statistic represents the total request count processed by the group, then the `Average` statistic for the specified metric must represent the average request count processed by each instance of the group.

If you publish your own metrics, you can aggregate the data points at a given interval and then publish the aggregated data points to CloudWatch. Before AWS Auto Scaling generates the forecast, it sums up all the metric data points that occurred within each hour to match the granularity period that is used in the forecast (60 minutes).

For information about terminology, available metrics, or how to publish new metrics, see Amazon CloudWatch Concepts in the *Amazon CloudWatch User Guide*.

After creating your scaling plan, you can use the AWS Auto Scaling console to visualize forecasts for the specified metric. For more information, see View Scaling Information for a Resource in the *AWS Auto Scaling User Guide*.

Contents

Dimensions

The dimensions of the metric.

Conditional: If you published your metric with dimensions, you must specify the same dimensions in your customized load metric specification.

Type: Array of MetricDimension (p. 28) objects

Required: No

MetricName

The name of the metric.

Type: String

Required: Yes

Namespace

The namespace of the metric.

Type: String

Required: Yes

Statistic

The statistic of the metric. The only valid value is `Sum`.

Type: String

Valid Values: `Sum`

Required: Yes

Unit

The unit of the metric.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

CustomizedScalingMetricSpecification

Represents a CloudWatch metric of your choosing that can be used for dynamic scaling as part of a target tracking scaling policy.

To create your customized scaling metric specification:

- Add values for each required parameter from CloudWatch. You can use an existing metric, or a new metric that you create. To use your own metric, you must first publish the metric to CloudWatch. For more information, see Publish Custom Metrics in the *Amazon CloudWatch User Guide*.
- Choose a metric that changes proportionally with capacity. The value of the metric should increase or decrease in inverse proportion to the number of capacity units. That is, the value of the metric should decrease when capacity increases.

For information about terminology, available metrics, or how to publish new metrics, see Amazon CloudWatch Concepts in the *Amazon CloudWatch User Guide*.

Contents

Dimensions

The dimensions of the metric.

Conditional: If you published your metric with dimensions, you must specify the same dimensions in your customized scaling metric specification.

Type: Array of MetricDimension (p. 28) objects

Required: No

MetricName

The name of the metric.

Type: String

Required: Yes

Namespace

The namespace of the metric.

Type: String

Required: Yes

Statistic

The statistic of the metric.

Type: String

Valid Values: `Average | Minimum | Maximum | SampleCount | Sum`

Required: Yes

Unit

The unit of the metric.

Type: String

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Datapoint

Represents a single value in the forecast data used for predictive scaling.

Contents

Timestamp

The time stamp for the data point in UTC format.

Type: Timestamp

Required: No

Value

The value of the data point.

Type: Double

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

MetricDimension

Represents a dimension for a customized metric.

Contents

Name

The name of the dimension.

Type: String

Required: Yes

Value

The value of the dimension.

Type: String

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PredefinedLoadMetricSpecification

Represents a predefined metric that can be used for predictive scaling.

After creating your scaling plan, you can use the AWS Auto Scaling console to visualize forecasts for the specified metric. For more information, see View Scaling Information for a Resource in the *AWS Auto Scaling User Guide*.

Contents

PredefinedLoadMetricType

The metric type.

Type: String

Valid Values: `ASGTotalCPUUtilization | ASGTotalNetworkIn | ASGTotalNetworkOut | ALBTargetGroupRequestCount`

Required: Yes

ResourceLabel

Identifies the resource associated with the metric type. You can't specify a resource label unless the metric type is `ALBTargetGroupRequestCount` and there is a target group for an Application Load Balancer attached to the Auto Scaling group.

You create the resource label by appending the final portion of the load balancer ARN and the final portion of the target group ARN into a single value, separated by a forward slash (/). The format is app/<load-balancer-name>/<load-balancer-id>/targetgroup/<target-group-name>/<target-group-id>, where:

- app/<load-balancer-name>/<load-balancer-id> is the final portion of the load balancer ARN
- targetgroup/<target-group-name>/<target-group-id> is the final portion of the target group ARN.

This is an example: app/EC2Co-EcsEl-1TKLTMITMM0EO/f37c06a68c1748aa/targetgroup/EC2Co-Defau-LDNM7Q3ZH1ZN/6d4ea56ca2d6a18d.

To find the ARN for an Application Load Balancer, use the DescribeLoadBalancers API operation. To find the ARN for the target group, use the DescribeTargetGroups API operation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1023.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

PredefinedScalingMetricSpecification

Represents a predefined metric that can be used for dynamic scaling as part of a target tracking scaling policy.

Contents

PredefinedScalingMetricType

The metric type. The `ALBRequestCountPerTarget` metric type applies only to Auto Scaling groups, Spot Fleet requests, and ECS services.

Type: String

Valid Values: `ASGAverageCPUUtilization | ASGAverageNetworkIn | ASGAverageNetworkOut | DynamoDBReadCapacityUtilization | DynamoDBWriteCapacityUtilization | ECSServiceAverageCPUUtilization | ECSServiceAverageMemoryUtilization | ALBRequestCountPerTarget | RDSReaderAverageCPUUtilization | RDSReaderAverageDatabaseConnections | EC2SpotFleetRequestAverageCPUUtilization | EC2SpotFleetRequestAverageNetworkIn | EC2SpotFleetRequestAverageNetworkOut`

Required: Yes

ResourceLabel

Identifies the resource associated with the metric type. You can't specify a resource label unless the metric type is `ALBRequestCountPerTarget` and there is a target group for an Application Load Balancer attached to the Auto Scaling group, Spot Fleet request, or ECS service.

You create the resource label by appending the final portion of the load balancer ARN and the final portion of the target group ARN into a single value, separated by a forward slash (/). The format is app/<load-balancer-name>/<load-balancer-id>/targetgroup/<target-group-name>/<target-group-id>, where:

- app/<load-balancer-name>/<load-balancer-id> is the final portion of the load balancer ARN
- targetgroup/<target-group-name>/<target-group-id> is the final portion of the target group ARN.

This is an example: app/EC2Co-EcsEl-1TKLTMITMM0EO/f37c06a68c1748aa/targetgroup/EC2Co-Defau-LDNM7Q3ZH1ZN/6d4ea56ca2d6a18d.

To find the ARN for an Application Load Balancer, use the DescribeLoadBalancers API operation. To find the ARN for the target group, use the DescribeTargetGroups API operation.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1023.

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2

- AWS SDK for Ruby V3

ScalingInstruction

Describes a scaling instruction for a scalable resource in a scaling plan. Each scaling instruction applies to one resource.

AWS Auto Scaling creates target tracking scaling policies based on the scaling instructions. Target tracking scaling policies adjust the capacity of your scalable resource as required to maintain resource utilization at the target value that you specified.

AWS Auto Scaling also configures predictive scaling for your Amazon EC2 Auto Scaling groups using a subset of parameters, including the load metric, the scaling metric, the target value for the scaling metric, the predictive scaling mode (forecast and scale or forecast only), and the desired behavior when the forecast capacity exceeds the maximum capacity of the resource. With predictive scaling, AWS Auto Scaling generates forecasts with traffic predictions for the two days ahead and schedules scaling actions that proactively add and remove resource capacity to match the forecast.

Important
We recommend waiting a minimum of 24 hours after creating an Auto Scaling group to configure predictive scaling. At minimum, there must be 24 hours of historical data to generate a forecast. For more information, see Best Practices for AWS Auto Scaling in the *AWS Auto Scaling User Guide*.

Contents

CustomizedLoadMetricSpecification

The customized load metric to use for predictive scaling. This parameter or a **PredefinedLoadMetricSpecification** is required when configuring predictive scaling, and cannot be used otherwise.

Type: CustomizedLoadMetricSpecification (p. 23) object

Required: No

DisableDynamicScaling

Controls whether dynamic scaling by AWS Auto Scaling is disabled. When dynamic scaling is enabled, AWS Auto Scaling creates target tracking scaling policies based on the specified target tracking configurations.

The default is enabled (`false`).

Type: Boolean

Required: No

MaxCapacity

The maximum capacity of the resource. The exception to this upper limit is if you specify a non-default setting for **PredictiveScalingMaxCapacityBehavior**.

Type: Integer

Required: Yes

MinCapacity

The minimum capacity of the resource.

Type: Integer

Required: Yes

PredefinedLoadMetricSpecification

The predefined load metric to use for predictive scaling. This parameter or a **CustomizedLoadMetricSpecification** is required when configuring predictive scaling, and cannot be used otherwise.

Type: PredefinedLoadMetricSpecification (p. 29) object

Required: No

PredictiveScalingMaxCapacityBehavior

Defines the behavior that should be applied if the forecast capacity approaches or exceeds the maximum capacity specified for the resource. The default value is `SetForecastCapacityToMaxCapacity`.

The following are possible values:

- `SetForecastCapacityToMaxCapacity` - AWS Auto Scaling cannot scale resource capacity higher than the maximum capacity. The maximum capacity is enforced as a hard limit.
- `SetMaxCapacityToForecastCapacity` - AWS Auto Scaling may scale resource capacity higher than the maximum capacity to equal but not exceed forecast capacity.
- `SetMaxCapacityAboveForecastCapacity` - AWS Auto Scaling may scale resource capacity higher than the maximum capacity by a specified buffer value. The intention is to give the target tracking scaling policy extra capacity if unexpected traffic occurs.

Only valid when configuring predictive scaling.

Type: String

Valid Values: `SetForecastCapacityToMaxCapacity | SetMaxCapacityToForecastCapacity | SetMaxCapacityAboveForecastCapacity`

Required: No

PredictiveScalingMaxCapacityBuffer

The size of the capacity buffer to use when the forecast capacity is close to or exceeds the maximum capacity. The value is specified as a percentage relative to the forecast capacity. For example, if the buffer is 10, this means a 10 percent buffer, such that if the forecast capacity is 50, and the maximum capacity is 40, then the effective maximum capacity is 55.

Only valid when configuring predictive scaling. Required if the **PredictiveScalingMaxCapacityBehavior** is set to `SetMaxCapacityAboveForecastCapacity`, and cannot be used otherwise.

The range is 1-100.

Type: Integer

Required: No

PredictiveScalingMode

The predictive scaling mode. The default value is `ForecastAndScale`. Otherwise, AWS Auto Scaling forecasts capacity but does not create any scheduled scaling actions based on the capacity forecast.

Type: String

Valid Values: `ForecastAndScale | ForecastOnly`

Required: No

ResourceId

The ID of the resource. This string consists of the resource type and unique identifier.

- Auto Scaling group - The resource type is `autoScalingGroup` and the unique identifier is the name of the Auto Scaling group. Example: `autoScalingGroup/my-asg`.
- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/default/sample-webapp`.
- Spot Fleet request - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- DynamoDB table - The resource type is `table` and the unique identifier is the resource ID. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the resource ID. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension associated with the resource.

- `autoscaling:autoScalingGroup:DesiredCapacity` - The desired capacity of an Auto Scaling group.
- `ecs:service:DesiredCount` - The desired task count of an ECS service.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet request.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.

Type: String

Valid Values: `autoscaling:autoScalingGroup:DesiredCapacity | ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | rds:cluster:ReadReplicaCount | dynamodb:table:ReadCapacityUnits | dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits`

Required: Yes

ScalingPolicyUpdateBehavior

Controls whether a resource's externally created scaling policies are kept or replaced.

The default value is `KeepExternalPolicies`. If the parameter is set to `ReplaceExternalPolicies`, any scaling policies that are external to AWS Auto Scaling are deleted and new target tracking scaling policies created.

Only valid when configuring dynamic scaling.

Condition: The number of existing policies to be replaced must be less than or equal to 50. If there are more than 50 policies to be replaced, AWS Auto Scaling keeps all existing policies and does not create new ones.

Type: String

Valid Values: `KeepExternalPolicies | ReplaceExternalPolicies`

Required: No

ScheduledActionBufferTime

The amount of time, in seconds, to buffer the run time of scheduled scaling actions when scaling out. For example, if the forecast says to add capacity at 10:00 AM, and the buffer time is 5 minutes, then the run time of the corresponding scheduled scaling action will be 9:55 AM. The intention is to give resources time to be provisioned. For example, it can take a few minutes to launch an EC2 instance. The actual amount of time required depends on several factors, such as the size of the instance and whether there are startup scripts to complete.

The value must be less than the forecast interval duration of 3600 seconds (60 minutes). The default is 300 seconds.

Only valid when configuring predictive scaling.

Type: Integer

Valid Range: Minimum value of 0.

Required: No

ServiceNamespace

The namespace of the AWS service.

Type: String

Valid Values: `autoscaling | ecs | ec2 | rds | dynamodb`

Required: Yes

TargetTrackingConfigurations

The target tracking configurations (up to 10). Each of these structures must specify a unique scaling metric and a target value for the metric.

Type: Array of TargetTrackingConfiguration (p. 44) objects

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go

- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScalingPlan

Represents a scaling plan.

Contents

ApplicationSource

A CloudFormation stack or a set of tags. You can create one scaling plan per application source.

Type: ApplicationSource (p. 22) object

Required: Yes

CreationTime

The Unix time stamp when the scaling plan was created.

Type: Timestamp

Required: No

ScalingInstructions

The scaling instructions.

Type: Array of ScalingInstruction (p. 32) objects

Required: Yes

ScalingPlanName

The name of the scaling plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\p{Print}&&[^|:/]]+`

Required: Yes

ScalingPlanVersion

The version number of the scaling plan.

Type: Long

Required: Yes

StatusCode

The status of the scaling plan.

- `Active` - The scaling plan is active.
- `ActiveWithProblems` - The scaling plan is active, but the scaling configuration for one or more resources could not be applied.
- `CreationInProgress` - The scaling plan is being created.
- `CreationFailed` - The scaling plan could not be created.
- `DeletionInProgress` - The scaling plan is being deleted.
- `DeletionFailed` - The scaling plan could not be deleted.
- `UpdateInProgress` - The scaling plan is being updated.

- `UpdateFailed` - The scaling plan could not be updated.

Type: String

Valid Values: `Active | ActiveWithProblems | CreationInProgress | CreationFailed | DeletionInProgress | DeletionFailed | UpdateInProgress | UpdateFailed`

Required: Yes

StatusMessage

A simple message about the current status of the scaling plan.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

StatusStartTime

The Unix time stamp when the scaling plan entered the current status.

Type: Timestamp

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScalingPlanResource

Represents a scalable resource.

Contents

ResourceId

The ID of the resource. This string consists of the resource type and unique identifier.

- Auto Scaling group - The resource type is `autoScalingGroup` and the unique identifier is the name of the Auto Scaling group. Example: `autoScalingGroup/my-asg`.
- ECS service - The resource type is `service` and the unique identifier is the cluster name and service name. Example: `service/default/sample-webapp`.
- Spot Fleet request - The resource type is `spot-fleet-request` and the unique identifier is the Spot Fleet request ID. Example: `spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE`.
- DynamoDB table - The resource type is `table` and the unique identifier is the resource ID. Example: `table/my-table`.
- DynamoDB global secondary index - The resource type is `index` and the unique identifier is the resource ID. Example: `table/my-table/index/my-table-index`.
- Aurora DB cluster - The resource type is `cluster` and the unique identifier is the cluster name. Example: `cluster:my-db-cluster`.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 1600.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: Yes

ScalableDimension

The scalable dimension for the resource.

- `autoscaling:autoScalingGroup:DesiredCapacity` - The desired capacity of an Auto Scaling group.
- `ecs:service:DesiredCount` - The desired task count of an ECS service.
- `ec2:spot-fleet-request:TargetCapacity` - The target capacity of a Spot Fleet request.
- `dynamodb:table:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB table.
- `dynamodb:table:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB table.
- `dynamodb:index:ReadCapacityUnits` - The provisioned read capacity for a DynamoDB global secondary index.
- `dynamodb:index:WriteCapacityUnits` - The provisioned write capacity for a DynamoDB global secondary index.
- `rds:cluster:ReadReplicaCount` - The count of Aurora Replicas in an Aurora DB cluster. Available for Aurora MySQL-compatible edition and Aurora PostgreSQL-compatible edition.

Type: String

Valid Values: `autoscaling:autoScalingGroup:DesiredCapacity | ecs:service:DesiredCount | ec2:spot-fleet-request:TargetCapacity | rds:cluster:ReadReplicaCount | dynamodb:table:ReadCapacityUnits |`

`dynamodb:table:WriteCapacityUnits | dynamodb:index:ReadCapacityUnits | dynamodb:index:WriteCapacityUnits`

Required: Yes

ScalingPlanName

The name of the scaling plan.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\p{Print}&&[^|:/]]+`

Required: Yes

ScalingPlanVersion

The version number of the scaling plan.

Type: Long

Required: Yes

ScalingPolicies

The scaling policies.

Type: Array of ScalingPolicy (p. 42) objects

Required: No

ScalingStatusCode

The scaling status of the resource.

- `Active` - The scaling configuration is active.
- `Inactive` - The scaling configuration is not active because the scaling plan is being created or the scaling configuration could not be applied. Check the status message for more information.
- `PartiallyActive` - The scaling configuration is partially active because the scaling plan is being created or deleted or the scaling configuration could not be fully applied. Check the status message for more information.

Type: String

Valid Values: `Inactive | PartiallyActive | Active`

Required: Yes

ScalingStatusMessage

A simple message about the current scaling status of the resource.

Type: String

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

ServiceNamespace

The namespace of the AWS service.

Type: String

Valid Values: `autoscaling | ecs | ec2 | rds | dynamodb`

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

ScalingPolicy

Represents a scaling policy.

Contents

PolicyName

The name of the scaling policy.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `\p{Print}+`

Required: Yes

PolicyType

The type of scaling policy.

Type: String

Valid Values: `TargetTrackingScaling`

Required: Yes

TargetTrackingConfiguration

The target tracking scaling policy. Includes support for predefined or customized metrics.

Type: TargetTrackingConfiguration (p. 44) object

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TagFilter

Represents a tag.

Contents

Key

The tag key.

Type: String

Length Constraints: Minimum length of 1. Maximum length of 128.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

Values

The tag values (0 to 20).

Type: Array of strings

Length Constraints: Minimum length of 1. Maximum length of 256.

Pattern: `[\u0020-\uD7FF\uE000-\uFFFD\uD800\uDC00-\uDBFF\uDFFF\r\n\t]*`

Required: No

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

TargetTrackingConfiguration

Describes a target tracking configuration to use with AWS Auto Scaling. Used with ScalingInstruction (p. 32) and ScalingPolicy (p. 42).

Contents

CustomizedScalingMetricSpecification

A customized metric. You can specify either a predefined metric or a customized metric.

Type: CustomizedScalingMetricSpecification (p. 25) object

Required: No

DisableScaleIn

Indicates whether scale in by the target tracking scaling policy is disabled. If the value is `true`, scale in is disabled and the target tracking scaling policy doesn't remove capacity from the scalable resource. Otherwise, scale in is enabled and the target tracking scaling policy can remove capacity from the scalable resource.

The default value is `false`.

Type: Boolean

Required: No

EstimatedInstanceWarmup

The estimated time, in seconds, until a newly launched instance can contribute to the CloudWatch metrics. This value is used only if the resource is an Auto Scaling group.

Type: Integer

Required: No

PredefinedScalingMetricSpecification

A predefined metric. You can specify either a predefined metric or a customized metric.

Type: PredefinedScalingMetricSpecification (p. 30) object

Required: No

ScaleInCooldown

The amount of time, in seconds, after a scale-in activity completes before another scale-in activity can start. This property is not used if the scalable resource is an Auto Scaling group.

With the *scale-in cooldown period*, the intention is to scale in conservatively to protect your application's availability, so scale-in activities are blocked until the cooldown period has expired. However, if another alarm triggers a scale-out activity during the scale-in cooldown period, Auto Scaling scales out the target immediately. In this case, the scale-in cooldown period stops and doesn't complete.

Type: Integer

Required: No

ScaleOutCooldown

The amount of time, in seconds, to wait for a previous scale-out activity to take effect. This property is not used if the scalable resource is an Auto Scaling group.

With the *scale-out cooldown period*, the intention is to continuously (but not excessively) scale out. After Auto Scaling successfully scales out using a target tracking scaling policy, it starts to calculate the cooldown time. The scaling policy won't increase the desired capacity again unless either a larger scale out is triggered or the cooldown period ends.

Type: Integer

Required: No

TargetValue

The target value for the metric. Although this property accepts numbers of type Double, it won't accept values that are either too small or too large. Values must be in the range of -2^360 to 2^360.

Type: Double

Required: Yes

See Also

For more information about using this API in one of the language-specific AWS SDKs, see the following:

- AWS SDK for C++
- AWS SDK for Go
- AWS SDK for Java V2
- AWS SDK for Ruby V3

Common Parameters

The following list contains the parameters that all actions use for signing Signature Version 4 requests with a query string. Any action-specific parameters are listed in the topic for that action. For more information about Signature Version 4, see Signature Version 4 Signing Process in the *Amazon Web Services General Reference*.

Action

The action to be performed.

Type: string

Required: Yes

Version

The API version that the request is written for, expressed in the format YYYY-MM-DD.

Type: string

Required: Yes

X-Amz-Algorithm

The hash algorithm that you used to create the request signature.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Valid Values: `AWS4-HMAC-SHA256`

Required: Conditional

X-Amz-Credential

The credential scope value, which is a string that includes your access key, the date, the region you are targeting, the service you are requesting, and a termination string ("aws4_request"). The value is expressed in the following format: *access_key*/*YYYYMMDD*/*region*/*service*/aws4_request.

For more information, see Task 2: Create a String to Sign for Signature Version 4 in the *Amazon Web Services General Reference*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-Date

The date that is used to create the signature. The format must be ISO 8601 basic format (YYYYMMDD'T'HHMMSS'Z'). For example, the following date time is a valid X-Amz-Date value: `20120325T120000Z`.

Condition: X-Amz-Date is optional for all requests; it can be used to override the date used for signing requests. If the Date header is specified in the ISO 8601 basic format, X-Amz-Date is

not required. When X-Amz-Date is used, it always overrides the value of the Date header. For more information, see Handling Dates in Signature Version 4 in the *Amazon Web Services General Reference*.

Type: string

Required: Conditional

X-Amz-Security-Token

The temporary security token that was obtained through a call to AWS Security Token Service (AWS STS). For a list of services that support temporary security credentials from AWS Security Token Service, go to AWS Services That Work with IAM in the *IAM User Guide*.

Condition: If you're using temporary security credentials from the AWS Security Token Service, you must include the security token.

Type: string

Required: Conditional

X-Amz-Signature

Specifies the hex-encoded signature that was calculated from the string to sign and the derived signing key.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

X-Amz-SignedHeaders

Specifies all the HTTP headers that were included as part of the canonical request. For more information about specifying signed headers, see Task 1: Create a Canonical Request For Signature Version 4 in the *Amazon Web Services General Reference*.

Condition: Specify this parameter when you include authentication information in a query string instead of in the HTTP authorization header.

Type: string

Required: Conditional

Common Errors

This section lists the errors common to the API actions of all AWS services. For errors specific to an API action for this service, see the topic for that API action.

AccessDeniedException

You do not have sufficient access to perform this action.

HTTP Status Code: 400

IncompleteSignature

The request signature does not conform to AWS standards.

HTTP Status Code: 400

InternalFailure

The request processing has failed because of an unknown error, exception or failure.

HTTP Status Code: 500

InvalidAction

The action or operation requested is invalid. Verify that the action is typed correctly.

HTTP Status Code: 400

InvalidClientTokenId

The X.509 certificate or AWS access key ID provided does not exist in our records.

HTTP Status Code: 403

InvalidParameterCombination

Parameters that must not be used together were used together.

HTTP Status Code: 400

InvalidParameterValue

An invalid or out-of-range value was supplied for the input parameter.

HTTP Status Code: 400

InvalidQueryParameter

The AWS query string is malformed or does not adhere to AWS standards.

HTTP Status Code: 400

MalformedQueryString

The query string contains a syntax error.

HTTP Status Code: 404

MissingAction

The request is missing an action or a required parameter.

HTTP Status Code: 400

MissingAuthenticationToken

The request must contain either a valid (registered) AWS access key ID or X.509 certificate.

HTTP Status Code: 403

MissingParameter

A required parameter for the specified action is not supplied.

HTTP Status Code: 400

NotAuthorized

You do not have permission to perform this action.

HTTP Status Code: 400

OptInRequired

The AWS access key ID needs a subscription for the service.

HTTP Status Code: 403

RequestExpired

The request reached the service more than 15 minutes after the date stamp on the request or more than 15 minutes after the request expiration date (such as for pre-signed URLs), or the date stamp on the request is more than 15 minutes in the future.

HTTP Status Code: 400

ServiceUnavailable

The request has failed due to a temporary failure of the server.

HTTP Status Code: 503

ThrottlingException

The request was denied due to request throttling.

HTTP Status Code: 400

ValidationError

The input fails to satisfy the constraints specified by an AWS service.

HTTP Status Code: 400

Logging AWS Auto Scaling API Calls with AWS CloudTrail

AWS Auto Scaling is integrated with AWS CloudTrail, a service that provides a record of actions taken by a user, role, or an AWS service in AWS Auto Scaling. CloudTrail captures all API calls for AWS Auto Scaling as events. The calls captured include calls from the AWS Auto Scaling console and code calls to the AWS Auto Scaling API. If you create a trail, you can enable continuous delivery of CloudTrail events to an Amazon S3 bucket, including events for AWS Auto Scaling. If you don't configure a trail, you can still view the most recent events in the CloudTrail console in **Event history**. Using the information collected by CloudTrail, you can determine the request that was made to AWS Auto Scaling, the IP address from which the request was made, who made the request, when it was made, and additional details.

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

AWS Auto Scaling Information in CloudTrail

CloudTrail is enabled on your AWS account when you create the account. When AWS Auto Scaling activity occurs, that activity is recorded in a CloudTrail event along with other AWS service events in **Event history**. You can view, search, and download recent events in your AWS account. For more information, see Viewing Events with CloudTrail Event History.

For an ongoing record of events in your AWS account, including events for AWS Auto Scaling, create a trail. A *trail* enables CloudTrail to deliver log files to an Amazon S3 bucket. By default, when you create a trail in the console, the trail applies to all AWS Regions. The trail logs events from all Regions in the AWS partition and delivers the log files to the Amazon S3 bucket that you specify. Additionally, you can configure other Amazon Web Services to further analyze and act upon the event data collected in CloudTrail logs. For more information, see the following:

- Overview for Creating a Trail
- CloudTrail Supported Services and Integrations
- Configuring Amazon SNS Notifications for CloudTrail
- Receiving CloudTrail Log Files from Multiple Regions and Receiving CloudTrail Log Files from Multiple Accounts

All AWS Auto Scaling actions are logged by CloudTrail and are documented in the AWS Auto Scaling API Reference. For example, calls to the `CreateScalingPlan`, `DeleteScalingPlan`, and `DescribeScalingPlans` actions generate entries in the CloudTrail log files.

Every event or log entry contains information about who generated the request. The identity information helps you determine the following:

- Whether the request was made with root or AWS Identity and Access Management (IAM) user credentials.
- Whether the request was made with temporary security credentials for a role or federated user.
- Whether the request was made by another AWS service.

For more information, see the CloudTrail userIdentity Element.

Understanding AWS Auto Scaling Log File Entries

A trail is a configuration that enables delivery of events as log files to an Amazon S3 bucket that you specify. CloudTrail log files contain one or more log entries. An event represents a single request from any source and includes information about the requested action, the date and time of the action, request parameters, and so on. CloudTrail log files aren't an ordered stack trace of the public API calls, so they don't appear in any specific order.

The following example shows a CloudTrail log entry that demonstrates the `CreateScalingPlan` action.


```
{
    "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"
}
```