AWS Auto Scaling

API Reference

API Version 2018-01-06

AWS Auto Scaling: API Reference

Table of Contents

Welcome

Use AWS Auto Scaling to quickly discover all the scalable AWS resources for your application and configure dynamic scaling for your scalable resources.

To get started, 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 scaling policies for the scalable resources in your scaling plan. Target tracking scaling policies adjust the capacity of your scalable resource as required to maintain resource utilization at the target value that you specified.

For more information about AWS Auto Scaling, see the AWS Auto Scaling User Guide.

This document was last published on November 19, 2018.

Actions

The following actions are supported:

CreateScalingPlan

Creates a scaling plan.

A scaling plan contains a set of instructions used to configure dynamic scaling for the scalable resources in your application. AWS Auto Scaling creates target tracking scaling policies based on the scaling instructions in your scaling plan.

Request Syntax

```
{
   "ApplicationSource": {
      "CloudFormationStackARN": "string",
      "TagFilters": [
         {
            "Key": "string",
            "Values": [ "string" ]
         }
      ]
   },
   "ScalingInstructions": [
      {
         "MaxCapacity": number,
         "MinCapacity": number,
         "ResourceId": "string",
         "ScalableDimension": "string",
         "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. 34).

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. 19) object

Required: Yes

ScalingInstructions (p. 3)

The scaling instructions.

Type: Array of ScalingInstruction (p. 23) 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 of the scaling plan. This value is always 1.

Type: Long

Errors

For information about the errors that are common to all actions, see Common Errors (p. 36).

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
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V2

DeleteScalingPlan

Deletes the specified scaling plan.

Request Syntax

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

Request Parameters

For information about the parameters that are common to all actions, see Common Parameters (p. 34).

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 of the scaling plan.

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. 36).

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
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V2

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. 34).

The request accepts the following data in JSON format.

MaxResults (p. 8)

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. 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 of the scaling plan.

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. 27) objects

Errors

For information about the errors that are common to all actions, see Common Errors (p. 36).

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
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V2

DescribeScalingPlans

Describes the specified scaling plans or all 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. 34).

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. 19) 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 of the scaling plan. 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": [
            {
               "MaxCapacity": number,
               "MinCapacity": number,
               "ResourceId": "string",
               "ScalableDimension": "string",
               "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. 25) objects

Errors

For information about the errors that are common to all actions, see Common Errors (p. 36).

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
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V2

UpdateScalingPlan

Updates the scaling plan for 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": [
      {
         "MaxCapacity": number,
         "MinCapacity": number,
         "ResourceId": "string",
         "ScalableDimension": "string",
         "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. 34).

The request accepts the following data in JSON format.

ApplicationSource (p. 15)

A CloudFormation stack or set of tags.

Type: ApplicationSource (p. 19) object

Required: No

ScalingInstructions (p. 15)

The scaling instructions.

Type: Array of ScalingInstruction (p. 23) objects

Required: No

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.

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. 36).

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
- AWS SDK for JavaScript
- AWS SDK for PHP V3
- AWS SDK for Python
- AWS SDK for Ruby V2

Data Types

The AWS Auto Scaling Plans 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

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. 31) 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
- AWS SDK for Ruby V2

CustomizedScalingMetricSpecification

Represents a customized metric for a target tracking policy.

Contents

Dimensions

The dimensions of the metric.

Type: Array of MetricDimension (p. 21) 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
- AWS SDK for Ruby V2

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
- AWS SDK for Ruby V2

PredefinedScalingMetricSpecification

Represents a predefined metric for a target tracking 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.

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.

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
- AWS SDK for Ruby V2

ScalingInstruction

Specifies the scaling configuration for a scalable resource.

Contents

MaxCapacity

The maximum value to scale to in response to a scale out event.

Type: Integer

Required: Yes

MinCapacity

The minimum value to scale to in response to a scale in event.

Type: Integer

Required: Yes

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.

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

ServiceNamespace

The namespace of the AWS service.

Type: String

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

Required: Yes

TargetTrackingConfigurations

The target tracking scaling policies (up to 10).

Type: Array of TargetTrackingConfiguration (p. 32) 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
- AWS SDK for Ruby V2

ScalingPlan

Represents a scaling plan.

Contents

ApplicationSource

The application source.

Type: ApplicationSource (p. 19) object

Required: Yes

CreationTime

The Unix timestamp when the scaling plan was created.

Type: Timestamp

Required: No

ScalingInstructions

The scaling instructions.

Type: Array of ScalingInstruction (p. 23) 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 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 timestamp 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
- AWS SDK for Ruby V2

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.

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 of the scaling plan.

Type: Long

Required: Yes

ScalingPolicies

The scaling policies.

Type: Array of ScalingPolicy (p. 30) 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
- AWS SDK for Ruby V2

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.

Type: TargetTrackingConfiguration (p. 32) 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
- AWS SDK for Ruby V2

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
- AWS SDK for Ruby V2

TargetTrackingConfiguration

Represents a target tracking scaling policy.

Contents

CustomizedScalingMetricSpecification

A customized metric.

Type: CustomizedScalingMetricSpecification (p. 20) object

Required: No

DisableScaleIn

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

Type: PredefinedScalingMetricSpecification (p. 22) object

Required: No

ScaleInCooldown

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

The cooldown period is used to block subsequent scale in requests until it has expired. The intention is to scale in conservatively to protect your application's availability. However, if another alarm triggers a scale out policy during the cooldown period after a scale-in, AWS Auto Scaling scales out your scalable target immediately.

Type: Integer

Required: No

ScaleOutCooldown

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

While the cooldown period is in effect, the capacity that has been added by the previous scale out event that initiated the cooldown is calculated as part of the desired capacity for the next scale out. The intention is to continuously (but not excessively) scale out.

Type: Integer

Required: No

TargetValue

The target value for the metric. The range is 8.515920e-109 to 1.174271e+108 (Base 10) or 2e-360 to 2e360 (Base 2).

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
- AWS SDK for Ruby V2

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

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 AWS 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"
}
```