aws

用户指南

Application Auto Scaling

Application Auto Scaling: 用户指南

Table of Contents

什么是 Application Auto Scaling？

Application Auto Scaling 是一项面向开发人员和系统管理员的 Web 服务，他们需要一种解决方案来自动扩展其可扩展资源，用于超出 Amazon EC2 的各项 AWS 服务。使用 Application Auto Scaling，您可以为以下资源配置自动缩放：

- AppStream 2.0 支舰队
- Aurora 副本
- Amazon Comprehend 文档分类和实体识别程序终端节点
- DynamoDB 表和全局二级索引
- Amazon Elastic Container Service (ECS) 服务
- ElastiCache 适用于 Redis 集群（复制组）
- Amazon EMR 集群
- Amazon Keyspaces (for Apache Cassandra) 表
- Lambda 函数预置并发
- Amazon Managed Streaming for Apache Kafka (MSK) 代理存储
- Amazon Neptune 集群
- SageMaker 端点变体
- SageMaker 推理组件
- SageMaker 无服务器配置的并发性
- Spot 队列请求
- 由您自己的应用程序或服务提供的自定义资源。有关更多信息，请参阅GitHub存储库。

要查看上面列出的任何 AWS 服务的区域可用性，请参阅区域表 。

有关使用 Auto Scaling 组扩缩 Amazon EC2 实例队列的信息，请参阅 Amazon EC2 Auto Scaling 用户指南。

Application Auto Scaling 的功能

Application Auto Scaling 可以让您根据您定义的条件弹性伸缩可扩展资源。

- 目标跟踪扩展-根据特定 CloudWatch 指标的目标值扩展资源。

- 步进扩缩 - 根据一组扩缩调整来扩缩资源，这些调整因警报违例大小而异。
- 计划的扩缩— 仅扩展一次或按经常性计划扩缩资源。

使用 Application Auto Scaling

您可以使用以下界面配置扩缩，具体取决于要扩缩的资源：

- AWS Management Console – 提供可用于配置扩缩的 Web 界面。如果你已经注册了一个 AWS 账户，请通过登录来访问 Application Auto Scaling AWS Management Console。然后，打开服务控制台以查看简介中列出的资源之一。确保以与要使用的资源 AWS 区域 相同的方式打开控制台。

  > **Note**
  >
  > 并非所有资源都可以访问控制台。有关更多信息，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务。

- AWS Command Line Interface (AWS CLI) — 为大量用户提供命令，并在 Windows AWS 服务、macOS 和 Linux 上受支持。要开始使用，请参阅 设置 AWS CLI。有关更多信息，请参阅 AWS CLI 命令参考中的 application-autoscaling。
- AWS Tools for Windows PowerShell— 为那些在 PowerShell 环境中编写脚本的用户提供一系列 AWS 产品的命令。要开始使用，请参阅 AWS Tools for Windows PowerShell 用户指南。有关更多信息，请参阅 AWS Tools for PowerShell Cmdlet 参考。
- AWS 软件开发工具包 — 提供特定语言的 API 操作并处理许多连接细节，例如计算签名、处理请求重试和处理错误。有关更多信息，请参阅 AWS 软件开发工具包。
- HTTPS API – 提供了您使用 HTTPS 请求调用的低级别 API 操作。有关更多信息，请参阅 Application Auto Scaling API 参考。
- AWS CloudFormation— 支持使用 CloudFormation 模板配置缩放。有关更多信息，请参阅 使用 AWS CloudFormation 创建 Application Auto Scaling 资源。

要以编程方式连接到 AWS 服务，请使用终端节点。有关调用 Application Auto Scal ing 的终端节点的信息，请参阅《AWS 一般参考 中的 AWS 配额。

开始使用 Application Auto Scaling

本主题介绍关键概念，帮助您了解 Application Auto Scaling 并开始使用它。

可扩展目标

您创建的实体，用于指定要扩展的资源。每个可扩展目标都由服务命名空间、资源 ID 和可扩展维度（表示基础服务的一些容量维度）唯一标识。例如，Amazon ECS 服务支持弹性伸缩其任务计数，DynamoDB 表支持弹性伸缩该表及其全局二级索引的读写容量，Aurora 集群支持扩缩其副本计数。

> **Tip**
>
> 每个可扩展目标还具有最小容量和最大容量。扩缩策略永远不会高于或低于最小最大范围。您可以直接对超出此范围的基础资源进行带外更改，Application Auto Scaling 不知道这些资源。但是，无论何时调用扩缩策略或调用 `RegisterScalableTarget` API，Application Auto Scaling 都会检索当前容量并将其与最小容量和最大容量进行比较。如果该容量超出最小-最大范围，则会更新容量以符合设置的最小值和最大值。

横向缩减

当 Application Auto Scaling 自动减少可扩展目标的容量时，可扩展目标将横向缩减。设置扩缩策略后，将无法将可扩展目标横向缩减至最小容量以下。

扩展

当 Application Auto Scaling 自动增加可扩展目标的容量时，可扩展目标将横向扩展。设置扩缩策略后，将无法将可扩展目标横向扩展至最大容量以上。

扩缩策略

扩缩策略指示 Application Auto Scaling 跟踪特定的 CloudWatch 指标。然后，它确定当指标高于或低于某个阈值时要采取的扩缩操作。例如，如果集群中的 CPU 使用率开始上升，您可能希望横向扩展，而当其再次下降时横向缩减。

用于弹性伸缩的指标由目标服务发布，但您也可以将自己的指标发布到 CloudWatch，然后将其与扩缩策略一起使用。

扩缩活动之间的冷却时间可让资源在另一个扩缩活动开始之前稳定下来。Application Auto Scaling 在冷却时间内继续评估指标。冷却时间结束后，扩缩策略将根据需要启动另一个扩缩活动。在冷却时间生效时，如果需要根据当前指标值进行更大的横向扩展，则扩缩策略会立即横向扩展。

计划的操作

计划的操作在特定日期和时间弹性伸缩资源。它们通过修改可扩展目标的最小容量和最大容量来工作，因此可以通过设置更高的最小容量或更低的最大容量用于按计划横向缩减和横向扩展。例如，您可以使用计划的操作来扩缩周末不消耗资源的应用程序，方法是在星期五减少容量，并在下一个星期一增加容量。

您还可以使用计划的操作优化随时间推移的最小值和最大值，以适应预期流量高于正常流量的情况，例如营销活动或季节性波动。这样做可以帮助您在需要横向扩展更高以满足不断增加的使用量时提高性能，并在使用较少的资源时降低成本。

了解更多信息

AWS 可以与 Application Auto Scaling 一起使用的服务 - 本节向您介绍可以扩展的服务，并通过注册可扩展目标来帮助您设置弹性伸缩。本节还介绍 Application Auto Scaling 为访问目标服务中的资源而创建的每个 IAM 服务相关角色。

目标跟踪扩缩策略 - Application Auto Scaling 的主要功能之一是目标跟踪扩缩策略。了解目标跟踪策略如何根据配置的指标和目标值自动调整所需容量，使利用率保持在恒定水平。例如，您可以配置目标跟踪，使 Spot 实例集的平均 CPU 利用率保持在 50%。然后，Application Auto Scaling 根据需要启动或终止 EC2 实例，以使所有服务器的聚合 CPU 利用率保持在 50%。

AWS 可以与 Application Auto Scaling 一起使用的服务

Application Auto Scaling 与其他 AWS 服务集成，因此您可以添加扩展功能以满足应用程序的需求。Auto Scaling 是服务的一个可选功能，在几乎所有情况下都默认禁用该功能。

下表列出了可以与 Application Auto Scaling 配合使用的 AWS 服务，包括有关支持配置自动缩放的方法的信息。您还可以将 Application Auto Scaling 与自定义资源一起使用。

控制台访问 - 通过在目标服务的控制台中配置扩缩策略，您可以配置兼容的 AWS 服务以启动弹性伸缩。

CLI 访问 - 您可以配置兼容的 AWS 服务，以使用 AWS CLI启动弹性伸缩。

SDK 访问权限 — 您可以将兼容的 AWS 服务配置为使用 AWS SDK 启动自动扩展。

CloudFormation ac@@ ces s — 您可以使用 AWS CloudFormation 堆栈模板将兼容的 AWS 服务配置为启动自动扩展。有关更多信息，请参阅 使用 AWS CloudFormation 创建 Application Auto Scaling 资源。

AWS 服务	控制台访问[1]	CLI 访问	软件开发工具包访问	CloudFormation 访问
AppStream 2.0	是	是	是	是
Aurora	是	是	是	是
Amazon Comprehend	否	是	是	是
Amazon DynamoDB	是	是	是	是

AWS 服务	控制台访问[1]	CLI 访问	软件开发工具包访问	CloudFormation 访问
Amazon ECS	 是	 是	 是	 是
Amazon ElastiCache	 是	 是	 是	 是
Amazon EMR	 是	 是	 是	 是
Amazon Keyspaces	 是	 是	 是	 是
Lambda	 否	 是	 是	 是
Amazon MSK	 是	 是	 是	 是
Amazon Neptune	 否	 是	 是	 是
SageMaker	 是	 是	 是	 是

AWS 服务	控制台访问[1]	CLI 访问	软件开发工具包访问	CloudFormation 访问
竞价型实例集	 是	 是	 是	 是
自定义资源	 否	 是	 是	 是

[1] 用于配置扩展策略的控制台访问权限。大多数服务不支持从控制台配置定时扩展。目前，只有 Amazon AppStream 2.0 和 Spot 队列提供用于计划扩展的控制台访问权限。 ElastiCache

亚马逊 AppStream 2.0 和 Application Auto Scaling

您可以使用目标跟踪扩展策略、步进扩展策略和计划扩展来扩展 AppStream 2.0 舰队。

使用以下信息来帮助你将 AppStream 2.0 与 Application Auto Scaling 集成。

为 2.0 创建的 AppStream 服务相关角色

使用 Applicati on Auto Scaling 将 AppStream 2.0 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_AppStreamFleet`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `appstream.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将 AppStream 2.0 舰队注册为可扩展目标

Application Auto Scaling 需要一个可扩展的目标，然后才能为 AppStream 2.0 队列创建扩展策略或计划操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

如果您使用 AppStream 2.0 控制台配置 auto Scaling，则 AppStream 2.0 会自动为您注册可扩展目标。

如果要使用 AWS CLI 或其中一个 AWS SDK 配置自动扩展，则可以使用以下选项：

- AWS CLI:

  调用 AppStream 2.0 舰队的register-scalable-target指挥部。以下示例注册名为 `sample-fleet` 的队列的所需容量，最小容量为一个队列实例，最大容量为 5 个队列实例。

  ```
  aws application-autoscaling register-scalable-target \
     --service-namespace appstream \
     --scalable-dimension appstream:fleet:DesiredCapacity \
     --resource-id fleet/sample-fleet \
     --min-capacity 1 \
     --max-capacity 5
  ```

  如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 AppStream 2.0 资源的其他有用信息：

亚马逊 AppStream 2.0 管理指南中的 Fleet Auto Scaling for AppStream 2.0

Amazon Aurora 和 Application Auto Scaling

您可以使用目标跟踪扩缩策略、分步扩缩策略和计划的扩缩来扩展 Aurora 数据库集群。

使用以下信息可帮助您将 Aurora 与 Application Auto Scaling 集成。

为 Aurora 创建的服务相关角色

在 Application Auto Scaling 中将 Aurora 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_RDSCluster`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `rds.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将 Aurora 数据库集群注册为可扩展目标

Application Auto Scaling 需要一个可扩展目标，然后才能为 Aurora 集群创建扩缩策略或计划的操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

如果您使用 Aurora 控制台配置弹性伸缩，Aurora 会自动为您注册一个可扩展的目标。

如果要使用 AWS CLI 或其中一个 SD AWS K 来配置 auto Scaling，则可以使用以下选项：

- AWS CLI:

  为 Aurora 集群调用 register-scalable-target 命令。以下示例在名为 `my-db-cluster` 的集群中注册 Aurora 副本的计数，最小容量为一个 Aurora 副本，最大容量为 8 个 Aurora 副本。

```
aws application-autoscaling register-scalable-target \
```

```
  --service-namespace rds \
  --scalable-dimension rds:cluster:ReadReplicaCount \
  --resource-id cluster:my-db-cluster \
  --min-capacity 1 \
  --max-capacity 8
```

如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 Aurora 资源的其他有用信息：

Amazon RDS 用户指南中的将 Amazon Aurora Auto Scaling 与 Aurora 副本一起使用

Amazon Comprehend 和 Application Auto Scaling

您可以使用目标跟踪扩缩策略和计划的扩缩来扩展 Amazon Comprehend 文档分类和实体识别程序终端节点。

使用以下信息可帮助您将 Amazon Comprehend 与 Application Auto Scaling 集成。

为 Amazon Comprehend 创建的服务相关角色

在 Application Auto Scaling 中将 Amazon Comprehend 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_ComprehendEndpoint`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `comprehend.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将 Amazon Comprehend 资源注册为可扩展目标

Application Auto Scaling 需要一个可扩展目标，然后才能为 Amazon Comprehend 文档分类或实体识别程序终端节点创建扩缩策略或计划的操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

要使用 AWS CLI 或其中一个 AWS SDK 配置自动扩展，您可以使用以下选项：

- AWS CLI:

  为文档分类终端节点调用 register-scalable-target 命令。以下示例使用终端节点的 ARN 注册文档分类程序终端节点模型要使用的所需推理单位数，最小容量为一个推理单位，最大容量为三个推理单位。

```
aws application-autoscaling register-scalable-target \
  --service-namespace comprehend \
  --scalable-dimension comprehend:document-classifier-endpoint:DesiredInferenceUnits
 \
  --resource-id arn:aws:comprehend:us-west-2:123456789012:document-classifier-
endpoint/EXAMPLE \
  --min-capacity 1 \
  --max-capacity 3
```

  如果成功，该命令会返回可扩展目标的 ARN。

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

为实体识别程序终端节点调用 register-scalable-target 命令。以下示例使用终端节点的 ARN 注册实体识别程序终端节点模型要使用的所需推理单位数，最小容量为一个推理单位，最大容量为三个推理单位。

```
aws application-autoscaling register-scalable-target \
  --service-namespace comprehend \
  --scalable-dimension comprehend:entity-recognizer-endpoint:DesiredInferenceUnits \
  --resource-id arn:aws:comprehend:us-west-2:123456789012:entity-recognizer-
endpoint/EXAMPLE \
  --min-capacity 1 \
  --max-capacity 3
```

如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 Amazon Comprehend 资源的其他有用信息：

Amazon Comprehend Developer Guide 中的 Auto scaling with endpoints

Amazon DynamoDB 和 Application Auto Scaling

您可以使用目标跟踪扩缩策略和计划的扩缩来扩展 DynamoDB 表和全局二级索引。

使用以下信息可帮助您将 DynamoDB 与 Application Auto Scaling 集成。

为 DynamoDB 创建的服务相关角色

使用 Application Auto Scaling 将 DynamoDB 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_DynamoDBTable`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `dynamodb.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将 DynamoDB 资源注册为可扩展目标

Application Auto Scaling 需要一个可扩展目标，然后才能为 DynamoDB 表或全局二级索引创建扩缩策略或计划的操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

如果您使用 DynamoDB 控制台配置弹性伸缩，DynamoDB 会自动为您注册一个可扩展的目标。

如果要使用 AWS CLI 或其中一个 AWS SDK 配置自动扩展，则可以使用以下选项：

- AWS CLI:

  调用register-scalable-target命令获取表的写入容量。以下示例注册名为 `my-table` 的表的预置写入容量，最小容量为 5 个写入容量单位，最大容量为 10 个写入容量单位。

  ```
  aws application-autoscaling register-scalable-target \
    --service-namespace dynamodb \
    --scalable-dimension dynamodb:table:WriteCapacityUnits \
    --resource-id table/my-table \
    --min-capacity 5 \
    --max-capacity 10
  ```

  如果成功，该命令会返回可扩展目标的 ARN。

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

调用register-scalable-target命令获取表的读取容量。以下示例注册名为 my-table 的表的预置读取容量，最小容量为 5 个读取容量单位，最大容量为 10 个读取单位。

```
aws application-autoscaling register-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits \
  --resource-id table/my-table \
  --min-capacity 5 \
  --max-capacity 10
```

如果成功，该命令会返回可扩展目标的 ARN。

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

调用register-scalable-target命令获取全局二级索引的写入容量。以下示例注册名为 my-table-index 的全局二级索引的预置写入容量，最小容量为 5 个写入容量单位，最大容量为 10 个写入容量单位。

```
aws application-autoscaling register-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:index:WriteCapacityUnits \
  --resource-id table/my-table/index/my-table-index \
  --min-capacity 5 \
  --max-capacity 10
```

如果成功，该命令会返回可扩展目标的 ARN。

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

```
}
```

调用register-scalable-target命令获取全局二级索引的读取容量。以下示例注册名为 `my-table-index` 的全局二级索引的预置读取容量，最小容量为 5 个读取容量单位，最大容量为 10 个读取容量单位。

```
aws application-autoscaling register-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:index:ReadCapacityUnits \
  --resource-id table/my-table/index/my-table-index \
  --min-capacity 5 \
  --max-capacity 10
```

如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 DynamoDB 资源的其他有用信息：

- Amazon DynamoDB 开发人员指南中的使用 DynamoDB Auto Scaling 管理吞吐量
- 在《亚马逊 DynamoD B 开发者指南》中评估表的自动缩放设置
- AWS CloudFormation 如何使用博客上的 DynamoDB 表和索引配置自动缩放 AWS

您还可以在中找到有关定时扩展的教程教程：通过 AWS CLI开始使用计划扩缩。在该教程中，您将了解配置扩缩以便您的 DynamoDB 表按计划的时间扩展的基本步骤。

Amazon ECS 和 Application Auto Scaling

您可以使用目标跟踪扩缩策略、分步扩缩策略和计划的扩缩来扩展 ECS 服务。

使用以下信息可帮助您将 Amazon ECS 与 Application Auto Scaling 集成。

为 Amazon ECS 创建的服务相关角色

在 Application Auto Scaling 中将 Amazon ECS 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_ECSService`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `ecs.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将 ECS 服务注册为可扩展目标

Application Auto Scaling 需要一个可扩展目标，然后才能为 Amazon ECS 服务创建扩缩策略或计划的操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

如果您使用 Amazon ECS 控制台配置弹性伸缩，Amazon ECS 会自动为您注册一个可扩展的目标。

如果要使用 AWS CLI 或其中一个 AWS SDK 配置自动扩展，则可以使用以下选项：

- AWS CLI:

  为 Amazon ECS 服务调用 register-scalable-target 命令。以下示例为名为 `sample-app-service` 的服务（在 `default` 集群上运行）注册可扩展目标，最小任务计数为一个任务，最大任务计数为 10 个任务。

```
aws application-autoscaling register-scalable-target \
```

```
  --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/default/sample-app-service \
  --min-capacity 1 \
  --max-capacity 10
```

如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 Amazon ECS 资源的其他有用信息：

- Amazon 弹性容器服务开发者指南中的服务自动扩展
- 在《Amazon 弹性容器服务最佳实践指南》中配置服务自动扩展

> **Note**
>
> 有关在 Amazon ECS 部署过程中暂停扩展流程的说明，请参阅以下文档：Amazon 弹性容器服务开发者指南中的服务自动扩展和部署

ElastiCache 适用于 Redis 和应用程序 Auto Scaling

您可以使用目标跟踪扩展 ElastiCache 策略和计划扩展来扩展 Redis 复制组。

使用以下信息来帮助您 ElastiCache 与 Application Auto Scaling 集成。

为 ElastiCache 创建的服务相关角色

使用 Applicati on Auto Scaling 将 ElastiCache 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_ElastiCacheRG`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `elasticache.application-autoscaling.amazonaws.com`

ElastiCache 使用 Application Auto Scaling 将 Redis 复制组注册为可扩展目标

Application Auto Scaling 需要一个可扩展的目标，然后才能为 ElastiCache 复制组创建扩展策略或计划操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

如果您使用 ElastiCache 控制台配置 auto Scaling，则 ElastiCache 会自动为您注册可扩展目标。

如果要使用 AWS CLI 或其中一个 AWS SDK 配置自动扩展，则可以使用以下选项：

- AWS CLI:

  调用 ElastiCache复制组的register-scalable-target命令。以下示例注册名为 `mycluster` 的复制组的所需节点组数量，最小容量为一个，最大容量为 5 个。

```
aws application-autoscaling register-scalable-target \
   --service-namespace elasticache \
   --scalable-dimension elasticache:replication-group:NodeGroups \
   --resource-id replication-group/mycluster \
   --min-capacity 1 \
   --max-capacity 5
```

  如果成功，该命令会返回可扩展目标的 ARN。

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

以下示例注册名为 mycluster 的复制组每个节点组所需的副本数量，最小容量为 1，最大容量为 5。

```
aws application-autoscaling register-scalable-target \
   --service-namespace elasticache \
   --scalable-dimension elasticache:replication-group:Replicas \
   --resource-id replication-group/mycluster \
   --min-capacity 1 \
   --max-capacity 5
```

如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 ResourceId、ScalableDimension、ServiceNamespace、MinCapacity 和 MaxCapacity 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 ElastiCache 资源的其他有用信息：

A mazon f ElastiCache or Redis 用户指南中适用于 Redis 集群的 Auto ElastiCache S caling

Amazon Keyspaces（针对 Apache Cassandra）和 Application Auto Scaling

您可以使用目标跟踪扩缩策略和计划的扩缩来扩展 Amazon Keyspaces 表。

使用以下信息可帮助您将 Amazon Keyspaces 与 Application Auto Scaling 集成。

为 Amazon Keyspaces 创建的服务相关角色

在 Application Auto Scaling 中将 Amazon Keyspaces 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_CassandraTable`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `cassandra.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将 Amazon Keyspaces 表注册为可扩展目标

Application Auto Scaling 需要一个可扩展目标，然后才能为 Amazon Keyspaces 表创建扩缩策略或计划的操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

如果您使用 Amazon Keyspaces 控制台配置弹性伸缩，Amazon Keyspaces 会自动为您注册一个可扩展的目标。

如果要使用 AWS CLI 或其中一个 SD AWS K 来配置 auto Scaling，则可以使用以下选项：

- AWS CLI:

  为 Amazon Keyspaces 表调用 register-scalable-target 命令。以下示例注册名为 `mytable` 的表的预置写入容量，最小容量为 5 个写入容量单位，最大容量为 10 个写入容量单位。

```
aws application-autoscaling register-scalable-target \
```

```
  --service-namespace cassandra \
  --scalable-dimension cassandra:table:WriteCapacityUnits \
  --resource-id keyspace/mykeyspace/table/mytable \
  --min-capacity 5 \
  --max-capacity 10
```

如果成功，该命令会返回可扩展目标的 ARN。

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

以下示例注册名为 mytable 的表的预置读取容量，最小容量为 5 个读取容量单位，最大容量为 10 个读取容量单位。

```
aws application-autoscaling register-scalable-target \
  --service-namespace cassandra \
  --scalable-dimension cassandra:table:ReadCapacityUnits \
  --resource-id keyspace/mykeyspace/table/mytable \
  --min-capacity 5 \
  --max-capacity 10
```

如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 ResourceId、ScalableDimension、ServiceNamespace、MinCapacity 和 MaxCapacity 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 Amazon Keyspaces 资源的其他有用信息：

使用 Amazon Keyspaces 管理吞吐容量在《亚马逊密钥空间（适用于 Ap ache Cassandra）开发者指南》中自动扩展

AWS Lambda 和应用程序 Auto Scaling

您可以使用目标跟踪扩展策略和计划扩展来扩展 AWS Lambda 预配置的并发量。

使用以下信息可帮助您将 Lambda 与 Application Auto Scaling 集成。

为 Lambda 创建的服务相关角色

使用 Application Auto Scaling 将 Lambda 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_LambdaConcurrency`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `lambda.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将 Lambda 函数注册为可扩展目标

Application Auto Scaling 需要一个可扩展目标，然后才能为 Lambda 函数创建扩缩策略或计划的操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

要使用 AWS CLI 或其中一个 AWS SDK 配置自动扩展，您可以使用以下选项：

- AWS CLI:

  为 Lambda 函数调用 register-scalable-target 命令。以下示例为名为 `my-function` 的函数注册别名为 `BLUE` 的预置并发，最小容量为 0，最大容量为 100。

```
aws application-autoscaling register-scalable-target \
  --service-namespace lambda \
```

```
  --scalable-dimension lambda:function:ProvisionedConcurrency \
  --resource-id function:my-function:BLUE \
  --min-capacity 0 \
  --max-capacity 100
```

如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 Lambda 函数的其他有用信息：

- 在《开发者指南》中@@ 配置预配置的并发性 AWS Lambda
- 安排 Lambda 预配置并发以应对博客上反复出现的高峰使用量 AWS

Amazon Managed Streaming for Apache Kafka (MSK) 和 Application Auto Scaling

您可以使用目标跟踪扩缩策略横向扩展 Amazon MSK 集群存储。按目标跟踪策略横向缩减已禁用。

使用以下信息可帮助您将 Amazon MSK 与 Application Auto Scaling 集成。

为 Amazon MSK 创建的服务相关角色

在 Application Auto Scaling 中将 Amazon MSK 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- AWSServiceRoleForApplicationAutoScaling_KafkaCluster

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- kafka.application-autoscaling.amazonaws.com

使用 Application Auto Scaling 将 Amazon MSK 集群存储注册为可扩展目标

Application Auto Scaling 需要一个可扩展的目标，然后才能为 Amazon MSK 集群的每个代理的存储卷大小创建扩缩策略。可扩展目标是 Application Auto Scaling 可以扩展的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

如果您使用 Amazon MSK 控制台配置弹性伸缩，Amazon MSK 会自动为您注册一个可扩展的目标。

如果要使用 AWS CLI 或其中一个 SD AWS K 来配置 auto Scaling，则可以使用以下选项：

- AWS CLI:

  为 Amazon MSK 集群调用 register-scalable-target 命令。以下示例注册 Amazon MSK 集群每个代理的存储卷大小，最小容量为 100 GiB，最大容量为 800 GiB。

```
aws application-autoscaling register-scalable-target \
   --service-namespace kafka \
   --scalable-dimension kafka:broker-storage:VolumeSize \
   --resource-id arn:aws:kafka:us-east-1:123456789012:cluster/demo-
cluster-1/6357e0b2-0e6a-4b86-a0b4-70df934c2e31-5 \
   --min-capacity 100 \
   --max-capacity 800
```

  如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

> **Note**
>
> 当 Amazon MSK 集群是可扩展目标时，横向缩减将禁用且无法启用。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 Amazon MSK 资源的其他有用信息：

《适用于 A pache 的亚马逊托管流媒体 Kafka 开发者指南》中的@@ 自动扩展

Amazon Neptune 和 Application Auto Scaling

您可以使用目标跟踪扩缩策略和计划的扩缩来扩展 Neptune 集群。

使用以下信息可帮助您将 Neptune 与 Application Auto Scaling 集成。

为 Neptune 创建的服务关联角色

在 Application A uto Scaling 中将 Neptune 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_NeptuneCluster`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `neptune.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将 Neptune 集群注册为可扩展目标

Application Auto Scaling 需要一个可扩展目标，然后才能为 Neptune 集群创建扩缩策略或计划的操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

要使用 AWS CLI 或其中一个 AWS SDK 配置自动扩展，您可以使用以下选项：

- AWS CLI:

  调用 Neptune 集群的register-scalable-target命令。以下示例注册名为 `mycluster` 的集群的所需容量，最小容量为一个，最大容量为八个。

  ```
  aws application-autoscaling register-scalable-target \
    --service-namespace neptune \
    --scalable-dimension neptune:cluster:ReadReplicaCount \
    --resource-id cluster:mycluster \
    --min-capacity 1 \
    --max-capacity 8
  ```

  如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 Neptune 资源的其他有用信息：

Neptune 用户指南中的自动扩缩 Amazon Neptune 数据库集群中的副本数量

亚马逊 SageMaker 和 Application Auto Scaling

您可以使用目标跟踪扩展 SageMaker 策略、步进扩展策略和计划扩展来扩展终端节点变体、无服务器端点的预配置并发以及推理组件。

使用以下信息来帮助您 SageMaker 与 Application Auto Scaling 集成。

为 SageMaker 创建的服务相关角色

使用 Applicati on Auto Scaling 将 SageMaker 资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_SageMakerEndpoint`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `sagemaker.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将 SageMaker 端点变体注册为可扩展目标

Application Auto Scaling 需要一个可扩展的目标，然后才能为 SageMaker 模型（变体）创建扩展策略或计划操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

如果您使用 SageMaker 控制台配置 auto Scaling，则 SageMaker 会自动为您注册可扩展目标。

如果要使用 AWS CLI 或其中一个 SD AWS K 来配置 auto Scaling，则可以使用以下选项：

- AWS CLI:

  调用register-scalable-target命令获取产品变体。以下示例为名为 `my-variant` 的产品变体（在 `my-endpoint` 端点上运行）注册所需的实例计数，最小容量为一个实例，最大容量为八个实例。

```
aws application-autoscaling register-scalable-target \
```

```
  --service-namespace sagemaker \
  --scalable-dimension sagemaker:variant:DesiredInstanceCount \
  --resource-id endpoint/my-endpoint/variant/my-variant \
  --min-capacity 1 \
  --max-capacity 8
```

如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

使用 Application Auto Scaling 将无服务器端点的预置并发注册为可扩展目标

Application Auto Scaling 也需要一个可扩展目标，然后才能为无服务器端点预置并发创建扩缩策略或计划的操作。

如果您使用 SageMaker 控制台配置 auto Scaling，则 SageMaker 会自动为您注册可扩展目标。

如果没有自动注册，请使用以下方法之一注册可扩展目标：

- AWS CLI:

  调用register-scalable-target命令获取产品变体。以下示例为名为 `my-variant` 的产品变体（在 `my-endpoint` 端点上运行）注册预置并发，最小容量为一个实例，最大容量为十个实例。

```
aws application-autoscaling register-scalable-target \
  --service-namespace sagemaker \
  --scalable-dimension sagemaker:variant:DesiredProvisionedConcurrency \
  --resource-id endpoint/my-endpoint/variant/my-variant \
  --min-capacity 1 \
  --max-capacity 10
```

如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

使用 Application Auto Scaling 将推理组件注册为可扩展目标

Application Auto Scaling 也需要一个可扩展目标，然后才能为推理组件创建扩展策略或计划的操作。

- AWS CLI:

  调用推理组件的register-scalable-target命令。以下示例为名为 `my-inference-component` 的推理组件注册所需的副本计数，最小容量为零个副本，最大容量为三个副本。

```
aws application-autoscaling register-scalable-target \
  --service-namespace sagemaker \
  --scalable-dimension sagemaker:inference-component:DesiredCopyCount \
  --resource-id inference-component/my-inference-component \
  --min-capacity 0 \
  --max-capacity 3
```

  如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在亚马逊 SageMaker 开发者指南中找到有关扩展 SageMaker 资源的其他有用信息：

- 自动缩放 Amazon SageMaker 模型
- 自动扩展无服务器端点的预配置并发度
- 为多模型终端节点部署设置 auto Scaling 策略
- 自动缩放异步端点

> **Note**
>
> 2023 年， SageMaker 推出了基于实时推理端点的新推理功能。您可以使用终端节点配置创建 SageMaker 终端节点，该端点配置定义了终端节点的实例类型和初始实例数。然后，创建一个推理组件，这是一个 SageMaker 托管对象，可用于将模型部署到终端节点。有关扩展推理组件的信息，请参阅 Amazon SageMaker 添加了新的推理功能以帮助降低基础模型部署成本和延迟，以及使用博客上的 Ama SageMaker zon 最新功能将模型部署成本平均降低 50%。AWS

Amazon EC2 Spot 实例集和 Application Auto Scaling

您可以使用目标跟踪扩缩策略、分步扩缩策略和计划的扩缩来扩展 Spot 实例集。

使用以下信息可帮助您将 Spot 实例集与 Application Auto Scaling 集成。

为 Spot 实例集创建的服务相关角色

在 Application Auto Scaling 中将 Spot 队列资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `ec2.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将 Spot 实例集注册为可扩展目标

Application Auto Scaling 需要一个可扩展目标，然后才能为 Spot 实例集创建扩缩策略或计划的操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

如果您使用 Spot 实例集控制台配置弹性伸缩，Spot 实例集会自动为您注册一个可扩展的目标。

如果要使用 AWS CLI 或其中一个 SD AWS K 来配置 auto Scaling，则可以使用以下选项：

- AWS CLI:

  为 Spot 实例集调用 register-scalable-target 命令。以下示例使用其请求 ID 注册 Spot 实例集的目标容量，最小容量为两个实例，最大容量为 10 个实例。

  ```
  aws application-autoscaling register-scalable-target \
    --service-namespace ec2 \
    --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
    --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
    --min-capacity 2 \
    --max-capacity 10
  ```

  如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 `ResourceId`、`ScalableDimension`、`ServiceNamespace`、`MinCapacity` 和 `MaxCapacity` 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展 Spot 队列的其他有用信息：

Amazon EC2 用户指南中的 Spot 实例集的自动扩展

自定义资源和 Application Auto Scaling

您可以使用目标跟踪扩缩策略、分步扩缩策略和计划的扩缩来扩展自定义资源。

使用以下信息可帮助您将自定义资源与 Application Auto Scaling 集成。

为自定义资源创建服务相关角色

使用 Applicati on Auto Scaling 将自定义资源注册为可扩展目标 AWS 账户 时，将在您的中自动创建以下服务相关角色。此角色允许 Application Auto Scaling 在您的账户中执行受支持的操作。有关更多信息，请参阅Application Auto Scaling 的服务相关角色。

- `AWSServiceRoleForApplicationAutoScaling_CustomResource`

服务相关角色使用的服务委托人

上一节中的服务相关角色只能由为角色定义的信任关系授权的服务委托人担任。Application Auto Scaling 使用的服务相关角色为以下服务委托人授予访问权限：

- `custom-resource.application-autoscaling.amazonaws.com`

使用 Application Auto Scaling 将自定义资源注册为可扩展目标

Application Auto Scaling 需要一个可扩展目标，然后才能为自定义资源创建扩缩策略或计划的操作。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。可扩展目标由资源 ID、可扩展维度和命名空间的组合唯一标识。

要使用 AWS CLI 或其中一个 AWS SDK 配置自动扩展，您可以使用以下选项：

- AWS CLI:

为自定义资源调用 register-scalable-target 命令。以下示例将自定义资源注册为可扩展目标，最小所需计数为一个容量单位，最大所需计数为 10 个容量单位。custom-resource-id.txt 文件包含一个标识资源 ID 的字符串，它表示通过 Amazon API Gateway 终端节点到自定义资源的路径。

```
aws application-autoscaling register-scalable-target \
  --service-namespace custom-resource \
  --scalable-dimension custom-resource:ResourceType:Property \
  --resource-id file://~/custom-resource-id.txt \
  --min-capacity 1 \
  --max-capacity 10
```

custom-resource-id.txt 的内容：

```
https://example.execute-api.us-west-2.amazonaws.com/prod/
scalableTargetDimensions/1-23456789
```

如果成功，该命令会返回可扩展目标的 ARN。

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

- AWS 软件开发工具包：

  调用 RegisterScalableTarget 操作并提供 ResourceId、ScalableDimension、ServiceNamespace、MinCapacity 和 MaxCapacity 作为参数。

相关资源

如果您刚刚开始使用 Application Auto Scaling，可以在以下文档中找到有关扩展自定义资源的其他有用信息：

GitHub存储库

进行设置以开始使用 Application Auto Scaling

首次设置 Application Auto Scaling 时请完成这一部分中的任务：

主题

注册到 AWS

如果您还没有 AWS 账户，请完成以下步骤来创建一个。

注册 AWS 账户

1. 打开 https://portal.aws.amazon.com/billing/signup。
2. 按照屏幕上的说明进行操作。

   在注册时，您将接到一通电话，要求您使用电话键盘输入一个验证码。

   当您注册 AWS 账户 时，系统将会创建一个 AWS 账户根用户。根用户有权访问该账户中的所有 AWS 服务和资源。作为安全最佳实践，请 为管理用户分配管理访问权限，并且只使用根用户执行需要根用户访问权限的任务。

在 AWS 区域 中使用 Application Auto Scaling

Application Auto Scaling 在多个 AWS 区域 中可用。利用全球 AWS 账户，您可以在大多数区域中使用资源。当利用中国区域的资源使用 Application Auto Scaling 时，请记住您必须拥有单独的 Amazon Web Services（中国）账户。此外，Application Auto Scaling 的实现方式上存在一些差异。有关在中国区域中使用 Application Auto Scaling 的更多信息，请参阅中国的 Application Auto Scaling。

设置好您的 AWS 账户后，继续阅读下一个主题：设置 AWS CLI。

设置 AWS CLI

AWS Command Line Interface (AWS CLI) 是一款用于管理 AWS 服务的统一开发工具，包括 Application Auto Scaling。按照以下步骤下载和配置 AWS CLI。

设置 AWS CLI

1. 下载、安装和配置 AWS CLI 版本 1 或 2。版本 1 和版本 2 中的 Application Auto Scaling 功能完全相同。有关说明，请参阅《AWS Command Line Interface 用户指南》中的以下主题：

   AWS CLI 版本 1

   - 安装、更新和卸载 AWS CLI
   - 配置 AWS CLI

   AWS CLI 版本 2

   - 安装或更新最新版本的 AWS CLI
   - 快速设置

   > **Note**
   >
   > 要进行 CLI 访问，您需要访问密钥 ID 和秘密访问密钥。如果可能，请使用临时凭证代替长期访问密钥。临时凭证包括访问密钥 ID、秘密访问密钥，以及一个指示凭证何时到期的安全令牌。为提高 AWS 账户的安全性，我们强烈建议您不要使用与您的 AWS 账户根用户关联的访问凭证。有关更多信息，请参阅《AWS 一般参考》中的 编程访问 和《IAM 用户指南》中的 IAM 中的安全最佳实践。

2. 为了确认 AWS CLI 配置文件的配置正确无误，请在命令窗口中运行以下命令：

   ```
   aws configure
   ```

   如果您的配置文件已正确配置，您应该看到类似于以下内容的输出。

   ```
   AWS Access Key ID [****************52FQ]:
   AWS Secret Access Key [****************xgyZ]:
   Default region name [us-east-1]:
   Default output format [json]:
   ```

3. 运行以下命令确认是否安装了 AWS CLI 的 Application Auto Scaling 命令。

   ```
   aws application-autoscaling help
   ```

通过命令行将 AWS CloudShell 与 Application Auto Scaling 结合使用

AWS CloudShell 允许您跳过在开发环境中安装 AWS CLI 的过程，直接在 AWS Management Console 中使用它。除无需安装外，您还无需配置凭证，也不需要指定区域。您的 AWS Management Console 会话会将此上下文提供给 AWS CLI。您可以在 支持的 AWS 区域 中使用 AWS CloudShell。

您可以运行 AWS CLI 命令对服务使用您的首选 shell（Bash、PowerShell 或 Z shell）。

您可以用以下两种方法之一从 AWS Management Console 启动 AWS CloudShell：

- 选择控制台导航栏中的 AWS CloudShell 图标。它位于搜索框的右侧。
- 使用控制台导航栏上的搜索框搜索 CloudShell，然后选择 CloudShell 选项。

首次在新的浏览器窗口中启动 AWS CloudShell 时，欢迎面板将显示并列出主要功能。关闭此面板后，系统会在 shell 配置和转发控制台凭证的同时提供状态更新。当系统显示命令提示符时，表示 shell 已经准备就绪，可以进行交互。

有关此服务的更多信息，请参阅 AWS CloudShell 用户指南。

使用 AWS CloudFormation 创建 Application Auto Scaling 资源

Application Auto Scaling 与 AWS CloudFormation一项服务集成，可帮助您对 AWS 资源进行建模和设置，从而减少创建和管理资源和基础架构所花费的时间。您可以创建一个描述所需所有 AWS 资源的模板，并为您预 AWS CloudFormation 置和配置这些资源。

使用时 AWS CloudFormation，您可以重复使用模板来一致且重复地设置 Application Auto Scaling 资源。只需描述一次您的资源，然后在多个 AWS 账户 区域中一遍又一遍地配置相同的资源。

Application Auto Scaling 和 AWS CloudFormation 模板

要为 Application Auto Scaling 和相关服务预置和配置资源，您必须了解 AWS CloudFormation 模板。模板是 JSON 或 YAML 格式的文本文件。这些模板描述了您要在 AWS CloudFormation 堆栈中配置的资源。如果你不熟悉 JSON 或 YAML，可以使用 D AWS CloudFormation esigner 来帮助你开始使用 AWS CloudFormation 模板。有关更多信息，请参阅《AWS CloudFormation 用户指南》中的什么是 AWS CloudFormation Designer？。

为 Application Auto Scaling 资源创建堆栈模板时，必须提供以下内容：

- 目标服务的命名空间（例如 **appstream**）。要获取服务命名空间，请参阅AWS::ApplicationAutoScaling::ScalableTarget参考资料。
- 与目标资源关联的可扩展维度（例如 **appstream:fleet:DesiredCapacity**）。请参阅AWS::ApplicationAutoScaling::ScalableTarget参考资料以获取可缩放的维度。
- 目标资源的资源 ID（例如 **fleet/sample-fleet**）。有关特定资源 ID 的语法和示例的信息，请参阅AWS::ApplicationAutoScaling::ScalableTarget参考资料。
- 目标资源的服务相关角色（例如 **arn:aws:iam::012345678910:role/aws-service-role/appstream.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_AppStreamFleet**）。请参阅 服务相关角色 ARN 参考 表以获取角色 ARN。

要了解有关 Application Auto Scaling 资源的更多信息，请参阅 AWS CloudFormation 用户指南中的 Application Auto Scaling 参考。

示例模板代码段

您可以在《AWS CloudFormation 用户指南》的以下章节中找到要包含在 AWS CloudFormation 模板中的示例片段：

- 有关扩展策略和计划操作的示例，请参阅使用配置应用程序 Auto Scaling 资源 AWS CloudFormation。
- 有关扩展策略的更多示例，请参阅AWS::ApplicationAutoScaling::ScalingPolicy。

了解更多关于 AWS CloudFormation

要了解更多信息 AWS CloudFormation，请参阅以下资源：

- AWS CloudFormation
- AWS CloudFormation 用户指南
- AWS CloudFormation API 引用
- AWS CloudFormation 命令行界面用户指南

计划扩展

通过计划扩缩，您可以根据可预测的负载变化，通过创建在特定时间增加或减少容量的计划操作，为应用程序设置自动扩缩。这使您可以主动扩缩应用程序，以适应可预测的负载变化。

例如，假设您每周遇到规律的流量模式，即负载在一周的中间增加，而在接近周末时会下降。您可以在 Application Auto Scaling 中配置与此模式一致的以下扩缩计划：

- 周三上午，一项计划操作通过增加先前设置的可扩展目标的最小容量来增加容量。
- 周五晚上，另一项计划操作通过降低先前设置的可扩展目标的最大容量来减少容量。

利用这些计划的扩缩操作，您可以优化成本和性能。您的应用程序有足够的容量来处理一周中间的流量高峰，但不会在其他时间过度配置不需要的容量。

您可以同时使用计划的扩缩和扩缩策略，以获得主动和被动扩缩方法的优势。运行计划的扩缩操作后，扩缩策略可以继续决定是否进一步扩缩容量。这有助于确保您有足够的容量来处理应用程序的负载。当您的应用程序扩展以满足需求时，当前容量必须在计划操作设置的最小容量和最大容量范围内。

主题

计划扩缩的工作原理

本主题描述了定时扩展的工作原理，并介绍了有效使用定时扩展所需的关键注意事项。

内容

工作方式

要使用计划扩缩，请创建指示 Application Auto Scaling 在特定时间执行扩缩活动的计划操作。创建计划的操作时，请指定可扩展目标、应进行扩缩活动的时间以及最小和最大容量。您可以创建仅扩展一次或按重复计划扩展的计划操作。

在指定的时间，Application Auto Scaling 通过将当前容量与指定的最小容量和最大容量进行比较，根据新容量值进行扩展。

- 如果当前容量小于指定的最小容量，Application Auto Scaling 将横向扩展（增加容量）到指定的最小容量。
- 如果当前容量大于指定的最大容量，Application Auto Scaling 将横向缩减（减少容量）到指定的最大容量。

注意事项

创建计划的操作时，请记住以下内容：

- 计划操作将 `MinCapacity` 和 `MaxCapacity` 设置为由计划操作在指定的日期和时间指定的内容。请求可以选择只包含这些大小中的一个。例如，您可以创建仅指定最小容量的计划操作。但是，在某些情况下，您必须包括两种大小，以确保新的最小容量不大于最大容量，或者新的最大容量不小于最小容量。
- 预设情况下，您设置的重复计划采用协调世界时 (UTC)。您可以更改时间以符合本地时区或您的网络中其他部分的时区。如果您指定的时区遵守夏令时，则操作会自动调整夏令时 (DST)。有关更多信息，请参阅 使用 cron 表达式安排重复发生的扩缩操作。
- 您可以临时关闭可扩展目标的计划扩缩。这有助于防止计划操作处于活动状态，而无需将其删除。然后，当您想要再次使用时，您可以恢复计划的扩展。有关更多信息，请参阅 暂停和恢复 Application Auto Scaling 扩缩。
- 将会保证同一可扩展目标的计划操作运行顺序，但不保证不同可扩展目标中的计划操作的执行顺序。
- 要成功完成计划操作，目标服务中的指定资源必须位于可扩展状态。如果不是此状态，则请求将会失败，并返回错误消息，例如：`Resource Id [ActualResourceId] is not scalable. Reason: The status of all DB instances must be 'available' or 'incompatible-parameters'`。
- 由于 Application Auto Scaling 和目标服务的分布式特性，计划操作触发时间与目标服务实际执行扩缩操作的时间之间的延迟可能有几秒钟。因为系统将按照指定操作的顺序来运行计划的操作，所以开始时间彼此接近的计划操作可能需要更长时间才能运行。

计划操作创建、管理和删除的常用命令

使用计划扩缩的常用命令包括：

- register-scalable-target注册 AWS 或自定义资源作为可扩展目标（Application Auto Scaling 可以扩展的资源），以及暂停和恢复扩展。
- put-scheduled-action添加或修改现有可扩展目标的计划操作。
- describe-scaling-activities返回有关某个 AWS 区域中扩展活动的信息。
- describe-scheduled-actions返回有关某个 AWS 区域中计划操作的信息。
- delete-scheduled-action删除预设操作。

相关资源

有关使用定时扩展的详细示例，请参阅 C AWS ompute Blog 上的博客文章 “计划 AWS Lambda 预配置并发以了解反复出现的峰值使用量”。

有关演练如何使用示例 AWS 资源创建计划操作的教程，请参阅 教程：通过 AWS CLI开始使用计划扩缩。

有关为自动扩缩组创建计划操作的信息，请参阅《Amazon EC2 Auto Scaling 用户指南》中的 Amazon EC2 Auto Scaling 的计划扩缩。

限制

以下是使用计划的扩缩时的限制：

- 每个可扩展目标的计划操作的名称必须是唯一的。
- Application Auto Scaling 不在计划表达式中提供二级精度。使用 Cron 表达式的最高解析精度是一分钟。
- 可扩展目标不能是 Amazon MSK 集群。Amazon MSK 不支持计划的扩缩。
- 在可扩展资源上查看、添加、更新或移除计划操作的控制台访问权限取决于您使用的资源。有关更多信息，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务。

使用 cron 表达式安排重复发生的扩缩操作

> **Important**
>
> 如需 Amazon EC2 Auto Scaling 的 cron 表达式的帮助，请参阅 Amazon EC2 Auto Scaling 用户指南中的定期计划主题。使用 Amazon EC2 Auto Scaling，您可以使用传统的 cron 语法，而不是 Application Auto Scaling 使用的自定义 cron 语法。

您可以使用 cron 表达式创建定期运行的计划操作。

若要创建重复计划，请指定 cron 表达式和时区来描述何时重复执行该计划操作。支持的时区值为 Joda-Time支持的 IANA 时区的规范名（例如 `Etc/GMT+9` 或 `Pacific/Tahiti`）。您可以选择指定开始时间和/或结束时间的日期和时间。有关使用创建计划操作 AWS CLI 的命令示例，请参阅创建指定时区的重复计划操作。

受支持的 cron 表达式格式由用空格分隔的六个字段组成：[Minutes] [Hours] [Day_of_Month] [Month] [Day_of_Week] [Year]。例如，cron 表达式 `30 6 ? * MON *` 会配置一个将于每周一早上 6:30 重复执行的计划操作。星号用作通配符，以匹配字段的所有值。

有关 Application Auto Scaling 计划操作的 cron 语法的更多信息，请参阅亚马逊 EventBridge 用户指南中的 Cron 表达式参考。

当您创建重复性计划时，请谨慎选择开始时间和结束时间。记住以下内容：

- 如果您指定开始时间，则 Application Auto Scaling 将在此时间执行操作，然后根据指定的重复执行操作。
- 如果指定结束时间，则操作在此时间之后停止重复。Application Auto Scaling 不会一直跟踪以前的值，并在结束时间后恢复为以前的值。
- 使用或 AWS SDK 创建 AWS CLI 或更新计划操作时，必须以 UTC 格式设置开始时间和结束时间。

示例

为 Application Auto Scaling 可扩展目标创建重复性计划时，可您可以参考下表。以下示例展示了使用 Application Auto Scaling 创建或更新计划操作的正确语法。

分钟	小时	日期	月份	星期几	年	含义
0	10	*	*	?	*	每天上午的 10:00 (UTC) 运行
15	12	*	*	?	*	每天在下午 12:15 (UTC) 运行
0	18	?	*	MON-FRI	*	每星期一到星期五的下午 6:00 (UTC) 运行
0	8	1	*	?	*	每月第 1 天的上午 8:00 (UTC) 运行
0/15	*	*	*	?	*	每 15 分钟运行一次
0/10	*	?	*	MON-FRI	*	从星期一到星期五，每 10 分钟运行一次
0/5	8-17	?	*	MON-FRI	*	每星期一到星期五的上午 8:00 和下午 5:55 (UTC) 之间，每 5 分钟运行一次

例外

此外，您还可以使用包含七个字段的字符串值创建 cron 表达式。在这种情况下，您可以使用前三个字段来指定运行计划操作的时间，包括秒数。完整的 cron 表达式包含以下以空格分隔的字段：[Seconds] [Minutes] [Hours] [Day_of_Month] [Month] [Day_of_Week] [Year]。但是，这种方法不能保证计划操作会在您指定的准确秒数运行。此外，某些服务控制台可能不支持 cron 表达式中的秒数字段。

Application Auto Scaling 的计划操作示例

以下示例说明如何使用该 AWS CLI put-scheduled-action命令创建计划操作。当您指定新容量时，可指定最小容量和/或最大容量。

为简洁起见，本主题中的示例说明与 Application Auto Scaling 集成的一些服务的 CLI 命令。要指定不同的可扩展目标，请在 `--service-namespace` 中指定其命名空间，在 `--scalable-dimension` 中指定其可扩展维度，并在 `--resource-id` 中指定其资源 ID。有关每项服务的更多信息和示例，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务 中的主题。

使用时 AWS CLI，请记住您的命令在 AWS 区域 配置文件中运行。如果您想要在不同的区域中运行命令，可以为配置文件更改默认区域，或者与命令一起使用 `--region` 参数。

内容

创建仅发生一次的计划操作

要在指定的日期和时间仅弹性伸缩可扩展目标一次，请使用 `--schedule "at(`*`yyyy-mm-ddThh:mm:ss`*`)"` 选项。

Example 示例：仅向外扩展一次

以下是创建计划操作以在特定日期和时间横向扩展容量的示例。

在为 `--schedule` 指定的日期和时间（UTC 时间 2021 年 3 月 31 日晚上 10:00），如果为 `MinCapacity` 指定的值高于当前容量，则 Application Auto Scaling 将横向扩展到 `MinCapacity`。

Linux、macOS 或 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource \
  --scalable-dimension custom-resource:ResourceType:Property \
  --resource-id file://~/custom-resource-id.txt \
  --scheduled-action-name scale-out \
  --schedule "at(2021-03-31T22:00:00)" \
  --scalable-target-action MinCapacity=3
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource --
scalable-dimension custom-resource:ResourceType:Property --resource-id file://~/custom-
resource-id.txt --scheduled-action-name scale-out --schedule "at(2021-03-31T22:00:00)"
 --scalable-target-action MinCapacity=3
```

> **Note**
>
> 运行此计划操作时，如果最大容量小于为最小容量指定的值，则必须指定新的最小容量和最大容量，而不仅仅是最小容量。

Example 示例：仅向内扩展一次

以下是创建计划操作以在特定日期和时间横向缩减容量的示例。

在为 `--schedule` 指定的日期和时间（UTC 时间 2021 年 3 月 31 日晚上 10:30），如果为 `MaxCapacity` 指定的值低于当前容量，则 Application Auto Scaling 将横向缩减到 `MaxCapacity`。

Linux、macOS 或 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource \
  --scalable-dimension custom-resource:ResourceType:Property \
  --resource-id file://~/custom-resource-id.txt \
  --scheduled-action-name scale-in \
  --schedule "at(2021-03-31T22:30:00)" \
  --scalable-target-action MinCapacity=0,MaxCapacity=0
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace custom-resource --
scalable-dimension custom-resource:ResourceType:Property --resource-id file://~/custom-
```

```
resource-id.txt --scheduled-action-name scale-in --schedule "at(2021-03-31T22:30:00)"
 --scalable-target-action MinCapacity=0,MaxCapacity=0
```

创建按重复间隔运行的计划操作

要按重复间隔计划扩缩，请使用 `--schedule "rate(value unit)"` 选项。该值必须为正整数。单位可以是 `minute`、`minutes`、`hour`、`hours`、`day` 或 `days`。有关更多信息，请参阅 Amazon Ev CloudWatch ents 用户指南中的费率表达式。

以下是使用 rate 表达式的计划操作的示例。

根据指定的计划（从 UTC 时间 2021 年 1 月 30 日中午 12:00 PM 开始，到 UTC 时间 2021 年 1 月 31 日晚上 10:00 结束，每 5 个小时一次），如果为 `MinCapacity` 指定的值高于当前容量，则 Application Auto Scaling 横向扩展到 `MinCapacity`。如果为 `MaxCapacity` 指定的值低于当前容量，则 Application Auto Scaling 将横向缩减到 `MaxCapacity`。

Linux、macOS 或 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --scheduled-action-name my-recurring-action \
  --schedule "rate(5 hours)" \
  --start-time 2021-01-30T12:00:00 \
  --end-time 2021-01-31T22:00:00 \
  --scalable-target-action MinCapacity=3,MaxCapacity=10
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace ecs --scalable-
dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
 --scheduled-action-name my-recurring-action --schedule "rate(5 hours)" --start-
time 2021-01-30T12:00:00 --end-time 2021-01-31T22:00:00 --scalable-target-action
 MinCapacity=3,MaxCapacity=10
```

创建按重复计划运行的计划操作

要按重复计划来计划扩缩，请使用 `--schedule "cron(fields)"` 选项。有关更多信息，请参阅 使用 cron 表达式安排重复发生的扩缩操作。

以下是使用 cron 表达式的计划操作的示例。

根据指定的计划（UTC 时间每天上午 9:00），如果为 MinCapacity 指定的值高于当前容量，则 Application Auto Scaling 将横向扩展到 MinCapacity。如果为 MaxCapacity 指定的值低于当前容量，则 Application Auto Scaling 将横向缩减到 MaxCapacity。

Linux、macOS 或 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace appstream \
  --scalable-dimension appstream:fleet:DesiredCapacity \
  --resource-id fleet/sample-fleet \
  --scheduled-action-name my-recurring-action \
  --schedule "cron(0 9 * * ? *)" \
  --scalable-target-action MinCapacity=10,MaxCapacity=50
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace appstream --
scalable-dimension appstream:fleet:DesiredCapacity --resource-id fleet/sample-fleet --
scheduled-action-name my-recurring-action --schedule "cron(0 9 * * ? *)" --scalable-
target-action MinCapacity=10,MaxCapacity=50
```

创建指定时区的一次性计划操作

默认情况下，计划操作设置为 UTC 时区。要指定不同的时区，请包含 --timezone 选项并指定时区的规范名称（例如，America/New_York）。有关更多信息，请参阅 https://www.joda.org/joda-time/timezones.html，其中提供了有关致电put-scheduled-action时支持的 IANA 时区的信息。

以下是创建计划操作以在特定日期和时间扩展容量时使用 --timezone 选项的示例。

在为 --schedule 指定的日期和时间（当地时间 2021 年 1 月 31 日下午 5:00），如果为 MinCapacity 指定的值高于当前容量，则 Application Auto Scaling 将横向扩展到 MinCapacity。如果为 MaxCapacity 指定的值低于当前容量，则 Application Auto Scaling 将横向缩减到 MaxCapacity。

Linux、macOS 或 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace comprehend \
  --scalable-dimension comprehend:document-classifier-endpoint:DesiredInferenceUnits \
  --resource-id arn:aws:comprehend:us-west-2:123456789012:document-classifier-endpoint/
EXAMPLE \
  --scheduled-action-name  my-one-time-action \
  --schedule "at(2021-01-31T17:00:00)" --timezone "America/New_York" \
```

```
  --scalable-target-action MinCapacity=1,MaxCapacity=3
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace comprehend --
scalable-dimension comprehend:document-classifier-endpoint:DesiredInferenceUnits
 --resource-id arn:aws:comprehend:us-west-2:123456789012:document-classifier-
endpoint/EXAMPLE --scheduled-action-name  my-one-time-action --schedule
 "at(2021-01-31T17:00:00)" --timezone "America/New_York" --scalable-target-action
 MinCapacity=1,MaxCapacity=3
```

创建指定时区的重复计划操作

以下是创建重复计划操作以扩展容量时使用 `--timezone` 选项的示例。有关更多信息，请参阅 使用 cron 表达式安排重复发生的扩缩操作。

根据指定的计划（当地时间每个星期一到星期五晚上 6:00），如果为 `MinCapacity` 指定的值高于当前容量，则 Application Auto Scaling 将横向扩展到 `MinCapacity`。如果为 `MaxCapacity` 指定的值低于当前容量，则 Application Auto Scaling 将横向缩减到 `MaxCapacity`。

Linux、macOS 或 Unix

```
aws application-autoscaling put-scheduled-action --service-namespace lambda \
  --scalable-dimension lambda:function:ProvisionedConcurrency \
  --resource-id function:my-function:BLUE \
  --scheduled-action-name my-recurring-action \
  --schedule "cron(0 18 ? * MON-FRI *)" --timezone "Etc/GMT+9" \
  --scalable-target-action MinCapacity=10,MaxCapacity=50
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace lambda
 --scalable-dimension lambda:function:ProvisionedConcurrency --resource-
id function:my-function:BLUE --scheduled-action-name my-recurring-action --schedule
 "cron(0 18 ? * MON-FRI *)" --timezone "Etc/GMT+9" --scalable-target-action
 MinCapacity=10,MaxCapacity=50
```

管理 Application Auto Scaling 的计划扩缩

AWS CLI 包括其他几个可帮助您管理计划操作的命令。

为简洁起见，本主题中的示例说明与 Application Auto Scaling 集成的一些服务的 CLI 命令。要指定不同的可扩展目标，请在 `--service-namespace` 中指定其命名空间，在 `--scalable-dimension` 中指定其可扩展维度，并在 `--resource-id` 中指定其资源 ID。有关每项服务的更多信息和示例，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务 中的主题。

使用时 AWS CLI，请记住您的命令在 AWS 区域 配置文件中运行。如果您想要在不同的区域中运行命令，可以为配置文件更改默认区域，或者与命令一起使用 `--region` 参数。

内容

查看指定服务的扩缩活动

要查看指定服务命名空间中所有可扩展目标的扩展活动，请使用describe-scaling-activities命令。

以下示例检索与 dynamodb 服务命名空间关联的扩缩活动。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scaling-activities --service-namespace dynamodb
```

Windows

```
aws application-autoscaling describe-scaling-activities --service-namespace dynamodb
```

如果命令成功，则将显示类似于以下内容的输出。

```
{
    "ScalingActivities": [
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting write capacity units to 10.",
            "ResourceId": "table/my-table",
            "ActivityId": "4d1308c0-bbcf-4514-a673-b0220ae38547",
            "StartTime": 1561574415.086,
```

```
            "ServiceNamespace": "dynamodb",
            "EndTime": 1561574449.51,
            "Cause": "maximum capacity was set to 10",
            "StatusMessage": "Successfully set write capacity units to 10. Change
successfully fulfilled by dynamodb.",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting min capacity to 5 and max capacity to 10",
            "ResourceId": "table/my-table",
            "ActivityId": "f2b7847b-721d-4e01-8ef0-0c8d3bacc1c7",
            "StartTime": 1561574414.644,
            "ServiceNamespace": "dynamodb",
            "Cause": "scheduled action name my-second-scheduled-action was triggered",
            "StatusMessage": "Successfully set min capacity to 5 and max capacity to
10",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting write capacity units to 15.",
            "ResourceId": "table/my-table",
            "ActivityId": "d8ea4de6-9eaa-499f-b466-2cc5e681ba8b",
            "StartTime": 1561574108.904,
            "ServiceNamespace": "dynamodb",
            "EndTime": 1561574140.255,
            "Cause": "minimum capacity was set to 15",
            "StatusMessage": "Successfully set write capacity units to 15. Change
successfully fulfilled by dynamodb.",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting min capacity to 15 and max capacity to 20",
            "ResourceId": "table/my-table",
            "ActivityId": "3250fd06-6940-4e8e-bb1f-d494db7554d2",
            "StartTime": 1561574108.512,
            "ServiceNamespace": "dynamodb",
            "Cause": "scheduled action name my-first-scheduled-action was triggered",
            "StatusMessage": "Successfully set min capacity to 15 and max capacity to
20",
            "StatusCode": "Successful"
        }
```

```
    ]
}
```

要更改此命令以使其仅检索其中一个可扩展目标的扩缩活动，请添加 `--resource-id` 选项。

描述指定服务的所有计划操作

要描述指定服务命名空间中所有可扩展目标的计划操作，请使用describe-scheduled-actions命令。

以下示例检索与 `ec2` 服务命名空间关联的计划操作。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scheduled-actions --service-namespace ec2
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-namespace ec2
```

如果成功，该命令返回类似以下内容的输出。

```
{
    "ScheduledActions": [
        {
            "ScheduledActionName": "my-one-time-action",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:493a6261-fbb9-432d-855d-3c302c14bdb9:resource/ec2/
spot-fleet-request/sfr-107dc873-0802-4402-a901-37294EXAMPLE:scheduledActionName/my-one-
time-action",
            "ServiceNamespace": "ec2",
            "Schedule": "at(2021-01-31T17:00:00)",
            "Timezone": "America/New_York",
            "ResourceId": "spot-fleet-request/sfr-107dc873-0802-4402-
a901-37294EXAMPLE",
            "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
            "ScalableTargetAction": {
                "MaxCapacity": 1
            },
            "CreationTime": 1607454792.331
        },
        {
```

```
            "ScheduledActionName": "my-recurring-action",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:493a6261-fbb9-432d-855d-3c302c14bdb9:resource/ec2/
spot-fleet-request/sfr-107dc873-0802-4402-a901-37294EXAMPLE:scheduledActionName/my-
recurring-action",
            "ServiceNamespace": "ec2",
            "Schedule": "rate(5 minutes)",
            "ResourceId": "spot-fleet-request/sfr-107dc873-0802-4402-
a901-37294EXAMPLE",
            "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
            "StartTime": 1604059200.0,
            "EndTime": 1612130400.0,
            "ScalableTargetAction": {
                "MinCapacity": 3,
                "MaxCapacity": 10
            },
            "CreationTime": 1607454949.719
        },
        {
            "ScheduledActionName": "my-one-time-action",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:4bce34c7-bb81-4ecf-b776-5c726efb1567:resource/ec2/
spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE:scheduledActionName/my-one-
time-action",
            "ServiceNamespace": "ec2",
            "Schedule": "at(2020-12-08T9:36:00)",
            "Timezone": "America/New_York",
            "ResourceId": "spot-fleet-request/sfr-40edeb7b-9ae7-44be-
bef2-5c4c8EXAMPLE",
            "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
            "ScalableTargetAction": {
                "MinCapacity": 1,
                "MaxCapacity": 3
            },
            "CreationTime": 1607456031.391
        }
    ]
}
```

描述可扩展目标的一个或多个计划操作

要检索有关指定可扩展目标的计划操作的信息，请在使用describe-scheduled-actions命令描述计划操作时添加该--resource-id选项。

如果您包含 --scheduled-action-names 选项并将计划操作的名称指定为其值，则该命令仅返回名称匹配的计划操作，如以下示例所示。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scheduled-actions --service-namespace ec2 \
  --resource-id spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE \
  --scheduled-action-names my-one-time-action
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-namespace ec2 --
resource-id spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE --scheduled-
action-names my-one-time-action
```

下面是示例输出。

```
{
    "ScheduledActions": [
        {
            "ScheduledActionName": "my-one-time-action",
            "ScheduledActionARN": "arn:aws:autoscaling:us-
west-2:123456789012:scheduledAction:4bce34c7-bb81-4ecf-b776-5c726efb1567:resource/ec2/
spot-fleet-request/sfr-40edeb7b-9ae7-44be-bef2-5c4c8EXAMPLE:scheduledActionName/my-one-
time-action",
            "ServiceNamespace": "ec2",
            "Schedule": "at(2020-12-08T9:36:00)",
            "Timezone": "America/New_York",
            "ResourceId": "spot-fleet-request/sfr-40edeb7b-9ae7-44be-
bef2-5c4c8EXAMPLE",
            "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
            "ScalableTargetAction": {
                "MinCapacity": 1,
                "MaxCapacity": 3
            },
            "CreationTime": 1607456031.391
        }
    ]
}
```

如果为 --scheduled-action-names 选项提供多个值，则名称匹配的所有计划操作都包含在输出中。

关闭可扩展目标的计划扩缩

您可以暂时关闭计划扩缩而不删除您的计划操作。有关更多信息，请参阅 暂停和恢复 Application Auto Scaling 扩缩。

使用带--suspended-state选项的register-scalable-target命令并指定true为ScheduledScalingSuspended属性的值，在可扩展目标上暂停定时缩放，如以下示例所示。

Linux、macOS 或 Unix

```
aws application-autoscaling register-scalable-target --service-namespace rds \
  --scalable-dimension rds:cluster:ReadReplicaCount --resource-id cluster:my-db-cluster
 \
  --suspended-state '{"ScheduledScalingSuspended": true}'
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace rds --
scalable-dimension rds:cluster:ReadReplicaCount --resource-id cluster:my-db-cluster --
suspended-state "{\"ScheduledScalingSuspended\": true}"
```

如果成功，该命令会返回可扩展目标的 ARN。

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

要恢复计划扩缩，请再次运行此命令，指定 false 作为 ScheduledScalingSuspended 属性的值。

删除计划的操作

完成计划操作后，可以使用delete-scheduled-action命令将其删除。

Linux、macOS 或 Unix

```
aws application-autoscaling delete-scheduled-action --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
```

```
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-37294EXAMPLE \
  --scheduled-action-name my-recurring-action
```

Windows

```
aws application-autoscaling delete-scheduled-action --service-namespace ec2 --scalable-
dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-request/
sfr-73fbd2ce-aa30-494c-8788-37294EXAMPLE --scheduled-action-name my-recurring-action
```

如果成功，该命令将返回到提示符。

教程：通过 AWS CLI开始使用计划扩缩

以下教程向您展示了如何通过帮助您创建计划操作 AWS CLI 来扩展名为的 DynamoDB 表的示例，从而开始计划扩展。`TestTable`如果您的 DynamoDB 中没有用于测试的 `TestTable` 表，则可以通过运行 Amazon DynamoDB 开发人员指南中的步骤 1：创建 DynamoDB 表中所示的 create-table 命令来立即创建 一个。

使用时 AWS CLI，请记住您的命令在为您的个人资料配置的 AWS 区域中运行。如果您想要在不同的区域中运行命令，可以为配置文件更改默认区域，或者与命令一起使用 `--region` 参数。

> **Note**
>
> 作为本教程的一部分，您可能会产生 AWS 费用。请监控免费套餐使用情况，并确保您了解与 DynamoDB 数据库使用的读取和写入容量单位数关联的成本。

内容

步骤 1：注册您的可扩展目标

首先使用 Application Auto Scaling 将您的 DynamoDB 表注册为可扩展目标。

向 Application Auto Scaling 注册您的可扩展目标

1. 首先，使用describe-scalable-targets命令检查是否已注册任何 DynamoDB 资源。这可以让您验证 TestTable 表是否已取消注册，以防它不是新表。

   Linux、macOS 或 Unix

   ```
   aws application-autoscaling describe-scalable-targets \
     --service-namespace dynamodb
   ```

   Windows

   ```
   aws application-autoscaling describe-scalable-targets --service-namespace dynamodb
   ```

   如果没有现有的可扩展目标，则这是系统的响应。

   ```
   {
       "ScalableTargets": []
   }
   ```

2. 使用以下register-scalable-target命令注册名为的 DynamoDB 表的写入容量。TestTable将最小所需容量设置为 5 个写入容量单位，将最大所需容量设置为 10 个写入容量单位。

   Linux、macOS 或 Unix

   ```
   aws application-autoscaling register-scalable-target \
     --service-namespace dynamodb \
     --scalable-dimension dynamodb:table:WriteCapacityUnits \
     --resource-id table/TestTable \
     --min-capacity 5 --max-capacity 10
   ```

   Windows

   ```
   aws application-autoscaling register-scalable-target --service-namespace dynamodb
    --scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/
   TestTable --min-capacity 5 --max-capacity 10
   ```

   如果成功，该命令会返回可扩展目标的 ARN。

   ```
   {
   ```

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

步骤 2：创建两个计划操作

Application Auto Scaling 可让您安排扩缩操作应发生的时间。您可以指定可扩展目标、扩展计划、最小容量和最大容量。在指定的时间，Application Auto Scaling 会更新可扩展目标的最小值和最大值。如果当前容量超出此范围，这会导致一个扩展活动。

如果您决定创建扩展策略，计划更新最小和最大容量也会有所帮助。扩展策略允许基于当前资源利用率来动态扩展您的资源。扩展策略的一种常见保护措施是设置适当的最小和最大容量值。

在本练习中，我们将创建两个一次性操作来分别进行扩展和缩减。

创建和查看计划操作

1. 要创建第一个计划操作，请使用以下put-scheduled-action命令。

   `--schedule` 中的 at 命令计划在将来的指定日期和时间要运行一次的操作。小时采用世界标准时间 24 小时格式。将操作安排在当前时间开始约 5 分钟后发生。

   在指定的日期和时间，Application Auto Scaling 将更新 `MinCapacity` 和 `MaxCapacity` 的值。假设表当前有 5 个写入容量单位，Application Auto Scaling 横向扩展到 `MinCapacity`，以使表拥有 15-20 个写入容量单位的新所需范围。

   Linux、macOS 或 Unix

   ```
   aws application-autoscaling put-scheduled-action \
     --service-namespace dynamodb \
     --scalable-dimension dynamodb:table:WriteCapacityUnits \
     --resource-id table/TestTable \
     --scheduled-action-name my-first-scheduled-action \
     --schedule "at(2019-05-20T17:05:00)" \
     --scalable-target-action MinCapacity=15,MaxCapacity=20
   ```

   Windows

   ```
   aws application-autoscaling put-scheduled-action --service-namespace dynamodb
    --scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/
   ```

```
TestTable --scheduled-action-name my-first-scheduled-action --schedule
 "at(2019-05-20T17:05:00)" --scalable-target-action MinCapacity=15,MaxCapacity=20
```

如果此命令成功执行，将不会返回任何输出。

2. 要创建 Application Auto Scaling 用来缩减的第二个计划操作，请使用以下put-scheduled-action命令。

   将操作安排在当前时间开始约 10 分钟后发生。

   在指定的日期和时间，Application Auto Scaling 将更新表的 `MinCapacity` 和 `MaxCapacity`，并横向缩减到 `MaxCapacity` 以将表恢复为 5-10 个写入容量单位的初始所需范围。

   Linux、macOS 或 Unix

```
aws application-autoscaling put-scheduled-action \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:WriteCapacityUnits \
  --resource-id table/TestTable \
  --scheduled-action-name my-second-scheduled-action \
  --schedule "at(2019-05-20T17:10:00)" \
  --scalable-target-action MinCapacity=5,MaxCapacity=10
```

   Windows

```
aws application-autoscaling put-scheduled-action --service-namespace dynamodb
 --scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/
TestTable --scheduled-action-name my-second-scheduled-action --schedule
 "at(2019-05-20T17:10:00)" --scalable-target-action MinCapacity=5,MaxCapacity=10
```

3. （可选）使用以下describe-scheduled-actions命令获取指定服务命名空间的计划操作列表。

   Linux、macOS 或 Unix

```
aws application-autoscaling describe-scheduled-actions \
  --service-namespace dynamodb
```

   Windows

```
aws application-autoscaling describe-scheduled-actions --service-namespace dynamodb
```

下面是示例输出。

```
{
    "ScheduledActions": [
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Schedule": "at(2019-05-20T18:35:00)",
            "ResourceId": "table/TestTable",
            "CreationTime": 1561571888.361,
            "ScheduledActionARN": "arn:aws:autoscaling:us-
east-1:123456789012:scheduledAction:2d36aa3b-cdf9-4565-b290-81db519b227d:resource/
dynamodb/table/TestTable:scheduledActionName/my-first-scheduled-action",
            "ScalableTargetAction": {
                "MinCapacity": 15,
                "MaxCapacity": 20
            },
            "ScheduledActionName": "my-first-scheduled-action",
            "ServiceNamespace": "dynamodb"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Schedule": "at(2019-05-20T18:40:00)",
            "ResourceId": "table/TestTable",
            "CreationTime": 1561571946.021,
            "ScheduledActionARN": "arn:aws:autoscaling:us-
east-1:123456789012:scheduledAction:2d36aa3b-cdf9-4565-b290-81db519b227d:resource/
dynamodb/table/TestTable:scheduledActionName/my-second-scheduled-action",
            "ScalableTargetAction": {
                "MinCapacity": 5,
                "MaxCapacity": 10
            },
            "ScheduledActionName": "my-second-scheduled-action",
            "ServiceNamespace": "dynamodb"
        }
    ]
}
```

步骤 3：查看扩缩活动

在此步骤中，您将查看计划操作触发的扩缩活动，并验证 DynamoDB 是否已更改表的写入容量。

查看扩展活动

1. 等待您选择的时间，然后使用以下describe-scaling-activities命令验证您的计划操作是否正常运行。

   Linux、macOS 或 Unix

```
aws application-autoscaling describe-scaling-activities \
  --service-namespace dynamodb
```

   Windows

```
aws application-autoscaling describe-scaling-activities --service-
namespace dynamodb
```

   以下是第一个计划操作在计划操作正在执行时的示例输出。

   扩展活动按创建日期排序，首先返回最新的扩展活动。

```
{
    "ScalingActivities": [
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting write capacity units to 15.",
            "ResourceId": "table/TestTable",
            "ActivityId": "d8ea4de6-9eaa-499f-b466-2cc5e681ba8b",
            "StartTime": 1561574108.904,
            "ServiceNamespace": "dynamodb",
            "Cause": "minimum capacity was set to 15",
            "StatusMessage": "Successfully set write capacity units to 15. Waiting
 for change to be fulfilled by dynamodb.",
            "StatusCode": "InProgress"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting min capacity to 15 and max capacity to 20",
            "ResourceId": "table/TestTable",
            "ActivityId": "3250fd06-6940-4e8e-bb1f-d494db7554d2",
            "StartTime": 1561574108.512,
            "ServiceNamespace": "dynamodb",
            "Cause": "scheduled action name my-first-scheduled-action was
 triggered",
```

```
            "StatusMessage": "Successfully set min capacity to 15 and max capacity
 to 20",
            "StatusCode": "Successful"
        }
    ]
}
```

以下是两个计划操作都运行完成后的示例输出。

```
{
    "ScalingActivities": [
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting write capacity units to 10.",
            "ResourceId": "table/TestTable",
            "ActivityId": "4d1308c0-bbcf-4514-a673-b0220ae38547",
            "StartTime": 1561574415.086,
            "ServiceNamespace": "dynamodb",
            "EndTime": 1561574449.51,
            "Cause": "maximum capacity was set to 10",
            "StatusMessage": "Successfully set write capacity units to 10. Change
 successfully fulfilled by dynamodb.",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting min capacity to 5 and max capacity to 10",
            "ResourceId": "table/TestTable",
            "ActivityId": "f2b7847b-721d-4e01-8ef0-0c8d3bacc1c7",
            "StartTime": 1561574414.644,
            "ServiceNamespace": "dynamodb",
            "Cause": "scheduled action name my-second-scheduled-action was
 triggered",
            "StatusMessage": "Successfully set min capacity to 5 and max capacity
 to 10",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting write capacity units to 15.",
            "ResourceId": "table/TestTable",
            "ActivityId": "d8ea4de6-9eaa-499f-b466-2cc5e681ba8b",
            "StartTime": 1561574108.904,
```

```
            "ServiceNamespace": "dynamodb",
            "EndTime": 1561574140.255,
            "Cause": "minimum capacity was set to 15",
            "StatusMessage": "Successfully set write capacity units to 15. Change
 successfully fulfilled by dynamodb.",
            "StatusCode": "Successful"
        },
        {
            "ScalableDimension": "dynamodb:table:WriteCapacityUnits",
            "Description": "Setting min capacity to 15 and max capacity to 20",
            "ResourceId": "table/TestTable",
            "ActivityId": "3250fd06-6940-4e8e-bb1f-d494db7554d2",
            "StartTime": 1561574108.512,
            "ServiceNamespace": "dynamodb",
            "Cause": "scheduled action name my-first-scheduled-action was
 triggered",
            "StatusMessage": "Successfully set min capacity to 15 and max capacity
 to 20",
            "StatusCode": "Successful"
        }
    ]
}
```

2. 成功运行计划操作后，请打开 DynamoDB 控制台并选择要处理的表。查看 Capacity（容量）选项卡下的 Write capacity units（写入容量单位）。在第二个扩展操作运行后，写入容量单位应已从 15 变为 10。

   此外，您还可以使用以下 describe-table 命令验证表的当前写入容量。包含 `--query` 选项以筛选输出。有关输出筛选功能的更多信息 AWS CLI，请参阅《AWS Command Line Interface 用户指南》 AWS CLI中的控制命令输出。

   Linux、macOS 或 Unix

   ```
   aws dynamodb describe-table --table-name TestTable \
     --query 'Table.[TableName,TableStatus,ProvisionedThroughput]'
   ```

   Windows

   ```
   aws dynamodb describe-table --table-name TestTable --query "Table.
   [TableName,TableStatus,ProvisionedThroughput]"
   ```

   下面是示例输出。

```
[
    "TestTable",
    "ACTIVE",
    {
        "NumberOfDecreasesToday": 1,
        "WriteCapacityUnits": 10,
        "LastIncreaseDateTime": 1561574133.264,
        "ReadCapacityUnits": 5,
        "LastDecreaseDateTime": 1561574435.607
    }
]
```

步骤 4：后续步骤

如果您想尝试同时使用计划扩展和扩展策略进行扩展，请按照中的步骤操作教程：配置自动扩缩以处理繁重的工作负载。

第 5 步：清理

当您完成入门练习后，可以按照如下步骤清除关联的资源。

删除计划的操作

以下delete-scheduled-action命令删除指定的计划操作。如果您要将此计划操作保留供将来使用，您可以跳过此操作。

Linux、macOS 或 Unix

```
aws application-autoscaling delete-scheduled-action \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:WriteCapacityUnits \
  --resource-id table/TestTable \
  --scheduled-action-name my-second-scheduled-action
```

Windows

```
aws application-autoscaling delete-scheduled-action --service-namespace dynamodb --
scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/TestTable --
scheduled-action-name my-second-scheduled-action
```

撤消可扩展目标的注册

使用以下deregister-scalable-target命令取消注册可扩展目标。如果您有任何您创建的扩展策略或尚未删除的计划操作，这条命令会将它们删除。如果您要将此可扩展目标保留供将来使用，您可以跳过此操作。

Linux、macOS 或 Unix

```
aws application-autoscaling deregister-scalable-target \
  --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:WriteCapacityUnits \
  --resource-id table/TestTable
```

Windows

```
aws application-autoscaling deregister-scalable-target --service-namespace dynamodb --
scalable-dimension dynamodb:table:WriteCapacityUnits --resource-id table/TestTable
```

删除 DynamoDB 表

使用以下 delete-table 命令可删除本教程中使用的表。如果您要保留该表供将来使用，可以跳过此步骤。

Linux、macOS 或 Unix

```
aws dynamodb delete-table --table-name TestTable
```

Windows

```
aws dynamodb delete-table --table-name TestTable
```

目标跟踪扩缩策略

目标跟踪扩缩策略根据目标指标值扩缩您的应用程序。这使您的应用程序无需人工干预即可保持最佳性能和成本效益。

通过目标跟踪，您可以选择一个指标和一个目标值，目标值用来表示应用程序的理想平均利用率或吞吐量水平。Application Auto Scaling 创建并管理在指标偏离目标时触发扩展事件的 CloudWatch警报。这与恒温器保持目标温度的方式类似。

例如，假设您当前有一个在竞价型实例集上运行的应用程序，并希望在应用程序负载变化时将该实例集的 CPU 利用率保持在 50% 左右。这为您提供额外容量以处理流量高峰，而无需维护过多的空闲资源。

创建一个将目标平均 CPU 利用率设置为 50% 的目标跟踪扩缩策略即可满足此需求。然后，当 CPU 使用率超过 50% 时，Application Auto Scaling 将横向扩展（增加容量），以处理增加的负载。当 CPU 利用率降至 50% 以下时，Application Auto Scaling 将横向缩减（减少容量），以便在利用率低的时期优化成本。

目标跟踪策略无需手动定义 CloudWatch 警报和缩放调整。Application Auto Scaling 会根据您设定的目标自动处理这个问题。

您可以使用预定义的指标或自定义指标，设定目标跟踪策略：

- 预定义指标：Application Auto Scaling 提供的指标，例如平均 CPU 利用率或每个目标的平均请求数。
- 自定义指标-您可以使用指标数学来组合指标、利用现有指标或使用自己发布到 CloudWatch的自定义指标。

选择一个与可扩展目标容量的变化成反比的指标。因此，如果将容量翻一番，则该指标将降低 50%。这使指标数据能够准确触发按比例扩缩事件。

主题

- 目标跟踪缩放的工作原理
- 使用创建目标跟踪扩展策略 AWS CLI
- 使用指标数学为 Application Auto Scaling 创建目标跟踪扩展策略

目标跟踪缩放的工作原理

本主题描述了目标跟踪扩展的工作原理，并介绍了目标跟踪扩展策略的关键要素。

内容

工作方式

要使用目标跟踪缩放，请创建目标跟踪扩展策略并指定以下内容：

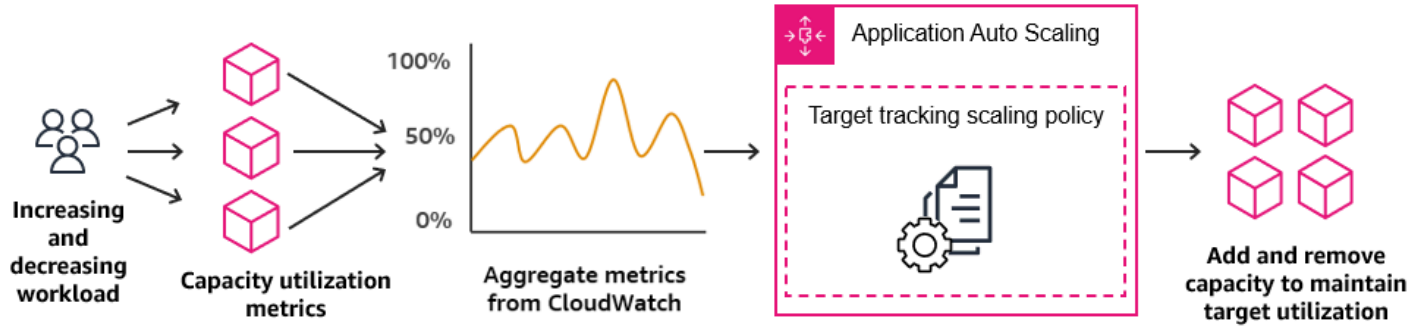
- 指标-要跟踪的 CloudWatch 指标，例如平均 CPU 利用率或每个目标的平均请求数。
- 目标值：指标的目标值，例如 50% 的 CPU 利用率或每个目标每分钟 1000 个请求。

Application Auto Scaling 创建和管理调用扩展策略的 CloudWatch 警报，并根据指标和目标值计算扩展调整。扩缩策略将根据需要增加或减少容量，将指标保持在指定的目标值或接近指定的目标值。

当指标高于目标值时，Application Auto Scaling 会通过增加容量来缩小指标值和目标值之间的差异，从而横向扩展。当指标低于目标值时，Application Auto Scaling 会通过减少容量来横向缩减。

扩缩活动在两者之间有冷却时间，以防止容量快速波动。您可以选择为扩缩策略配置冷却时间。

下图概述显示设置完成时目标跟踪扩缩策略的工作原理。



请注意，目标跟踪扩缩策略在利用率提高时添加容量比在利用率降低时删除容量更为积极。例如，如果策略的指定指标达到其目标值，则策略假定您的应用程序已达到高负载。因此，它通过尽可能快地添加与指标值成比例的容量来进行响应。指标越高，添加的容量就越多。

当指标低于目标值时，策略预计利用率最终会再次增加。在此场景中，只有当利用率超过远低于目标值（通常比目标值低 10% 以上）的阈值时，它才会通过删除容量来减慢扩缩速度，从而认为利用率已放缓。这种更保守的行为旨在确保只有当应用程序不再遇到与之前相同的高级别需求时，才会删除容量。

选择指标

您可以使用预定义的指标或自定义指标，创建目标跟踪扩展策略。

使用预定义指标类型创建目标跟踪扩展策略时，您可以从 目标跟踪扩展策略的预定义目标 中的预定义指标列表中选择一个指标。

选择指标时请记住原则：

- 并非所有自定义指标都适用于目标跟踪。指标必须是有效的使用率指标并且描述可扩展目标的繁忙程度。指标值必须根据可扩展目标的容量按比例增加或减少，以便指标数据可用于按比例扩展可扩展目标。
- 要使用 `ALBRequestCountPerTarget` 指标，您必须指定 `ResourceLabel` 参数以标识与该指标关联的目标组。
- 当某个指标将实数 0 值发送到 CloudWatch（例如`ALBRequestCountPerTarget`）时，当您的应用程序持续一段时间内没有流量时，Application Auto Scaling 可以缩减到 0。要在没有请求路由时将可扩展目标横向缩减到 0，可扩展目标的最小容量必须设置为 0。
- 您可以使用指标数学组合现有指标，而不必发布要在扩缩策略中使用的新指标。有关更多信息，请参阅 使用指标数学为 Application Auto Scaling 创建目标跟踪扩展策略。
- 要查看您使用的服务是否支持在服务控制台中指定自定义指标，请参阅该服务的文档。
- 我们建议您使用每隔一分钟可用的指标，以帮助您更快地扩展以响应利用率变化。目标跟踪将评估所有预定义指标和自定义指标的以一分钟为粒度聚合的指标，但底层指标发布数据的频率可能会降低。

例如，默认情况下，所有 Amazon EC2 指标都以五分钟为间隔发送，但可配置为每隔一分钟（即详细监控）发送。是否选择此配置取决于单个服务。大部分情况下可尝试使用尽可能小的间隔。

定义目标值

创建目标跟踪扩缩策略时，必须指定一个目标值。目标值表示应用程序的最佳平均利用率或吞吐量。为了经济高效地使用资源，目标值的设置应尽可能高，并为流量的意外增加提供合的理缓冲。当应用程序针对正常流量进行最佳横向扩展时，实际指标值应等于或略低于目标值。

当扩缩策略基于吞吐量（例如，每目标的应用程序负载均衡器请求计数、网络 I/O 或其他计数指标）时，目标值表示单个实体（例如 Application Load Balancer 目标组的单个目标）在一分钟内的最佳平均吞吐量。

定义冷却时间

您可以选择在目标跟踪扩展策略中定义冷却时间。

冷却时间指定了扩展策略等待上一个扩展活动生效的时间量。

冷却时间有两种类型：

- 使用 scale-out cooldown period (向外扩展冷却时间)，目的是持续（但不过度）向外扩展。Application Auto Scaling 使用扩展策略成功横向扩展后，它将开始计算冷却时间。除非触发更大的横向扩展或冷却时间结束，否则扩展策略不会再次增加所需容量。尽管此向外扩展冷却时间有效，但启动向外扩展活动所添加的容量将计算为下一个向外扩展活动所需容量的一部分。
- 使用横向缩减冷却时间，目的是以保守方式进行横向缩减，以保护应用程序的可用性，因此在冷却时间过期之前，横向缩减活动会被阻止。但是，如果另一个警报在缩减冷却时间内触发了向外扩展活动，Application Auto Scaling 将立即向外扩展目标。在这种情况下，横向缩减冷却时间会停止而不完成。

每个冷却时间以秒为单位进行度量，仅适用于与扩展策略相关的扩展活动。在冷却时间内，当计划的操作在计划的时间开始时，它可以立即触发扩展活动，而无需等待冷却时间到期。

您可以从默认值开始，稍后可对其进行微调。例如，您可能需要延长冷却时间，以防止目标跟踪扩展策略对短时间内发生的更改过于激进。

默认值

Application Auto Scaling 为 ElastiCache 复制组提供默认值 600，为以下可扩展目标提供默认值 300：

- AppStream 2.0 支舰队
- Aurora 数据库集群
- ECS 服务
- Neptune 集群
- SageMaker 端点变体
- SageMaker 推理组件
- SageMaker 无服务器配置的并发性
- Spot Fleets
- 自定义资源

对于所有其他可扩展目标，默认值为 0 或 null：

- Amazon Comprehend 文档分类和实体识别程序终端节点
- DynamoDB 表和全局二级索引
- Amazon Keyspaces 表
- Lambda 预配置并发
- Amazon MSK 代理存储

Application Auto Scaling 评估冷却时间时，会将 null 值视为零值。

您可以更新任何默认值（包括 null 值），以设置自己的冷却时间。

注意事项

使用目标跟踪扩缩策略时，需要注意以下事项：

- 请勿创建、编辑或删除与目标跟踪扩展策略一起使用的 CloudWatch 警报。Application Auto Scaling 创建和管理与目标跟踪扩展策略关联的 CloudWatch 警报，并在不再需要时将其删除。
- 如果指标缺少数据点，则会导致 CloudWatch 警报状态更改为`INSUFFICIENT_DATA`。发生这种情况时，在找到新的数据点之前，Application Auto Scaling 无法扩展您的可扩展目标。有关在数据不足时创建警报的信息，请参阅 使用 CloudWatch 警报进行监控。
- 如果设计为很少报告指标，则指标数学可能会很有帮助。例如，要使用最新的值，则使用`FILL(m1,REPEAT)` 函数，其中 `m1` 是指标。

- 您可能会看到目标值与实际指标数据点之间存在差距。这是因为 Application Auto Scaling 在确定要添加或删除多少容量时将始终通过向上或向下舍入保守地进行操作，以免添加的容量不足或删除的容量过多。但是，对于具有小容量的可扩展目标，实际指标数据点可能看起来与目标值差距很大。

  对于容量更高的可扩展目标，添加或删除容量将缩小目标值与实际指标数据点之间的差距。
- 目标跟踪扩展策略假设它应该在指定指标高于目标值时执行向外扩展。因此，不能使用目标跟踪扩展策略在指定指标低于目标值时向外扩展。

多个扩缩策略

一个可扩展目标可以具有多个目标跟踪扩展策略，前提是它们分别使用不同的指标。Application Auto Scaling 的目的是始终优先考虑可用性，因此其行为会有所不同，具体取决于目标跟踪策略是否已准备好横向扩展或横向缩减。如果任何目标跟踪策略已准备好进行向外扩展，它将向外扩展可扩展目标，但仅在所有目标跟踪策略（启用了缩减部分）准备好缩减时才执行缩减。

如果多个扩缩策略指示可扩展目标同时横向扩展或横向缩减，则 Application Auto Scaling 会根据为横向缩减和横向扩展提供最大容量的策略进行扩展。这让您能够更灵活地覆盖多种场景，并确保始终有足够的容量来处理工作负载。

您可以禁用目标跟踪扩展策略的横向缩减部分，以便使用与横向扩展不同的方法进行横向缩减。例如，您可以使用步进扩展策略进行缩减，同时使用目标跟踪扩展策略进行横向扩展。

不过，在将目标跟踪扩展策略与步进扩展策略结合使用时，我们建议您务必谨慎，因为这些策略之间的冲突可能会导致意外的行为。例如，如果步进扩展策略在目标跟踪策略准备执行缩减之前启动缩减活动，则不会阻止缩减活动。在缩减活动完成后，目标跟踪策略可能会指示可扩展目标重新横向扩展。

对于具有周期性质的工作负载，您还可以选择使用计划扩展按计划自动更改容量。对于每个计划的操作，可以定义新的最小容量值和新的最大容量值。这些值构成扩展策略的边界。当立即需要容量时，计划扩展和目标跟踪扩展的组合有助于减少利用率级别急剧增加的影响。

扩缩策略创建、管理和删除的常用命令

使用扩缩策略的常用命令包括：

- register-scalable-target注册 AWS 或自定义资源作为可扩展目标（Application Auto Scaling 可以扩展的资源），以及暂停和恢复扩展。
- put-scaling-policy为现有可扩展目标添加或修改扩展策略。
- describe-scaling-activities返回有关某个 AWS 区域中扩展活动的信息。

- describe-scaling-policies返回有关某个 AWS 区域中扩展策略的信息。
- delete-scaling-policy删除扩展策略。

相关资源

有关为自动扩缩组创建目标跟踪扩缩策略的信息，请参阅《Amazon EC2 Auto Scaling 用户指南》中的 Amazon EC2 Auto Scaling 的目标跟踪扩缩策略。

限制

以下是使用目标跟踪扩缩策略时的限制：

- 可扩展目标不能是 Amazon EMR 集群。Amazon EMR 不支持目标跟踪扩缩策略。
- 当 Amazon MSK 集群是可扩展目标时，横向缩减将禁用且无法启用。
- 您不能使用`RegisterScalableTarget`或 `PutScalingPolicy` API 操作来更新 AWS Auto Scaling 扩展计划。有关如何使用扩缩计划的更多信息，请参阅 AWS Auto Scaling 文档。
- 在可扩展资源上查看、添加、更新或移除目标跟踪扩缩策略的控制台访问权限取决于您使用的资源。有关更多信息，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务。

使用创建目标跟踪扩展策略 AWS CLI

您可以通过使用来执行以下配置任务，为 Application Auto Scaling 创建目标跟踪扩展策略。 AWS CLI

1. 注册可扩展目标。
2. 在可扩展目标上添加目标跟踪扩缩策略。

为简洁起见，本主题中的示例说明了用于 Amazon EC2 Spot 实例集的 CLI 命令。要指定不同的可扩展目标，请在 `--service-namespace` 中指定其命名空间，在 `--scalable-dimension` 中指定其可扩展维度，并在 `--resource-id` 中指定其资源 ID。有关每项服务的更多信息和示例，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务 中的主题。

使用时 AWS CLI，请记住您的命令在 AWS 区域 配置文件中运行。如果您想要在不同的区域中运行命令，可以为配置文件更改默认区域，或者与命令一起使用 `--region` 参数。

内容

- 注册可扩展目标

- 创建目标跟踪扩缩策略
- 描述目标跟踪扩缩策略
- 删除目标跟踪扩缩策略

注册可扩展目标

如果您尚未注册，请注册可扩展目标。使用register-scalable-target命令将目标服务中的特定资源注册为可扩展目标。以下示例使用 Application Auto Scaling 注册 Spot 实例集请求。Application Auto Scaling 可以扩展 Spot 实例集中的实例数，最少为 2 个实例，最多为 10 个实例。将每个*******替换为您自己的信息。

Linux、macOS 或 Unix

```
aws application-autoscaling register-scalable-target --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
  --min-capacity 2 --max-capacity 10
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace ec2 --
scalable-dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-
request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE --min-capacity 2 --max-capacity 10
```

如果成功，该命令会返回可扩展目标的 ARN。

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

创建目标跟踪扩缩策略

要创建目标跟踪扩展策略，您可以使用以下示例来帮助您入门。

创建目标跟踪扩展策略

1. 使用以下cat命令将扩展策略的目标值和预定义的指标规范存储在主目录中名为config.json的 JSON 文件中。以下是将平均 CPU 利用率保持在 50% 的目标跟踪配置示例。

```
$ cat ~/config.json
{
  "TargetValue": 50.0,
  "PredefinedMetricSpecification":
    {
      "PredefinedMetricType": "EC2SpotFleetRequestAverageCPUUtilization"
    }
}
```

有关更多信息，请参阅《App licati PredefinedMetricSpecificationon Auto Scaling API 参考》中的。

或者，您可以使用自定义指标进行扩展，方法是创建自定义指标规范，并为中的每个参数添加值 CloudWatch。以下是将指定指标的平均利用率保持在 100 的目标跟踪配置示例。

```
$ cat ~/config.json
{
   "TargetValue": 100.0,
   "CustomizedMetricSpecification":{
      "MetricName": "MyUtilizationMetric",
      "Namespace": "MyNamespace",
      "Dimensions": [
         {
            "Name": "MyOptionalMetricDimensionName",
            "Value": "MyOptionalMetricDimensionValue"
         }
      ],
      "Statistic": "Average",
      "Unit": "Percent"
   }
}
```

有关更多信息，请参阅《App licati CustomizedMetricSpecificationon Auto Scaling API 参考》中的。

2. 使用以下put-scaling-policy命令以及您创建config.json的文件来创建名为的扩展策略cpu50-target-tracking-scaling-policy。

   Linux、macOS 或 Unix

```
aws application-autoscaling put-scaling-policy --service-namespace ec2 \
```

```
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
  --policy-name cpu50-target-tracking-scaling-policy --policy-type
 TargetTrackingScaling \
  --target-tracking-scaling-policy-configuration file://config.json
```

Windows

```
aws application-autoscaling put-scaling-policy --service-namespace ec2 --scalable-
dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-request/
sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE --policy-name cpu50-target-tracking-
scaling-policy --policy-type TargetTrackingScaling --target-tracking-scaling-
policy-configuration file://config.json
```

如果成功，此命令将返回代表您创建的两个 CloudWatch 警报的 ARN 和名称。

```
{
    "PolicyARN": "arn:aws:autoscaling:region:account-
id:scalingPolicy:policy-id:resource/ec2/spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE:policyName/cpu50-target-tracking-scaling-policy",
    "Alarms": [
        {
            "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-
spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-
b46e-434a-a60f-3b36d653feca",
            "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-b46e-434a-a60f-3b36d653feca"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-
spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-
d19b-4a63-a812-6c67aaf2910d",
            "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-d19b-4a63-a812-6c67aaf2910d"
        }
    ]
}
```

描述目标跟踪扩缩策略

您可以使用以下describe-scaling-policies命令描述指定服务命名空间的所有扩展策略。

```
aws application-autoscaling describe-scaling-policies --service-namespace ec2
```

您可使用 `--query` 参数筛选结果以仅显示目标跟踪扩展策略。有关 `query` 的语法的更多信息，请参阅 AWS Command Line Interface 用户指南中的控制 AWS CLI的命令输出。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scaling-policies --service-namespace ec2 \
  --query 'ScalingPolicies[?PolicyType==`TargetTrackingScaling`]'
```

Windows

```
aws application-autoscaling describe-scaling-policies --service-namespace ec2 --query
 "ScalingPolicies[?PolicyType==`TargetTrackingScaling`]"
```

下面是示例输出。

```
[
    {
        "PolicyARN": "PolicyARN",
        "TargetTrackingScalingPolicyConfiguration": {
            "PredefinedMetricSpecification": {
                "PredefinedMetricType": "EC2SpotFleetRequestAverageCPUUtilization"
            },
            "TargetValue": 50.0
        },
        "PolicyName": "cpu50-target-tracking-scaling-policy",
        "ScalableDimension": "ec2:spot-fleet-request:TargetCapacity",
        "ServiceNamespace": "ec2",
        "PolicyType": "TargetTrackingScaling",
        "ResourceId": "spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE",
        "Alarms": [
            {
                "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-
spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-
b46e-434a-a60f-3b36d653feca",
                "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmHigh-d4f0770c-b46e-434a-a60f-3b36d653feca"
            },
            {
```

```
                "AlarmARN": "arn:aws:cloudwatch:region:account-id:alarm:TargetTracking-
spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-
d19b-4a63-a812-6c67aaf2910d",
                "AlarmName": "TargetTracking-spot-fleet-request/sfr-73fbd2ce-
aa30-494c-8788-1cee4EXAMPLE-AlarmLow-1b437334-d19b-4a63-a812-6c67aaf2910d"
            }
        ],
        "CreationTime": 1515021724.807
    }
]
```

删除目标跟踪扩缩策略

当您完成目标跟踪扩展策略后，可以使用delete-scaling-policy命令将其删除。

以下命令将删除指定 Spot 队组请求的目标跟踪扩展策略。它还会删除 Application Auto Scaling 代表你创建的 CloudWatch 警报。

Linux、macOS 或 Unix

```
aws application-autoscaling delete-scaling-policy --service-namespace ec2 \
  --scalable-dimension ec2:spot-fleet-request:TargetCapacity \
  --resource-id spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE \
  --policy-name cpu50-target-tracking-scaling-policy
```

Windows

```
aws application-autoscaling delete-scaling-policy --service-namespace ec2 --scalable-
dimension ec2:spot-fleet-request:TargetCapacity --resource-id spot-fleet-request/
sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE --policy-name cpu50-target-tracking-scaling-
policy
```

使用指标数学为 Application Auto Scaling 创建目标跟踪扩展策略

使用指标数学，您可以查询多个 CloudWatch 指标，并使用数学表达式根据这些指标创建新的时间序列。您可以在 CloudWatch 控制台中可视化生成的时间序列并将其添加到仪表板中。有关指标数学的更多信息，请参阅 Amazon CloudWatch 用户指南中的使用指标数学。

以下考虑因素适用于指标数学表达式：

- 您可以查询任何可用的 CloudWatch 指标。每个指标都是指标名称、命名空间和零个或多个维度的唯一组合。
- 您可以使用任何算术运算符 (+-*/^)、统计函数（例如 AVG 或 SUM）或其他支持的函数。CloudWatch
- 您可以在数学表达式的公式中同时使用指标和其他数学表达式的结果。
- 指标规范中使用的任何表达式最终都必须返回一个单个时间序列。
- 您可以使用 CloudWatch 控制台或 CloudWatch GetMetricDataAPI 验证指标数学表达式是否有效。

主题

示例：每个任务的 Amazon SQS 队列积压

要计算每个任务的 Amazon SQS 队列积压，请获取可用于从队列中检索的消息的大致数量，然后将该数字除以服务中运行的 Amazon ECS 任务的数量。有关更多信息，请参阅计算博客上使用自定义指标的亚马逊弹性容器服务 (ECS) A AWS uto Sc aling。

表达式的逻辑如下：

```
sum of (number of messages in the queue)/(number of tasks that are
currently in the RUNNING state)
```

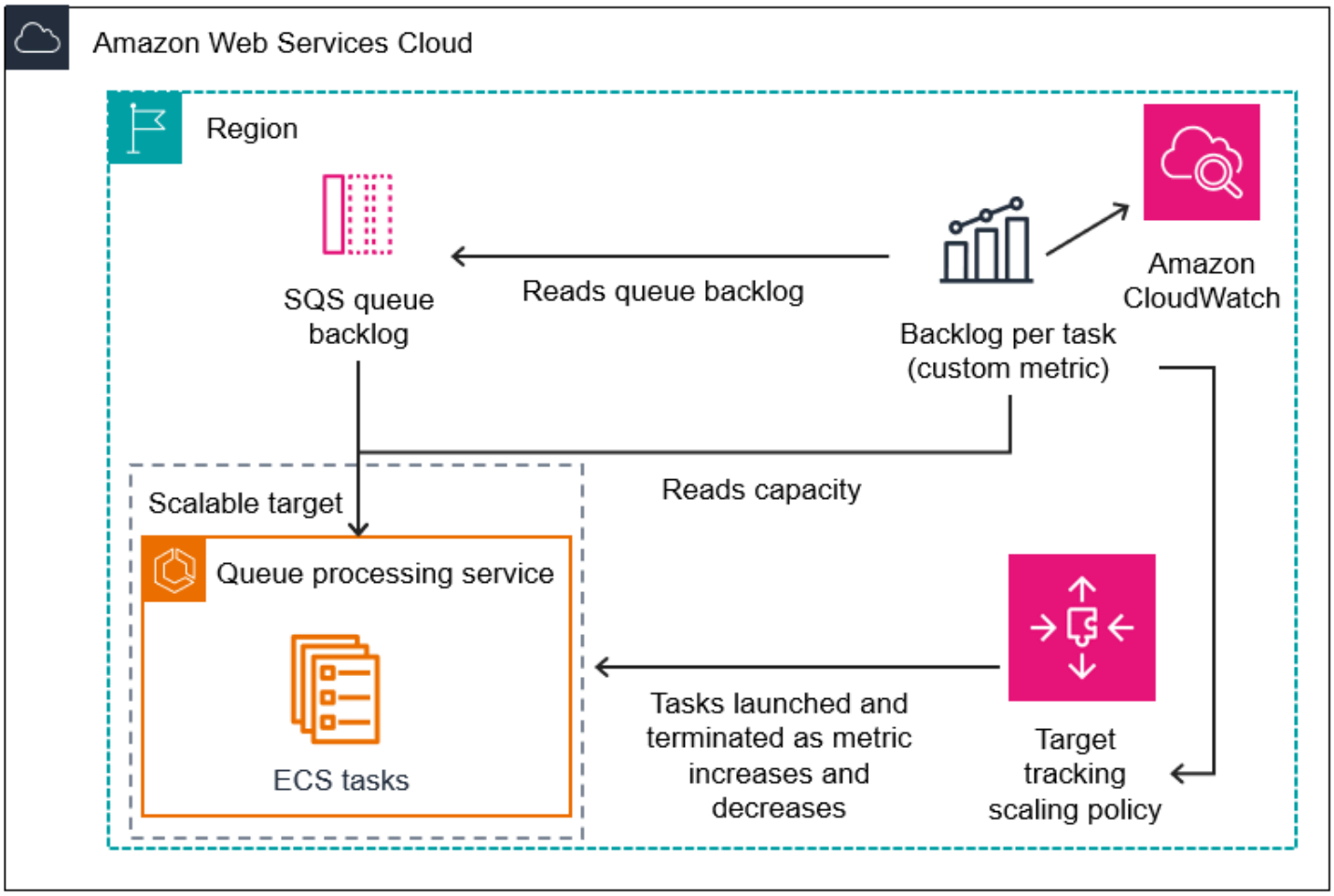
那么您的 CloudWatch 指标信息如下所示。

ID	CloudWatch 公制	Statistic	周期
m1	ApproximateNumberOfMessagesVisible	Sum	1 minute
m2	RunningTaskCount	平均值	1 minute

您的指标数学 ID 和表达式如下所示。

ID	Expression
e1	(m1)/(m2)

下图说明了该指标的架构：



使用该指标数学来创建目标跟踪扩展策略 (AWS CLI)

1. 将指标数学表达式作为自定义指标规范的一部分存储在名为 `config.json` 的 JSON 文件中。

   使用下面的示例帮助您快速开始。将每个`#######`替换为您自己的信息。

```
{
    "CustomizedMetricSpecification": {
        "Metrics": [
            {
```

```
                "Label": "Get the queue size (the number of messages waiting to be
processed)",
                "Id": "m1",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "ApproximateNumberOfMessagesVisible",
                        "Namespace": "AWS/SQS",
                        "Dimensions": [
                            {
                                "Name": "QueueName",
                                "Value": "my-queue"
                            }
                        ]
                    },
                    "Stat": "Sum"
                },
                "ReturnData": false
            },
            {
                "Label": "Get the ECS running task count (the number of currently
running tasks)",
                "Id": "m2",
                "MetricStat": {
                    "Metric": {
                        "MetricName": "RunningTaskCount",
                        "Namespace": "ECS/ContainerInsights",
                        "Dimensions": [
                            {
                                "Name": "ClusterName",
                                "Value": "my-cluster"
                            },
                            {
                                "Name": "ServiceName",
                                "Value": "my-service"
                            }
                        ]
                    },
                    "Stat": "Average"
                },
                "ReturnData": false
            },
            {
                "Label": "Calculate the backlog per instance",
                "Id": "e1",
```

```
                "Expression": "m1 / m2",
                "ReturnData": true
            }
        ]
    },
    "TargetValue": 100
}
```

有关更多信息，请参阅《App licati TargetTrackingScalingPolicyConfigurationon Auto Scaling API 参考》中的。

> **Note**
>
> 以下是一些其他资源，可以帮助您查找指标名称、命名空间、维度和指标 CloudWatch 统计信息：
>
> - 有关 AWS 服务的可用指标的信息，请参阅《亚马逊 CloudWatch 用户指南》中发布 CloudWatch 指标的AWS 服务。
> - 要使用获取指标的确切指标名称、命名空间和维度（如果适用）AWS CLI，请参阅列表 CloudWatch 指标。

2. 要创建此策略，请使用 JSON 文件作为输入运行put-scaling-policy命令，如以下示例所示。

```
aws application-autoscaling put-scaling-policy --policy-name sqs-backlog-target-tracking-scaling-policy \
  --service-namespace ecs --scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service \
  --policy-type TargetTrackingScaling --target-tracking-scaling-policy-configuration file://config.json
```

如果成功，此命令将返回策略的 Amazon 资源名称 (ARN) 和代表您创建的两个 CloudWatch 警报的 ARN。

```
{
    "PolicyARN": "arn:aws:autoscaling:us-west-2:012345678910:scalingPolicy:
 8784a896-b2ba-47a1-b08c-27301cc499a1:resource/ecs/service/my-cluster/my-service:policyName/sqs-backlog-target-tracking-scaling-policy",
    "Alarms": [
        {
```

```
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:012345678910:alarm:TargetTracking-service/my-cluster/my-service-
AlarmHigh-9bc77b56-0571-4276-ba0f-d4178882e0a0",
            "AlarmName": "TargetTracking-service/my-cluster/my-service-
AlarmHigh-9bc77b56-0571-4276-ba0f-d4178882e0a0"
        },
        {
            "AlarmARN": "arn:aws:cloudwatch:us-
west-2:012345678910:alarm:TargetTracking-service/my-cluster/my-service-
AlarmLow-9b6ad934-6d37-438e-9e05-02836ddcbdc4",
            "AlarmName": "TargetTracking-service/my-cluster/my-service-
AlarmLow-9b6ad934-6d37-438e-9e05-02836ddcbdc4"
        }
    ]
}
```

> **Note**
>
> 如果此命令引发错误，请确保已将 AWS CLI 本地版本更新到最新版本。

限制

- 最大请求大小为 50KB。这是您在策略定义中使用公制数学时 PutScalingPolicyAPI 请求的总有效负载大小。如果您超过此限制，Application Auto Scaling 会拒绝该请求。
- 结合使用指标数学与目标跟踪扩缩策略时，不支持以下服务：
  - Amazon Keyspaces（Apache Cassandra 兼容）
  - DynamoDB
  - Amazon EMR
  - Amazon MSK
  - Amazon Neptune

分步扩展策略

分步扩展策略根据 CloudWatch 警报以预定义的增量扩展应用程序的容量。您可以定义单独的扩缩策略，以便在超过警报阈值时处理横向扩展（增加容量）和横向缩减（减少容量）。

使用分步扩展策略，您可以创建和管理调用扩展过程的 CloudWatch 警报。当警报被触发时，Application Auto Scaling 会启动与该警报关联的扩缩策略。

分步扩缩策略使用一组调整（称为分步调整）来扩缩容量。调整的大小将根据超出警报阈值的规模而变化。

- 如果违例超过第一个阈值，Application Auto Scaling 将应用第一步调整。
- 如果违例超过第二个阈值，Application Auto Scaling 将应用第二步调整，以此类推。

这使扩缩策略能够针对警报指标的微小和重大变化作出适当响应。

当扩缩活动正在进行中时，该策略将继续响应其他警报。这意味着 Application Auto Scaling 将在所有警报发生时对其进行评估。冷却时间用于防止由于快速连续发生多个警报而导致的过度扩缩。

与目标跟踪一样，分步扩缩可以帮助在流量发生变化时自动扩缩应用程序的容量。但是，目标跟踪策略往往更易于实施和管理，以满足稳定的扩缩需求。

您可以将分步扩缩策略与以下可扩展目标配合使用：

- AppStream 2.0 支舰队
- Aurora 数据库集群
- ECS 服务
- EMR 集群
- SageMaker 端点变体
- SageMaker 推理组件
- SageMaker 无服务器配置的并发性
- Spot Fleets
- 自定义资源

主题

步进缩放的工作原理

本主题描述了步进缩放的工作原理，并介绍了步进扩展策略的关键要素。

内容

工作方式

要使用步进缩放，您需要创建一个 CloudWatch 警报，用于监控可扩展目标的指标。定义确定触发警报的指标、阈值和评估周期数。您还可以创建分步扩缩策略，在其中定义在突破警报阈值时如何扩缩容量并将其与可扩展目标相关联。

在策略中添加分步调整。您可以根据警报的阈值突破大小定义不同的分步调整。例如：

- 如果警报指标达到 60%，则横向扩展 10 个容量单位
- 如果警报指标达到 75%，则横向扩展 30 个容量单位
- 如果警报指标达到 85%，则横向扩展 40 个容量单位

当在指定的评估周期数内超过警报阈值时，Application Auto Scaling 将应用策略中定义的分步调整。针对其他警报触发情况，可以继续进行调整，直到警报状态恢复为 OK。

扩缩活动在两者之间有冷却时间，以防止容量快速波动。您可以选择为扩缩策略配置冷却时间。

分步调整

在创建分步扩缩策略时，您可以指定一个或多个分步调整，它们会动态地根据警报违规的大小自动扩缩目标容量。每个分步调整指定以下内容：

- 指标值的下限
- 指标值的上限
- 要扩展的数量（基于扩展调整类型）

CloudWatch 根据与 CloudWatch 警报关联的指标的统计数据聚合指标数据点。超过警报时，将调用相应的扩缩策略。Application Auto Scaling 将您指定的聚合类型应用于来自的最新指标数据点 CloudWatch （而不是原始指标数据）。它将此聚合指标值与步进调整定义的上限和下限进行比较，以确定执行哪个步进调整。

您可以指定相对于违例阈值的上限和下限。例如，假设您针对指标高于 50% 时发出了 CloudWatch 警报并制定了扩展策略。然后，当指标低于 50% 时，又发出了另一个警报并采取横向缩减策略。您进行了一组分步调整，每个策略的调整类型为 `PercentChangeInCapacity`：

示例：扩展策略的步进调整

下限	上限	调整
0	10	0
10	20	10
20	null	30

示例：缩放策略的步进调整

下限	上限	调整
-10	0	0
-20	-10	-10
null	-20	-30

这将创建以下扩展配置。

```
Metric value

-infinity          30%    40%          60%     70%             infinity
-------------------------------------------------------------------------
          -30%      | -10% | Unchanged  | +10%  |       +30%
-------------------------------------------------------------------------
```

现在，假设您有一个容量为 10 的可扩展目标，使用基于此目标的扩缩配置。以下几点总结了扩展配置相对于可扩展目标的容量的行为：

- 在聚合指标值大于 40 且小于 60 时，将保留原始容量。
- 如果指标值达到 60，则 Application Auto Scaling 将可扩展目标的容量增加 1，达到 11。这基于向外扩展策略的第二个步进调整（增加 10 的 10%）。在增加新容量后，Application Auto Scaling 将当前容量增加到 11。如果即使在增加该容量后指标值仍增加到 70，Application Auto Scaling 会将目标容量增加 3 以达到 14。这基于向外扩展策略的第三个步进调整（增加 11 的 30%，即 3.3，向下舍入到 3）。
- 如果指标值达到 40，根据横向缩减策略的第二个分步调整（减去 14 的 10%，即 1.4，向下舍入到 1），Application Auto Scaling 将可扩展目标的容量减少 1，达到 13。如果在减少该容量后指标值仍降到 30，根据横向缩减策略的第三个步骤调整（减去 13 的 30%，即 3.9，向下舍入到 3），Application Auto Scaling 将目标容量减小 3 以达到 10。

为扩展策略指定步进调整时，请注意以下事项：

- 分步调整范围不能重叠或有间隙。
- 只有一个分步调整可以有空下限 (负无穷)。如果一个分步调整有负下限，则必须有一个分步调整有空下限。
- 只有一个分步调整可以有空上限 (正无穷)。如果一个分步调整有正上限，则必须有一个分步调整有空上限。
- 同一分步调整中的上限和下限不能为空。
- 如果指标值高于违例阈值，则含下限而不含上限。如果指标值低于违例阈值，则不含下限而含上限。

扩展调整类型

您可以根据您选择的扩展调整类型来定义执行最佳扩展操作的扩展策略。您可以将调整类型指定为可扩展目标的当前容量的百分比或绝对数。

对于分步扩缩策略，Application Auto Scaling 支持以下调整类型：

- ChangeInCapacity—按指定值增加或减少可扩展目标的当前容量。正值将增加容量，负值将减少容量。例如：如果当前容量为 3 且调整值为 5，则 Application Auto Scaling 将为容量增加 5（总量为 8）。
- ExactCapacity— 将可扩展目标的当前容量更改为指定值。为此调整类型指定一个非负值。例如：如果当前容量为 3 且调整值为 5，则 Application Auto Scaling 将容量更改为 5。
- PercentChangeInCapacity—按指定的百分比增加或减少可扩展目标的当前容量。正值将增加容量，负值将减少容量。例如：如果当前容量为 10 且调整值为 10%，则 Application Auto Scaling 将为容量增加 1（总量为 11）。

> **Note**
>
> 如果得出的值不是整数，Application Auto Scaling 将进行舍入，如下所示：
>
> - 大于 1 的值向下取整。例如，12.7 取整为 12。
> - 0 和 1 之间的值舍入到 1。例如，.67 取整为 1。
> - 0 和 -1 之间的值舍入到 -1。例如，-.58 取整为 -1。
> - 小于 -1 的值向上取整。例如，-6.67 取整为 -6。

使用 PercentChangeInCapacity，您还可以使用`MinAdjustmentMagnitude`参数指定最小缩放量。例如，假定您创建一个增加 25% 的策略，并指定最小扩展量为 2。如果可扩展目标的容量为 4 并执行该扩展策略，4 的 25% 为 1。不过，由于您指定最小扩展量为 2，Application Auto Scaling 将增加 2。

冷却时间

您可以选择在分步扩展策略中定义冷却时间。

冷却时间指定了扩展策略等待上一个扩展活动生效的时间量。

有两种方法可以计划分步扩展配置的冷却时间使用：

- 使用横向扩展策略的冷却时间，目的是持续（但不过度）横向扩展。Application Auto Scaling 使用扩展策略成功横向扩展后，它将开始计算冷却时间。除非触发更大的横向扩展或冷却时间结束，否则扩展策略不会再次增加所需容量。尽管此向外扩展冷却时间有效，但启动向外扩展活动所添加的容量将计算为下一个向外扩展活动所需容量的一部分。

- 使用横向缩减策略冷却时间，目的是以保守方式进行横向缩减，以保护应用程序的可用性，因此在横向缩减冷却时间过期之前，横向缩减活动会被阻止。但是，如果另一个警报在缩减冷却时间内触发了向外扩展活动，Application Auto Scaling 将立即向外扩展目标。在这种情况下，横向缩减冷却时间会停止而不完成。

例如，出现流量高峰时会触发警报，Application Auto Scaling 会自动增加容量以帮助处理增加的负载。如果为横向扩展策略设置冷却时间，当警报触发策略以将容量增加 2 时，扩展活动成功完成，横向扩展冷却时间开始。如果在冷却时间内再次触发警报，但进行了 3 这样的更大幅度的步进调整，之前增加的 2 将视为当前容量的一部分。因此，仅在容量中增加 1。与等待冷却时间过期相比，这可以实现更快的扩展，但不会增加超出您需求的容量。

冷却时间以秒为单位进行度量，仅适用于与扩展策略相关的扩展活动。在冷却时间内，当计划的操作在计划的时间开始时，它可以立即触发扩展活动，而无需等待冷却时间到期。

如果未指定值，则默认值为 300。

扩缩策略创建、管理和删除的常用命令

使用扩缩策略的常用命令包括：

- register-scalable-target注册 AWS 或自定义资源作为可扩展目标（Application Auto Scaling 可以扩展的资源），以及暂停和恢复扩展。
- put-scaling-policy为现有可扩展目标添加或修改扩展策略。
- describe-scaling-activities返回有关某个 AWS 区域中扩展活动的信息。
- describe-scaling-policies返回有关某个 AWS 区域中扩展策略的信息。
- delete-scaling-policy删除扩展策略。

注意事项

使用分步扩缩策略时，需要注意以下事项：

- 考虑是否可以足够准确地预测应用程序上的分步调整，以便使用分步扩缩。如果您的扩缩指标的升高或降低与可扩展目标的容量成比例，则建议您使用目标跟踪扩缩策略。您仍然可以选择使用步进扩展作为附加策略来实现更高级的配置。例如，您可以在利用率达到特定级别时配置更积极的响应。
- 确保在横向扩展和横向缩减之间选择足够的余量，以防止摇摆。摆动是横向缩减和横向扩展的无限循环。也就是说，如果采取扩展操作，则指标值将更改并启动另一个相反方向的扩展操作。

相关资源

有关为自动扩缩组创建分步扩缩策略的信息，请参阅《Amazon EC2 Auto Scaling 用户指南》中的 Amazon EC2 Auto Scaling 的步进和简单扩展策略。

限制

- 在可扩展资源上查看、添加、更新或移除分步扩缩策略的控制台访问权限取决于您使用的资源。有关更多信息，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务。

使用 AWS CLI创建分步扩缩策略

您可以通过使用来执行以下配置任务，为 Application Auto Scaling 创建步进扩展策略。 AWS CLI

1. 注册可扩展目标。
2. 在可扩展目标上添加分步扩缩策略。
3. 为策略创建 CloudWatch 警报。

为简洁起见，本主题中的示例说明了 Amazon ECS 服务的 CLI 命令。要指定不同的可扩展目标，请在 `--service-namespace` 中指定其命名空间，在 `--scalable-dimension` 中指定其可扩展维度，并在 `--resource-id` 中指定其资源 ID。有关每项服务的更多信息和示例，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务 中的主题。

使用时 AWS CLI，请记住您的命令在 AWS 区域 配置文件中运行。如果您想要在不同的区域中运行命令，可以为配置文件更改默认区域，或者与命令一起使用 `--region` 参数。

内容

注册可扩展目标

如果您尚未注册，请注册可扩展目标。使用register-scalable-target命令将目标服务中的特定资源注册为可扩展目标。以下示例使用 Application Auto Scaling 注册 Amazon ECS 服务。Application Auto Scaling 可扩展任务的数量，最少 2 个任务，最多 10 个任务。将每个*******替换为您自己的信息。

Linux、macOS 或 Unix

```
aws application-autoscaling register-scalable-target --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --min-capacity 2 --max-capacity 10
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace ecs --
scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
 --min-capacity 2 --max-capacity 10
```

如果成功，该命令会返回可扩展目标的 ARN。

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

创建分步扩缩策略

要为可扩展目标创建分步扩展策略，您可以使用以下示例来帮助您入门。

Scale out

为横向扩展（增加容量）创建分步扩展策略

1. 使用以下cat命令将步进扩展策略配置存储在主目录中名为config.json的 JSON 文件中。以下是一个配置示例，其调整类型为PercentChangeInCapacity，该配置根据以下步骤调整（假设 CloudWatch 警报阈值为 70）来增加可扩展目标的容量：

   - 当指标值大于或等于 70 但小于 85 时，将容量增加 10%
   - 当指标值大于或等于 85 但小于 95 时，将容量增加 20%

- 当指标值大于或等于 95 时，将容量增加 30%

```
$ cat ~/config.json
{
  "AdjustmentType": "PercentChangeInCapacity",
  "MetricAggregationType": "Average",
  "Cooldown": 60,
  "MinAdjustmentMagnitude": 1,
  "StepAdjustments": [
    {
      "MetricIntervalLowerBound": 0.0,
      "MetricIntervalUpperBound": 15.0,
      "ScalingAdjustment": 10
    },
    {
      "MetricIntervalLowerBound": 15.0,
      "MetricIntervalUpperBound": 25.0,
      "ScalingAdjustment": 20
    },
    {
      "MetricIntervalLowerBound": 25.0,
      "ScalingAdjustment": 30
    }
  ]
}
```

有关更多信息，请参阅《App licati StepScalingPolicyConfigurationon Auto Scaling API 参考》中的。

2. 使用以下put-scaling-policy命令以及您创建config.json的文件来创建名为的扩展策略my-step-scaling-policy。

   Linux、macOS 或 Unix

```
aws application-autoscaling put-scaling-policy --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --policy-name my-step-scaling-policy --policy-type StepScaling \
  --step-scaling-policy-configuration file://config.json
```

   Windows

```
aws application-autoscaling put-scaling-policy --service-namespace ecs --
scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-
service --policy-name my-step-scaling-policy --policy-type StepScaling --step-
scaling-policy-configuration file://config.json
```

输出包括作为策略唯一名称的 ARN。你需要它来为你的策略创建 CloudWatch 警报。

```
{
    "PolicyARN":
 "arn:aws:autoscaling:region:123456789012:scalingPolicy:ac542982-cbeb-4294-891c-
a5a941dfa787:resource/ecs/service/my-cluster/my-service:policyName/my-step-
scaling-policy"
}
```

Scale in

为缩小规模（减少容量）创建分步扩展策略

1. 使用以下cat命令将步进扩展策略配置存储在主目录中名为config.json的 JSON 文件中。以下是一个配置示例，其调整类型为ChangeInCapacity，根据以下步骤调整（假设 CloudWatch 警报阈值为 50），该配置会降低可扩展目标的容量：

   - 当指标值小于或等于 50 但大于 40 时，将容量减少 1
   - 当指标值小于或等于 40 但大于 30 时，将容量减少 2
   - 当指标值小于或等于 30 时，将容量减少 3

```
$ cat ~/config.json
{
  "AdjustmentType": "ChangeInCapacity",
  "MetricAggregationType": "Average",
  "Cooldown": 60,
  "StepAdjustments": [
    {
      "MetricIntervalUpperBound": 0.0,
      "MetricIntervalLowerBound": -10.0,
      "ScalingAdjustment": -1
    },
    {
```

```
      "MetricIntervalUpperBound": -10.0,
      "MetricIntervalLowerBound": -20.0,
      "ScalingAdjustment": -2
    },
    {
      "MetricIntervalUpperBound": -20.0,
      "ScalingAdjustment": -3
    }
  ]
}
```

有关更多信息，请参阅《App licati StepScalingPolicyConfigurationon Auto Scaling API 参考》中的。

2. 使用以下put-scaling-policy命令以及您创建config.json的文件来创建名为的扩展策略my-step-scaling-policy。

   Linux、macOS 或 Unix

```
aws application-autoscaling put-scaling-policy --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --policy-name my-step-scaling-policy --policy-type StepScaling \
  --step-scaling-policy-configuration file://config.json
```

   Windows

```
aws application-autoscaling put-scaling-policy --service-namespace ecs --
scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-
service --policy-name my-step-scaling-policy --policy-type StepScaling --step-
scaling-policy-configuration file://config.json
```

   输出包括作为策略唯一名称的 ARN。你需要它来为你的策略创建 CloudWatch 警报。

```
{
    "PolicyARN":
 "arn:aws:autoscaling:region:123456789012:scalingPolicy:ac542982-cbeb-4294-891c-
a5a941dfa787:resource/ecs/service/my-cluster/my-service:policyName/my-step-
scaling-policy"
}
```

创建调用扩缩策略的警报

最后，使用以下 CloudWatch put-metric-alarm命令创建警报，以便与步进缩放策略一起使用。在本示例中，您将根据平均 CPU 利用率发出警报。如果警报在至少两个连续 60 秒的评估期间达到 70% 的阈值，则它将被配置为处于 ALARM 状态。要指定其他 CloudWatch 指标或使用您自己的自定义指标，请在中指定其名称，在中`--metric-name`指定其命名空间`--namespace`。

Linux、macOS 或 Unix

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmHigh-ECS:service/my-
cluster/my-service \
  --metric-name CPUUtilization --namespace AWS/ECS --statistic Average \
  --period 60 --evaluation-periods 2 --threshold 70 \
  --comparison-operator GreaterThanOrEqualToThreshold \
  --dimensions Name=ClusterName,Value=default Name=ServiceName,Value=sample-app-service
 \
  --alarm-actions PolicyARN
```

Windows

```
aws cloudwatch put-metric-alarm --alarm-name Step-Scaling-AlarmHigh-ECS:service/
my-cluster/my-service --metric-name CPUUtilization --namespace AWS/ECS --statistic
 Average --period 60 --evaluation-periods 2 --threshold 70 --comparison-operator
 GreaterThanOrEqualToThreshold --dimensions Name=ClusterName,Value=default
 Name=ServiceName,Value=sample-app-service --alarm-actions PolicyARN
```

描述分步扩缩策略

您可以使用以下describe-scaling-policies命令描述指定服务命名空间的所有扩展策略。

```
aws application-autoscaling describe-scaling-policies --service-namespace ecs
```

您可以使用 `--query` 参数将结果筛选为仅步进扩展策略。有关 `query` 的语法的更多信息，请参阅 AWS Command Line Interface 用户指南中的控制 AWS CLI的命令输出。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scaling-policies --service-namespace ecs \
  --query 'ScalingPolicies[?PolicyType==`StepScaling`]'
```

Windows

```
aws application-autoscaling describe-scaling-policies --service-namespace ecs --query
 "ScalingPolicies[?PolicyType==`StepScaling`]"
```

下面是示例输出。

```
[
    {
        "PolicyARN": "PolicyARN",
        "StepScalingPolicyConfiguration": {
            "MetricAggregationType": "Average",
            "Cooldown": 60,
            "StepAdjustments": [
                {
                    "MetricIntervalLowerBound": 0.0,
                    "MetricIntervalUpperBound": 15.0,
                    "ScalingAdjustment": 1
                },
                {
                    "MetricIntervalLowerBound": 15.0,
                    "MetricIntervalUpperBound": 25.0,
                    "ScalingAdjustment": 2
                },
                {
                    "MetricIntervalLowerBound": 25.0,
                    "ScalingAdjustment": 3
                }
            ],
            "AdjustmentType": "ChangeInCapacity"
        },
        "PolicyType": "StepScaling",
        "ResourceId": "service/my-cluster/my-service",
        "ServiceNamespace": "ecs",
        "Alarms": [
            {
                "AlarmName": "Step-Scaling-AlarmHigh-ECS:service/my-cluster/my-
service",
                "AlarmARN": "arn:aws:cloudwatch:region:012345678910:alarm:Step-Scaling-
AlarmHigh-ECS:service/my-cluster/my-service"
            }
        ],
        "PolicyName": "my-step-scaling-policy",
```

```
        "ScalableDimension": "ecs:service:DesiredCount",
        "CreationTime": 1515024099.901
    }
]
```

删除分步扩缩策略

当您不再需要某个步进扩展策略时，可将其删除。要同时删除扩展策略和 CloudWatch 警报，请完成以下任务。

删除您的扩展策略

使用以下 delete-scaling-policy 命令。

Linux、macOS 或 Unix

```
aws application-autoscaling delete-scaling-policy --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service \
  --policy-name my-step-scaling-policy
```

Windows

```
aws application-autoscaling delete-scaling-policy --service-namespace ecs --scalable-
dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service --
policy-name my-step-scaling-policy
```

删除 CloudWatch 警报

使用 delete-alarms 命令。您可以一次删除一个或多个警报。例如，使用以下命令可删除 Step-Scaling-AlarmHigh-ECS:service/my-cluster/my-service 和 Step-Scaling-AlarmLow-ECS:service/my-cluster/my-service 警报：

```
aws cloudwatch delete-alarms --alarm-name Step-Scaling-AlarmHigh-ECS:service/my-
cluster/my-service Step-Scaling-AlarmLow-ECS:service/my-cluster/my-service
```

教程：配置自动扩缩以处理繁重的工作负载

> **Important**
>
> 在开始学习本教程之前，我们建议您首先阅读以下介绍性教程：教程：通过 AWS CLI开始使用计划扩缩。

在本教程中，您将了解当应用程序的工作负载比正常工作负载重的情况下，如何根据时间范围进行横向扩展和横向缩减。当您的应用程序可能会突然有大量定期或季节性的访问者时，这会很有帮助。

您可以将目标跟踪扩缩策略与计划的扩缩一起使用来处理额外负载。计划的扩缩会根据您指定的计划代表您自动启动对 `MinCapacity` 和 `MaxCapacity` 的更改。当目标跟踪扩缩策略在资源上处于活动状态时，它可以根据当前资源利用率在新的最小容量和最大容量范围内动态扩展。

完成本教程后，您将了解如何：

- 使用计划的扩缩添加额外容量，以在达到限值之前满足重负载，然后在不再需要时删除额外容量。
- 使用目标跟踪扩缩策略根据当前资源利用率扩展您的应用程序。

目录

先决条件

本教程假定您已执行以下操作：

- 您已创建 AWS 账户。有关更多信息，请参阅进行设置以开始使用 Application Auto Scaling。
- 您已经安装并配置 AWS CLI。有关更多信息，请参阅设置 AWS CLI。

- 您的账户具有使用 Application Auto Scaling 将资源注册和取消注册为可扩展目标的所有必要权限。它还具有创建扩缩策略和计划的操作的所有必要权限。有关更多信息，请参阅适用于 Application Auto Scaling 的 Identity and Access Management。
- 您在非生产环境中拥有可用于本教程的受支持资源。如果您还没有，请立即创建一个。有关使用 Application Auto Scaling 的 AWS 服务和资源的信息，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务 部分。

> **Note**
>
> 完成本教程时有两个步骤，您可在这两个步骤中将最大和最小容量值设置为 0，以将当前容量重置为 0。根据您当前通过 Application Auto Scaling 使用的资源，您可能无法在这些步骤中将当前容量设置为 0。为帮助您解决此问题，输出中将会显示一条消息，指示最小容量不能小于指定的值，并将提供 AWS 资源可接受的最小容量值。

步骤 1：注册您的可扩展目标

首先使用 Application Auto Scaling 将您的资源注册为可扩展目标。可扩展目标是 Application Auto Scaling 可以横向扩展或横向缩减的资源。

向 Application Auto Scaling 注册您的可扩展目标

- 使用以下 register-scalable-target 命令注册新的可扩展目标。将 `--min-capacity` 和 `--max-capacity` 值设置为 0 以将当前容量重置为 0。

  将 `--service-namespace` 中的示例文本替换为您通过 Application Auto Scaling 使用的 AWS 服务的命名空间，将 `--scalable-dimension` 替换为与您注册的资源关联的可扩展维度，并将 `--resource-id` 替换为资源的标识符。这些值因使用的资源以及资源 ID 的构造方式而异。有关更多信息，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务 部分中的主题。这些主题包括向您显示如何使用 Application Auto Scaling 注册可扩展目标的示例命令。

  Linux、macOS 或 Unix

```
aws application-autoscaling register-scalable-target \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --min-capacity 0 --max-capacity 0
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace namespace
 --scalable-dimension dimension --resource-id identifier --min-capacity 0 --max-
capacity 0
```

如果成功，该命令会返回可扩展目标的 ARN。

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

步骤 2：根据您的要求设置计划的操作

您可以使用 put-scheduled-action 命令创建配置为满足业务需求的计划操作。在本教程中，我们重点介绍一种配置，通过将容量减少到 0 来停止在工作时间以外消耗资源。

创建在噪声横向扩展的计划操作

1. 要横向扩展可扩展目标，请使用以下 put-scheduled-action 命令。使用 cron 表达式将 `--schedule` 参数包含在采用 UTC 时间的定期计划中。

   根据指定的计划（UTC 时间每天上午 9:00），Application Auto Scaling 会将 `MinCapacity` 和 `MaxCapacity` 值更新为 1-5 个容量单位的所需范围。

   Linux、macOS 或 Unix

   ```
   aws application-autoscaling put-scheduled-action \
     --service-namespace namespace \
     --scalable-dimension dimension \
     --resource-id identifier \
     --scheduled-action-name my-first-scheduled-action \
     --schedule "cron(0 9 * * ? *)" \
     --scalable-target-action MinCapacity=1,MaxCapacity=5
   ```

   Windows

```
aws application-autoscaling put-scheduled-action --service-namespace namespace --
scalable-dimension dimension --resource-id identifier --scheduled-action-name my-
first-scheduled-action --schedule "cron(0 9 * * ? *)" --scalable-target-action
 MinCapacity=1,MaxCapacity=5
```

如果此命令成功执行，将不会返回任何输出。

2. 要确认您的计划操作是否存在，请使用以下 describe-scheduled-actions 命令。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scheduled-actions \
  --service-namespace namespace \
  --query 'ScheduledActions[?ResourceId==`identifier`]'
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-
namespace namespace --query "ScheduledActions[?ResourceId==`identifier`]"
```

下面是示例输出。

```
[
    {
        "ScheduledActionName": "my-first-scheduled-action",
        "ScheduledActionARN": "arn",
        "Schedule": "cron(0 9 * * ? *)",
        "ScalableTargetAction": {
            "MinCapacity": 1,
            "MaxCapacity": 5
        },
        ...
    }
]
```

创建在夜间横向缩减的计划操作

1. 重复上述过程以创建另一个计划的操作，Application Auto Scaling 使用该操作在当天结束时进行横向缩减。

根据指定的计划（UTC 时间每天晚上 8:00），Application Auto Scaling 会将目标的 `MinCapacity` 和 `MaxCapacity` 更新为 0，如以下 put-scheduled-action 命令所示。

Linux、macOS 或 Unix

```
aws application-autoscaling put-scheduled-action \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --scheduled-action-name my-second-scheduled-action \
  --schedule "cron(0 20 * * ? *)" \
  --scalable-target-action MinCapacity=0,MaxCapacity=0
```

Windows

```
aws application-autoscaling put-scheduled-action --service-namespace namespace --
scalable-dimension dimension --resource-id identifier --scheduled-action-name my-
second-scheduled-action --schedule "cron(0 20 * * ? *)" --scalable-target-action
 MinCapacity=0,MaxCapacity=0
```

2. 要确认您的计划操作是否存在，请使用以下 describe-scheduled-actions 命令。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scheduled-actions \
  --service-namespace namespace \
  --query 'ScheduledActions[?ResourceId==`identifier`]'
```

Windows

```
aws application-autoscaling describe-scheduled-actions --service-
namespace namespace --query "ScheduledActions[?ResourceId==`identifier`]"
```

下面是示例输出。

```
[
    {
        "ScheduledActionName": "my-first-scheduled-action",
        "ScheduledActionARN": "arn",
        "Schedule": "cron(0 9 * * ? *)",
```

```
        "ScalableTargetAction": {
            "MinCapacity": 1,
            "MaxCapacity": 5
        },
        ...
    },
    {
        "ScheduledActionName": "my-second-scheduled-action",
        "ScheduledActionARN": "arn",
        "Schedule": "cron(0 20 * * ? *)",
        "ScalableTargetAction": {
            "MinCapacity": 0,
            "MaxCapacity": 0
        },
        ...
    }
]
```

步骤 3：添加目标跟踪扩缩策略

现在，您已经准备好基本计划，请添加一个目标跟踪扩缩策略，以根据当前资源利用率进行扩展。

通过目标跟踪，Application Auto Scaling 会将策略中的目标值与指定指标的当前值进行比较。如果两个值在一段时间内不相等，Application Auto Scaling 会添加或删除容量以保持稳定的性能。随着应用程序负载和指标值的增加，Application Auto Scaling 会尽可能快地增加容量，而不会超过 `MaxCapacity`。当 Application Auto Scaling 由于负载最小而删除容量时，它会在不低于 `MinCapacity` 时执行此操作。通过根据使用情况调整容量，您只需支付应用程序所需的费用。

如果由于您的应用程序没有任何负载而导致指标数据不足，则 Application Auto Scaling 不会添加或删除容量。换句话说，Application Auto Scaling 在信息不足的情况下会优先考虑可用性。

您可以添加多个扩缩策略，但请确保不会添加冲突的分步扩缩策略，这可能会导致不良行为。例如，如果步进扩展策略在目标跟踪策略准备执行缩减之前启动缩减活动，则不会阻止缩减活动。在横向缩减活动完成后，目标跟踪策略可能会指示 Application Auto Scaling 再次横向扩展。

创建目标跟踪扩展策略

1. 使用以下 put-scaling-policy 命令创建策略。

   最常用于目标跟踪的指标是预定义的，您可以在不提供 CloudWatch 的完整指标规范的情况下使用这些指标。有关可用预定义指标的更多信息，请参阅 目标跟踪扩缩策略。

在运行此命令之前，请确保您的预定义指标需要目标值。例如，要在 CPU 利用率达到 50% 时横向扩展，请将目标值指定为 50.0。或者，要在使用率达到 70% 时横向扩展 Lambda 预置并发，请将目标值指定为 0.7。有关特定资源目标值的信息，请参阅服务提供的有关如何配置目标跟踪的文档。有关更多信息，请参阅AWS 可以与 Application Auto Scaling 一起使用的服务。

Linux、macOS 或 Unix

```
aws application-autoscaling put-scaling-policy \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --policy-name my-scaling-policy --policy-type TargetTrackingScaling \
  --target-tracking-scaling-policy-configuration '{ "TargetValue": 50.0,
 "PredefinedMetricSpecification": { "PredefinedMetricType": "predefinedmetric" }}'
```

Windows

```
aws application-autoscaling put-scaling-policy --service-namespace namespace --
scalable-dimension dimension --resource-id identifier --policy-name my-scaling-
policy --policy-type TargetTrackingScaling --target-tracking-scaling-policy-
configuration "{ \"TargetValue\": 50.0, \"PredefinedMetricSpecification\":
 { \"PredefinedMetricType\": \"predefinedmetric\" }}"
```

如果成功，此命令将返回包含代表您创建的两个 CloudWatch 警报的 ARN 和名称。

2. 要确认您的计划操作是否存在，请使用以下 describe-scaling-policies 命令。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scaling-policies --service-namespace namespace
 \
  --query 'ScalingPolicies[?ResourceId==`identifier`]'
```

Windows

```
aws application-autoscaling describe-scaling-policies --service-namespace namespace
 --query "ScalingPolicies[?ResourceId==`identifier`]"
```

下面是示例输出。

```
[
    {
        "PolicyARN": "arn",
        "TargetTrackingScalingPolicyConfiguration": {
            "PredefinedMetricSpecification": {
                "PredefinedMetricType": "predefinedmetric"
            },
            "TargetValue": 50.0
        },
        "PolicyName": "my-scaling-policy",
        "PolicyType": "TargetTrackingScaling",
        "Alarms": [],
        ...
    }
]
```

步骤 4：后续步骤

发生扩缩活动时，您可以在可扩展目标的扩缩活动输出中看到该活动的记录，例如：

```
Successfully set desired count to 1. Change successfully fulfilled by ecs.
```

要使用 Application Auto Scaling 监控扩缩活动，您可以使用以下 describe-scaling-activities 命令。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scaling-activities
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier
```

Windows

```
aws application-autoscaling describe-scaling-activities --service-namespace namespace
 --scalable-dimension dimension --resource-id identifier
```

第 5 步：清除

为防止您的账户对主动扩缩时创建的资源产生费用，您可以按如下方式清理关联的扩缩配置。

删除扩缩配置不会删除您基础的 AWS 资源。该操作也不会恢复为其原来的容量。您可以使用创建资源的服务的控制台来删除该资源或调整其容量。

删除计划的操作

以下 delete-scheduled-action 命令可删除指定的计划操作。如果您要保留创建的计划操作，您可以跳过此步骤。

Linux、macOS 或 Unix

```
aws application-autoscaling delete-scheduled-action \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --scheduled-action-name my-second-scheduled-action
```

Windows

```
aws application-autoscaling delete-scheduled-action --service-namespace namespace
 --scalable-dimension dimension --resource-id identifier --scheduled-action-name my-
second-scheduled-action
```

删除扩缩策略

以下 delete-scaling-policy 命令可删除指定的目标跟踪扩缩策略。如果您要保留创建的扩缩策略，您可以跳过此步骤。

Linux、macOS 或 Unix

```
aws application-autoscaling delete-scaling-policy \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --policy-name my-scaling-policy
```

Windows

```
aws application-autoscaling delete-scaling-policy --service-namespace namespace --
scalable-dimension dimension --resource-id identifier --policy-name my-scaling-policy
```

撤消可扩展目标的注册

使用以下 deregister-scalable-target 命令可取消注册可扩展目标。如果您有任何您创建的扩展策略或尚未删除的计划操作，这条命令会将它们删除。如果您要将此可扩展目标保留供将来使用，您可以跳过此操作。

Linux、macOS 或 Unix

```
aws application-autoscaling deregister-scalable-target \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier
```

Windows

```
aws application-autoscaling deregister-scalable-target --service-namespace namespace --
scalable-dimension dimension --resource-id identifier
```

暂停和恢复 Application Auto Scaling 扩缩

本主题说明如何暂停然后恢复应用程序中可扩展目标的一个或多个扩展活动。暂停-恢复功能用于临时暂停由您的扩展策略和计划操作触发的扩展活动。暂停-恢复功能非常有用，例如，当您更改或调查配置问题时，不希望自动扩展潜在产生干扰。您可以保留您的扩展策略和计划操作，在您准备就绪时，可以恢复扩展活动。

在随后的示例 CLI 命令中，您可在 config.json 文件中传递 JSON 格式的参数。您还可以通过使用引号将 JSON 数据结构括起来，在命令行上传递这些参数。有关更多信息，请参阅 AWS Command Line Interface 用户指南中的在 AWS CLI 中将引号和字符串结合使用。

内容

> **Note**
>
> 有关在 Amazon ECS 部署过程中暂停扩展流程的说明，请参阅以下文档：Amazon 弹性容器服务开发者指南中的服务自动扩展和部署

扩缩活动

Application Auto Scaling 支持将以下扩缩活动置于暂停状态：

- 由扩展策略触发的所有缩减活动。
- 由扩展策略触发的所有横向扩展活动。
- 涉及计划操作的所有扩展活动。

以下描述说明了暂停各个扩展活动时会发生什么。每个扩展活动都可以单独暂停和恢复。根据暂停扩展活动的原因，您可能需要一起暂停多个扩展活动。

`DynamicScalingInSuspended`

- 在触发目标跟踪扩缩策略或分步扩缩策略时，Application Auto Scaling 不会删除容量。这使您可以暂时禁用与扩展策略关联的缩减活动，而不删除扩展策略或其关联的 CloudWatch 警报。当您恢复横向缩减时，Application Auto Scaling 会评估具有当前违反的警报阈值的策略。

`DynamicScalingOutSuspended`

- 在触发目标跟踪扩缩策略或分步扩缩策略时，Application Auto Scaling 不会增加容量。这使您可以暂时禁用与扩展策略关联的横向扩展活动，而不删除扩展策略或其关联的 CloudWatch 警报。当您恢复横向扩展时，Application Auto Scaling 会评估具有当前违反的警报阈值的策略。

`ScheduledScalingSuspended`

- 在暂停期间，Application Auto Scaling 不启动计划要运行的扩缩操作。当您恢复计划的扩缩时，Application Auto Scaling 仅评估尚未经过执行时间的计划操作。

暂停和恢复扩展活动

您可以暂停和恢复 Application Auto Scaling 可扩展目标的单个或所有扩缩活动。

> **Note**
>
> 为简洁起见，这些示例说明了如何暂停和恢复 DynamoDB 表的扩缩。要指定不同的可扩展目标，请在 `--service-namespace` 中指定其命名空间，在 `--scalable-dimension` 中指定其可扩展维度，并在 `--resource-id` 中指定其资源 ID。有关每项服务的更多信息和示例，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务 中的主题。

暂停扩展活动

打开一个命令行窗口，然后使用 register-scalable-target 命令和 `--suspended-state` 选项，如下所示。

Linux、macOS 或 Unix

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table \
  --suspended-state file://config.json
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb --
scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table --
suspended-state file://config.json
```

如果成功，该命令会返回可扩展目标的 ARN。

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

要仅暂停某个扩展策略触发的缩减活动，请在 config.json 中指定以下内容。

```
{
    "DynamicScalingInSuspended":true
}
```

要仅暂停某个扩展策略触发的横向扩展活动，请在 config.json 中指定以下内容。

```
{
    "DynamicScalingOutSuspended":true
}
```

要仅暂停涉及计划操作的扩展活动，请在 config.json 中指定以下内容。

```
{
    "ScheduledScalingSuspended":true
}
```

暂停所有扩展活动

将 register-scalable-target 命令与 `--suspended-state` 选项一起使用，如下所示。

Linux、macOS 或 Unix

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table \
  --suspended-state file://config.json
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb --
scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table --
suspended-state file://config.json
```

此示例假定文件 config.json 包含以下 JSON 格式的参数。

```
{
    "DynamicScalingInSuspended":true,
    "DynamicScalingOutSuspended":true,
    "ScheduledScalingSuspended":true
}
```

如果成功，该命令会返回可扩展目标的 ARN。

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

查看暂停的扩缩活动

使用 describe-scalable-targets 命令可确定可扩展目标处于暂停状态的扩展活动。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scalable-targets --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table
```

Windows

```
aws application-autoscaling describe-scalable-targets --service-namespace dynamodb --
scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table
```

下面是示例输出。

```
{
    "ScalableTargets": [
        {
```

```
            "ServiceNamespace": "dynamodb",
            "ScalableDimension": "dynamodb:table:ReadCapacityUnits",
            "ResourceId": "table/my-table",
            "MinCapacity": 1,
            "MaxCapacity": 20,
            "SuspendedState": {
                "DynamicScalingOutSuspended": true,
                "DynamicScalingInSuspended": true,
                "ScheduledScalingSuspended": true
            },
            "CreationTime": 1558125758.957,
            "RoleARN": "arn:aws:iam::123456789012:role/aws-
service-role/dynamodb.application-autoscaling.amazonaws.com/
AWSServiceRoleForApplicationAutoScaling_DynamoDBTable"
        }
    ]
}
```

恢复扩缩活动

当您准备好恢复扩展活动时，可以使用 register-scalable-target 命令恢复它。

以下示例命令恢复指定的可扩展目标的所有扩展活动。

Linux、macOS 或 Unix

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb \
  --scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table \
  --suspended-state file://config.json
```

Windows

```
aws application-autoscaling register-scalable-target --service-namespace dynamodb --
scalable-dimension dynamodb:table:ReadCapacityUnits --resource-id table/my-table --
suspended-state file://config.json
```

此示例假定文件 config.json 包含以下 JSON 格式的参数。

```
{
    "DynamicScalingInSuspended":false,
    "DynamicScalingOutSuspended":false,
    "ScheduledScalingSuspended":false
```

```
}
```

如果成功，该命令会返回可扩展目标的 ARN。

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

Application Auto Scaling 的扩展活动

Application Auto Scaling 会监控您的扩展策略的 CloudWatch 指标，并在超过阈值时启动扩展活动。当您手动或按照计划修改可扩展目标的最大大小或最小大小时，它也会启动扩展活动。

进行扩展活动时，Application Auto Scaling 会执行以下操作之一：

- 增加可扩展目标的容量（称为横向扩展）
- 减少可扩展目标的容量（称为横向缩减）

您可以查看过去六周的扩展活动。

按可扩展目标查找扩展活动

要查看特定可扩展目标的扩展活动，请使用以下describe-scaling-activities命令。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scaling-activities --service-namespace ecs \
  --scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
```

Windows

```
aws application-autoscaling describe-scaling-activities --service-namespace ecs --
scalable-dimension ecs:service:DesiredCount --resource-id service/my-cluster/my-service
```

在以下示例响应中，StatusCode 包含活动的当前状态，StatusMessage 包含有关扩展活动状态的信息。

```
{
    "ScalingActivities": [
        {
            "ScalableDimension": "ecs:service:DesiredCount",
            "Description": "Setting desired count to 1.",
            "ResourceId": "service/my-cluster/my-service",
            "ActivityId": "e6c5f7d1-dbbb-4a3f-89b2-51f33e766399",
```

```
            "StartTime": 1462575838.171,
            "ServiceNamespace": "ecs",
            "EndTime": 1462575872.111,
            "Cause": "monitor alarm web-app-cpu-lt-25 in state ALARM triggered policy
 web-app-cpu-lt-25",
            "StatusMessage": "Successfully set desired count to 1. Change successfully
 fulfilled by ecs.",
            "StatusCode": "Successful"
        }
    ]
}
```

有关响应中字段的描述，请参阅ScalingActivity《App lication Auto Scaling API 参考》。

以下状态代码指示引发扩展活动的扩展事件何时达到完成状态：

- `Successful` – 扩展已成功完成
- `Overridden` – 所需容量已由较新的扩展事件进行更新
- `Unfulfilled` – 扩展超时或目标服务无法满足请求
- `Failed` – 扩展失败，出现异常

> **Note**
>
> 扩展活动的状态也可能为 `Pending` 或 `InProgress`。在目标服务响应之前，所有扩展活动都具有 `Pending` 状态。目标响应后，扩展活动的状态更改为 `InProgress`。

包括未扩展的活动

默认情况下，扩展活动不反映 Application Auto Scaling 决定是否不扩展的时间。

例如，假设 Amazon ECS 服务超出给定指标的最大阈值，但任务数已达到允许的最大任务数。在这种情况下，Application Auto Scaling 不会横向扩展所需的任务数。

要在响应中包括未缩放的活动（不是按比例缩放的活动），请在describe-scaling-activities命令中添加`--include-not-scaled-activities`选项。

Linux、macOS 或 Unix

```
aws application-autoscaling describe-scaling-activities --include-not-scaled-activities
 \
  --service-namespace ecs --scalable-dimension ecs:service:DesiredCount \
  --resource-id service/my-cluster/my-service
```

Windows

```
aws application-autoscaling describe-scaling-activities --include-not-scaled-activities
 --service-namespace ecs --scalable-dimension ecs:service:DesiredCount --resource-
id service/my-cluster/my-service
```

> **Note**
>
> 如果此命令引发错误，请确保已将 AWS CLI 本地版本更新到最新版本。

为了确认响应包含未扩展的活动，输出中会显示部分失败的扩展活动的 NotScaledReasons 元素（如果不是全部失败的话）。

```
{
    "ScalingActivities": [
        {
            "ScalableDimension": "ecs:service:DesiredCount",
            "Description": "Attempting to scale due to alarm triggered",
            "ResourceId": "service/my-cluster/my-service",
            "ActivityId": "4d759079-a31f-4d0c-8468-504c56e2eecf",
            "StartTime": 1664928867.915,
            "ServiceNamespace": "ecs",
            "Cause": "monitor alarm web-app-cpu-gt-75 in state ALARM triggered policy
 web-app-cpu-gt-75",
            "StatusCode": "Failed",
            "NotScaledReasons": [
                {
                    "Code": "AlreadyAtMaxCapacity",
                    "MaxCapacity": 4
                }
            ]
        }
    ]
}
```

有关响应中字段的描述，请参阅ScalingActivity《App lication Auto Scaling API 参考》。

如果返回未扩展的活动，根据 Code 中列出的原因代码，响应中可能会出现 CurrentCapacity、MaxCapacity 和 MinCapacity 等属性。

为防止出现大量重复条目，只有第一个未按比例缩放的活动才会记录在扩展活动历史记录中。除非不按比例缩放的原因发生变化，否则任何后续未按比例缩放的活动都不会生成新条目。

了解未扩展的原因代码

以下是未扩展的活动的原因代码。

原因代码	定义			
AutoScalingAnticipatedFlapping	Auto Scaling 算法决定不采取扩展操作，因为这会导致摆动。摆动是横向缩减和横向扩展的无限循环。也就是说，如果采取扩展操作，则指标值将更改以启动另一个相反方向的扩展操作。			
TargetServicePutResourceAsUnscalable	目标服务暂时将资源置于不可扩展状态。如果满足扩展策略中配置的自动扩展条件，Application Auto Scaling 将重试。			
AlreadyAtMaxCapacity	扩展被指定的最大容量阻止。如果您希望 Application			

原因代码	定义			
	Auto Scaling 进行横向扩展，则需要增加最大容量。			
AlreadyAtMinCapacity	扩展被指定的最小容量阻止。如果您希望 Application Auto Scaling 进行横向缩减，则需要减少最小容量。			
AlreadyAtDesiredCapacity	Auto Scaling 算法计算得出，修改后的容量符合当前容量。			

Application Auto Scaling 监控

监控是保持 Application Auto Scaling 和其他 AWS 解决方案的可靠性、可用性和性能的重要方面。您应该从 AWS 解决方案的各个部分收集监控数据，以便您可以更轻松地调试多点故障（如果发生）。AWS 提供监控工具以监控 Application Auto Scaling，当出现问题时报告并在适当时采取自动措施。

您可以使用以下功能来帮助您管理 AWS 资源：

AWS CloudTrail

使用 AWS CloudTrail，您可以跟踪由您或代表您的 AWS 账户 对 Application Auto Scaling API 发出的调用。CloudTrail 将信息存储在您指定的 Amazon S3 存储桶中的日志文件中。您可以标识调用 Application Auto Scaling 的具体用户和账户、发出调用的源 IP 地址以及调用的发生时间。有关更多信息，请参阅使用 AWS CloudTrail 记录 Application Auto Scaling API 调用。

> **Note**
>
> 有关可助您记录和收集工作负载相关数据的其他 AWS 服务的信息，请参阅《AWS 规范性指南》中的 面向应用程序所有者的日志记录和监控指南。

Amazon CloudWatch

Amazon CloudWatch 可帮助您分析日志并实时监控您的 AWS 资源和托管应用程序指标。您可以收集和跟踪指标，创建自定义的控制面板，以及设置警报以在指定的指标达到您指定的阈值时通知您或采取措施。例如，您可以通过 CloudWatch 跟踪资源利用率，并在利用率非常高或指标警报进入 `INSUFFICIENT_DATA` 状态时通知您。有关更多信息，请参阅使用 CloudWatch 监控资源。

CloudWatch 还可跟踪 Application Auto Scaling 的 AWS API 使用情况指标。您可以使用这些指标来配置警报，以在 API 调用量超过您定义的阈值时提醒您。有关更多信息，请参阅《Amazon CloudWatch 用户指南》中的 AWS 使用情况指标。

Amazon EventBridge

Amazon EventBridge 是一种无服务器事件总线服务，可以轻松地将应用程序与来自各种来源的数据相连接。EventBridge 可以从您自己的应用程序、软件即服务 (SaaS) 应用程序和AWS服务传输实时数据流，然后将该数据路由到诸如 Lambda 之类的目标。这让您可以监控服务中发生的事件，

并构建事件驱动型架构。有关更多信息，请参阅使用 Amazon EventBridge 监控 Application Auto Scaling 事件。

AWS Health Dashboard

AWS Health Dashboard（PHD）会显示相关信息，并提供因 AWS 资源的运行状况变化所触发的通知。信息会以两种方式显示：在显示按类别组织的最近和未来事件的控制面板上，以及在显示过去 90 天内所有事件的完整事件日志中。有关更多信息，请参阅Application Auto Scaling 的 AWS Health Dashboard 通知。

使用 AWS CloudTrail 记录 Application Auto Scaling API 调用

Application Auto Scaling 已与 AWS CloudTrail 服务集成，后者可记录用户、角色或使用 Application Auto Scaling API 的 AWS 服务所执行的操作。CloudTrail 可将所有对 Application Auto Scaling 的 API 调用作为事件捕获。所捕获的调用包含来自 AWS Management Console的调用，以及对 Application Auto Scaling API 的代码调用。如果您创建了跟踪，则可以将 CloudTrail 事件持续传送到 Amazon S3 存储桶，包括 Application Auto Scaling 事件。如果您不配置跟踪，则仍可在 CloudTrail 控制台中的事件历史记录中查看最新事件。使用由 CloudTrail 收集的信息，您可以确定向 Application Auto Scaling 发出的具体请求、发出请求的 IP 地址、请求的发出者、请求的发出时间以及其他详细信息。

要了解有关 CloudTrail 的更多信息，请参阅《AWS CloudTrail 用户指南》。

CloudTrail 中的 Application Auto Scaling 信息

在您创建 AWS 账户 时，将在该账户上启用 CloudTrail。发生 Application Auto Scaling 活动时，该活动将记录在 CloudTrail 事件中，并与其他 AWS 服务事件一起保存在事件历史记录中。您可以在 AWS 账户中查看、搜索和下载最新事件。有关更多信息，请参阅使用 CloudTrail 事件历史记录查看事件。

要持续记录 AWS 账户中的事件（包括 Application Auto Scaling 事件），请创建跟踪。通过跟踪记录，CloudTrail 可将日志文件传送至 Simple Storage Service（Amazon S3）存储桶。预设情况下，在控制台中创建跟踪记录时，此跟踪记录应用于所有AWS 区域。此跟踪记录在 AWS 分区中记录所有区域中的事件，并将日志文件传送至您指定的 Simple Storage Service（Amazon S3）桶。此外，您可以配置其他 Amazon Web Services 服务，进一步分析在 CloudTrail 日志中收集的事件数据并采取行动。有关更多信息，请参阅下列内容：

- 创建跟踪概览
- CloudTrail 支持的服务和集成
- 为 CloudTrail 配置 Amazon SNS 通知

- 从多个区域接收 CloudTrail 日志文件和从多个账户接收 CloudTrail 日志文件

所有 Application Auto Scaling 操作都由 CloudTrail 记录，相关文档请参阅 Application Auto Scaling API 参考。例如，对 `PutScalingPolicy`、`DeleteScalingPolicy` 和 `DescribeScalingPolicies` 操作的调用会在 CloudTrail 日志文件中生成条目。

每个事件或日志条目都包含有关生成请求的人员信息。身份信息可帮助您确定以下内容：

- 请求是使用根用户凭证还是 AWS Identity and Access Management（IAM）用户凭证发出。
- 请求是使用角色还是联合身份用户的临时安全凭证发出的。
- 请求是否由其它 AWS 服务发出。

有关更多信息，请参阅 CloudTrail userIdentity 元素。

了解 Application Auto Scaling 日志文件条目

跟踪是一种配置，可用于将事件作为日志文件传送到您指定的 Amazon S3 桶。CloudTrail 日志文件包含一个或多个日志条目。一个事件表示来自任何源的一个请求，包括有关所请求的操作、操作的日期和时间、请求参数等方面的信息。CloudTrail 日志文件不是公用 API 调用的有序堆栈跟踪，因此它们不会按任何特定顺序显示。

下面的示例显示了一个 CloudTrail 日志条目，该条目说明了 `DescribeScalableTargets` 操作。

```
{
    "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": "2018-08-16T23:20:32Z",
```

```
    "eventSource": "autoscaling.amazonaws.com",
    "eventName": "DescribeScalableTargets",
    "awsRegion": "us-west-2",
    "sourceIPAddress": "72.21.196.68",
    "userAgent": "EC2 Spot Console",
    "requestParameters": {
        "serviceNamespace": "ec2",
        "scalableDimension": "ec2:spot-fleet-request:TargetCapacity",
        "resourceIds": [
            "spot-fleet-request/sfr-05ceaf79-3ba2-405d-e87b-612857f1357a"
        ]
    },
    "responseElements": null,
    "additionalEventData": {
        "service": "application-autoscaling"
    },
    "requestID": "0737e2ea-fb2d-11e3-bfd8-99133058e7bb",
    "eventID": "3fcfb182-98f8-4744-bd45-b38835ab61cb",
    "eventType": "AwsApiCall",
    "recipientAccountId": "123456789012"
}
```

相关资源

借助 CloudWatch Logs，您可以监控和接收由 CloudTrail 捕获的特定事件的警报。发送到 CloudWatch Logs 的事件，是配置为由您的跟踪记录的事件，因此，请确保您已配置一个或多个跟踪来记录您需要监控的事件类型。CloudWatch Logs 可以监控日志文件中的信息，并在达到特定阈值时通知您。您还可以在高持久性存储中检索您的日志数据。有关更多信息，请参阅 Amazon CloudWatch Logs 用户指南 以及《AWS CloudTrail 用户指南》中的 使用 Amazon CloudWatch Logs 监控 CloudTrail 日志文件 主题。

使用 CloudWatch 监控资源

这一部分提供有关使用 CloudWatch 监控可扩展资源的指标的信息。

主题

- 使用 CloudWatch 构建控制面板
- 使用 CloudWatch 警报进行监控
- 使用 CloudWatch 监控资源使用情况

使用 CloudWatch 构建控制面板

您可以使用 Amazon CloudWatch（它将生成有关您的使用情况和性能的指标）监控应用程序使用资源的方式。CloudWatch 从您的 AWS 资源和您在 AWS 上运行的应用程序收集原始数据，并将其处理为可读的近实时指标。这些指标保留 15 个月，以便您可以访问历史信息，从而更好地了解应用程序的执行情况。有关更多信息，请参阅 Amazon CloudWatch 用户指南。

CloudWatch 控制面板是 CloudWatch 控制台中的可自定义主页，可用于在单个视图中监控资源，即便是分布到不同区域的资源，也能对其进行监控。您可以使用 CloudWatch 控制面板创建 AWS 资源的所选指标的自定义视图。您可以在每个图表上选择用于每个指标的颜色，以便更轻松地跨多个图表跟踪同一指标。

创建 CloudWatch 控制面板

1. 通过以下网址打开 CloudWatch 控制台：https://console.aws.amazon.com/cloudwatch/。
2. 在导航窗格中，选择控制面板，然后选择创建新的控制面板。
3. 输入控制面板的名称，例如要查看其 CloudWatch 数据的服务的名称。
4. 请选择创建控制面板。
5. 选择要添加到控制面板的小部件类型，例如折线图。然后选择配置，并选择要添加到控制面板的指标。有关更多信息，请参阅 Amazon CloudWatch 用户指南中的在 CloudWatch 控制面板中添加或删除图表。

默认情况下，您在 CloudWatch 控制面板中创建的指标为平均值。虽然 CloudWatch 允许您为每个指标选择任何统计数据，但并非所有的组合都有用。例如，CPU 利用率的平均、最小和最大统计数据均有用，但求和统计数据却无用。

衡量应用程序性能的常用方法是平均 CPU 利用率。如果 CPU 利用率增加，而您没有足够的容量来处理它，则应用程序可能无响应。另一方面，如果在利用率低时您有太多的容量和资源正在运行，这会增加使用该服务的成本。

根据服务的不同，您还拥有跟踪可用预配置吞吐量的指标。例如，对于在具有预置并发性的函数别名或版本上正在处理的调用数，Lambda 会发出 `ProvisionedConcurrencyUtilization` 指标。如果您正在启动大型作业并同时多次调用同一函数，则当作业超出可用的预配置并发性时，该作业可能会遇到延迟。另一方面，如果您的预配置并发性比您需要的多，则您的成本可能高于应有的成本。

在完全设置资源之前，不会显示指标。此外，如果某个指标在过去 14 天内未发布数据，在搜索要添加到 CloudWatch 控制面板上的图表的指标时，将找不到该指标。有关如何手动添加任何指标的信息，请参阅 Amazon CloudWatch 用户指南中的在 CloudWatch 控制面板上手动绘制指标。

有关更多信息，请参阅 使用 CloudWatch 监控资源使用情况 表格内可用的服务文档。

使用 CloudWatch 警报进行监控

您可以创建警报，以在 Amazon CloudWatch 检测到可能需要您注意的任何问题时通知您。

CloudWatch 警报会监控一个指标。该警报仅当状态发生变化并且已持续您指定的时间段时才会触发一个或多个操作。例如，您可以设置一个警报以在指标值低于或超过特定水平时通知您，从而确保在潜在问题出现之前您就得到通知。

CloudWatch 还允许您设置警报，当指标处于 `INSUFFICIENT_DATA` 状态时通知您。任何 AWS 服务的任何指标均可对 `INSUFFICIENT_DATA` 发出警报。这是新警报的初始状态，但如果 CloudWatch 指标变为不可用，或没有足够的数据可用于指标以确定警报状态时，警报状态也会变为 `INSUFFICIENT_DATA`。例如，仅当 Lambda 函数处于活动状态时，AWS Lambda 才会每分钟向 CloudWatch 发出一次 `ProvisionedConcurrencyUtilization` 指标。如果函数处于非活动状态，则会导致警报在等待指标时进入 `INSUFFICIENT_DATA` 状态。这是正常的，可能不一定意味着存在问题，但如果您预期在一段时间内进行活动，但没有任何活动，则可能表明存在问题。

本主题介绍如何创建警报，以在指标处于您定义的阈值范围之内或之外时或数据不足时发送通知。有关更多详细信息，请参阅 Amazon CloudWatch 用户指南中的使用 Amazon CloudWatch 警报。

创建发送电子邮件的警报

1. 通过以下网址打开 CloudWatch 控制台：https://console.aws.amazon.com/cloudwatch/。
2. 在导航窗格中，依次选择 Alarms 和 Create Alarm。
3. 选择 Select Metric（选择指标）。

   系统会将您导向到可在其中找到所有指标的页面。可用指标的类型取决于您使用的服务和功能。指标的分组首先依据服务命名空间，然后依据每个命名空间内的各种维度组合。
4. 选择一个指标命名空间（例如 Lambda），然后选择一个指标维度（例如 By Function Name [按函数名称]）。

   All metrics（所有指标）选项卡显示所选维度和命名空间的所有指标。
5. 选中您要为其创建警报的指标旁边的复选框，然后选择 Select metric（选择指标）。
6. 按如下所示配置警报，然后选择 Next（下一步）：

   - 在 Metric（指标）下，选择 `1 minute` 或 `5 minutes` 的汇总期。如果您使用一分钟作为某个指标的汇总期，则每分钟具有一个数据点。周期越短，创建的警报越敏感。
   - 在 Conditions（条件）下，配置您的阈值，例如，生成通知之前指标必须超过的值。

- 在 Additional configuration（其他配置）下，对于 Datapoints to alarm（触发警报的数据点数），输入指标值必须满足阈值条件才会触发警报的数据点（评估时间段）数。例如，2 个连续的 5 分钟时间段需要花 10 分钟才会触发警报。
- 对于 Missing data treatment（缺失数据处理），保留默认值并将缺失的数据点处理为缺失。

  某些指标仅在发生活动时报告。这可能会导致报告稀疏的指标。如果指标在设计上经常缺少数据点，则在这些期间警报的状态为 INSUFFICIENT_DATA。要强制警报保持之前的 ALARM 或 OK 状态以防止警报摆动，您可以选择忽略缺少的数据。

7. 在 Notification（通知）下，选择警报处于 ALARM、OK 或 INSUFFICIENT_DATA 状态时通知的 SNS 主题。要使告警为相同告警状态或不同告警状态发送多个通知，请选择 Add notification（添加通知）。
8. 在完成后，选择下一步。
9. 输入警报的名称和描述（可选），然后选择 Next (下一步)。
10. 选择Create alarm（创建警报）。

检查警报的状态

1. 通过以下网址打开 CloudWatch 控制台：https://console.aws.amazon.com/cloudwatch/。
2. 在导航窗格中，选择 Alarms（告警）以查看警报列表。
3. 要筛选警报，请使用搜索字段旁边的下拉筛选器，然后选择要应用的筛选选项。
4. 要编辑或删除警报，请选择警报，然后选择 Actions（操作）、Edit（编辑） 或 Actions（操作）、Delete（删除）。

使用 CloudWatch 监控资源使用情况

借助 Amazon CloudWatch，您可以更清楚地查看您在可扩展资源中的应用程序。CloudWatch 是一项针对 AWS 资源的监控服务。您可以使用 CloudWatch 收集和跟踪指标，设置警报，并自动应对您的 AWS 资源的变化。您还可以创建控制面板来监控所需的特定指标或指标集。

当您和与 Application Auto Scaling 集成的服务进行交互时，它们会将下表中显示的指标发送到 CloudWatch。在 CloudWatch 中，指标的分组首先依据服务命名空间，然后依据每个命名空间内的各种维度组合。这些指标可以帮助您监控资源使用量并计划应用程序的容量。如果您的应用程序的工作负载不稳定，则表明您应该考虑使用 Auto Scaling。有关这些指标的详细描述，请参阅相关指标的文档。

目录

- 用于监控资源使用量的 CloudWatch 指标
- 目标跟踪扩展策略的预定义目标

用于监控资源使用量的 CloudWatch 指标

下表列出了可用于支持监控资源使用量的 CloudWatch 指标。此列表并不详尽，但能为您提供一个好起点。如果您在 CloudWatch 控制台中未看到这些指标，请确保您已完成资源的设置。有关更多信息，请参阅 Amazon CloudWatch 用户指南。

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
AppStream 2.0			
队列	AWS/AppStream	名称：AvailableCapacity 维度：实例集	AppStream 2.0 指标
队列	AWS/AppStream	名称：CapacityUtilization 维度：实例集	AppStream 2.0 指标
Aurora			
副本	AWS/RDS	名称：CPUUtilization 维度：DBClusterIdenti	Aurora 集群级指标

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
		fier、Role（读取器）	
副本	AWS/RDS	名称：DatabaseConnections 维度：DBClusterIdentifier、Role（读取器）	Aurora 集群级指标
Amazon Comprehend			
文档分类端点	AWS/Comprehend	名称：InferenceUtilization 维度：EndpointArn	Amazon Comprehend 端点指标
实体识别程序端点	AWS/Comprehend	名称：InferenceUtilization 维度：EndpointArn	Amazon Comprehend 端点指标

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
DynamoDB			
表和全局二级索引	AWS/DynamoDB	名称：ProvisionedReadCapacityUnits 维度：TableName、GlobalSecondaryIndexName	DynamoDB 指标
表和全局二级索引	AWS/DynamoDB	名称：ProvisionedWriteCapacityUnits 维度：TableName、GlobalSecondaryIndexName	DynamoDB 指标

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
表和全局二级索引	AWS/DynamoDB	名称：ConsumedReadCapacityUnits 维度：TableName、GlobalSecondaryIndexName	DynamoDB 指标
表和全局二级索引	AWS/DynamoDB	名称：ConsumedWriteCapacityUnits 维度：TableName、GlobalSecondaryIndexName	DynamoDB 指标
Amazon ECS			
服务	AWS/ECS	名称：CPUUtilization 维度：ClusterName、ServiceName	Amazon ECS 指标

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
服务	AWS/ECS	名称：MemoryUtilization 维度：ClusterName、ServiceName	Amazon ECS 指标
服务	AWS/ApplicationELB	名称：RequestCountPerTarget 维度：TargetGroup	应用程序负载均衡器指标
ElastiCache			
集群（复制组）	AWS/ElastiCache	名称：DatabaseMemoryUsageCountedForEvictPercentage 维度：ReplicationGroupId	ElastiCache for Redis 指标

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
集群（复制组）	AWS/ElastiCache	名称：DatabaseCapacityUsageCountedForEvictPercentage 维度：ReplicationGroupId	ElastiCache for Redis 指标
集群（复制组）	AWS/ElastiCache	名称：EngineCPUUtilization 维度：ReplicationGroupId、角色（主要）	ElastiCache for Redis 指标

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
集群（复制组）	AWS/ElastiCache	名称：EngineCPUUtilization 维度：ReplicationGroupId、角色（副本）	ElastiCache for Redis 指标
Amazon EMR			
集群	AWS/ElasticMapReduce	名称：YARNMemoryAvailablePercentage 维度：ClusterId	Amazon EMR 指标
Amazon Keyspaces			
表	AWS/Cassandra	名称：ProvisionedReadCapacityUnits 维度：键空间、TableName	Amazon Keyspaces 指标

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
表	AWS/Cassandra	名称：ProvisionedWriteCapacityUnits 维度：键空间、TableName	Amazon Keyspaces 指标
表	AWS/Cassandra	名称：ConsumedReadCapacityUnits 维度：键空间、TableName	Amazon Keyspaces 指标
表	AWS/Cassandra	名称：ConsumedWriteCapacityUnits 维度：键空间、TableName	Amazon Keyspaces 指标
Lambda			

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
预配置并发	AWS/Lambda	名称：ProvisionedConcurrencyUtilization 维度：FunctionName、资源	Lambda 函数指标
Amazon MSK			
代理存储	AWS/Kafka	名称：KafkaDataLogsDiskUsed 维度：集群名称	Amazon MSK 指标
代理存储	AWS/Kafka	名称：KafkaDataLogsDiskUsed 维度：集群名称、代理 ID	Amazon MSK 指标
Neptune			

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
集群	AWS/Neptune	名称：CPUUtilization 维度：DBClusterIdentifier、Role（读取器）	Neptune 指标
SageMaker			
端点变体	AWS/SageMaker	名称：InvocationsPerInstance 维度：EndpointName、VariantName	调用指标
推理组件	AWS/SageMaker	名称：InvocationsPerCopy 维度：InferenceComponentName	调用指标

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
无服务器端点的预置并发	AWS/SageMaker	名称：ServerlessProvisionedConcurrencyUtilization 维度：EndpointName、VariantName	无服务器端点指标
Spot 实例集（Amazon EC2）			
Spot Fleets	AWS/EC2Spot	名称：CPUUtilization 维度：FleetRequestId	竞价型实例集指标
Spot Fleets	AWS/EC2Spot	名称：NetworkIn 维度：FleetRequestId	竞价型实例集指标

可扩展资源	命名空间	CloudWatch 指标	指向文档的链接
Spot Fleets	AWS/EC2Spot	名称：NetworkOut 维度：FleetRequestId	竞价型实例集指标
Spot Fleets	AWS/ApplicationELB	名称：RequestCountPerTarget 维度：TargetGroup	应用程序负载均衡器指标

目标跟踪扩展策略的预定义目标

下表列出了 Application Auto Scaling API 参考中的预定义指标类型及其相应的 CloudWatch 指标名称。每个预定义指标代表底层 CloudWatch 指标值的聚合。除非另有说明，否则结果是一分钟内基于百分比的平均资源使用量。预定义指标仅在设置目标跟踪扩展策略的情况下使用。

有关这些指标的更多信息，请参阅 用于监控资源使用量的 CloudWatch 指标 中的表格内可用的服务文档。

预定义指标类型	CloudWatch 指标名称
AppStream 2.0	
`AppStreamAverageCapacityUtilization`	CapacityUtilization
Aurora	

预定义指标类型	CloudWatch 指标名称
RDSReaderAverageCPUUtilization	CPU 利用率
RDSReaderAverageDatabaseConnections	DatabaseConnections[1]
Amazon Comprehend	
ComprehendInferenceUtilization	InferenceUtilization
DynamoDB	
DynamoDBReadCapacityUtilization	ProvisionedReadCapacityUnits、ConsumedReadCapacityUnits[2]
DynamoDBWriteCapacityUtilization	ProvisionedWriteCapacityUnits、ConsumedWriteCapacityUnits[2]
Amazon ECS	
ECSServiceAverageCPUUtilization	CPU 利用率
ECSServiceAverageMemoryUtilization	MemoryUtilization
ALBRequestCountPerTarget	RequestCountPerTarget[1]
ElastiCache	
ElastiCacheDatabaseMemoryUsageCountedForEvictPercentage	DatabaseMemoryUsageCountedForEvictPercentage
ElastiCacheDatabaseCapacityUsageCountedForEvictPercentage	DatabaseCapacityUsageCountedForEvictPercentage
ElastiCachePrimaryEngineCPUUtilization	EngineCPUUtilization
ElastiCacheReplicaEngineCPUUtilization	EngineCPUUtilization

预定义指标类型	CloudWatch 指标名称
Amazon Keyspaces	
`CassandraReadCapacityUtilization`	ProvisionedReadCapacityUnits、ConsumedReadCapacityUnits[2]
`CassandraWriteCapacityUtilization`	ProvisionedWriteCapacityUnits、ConsumedWriteCapacityUnits[2]
Lambda	
`LambdaProvisionedConcurrencyUtilization`	ProvisionedConcurrencyUtilization
Amazon MSK	
`KafkaBrokerStorageUtilization`	KafkaDataLogsDiskUsed
Neptune	
`NeptuneReaderAverageCPUUtilization`	CPU 利用率
SageMaker	
`SageMakerVariantInvocationsPerInstance`	InvocationsPerInstance[1]
`SageMakerInferenceComponentInvocationsPerCopy`	InvocationsPerCopy[1]
`SageMakerVariantProvisionedConcurrencyUtilization`	ServerlessProvisionedConcurrencyUtilization
竞价型实例集	
`EC2SpotFleetRequestAverageCPUUtilization`	CPUUtilization[3]

预定义指标类型	CloudWatch 指标名称
`EC2SpotFleetRequestAverageNetworkIn`[3]	NetworkIn[1] [3]
`EC2SpotFleetRequestAverageNetworkOut`[3]	NetworkOut[1] [3]
`ALBRequestCountPerTarget`	RequestCountPerTarget[1]

[1] 指标基于计数，而不是百分比。

[2] 对于 DynamoDB 和 Amazon Keyspaces，预定义指标是两个 CloudWatch 指标之和，以支持基于预置吞吐量消耗的扩展。

[3] 为了获得最佳扩展性能，应使用 Amazon EC2 详细监控。

使用 Amazon EventBridge 监控 Application Auto Scaling 事件

Amazon EventBridge（以前称为 CloudWatch Events）可帮助您监控 Application Auto Scaling 特定的事件，并启动将会使用其他 AWS 服务的目标操作。来自 AWS 服务服务的事件将近乎实时传输到 EventBridge。

借助 EventBridge，您可以创建用于匹配传入事件的规则并将事件路由到目标以进行处理。

有关更多信息，请参阅 Amazon EventBridge 用户指南中的 Amazon EventBridge 入门。

Application Auto Scaling 事件

以下是 Application Auto Scaling 的示例事件。事件会尽可能生成。

当前，只有特定扩展到最大值的事件和通过 CloudTrail 的 API 调用才可用于 Application Auto Scaling。

事件类型

- 状态更改的事件：扩展到最大容量
- 通过 CloudTrail 调用 API 的事件

状态更改的事件：扩展到最大容量

以下示例事件显示 Application Auto Scaling 将可扩展目标的容量增加（横向扩展）至其最大容量限制。如果需求再次增加，Application Auto Scaling 将无法进一步扩展目标，因为它已经扩展到其最大容量。

在 `detail` 对象中，`resourceId`、`serviceNamespace` 和 `scalableDimension` 属性的值用来标识可扩展目标。`newDesiredCapacity` 和 `oldDesiredCapacity` 属性的值用于指横向扩展事件完成后的新容量和横向扩展事件开始前的原始容量。`maxCapacity` 是可伸缩目标的最大容量限制。

```
{
  "version": "0",
  "id": "11112222-3333-4444-5555-666677778888",
  "detail-type": "Application Auto Scaling Scaling Activity State Change",
  "source": "aws.application-autoscaling",
  "account": "123456789012",
  "time": "2019-06-12T10:23:40Z",
  "region": "us-west-2",
  "resources": [],
  "detail": {
    "startTime": "2022-06-12T10:20:43Z",
    "endTime": "2022-06-12T10:23:40Z",
    "newDesiredCapacity": 8,
    "oldDesiredCapacity": 5,
    "minCapacity": 2,
    "maxCapacity": 8,
    "resourceId": "table/my-table",
    "scalableDimension": "dynamodb:table:WriteCapacityUnits",
    "serviceNamespace": "dynamodb",
    "statusCode": "Successful",
    "scaledToMax": true,
    "direction": "scale-out"
}
```

要创建捕获所有可扩展目标的所有 `scaledToMax` 状态更改事件的规则，请使用以下示例事件模式。

```
{
  "source": [
    "aws.application-autoscaling"
  ],
  "detail-type": [
    "Application Auto Scaling Scaling Activity State Change"
```

```
  ],
  "detail": {
    "scaledToMax": [
      true
    ]
  }
}
```

通过 CloudTrail 调用 API 的事件

跟踪是一种配置，让 AWS CloudTrail 能够将事件作为日志文件传输到某个 Amazon S3 存储桶。CloudTrail 日志文件包含若干日志条目。一个事件代表一个日志条目，包括有关所请求操作的信息、操作的日期和时间以及请求参数。要了解如何开始使用 CloudTrail，请参阅《AWS CloudTrail 用户指南》中的创建跟踪。

通过 CloudTrail 传输的所有事件都将 `AWS API Call via CloudTrail` 作为 `detail-type` 的值。

以下示例事件代表一个 CloudTrail 日志文件条目，该条目显示某个控制台用户调用了 Application Auto Scaling RegisterScalableTarget 操作。

```
{
  "version": "0",
  "id": "99998888-7777-6666-5555-444433332222",
  "detail-type": "AWS API Call via CloudTrail",
  "source": "aws.autoscaling",
  "account": "123456789012",
  "time": "2022-07-13T16:50:15Z",
  "region": "us-west-2",
  "resources": [],
  "detail": {
    "eventVersion": "1.08",
    "userIdentity": {
      "type": "IAMUser",
      "principalId": "123456789012",
      "arn": "arn:aws:iam::123456789012:user/Bob",
      "accountId": "123456789012",
      "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
      "sessionContext": {
        "sessionIssuer": {
          "type": "Role",
          "principalId": "123456789012",
```

```
          "arn": "arn:aws:iam::123456789012:role/Admin",
          "accountId": "123456789012",
          "userName": "Admin"
        },
        "webIdFederationData": {},
        "attributes": {
          "creationDate": "2022-07-13T15:17:08Z",
          "mfaAuthenticated": "false"
        }
      }
    },
    "eventTime": "2022-07-13T16:50:15Z",
    "eventSource": "autoscaling.amazonaws.com",
    "eventName": "RegisterScalableTarget",
    "awsRegion": "us-west-2",
    "sourceIPAddress": "AWS Internal",
    "userAgent": "EC2 Spot Console",
    "requestParameters": {
      "resourceId": "spot-fleet-request/sfr-73fbd2ce-aa30-494c-8788-1cee4EXAMPLE",
      "serviceNamespace": "ec2",
      "scalableDimension": "ec2:spot-fleet-request:TargetCapacity",
      "minCapacity": 2,
      "maxCapacity": 10
    },
    "responseElements": null,
    "additionalEventData": {
      "service": "application-autoscaling"
    },
    "requestID": "e9caf887-8d88-11e5-a331-3332aa445952",
    "eventID": "49d14f36-6450-44a5-a501-b0fdcdfaeb98",
    "readOnly": false,
    "eventType": "AwsApiCall",
    "managementEvent": true,
    "recipientAccountId": "123456789012",
    "eventCategory": "Management",
    "sessionCredentialFromConsole": "true"
  }
}
```


要创建基于所有可扩展目标的 DeleteScalingPolicy 和 DeregisterScalableTarget API 调用的规则，请使用以下示例事件模式：


```
{
```

```
  "source": [
    "aws.autoscaling"
  ],
  "detail-type": [
    "AWS API Call via CloudTrail"
  ],
  "detail": {
    "eventSource": [
      "autoscaling.amazonaws.com"
    ],
    "eventName": [
       "DeleteScalingPolicy",
       "DeregisterScalableTarget"
    ],
    "additionalEventData": {
      "service": [
        "application-autoscaling"
      ]
    }
  }
}
```

有关使用 CloudTrail 的更多信息，请参阅使用 AWS CloudTrail 记录 Application Auto Scaling API 调用。

Application Auto Scaling 的 AWS Health Dashboard 通知

为帮助您管理失败的扩缩事件，AWS Health Dashboard 为 Application Auto Scaling 发出的通知提供支持。当前只有特定于您的 DynamoDB 资源的横向扩展事件可用。

AWS Health Dashboard 是 AWS Health 服务的一部分。它不需要设置，您的账户中通过身份验证的任何用户都可以查看。有关更多信息，请参阅 AWS Health Dashboard 入门。

如果您的 DynamoDB 资源由于 DynamoDB 服务配额限制而未横向扩展，您将收到类似于以下内容的消息。如果您收到此消息，则应将其视为采取措施的警报。

```
  Hello,

  A scaling action has attempted to scale out your DynamoDB resources in the
  eu-west-1 region. This operation has been prevented because it would have exceeded a
  table-level write throughput limit (Provisioned mode). This limit restricts the
  provisioned write capacity of the table and all of its associated global secondary
```

```
  indexes. To address the issue, refer to the Amazon DynamoDB Developer Guide for
  current limits and how to request higher limits [1].

  To identify your DynamoDB resources that are impacted, use the
  describe-scaling-activities command or the DescribeScalingActivities operation [2]
[3].
  Look for a scaling activity with StatusCode "Failed" and a StatusMessage similar to
  "Failed to set write capacity units to 45000. Reason: The requested
 WriteCapacityUnits,
  45000, is above the per table maximum for the account in eu-west-1. Per table
 maximum:
  40000." You can also view these scaling activities from the Capacity tab of your
 tables
  in the AWS Management Console for DynamoDB.

  We strongly recommend that you address this issue to ensure that your tables
  are prepared to handle increases in traffic. This notification is sent only once in
  each 12 hour period, even if another failed scaling action occurs.

  [1] https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/
Limits.html#default-limits-throughput-capacity-modes
  [2] https://docs.aws.amazon.com/cli/latest/reference/application-autoscaling/
describe-scaling-activities.html
  [3] https://docs.aws.amazon.com/autoscaling/application/APIReference/
API_DescribeScalingActivities.html

  Sincerely,
  Amazon Web Services
```

Application Auto Scaling 的标签支持

您可以使用 AWS CLI 或 SDK 标记应用程序 Application Auto Scaling 可扩展目标。可扩展目标是表示 Application Auto Scaling 可以扩展的 AWS 或自定义资源的实体。

每个标签都包含游湖使用 Application Auto Scaling API 定义的键和值。标签可以帮助您根据组织的需求配置对特定可扩展目标的精细访问权限。有关更多信息，请参阅结合使用 ABAC 与 Application Auto Scaling。

您可以在注册新的可扩展目标时向目标添加标签，也可以将向现有的可扩展目标添加标签。

用于管理标签的常用命令包括：

- register-scalable-target，用于在注册新的可扩展目标时对其进行标记。
- tag-resource，用于向现有的可扩展目标添加标签。
- list-tags-for-resource，用于返回可扩展目标上的标签。
- untag-resource，用于删除标签。

标签示例

使用以下 register-scalable-target 命令和 `--tags` 选项，如下所示。该示例可用两个标签来标记可扩展目标：一个名称为 **`environment`**、标签值为 **`production`** 的标签键；一个名称为 **`iscontainerbased`**、标签值为 **`true`** 的标签键。

将 `--min-capacity` 和 `--max-capacity` 中的示例值以及 `--service-namespace` 中的示例文本替换为您与 Application Auto Scaling 搭配使用的 AWS 服务的命名空间，将 `--scalable-dimension` 替换为与您注册的资源关联的可扩展维度，并将 `--resource-id` 替换为资源的标识符。有关每项服务的更多信息和示例，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务 中的主题。

```
aws application-autoscaling register-scalable-target \
  --service-namespace namespace \
  --scalable-dimension dimension \
  --resource-id identifier \
  --min-capacity 1 --max-capacity 10 \
  --tags environment=production,iscontainerbased=true
```

如果成功，该命令会返回可扩展目标的 ARN。

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

> **Note**
>
> 如果此命令引发错误，请确保您已在本地将 AWS CLI 更新到最新版本。

安全性标签

使用标签验证请求者（例如 IAM 用户或角色）是否有权限执行某些操作。使用下面的一个或多个条件键，在 IAM policy 的条件元素中提供标签信息：

- 使用 `aws:ResourceTag/tag-key: tag-value` 可允许（或拒绝）带特定标签的可扩展目标上的用户操作。
- 使用 `aws:RequestTag/tag-key: tag-value` 要求在请求中存在（或不存在）特定标签。
- 使用 `aws:TagKeys [tag-key, ...]` 要求在请求中存在（或不存在）特定标签键。

例如，以下 IAM policy 授予执行 DeregisterScalableTarget、DeleteScalingPolicy 和 DeleteScheduledAction 操作的权限。但如果对其执行操作的可扩展目标组具有标签 **environment=production**，在此策略也会拒绝操作。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "application-autoscaling:DeregisterScalableTarget",
                "application-autoscaling:DeleteScalingPolicy",
                "application-autoscaling:DeleteScheduledAction"
            ],
            "Resource": "*"
            }
        },
        {
            "Effect": "Deny",
```

```
            "Action": [
                "application-autoscaling:DeregisterScalableTarget",
                "application-autoscaling:DeleteScalingPolicy",
                "application-autoscaling:DeleteScheduledAction"
            ],
            "Resource": "*",
            "Condition": {
                "StringEquals": {"aws:ResourceTag/environment": "production"}
            }
        }
    ]
}
```

控制对标签的访问

使用标签来验证请求者（例如 IAM 用户或角色）是否有权限添加、修改或删除可扩展目标的标签。

例如，您可以创建一个 IAM policy，以仅允许从可扩展目标中删除具有 **temporary** 键的标签。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "application-autoscaling:UntagResource",
            "Resource": "*",
            "Condition": {
                "ForAllValues:StringEquals": { "aws:TagKeys": ["temporary"] }
            }
        }
    ]
}
```

Application Auto Scaling 中的安全性

云安全 AWS 是重中之重。作为 AWS 客户，您可以受益于专为满足大多数安全敏感型组织的要求而构建的数据中心和网络架构。

安全是双方共同承担 AWS 的责任。责任共担模式将其描述为云的安全性和云中的安全性：

- 云安全 — AWS 负责保护在 AWS 云中运行 AWS 服务的基础架构。 AWS 还为您提供可以安全使用的服务。作为AWS 合规计划合规计划合规计划合的一部分，第三方审计师定期测试和验证我们安全的有效性。要了解适用于 Appl AWS ication Auto Scaling 的合规计划，请按合规计划查看。
- 云端安全-您的责任由您使用的 AWS 服务决定。您还需要对其它因素负责，包括您的数据的敏感性、您公司的要求以及适用的法律法规。

该文档帮助您了解如何在使用 Application Auto Scaling 时应用责任共担模式。以下主题说明如何配置 Application Auto Scaling 以实现您的安全性和合规性目标。您还将学习如何使用其他 AWS 服务来帮助您监控和保护您的 Application Auto Scaling 资源。

主题

- Application Auto Scaling 和接口 VPC 终端节点
- Application Auto Scaling 和数据保护
- 适用于 Application Auto Scaling 的 Identity and Access Management
- Application Auto Scaling 的合规性验证
- Application Auto Scaling 中的恢复功能
- Application Auto Scaling 中的基础设施安全性

Application Auto Scaling 和接口 VPC 终端节点

您可以通过将 Application Auto Scaling 配置为使用接口 VPC 端点，以改善 VPC 的安保状况。接口终端节点由一项技术提供支持 AWS PrivateLink，通过将您的 VPC 和应用程序 Auto Scaling 之间的所有网络流量限制在网络上，使您能够私下访问应用程序 Auto Scaling API。 AWS 借助接口终端节点，您也不需要 Internet 网关、NAT 设备或虚拟专用网关。

您无需进行配置 AWS PrivateLink，但建议您这样做。有关 AWS PrivateLink 和 VPC 终端节点的更多信息，请参阅什么是 AWS PrivateLink？ 在AWS PrivateLink 指南中。

主题

创建接口 VPC 终端节点

使用以下服务名称为 Application Auto Scaling 创建端点：

```
com.amazonaws.region.application-autoscaling
```

有关更多信息，请参阅AWS PrivateLink 指南中的使用接口 VPC 终端节点访问 AWS 服务。

您不需要更改任何其他设置。Application Auto Scaling 使用 AWS 服务终端节点或私有接口 VPC 终端节点调用其他服务，以使用中者为准。

创建 VPC 端点策略

您可以向 VPC 终端节点附加策略来控制对 Application Auto Scaling API 的访问。该策略指定：

- 可执行操作的主体。
- 可执行的操作。
- 可对其执行操作的资源。

以下示例显示了一个 VPC 终端节点策略，该策略拒绝所有人通过终端节点删除扩展策略的权限。示例策略还授予所有人执行所有其他操作的权限。

```
{
   "Statement": [
        {
            "Action": "*",
            "Effect": "Allow",
            "Resource": "*",
            "Principal": "*"
        },
        {
            "Action": "application-autoscaling:DeleteScalingPolicy",
            "Effect": "Deny",
            "Resource": "*",
            "Principal": "*"
        }
```

```
    ]
}
```

有关更多信息，请参阅《AWS PrivateLink 指南》中的 VPC 终端节点策略。

Application Auto Scaling 和数据保护

分担责任模型 AWS 分担责任模型适用于 Application Auto Scaling 中的数据保护。如本模型所述 AWS ，负责保护运行所有内容的全球基础架构 AWS Cloud。您负责维护对托管在此基础设施上的内容的控制。您还负责您所使用的 AWS 服务 的安全配置和管理任务。有关数据隐私的更多信息，请参阅数据隐私常见问题。有关欧洲数据保护的信息，请参阅 AWS 安全性博客 上的 AWS 责任共担模式和 GDPR 博客文章。

出于数据保护目的，我们建议您保护 AWS 账户 凭证并使用 AWS IAM Identity Center 或 AWS Identity and Access Management (IAM) 设置个人用户。这样，每个用户只获得履行其工作职责所需的权限。我们还建议您通过以下方式保护数据：

- 对每个账户使用 multi-factor authentication（MFA）。
- 使用 SSL/TLS 与资源通信。 AWS 我们要求使用 TLS 1.2，建议使用 TLS 1.3。
- 使用设置 API 和用户活动日志 AWS CloudTrail。
- 使用 AWS 加密解决方案以及其中的所有默认安全控件 AWS 服务。
- 使用高级托管安全服务（例如 Amazon Macie），它有助于发现和保护存储在 Amazon S3 中的敏感数据。
- 如果您在 AWS 通过命令行界面或 API 进行访问时需要经过 FIPS 140-2 验证的加密模块，请使用 FIPS 端点。有关可用的 FIPS 端点的更多信息，请参阅《美国联邦信息处理标准（FIPS）第 140-2 版》。

我们强烈建议您切勿将机密信息或敏感信息（如您客户的电子邮件地址）放入标签或自由格式文本字段（如名称字段）。这包括使用控制台、API 或软件开发工具包 AWS 服务 使用 Application Auto AWS Scaling 或其他软件开发工具包的情况。 AWS CLI在用于名称的标签或自由格式文本字段中输入的任何数据都可能会用于计费或诊断日志。如果您向外部服务器提供网址，强烈建议您不要在网址中包含凭证信息来验证对该服务器的请求。

适用于 Application Auto Scaling 的 Identity and Access Management

AWS Identity and Access Management (IAM) AWS 服务 可帮助管理员安全地控制对 AWS 资源的访问权限。IAM 管理员控制谁可以通过身份验证（登录）和授权（具有权限）使用 Application Auto Scaling 资源。您可以使用 IAM AWS 服务 ，无需支付额外费用。

要使用 Application Auto Scaling，你需要一个 AWS 账户 和用于登录账户的安全证书。有关更多信息，请参阅 进行设置以开始使用 Application Auto Scaling。

有关完整的 IAM 文档，请参阅 IAM 用户指南。

访问控制

您可以使用有效的凭证来对自己的请求进行身份验证，但您还必须拥有权限才能创建或访问 Application Auto Scaling 资源。例如，您必须拥有相应的权限才能执行创建扩展策略、配置计划扩展等操作。

以下各节详细介绍 IAM 管理员如何使用 IAM 通过控制谁可以执行 Application Auto Scaling API 操作来帮助保护您的 AWS 资源。

主题

Application Auto Scaling 如何与 IAM 一起使用

> **Note**
>
> 2017 年 12 月，对 Application Auto Scaling 进行了更新，同时为 Application Auto Scaling 集成服务启用了多个服务相关角色。需要特定的 IAM 权限和 Application Auto Scaling 服务相关角色（或用于 Amazon EMR 弹性伸缩的服务角色），以便用户可以配置扩缩。

在使用 IAM 管理对 Application Auto Scaling 的访问权限之前，您需要了解哪些 IAM 功能可用于 Application Auto Scaling。

可以与 Application Auto Scaling 结合使用的 IAM 功能

IAM 功能	Application Auto Scaling 支持
基于身份的策略	是
策略操作	是
策略资源	支持
策略条件键（特定于服务）	是
基于资源的策略	否
ACL	否
ABAC（策略中的标签）	部分
临时凭证	支持
服务角色	支持
服务相关角色	支持

要全面了解 Application Auto Scaling 和其他功能如何 AWS 服务 与大多数 IAM 功能配合使用 AWS 服务 ，请在 IAM 用户指南中查看如何与 IAM 配合使用。

Application Auto Scaling 基于身份的策略

支持基于身份的策略	是

基于身份的策略是可附加到身份（如 IAM 用户、用户组或角色）的 JSON 权限策略文档。这些策略控制用户和角色可在何种条件下对哪些资源执行哪些操作。要了解如何创建基于身份的策略，请参阅 IAM 用户指南中的创建 IAM policy。

通过使用 IAM 基于身份的策略，您可以指定允许或拒绝的操作和资源以及允许或拒绝操作的条件。您无法在基于身份的策略中指定主体，因为它适用于其附加的用户或角色。要了解可在 JSON 策略中使用的所有元素，请参阅《IAM 用户指南》中的 IAM JSON 策略元素引用。

Application Auto Scaling 基于身份的策略示例

要查看 Application Auto Scaling 基于身份的策略的示例，请参阅 Application Auto Scaling 基于身份的策略示例。

操作

支持策略操作	支持

在 IAM policy 语句中，您可以从支持 IAM 的任何服务中指定任何 API 操作。对于 Application Auto Scaling，请使用以下前缀为 API 操作命名：`application-autoscaling:`。例如：`application-autoscaling:RegisterScalableTarget`、`application-autoscaling:PutScalingPolicy` 和 `application-autoscaling:DeregisterScalableTarget`。

要在单个语句中指定多项操作，请使用逗号将它们隔开，如下例所示。

```
"Action": [
      "application-autoscaling:DescribeScalingPolicies",
      "application-autoscaling:DescribeScalingActivities"
```

您也可以使用通配符 (*) 指定多个操作。例如，要指定以单词 `Describe` 开头的所有操作，请包括以下操作。

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

有关应用程序 Auto Scaling 操作的列表，请参阅《服务授权参考》中的 App AWS lication Auto Scaling 定义的操作。

资源

支持策略资源	支持

在 IAM policy 声明中，`Resource` 元素指定了该声明涵盖的一个或多个对象。对于 Application Auto Scaling，每个 IAM policy 语句都适用于您使用可扩展目标的 Amazon 资源名称（ARN）指定的可扩展目标。

可扩展目标的 ARN 资源格式：

```
arn:aws:application-autoscaling:region:account-id:scalable-target/unique-identifier
```

例如，您可以在语句中使用 ARN 指示特定的可扩展目标，如下所示。唯一 ID（1234abcd56ab78cd901ef1234567890ab123）是 Application Auto Scaling 分配给可扩展目标的值。

```
"Resource": "arn:aws:application-autoscaling:us-east-1:123456789012:scalable-
target/1234abcd56ab78cd901ef1234567890ab123"
```

您可以使用通配符（*）指定属于特定账户的所有实例，如下所示。

```
"Resource": "arn:aws:application-autoscaling:us-east-1:123456789012:scalable-target/*"
```

要指定所有资源，或者如果特定 API 操作不支持 ARN，请在 Resource 元素中使用 * 通配符，如下所示。

```
"Resource": "*"
```

有关更多信息，请参阅《服务授权参考》中的 App AWS lication Auto Scaling 定义的资源类型。

条件键

支持特定于服务的策略条件键	支持

您可以在控制 Application Auto Scaling 资源访问的 IAM policy 中指定条件。仅当条件为 True 时，策略语句才有效。

Application Auto Scaling 支持以下服务定义条件键，您可以在基于身份的策略中使用这些条件键来确定谁可以执行 Application Auto Scaling API 操作。

- `application-autoscaling:scalable-dimension`
- `application-autoscaling:service-namespace`

要了解可以将条件键与哪些应用程序 Auto Scaling API 操作一起使用，请参阅《服务授权参考》中的 App AWS lication Auto Scaling 定义的操作。有关使用应用程序 Auto Scaling 条件键的更多信息，请参阅AWS 应用程序 Auto Scaling 的条件键。

要查看对所有服务都可用的全局条件键，请参阅《IAM 用户指南》中的 AWS 全局条件上下文键。

基于资源的策略

支持基于资源的策略	不支持

其他 AWS 服务，例如 Amazon 简单存储服务，支持基于资源的权限策略。例如，您可以将权限策略挂载到 S3 存储桶以管理对该存储桶的访问权限。

Application Auto Scaling 不支持基于资源的策略。

访问控制列表 (ACL)

支持 ACL	不支持

Application Auto Scaling 不支持访问控制列表 (ACL)。

结合使用 ABAC 与 Application Auto Scaling

支持 ABAC（策略中的标签）	部分

基于属性的访问权限控制（ABAC）是一种授权策略，该策略基于属性来定义权限。在中 AWS，这些属性称为标签。您可以将标签附加到 IAM 实体（用户或角色）和许多 AWS 资源。标记实体和资源是 ABAC 的第一步。然后设计 ABAC 策略，以在主体的标签与他们尝试访问的资源标签匹配时允许操作。

ABAC 在快速增长的环境中非常有用，并在策略管理变得繁琐的情况下可以提供帮助。

要基于标签控制访问，您需要使用 `aws:ResourceTag/key-name`、`aws:RequestTag/key-name` 或 `aws:TagKeys` 条件键在策略的条件元素中提供标签信息。

ABAC 可用于支持标签的资源，但并非所有资源都支持标签。计划操作和扩缩策略不支持标签，但可扩展目标支持标签。有关更多信息，请参阅 Application Auto Scaling 的标签支持。

有关 ABAC 的更多信息,请参阅《IAM 用户指南》中的什么是 ABAC？。要查看设置 ABAC 步骤的教程，请参阅《IAM 用户指南》中的使用基于属性的访问权限控制（ABAC）。

将临时凭证与 Application Auto Scaling 一起使用

支持临时凭证	支持

当你使用临时证书登录时，有些 AWS 服务 不起作用。有关更多信息，包括哪些 AWS 服务 适用于临时证书，请参阅 IAM 用户指南中的AWS 服务 与 IAM 配合使用的信息。

如果您使用除用户名和密码之外的任何方法登录，则 AWS Management Console 使用的是临时证书。例如，当您 AWS 使用公司的单点登录 (SSO) 链接进行访问时，该过程会自动创建临时证书。当您以用户身份登录控制台，然后切换角色时，您还会自动创建临时凭证。有关切换角色的更多信息，请参阅《IAM 用户指南》中的 切换到角色（控制台）。

您可以使用 AWS CLI 或 AWS API 手动创建临时证书。然后，您可以使用这些临时证书进行访问 AWS。 AWS 建议您动态生成临时证书，而不是使用长期访问密钥。有关更多信息，请参阅 IAM 中的临时安全凭证。

服务角色

支持服务角色	支持

如果您的 Amazon EMR 集群使用弹性伸缩，则此功能允许 Application Auto Scaling 代表您担任服务角色。与服务相关角色类似，服务角色允许此服务访问其他服务中的资源以代表您完成操作。服务角色显示在 IAM 账户中，并归该账户所有。这意味着，IAM 管理员可以更改该角色的权限。但是，这样做可能会中断服务的功能。

Application Auto Scaling 仅支持 Amazon EMR 的服务角色。有关 EMR 服务角色的文档，请参阅 Amazon EMR Management Guide 中的 Using automatic scaling with a custom policy for instance groups。

> **Note**
>
> 引入服务相关角色之后，不再需要旧式服务角色，例如，适用于 Amazon ECS 和竞价型实例集的服务角色。

服务相关角色

支持服务相关角色	支持

服务相关角色是一种与服务相关联的 AWS 服务服务角色。服务可以代入代表您执行操作的角色。服务相关角色出现在您的中 AWS 账户 ，并且归服务所有。IAM 管理员可以查看但不能编辑服务相关角色的权限。

有关 Application Auto Scaling 服务相关角色的信息，请参阅 Application Auto Scaling 的服务相关角色。

AWS Application Auto Scaling 的托管策略

AWS 托管策略是由创建和管理的独立策略 AWS。 AWS 托管策略旨在为许多常见用例提供权限，以便您可以开始为用户、组和角色分配权限。

请记住， AWS 托管策略可能不会为您的特定用例授予最低权限权限，因为它们可供所有 AWS 客户使用。我们建议通过定义特定于您的使用场景的客户管理型策略来进一步减少权限。

您无法更改 AWS 托管策略中定义的权限。如果 AWS 更新 AWS 托管策略中定义的权限，则更新会影响该策略所关联的所有委托人身份（用户、组和角色）。 AWS 最有可能在启动新的 API 或现有服务可以使用新 AWS 服务 的 API 操作时更新 AWS 托管策略。

有关更多信息，请参阅《IAM 用户指南》中的 AWS 托管式策略。

内容

- AWS 托管策略授予对 SageMaker 和的访问权限 CloudWatch
- AWS 授予对 EC2 Spot 队列访问权限的托管策略以及 CloudWatch
- AWS 托管策略授予对您的自定义资源的访问权限以及 CloudWatch
- Application Auto Scaling 更新 AWS 了托管策略

AWS 托管策略授予对 AppStream 2.0 的访问权限和 CloudWatch

策略名称：AWSApplicationAutoscalingAppStreamFleetPolicy

您无法将 AWSApplicationAutoscalingAppStreamFleetPolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 调用亚马逊 AppStream CloudWatch 并代表您执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_AppStreamFleet 服务相关角色权限策略允许 Application Auto Scaling 对所有相关资源 ("Resource": "*") 完成以下操作：

- 操作：appstream:DescribeFleets
- 操作：appstream:UpdateFleet
- 操作：cloudwatch:DescribeAlarms
- 操作：cloudwatch:PutMetricAlarm
- 操作：cloudwatch:DeleteAlarms

AWS 授予对 Aurora 的访问权限的托管策略和 CloudWatch

策略名称：AWSApplicationAutoscalingRDSClusterPolicy

您无法将 AWSApplicationAutoscalingRDSClusterPolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Auto Scaling 调用 Aurora CloudWatch 并代表您执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_RDSCluster 服务相关角色权限策略允许 Application Auto Scaling 对所有相关资源 ("Resource": "*") 完成以下操作：

- 操作：rds:AddTagsToResource

- 操作：rds:CreateDBInstance
- 操作：rds:DeleteDBInstance
- 操作：rds:DescribeDBClusters
- 操作：rds:DescribeDBInstance
- 操作：cloudwatch:DescribeAlarms
- 操作：cloudwatch:PutMetricAlarm
- 操作：cloudwatch:DeleteAlarms

AWS 授予访问亚马逊 Comprehend 权限的托管策略和 CloudWatch

策略名称：AWSApplicationAutoscalingComprehendEndpointPolicy

您无法将 AWSApplicationAutoscalingComprehendEndpointPolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 调用 Amazon CloudWatch Comprehend 并代表您执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_ComprehendEndpoint 服务相关角色权限策略允许 Application Auto Scaling 对所有相关资源 ("Resource": "*") 完成以下操作：

- 操作：comprehend:UpdateEndpoint
- 操作：comprehend:DescribeEndpoint
- 操作：cloudwatch:DescribeAlarms
- 操作：cloudwatch:PutMetricAlarm
- 操作：cloudwatch:DeleteAlarms

AWS 授予对 DynamoDB 的访问权限的托管策略以及 CloudWatch

策略名称：AWSApplicationAutoscalingDynamoDBTablePolicy

您无法将 AWSApplicationAutoscalingDynamoDBTablePolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 代表您调用 DynamoDB CloudWatch 并执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_DynamoDBTable 服务相关角色权限策略允许 Application Auto Scaling 对所有相关资源 ("Resource": "*") 完成以下操作：

- 操作：dynamodb:DescribeTable
- 操作：dynamodb:UpdateTable
- 操作：cloudwatch:DescribeAlarms
- 操作：cloudwatch:PutMetricAlarm
- 操作：cloudwatch:DeleteAlarms

AWS 托管策略授予对 Amazon ECS 的访问权限和 CloudWatch

策略名称：AWSApplicationAutoscalingECSServicePolicy

您无法将 AWSApplicationAutoscalingECSServicePolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 调用 Amazon ECS CloudWatch 并代表您执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_ECSService 服务相关角色权限策略允许 Application Auto Scaling 对所有相关资源 ("Resource": "*") 完成以下操作：

- 操作：ecs:DescribeServices
- 操作：ecs:UpdateService
- 操作：cloudwatch:DescribeAlarms
- 操作：cloudwatch:PutMetricAlarm
- 操作：cloudwatch:DeleteAlarms

AWS 托管策略授予对 ElastiCache 和的访问权限 CloudWatch

策略名称：AWSApplicationAutoscalingElastiCacheRGPolicy

您无法将 AWSApplicationAutoscalingElastiCacheRGPolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 代表您调用 ElastiCache CloudWatch 和执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_ElastiCacheRG 服务相关角色权限策略允许 Application Auto Scaling 对指定的资源完成以下操作：

- 操作：所有资源上的 elasticache:DescribeReplicationGroups
- 操作：所有资源上的 elasticache:ModifyReplicationGroupShardConfiguration
- 操作：所有资源上的 elasticache:IncreaseReplicaCount
- 操作：所有资源上的 elasticache:DecreaseReplicaCount
- 操作：所有资源上的 elasticache:DescribeCacheClusters
- 操作：所有资源上的 elasticache:DescribeCacheParameters
- 操作：所有资源上的 cloudwatch:DescribeAlarms
- 操作：资源 arn:*:cloudwatch:*:*:alarm:TargetTracking* 上的 cloudwatch:PutMetricAlarm
- 操作：资源 arn:*:cloudwatch:*:*:alarm:TargetTracking* 上的 cloudwatch:DeleteAlarms
- 操作：cloudwatch:DeleteAlarms

AWS 托管策略授予对 Amazon Keyspaces 的访问权限和 CloudWatch

策略名称：AWSApplicationAutoscalingCassandraTablePolicy

您无法将 AWSApplicationAutoscalingCassandraTablePolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 调用 Amazon Keyspaces CloudWatch 并代表您执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_CassandraTable 服务相关角色权限策略允许 Application Auto Scaling 对指定的资源完成以下操作：

- 操作：资源 arn:*:cassandra:*:*:/keyspace/system/table/* 上的 cassandra:Select
- 操作：资源 arn:*:cassandra:*:*:/keyspace/system_schema/table/* 上的 cassandra:Select
- 操作：资源 arn:*:cassandra:*:*:/keyspace/system_schema_mcs/table/* 上的 cassandra:Select
- 操作：资源 arn:*:cassandra:*:*:"*" 上的 cassandra:Alter
- 操作：cloudwatch:DescribeAlarms

- 操作：cloudwatch:PutMetricAlarm
- 操作：cloudwatch:DeleteAlarms

AWS 授予对 Lambda 的访问权限的托管策略和 CloudWatch

策略名称：AWSApplicationAutoscalingLambdaConcurrencyPolicy

您无法将 AWSApplicationAutoscalingLambdaConcurrencyPolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Auto Scaling 代表您调用 Lambda CloudWatch 并执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_LambdaConcurrency 服务相关角色权限策略允许 Application Auto Scaling 对所有相关资源 ("Resource": "*") 完成以下操作：

- 操作：lambda:PutProvisionedConcurrencyConfig
- 操作：lambda:GetProvisionedConcurrencyConfig
- 操作：lambda:DeleteProvisionedConcurrencyConfig
- 操作：cloudwatch:DescribeAlarms
- 操作：cloudwatch:PutMetricAlarm
- 操作：cloudwatch:DeleteAlarms

AWS 托管策略授予对 Amazon MSK 的访问权限和 CloudWatch

策略名称：AWSApplicationAutoscalingKafkaClusterPolicy

您无法将 AWSApplicationAutoscalingKafkaClusterPolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 调用 Amazon MSK CloudWatch 并代表您执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_KafkaCluster 服务相关角色权限策略允许 Application Auto Scaling 对所有相关资源 ("Resource": "*") 完成以下操作：

- 操作：kafka:DescribeCluster
- 操作：kafka:DescribeClusterOperation

- 操作：kafka:UpdateBrokerStorage
- 操作：cloudwatch:DescribeAlarms
- 操作：cloudwatch:PutMetricAlarm
- 操作：cloudwatch:DeleteAlarms

AWS 授予对 Neptune 的访问权限的托管策略和 CloudWatch

策略名称：AWSApplicationAutoscalingNeptuneClusterPolicy

您无法将 AWSApplicationAutoscalingNeptuneClusterPolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 调用 Neptune CloudWatch 并代表您执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_NeptuneCluster 服务相关角色权限策略允许 Application Auto Scaling 对指定的资源完成以下操作：

- Amazon Neptune 数据库引擎 ("Condition":{"StringEquals": {"rds:DatabaseEngine":"neptune"}) 中在带有前缀 autoscaled-reader 的资源上的操作：rds:AddTagsToResource
- 操作：所有资源上的 rds:ListTagsForResource
- Amazon Neptune 数据库引擎 ("Condition":{"StringEquals": {"rds:DatabaseEngine":"neptune"}) 中在所有数据库集群 ("Resource":"arn:*:rds:*:*:db:autoscaled-reader*", "arn:aws:rds:*:*:cluster:*") 中带有前缀 autoscaled-reader 的资源上的操作：rds:CreateDBInstance
- 操作：所有资源上的 rds:DescribeDBInstances
- 操作：所有资源上的 rds:DescribeDBClusters
- 操作：所有资源上的 rds:DescribeDBClusterParameters
- 操作：资源 arn:*:rds:*:*:db:autoscaled-reader* 上的 rds:DeleteDBInstance
- 操作：所有资源上的 cloudwatch:DescribeAlarms
- 操作：资源 arn:*:cloudwatch:*:*:alarm:TargetTracking* 上的 cloudwatch:PutMetricAlarm
- 操作：资源 arn:*:cloudwatch:*:*:alarm:TargetTracking* 上的 cloudwatch:DeleteAlarms

- 操作：cloudwatch:DeleteAlarms

AWS 托管策略授予对 SageMaker 和的访问权限 CloudWatch

策略名称：AWSApplicationAutoscalingSageMakerEndpointPolicy

您无法将 AWSApplicationAutoscalingSageMakerEndpointPolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 代表您调用 SageMaker CloudWatch 和执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_SageMakerEndpoint 服务相关角色权限策略允许 Application Auto Scaling 对指定的资源完成以下操作：

- 操作：所有资源上的 sagemaker:DescribeEndpoint
- 操作：所有资源上的 sagemaker:DescribeEndpointConfig
- 操作：所有资源上的 sagemaker:DescribeInferenceComponent
- 操作：所有资源上的 sagemaker:UpdateEndpointWeightsAndCapacities
- 操作：所有资源上的 sagemaker:UpdateInferenceComponentRuntimeConfig
- 操作：所有资源上的 cloudwatch:DescribeAlarms
- 操作：资源 arn:*:cloudwatch:*:*:alarm:TargetTracking* 上的 cloudwatch:PutMetricAlarm
- 操作：资源 arn:*:cloudwatch:*:*:alarm:TargetTracking* 上的 cloudwatch:DeleteAlarms

AWS 授予对 EC2 Spot 队列访问权限的托管策略以及 CloudWatch

策略名称：AWSApplicationAutoscalingEC2SpotFleetRequestPolicy

您无法将 AWSApplicationAutoscalingEC2SpotFleetRequestPolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 调用 Amazon EC2 CloudWatch 并代表您执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest 服务相关角色权限策略允许 Application Auto Scaling 对所有相关资源 ("Resource": "*") 完成以下操作：

- 操作：ec2:DescribeSpotFleetRequests
- 操作：ec2:ModifySpotFleetRequest
- 操作：cloudwatch:DescribeAlarms
- 操作：cloudwatch:PutMetricAlarm
- 操作：cloudwatch:DeleteAlarms

AWS 托管策略授予对您的自定义资源的访问权限以及 CloudWatch

策略名称：AWSApplicationAutoScalingCustomResourcePolicy

您无法将 AWSApplicationAutoScalingCustomResourcePolicy 附加到您的 IAM 身份（用户或角色）。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 调用通过 API Gateway 提供的自定义资源 CloudWatch并代表您执行扩展。

权限详细信息

AWSServiceRoleForApplicationAutoScaling_CustomResource 服务相关角色权限策略允许 Application Auto Scaling 对所有相关资源 ("Resource": "*") 完成以下操作：

- 操作：execute-api:Invoke
- 操作：cloudwatch:DescribeAlarms
- 操作：cloudwatch:PutMetricAlarm
- 操作：cloudwatch:DeleteAlarms

Application Auto Scaling 更新 AWS 了托管策略

查看自该服务开始跟踪这些更改以来 Application Auto Scaling AWS 托管策略更新的详细信息。有关此页面更改的自动提醒，请订阅 Application Auto Scaling Document history（文档历史记录）页面上的 RSS 源。

更改	描述	日期
Application Auto Scaling 为其 SageMaker 服务相关角色添加权限	现在，此策略向服务授予调用 SageMakerDescribeInferenceComponent 和 UpdateInferenceCom	2023 年 11 月 13 日

更改	描述	日期
	`ponentRuntimeConfig` API 操作的权限，以支持为即将到来的集成自动缩放 SageMaker 资源提供兼容性。现在，该策略还将 CloudWatch `PutMetricAlarm` 和 `DeleteAlarms` API 操作限制为与目标跟踪扩展策略一起使用的 CloudWatch 警报。	
Application Auto Scaling 添加 Neptune 策略	Application Auto Scaling 为 Neptune 添加了一个新的托管式策略。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 调用 Neptune CloudWatch 并代表您执行扩展。	2021 年 10 月 6 日
Application Auto Scaling ElastiCache 为 Redis 策略添加了	Application Auto Scaling 为添加了一个新的托管策略 ElastiCache。此策略附加到服务相关角色，该角色允许 Application Auto Scaling 代表您调用 ElastiCache CloudWatch 和执行扩展。	2021 年 8 月 19 日
Application Auto Scaling 已开启跟踪更改	Application Auto Scaling 开始跟踪其 AWS 托管策略的更改。	2021 年 8 月 19 日

Application Auto Scaling 的服务相关角色

Application Auto Scaling 使用服务相关角色来获得代表您调用其他 AWS 服务所需的权限。服务相关角色是一种独特的 AWS Identity and Access Management (IAM) 角色，直接链接到 AWS 服务。服务相关角色提供了一种向服务委派权限的安全方式， AWS 因为只有关联的服务才能担任服务相关角色。

内容

概述

对于与 Application Auto Scaling 集成的服务，Application Auto Scaling 将为您创建服务相关角色。每个服务都有一个服务相关角色。每个服务相关角色信任指定的服务委托人来代入该角色。有关更多信息，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务。

Application Auto Scaling 包含每个服务相关角色的所有必要权限。这些托管式权限由 Application Auto Scaling 创建和管理，它们定义每种资源类型允许的操作。有关每个角色授予的权限的详细信息，请参阅 AWS Application Auto Scaling 的托管策略。

下面的部分介绍如何创建和管理 Application Auto Scaling 服务相关角色。首先配置权限以允许 IAM 实体（如用户、组或角色）创建、编辑或删除服务相关角色。

创建服务相关角色所需的权限

Application Auto Scaling 需要权限才能在您中的任何用户首次 AWS 账户 调用`RegisterScalableTarget`给定服务时创建服务相关角色。如果服务相关角色不存在，Application Auto Scaling 会为您账户中的目标服务创建该角色。此服务相关角色向 Application Auto Scaling 授予权限，以便它能代表您调用目标服务。

为使自动角色创建成功，用户必须具有 `iam:CreateServiceLinkedRole` 操作的权限。

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

以下是一个基于身份的策略，该策略授予为竞价型实例集创建服务相关角色的权限。您可以在策略的 `Resource` 字段中将服务相关角色指定为 ARN，并将服务相关角色的服务委托人指定为条件，如下所示。有关每种服务的 ARN，请参阅 服务相关角色 ARN 参考。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "iam:CreateServiceLinkedRole",
            "Resource": "arn:aws:iam::*:role/aws-service-role/ec2.application-
autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest",
            "Condition": {
                "StringLike": {
                    "iam:AWSServiceName":"ec2.application-autoscaling.amazonaws.com"
                }
            }
        }
    ]
}
```

> **Note**
>
> iam:AWSServiceName IAM 条件键将指定角色附加到的服务委托人，在本示例策略中指示为 *ec2.application-autoscaling*.amazonaws.com。不要尝试猜测服务委托人。要查看服务的服务委托人，请参阅 AWS 可以与 Application Auto Scaling 一起使用的服务。

创建服务相关角色（自动）

您无需手动创建服务相关角色。Application Auto Scaling 将在您调用 RegisterScalableTarget 时为您创建相应的服务相关角色。例如，如果您已为 Amazon ECS 服务设置弹性伸缩，则 Application Auto Scaling 会创建 AWSServiceRoleForApplicationAutoScaling_ECSService 角色。

创建服务相关角色（手动）

要创建服务相关角色，您可以使用 IAM 控制台或 IAM API。 AWS CLI有关更多信息，请参阅 IAM 用户指南中的创建服务相关角色。

创建服务相关角色 (AWS CLI)

使用以下 create-service-linked-roleCLI 命令创建 Application Auto Scaling 服务相关角色。在请求中，指定服务名称“前缀”。

要查找服务名称前缀，请参阅关于 AWS 可以与 Application Auto Scaling 一起使用的服务 部分中每个服务的服务相关角色的服务委托人的信息。服务名称和服务委托人共享相同的前缀。例如，要创建 AWS Lambda 服务相关角色，请使用`lambda.application-autoscaling.amazonaws.com`。

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

编辑服务相关角色

对于 Application Auto Scaling 创建的服务相关角色，您只能编辑其描述。有关更多信息，请参阅《IAM 用户指南》中的编辑服务相关角色。

删除服务相关角色

如果您不再将 Application Auto Scaling 用于支持的服务，我们建议您删除相应的服务相关角色。

只有在先删除相关 AWS 资源后，才能删除服务相关角色。这可以防止您无意中撤销 Application Auto Scaling 对您的资源的权限。有关更多信息，请参阅有关可扩展资源的文档。例如，要删除 Amazon ECS 服务，请参阅 Amazon Elastic Container Service 开发者指南中的删除服务。

您可以使用 IAM 删除服务相关角色。有关更多信息，请参阅《IAM 用户指南》中的删除服务相关角色。

在删除某个服务相关角色后，当您调用 `RegisterScalableTarget` 时，Application Auto Scaling 将重新创建该角色。

Application Auto Scaling 服务相关角色支持的区域

Application Auto Scaling 支持在提供服务的所有 AWS 区域中使用服务相关角色。

服务相关角色 ARN 参考

服务	ARN
AppStream 2.0	`arn:aws:iam:: 012345678910 :role/aws-service-role/appstream.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_AppStreamFleet`

服务	ARN
Aurora	arn:aws:iam:: *012345678910* :role/aws-service-role/rds.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_RDSCluster
Comprehend	arn:aws:iam:: *012345678910* :role/aws-service-role/comprehend.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_ComprehendEndpoint
DynamoDB	arn:aws:iam:: *012345678910* :role/aws-service-role/dynamodb.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_DynamoDBTable
ECS	arn:aws:iam:: *012345678910* :role/aws-service-role/ecs.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_ECSService
ElastiCache	arn:aws:iam:: *012345678910* :role/aws-service-role/elasticache.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_ElastiCacheRG
Keyspaces	arn:aws:iam:: *012345678910* :role/aws-service-role/cassandra.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_CassandraTable
Lambda	arn:aws:iam:: *012345678910* :role/aws-service-role/lambda.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_LambdaConcurrency
MSK	arn:aws:iam:: *012345678910* :role/aws-service-role/kafka.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_KafkaCluster

服务	ARN
Neptune	arn:aws:iam:: *012345678910* :role/aws-service-role/neptune.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_NeptuneCluster
SageMaker	arn:aws:iam:: *012345678910* :role/aws-service-role/sagemaker.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_SageMakerEndpoint
Spot Fleets	arn:aws:iam:: *012345678910* :role/aws-service-role/ec2.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest
自定义资源	arn:aws:iam:: *012345678910* :role/aws-service-role/custom-resource.application-autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_CustomResource

> **Note**
>
> 即使指定的服务相关角色尚不存在，您也可以在 AWS CloudFormation 堆栈模板中为AWS::ApplicationAutoScaling::ScalableTarget资源的RoleARN属性指定服务相关角色的 ARN。Application Auto Scaling 将自动为您创建该角色。

Application Auto Scaling 基于身份的策略示例

默认情况下，您中的全新用户 AWS 账户 无权执行任何操作。IAM 管理员必须创建并分配 IAM policy，以便为 IAM 身份（例如用户或角色）授予执行 Application Auto Scaling API 操作的权限。

要了解如何使用以下示例 JSON 策略文档创建 IAM policy，请参阅《IAM 用户指南》中的在 JSON 选项卡上创建策略。

内容

- Application Auto Scaling API 操作所需的权限
- 对目标服务进行 API 操作所需的权限以及 CloudWatch
- 在中工作的权限 AWS Management Console

Application Auto Scaling API 操作所需的权限

以下策略为调用 Application Auto Scaling API 时的常见使用案例授予权限。编写基于身份的策略时，请参阅本节。每个策略授予执行全部或部分 Application Auto Scaling API 操作的权限。您还需要确保最终用户拥有目标服务的权限，以及 CloudWatch （有关详细信息，请参阅下一节）。

以下基于身份的策略授予执行全部 Application Auto Scaling API 操作的权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "application-autoscaling:*"
            ],
            "Resource": "*"
        }
    ]
}
```

以下基于身份的策略授予执行配置扩展策略而非计划操作所需的全部 Application Auto Scaling API 操作的权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "application-autoscaling:RegisterScalableTarget",
              "application-autoscaling:DescribeScalableTargets",
              "application-autoscaling:DeregisterScalableTarget",
              "application-autoscaling:PutScalingPolicy",
              "application-autoscaling:DescribeScalingPolicies",
              "application-autoscaling:DescribeScalingActivities",
              "application-autoscaling:DeleteScalingPolicy"
```

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

以下基于身份的策略授予执行配置计划操作而非扩展策略所需的全部 Application Auto Scaling API 操作的权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "application-autoscaling:RegisterScalableTarget",
              "application-autoscaling:DescribeScalableTargets",
              "application-autoscaling:DeregisterScalableTarget",
              "application-autoscaling:PutScheduledAction",
              "application-autoscaling:DescribeScheduledActions",
              "application-autoscaling:DescribeScalingActivities",
              "application-autoscaling:DeleteScheduledAction"
            ],
            "Resource": "*"
        }
    ]
}
```

对目标服务进行 API 操作所需的权限以及 CloudWatch

要成功配置并将 Application Auto Scaling 与目标服务一起使用，必须向最终用户授予访问亚马逊 CloudWatch 以及他们将为其配置扩展的每项目标服务的权限。使用以下策略授予使用目标服务和所需的最低权限 CloudWatch。

内容

- AppStream 2.0 支舰队
- Aurora 副本
- Amazon Comprehend 文档分类和实体识别程序终端节点
- DynamoDB 表和全局二级索引
- ECS 服务

- ElastiCache 复制组
- Amazon EMR 集群
- Amazon Keyspaces 表
- Lambda 函数
- Amazon Managed Streaming for Apache Kafka (MSK) 代理存储
- Neptune 集群
- SageMaker 端点
- Spot 实例集（Amazon EC2）
- 自定义资源

AppStream 2.0 支舰队

以下基于身份的策略授予所需的所有 AppStream 2.0 和 CloudWatch API 操作的权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "appstream:DescribeFleets",
              "appstream:UpdateFleet",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Aurora 副本

以下基于身份的策略授予对所有 Aurora 和 CloudWatch API 所需操作的权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
```

```
        {
            "Effect": "Allow",
            "Action": [
              "rds:AddTagsToResource",
              "rds:CreateDBInstance",
              "rds:DeleteDBInstance",
              "rds:DescribeDBClusters",
              "rds:DescribeDBInstances",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Amazon Comprehend 文档分类和实体识别程序终端节点

以下基于身份的策略向所有必需的 Amazon Com CloudWatch prehend 和 API 操作授予权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "comprehend:UpdateEndpoint",
              "comprehend:DescribeEndpoint",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

DynamoDB 表和全局二级索引

以下基于身份的策略向所有必需的 DynamoDB 和 API 操作授予权限。 CloudWatch

```
{
```

```
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "dynamodb:DescribeTable",
              "dynamodb:UpdateTable",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

ECS 服务

以下基于身份的策略向所有必需的 ECS 和 CloudWatch API 操作授予权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "ecs:DescribeServices",
              "ecs:UpdateService",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

ElastiCache 复制组

以下基于身份的策略向所有 ElastiCache 必需的 CloudWatch API 操作授予权限。

```
{
    "Version": "2012-10-17",
```

```
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "elasticache:ModifyReplicationGroupShardConfiguration",
              "elasticache:IncreaseReplicaCount",
              "elasticache:DecreaseReplicaCount",
              "elasticache:DescribeReplicationGroups",
              "elasticache:DescribeCacheClusters",
              "elasticache:DescribeCacheParameters",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Amazon EMR 集群

以下基于身份的策略向所有必需的 Amazon EMR 和 CloudWatch API 操作授予权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "elasticmapreduce:ModifyInstanceGroups",
              "elasticmapreduce:ListInstanceGroups",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Amazon Keyspaces 表

以下基于身份的策略向所有 Amazon Keyspaces 和 CloudWatch API 操作授予权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "cassandra:Select",
              "cassandra:Alter",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Lambda 函数

以下基于身份的策略向所有必需的 Lambda 和 CloudWatch API 操作授予权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "lambda:PutProvisionedConcurrencyConfig",
              "lambda:GetProvisionedConcurrencyConfig",
              "lambda:DeleteProvisionedConcurrencyConfig",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Amazon Managed Streaming for Apache Kafka (MSK) 代理存储

以下基于身份的策略向所有必需的 Amazon MSK 和 CloudWatch API 操作授予权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "kafka:DescribeCluster",
              "kafka:DescribeClusterOperation",
              "kafka:UpdateBrokerStorage",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Neptune 集群

以下基于身份的策略向所有必需的 Neptune 和 CloudWatch API 操作授予权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "rds:AddTagsToResource",
              "rds:CreateDBInstance",
              "rds:DescribeDBInstances",
              "rds:DescribeDBClusters",
              "rds:DescribeDBClusterParameters",
              "rds:DeleteDBInstance",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

SageMaker 端点

以下基于身份的策略向所有 SageMaker 必需的 CloudWatch API 操作授予权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "sagemaker:DescribeEndpoint",
              "sagemaker:DescribeEndpointConfig",
              "sagemaker:DescribeInferenceComponent",
              "sagemaker:UpdateEndpointWeightsAndCapacities",
              "sagemaker:UpdateInferenceComponentRuntimeConfig",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

Spot 实例集（Amazon EC2）

以下基于身份的策略向所有必需的 Spot 队列和 CloudWatch API 操作授予权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "ec2:DescribeSpotFleetRequests",
              "ec2:ModifySpotFleetRequest",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
```

```
}
```

自定义资源

以下基于身份的策略授予执行 API Gateway API 操作的权限。该策略还授予 CloudWatch 执行所有必需操作的权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
              "execute-api:Invoke",
              "cloudwatch:DescribeAlarms",
              "cloudwatch:PutMetricAlarm",
              "cloudwatch:DeleteAlarms"
            ],
            "Resource": "*"
        }
    ]
}
```

在中工作的权限 AWS Management Console

没有独立的 Application Auto Scaling 控制台。与 Application Auto Scaling 集成的大多数服务都具有专用于帮助您通过控制台配置扩缩的功能。

在大多数情况下，每项服务都提供 AWS 托管（预定义）IAM 策略，这些策略定义了对其控制台的访问权限，其中包括对 Application Auto Scaling API 操作的权限。有关详细信息，请参阅要使用其控制台的服务的文档。

您还可以创建自己的自定义 IAM policy，为用户授予在 AWS Management Console中查看和处理特定 Application Auto Scaling API 操作的精细权限。您可以使用前面部分中的示例策略；但是，它们是为使用 AWS CLI 或 SDK 发出的请求而设计的。控制台使用其他 API 操作实现其功能，因此这些策略可能不会按预期方式起作用。例如，要配置分步缩放，用户可能需要额外的权限才能创建和管理 CloudWatch 警报。

> **Tip**
>
> 为帮助您了解在控制台中执行任务所需的相应 API 操作，您可以使用 AWS CloudTrail等服务。有关更多信息，请参阅《AWS CloudTrail 用户指南》。

以下基于身份的策略授予为竞价型实例集配置扩展策略的权限。除了竞价型实例集的 IAM 权限之外，从 Amazon EC2 控制台访问实例集扩展设置的控制台用户必须具有使用支持动态扩展的服务的适当权限。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "application-autoscaling:*",
                "ec2:DescribeSpotFleetRequests",
                "ec2:ModifySpotFleetRequest",
                "cloudwatch:DeleteAlarms",
                "cloudwatch:DescribeAlarmHistory",
                "cloudwatch:DescribeAlarms",
                "cloudwatch:DescribeAlarmsForMetric",
                "cloudwatch:GetMetricStatistics",
                "cloudwatch:ListMetrics",
                "cloudwatch:PutMetricAlarm",
                "cloudwatch:DisableAlarmActions",
                "cloudwatch:EnableAlarmActions",
                "sns:CreateTopic",
                "sns:Subscribe",
                "sns:Get*",
                "sns:List*"
            ],
            "Resource": "*"
        },
        {
            "Effect": "Allow",
            "Action": "iam:CreateServiceLinkedRole",
            "Resource": "arn:aws:iam::*:role/aws-service-role/ec2.application-
autoscaling.amazonaws.com/AWSServiceRoleForApplicationAutoScaling_EC2SpotFleetRequest",
            "Condition": {
                "StringLike": {
```

```
                    "iam:AWSServiceName":"ec2.application-autoscaling.amazonaws.com"
                }
            }
        }
    ]
}
```

该策略允许控制台用户在 Amazon EC2 控制台中查看和修改扩展策略，并在控制 CloudWatch 台中创建和管理 CloudWatch 警报。

您可以调整 API 操作以限制用户访问权限。例如，将 `application-autoscaling:Describe*` 替换为 `application-autoscaling:*` 意味着用户具有只读访问权限。

您也可以根据需要调整 CloudWatch 权限，以限制用户对 CloudWatch功能的访问权限。有关更多信息，请参阅 Amazon CloudWatch 用户指南中的使用 CloudWatch 控制台所需的权限。

Application Auto Scaling 访问故障排除

如果使用 Application Auto Scaling 时遇到 `AccessDeniedException` 或类似的困难，请参阅本节中的信息。

我无权在 Application Auto Scaling 中执行操作

如果您`AccessDeniedException`在调用 AWS API 操作时收到，则表示您正在使用的 AWS Identity and Access Management (IAM) 证书没有进行该调用所需的权限。

如果 `mateojackson` 用户尝试查看有关可扩展目标的详细信息，但没有 `application-autoscaling:DescribeScalableTargets` 权限，则会出现以下示例错误。

```
An error occurred (AccessDeniedException) when calling the DescribeScalableTargets
 operation: User: arn:aws:iam::123456789012:user/mateojackson is not authorized to
 perform: application-autoscaling:DescribeScalableTargets
```

如果您收到此错误或类似错误，则必须联系您的管理员寻求帮助。

您的账户管理员需要确保您拥有访问所有 API 操作的权限，Application Auto Scaling 使用这些操作来访问目标服务中的资源和 CloudWatch。根据您使用的资源，需要不同的权限。用户初次配置指定资源的扩缩时，Application Auto Scaling 还需要创建服务相关角色的权限。

我是管理员，我的 IAM policy 返回错误或未按预期工作

除了 Application Auto Scaling 操作外，您的 IAM 策略还必须授予调用目标服务的权限和 CloudWatch。如果用户或应用程序没有这些额外的权限，其访问可能会被意外拒绝。要为账户中的用户和应用程序编写 IAM policy，请参阅 Application Auto Scaling 基于身份的策略示例 中的信息。

有关如何执行验证的信息，请参阅 对目标资源进行 API 调用的权限验证。

请注意，某些权限问题也可能是由于创建 Application Auto Scaling 所使用的服务相关角色时出现问题所致。有关创建这些服务相关角色的信息，请参阅 Application Auto Scaling 的服务相关角色。

对目标资源进行 API 调用的权限验证

向 Application Auto Scaling API 操作发出授权请求要求 API 调用者必须具有访问目标服务中和中的 AWS 资源的权限 CloudWatch。在继续处理请求 CloudWatch 之前，Application Auto Scaling 会验证与目标服务关联的请求的权限。为此，我们将发出一系列调用来验证目标资源的 IAM 权限。返回响应时，Application Auto Scaling 会读取该响应。如果 IAM 权限不允许指定的操作，则 Application Auto Scaling 将使请求失败，并将错误返回给用户，其中包含有关缺少权限的信息。这可确保用户想要部署的扩缩配置按预期工作，并且在请求失败时返回有用的错误。

作为其工作原理的示例，以下信息提供了有关应用程序 Auto Scaling 如何使用 Aurora 和 CloudWatch 执行权限验证的详细信息。

当用户针对 Aurora 数据库集群调用 `RegisterScalableTarget` API 时，Application Auto Scaling 会执行以下所有检查以验证用户是否具有所需的权限（以粗体显示）。

- rds:CreateDBInstance：为确定用户是否具有此权限，我们将向 `CreateDBInstance` API 操作发送请求，尝试在用户指定的 Aurora 数据库集群中创建具有无效参数（空实例 ID）的数据库实例。对于授权用户，该 API 将在审计请求后返回 `InvalidParameterValue` 错误代码响应。但是，对于未经授权的用户，我们会收到 `AccessDenied` 错误，使 Application Auto Scaling 请求失败并显示 `ValidationException` 错误，向用户列出缺少的权限。
- rds:DeleteDBInstance：我们将向 `DeleteDBInstance` API 操作发出一个空实例 ID。对于授权用户，此请求会导致 `InvalidParameterValue` 错误。对于未经授权的用户，它会导致 `AccessDenied` 并向用户发送验证异常（与第一个要点中描述的处理相同）。
- rdsAddTagsToResource:: 由于 `AddTagsToResource` API 操作需要亚马逊资源名称 (ARN)，因此必须使用无效的账户 ID (12345) 和虚拟实例 ID () 指定 “虚拟” 资源non-existing-db来构建 ARN ()。`arn:aws:rds:us-east-1:12345:db:non-existing-db`对于授权用户，此请求会导致 `InvalidParameterValue` 错误。对于未经授权的用户，它会导致 `AccessDenied` 并向用户发送验证异常。

- rds:DescribeDBCluster：我们描述为弹性伸缩注册的资源的集群名称。对于授权用户，我们将得到一个有效的描述结果。对于未经授权的用户，它会导致 AccessDenied 并向用户发送验证异常。
- rds:DescribeDBInstance。我们使用 db-cluster-id 筛选条件调用 DescribeDBInstance API，筛选用户提供的集群名称以注册可扩展目标。对于授权用户，我们可以描述数据库集群中的所有数据库实例。对于未经授权的用户，此调用会导致 AccessDenied 并向用户发送验证异常。
- cloudwatchPutMetricAlarm:: 我们在调用 PutMetricAlarm API 时不带任何参数。由于缺少警报名称，对于授权用户，请求会导致 ValidationError。对于未经授权的用户，它会导致 AccessDenied 并向用户发送验证异常。
- cloudwatchDescribeAlarms:: 我们在调用 DescribeAlarms API 时将最大记录数值设置为 1。对于授权用户，我们预期响应中有一个警报的信息。对于未经授权的用户，此调用会导致 AccessDenied 并向用户发送验证异常。
- cloudwatchDeleteAlarms:: 与PutMetricAlarm上述类似，我们不提供任何参数可供DeleteAlarms请求。由于请求中缺少警报名称，对于授权用户，此调用将失败并显示 ValidationError。对于未经授权的用户，它会导致 AccessDenied 并向用户发送验证异常。

只要发生任何一个验证异常，它就会被记录下来。您可以使用采取措施手动识别哪些呼叫未通过验证 AWS CloudTrail。有关更多信息，请参阅《AWS CloudTrail 用户指南》。

> **Note**
>
> 如果您使用收到有关应用程序 Auto Scaling 事件的警报 CloudTrail，则默认情况下，这些警报将包括用于验证用户权限的应用程序 Auto Scaling 调用。要过滤掉这些提示，请使用 invokedBy 字段，它们包含用于这些验证检查的 application-autoscaling.amazonaws.com。

Application Auto Scaling 的合规性验证

要了解是否属于特定合规计划的范围，请参阅AWS 服务 “按合规计划划分的范围” ”，然后选择您感兴趣的合规计划。 AWS 服务 有关一般信息，请参阅AWS 合规计划AWS。

您可以使用下载第三方审计报告 AWS Artifact。有关更多信息，请参阅中的 “下载报告” 中的 “ AWS Artifact。

您在使用 AWS 服务 时的合规责任取决于您的数据的敏感性、贵公司的合规目标以及适用的法律和法规。 AWS 提供了以下资源来帮助实现合规性：

- 安全与合规性快速入门指南 — 这些部署指南讨论了架构注意事项，并提供了部署以安全性和合规性为重点 AWS 的基准环境的步骤。
- 在 A@@ mazon Web Services 上构建 HIPAA 安全与合规性 — 本白皮书描述了各公司如何使用 AWS 来创建符合 HIPAA 资格的应用程序。

> **Note**
>
> 并非所有 AWS 服务 人都符合 HIPAA 资格。有关更多信息，请参阅符合 HIPAA 要求的服务参考。

- AWS 合AWS 规资源 — 此工作簿和指南集可能适用于您的行业和所在地区。
- AWS 客户合规指南 — 从合规角度了解责任共担模式。这些指南总结了保护的最佳实践， AWS 服务并将指南映射到跨多个框架（包括美国国家标准与技术研究院 (NIST)、支付卡行业安全标准委员会 (PCI) 和国际标准化组织 (ISO)）的安全控制。
- 使用AWS Config 开发人员指南中的规则评估资源 — 该 AWS Config 服务评估您的资源配置在多大程度上符合内部实践、行业准则和法规。
- AWS Security Hub— 这 AWS 服务 可以全面了解您的安全状态 AWS。Security Hub 通过安全控件评估您的 AWS 资源并检查其是否符合安全行业标准和最佳实践。有关受支持服务及控件的列表，请参阅 Security Hub 控件参考。
- AWS Audit Manager— 这 AWS 服务 可以帮助您持续审计 AWS 使用情况，从而简化风险管理以及对法规和行业标准的合规性。

Application Auto Scaling 中的恢复功能

AWS 全球基础设施是围绕 AWS 区域和可用区构建的。

AWS 区域提供多个物理隔离和隔离的可用区，这些可用区通过低延迟、高吞吐量和高度冗余的网络相连。

利用可用区，您可以设计和操作在可用区之间无中断地自动实现失效转移的应用程序和数据库。与传统的单个或多个数据中心基础设施相比，可用区具有更高的可用性、容错性和可扩展性。

有关 AWS 区域和可用区的更多信息，请参阅AWS 全球基础设施。

Application Auto Scaling 中的基础设施安全性

作为一项托管服务，Application Auto Scaling 受 AWS 全球网络安全的保护。有关 AWS 安全服务以及如何 AWS 保护基础设施的信息，请参阅AWS 云安全。要使用基础设施安全的最佳实践来设计您的 AWS 环境，请参阅 S AWS ecurity Pillar Well-Architected Fram ework 中的基础设施保护。

您可以使用 AWS 已发布的 API 调用通过网络访问 Application Auto Scaling。客户端必须支持以下内容：

- 传输层安全性协议 (TLS) 我们要求使用 TLS 1.2，建议使用 TLS 1.3。
- 具有完全向前保密 (PFS) 的密码套件，例如 DHE（临时 Diffie-Hellman）或 ECDHE（临时椭圆曲线 Diffie-Hellman）。大多数现代系统（如 Java 7 及更高版本）都支持这些模式。

此外，必须使用访问密钥 ID 和与 IAM 委托人关联的秘密访问密钥来对请求进行签名。或者，您可以使用 AWS Security Token Service（AWS STS）生成临时安全凭证来对请求进行签名。

Application Auto Scaling 配额

您的 AWS 账户 对于每项 AWS 服务都具有默认配额（以前称为限制）。除非另有说明，否则，每个限额都特定于 区域。您可以请求增加某些限额，但其他一些限额无法增加。

要查看 Application Auto Scaling 配额，请打开 Service Quotas 控制台。在导航窗格中，选择 AWS 服务，然后选择 Application Auto Scaling。

要请求提高配额，请参阅《Service Quotas 用户指南》中的请求提高配额。如果配额在 Service Quotas 中尚不可用，请使用 Application Auto Scaling 限制表。确保在增加请求中指定资源的类型，例如 Amazon ECS 或 DynamoDB。

您的 AWS 账户 具有以下 Application Auto Scaling 相关配额。

每个账户每个区域的默认配额

物品	默认	可调整
每个资源类型的最大可扩展目标数	默认配额因资源类型而异。 对于所有其他资源类型，最多 5000 个 Amazon DynamoDB 可扩展目标、3000 个 ECS 可扩展目标、1500 个 Amazon Keyspaces 可扩展目标和 500 个可扩展目标。	是
每个可扩展目标的最大扩展策略数	50 包含步进扩展策略和目标跟踪策略。	否
每个可扩展目标的最大计划操作数	200	否
每个步进扩展策略的最大步进调整数	20	否

在扩展工作负载时，请牢记服务配额。例如，当您达到某个服务允许的最大容量单位数时，向外扩展操作将会停止。如果需求下降并且当前容量下降，则 Application Auto Scaling 会再次横向扩展。为避免再次达到此容量限制，您可以请求增加配额限制。对于最大资源容量，每个服务都有各自的默认配额。有关其他 AWS 服务默认配额的信息，请参阅 Amazon Web Services 一般参考 中的服务端点和配额。

文档历史记录

下表介绍了自 2018 年 1 月以来对 Application Auto Scaling 文档的重要补充。如需对此文档更新的通知，您可以订阅 RSS 源。

变更	说明	日期
指南更改	更新了配额文档中的每个资源类型的最大可扩展目标数。请参阅 Application Auto Scaling 配额。	2024 年 1 月 16 日
对 SageMaker 推理组件的支持	使用 Application Auto Scaling 扩展推理组件的副本。	2023 年 11 月 29 日
IAM 服务相关角色权限更新	Application Auto Scaling 更新了 `AWSApplicationAutoscalingSageMakerEndpointPolicy` 策略。有关更多信息，请参阅 AWS 托管式策略的 Application Auto Scaling 更新。	2023 年 11 月 13 日
Suppor SageMaker t 支持无服务器配置的并发	使用 Application Auto Scaling 扩展无服务器端点的预置并发。	2023 年 5 月 9 日
使用标签对您的可扩展目标进行分类	您可以将自己的元数据以标签的形式分配给 Application Auto Scaling 可扩展目标。请参阅 Application Auto Scaling 标签支持。	2023 年 3 月 20 日
Support 支持 CloudWatch 公制数学	创建目标跟踪扩展策略时，您现在可使用指标数学。使用指标数学，您可以查询多个 CloudWatch 指标，并使用数学表达式根据这些指标创建新的	2023 年 3 月 14 日

	时间序列。请参阅使用指标数学为 Application Auto Scaling 创建目标跟踪扩展策略	
指南更改	《Application Auto Scaling 用户指南》中的新主题可帮助您了解如何开始将 AWS CloudShell 与 Application Auto Scaling 结合使用。请参阅 通过命令行将 AWS CloudShell 与 Application Auto Scaling 结合使用。	2023 年 2 月 17 日
不扩展的原因	现在，您可以使用 Application Auto Scaling API，检索 Application Auto Scaling 不扩展资源的机器可读原因。请参阅 Scaling activities for Application Auto Scaling（Application Auto Scaling 的扩展活动）。	2023 年 1 月 4 日
指南更改	更新了配额文档中的每个资源类型的最大可扩展目标数。请参阅 Application Auto Scaling 配额。	2022 年 5 月 6 日
添加对 Amazon Neptune 集群的支持	使用 Application Auto Scaling 来扩展 Amazon Neptune 数据库集群中的副本数量。有关更多信息，请参阅 Amazon Neptune 和 Application Auto Scaling。主题对 AWS 托管策略的 Application Auto Scaling 更新已更新以列出与 Neptune 集成的新托管策略。	2021 年 10 月 6 日

Application Auto Scaling 现在报告对其 AWS 托管式策略的更改	从 2021 年 8 月 19 日开始，对托管式策略的更改将在主题 AWS 托管式策略的 Application Auto Scaling 更新中报告。列出的第一个更改是添加了 Redis ElastiCache 所需的权限。	2021 年 8 月 19 日
添加对 Redi ElastiCache s 复制组的支持	使用 Application Auto Scaling 来扩展适用于 Redis 的复制组（集群）的节点组数量和每个节点组 ElastiCache 的副本数量。有关更多信息，请参阅 Redis 和 A ElastiCache pplication Auto Scaling。	2021 年 8 月 19 日
指南更改	Application Auto Scaling 用户指南中的新 IAM 主题可帮助您解决访问 Application Auto Scaling 的问题。有关更多信息，请参阅 Identity and Access Management for Application Auto Scaling。还为目标服务和亚马逊上的操作添加了新的示例 IAM 权限策略 CloudWatch。有关更多信息，请参阅有关使用 AWS CLI 或软件开发工具包的示例策略。	2021 年 2 月 23 日
添加对本地时区的支持	现在，您可以在本地时区创建计划的操作。如果您的时区遵守夏令时，它会自动调整夏令时 (DST)。有关更多信息，请参阅计划的扩缩。	2021 年 2 月 2 日

指南更改	Application Auto Scaling 用户指南中的新教程可帮助您了解在使用 Application Auto Scaling 时如何使用目标跟踪扩缩策略和计划的扩缩来提高应用程序的可用性。此外，还有一个新主题说明了在检测到任何可能需要您注意的问题时 CloudWatch 如何触发通知。	2020 年 10 月 15 日
添加对 Amazon Managed Streaming for Apache Kafka 集群存储的支持	使用目标跟踪扩缩策略以横向扩展与 Amazon MSK 集群关联的代理存储量。	2020 年 9 月 30 日
添加对 Amazon Comprehend 实体识别程序终端节点的支持	使用 Application Auto Scaling 可扩展为 Amazon Comprehend 实体识别程序终端节点预置的推理单位数量。	2020 年 9 月 28 日
添加对 Amazon Keyspaces (for Apache Cassandra) 表的支持	使用 Application Auto Scaling 扩展 Amazon Keyspaces 表的预置吞吐量（读取和写入容量）。	2020 年 4 月 23 日
新增“安全性”章节	Application Auto Scaling 用户指南中新的安全章节可帮助您了解如何在使用 Application Auto Scaling 时应用责任共担模式。作为此更新的一部分，已将用户指南的“身份验证和访问控制”一章替换为一个新的、更实用的部分，即 Identity and Access Management for Application Auto Scaling。	2020 年 1 月 16 日
次要更新	各种改进和更正。	2020 年 1 月 15 日

增加了通知功能	Application Auto Scaling 现在会向亚马逊发送事件，EventBridge 并在发生某些操作AWS Health Dashboard时向您发送通知。有关更多信息，请参阅 Application Auto Scaling 监控。	2019 年 12 月 20 日
添加对 AWS Lambda 函数的支持	使用 Application Auto Scaling 扩展 Lambda 函数的预置并发。	2019 年 12 月 3 日
添加对 Amazon Comprehend 文档分类终端节点的支持	使用 Application Auto Scaling 扩展 Amazon Comprehend 文档分类终端节点的吞吐量。	2019 年 11 月 25 日
添加 AppStream 2.0 对目标跟踪扩展策略的支持	使用目标跟踪扩展策略来扩展 AppStream 2.0 舰队的规模。	2019 年 11 月 25 日
对 Amazon VPC 终端节点的支持	您现在可以在 VPC 和 Application Auto Scaling 之间建立私有连接。有关迁移注意事项和说明，请参阅 Application Auto Scaling 和接口 VPC 终端节点。	2019 年 11 月 22 日
暂停和恢复扩缩	增加了对暂停和恢复扩展的支持。有关更多信息，请参阅暂停和恢复 Application Auto Scaling 的扩缩。	2019 年 8 月 29 日
新章节	设置部分已添加到 Application Auto Scaling 文档。对整个用户指南进行了少量改进和修复。	2019 年 6 月 28 日

指南更改	改进了 Application Auto Scaling 文档中的计划的扩展、分步扩缩策略和目标跟踪扩缩策略部分。	2019 年 3 月 11 日
添加对自定义资源的支持	使用 Application Auto Scaling 扩展由您自己的应用程序或服务提供的自定义资源。有关更多信息，请参阅我们的GitHub存储库。	2018 年 7 月 9 日
添加对 SageMaker 端点变体的支持	使用 Application Auto Scaling 扩展为变体预置的终端节点实例数。	2018 年 2 月 28 日

下表介绍了 2018 年 1 月之前对 Application Auto Scaling 文档的重要更改。

更改	描述	日期
添加对 Aurora 副本的支持	使用 Application Auto Scaling 扩展所需的计数。有关更多信息，请参阅 Amazon RDS 用户指南中的将 Amazon Aurora Auto Scaling 与 Aurora 副本一起使用。	2017 年 11 月 17 日
添加对计划扩展的支持	使用计划的扩展在特定预设时间或按照特定预设间隔扩展资源。有关更多信息，请参阅 Application Auto Scaling 的计划扩缩。	2017 年 11 月 8 日
添加对目标跟踪扩展策略的支持	使用目标跟踪扩展策略通过几个简单步骤为您的应用程序设置动态扩展。有关更多信息，请参阅 Application Auto Scaling 的目标跟踪扩缩策略。	2017 年 7 月 12 日

更改	描述	日期
添加对 DynamoDB 表和全局二级索引的预置读取和写入容量的支持	使用 Application Auto Scaling 扩展预置吞吐量（读取和写入容量）。有关更多信息，请参阅 Amazon DynamoDB 开发人员指南中的使用 DynamoDB Auto Scaling 管理吞吐量。	2017 年 6 月 14 日
添加对 AppStream 2.0 舰队的支持	使用 Application Auto Scaling 扩展队列的规模。有关更多信息，请参阅 Amazon AppStream 2.0 管理指南中的 Fleet Auto Scaling for AppStream 2.0。	2017 年 3 月 23 日
添加对 Amazon EMR 集群的支持	使用 Application Auto Scaling 扩展核心和任务节点。有关更多信息，请参阅 Amazon EMR 管理指南 中的在 Amazon EMR 中使用弹性伸缩。	2016 年 11 月 18 日
增加对 Spot 队列的支持	使用 Application Auto Scaling 扩展目标容量。有关更多信息，请参阅适用于 Linux 实例的 Amazon EC2 用户指南中的 Spot 实例集的弹性伸缩。	2016 年 9 月 1 日
添加对 Amazon ECS 服务的支持	使用 Application Auto Scaling 扩展所需的计数。有关更多信息，请参阅 Amazon Elastic Container Service 开发人员指南中的服务 Auto Scaling。	2016 年 8 月 9 日

本文属于机器翻译版本。若本译文内容与英语原文存在差异，则一律以英文原文为准。