aws

扩展计划用户指南

AWS Auto Scaling

AWS Auto Scaling: 扩展计划用户指南

Table of Contents

什么是扩缩计划？

使用扩缩计划在几分钟内为相关或关联的可扩展资源配置弹性伸缩。例如，您可以使用标签将资源分为生产、测试或开发等类别。然后，您可以搜索属于每个类别的可扩展资源并设置扩缩计划。或者，如果您的云基础架构包括 AWS CloudFormation，则可以定义用于创建资源集合的堆栈模板。然后为属于每个堆栈的可扩展资源创建扩缩计划。

支持的资源

AWS Auto Scaling 支持对以下服务和资源使用扩展计划：

- Amazon Aurora – 增减为 Aurora 数据库集群预置的 Aurora 只读副本数量。
- Amazon EC2 Auto Scaling – 通过增减 Auto Scaling 组的所需容量启动或终止 EC2 实例。
- Amazon Elastic Container Service – 在 Amazon ECS 中增减所需的任务数。
- Amazon DynamoDB – 增减 DynamoDB 表或全局二级索引的预置读取和写入容量。
- 竞价型实例集 – 通过增减竞价型实例集的目标容量来启动或终止 EC2 实例。

扩缩计划的功能和优势

扩缩计划具有以下功能和优势：

- 资源发现 — AWS Auto Scaling 提供自动资源发现，以帮助在应用程序中查找可扩展的资源。
- 动态扩缩 – 扩缩计划使用 Amazon EC2 Auto Scaling 和 Application Auto Scaling 服务来调整可扩展资源的容量，以适应流量或工作负载的变化。动态扩缩指标可以是标准的利用率或吞吐量指标，也可以是自定义指标。
- 内置扩缩建议 – AWS Auto Scaling 提供包含建议的扩缩策略，您可以使用这些建议来优化性能、成本或平衡性能与成本。
- 预测性扩缩 – 扩缩计划还支持 Auto Scaling 组的预测性扩缩。这有助于在定期出现峰值时更快地扩展 Amazon EC2 容量。

> **Important**
>
> 如果您仅将扩缩计划用于预测性扩缩，我们强烈建议您直接在自动扩缩组组上设置预测性扩缩策略。这一最近推出的选项提供了更多功能，例如通过指标聚合来创建新的自定义指标或跨蓝

绿部署保留历史指标数据。有关更多信息，请参阅 Amazon EC2 Auto Scaling 用户指南中的 Amazon EC2 Auto Scaling 的预测性扩展。
有关从扩展计划迁移到 Amazon EC2 Auto Scaling 预测性扩展策略的指南，请参阅迁移您的扩展计划。

如何开始

使用以下资源可帮助您创建和使用扩缩计划：

- 扩展计划的工作原理
- 扩展计划的最佳实践
- 扩缩计划入门

使用扩缩计划

您可以通过下面的任何一种方式来创建、访问和管理扩缩计划：

- AWS Management Console – 提供了可用来访问扩缩计划的 Web 界面。如果您已经注册了 AWS 账户，则可以通过登录来访问您的扩展计划 AWS Management Console，使用导航栏上的搜索框进行搜索 AWS Auto Scaling，然后选择AWS Auto Scaling。
- AWS Command Line Interface (AWS CLI) — 为各种各样的用户提供命令 AWS 服务，并在 Windows、macOS 和 Linux 上受支持。要开始使用，请参阅AWS Command Line Interface 《用户指南》。有关更多信息，请参阅《AWS CLI 命令参考》中的 autoscaling-plans。
- 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 AWS Auto Scaling API 参考。
- AWS CloudFormation— 支持使用 CloudFormation 模板创建扩展计划。有关更多信息，请参阅《AWS CloudFormation 用户指南》中的AWS::AutoScalingPlans::ScalingPlan参考资料。

区域可用性

该 AWS Auto Scaling API 有多个版本可用， AWS 区域 并且为每个区域都提供了一个终端节点。有关当前可用 API 的所有区域和终端节点的列表，请参阅中国亚马逊网络服务终端节点和配额 AWS。

定价

所有扩缩计划功能都已为您启用。除了服务费 CloudWatch 和您使用的其他 AWS Cloud 资源的服务费外，这些功能不收取任何额外费用。

> **Note**
>
> 预测性扩展功能依靠 CloudWatch GetMetric数据操作来收集历史指标数据以进行容量预测，这会产生成本。但是，如果您使用 Amazon EC2 Auto Scaling 扩展策略而不是扩展计划启用预测性扩展，则调用无需支付任何费用`GetMetricData`。

扩展计划的工作原理

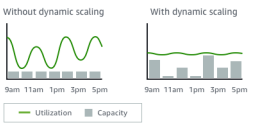
AWS Auto Scaling 允许您使用扩展计划来配置一组扩展资源的指令。如果您使用可扩展资源 AWS CloudFormation 或为其添加标签，则可以为每个应用程序的不同资源集设置扩展计划。 AWS Auto Scaling 控制台提供针对每种资源定制的扩展策略的建议。在创建扩缩计划后，它会将动态扩缩方法和预测性扩缩方法进行组合以支持您的扩缩策略。

什么是扩展策略？

扩展策略说明 AWS Auto Scaling 如何优化扩展计划中的资源利用率。您可以针对可用性、成本或这两者的平衡进行优化。或者，您也可以考虑根据自己定义的指标和阈值自行创建自定义策略。您可以为每种资源或资源类型设置单独的策略。

什么是动态扩展？

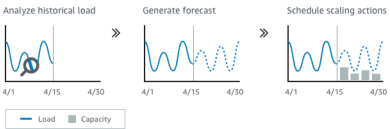
动态扩展为您的扩展计划中的资源创建目标跟踪扩展策略。这些扩展策略将调整资源容量以响应资源利用率的实时变化。其目的是提供足够的容量以将利用率保持在扩展策略指定的目标值。这与恒温器保持家里温度的方式类似。您选择温度，恒温器将完成剩下的工作。



例如，您可以配置扩缩计划以使 Amazon Elastic Container Service（Amazon ECS）服务运行的任务数能够保持 75% 的 CPU 利用率。当服务的 CPU 利用率超过 75%（这意味着，为服务预留的 CPU 已使用了超过 75%）时，您的扩缩策略会向您的服务添加另一个任务来帮助处理增加的负载。

什么是预测式扩展？

预测性扩缩使用机器学习来分析每个资源的历史工作负载，并定期预测未来的负载。这类似于天气预报的工作方式。利用预测，预测式扩展会生成计划的扩展操作，以确保在应用程序需要之前有资源容量可用。与动态扩展相似，预测式扩展的作用是将利用率保持在扩展策略指定的目标值。



例如，您可以启用预测式扩展并配置您的扩展策略，以将 Auto Scaling 组的平均 CPU 利用率保持在 50%。您的预测认为每天 8 点会出现流量峰值。您的扩展计划将创建未来的计划扩展操作，以确保您

的 Auto Scaling 组已做好提前处理该流量的准备。这有助于使应用程序性能保持不变，目的是始终拥有所需的容量来尽可能让资源利用率保持在接近 50%。

以下是理解预测性扩缩的关键概念：

- 负荷预测：针对指定负荷指标 AWS Auto Scaling 分析最多 14 天的历史记录，并预测未来两天的未来需求。此数据以一小时的间隔提供并且每天更新。
- 计划的扩展操作： AWS Auto Scaling 计划主动增加和减少容量的扩展操作，以匹配负载预测。在计划时间，使用计划的扩展操作指定的值 AWS Auto Scaling 更新最小容量。其目的是将资源利用率保持在扩展策略指定的目标值。如果您的应用程序需要的容量大于预期，则可以使用动态扩展来增加额外的容量。
- 最大容量行为：弹性伸缩的最小和最大容量限制适用于每个资源。但您可以控制在预测容量超过最大容量时应用程序是否可以增加容量以超过最大容量。

> **Note**
>
> 现在，您可以使用 Auto Scaling 组的预测性扩缩策略。有关更多信息，请参阅 Amazon EC2 Auto Scaling 用户指南中的 Amazon EC2 Auto Scaling 的预测性扩展。

扩展计划的最佳实践

以下最佳实践可帮助您充分利用扩展计划：

- 创建启动模板或启动配置时，请启用详细监控，以一分钟为频率获取 EC2 实例的 CloudWatch 指标数据，因为这样可以确保更快地响应负载变化。根据五分钟频率的指标进行扩缩会导致响应时间增加，并且扩缩所依据的指标数据可能过时。预设情况下，系统会为 EC2 实例启用基本监控，这意味着实例的指标数据将以五分钟为间隔提供。您可以启用详细监控，从而以一分钟的频率获取实例的指标数据，但这会产生额外的费用。有关更多信息，请参阅《Amazon EC2 Auto Scaling 用户指南》中的为 Auto Scaling 实例配置监控。
- 我们还建议您启用 Auto Scaling 组指标。否则，实际容量数据不会显示在完成“创建扩展计划”向导后提供的容量预测图中。有关更多信息，请参阅 Amazon EC2 Auto Scaling 用户指南中的 Auto Scaling 组和实例的监控 CloudWatch 指标。
- 检查您的 Auto Scaling 组使用的实例类型，并谨防使用具爆发能力的实例类型。具爆发能力的 Amazon EC2 实例（例如 T3 和 T2 实例）旨在提供基准水平的 CPU 性能，并且能够在您的工作负载需要时突增到更高的水平。根据扩展计划指定的目标利用率，您可以管理超出基准的风险，然后用完 CPU 积分，这将限制性能。有关更多信息，请参阅具爆发能力的实例的 CPU 积分和基准性能。要将这些实例配置为`unlimited`，请参阅 A mazon EC2 用户指南中的使用 Auto Scaling 组将突发性能实例启动为无限制。

其他考虑因素

> **Note**
>
> 预测性扩缩有更新的版本，于 2021 年 5 月发布。此版本中引入的某些功能在扩缩计划中不可用，您必须使用直接在自动扩缩组上设置的预测性扩缩策略才能访问这些功能。有关更多信息，请参阅 Amazon EC2 Auto Scaling 用户指南中的 Amazon EC2 Auto Scaling 的预测性扩展。

考虑以下其他注意事项：

- 预测性扩缩使用负载预测来计划未来的容量。预测质量因负载的周期性和所训练预测模型的适用性而异。可以在仅预测模式下运行预测式扩展，以评估预测的质量和预测创建的扩展操作。您可以在创建扩展计划时将预测式扩展模式设置为仅预测，然后在完成评估预测质量后将其更改为预测和缩放。有关更多信息，请参阅预测性扩展设置和监控和评估预测。

- 如果您选择为预测式扩展指定不同的指标，则必须确保扩展指标和负载指标密切相关。指标值必须随着 Auto Scaling 组中实例的数量按比例增加和缩小。这样可确保指标数据可用于随实例数量按比例扩展或缩减。例如，负载指标是请求计数总计，扩展指标是平均 CPU 利用率。如果请求计数总计增加了 50%，则还应将平均 CPU 利用率增加 50%，前提是容量保持不变。
- 在创建扩展计划之前，您应通过访问创建扩展计划时使用的控制台来删除任何先前计划的、不再需要的扩展操作。 AWS Auto Scaling 不会创建与现有计划扩展操作重叠的预测性扩展操作。
- 您的最小容量和最大容量的自定义设置，以及用于动态扩展的其他设置将显示在其他控制台中。但是，我们建议，您在创建扩展计划后，不要通过其他控制台修改这些设置，因为您的扩展计划不从其他控制台接收更新。
- 您的扩展计划可以包含来自多个服务的资源，但每个资源一次只能在一个扩展计划中。

避免 ActiveWithProblems 错误

创建扩展计划或向扩展计划添加资源时，可能会出现 “ActiveWith问题” 错误。当扩展计划处于活动状态，但一个或多个资源的扩展配置无法应用时，出现此错误。

通常情况下，这是因为资源已经具有扩展策略，或 Auto Scaling 组不符合预测式扩展的最低要求。

如果您的任何资源已经具有来自各种服务控制台的扩展策略，则默认情况下， AWS Auto Scaling 不会覆盖这些其他扩展策略或创建新的扩展策略。您可以选择删除现有的扩展策略，并将其替换为通过 AWS Auto Scaling 控制台创建的目标跟踪扩展策略。为此，您可以为每个具有要覆盖的扩展策略的资源启用 Replace external scaling policies (替换外部扩展策略) 设置。

对于预测性扩展，在创建新的 Auto Scaling 组来配置预测式扩展后，我们建议等待 24 小时时间。至少必须有 24 小时的历史数据才能生成初始预测。如果该组具有的历史数据少于 24 小时并且启用了预测性扩缩，则扩缩计划在该组收集所需数据量之后的下一个预测期之前无法生成预测数据。但是，您也可以编辑和保存扩展计划，以便在 24 小时的数据可用后立即重新启动预测过程。

扩缩计划入门

在创建将用于应用程序的扩缩计划之前，请全面考察应用程序在 AWS Cloud中运行时的情况。记录以下内容：

- 您是否具有通过其他控制台创建的现有扩展策略。您可以替换现有扩展策略，也可以在创建扩展计划时保留它们（不允许对其值进行任何更改）。
- 对您的应用程序中作为整体基于资源的每个可扩展资源有意义的目标利用率。例如，与 EC2 实例的可用 CPU 相比，Auto Scaling 组中的 EC2 实例预期使用的 CPU 数量。或者对于像 DynamoDB 这样使用预置吞吐量模型的服务，与可用吞吐量相比，表或索引预期使用的读写活动量。换言之，即使用的容量与预置容量的比值。创建扩缩计划后，您可以随时更改目标利用率。
- 启动和配置服务器需要多长时间。了解这一点有助于您为每个 EC2 实例配置一个在启动后预热的时间段，以确保前一台服务器仍在启动时不会启动新的服务器。
- 指标历史是否长到足够用于预测式扩展（如果使用的是新创建的 Auto Scaling 组）。一般而言，具有 14 整天的历史数据将转化为更准确的预测。最小值为 24 小时。

您越了解您的应用程序，您制定扩展计划的效率就越高。

以下任务可帮助您熟悉扩缩计划。您将为单个 Auto Scaling 组创建一个扩缩计划，并启用预测性扩缩和动态扩缩。

任务

步骤 1：查找您的可扩展资源

这一部分包括在 AWS Auto Scaling 控制台中创建扩缩计划的动手实践说明。如果这是您的第一个扩缩计划，我们建议您首先使用一个 Amazon EC2 Auto Scaling 组创建一个示例扩缩计划。

先决条件

要练习使用扩缩计划，请创建一个 Auto Scaling 组。在该 Auto Scaling 组中至少启动一个 Amazon EC2 实例。有关更多信息，请参阅《Amazon EC2 Auto Scaling 用户指南》中的 Amazon EC2 Auto Scaling 入门。

使用启用 CloudWatch 指标的 Auto Scaling 组，在完成 “创建扩展计划” 向导时可用的图表上显示容量数据。有关更多信息，请参阅《Amazon EC2 Auto Scaling 用户指南》中的启用 Auto Scaling 组指标。

如果可能，在几天或更长时间内生成一些负载，以便为预测性扩展功能提供 CloudWatch 指标数据。

验证您拥有使用扩缩计划所需的权限。有关更多信息，请参阅 扩展计划的身份和访问管理。

将您的 Auto Scaling 组添加到您的新扩缩计划

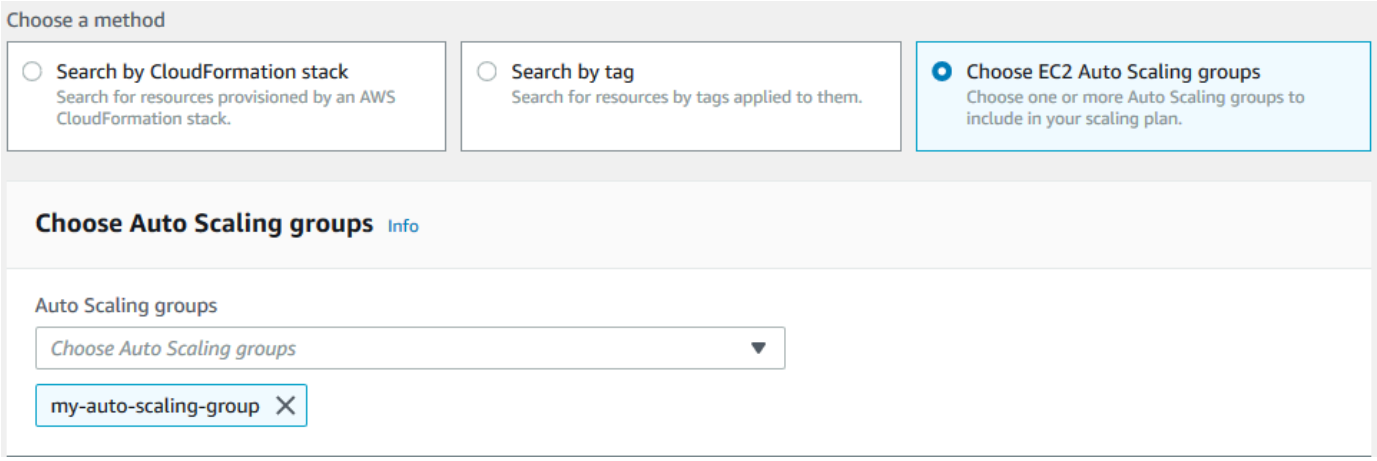
从控制台创建扩缩计划时，控制台首先会帮助您查找可扩展的资源。请首先确认您满足以下要求，然后再继续操作：

- 如上一部分所述，您创建了一个 Auto Scaling 组并至少启动了一个 EC2 实例。
- 您创建的 Auto Scaling 组至少已存在 24 小时。

开始创建扩缩计划

1. 打开 AWS Auto Scaling 控制台，网址为 https://console.aws.amazon.com/awsautoscaling/。
2. 在屏幕顶部的导航栏中，选择您在创建 Auto Scaling 组时使用的同一区域。
3. 从欢迎页面中，选择 Get started (开始使用)。
4. 在 Find scalable resources（查找可扩展资源）页面中，执行下面的一项操作：
   - 选择 “按 CloudFormation 堆栈搜索”，然后选择要使用的 AWS CloudFormation 堆栈。
   - 选择 Search by tag（按标签搜索）。然后对于每个标签，从 Key（键）中选择一个标签键，并从 Value（值）中选择标签值。要添加标签，请选择 Add another row (添加其他行)。要删除标签，请选择删除。
   - 选择 Choose EC2 Auto Scaling groups（选择 EC2 Auto Scaling 组），然后，选择一个或多个 Auto Scaling 组。

> **Note**
>
> 有关入门教程，请选择 Choose EC2 Auto Scaling groups（选择 EC2 Auto Scaling 组），然后选择您创建的 Auto Scaling 组。



5. 选择 Next（下一步）以继续扩缩计划的创建过程。

详细了解如何发现可扩展资源

如果您已经创建了示例扩展计划并想创建更多扩展计划，请更详细地查看以下使用 CloudFormation 堆栈或一组标签的场景。在使用控制台创建扩展计划时，您可以使用此部分来决定是选择 “按 CloudFormation 堆栈搜索” 还是 “按标签搜索” 选项来发现您的可扩展资源。

当您在创建扩展计划向导的步骤 1 中选择 “按 CloudFormation 堆栈搜索” 或 “按标签搜索” 选项时，这会使与堆栈或一组标签关联的可扩展资源可用于扩展计划。当您定义扩展计划时，您接着可以选择要包含或排除其中哪些资源。

使用 CloudFormation 堆栈发现可扩展的资源

使用时 CloudFormation，您可以使用堆栈来配置资源。堆栈中的所有资源均由堆栈的模板定义。您的扩展计划在堆栈顶部添加了一个业务流程层，从而可以更轻松地为多个资源配置扩展。如果没有扩展计划，则需要为每个可扩展资源单独设置扩展。这意味着要弄清楚预配置资源和扩展策略的顺序，并了解这些依赖项工作方式的精妙之处。

在 AWS Auto Scaling 控制台中，您可以选择现有堆栈对其进行扫描，寻找可以配置为自动扩展的资源。 AWS Auto Scaling 仅查找在选定堆栈中定义的资源。它不会遍历嵌套堆栈。

要在 CloudFormation 堆栈中发现您的 ECS 服务， AWS Auto Scaling 控制台必须知道哪个 ECS 集群正在运行该服务。这要求您的 ECS 服务与运行该服务的 ECS 集群位于同一个 CloudFormation 堆栈中。否则，它们必须是默认集群的一部分。为了正确识别服务，ECS 服务名称在每个 ECS 集群中也必须是唯一的。

有关的更多信息 CloudFormation，请参阅什么是 AWS CloudFormation？ 在《AWS CloudFormation 用户指南》中。

使用标签发现可扩展资源

标签提供的元数据可用于使用标签过滤器在 AWS Auto Scaling 控制台中发现相关的可扩展资源。

使用标签来查找以下任何资源：

- Aurora 数据库集群
- 自动扩缩组
- DynamoDB 表和全局二级索引

当您按多个标签搜索时，每个资源都必须发现所有列出的标签。

有关标记的更多信息，请参阅以下文档。

- 请参阅《Amazon Aurora 用户指南》以了解如何标记 Aurora 集群。
- 请参阅《Amazon EC2 Auto Scaling 用户指南》以了解如何标记 Auto Scaling 组。
- 请参阅《Amazon DynamoDB 开发人员指南》以了解如何标记 DynamoDB 资源。
- 要详细了解为AWS 资源添加标签的最佳实践，AWS 一般参考请参阅。

步骤 2：指定扩展策略

使用以下过程为上一步中发现的资源指定扩展策略。

对于每种类型的资源， AWS Auto Scaling 选择最常用于确定在任何给定时间使用了多少资源的指标。您应选择最合适的扩展策略以根据此指标优化性能。当您启用动态扩展功能和预测式扩展功能时，在它们之间共享扩展策略。有关更多信息，请参阅 扩展计划的工作原理。

有以下扩展策略可用：

- 优化可用性 — 自动AWS Auto Scaling 扩展和扩展资源，将资源利用率保持在 40%。当您的应用程序具有紧急且有时无法预测的扩展需求时，此选项很有用。

- 平衡可用性和成本 — 自动AWS Auto Scaling 扩展和扩展资源，将资源利用率保持在 50%。此选项可帮助您保持高可用性，同时降低成本。
- 针对成本进行优化 — 自动AWS Auto Scaling 扩展和扩展资源，将资源利用率保持在 70%。如果您的应用程序可以在需求出现意外更改时处理缓冲区容量减少的情况，则此选项可用于降低成本。

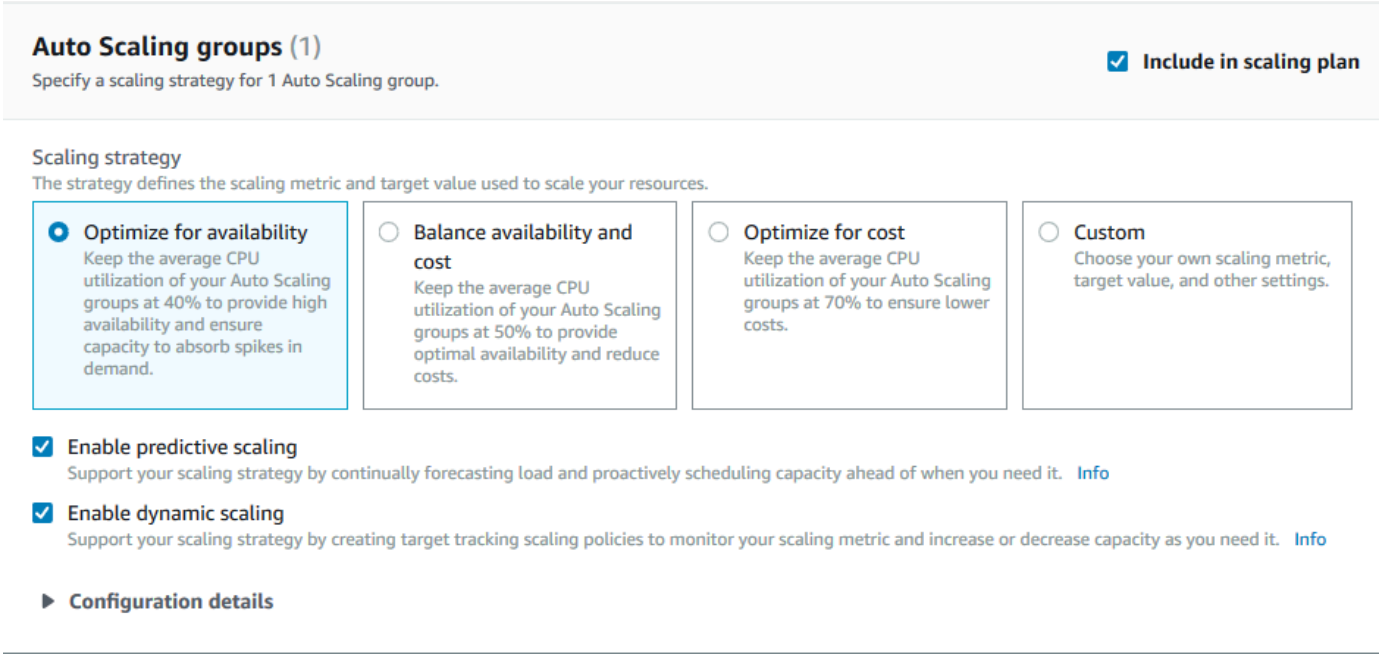
例如，扩展计划将您的 Auto Scaling 组配置为根据组中所有实例平均使用的 CPU 量来添加或删除 Amazon EC2 实例。您可选择是否通过更改扩展策略来针对可用性、成本或两者的组合优化使用率。

如果现成的策略不能满足您的需求，您也可以配置自定义策略。使用自定义策略，您可以更改目标利用率值，选择其他指标，或同时采用这两种方法。

> **Important**
>
> 对于入门教程，请仅完成以下过程的第一步，然后选择 Next（下一步）继续。

指定扩缩策略

1. 在 Specify scaling strategy (指定扩展策略) 页上，对于 Scaling plan details (扩展计划详细信息)、Name (名称)，输入扩展计划的名称。扩缩计划的名称在此区域的扩缩计划集中必须唯一。扩缩计划的名称最多可使用 128 个字符，并且不得包含竖线“|”、正斜杠“/”或冒号“:”。
2. 所有包含的资源都按资源类型列出。对于 Auto Scaling groups（Auto Scaling 组），执行以下操作：



a. 跳过此步骤以使用默认扩缩策略和指标。要使用其他扩缩策略或指标，请继续执行以下步骤：

   i. 对于 Scaling strategy（扩缩策略），选择所需的扩缩策略。

      对于入门教程，一定要选择 Optimize for availability（提高可用性）。这会指定将 Auto Scaling 组的平均 CPU 利用率保持在 40%。

   ii. 如果您选择 Custom（自定义），则展开 Configuration details（配置详细信息）以选择所需的指标和目标值。

      - 对于 Scaling metric (扩展指标)，请选择所需的扩展指标。
      - 对于 Target value（目标值），选择所需的目标值，例如在任意一分钟间隔内的目标利用率或目标吞吐量。
      - 对于 Load metric（负载指标）[仅限 Auto Scaling 组]，选择将用于预测性扩缩的负载指标。
      - 选择替换外部扩展策略以指定哪些策略 AWS Auto Scaling 可以删除先前从扩展计划外部（例如从其他控制台）创建的扩展策略，并将其替换为由扩展计划创建的新目标跟踪扩展策略。

b. （可选）预设情况下，系统已为 Auto Scaling 组启用预测性扩缩。要为 Auto Scaling 组关闭预测性扩缩，请清除 Enable predictive scaling（启用预测性扩缩）。

c. （可选）默认情况下，将为每个资源类型启用动态扩展。要为某种资源关闭动态扩缩，请清除 Enable dynamic scaling（启用动态扩缩）。

d. （可选）默认情况下，当您指定为其发现了多个可扩展资源的应用程序源时，所有资源类型自动包括到您的扩展计划中。要在扩展计划中忽略某种资源，请清除包含在扩展计划中。

3. （可选）要为其他资源类型指定扩缩策略，请重复上述步骤。

4. 完成后，选择 Next（下一步）以继续扩缩计划的创建过程。

步骤 3：配置高级设置（可选）

现在您已指定要用于每个资源类型的扩展策略，可以使用配置高级设置步骤，选择按资源自定义任何默认设置。对于每个资源类型，您可以定义多组设置。但在大多数情况下，默认设置应会更加高效，最小容量和最大容量的值也许可以例外，但应谨慎调整。

如果要保留默认设置，则跳过此过程。您可以通过编辑扩展计划随时更改这些设置。

> **Important**
>
> 对于入门教程，我们可以进行一些更改，以更新 Auto Scaling 组的最大容量并启用仅预测模式的预测性扩缩。虽然您不需要自定义教程的所有设置，我们可以简单查看一下各个部分中的设置。

常规设置

使用此过程可以按照各个资源，查看和自定义您在上一步中指定的设置。您还可以自定义每个资源的最小容量和最大容量。

查看和自定义常规设置

1. 在配置高级设置页面上，选择左侧任意部分标题的箭头以展开该部分。在本教程中，展开 Auto Scaling 组部分。
2. 从显示的表中，选择您在本教程中使用的 Auto Scaling 组。
3. 保留选中包含在扩展计划中选项。如果未选择此选项，则扩展计划中会忽略资源。如果您未包含至少一个资源，则无法创建扩展计划。
4. 要展开视图并查看常规设置部分的详细信息，请选择部分标题左侧的箭头。
5. 您可以选择以下任意项。在本教程中，找到最大容量设置并输入值 3 代替当前值。

   - Scaling strategy (扩展策略) – 允许您提高可用性、优化成本，或使可用性和成本达到平衡，或指定自定义策略。
   - Enable dynamic scaling (启用动态扩展) – 如果清除了此设置，则无法使用目标跟踪扩展配置扩展所选资源。
   - Enable predictive scaling (启用预测式扩展) – [仅限 Auto Scaling 组] 如果清除此设置，则无法使用预测式扩展来扩展所选组。
   - Scaling metric (扩展指标) – 指定要使用的扩展指标。如果您选择 Custom (自定义)，则可指定要使用的自定义指标而不是控制台中可用的负载指标。有关更多信息，请参阅此部分中的下一个主题。
   - Target value (目标值) – 指定要使用的目标利用率值。
   - Load metric (负载指标) – [仅限 Auto Scaling 组] 指定要使用的负载指标。如果您选择 Custom (自定义)，则可指定要使用的自定义指标而不是控制台中可用的负载指标。有关更多信息，请参阅此部分中的下一个主题。
   - 最小容量-指定资源的最小容量。 AWS Auto Scaling 确保您的资源永远不会低于此大小。

- 最大容量-指定资源的最大容量。 AWS Auto Scaling 确保您的资源永远不会超过此大小。

> **Note**
>
> 使用预测式扩展时，您也可以选择根据预测容量来使用其他最大容量行为。此设置位于预测式扩展设置部分中。

自定义指标

AWS Auto Scaling 提供了最常用的自动缩放指标。但是，根据您的需求，您可能偏爱从不同的指标而不是控制台中的指标获取数据。Amazon CloudWatch 有许多不同的指标可供选择。 CloudWatch 还允许您发布自己的指标。

您可以使用 JSON 来指定 CloudWatch 自定义指标。在按照这些说明进行操作之前，我们建议您先熟悉 Amazon CloudWatch 用户指南。

要指定自定义指标，您可使用模板中的一组必需参数构造 JSON 格式的负载。您可以为来自的每个参数添加值 CloudWatch。在扩展计划的高级设置中，我们提供模板作为扩展指标和负载指标的自定义选项的一部分。

JSON 通过两种方式表示数据：

- 对象，其是无序名称-值对集合。对象是在左大括号 ({) 和右大括号 (}) 内定义的。每个名称-值对以名称开头，后接一个冒号，再接值。名称-值对是用逗号隔开的。
- 数组，其是有序值集合。数组是在左方括号 ([) 和右方括号 (]) 内定义的。数组中的项目是用逗号隔开的。

下面是为每个参数提供示例值的 JSON 模板的示例：

```
{
  "MetricName": "MyBackendCPU",
  "Namespace": "MyNamespace",
  "Dimensions": [
    {
      "Name": "MyOptionalMetricDimensionName",
      "Value": "MyOptionalMetricDimensionValue"
    }
  ],
  "Statistic": "Sum"
```

```
}
```

有关更多信息，请参阅《AWS Auto Scaling API 参考》中的自定义扩展指标规范和自定义负载指标规范。

动态扩展设置

使用此过程可以查看和自定义 AWS Auto Scaling 创建的目标跟踪扩展策略的设置。

查看和自定义动态扩展的设置

1. 要展开视图并查看动态扩展设置部分的详细信息，请选择部分标题左侧的箭头。
2. 您可以为以下项进行选择。但是，默认设置非常适用于本教程。

   - Replace external scaling policies (替换外部扩展策略) – 如果清除此设置，则将保留在扩展计划外创建的现有扩展策略并且不会创建新的扩展策略。
   - Disable scale-in (禁用横向缩减) – 如果清除此设置，则在指定指标低于目标值时，允许自动横向缩减以减小资源的当前容量。
   - Cooldown (冷却) – 创建横向扩展和横向缩减冷却时间。冷却时间是指等待上一个扩展活动生效的时间量。有关更多信息，请参阅《Application Auto Scaling 用户指南》中的冷却时间。（如果资源是 Auto Scaling 组，则不使用此设置。）
   - 实例预热 — [仅限 Auto Scaling 组] 控制新启动的实例开始对指标做出贡献之前所经过的时间。CloudWatch 有关更多信息，请参阅《Amazon EC2 Auto Scaling 用户指南》中的实例预热。

预测性扩展设置

如果您的资源是 Auto Scaling 组，请使用此过程查看和自定义用于预测性扩展的设置 AWS Auto Scaling 。

查看和自定义预测式扩展的设置

1. 要展开视图并查看预测式扩展设置部分的详细信息，请选择部分标题左侧的箭头。
2. 您可以为以下项进行选择。在本教程中，请将预测式扩展模式更改为仅预测。

   - Predictive scaling mode (预测式扩展模式) – 指定扩展模式。默认值为 Forecast and scale (预测和扩展)。如果您将它更改为仅预测，则扩展计划将预测未来容量，但不会应用扩展操作。
   - Pre-launch instances (预启动实例) – 调整横向扩展时要提前运行的扩展操作。例如，预测表示在上午 10:00 点增加容量，缓冲时间为 5 分钟（300 秒）。这样，对应的扩展操作的运行时间

为上午 9:55。这对于 Auto Scaling 组很有帮助，这些组在从实例启动到服务可能需要几分钟。实际时间取决于诸多因素，如实例大小和是否有启动脚本要完成等。默认值为 300 秒。

- Max capacity behavior (最大容量行为) – 控制当预测容量接近或超过当前指定的最大容量时，所选资源是否可以纵向扩展到最大容量以上。默认值为强制实施最大容量设置。
  - 强制使用最大容量设置-AWS Auto Scaling 不能将资源容量扩展到高于最大容量的范围。最大容量是作为硬限制实施的。
  - 将@@ 最大容量设置为等于预测能力-AWS Auto Scaling 可以将资源容量扩展到高于最大容量以等于但不超过预测能力。
  - 将最大容量增加到预测容量之上 —AWS Auto Scaling 可以按指定的缓冲值将资源容量扩展到比最大容量更高的容量。目的是在出现意外流量时，为目标跟踪扩展策略提供额外的容量。
- Max capacity behavior buffer (最大容量行为缓冲区) – 如果您选择 Increase maximum capacity above forecast capacity (提高最大容量以超过预测容量)，选择在预测容量接近或超过最大容量时，所用容量缓冲区的大小。该值是作为相对于预测容量的百分比指定的。例如,使用 10% 的缓冲区，如果预测容量为 50，最大容量为 40，则有效的最大容量是 55。

3. 自定义完设置之后，选择 Next (下一步)。

> **Note**
>
> 要还原您的任何更改，请选择所需资源，然后选择 Revert to original (还原为最初设置)。这会将所选资源重置为扩展计划中的上一个已知状态。

步骤 4：创建您的扩展计划

在 Review and create (审核和创建) 页面上，审核您的扩展计划并选择 Create scaling plan (创建扩展计划)。您会定向到显示扩展计划状态的页面。在更新资源时，扩展计划的创建可能需要一点时间才能完成。

通过预测缩放， AWS Auto Scaling 可以分析过去 14 天内指定负荷指标的历史记录（至少需要 24 小时的数据），以生成未来两天的预测。然后，它将安排扩展操作来调整资源容量调整，使之与预测期内每小时的预测匹配。

在扩展计划创建完成之后，通过在扩展计划屏幕中选择其名称来查看扩展计划详细信息。

（可选）查看资源的扩展信息

使用此过程可以查看为资源创建的扩展信息。

数据通过以下方式提供：

- 图表显示了来自的最新指标历史数据 CloudWatch。
- 预测缩放图显示基于来自的数据的负荷预测和容量预测 AWS Auto Scaling。
- 表中列出了为资源计划的所有预测式扩展操作。

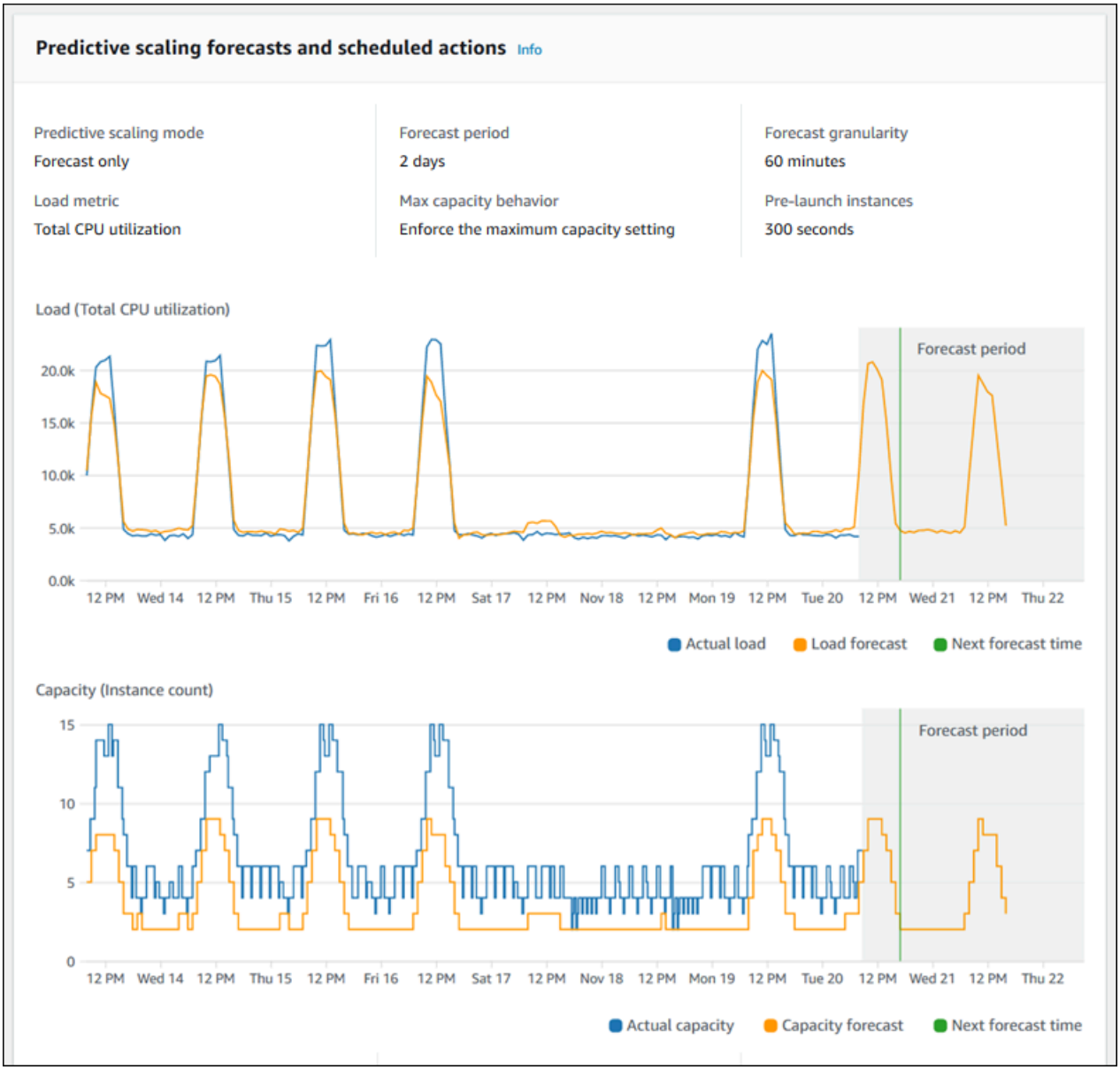
查看资源的扩展信息

1. 打开 AWS Auto Scaling 控制台，网址为 https://console.aws.amazon.com/awsautoscaling/。
2. 在 Scaling plans (扩展计划) 页面上，选择扩展计划。
3. 在 Scaling plan details (扩展计划详细信息) 页面上，选择要查看的资源。

监控和评估预测

当扩展计划启动运行时，您可以监控负载预测、容量预测和扩展操作，以检查预测式扩展的性能。所有启用预测性扩展的 Auto Scaling 组均可在 AWS Auto Scaling 控制台中查看所有这些数据。请记住，您的扩展计划需要至少 24 小时的历史负载数据来进行初次预测。

在以下示例中，每个图表的左侧都显示历史模式。右侧显示扩展计划在预测期间生成的预测。实际值和预测值（分别为蓝色和橙色）均绘制。



AWS Auto Scaling 自动从您的数据中学习。首先，它会进行负载预测。然后，容量预测计算确定支持应用程序所需的最小实例数。根据容量预测， AWS Auto Scaling 计划在预测的负载变化之前扩展 Auto Scaling 组的扩展操作。如果启用了动态扩展（推荐），则 Auto Scaling 组可以根据实例组的当前利用率横向扩展其他容量（或删除容量）。

当评估预测式扩展的执行情况时，可监控实际值和预测值在一段时间内 的接近程度。创建扩展计划时，会根据最新的实际数据 AWS Auto Scaling 提供图表。它还提供接下来 48 小时内的初始预测。但是，在创建扩展计划后，几乎没有可与实际数据进行比较的预测数据。请等到扩展计划已获取若干时间段的预测值，然后再将历史预测值与实际值进行比较。经过几天的每日预测后，您将有更多的预测值样本与实际值进行比较。

对于每天发生的模式，创建扩展计划和评估预测有效性之间的时间间隔可以短至为几天。但是，此时间长度不足以基于最近模式更改来评估预测。例如，假设您正在查看对某个 Auto Scaling 组的预测，该组在过去一周启动了一个新的市场营销活动。该活动显著增加了您在每周的相同两天的 Web 流量。在类似这样的情况下，我们建议您等待该组收集完整的一周或两周的新数据，然后再评估预测的有效性。对于仅仅开始收集指标数据的全新 Auto Scaling 组，上述建议同样适用。

如果您在监控实际值和预测值一段时间之后，发现它们并不匹配，则还应考虑负载指标的选择。若要有效发挥作用，负载指标必须表示对 Auto Scaling 组中所有实例的总负载的可靠而准确的度量。负载指标是预测性扩展的核心。如果您选择非最佳负载指标，则它可能会阻止预测性扩展，从而进行准确的负载和容量预测，并为您的 Auto Scaling 组安排正确的容量调整。

第 5 步：清理

在完成入门教程后，您可以选择保留您的扩展计划。但是，如果您的扩展计划未在活跃使用中，则应考虑将其删除以免您的账户产生不必要的费用。

删除扩展计划会删除目标跟踪扩展策略、其关联 CloudWatch 警报以及代表您 AWS Auto Scaling 创建的预测性扩展操作。

删除扩展计划不会删除您的 AWS CloudFormation 堆栈、Auto Scaling 组或其他可扩展资源。

删除扩展计划

1. 打开 AWS Auto Scaling 控制台，网址为 https://console.aws.amazon.com/awsautoscaling/。
2. 在扩展计划页面上，选择您为此教程创建的扩展计划，然后选择删除。
3. 当系统提示进行确认时，选择 Delete（删除）。

在您删除扩展计划后，您的资源不会恢复到其原始容量。例如，如果您的 Auto Scaling 组在您删除扩展计划时扩展到 10 个实例，则您的组在扩展计划删除后仍将扩展到 10 个实例。您可以通过分别访问各个服务的控制台，更新特定资源的容量。

删除 Auto Scaling 组

为了防止您的账户产生 Amazon EC2 费用，您还应删除为本教程创建的 Auto Scaling 组。

有关 step-by-step 说明，请参阅 Amazon EC2 Auto Scaling 用户指南中的删除您的 Auto Scaling 组。

步骤 6：后续步骤

前面您熟悉了扩缩计划以及它的一些功能，下面可以尝试使用 AWS CloudFormation创建自己的扩缩计划模板。

AWS CloudFormation 模板是一个 JSON 或 YAML 格式的文本文件，它描述了运行应用程序或服务所需的 Amazon Web Services 基础设施，以及基础设施组件之间的任何互连。使用 AWS CloudFormation，您可以将一组关联的资源作为堆栈进行部署和管理。 AWS CloudFormation 无需额外付费，而且您只需为运行应用程序所需的 AWS 资源付费。资源可以由您在模板中定义的任何 AWS 资源组成。有关更多信息，请参阅《AWS CloudFormation 开发人员指南》中的 AWS CloudFormation 概念。

在《AWS CloudFormation 用户指南》中，我们提供了一个简单的模板帮助您入门。示例模板可在 AWS CloudFormation 模板参考文档的AWS::AutoScalingPlans::ScalingPlan章节中找到。示例模板为单个 Auto Scaling 组创建扩展计划，并启用预测式扩展和动态扩展。

有关更多信息，请参阅《AWS CloudFormation 用户指南》中的 AWS CloudFormation入门。

迁移您的扩展计划

您可以从扩展计划迁移到 Amazon EC2 Auto Scaling 和 Application Auto Scaling 扩展策略。

迁移过程

> **Important**
>
> 要迁移扩展计划，必须按精确顺序完成多个步骤。迁移扩展计划时，请不要对其进行更新，因为这会破坏操作顺序并可能导致不良行为。

第 1 步：查看您的现有设置

要确定必须移动哪些缩放设置，请使用 desc ribe-scaling-p lans 命令。

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

记下要从现有扩展计划中保留的项目，其中可能包括以下内容：

- **MinCapacity**— 可扩展资源的最小容量。
- **MaxCapacity**— 可扩展资源的最大容量。

- **PredefinedLoadMetricType**— 用于预测性扩展的负载指标。
- **PredefinedScalingMetricType**— 用于目标跟踪（动态）扩展和预测性扩展的扩展指标。
- **TargetValue**— 缩放指标的目标值。

扩展计划和扩展策略之间的区别

扩展计划和扩展策略之间有一些重要的区别：

- 扩展策略只能启用一种扩展类型：目标跟踪扩展或预测扩展。要同时使用这两种扩展方法，必须创建单独的策略。
- 同样，您必须在各自的策略中分别定义预测性扩展的扩展指标和用于目标跟踪扩展的扩展指标。

步骤 2：创建预测性扩展策略

如果您不使用预测缩放，请直接跳至步骤 4：准备删除扩展计划。

为了留出时间来评估预测，我们建议您先创建预测性扩展策略，然后再创建其他扩展策略。

对于任何具有现有负载指标规范的 Auto Scaling 组，请执行以下操作将其转换为基于 Amazon EC2 自动扩展的预测性扩展策略。

创建预测性扩展策略

1. 在 JSON 文件中，定义一个`MetricSpecifications`结构，如以下示例所示：

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

2. 在`MetricSpecifications`结构中，针对扩展计划中的每个负载指标，`CustomizedLoadMetricSpecification`使用扩展计划中的等效设置创建`PredefinedLoadMetricSpecification`或。

   以下是载荷指标部分结构的示例。

With predefined metrics

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

有关更多信息，请参阅 Amazon EC2 Auto Scaling API 参考中的PredictiveScalingPredefinedLoad指标。

With custom metrics

```
{
  "MetricSpecifications":[
    {
      "CustomizedLoadMetricSpecification":{
        "MetricDataQueries":[
          {
            "Id":"load_metric",
            "MetricStat":{
              "Metric":{
                "MetricName":"MyLoadMetric",
                "Namespace":"MyNameSpace",
                "Dimensions":[
                  {
                    "Name":"MyOptionalMetricDimensionName",
                    "Value":"MyOptionalMetricDimensionValue"
                  }
                ]
              },
              "Stat":"Sum"
            }
          }
        ]
      },
      ...
```

```
    }
  ]
}
```

有关更多信息，请参阅 Amazon EC2 Auto Scaling API 参考中的PredictiveScalingCustomizedLoad指标。

3. 将缩放指标规范添加到中MetricSpecifications并定义目标值。

以下是扩展指标和目标值部分的结构示例。

With predefined metrics

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

有关更多信息，请参阅 Amazon EC2 Auto Scaling API 参考中的PredictiveScalingPredefinedScaling指标。

With custom metrics

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

```
                "Namespace":"MyNameSpace",
                "Dimensions":[
                  {
                    "Name":"MyOptionalMetricDimensionName",
                    "Value":"MyOptionalMetricDimensionValue"
                  }
                ]
              },
              "Stat":"Sum"
            }
          }
        ]
      },
      "CustomizedScalingMetricSpecification":{
        "MetricDataQueries":[
          {
            "Id":"scaling_metric",
            "MetricStat":{
              "Metric":{
                "MetricName":"MyUtilizationMetric",
                "Namespace":"MyNameSpace",
                "Dimensions":[
                  {
                    "Name":"MyOptionalMetricDimensionName",
                    "Value":"MyOptionalMetricDimensionValue"
                  }
                ]
              },
              "Stat":"Average"
            }
          }
        ]
      },
      "TargetValue":50
    }
  ],
  ...
}
```

有关更多信息，请参阅 Amazon EC2 Auto Scaling API 参考中的PredictiveScalingCustomizedScaling指标。

4. 要仅进行预测，请添加值为Mode的属性ForecastOnly。在完成预测性缩放迁移并确保预测准确可靠之后，您可以更改模式以允许扩展。有关更多信息，请参阅 步骤 7：重新激活预测扩展。

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

有关更多信息，请参阅 Amazon EC2 Auto Scaling API 参考中的PredictiveScaling配置。

5. 如果您的扩展计划中存在该**ScheduledActionBufferTime**属性，则将其值复制到预测性扩展策略中的SchedulingBufferTime属性中。

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

有关更多信息，请参阅 Amazon EC2 Auto Scaling API 参考中的PredictiveScaling配置。

6. 如果扩展计划中存在**PredictiveScalingMaxCapacityBehavior**和**PredictiveScalingMaxCapacityBuffer**属性，则可以在预测性扩展策略中配置MaxCapacityBreachBehavior和MaxCapacityBuffer属性。这些属性定义了当预测容量接近或超过为 Auto Scaling 组指定的最大容量时应发生的情况。

> **Warning**
>
> 如果将该MaxCapacityBreachBehavior属性设置为IncreaseMaxCapacity，则启动的实例数可能会超过预期值，除非您监控和管理增加的最大容量。增加的最大容量将变为 Auto Scaling 组新的正常最大容量，直到您手动对其进行更新。最大容量不会自动减少到原始的最大容量。

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

有关更多信息，请参阅 Amazon EC2 Auto Scaling API 参考中的PredictiveScaling配置。

7. 使用唯一名称保存 JSON 文件。记下文件名。在下一步中，当你重新激活预测性扩展策略时，你需要它，然后在迁移过程结束时再次需要它。有关更多信息，请参阅 步骤 7：重新激活预测扩展。
8. 保存 JSON 文件后，运行 put-s caling-policy 命令。在以下示例中，将每个#######替换为您自己的信息。

```
aws autoscaling put-scaling-policy --policy-name my-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://my-predictive-scaling-config.json
```

如果成功，此命令将返回策略的 Amazon Resource Name (ARN)。

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

9. 对于要迁移到基于 Amazon EC2 自动扩展的预测性扩展策略的每个负载指标规范，重复这些步骤。

步骤 3：查看预测性扩展策略生成的预测

如果您不使用预测缩放，请跳过以下步骤。

在您创建预测性扩展策略后不久就会提供预测。在 Amazon EC2 Auto Scaling 生成预测后，您可以通过 Amazon EC2 Auto Scaling 控制台查看策略的预测并根据需要进行调整。

查看预测性扩展策略的预测

1. 通过以下网址打开 Amazon EC2 控制台：https://console.aws.amazon.com/ec2/。
2. 在导航窗格中，选择 Auto Scaling Groups，然后从列表中选择您的 Auto Scaling 组的名称。
3. 在自动扩展选项卡的预测扩展策略中，选择您的策略。
4. 在监控部分中，您可以根据实际值查看策略对过去和未来负载和容量的预测。

   有关更多信息，请参阅 Amazon EC2 Auto Scaling 用户指南中的查看预测性扩展监控图表。
5. 对您创建的每个预测性扩展策略重复这些步骤。

步骤 4：准备删除扩展计划

对于任何具有现有目标跟踪扩展配置的资源，请在删除扩展计划之前执行以下操作以从扩展计划中收集所需的任何其他信息。

要描述扩展计划中的扩展策略信息，请使用 desc ribe-scaling-plan- resources 命令。在以下示例命令中，用您自己的信息替换 *my-scaling-plan*。

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

查看输出并确认您要迁移所述的扩展策略。使用此信息在中创建新的 Amazon EC2 Auto Scaling 和基于应用程序自动扩展的目标跟踪扩展策略。步骤 6：重新激活动态缩放

步骤 5：删除扩展计划

在创建新的目标跟踪扩展策略之前，必须删除扩展计划才能删除其创建的扩展策略。

要删除您的扩展计划，请使用 delete-scaling-pl an 命令。在以下示例命令中，用您自己的信息替换 *my-scaling-plan*。

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

删除扩展计划后，动态伸缩将停用。因此，如果流量或工作负载突然激增，则每个可扩展资源的可用容量不会自行增加。作为预防措施，您可能需要在短期内手动增加可扩展资源的容量。

增加 Auto Scaling 组的容量

1. 通过以下网址打开 Amazon EC2 控制台：https://console.aws.amazon.com/ec2/。
2. 在导航窗格中，选择 Auto Scaling Groups，然后从列表中选择您的 Auto Scaling 组的名称。
3. 在 Details（详细信息）选项卡上，选择 Group details（组详细信息）、Edit (编辑)。
4. 对于所需容量，请增加所需容量。
5. 完成后，选择 “更新”。

将 Aurora 副本添加到数据库集群

1. 通过以下网址打开 Amazon RDS 控制台：https://console.aws.amazon.com/rds/。
2. 在导航窗格中，选择数据库，然后选择您的数据库集群。
3. 确保集群和主实例都处于可用状态。
4. 选择 “操作”、“添加读者”。
5. 在添加读取器页面上，为您的新 Aurora 副本指定选项。
6. 选择 “添加阅读器”。

增加 DynamoDB 表或全局二级索引的预配置读取和写入容量

1. 打开 DynamoDB 控制台：https://console.aws.amazon.com/dynamodb/。
2. 在导航窗格中，选择表，然后从列表中选择表的名称。
3. 在 “其他设置” 选项卡上，选择 “读/写容量”、“编辑”。
4. 在编辑读/写容量页面上，对于读取容量、预配置容量单位，增加表的预配置读取容量。
5. （可选）如果您希望全局二级索引使用与基表相同的读取容量设置，请选中 “对所有全局二级索引使用相同的读取容量设置” 复选框。
6. 对于写入容量，即预置容量单位，请增加表的预配置写入容量。
7. （可选）如果您希望全局二级索引使用与基表相同的写入容量设置，请选中 “对所有全局二级索引使用相同的写入容量设置” 复选框。
8. 如果您没有在步骤 5 或 7 中选中复选框，请向下滚动页面以更新所有全局二级索引的读取和写入容量。
9. 选择 “保存更改” 以继续。

增加您的 Amazon ECS 服务的运行任务数

1. 在 https://console.aws.amazon.com/ecs/v2 打开控制台。
2. 在导航窗格中，选择 Clusters，然后从列表中选择您的集群名称。
3. 在 “服务” 部分，选中服务旁边的复选框，然后选择 “更新”。
4. 对于预期任务，请输入要为服务运行的任务数量。
5. 选择更新。

增加 Spot 队列的容量

1. 通过以下网址打开 Amazon EC2 控制台：https://console.aws.amazon.com/ec2/。
2. 在导航窗格中，选择竞价请求，然后选择您的竞价型队列请求。
3. 依次选择 Actions (操作) 和 Modify target capacity (修改目标容量)。
4. 在修改目标容量中，输入新的目标容量和按需实例部分。
5. 选择提交。

步骤 6：重新激活动态缩放

通过创建目标跟踪扩展策略来重新激活动态扩展。

在为 Auto Scaling 组创建目标跟踪扩展策略时，可以将其直接添加到该组中。在为其他可扩展资源创建目标跟踪扩展策略时，首先要将该资源注册为可扩展目标，然后向可扩展目标添加目标跟踪扩展策略。

主题

- 为 Auto Scaling 群组创建目标跟踪扩展策略
- 为其他可扩展资源创建目标跟踪扩展策略

为 Auto Scaling 群组创建目标跟踪扩展策略

为 Auto Scaling 组创建目标跟踪扩展策略

1. 在 JSON 文件中，`CustomizedMetricSpecification`使用扩展计划中的等效设置创建`PredefinedMetricSpecification`或。

以下是目标跟踪配置的示例。在这些示例中，用您自己的信息替换每个########。

With predefined metrics

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

有关更多信息，请参阅 Amazon EC2 Auto Scaling API 参考中的PredefinedMetric规范。

With custom metrics

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

有关更多信息，请参阅 Amazon EC2 Auto Scaling API 参考中的CustomizedMetric规范。

2. 要创建扩展策略，请使用 put-scaling-polic y 命令以及您在上一步中创建的 JSON 文件。在以下示例中，将每个#######替换为您自己的信息。

```
aws autoscaling put-scaling-policy --policy-name my-target-tracking-scaling-policy
 \
  --auto-scaling-group-name my-asg --policy-type TargetTrackingScaling \
  --target-tracking-configuration file://config.json
```

3. 对于要迁移到基于 Amazon EC2 Auto Scaling 的目标跟踪扩展策略的每个基于扩展计划的扩展策略，重复此过程。

为其他可扩展资源创建目标跟踪扩展策略

接下来，通过执行以下配置任务，为其他可扩展资源创建目标跟踪扩展策略。

- 使用 Application Auto Scaling 服务注册用于自动缩放的可扩展目标。
- 在可扩展目标上添加目标跟踪扩缩策略。

为其他可扩展资源创建目标跟踪扩展策略

1. 使用 register-scalable-tar get 命令将资源注册为可扩展目标并定义扩展策略的扩展限制。

   在以下示例中，将每个*#######*替换为您自己的信息。对于命令选项，请提供以下信息：

   - `--service-namespace`— 目标服务的命名空间（例如，**`ecs`**）。要获取服务命名空间，请参阅 Targe RegisterScalablet 参考文档。
   - `--scalable-dimension`— 与目标资源关联的可扩展维度（例如**`ecs:service:DesiredCount`**）。要获得可缩放的维度，请参阅 RegisterScalableTarget 参考文档。
   - `--resource-id`— 目标资源的资源 ID（例如，**`service/`*`my-cluster`*`/`*`my-service`***）。有关特定资源 ID 的语法和示例的信息，请参阅 Targe RegisterScalablet 参考资料。

   ```
   aws application-autoscaling register-scalable-target --service-namespace namespace
    \
     --scalable-dimension dimension \
     --resource-id identifier \
     --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"
   }
   ```

2. 在 JSON 文件中，`CustomizedMetricSpecification`使用扩展计划中的等效设置创建`PredefinedMetricSpecification`或。

   以下是目标跟踪配置的示例。

With predefined metrics

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

有关更多信息，请参阅 App lication Auto Scaling API 参考中的PredefinedMetric规范。

With custom metrics

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

有关更多信息，请参阅 App lication Auto Scaling API 参考中的CustomizedMetric规范。

3. 要创建扩展策略，请使用 put-scaling-polic y 命令以及您在上一步中创建的 JSON 文件。

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

4. 对于要迁移到基于应用程序 Auto Scaling 的目标跟踪扩展策略的每个基于扩展计划的扩展策略，重复此过程。

步骤 7：重新激活预测扩展

如果您不使用预测扩展，请跳过此步骤。

通过将预测缩放切换到预测和缩放来重新激活预测性扩展。

要进行此更改，请更新您在中创建的 JSON 文件，步骤 2：创建预测性扩展策略并将该Mode选项的值更改ForecastAndScale为，如下例所示：

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

然后，使用 put-scaling -policy 命令更新每个预测性扩展策略。在此示例中，用您自己的信息替换每个########。

```
aws autoscaling put-scaling-policy --policy-name my-predictive-scaling-policy \
  --auto-scaling-group-name my-asg --policy-type PredictiveScaling \
  --predictive-scaling-configuration file://my-predictive-scaling-config.json
```

或者，您可以在 Amazon EC2 Auto Scaling 控制台中进行此更改，方法是打开基于预测设置的比例。有关更多信息，请参阅 Amazon EC2 Auto Scaling 用户指南中的 Amazon EC2 Auto Scaling 的预测性扩展。

用于迁移目标跟踪扩展策略的 Amazon EC2 Auto Scaling 参考

为了便于参考，下表列出了扩展计划中的所有目标跟踪配置属性及其在 Amazon EC2 Auto Scaling PutScalingPolicy API 操作中的相应属性。

扩展计划源属性	Amazon EC2 Auto Scaling 目标属性
PolicyName	PolicyName
PolicyType	PolicyType
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Name	TargetTrackingConfiguration.CustomizedMetricSpecification.Dimensions.Name

扩展计划源属性	Amazon EC2 Auto Scaling 目标属性
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Value	TargetTrackingConfiguration.CustomizedMetricSpecification.Dimensions.Value
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.MetricName	TargetTrackingConfiguration.CustomizedMetricSpecification.MetricName
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Namespace	TargetTrackingConfiguration.CustomizedMetricSpecification.Namespace
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Statistic	TargetTrackingConfiguration.CustomizedMetricSpecification.Statistic
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Unit	TargetTrackingConfiguration.CustomizedMetricSpecification.Unit
TargetTrackingConfiguration.DisableScaleIn	TargetTrackingConfiguration.DisableScaleIn
TargetTrackingConfiguration.EstimatedInstanceWarmup	TargetTrackingConfiguration.EstimatedInstanceWarmup [1]
TargetTrackingConfiguration.PredefinedScalingMetricSpecification.PredefinedScalingMetricType	TargetTrackingConfiguration.PredefinedMetricSpecification.PredefinedMetricType
TargetTrackingConfiguration.PredefinedScalingMetricSpecification.ResourceLabel	TargetTrackingConfiguration.PredefinedMetricSpecification.ResourceLabel
TargetTrackingConfiguration.ScaleInCooldown	Not available

扩展计划源属性	Amazon EC2 Auto Scaling 目标属性
`TargetTrackingConfiguration.ScaleOutCooldown`	`Not available`
`TargetTrackingConfiguration.TargetValue`	`TargetTrackingConfiguration.TargetValue`

[1] 实例预热是 Auto Scaling 组的一项功能，它有助于确保新启动的实例在将其使用数据提供给扩展指标之前准备好接收流量。当实例仍在预热时，Amazon EC2 Auto Scaling 会减慢向组中添加或删除实例的过程。我们建议您使用 Auto Scaling 组的默认实例预热设置来确保所有实例启动都使用相同的实例预热时间，而不是为扩展策略指定预热时间。有关更多信息，请参阅《Amazon EC2 Auto Scaling 用户指南》中的设置 Auto Scaling 组的原定设置实例预热。

用于迁移目标跟踪扩展策略的 Auto Scaling 参考

为了便于参考，下表列出了扩展计划中的所有目标跟踪配置属性及其在 Application Auto Scaling `PutScalingPolicy` API 操作中的相应属性。

扩展计划源属性	Application Auto Scaling 目标属性
`PolicyName`	`PolicyName`
`PolicyType`	`PolicyType`
`TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Name`	`TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Dimensions.Name`
`TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Dimensions.Value`	`TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Dimensions.Value`

扩展计划源属性	Application Auto Scaling 目标属性
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.MetricName	TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.MetricName
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Namespace	TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Namespace
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Statistic	TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Statistic
TargetTrackingConfiguration.CustomizedScalingMetricSpecification.Unit	TargetTrackingScalingPolicyConfiguration.CustomizedMetricSpecification.Unit
TargetTrackingConfiguration.DisableScaleIn	TargetTrackingScalingPolicyConfiguration.DisableScaleIn
TargetTrackingConfiguration.EstimatedInstanceWarmup	Not available
TargetTrackingConfiguration.PredefinedScalingMetricSpecification.PredefinedScalingMetricType	TargetTrackingScalingPolicyConfiguration.PredefinedMetricSpecification.PredefinedMetricType
TargetTrackingConfiguration.PredefinedScalingMetricSpecification.ResourceLabel	TargetTrackingScalingPolicyConfiguration.PredefinedMetricSpecification.ResourceLabel
TargetTrackingConfiguration.ScaleInCooldown [1]	TargetTrackingScalingPolicyConfiguration.ScaleInCooldown
TargetTrackingConfiguration.ScaleOutCooldown [1]	TargetTrackingScalingPolicyConfiguration.ScaleOutCooldown

扩展计划源属性	Application Auto Scaling 目标属性
`TargetTrackingConfiguration.TargetValue`	`TargetTrackingScalingPolicyConfiguration.TargetValue`

[1] 当您的可扩展资源横向扩展（增加容量）和向内扩展（减少容量）时，Application Auto Scaling 会使用冷却时间来减慢扩展速度。有关更多信息，请参阅《App lication Auto Scaling 用户指南》中的定义冷却时间。

其他信息

要了解如何通过控制台创建新的预测性扩展策略，请参阅以下主题：

- Amazon EC2 Auto Scalin g — 在 Amazon EC2 Auto Scaling 用户指南中创建预测性扩展策略。

要了解如何使用控制台创建新的目标跟踪扩展策略，请参阅以下主题：

- 亚马逊 Aurora — 亚马逊 RDS 用户指南中的 Amazon Aurora Auto Scaling 与 A urora 副本一起使用。
- DynamoDB — 使用《亚马逊 D AWS Management Console ynamoDB 开发者指南》中的 dynamoDB 自动缩放。
- Amazon EC2 Auto Scalin g — 在 Amazon EC2 Auto Scaling 用户指南中创建目标跟踪扩展策略。
- Amazon ECS — 使用亚马逊弹性容器服务开发者指南中的控制台更新服务。
- Spot 队列 — 使用 A mazon EC2 用户指南中的目标跟踪策略扩展竞价型队列。

扩缩计划的安全性

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。要了解适用的合规计划 AWS Auto Scaling，请参阅AWS 按合规计划划分的范围内AWS 服务（按合分）。
- 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。

此文档可帮助您了解如何在使用扩缩计划时应用责任共担模式，并帮助您了解如何管理扩缩计划的访问权限。

主题

- 使用接口 VPC 终端节点访问扩展计划
- 扩展计划的数据保护
- 扩展计划的身份和访问管理
- 扩展计划的合规性验证
- 扩展计划的基础设施安全

使用接口 VPC 终端节点访问扩展计划

您可以使用 AWS PrivateLink 在您的 VPC 和之间创建私有连接 AWS Auto Scaling。您可以像在 VPC 中 AWS Auto Scaling 一样进行访问，无需使用互联网网关、NAT 设备、VPN 连接或 AWS Direct Connect 连接。VPC 中的实例不需要公有 IP 地址即可访问 AWS Auto Scaling。

您可以通过创建由 AWS PrivateLink提供支持的接口端点来建立此私有连接。我们将在您为接口端点启用的每个子网中创建一个端点网络接口。这些是请求者托管的网络接口，用作发往 AWS Auto Scaling 的流量的入口点。

有关更多信息，请参阅AWS PrivateLink 指南 AWS PrivateLink中的AWS 服务 通过访问。

主题

为扩缩计划创建接口 VPC 终端节点

使用以下服务名称为 AWS Auto Scaling 扩展计划创建终端节点：

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

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。

您无需更改任何其他设置。 AWS Auto Scaling API AWS 服务 使用服务终端节点或私有接口 VPC 终端节点（以正在使用哪个终端节点为准）调用其他终端节点。

为扩缩计划创建 VPC 终端节点策略

您可以将策略附加到您的 VPC 终端节点以控制对 AWS 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计划的权限。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": "autoscaling-plans:DeleteScalingPlan",
```

```
            "Effect": "Deny",
            "Resource": "*",
            "Principal": "*"
        }
    ]
}
```

有关更多信息，请参阅 AWS PrivateLink 指南中的 VPC 端点策略。

终端节点迁移

2019 年 11 月 22 日，我们推出了autoscaling-plans.*region*.amazonaws.com用于调用 AWS Auto Scaling API 的新默认 DNS 主机名和终端节点。新的终端节点与最新版本的 AWS CLI 和软件开发工具包兼容。如果您尚未这样做，请安装最新版本 AWS CLI 和 SDK 以使用新的终端节点。要更新 AWS CLI，请参阅AWS Command Line Interface 用户指南中的 AWS CLI 使用 pip 安装。有关软件开发工具包的信息，请参阅适用于 Amazon Web AWS Services 的工具。

> **Important**
>
> 为了向后兼容，将继续支持调用 AWS Auto Scaling API 的现有autoscaling.*region*.amazonaws.com端点。设置 autoscaling.*region*.amazonaws.com 终端节点作为私有接口 VPC 终端节点，请参阅《Amazon EC2 Auto Scaling 用户指南》中的 Amazon EC2 Auto Scaling 和接口 VPC 终端节点。

使用 CLI 或 AWS Auto Scaling API 时要调用的终端节点

对于的当前版本 AWS Auto Scaling，您对 AWS Auto Scaling API 的调用会自动转到autoscaling-plans.*region*.amazonaws.com终端节点，而不是autoscaling.*region*.amazonaws.com。

通过在每个命令中使用以下参数来指定新的终端节点，可以在 CLI 中调用此终端节点：--endpoint-url https://autoscaling-plans.*region*.amazonaws.com。

您还可以在 CLI 中调用旧终端节点，方法是在每个命令中使用 --endpoint-url https://autoscaling.*region*.amazonaws.com 参数来指定该终端节点，但不建议这样做。

有关用于调用 API 的各种软件开发工具包，请参阅相关软件开发工具包的文档，以了解如何将请求转到特定终端节点。有关更多信息，请参阅用于 Amazon Web Services 的工具。

扩展计划的数据保护

分 AWS 担责任模型适用于中的数据保护 AWS Auto Scaling。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：

- 对每个账户使用多重身份验证 (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）。这包括您使用控制台、API AWS Auto Scaling 或 SDK 或以其他 AWS 服务 方式使用控制台 AWS CLI、API 或 AWS SDK 的情况。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。

扩展计划的身份和访问管理

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 管理员控制谁可以进行身份验证（登录）和授权（有权限）使用 AWS Auto Scaling 资源。您可以使用 IAM AWS 服务 ，无需支付额外费用。

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，但除非您拥有权限，否则您不能创建或访问扩缩计划。例如，您必须具有创建扩展计划、配置预测式扩展等的权限。

以下部分详细说明了 IAM 管理员如何使用 IAM 来控制可以使用扩缩计划的用户，从而帮助保护您的扩缩计划。

主题

扩缩计划如何与 IAM 结合使用

在使用 IAM 管理谁可以创建、访问和管理 AWS Auto Scaling 扩展计划之前，您应该了解扩展计划中可以使用哪些 IAM 功能。

主题

基于身份的策略

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。扩缩计划支持特定的操作、资源和条件键。要了解在 JSON 策略中使用的所有元素，请参阅《IAM 用户指南》中的 IAM JSON 策略元素参考。

操作

管理员可以使用 AWS JSON 策略来指定谁有权访问什么。也就是说，哪个主体 可以对什么资源 执行操作，以及在什么 条件 下执行。

JSON 策略的 Action 元素描述可用于在策略中允许或拒绝访问的操作。策略操作通常与关联的 AWS API 操作同名。有一些例外情况，例如没有匹配 API 操作的仅限权限 操作。还有一些操作需要在策略中执行多个操作。这些附加操作称为相关操作。

在策略中包含操作以授予执行关联操作的权限。

IAM policy 语句中的扩缩计划操作在操作前使用以下前缀：autoscaling-plans:。策略语句必须包含 Action 或 NotAction 元素。扩缩计划有一组自己的操作，这些操作描述了您可以使用此服务执行的任务。

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

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

要查看您可以在策略语句中使用的扩缩计划操作的完整列表，请参阅《服务授权参考》中的 AWS Auto Scaling的操作、资源和条件键。

资源

Resource 元素指定要向其应用操作的对象。

扩缩计划没有可用作 IAM policy 语句的 Resource 元素的服务定义资源。因此没有可在 IAM policy 中使用的 Amazon 资源名称（ARN）。要控制对扩缩计划操作的访问权限，请在编写 IAM policy 时始终使用 *（星号）作为资源。

条件键

在 Condition 元素（或 Condition 块）中，可以指定语句生效的条件。例如，您可能希望策略仅在特定日期后应用。要表示条件，请使用预定义的条件键。

扩缩计划不提供任何服务特定的条件键，但支持使用某些全局条件键。要查看所有 AWS 全局条件键，请参阅 IAM 用户指南中的AWS 全局条件上下文密钥。

Condition 元素是可选的。

示例

要查看基于身份的扩缩计划策略示例，请参阅基于身份的扩缩计划策略示例。

基于资源的策略

其他 Amazon Web Services（如 Amazon Simple Storage Service）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。

扩缩计划不支持基于资源的策略。

访问控制列表 (ACL)

扩缩计划不支持访问控制列表（ACL）。

基于标签的授权

无法标记扩缩计划。此外也没有可以标记的服务定义资源。因此，扩缩计划不支持基于资源中标签的访问控制。

扩缩计划可能包含可标记的资源，例如 Auto Scaling 组，这些资源支持基于标签的访问控制。有关更多信息，请参阅该 AWS 服务的文档。

IAM 角色

IAM 角色是 AWS 账户 中具有特定权限的实体。

使用临时凭证

您可以使用临时凭证进行联合身份登录，来担任 IAM 角色或担任跨账户角色。您可以通过调用 AWS STS API 操作（例如AssumeRole或GetFederation令牌）来获取临时安全证书。

扩缩计划支持使用临时凭证。

扩缩计划的服务相关角色

AWS Auto Scaling 使用服务相关角色获得代表您调用其他 AWS 服务所需的权限。服务相关角色让您能够更轻松地设置扩缩计划，因为您不必手动添加所需的权限。有关更多信息，请参阅《IAM 用户指南》中的使用服务相关角色。

AWS Auto Scaling 当您使用扩展计划时，使用几种类型的服务相关角色 AWS 服务 代表您呼叫其他角色：

- 预测性扩展服务相关角色- AWS Auto Scaling 允许访问来自 CloudWatch的历史指标数据。此角色还允许根据负载预测和容量预测为 Auto Scaling 组创建计划操作。有关更多信息，请参阅 预测性扩缩服务相关角色。
- Amazon EC2 Auto Scaling 服务相关角色 — AWS Auto Scaling 允许访问和管理 Auto Scaling 群组的目标跟踪扩展策略。有关更多信息，请参阅《Amazon EC2 Auto Scaling 用户指南中的 Amazon EC2 Auto Scaling 的服务相关角色。
- Application Auto Scaling 服务相关角色 — 允许 AWS Auto Scaling 访问和管理其他可扩展资源的目标跟踪扩展策略。每个服务都有一个服务相关角色。有关更多信息，请参阅《Application Auto Scaling 用户指南》中的 Application Auto Scaling 服务相关角色。

您可以使用以下过程来确定您的账户是否已经具有服务相关角色。

确定服务相关角色是否已存在

1. 通过以下网址打开 IAM 控制台：https://console.aws.amazon.com/iam/。
2. 在导航窗格中，选择角色。
3. 在列表中搜索 `AWSServiceRole` 以查找您账户中存在的服务相关角色。查找您要检查的服务相关角色的名称。

服务角色

AWS Auto Scaling 没有用于扩展计划的服务角色。

预测性扩缩服务相关角色

AWS Auto Scaling 使用服务相关角色获得在您处理扩展计划时 AWS 代表您呼叫他人所需的权限。有关更多信息，请参阅 扩缩计划的服务相关角色。

以下部分介绍了如何创建和管理预测性扩缩服务相关角色。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。

服务相关角色授予的权限

AWS Auto Scaling 启用预测性扩展时，使用名AutoScaling为 AWSServiceRoleForAutoScalingPlans_EC2 的服务相关角色代表您调用以下操作：

- `cloudwatch:GetMetricData`
- `autoscaling:DescribeAutoScalingGroups`

- autoscaling:DescribeScheduledActions
- autoscaling:BatchPutScheduledUpdateGroupAction
- autoscaling:BatchDeleteScheduledAction

AWSServiceRoleForAutoScalingPlans_EC2 AutoScaling 信任该autoscaling-plans.amazonaws.com服务会担任该角色。

创建服务相关角色（自动）

您无需手动创建 AWSServiceRoleForAutoScalingPlans_EC2 AutoScaling 角色。 AWS 当您在账户中创建扩展计划并启用预测性扩展时，将为您创建此角色。

AWS 要代表您创建服务相关角色，您必须拥有所需的权限。有关更多信息，请参阅《IAM 用户指南》中的服务相关角色权限。

创建服务相关角色（手动）

要手动创建服务相关角色，您可以使用 IAM 控制台、IAM CLI 或 IAM API。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-role CLI 命令创建此服务相关角色。

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

编辑服务相关角色

您可以AutoScaling使用 IAM 编辑 AWSServiceRoleForAutoScalingPlans_EC2 的描述。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

如果您不再需要使用扩展计划，我们建议您删除 AWSServiceRoleForAutoScalingPlans_EC2 AutoScaling。

只有在删除您的 AWS 账户 中启用了预测性扩缩的所有扩缩计划后，才能删除服务相关角色。这可确保您不会无意中删除用于访问您的扩缩计划的权限。

您可以使用 IAM 控制台、IAM CLI 或 IAM API 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。

删除 AWSServiceRoleForAutoScalingPlans_EC2 AutoScaling 服务相关角色后，如果您 AWS Auto Scaling 创建的扩展计划启用了预测性扩展，则会再次创建该角色。

支持的区域

AWS Auto Scaling 支持在所有可用的扩展计划中使用服务相关角色。 AWS 区域 要了解支持扩缩计划的区域，请参阅《AWS 一般参考》中的 AWS Auto Scaling 端点和限额。

基于身份的扩缩计划策略示例

默认情况下，全新的 IAM 用户没有执行任何操作的权限。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）授予使用扩展计划的权限。

要了解如何使用这些示例 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。

主题

- 策略最佳实践
- 允许用户创建扩展计划
- 允许用户启用预测式扩展
- 其他必需的权限
- 创建服务相关角色所需的权限

策略最佳实践

基于身份的策略决定了某人是否可以在您的账户中创建、访问或删除 AWS Auto Scaling 资源。这些操作可能会使 AWS 账户产生成本。创建或编辑基于身份的策略时，请遵循以下准则和建议：

- 开始使用 AWS 托管策略并转向最低权限权限 — 要开始向用户和工作负载授予权限，请使用为许多常见用例授予权限的AWS 托管策略。它们在你的版本中可用 AWS 账户。我们建议您通过定义针对您的用例的 AWS 客户托管策略来进一步减少权限。有关更多信息，请参阅《IAM 用户指南》中的 AWS 托管策略或工作职能的AWS 托管策略。
- 应用最低权限 – 在使用 IAM 策略设置权限时，请仅授予执行任务所需的权限。为此，您可以定义在特定条件下可以对特定资源执行的操作，也称为最低权限许可。有关使用 IAM 应用权限的更多信息，请参阅《IAM 用户指南》中的 IAM 中的策略和权限。
- 使用 IAM 策略中的条件进一步限制访问权限 – 您可以向策略添加条件来限制对操作和资源的访问。例如，您可以编写策略条件来指定必须使用 SSL 发送所有请求。如果服务操作是通过特定的方式使

用的，则也可以使用条件来授予对服务操作的访问权限 AWS 服务，例如 AWS CloudFormation。有关更多信息，请参阅《IAM 用户指南》中的 IAM JSON 策略元素：条件。

- 使用 IAM Access Analyzer 验证您的 IAM 策略，以确保权限的安全性和功能性 – IAM Access Analyzer 会验证新策略和现有策略，以确保策略符合 IAM 策略语言 (JSON) 和 IAM 最佳实践。IAM Access Analyzer 提供 100 多项策略检查和可操作的建议，以帮助您制定安全且功能性强的策略。有关更多信息，请参阅《IAM 用户指南》中的 IAM Acess Analyzer 策略验证。
- 需要多重身份验证 (MFA)-如果 AWS 账户您的场景需要 IAM 用户或根用户，请启用 MFA 以提高安全性。若要在调用 API 操作时需要 MFA，请将 MFA 条件添加到您的策略中。有关更多信息，请参阅《IAM 用户指南》中的 配置受 MFA 保护的 API 访问。

有关 IAM 中的最佳实操的更多信息，请参阅《IAM 用户指南》中的 IAM 中的安全最佳实操。

允许用户创建扩展计划

以下示例所示为用于授予创建扩展计划权限的基于身份的策略。

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

要使用扩缩计划，终端用户必须拥有额外的权限，以允许他们使用其账户中的特定资源。这些权限在其他必需的权限中列出。

每个控制台用户还需要权限，允许他们发现其账户中的可扩展资源并从 AWS Auto Scaling 控制台查看 CloudWatch 指标数据图表。下面列出了使用 AWS Auto Scaling 控制台所需的其他权限集：

- `cloudformation:ListStacks`：列出堆栈。

- tag:GetTagKeys：查找包含特定标签键的可扩展资源。
- tag:GetTagValues：查找包含特定标签值的资源。
- autoscaling:DescribeTags：查找包含特定标签的 Auto Scaling 组。
- cloudwatch:GetMetricData：查看指标图表中的数据。

允许用户启用预测式扩展

以下示例所示为用于授予启用预测式扩展权限的基于身份的策略。这些权限可扩展设置为扩展 Auto Scaling 组的扩展计划的功能。

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

其他必需的权限

要成功配置扩展计划，必须授予最终用户要为其配置扩展的每个目标服务的权限。要授予使用目标服务所需的最低权限，请阅读本章节中的信息，并在 IAM 策略语句的 Action 元素中指定相关操作。

自动扩缩组

要将 Auto Scaling 组添加到扩展计划，用户必须具有来自 Amazon EC2 Auto Scaling 的以下权限：

- autoscaling:UpdateAutoScalingGroup
- autoscaling:DescribeAutoScalingGroups
- autoscaling:PutScalingPolicy
- autoscaling:DescribePolicies

- autoscaling:DeletePolicy

ECS 服务

要将 ECS 服务添加到扩展计划，用户必须具有来自 Amazon ECS 和 Application Auto Scaling 的以下权限：

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

竞价型实例集

要将 Spot 队列添加到扩展计划，用户必须具有来自 Amazon EC2 和 Application Auto Scaling 的以下权限：

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

DynamoDB 表或全局索引

要将 DynamoDB 表或全局索引添加到扩展计划，用户必须具有来自 DynamoDB 和 Application Auto Scaling 的以下权限：

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

Aurora 数据库集群

要将 Aurora 数据库集群添加到扩展计划，用户必须具有来自 Amazon Aurora 和 Application Auto Scaling 的以下权限：

- rds:AddTagsToResource
- rds:CreateDBInstance
- rds:DeleteDBInstance
- rds:DescribeDBClusters
- rds:DescribeDBInstances
- application-autoscaling:RegisterScalableTarget
- application-autoscaling:DescribeScalableTargets
- application-autoscaling:DeregisterScalableTarget
- application-autoscaling:PutScalingPolicy
- application-autoscaling:DescribeScalingPolicies
- application-autoscaling:DeleteScalingPolicy

创建服务相关角色所需的权限

AWS Auto Scaling 当您的任何用户首次创建启用预测性扩展的扩展计划时，需要权限才能 AWS 账户创建服务相关角色。如果服务相关角色尚不存在，请在您的账户中 AWS Auto Scaling 创建该角色。服务相关角色向授予权限， AWS Auto Scaling 以便它可以代表您调用其他服务。

为使自动角色创建操作成功，用户必须具有 iam:CreateServiceLinkedRole 操作的权限。

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

以下示例所示为用于授予创建服务相关角色权限的基于身份的策略。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": "iam:CreateServiceLinkedRole",
            "Resource": "arn:aws:iam::*:role/aws-service-role/autoscaling-
plans.amazonaws.com/AWSServiceRoleForAutoScalingPlans_EC2AutoScaling",
            "Condition": {
                "StringLike": {
                    "iam:AWSServiceName":"autoscaling-plans.amazonaws.com"
                }
            }
        }
    ]
}
```

有关更多信息，请参阅 预测性扩缩服务相关角色。

扩展计划的合规性验证

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，并提供了在这些基础上 AWS 部署以安全性和合规性为重点的基准环境的步骤。
- 在 A@@ mazon Web Services 上构建 HIPAA 安全与合规架构 — 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 合规资源AWS — 此工作簿和指南集可能适用于您所在的行业和所在地区。
- 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。
- Amazon GuardDuty — 它通过监控您的 AWS 账户环境中是否存在可疑和恶意活动，来 AWS 服务检测您的工作负载、容器和数据面临的潜在威胁。 GuardDuty 通过满足某些合规性框架规定的入侵检测要求，可以帮助您满足各种合规性要求，例如 PCI DSS。
- 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。

扩展计划的基础设施安全

作为一项托管服务 AWS 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 AWS Auto Scaling 通过网络进行访问。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。

扩缩计划的配额

您 AWS 账户 拥有与扩展计划相关的默认配额（以前称为限制）。除非另有说明，否则，每个配额都特定于区域。您可以请求增加某些配额，但其他一些配额无法增加。

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 服务并选择 AWS Auto Scaling 计划。

要请求提高限额，请参阅《服务限额用户指南》中的请求提高限额。

您 AWS 账户 有以下与扩展计划相关的配额。

名称	默认值	可调整
每种资源类型的可扩展资源	亚马逊 DynamoDB：3,000 | Amazon EC2 Auto Scaling 群组：200 | 所有其他资源类型：500	是
扩展计划	100	是
每个扩展计划的扩展指令数	500	否
每个扩展指令的目标跟踪配置	10	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。如果需求下降而当前容量减少，则 AWS Auto Scaling 可以再次扩展。为避免再次达到此服务配额限制，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。有关其他亚马逊云科技默认配额的信息，请参阅 Amazon Web Services 一般参考 中的服务端点和配额。

扩缩计划文档历史记录

下表描述了 AWS 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
用于迁移 AWS Auto Scaling 到备选选项的新内容	现在，您可以从迁移 AWS Auto Scaling 到 Amazon EC2 Auto Scaling 预测扩展，它提供了更多功能。有关更多信息，请参阅迁移您的扩展计划。	2024 年 4 月 5 日
新的安全内容	我们发布了更新的安全章节。作为本次更新的一部分，我们将“身份验证和访问控制”替换为的身份和访问管理 AWS Auto Scaling。	2020 年 3 月 12 日
对 Amazon VPC 终端节点的支持	现在，您可以在 VPC 和之间建立私有连接 AWS Auto Scaling。有关迁移注意事项和说明，请参阅扩缩计划和接口 VPC 端点。	2019 年 11 月 22 日
Support 支持将最大容量增加到高于预测容量	添加控制台支持，允许扩展计划按指定的缓冲区值增大高于预测容量的最大容量。有关更多信息，请参阅预测缩放设置。	2019 年 3 月 9 日
预测式扩展和增强	现在，您可以使用预测性扩展来主动扩展您的 Amazon EC2 Auto Scaling 组。此版本还增加了以下支持：替换在扩展计划之外创建的扩展策略（例如来自其他控制台的策略）以及	2018 年 11 月 20 日

	控制是否启用您的计划的动态扩展功能。	
支持自定义资源设置	添加了对每个单独资源或多个资源同时自定义各种设置的支持。	2018 年 10 月 9 日
将标签作为应用程序源	此版本增加了对指定一组标签作为应用程序源的支持。	2018 年 4 月 23 日
新增服务	的初始版本 AWS Auto Scaling.	2018 年 1 月 16 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。