

\*\*\*Unable to locate subtitle\*\*\*

# AWS Well-Architected Framework



# AWS Well-Architected Framework: \*\*\*Unable to locate subtitle\*\*\*

Copyright © 2025 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

# Table of Contents

|              |    |
|--------------|----|
| 摘要和简介 .....  | 1  |
| 简介 .....     | 1  |
| 定义 .....     | 2  |
| 关于架构 .....   | 3  |
| 一般设计原则 ..... | 4  |
| 框架的支柱 .....  | 6  |
| 卓越运营 .....   | 6  |
| 设计原则 .....   | 6  |
| 定义 .....     | 7  |
| 最佳实践 .....   | 8  |
| 资源 .....     | 14 |
| 安全性 .....    | 15 |
| 设计原则 .....   | 15 |
| 定义 .....     | 16 |
| 最佳实践 .....   | 16 |
| 资源 .....     | 21 |
| 可靠性 .....    | 22 |
| 设计原则 .....   | 22 |
| 定义 .....     | 23 |
| 最佳实践 .....   | 23 |
| 资源 .....     | 28 |
| 性能效率 .....   | 28 |
| 设计原则 .....   | 28 |
| 定义 .....     | 29 |
| 最佳实践 .....   | 29 |
| 资源 .....     | 34 |
| 成本优化 .....   | 34 |
| 设计原则 .....   | 35 |
| 定义 .....     | 35 |
| 最佳实践 .....   | 36 |
| 资源 .....     | 40 |
| 可持续性 .....   | 41 |
| 设计原则 .....   | 41 |
| 定义 .....     | 42 |

|                  |     |
|------------------|-----|
| 最佳实践 .....       | 42  |
| 审查流程 .....       | 48  |
| 总结 .....         | 50  |
| 贡献者 .....        | 51  |
| 延伸阅读 .....       | 52  |
| 文档修订 .....       | 53  |
| 附录：问题和最佳实践 ..... | 55  |
| 卓越运营 .....       | 55  |
| 组织 .....         | 55  |
| 准备 .....         | 83  |
| 运营 .....         | 138 |
| 演进 .....         | 166 |
| 安全性 .....        | 180 |
| 安全基础知识 .....     | 180 |
| 身份与访问权限管理 .....  | 195 |
| 检测 .....         | 236 |
| 基础设施保护 .....     | 244 |
| 数据保护 .....       | 259 |
| 事件响应 .....       | 284 |
| 应用程序安全性 .....    | 301 |
| 可靠性 .....        | 317 |
| 基础 .....         | 317 |
| 工作负载架构 .....     | 350 |
| 变更管理 .....       | 386 |
| 故障管理 .....       | 417 |
| 性能效率 .....       | 498 |
| 架构选择 .....       | 499 |
| 计算和硬件 .....      | 510 |
| 数据管理 .....       | 524 |
| 联网和内容分发 .....    | 543 |
| 流程和文化 .....      | 566 |
| 成本优化 .....       | 578 |
| 践行云财务管理 .....    | 578 |
| 支出和使用情况意识 .....  | 597 |
| 具有成本效益的资源 .....  | 630 |
| 管理需求和供应资源 .....  | 661 |



|             |     |
|-------------|-----|
| 持续优化 .....  | 671 |
| 可持续性 .....  | 677 |
| 区域选择 .....  | 677 |
| 符合需求 .....  | 679 |
| 软件和架构 ..... | 690 |
| 数据 .....    | 698 |
| 硬件和服务 ..... | 714 |
| 流程和文化 ..... | 722 |
| 声明 .....    | 729 |

# AWS Well-Architected Framework

发布日期：2023 年 10 月 3 日 ([文档修订](#))

AWS Well-Architected Framework 能够帮助您认识到您在 AWS 上构建系统时所做决策的优缺点。通过使用此框架，您将了解在云中设计和运行可靠、安全、高效、经济实惠且可持续的系统的架构最佳实践。

## 简介

AWS Well-Architected Framework 能够帮助您认识到您在 AWS 上构建系统时所做决策的优缺点。使用该框架有助于您了解在 AWS Cloud 中设计和运行安全、可靠、高效且经济实惠的可持续工作负载的架构最佳实践。它提供了一种方法，使您能够根据最佳实践持续衡量架构，并确定需要改进的方面。审查架构的流程是关于架构决策的建设性对话，不是一种审核机制。我们相信，拥有架构完善的系统能够大大提高实现业务成功的可能性。

AWS 解决方案架构师拥有多年为各种垂直行业和使用案例设计解决方案的经验。我们也已经帮助成千上万客户对其 AWS 之上的架构进行设计与审查。从这些经验中，我们得以总结出在云中设计系统的最佳实践和核心策略。

AWS Well-Architected Framework 囊括了一系列基础性问题，有助于您了解某种架构是否符合云最佳实践。该框架为您提供了一种一致的方法，来对标您所期望的现代云端系统能力，建立一整套质量评估体系，以及评估实现这样的质量需要采取的具体措施。随着 AWS 不断发展，我们将继续与客户协作并增进了解，同时将实际经验融入到 Well-Architected 定义的持续完善当中。

此框架面向各类技术性角色，例如首席技术官 (CTO)、架构师、开发人员和运维团队成员。它介绍了可在设计和运行云工作负载时使用的 AWS 最佳实践和策略，提供了进一步实施细节和架构模式的链接。有关更多信息，请参阅 [AWS Well-Architected 主页](#)。

AWS 还提供可用于审查您的工作负载的免费服务。此 [AWS Well-Architected Tool](#) (AWS WA Tool) 是一种云服务，它提供统一的流程，有助于使用 AWS Well-Architected Framework 对自己的架构进行审核和衡量。借助 AWS WA Tool 提供的建议，可让您的工作负载变得更加可靠、安全、高效和经济实惠。

为了帮助您应用最佳实践，我们创建了 [AWS Well-Architected 实验室](#)，它可以为您提供代码和文档的存储库，让您亲自体验最佳实践的实施。我们还与 AWS 合作伙伴网络 (APN, AWS Partner Network) 中的精选合作伙伴开展合作，他们是 [AWS Well-Architected 合作伙伴计划](#) 的成员。这些 AWS 合作伙伴拥有丰富的 AWS 知识，可以帮助您审查并改进工作负载。

## 定义

AWS 的专家每天都在帮助客户设计系统，以利用云中的最佳实践。在设计过程中，我们与您一起对架构进行权衡调整。当您在真实环境中部署这些系统时，我们将关注这些系统的运作状况，同时衡量上述调整的效果。

依托于实践经验，我们构建了 AWS Well-Architected Framework，它为客户和合作伙伴评估架构提供了一系列最佳实践，并提供了相应的可用于评估架构是否符合 AWS 最佳实践的问题。

AWS Well-Architected Framework 建立在六个支柱的基础上，它们分别是卓越运营、安全性、可靠性、性能效率、成本优化和可持续性。

表 1.AWS Well-Architected Framework 的支柱

| 名称   | 描述                                                                                           |
|------|----------------------------------------------------------------------------------------------|
| 卓越运营 | 能够有效地支持发展和运行工作负载，获取对运营的洞察，以及不断改进支持流程和程序以实现业务价值。                                              |
| 安全性  | 安全性支柱描述了如何利用云技术来保护数据、系统和资产，以改善您的安全状况。                                                        |
| 可靠性  | 可靠性支柱涵盖相关工作负载按照计划正确而稳定执行其预期功能的能力。它包括在其全部生命周期内运行和测试工作负载的能力。本白皮书深度介绍了有关在 AWS 中实施可靠工作负载的最佳实践指导。 |
| 性能效率 | 有效利用计算资源来满足系统要求，并随着需求变化和技术发展保持这种效率的能力。                                                       |
| 成本优化 | 以最低价格运行系统来交付业务价值的能力。                                                                         |
| 可持续性 | 这指的是通过最大程度提高预设资源的回报，同时最大程度减少总体资源需求，提高工作负载各组件的效率以减低能源消耗，从而持续改善可持续性影响的能力。                      |

在 AWS Well-Architected Framework 中，我们使用了以下术语：

- 这里的 **组件** 是指针对相关需求提供的代码、配置和 AWS 资源的组合。组件通常是技术处理单元，与其他组件分离。
- 术语 **工作负载** 指的是共同提供业务价值的组件集合。工作负载通常是业务和技术领导者沟通的细节层次。
- 我们将 **架构** 定义为组件在工作负载中协同工作的方式。架构图的重点通常是组件如何通信和交互。
- **里程碑** 将随着架构在整个产品生命周期内（设计、实施、测试、上线和生产）的演进记录架构中的关键变更。
- 组织内的 **技术产品组合** 是业务运营所需的工作负载集合。
- 这里的 **工作量** 用于对执行任务所需的时间、精力和任务本身的复杂性进行分类。每个组织都需要考虑团队的规模和专业知识和工作负载的复杂性，了解更多背景信息，以便对组织的工作量进行正确分类。
  - **高**：这项工作可能需要数周或数月。它可以分解为若干案例、发布和任务。
  - **中**：这项工作可能需要数天或数周。它可以分解为若干发布和任务。
  - **低**：这项工作可能需要数小时或数天。它可以分解为若干任务。

在设计工作负载时，您会基于您的业务环境在各个支柱之间做出权衡。这些业务决策可以确定设计优先事项。在开发环境中，您可能会进行优化，牺牲一部分可靠性来改进可持续性影响并降低成本；而对于任务关键型解决方案，您可能会在成本和可持续性影响方面做出妥协，来提高可靠性。在电子商务解决方案中，性能可能会影响收入和客户的购买偏好。对于安全性和卓越运营，一般不会对它们和其他支柱之间进行权衡。

## 关于架构

在本地环境中，客户通常有一个技术架构中心团队，来监督其他产品或功能团队，从而确保他们遵循最佳实践。技术架构团队通常包含一组角色，比如技术架构师（基础设施）、解决方案架构师（软件）、数据架构师、网络架构师和安全架构师。这些团队一般将 [TOGAF](#) 或 [Zachman Framework](#) 用作企业架构能力的一部分。

在 AWS，我们倾向于将能力分配到多个团队，而不是只让一个核心团队具有这种能力。当您选择分配决策权限时，会存在一定的风险，例如，确保团队达到内部标准。我们通过两种方法降低这些风险。第一，我们采取（即行为方式、流程、标准和公认的规范），专注于让每个团队都具有这种能力，并且我们通过设立专家团队，来确保团队不断提高以突破新的标准。第二，我们实施了各种机制，来自动执行检查，以确保满足各项标准。

 “徒有良好的心愿没有用，需要良好的机制来实现它们”– 杰夫·贝索斯 ( Jeff Bezos ) 。

这意味着用机制 ( 通常是自动的 ) 来替代人类工作，检查是否遵守了规则或流程。这种分布式方法由[亚马逊的领导力原则提供支持](#)，在所有角色中建立一种 从客户出发 的工作文化。逆向思维是我们的创新过程的基本组成部分。我们从客户和客户需求出发，定义和指导我们的工作。只有以客户为中心的团队才能开发出真正满足客户需求的产品。

对于架构，这意味着我们希望每个团队都有能力创建架构并遵循最佳实践。为了帮助新团队获得这些能力或帮助现有团队提高其标准，我们创建了一个由首席工程师组成的虚拟社群，这些工程师可以检查现有团队的设计，帮助他们了解 AWS 最佳实践。首席工程师社群旨在让您能够接触和了解最佳实践。例如，通过午间谈话交流如何将最佳实践应用到实例中。这些谈话会被记录下来，用作新团队成员入门材料的一部分。

AWS 最佳实践源于我们在互联网规模上运行成千上万个系统的经验。我们倾向于使用数据定义最佳实践，同时我们还通过首席工程师等主题专家来设定最佳实践。当首席工程师发现新的最佳实践时，他们将以社群的形式确保所有团队遵循这些最佳实践。同时，这些最佳实践还会被正式纳入我们的内部审查流程以及强制性合规机制中。Well-Architected Framework 是面向客户实施我们的内部审查流程，其中将我们在不同领域角色 ( 例如解决方案架构和内部工程团队 ) 中的主要设计思维编制成文。Well-Architected Framework 是一种可扩展的机制，使您能够有效利用现有的经验。

通过首席工程师在社群内分散架构责任的方法，我们相信设计良好的企业架构是由客户的需求驱动的，并且可以付诸实现。通过让技术主管 ( 例如首席信息官或开发经理 ) 针对所有工作负载执行良好架构审查，您能够更好地了解技术栈存在的风险。以此方法，您可以确定不同团队间可以使用主题，通过机制、培训或午间谈话等方式，让首席工程师可以与多个团队分享他们在特定领域的想法。

## 一般设计原则

Well-Architected Framework 定义了一系列一般性设计原则，以促进良好的云端设计：

- 停止猜测您的容量需求：如果您在部署工作负载时作出糟糕的容量决策，结果常常造成昂贵的资源闲置或因容量不足而影响性能。利用云计算，这些问题都不复存在。您可以按需使用容量，并自动对容量规模进行扩缩。
- 以生产规模进行系统测试：在云中，您可以根据需要创建一套生产规模等级的测试环境，完成测试，然后停用资源。由于测试环境只需在运行时付费，您模拟真实环境的成本仅为本地测试成本的一小部分。

- 在考虑架构实验的情况下实现自动化：通过自动化操作，您可以低成本创建和复制工作负载，避免人力支出。您可以跟踪自动化变更，审核所产生的影响，并在必要时恢复到以前的参数。
- 考虑架构演进：在传统环境中，架构决策通常作为静态的一次性事件实现，在其生命周期内包含几个重要的系统版本。随着业务及其环境继续演进，这些初始决策可能无法适应不断变化的业务能力需求。在云中，自动化和按需测试能力将显著降低设计变更所带来影响的风险。这使系统能够随时间推移不断演进，以便企业能够不断地发展创新。
- 利用数据驱动架构：在云中，您可以收集有关您的架构选择如何影响工作负载表现的数据。这使您能够基于事实做出如何改进工作负载的决策。您的云基础设施以代码形式存在，因此您可以随着时间的推移，基于这些数据做出明智的架构选择和改进。
- 通过实际演练不断改进：通过定期安排实际演练来模拟生产中的各种事件，测试架构和流程的性能。这将帮助您了解可以从哪些方面作出改进，并有助于培养组织处理各种事件的经验。

# 框架的支柱

构建软件系统与建楼很像。如果基础不牢固，结构问题将会破坏整栋大楼的完整性和功能。在设计技术解决方案时，如果忽视卓越运营、安全性、可靠性、性能效率、成本优化和可持续性这六大支柱，就很难构建一个能够满足您的期望和需求的系统。通过把这些支柱整合到架构中，您将能构建稳定而高效的系统，这将使您能够专注于设计的其他方面，例如功能性需求。

## 要素

- [卓越运营](#)
- [安全性](#)
- [可靠性](#)
- [性能效率](#)
- [成本优化](#)
- [可持续性](#)

## 卓越运营

卓越运营支柱能够有效地支持发展和运行工作负载，获取对运营的洞察，以及不断改进支持流程和程序以实现业务价值。

卓越运营支柱概述了各种设计原则、最佳实践和问题。如需有关具体实施的说明性指导，请参阅 [《卓越运营支柱》白皮书](#)。

## 主题

- [设计原则](#)
- [定义](#)
- [最佳实践](#)
- [资源](#)

## 设计原则

以下是在云中实现卓越运营的设计原则：

- 执行运营即代码：在云中，您可以将用于应用程序代码的工程规范应用于整个环境。您可以将整个工作负载（应用程序、基础设施等）定义为代码，并使用该代码进行更新。您可以为运营流程编写脚



本，并通过启动这些脚本来自动执行流程，以响应事件。通过执行运营即代码，您可以减少人为错误并实现对事件的一致响应。

- 频繁进行小型、可回滚的变更：将工作负载设计为可扩展且松耦合，以允许定期更新组件。自动部署技术加上小型增量变更可缩小影响范围，并能够在发生故障时更快地进行回滚。这将增强您的信心，在保持质量和快速适应市场条件变化的同时，为您的工作负载提供有益的变化。
- 经常完善操作程序：随着工作负载的演变，应相应地改进操作程序。在使用运营程序时，要寻找机会改进它们。定期审查并验证所有流程是否有效，以及团队是否熟悉这些流程。在发现差距时，相应地更新程序。向所有利益相关者和团队传达程序更新。将运营游戏化，以分享最佳实践并向团队传授知识。
- 预测故障：执行“故障演练”，找出潜在的故障源，以便消除或减少故障。测试您的故障场景，并确认您了解相应影响。测试您的响应程序，以确保它们的有效性，以及团队熟练他们的流程。设置定期的实际演练，以测试工作负载和团队对模拟事件的响应。
- 从所有运营故障中吸取经验教训：从所有运营事件和故障中吸取经验教训，推动改进。在多个团队乃至组织范围中分享经验教训。
- 使用托管服务：尽可能使用 AWS 托管服务，减少运营负担。围绕与这些服务的交互制定操作程序。
- 实施可观测性以获得切实可行的见解：全面了解工作负载行为、性能、可靠性、成本和运行状况。建立关键绩效指标（KPI），利用可观测性遥测来作出明智的决策，并在业务结果面临风险时迅速采取行动。基于可操作的可观测性数据，主动提高性能和可靠性，降低成本。

## 定义

在云中实现卓越运营有四个领域的最佳实践：

- 组织
- 准备
- 运营
- 演进

您的组织领导层负责定义业务目标。您的组织必须了解各种要求和重点，并利用它们来组织和开展工作，从而为获得业务成果提供支持。您的工作负载必须发出所需信息以提供支持。采用多种服务来实现工作负载的集成、部署和交付，这将通过自动执行重复流程，为生产环境带来更多有益更改。

工作负载的运营可能存在固有风险。了解这些风险并作出明智的生产决策。您的团队必须能够支持您的工作负载。从预期业务成果中得出的业务和运营指标将有助于您了解工作负载的运行状况、运营活动以



及对事件的响应。您的重点将随着您的业务需求和业务环境的变化而变化。将这些作为反馈循环，持续推动组织和工作负载运营的改进。

## 最佳实践

### 主题

- [组织](#)
- [准备](#)
- [运营](#)
- [演进](#)

### 组织

您的团队必须对整个人工负载、他们在其中的角色以及共同的业务目标有一致的理解，以便确立运营重点以实现业务成功。明确运营重点可以让您的工作效益最大化。评估内部和外部客户需求，让包括业务、开发和运营团队在内的主要利益相关方参与进来，以便确定工作重心。评估客户需求将确保您充分了解实现业务成果所需的支持。确保了解组织监管规定的指导原则或义务，以及监管合规性要求和行业标准等可能需要遵循或重视的外部因素。验证您是否具有确定内部监管和外部合规性要求更改的机制。如果未确定要求，请确保您已对此决定进行尽职调查。定期审查您的运营重点，以便在需求发生变化时对其进行更新。

评估业务面临的威胁（例如业务风险和负债以及信息安全威胁），并在风险注册表中维护这些信息。评估风险的影响，在有冲突的利益或替代方法之间做出权衡。例如，新功能的加速上市可能会比成本优化更重要，或者您可以为非关系数据选择关系数据库来简化系统迁移工作，而无需重构。管理收益和风险，以便在确定工作重心时作出明智的决策。有些风险或选择可能在一段时间内可以接受，或许可以降低相关风险，或者允许风险继续存在可能会令人无法接受，在这种情况下，您将采取措施来化解风险。

您的团队必须了解他们在实现业务成果方面所发挥的作用。团队必须了解自己在其他团队获得成功的过程中所扮演的角色、其他团队在他们获得成功的过程中所扮演的角色，并设定共同的目标。了解责任分配、所有权归属、决策制定方式以及决策者将有助于集中精力，最大限度地发挥团队的优势。团队的需求将由其所支持的客户、所在组织、团队的组成以及工作负载的特征决定。期望单个运营模式能够支持组织中的所有团队及其工作负载是不合理的。

确保每个应用程序、工作负载、平台和基础设施组件都有确定的负责人，并且每个流程和程序都有确定的负责人负责其定义，有负责人负责其性能。

了解每个组件、流程和程序的业务价值，了解为什么要配置这些资源或为什么要执行这些活动，以及为什么要拥有该所有权，这些都有助于确定团队成员的行动。清晰定义团队成员的责任以便他们可以适当

地采取行动，并制定相关机制，确定责任和所有权。制定用于请求添加、更改和例外的机制，以免限制创新。在团队之间定义协议，描述团队之间如何开展合作以相互支持以及您的业务成果。

为您的团队成员提供支持，以便他们可以更有效地采取行动并为您的业务成果提供支持。参与其中的高层领导应设定期望并衡量是否成功。高层领导应是采用最佳实践和组织发展的发起人、倡导者和推动者。允许团队成员在成果面临风险时采取行动以尽可能减少影响，并鼓励他们在认为存在风险时向决策者和利益相关者上报，以便解决问题并避免事件。及时、清晰、可行地传达已知风险和计划内事件，以便团队成员可以及时采取适当行动。

鼓励进行试验，以加快学习速度，并使团队成员保持兴趣和参与热情。团队必须增强自己的技能组合，以采用新技术，并随需求和责任的变化继续提供支持。专门安排学习时间，以提供支持并鼓励参与其中。确保您的团队成员拥有取得成功所需的资源（包括工具和团队成员），并具有支持您的业务成果的规模。利用跨组织的多样性来寻求多种独特的见解。利用这种见解提高创新能力、对您的假设提出质疑，并降低确认偏差的风险。在团队内部提升包容性、多样性和可达性有助于获取有益的见解。

如果存在适用于您组织的外部法规或合规性要求，则应使用 [AWS Cloud 合规性](#) 提供的资源来协助培训您的团队，以便他们能够确定运营重点会受到的影响。Well-Architected Framework 强调学习、衡量和改进。它为您提供了一种一致的方法来评估架构，并实施将随着时间推移而扩展的设计。AWS 提供了 AWS Well-Architected Tool，可协助您在开发之前审查方法，在生产之前审查工作负载状态，以及在生产过程中审查工作负载状态。您可以将其与最新的 AWS 架构最佳实践进行比较，监控整体状态，并深入了解潜在风险。AWS Trusted Advisor 是一种工具，让您可以访问一组核心检查，这些检查会提出优化建议，有助于确定您的运维重点。商业支持和企业支持客户可以访问其他检查，这些检查重点关注安全性、可靠性、性能、成本优化和可持续性，可进一步协助他们确定运营重点。

AWS 有助于您向团队介绍 AWS 及其服务，让他们深入了解自己的选择会如何影响工作负载。使用由 AWS 支持 提供的资源（AWS 知识中心、AWS 开发论坛和 AWS 支持 中心）和 AWS 文档来培训您的团队。请通过 AWS 支持 中心联系 AWS 支持，获取与 AWS 问题有关的协助。AWS 还分享了我们通过在 Amazon Builders' Library 中的 AWS 运维学到的最佳实践和模式。您可以通过 AWS 博客和 AWS 官方播客，获得各种其他有用信息。AWS Training and Certification 提供了一些培训，可以通过自定进度的数字课程，学习 AWS 的基础知识。您还可以报名参加讲师指导培训，进一步培养您团队的 AWS 技能。

使用能够跨 AWS Organizations 等账户集中监管环境的工具或服务，来协助管理运营模式。AWS Control Tower 等服务扩展了这一管理功能，使您能够定义账户设置的蓝图（支持您的运营模式），使用 AWS Organizations 进行持续监管以及自动预置新账户。托管服务提供商（如 AWS Managed Services、AWS Managed Services 合作伙伴或 AWS 合作伙伴网络中的托管服务提供商）会提供实施云环境的专业知识，并为您的安全性和合规性要求以及业务目标提供支持。将托管服务添加到您的运营模式可以节省您的时间和资源，并使您的内部团队保持精干，专注于凸显业务优势的战略成果，而不是开发新的技能和功能。

以下问题主要针对卓越运营的准备阶段。（有关卓越运营问题的列表和最佳实践，请参阅 [Appendix](#)。）

#### OPS 1: How do you determine what your priorities are?

Everyone must understand their part in achieving business success. Have shared goals in order to set priorities for resources. This will maximize the benefits of your efforts.

#### OPS 2: How do you structure your organization to support your business outcomes?

Your teams must understand their part in achieving business outcomes. Teams must understand their roles in the success of other teams, the role of other teams in their success, and have shared goals. Understanding responsibility, ownership, how decisions are made, and who has authority to make decisions will help focus efforts and maximize the benefits from your teams.

#### OPS 3: How does your organizational culture support your business outcomes?

Provide support for your team members so that they can be more effective in taking action and supporting your business outcome.

您可能会发现，您需要在某个时间点侧重点于一小部分运营重点。长期使用平衡的方法来确保培养所需能力和管理风险。定期回顾运营重点，并根据需求变化进行更新。当责任和所有权不确定或未知时，您将面临以下风险：没有及时执行必要的活动，以及在处理这些需求时可能出现工作冗余和潜在冲突。组织文化会直接影响团队成员的工作满意度和保留率。提升团队成员的参与度和能力，取得业务成功。创新必须进行试验，才能将创意转化为成果。应认识到，取得非预期结果也算试验成功，因为这种试验发现了无法实现成功的途径。

## 准备

要为卓越运营做好准备，您必须了解您的工作负载及其预期行为。然后，您需要能够针对它们进行设计，以提供对其状态的洞察并构建程序以提供支持。

将工作负载设计成能够提供必要的信息，以便您了解其所有组件的内部状态（例如指标、日志、事件和跟踪信息），为可观测性和调查问题提供支持。可观测性不仅仅是简单的监控，它让您可以根据系统的外部输出全面了解系统的内部运作。可观测性源于指标、日志和跟踪数据，可提供对系统行为和动态的

深刻见解。通过有效的可观测性，团队可以识别模式、异常和趋势，从而能够主动解决潜在问题并保持最佳系统运行状况。要想确保监控活动与业务目标协调一致，确定关键绩效指标 ( KPI ) 至关重要。这种一致性可确保团队使用真正重要的指标作出数据驱动型决策，从而优化系统性能和业务成果。此外，可观测性使企业能够积极采取行动，而不是被动做出反应。团队可以了解其系统中的因果关系，以此预测和预防问题，而不仅仅是对问题做出反应。随着工作负载的变化，必须重新审视和完善可观测性策略，确保其仍然适用且有效。

采用改进生产调整流程并支持重构、快速质量反馈和错误修复的方法。这些方法可以加快有益更改进入生产环境的速度、减少产生的问题，并能够快速识别和修复通过部署活动引入的问题或在环境中发现的问题。

采用可提供快速质量反馈，并在更改没有达到目标成效时实现快速恢复的方法。使用这些实践可以减轻因部署更改而产生的问题的影响。制定计划以防更改不成功，这样在必要时能够更快速的响应，并测试和验证所做的更改。了解环境中的计划活动，以便管理更改风险，避免影响计划活动。强调频繁、小规模、可逆更改，以限制更改范围。这样可以加快故障排除和修复速度，并支持回滚更改。此外，还意味着能够更频繁地从有价值的更改中获益。

评估工作负载、流程和程序以及工作人员的运营准备就绪情况，以了解与工作负载相关的运营风险。使用一致的流程（包括手动或自动化检查清单）来了解何时可运营工作负载或进行更改。这也有助于您发现必须制定计划予以解决的任何问题。准备好记录日常活动的运行手册和指导问题解决流程的行动手册。了解收益和风险，以便作出明智的决策，从而将更改应用于生产环境。

AWS 使您能够将整个工作负载（应用程序、基础设施、策略、监管和运维）视为代码。这意味着，您可以将用于应用程序代码的工程规范应用于堆栈的每个元素，并在团队或组织之间共享，提高开发工作的效益。使用云中的运营即代码功能和安全测试功能开发工作负载、运营流程和故障演练。使用 AWS CloudFormation，您可以实现一致的模板化沙盒开发、测试和生产环境，提高运营管理水平。

以下问题主要针对卓越运营的准备阶段。

#### OPS 4: How do you implement observability in your workload?

Implement observability in your workload so that you can understand its state and make data-driven decisions based on business requirements.

#### OPS 5: How do you reduce defects, ease remediation, and improve flow into production?

Adopt approaches that improve flow of changes into production that achieve refactoring fast feedback on quality, and bug fixing. These accelerate beneficial changes entering production, limit

## OPS 5: How do you reduce defects, ease remediation, and improve flow into production?

issues deployed, and achieve rapid identification and remediation of issues introduced through deployment activities.

## OPS 6: How do you mitigate deployment risks?

Adopt approaches that provide fast feedback on quality and achieve rapid recovery from changes that do not have desired outcomes. Using these practices mitigates the impact of issues introduced through the deployment of changes.

## OPS 7: How do you know that you are ready to support a workload?

Evaluate the operational readiness of your workload, processes and procedures, and personnel to understand the operational risks related to your workload.

对代码化运营进行投资，以最大限度地提高运营人员的工作效率，最大限度地降低错误率，并实现自动响应。使用“故障演练”来预测故障，并在适当的时候创建程序。使用资源标签和 AWS Resource Groups，按照一致的标记策略应用元数据，以标识您的资源。标记您的资源，以便进行整理、成本核算、访问控制并有针对性地自动执行运营活动。利用云的弹性特点结合相应部署实践，来推动开发活动和系统的预部署，以加快部署速度。当您对用于评估工作负载的检查清单进行更改时，请计划要对不再符合条件的活动系统执行哪些操作。

## 运营

可观测性使您可以专注于有意义的数据，并了解工作负载的交互和输出。通过专注于基本见解并消除不必要的的数据，您可以直接了当地了解工作负载性能。这不仅对收集数据至关重要，对正确解读数据也至关重要。定义明确的基准，设置适当的警报阈值，并主动监控任何偏差。关键指标的改变，尤其是与其他数据关联时，可以精确定位特定的问题领域。借助可观测性，您可以更好地预见和应对潜在挑战，从而确保您的工作负载平稳运行并满足业务需求。

工作负载运营是否成功通过业务成果和客户结果的实现情况加以衡量。定义预期结果、确定成功的衡量方式，并确定将在这些计算中使用的指标，以确定工作负载和运营是否成功。运营状况包括工作负载的运行状况，以及为支持工作负载而执行的操作的运行状况和成败（例如，部署和事件响应）。设立改



进、调查和介入的指标基线，收集和分析您的指标，然后验证您对运营成功的理解及其随时间变化的规律。使用收集的指标来确定您是否可以满足客户需求和业务需求，并确定需要改进的领域。

要实现卓越运营，您需要进行有效且高效的运营事件管理。这适用于计划内和计划外的运营事件。使用已确定的运行手册解决易于理解的事件，并使用行动手册来帮助调查和解决问题。您需要根据事件对业务和客户的影响排定其优先级。确保在出现事件警报时，会有指定负责人运行相关流程。事先定义解决事件所需的人员，并配备一个上报流程，以便根据紧急程度和影响在必要时引入额外人员。确定并引入有权决定行动方案的人员，这些行动方案将对之前未解决的事件响应产生业务影响。

通过为目标受众（例如，客户、业务人员、开发人员、运营人员）定制的控制面板和通知来发布工作负载的运行状态，以便他们可以采取相应措施、管理预期，并在恢复正常运营时收到通知。

在 AWS 中，您可以为收集的工作负载指标和 AWS 自带指标生成控制面板视图。您可以利用 CloudWatch 或第三方应用程序来汇总和呈现运维活动的业务、工作负载和运营级别视图。AWS 通过日志记录功能（包括 AWS X-Ray、CloudWatch、CloudTrail 和 VPC 流日志）提供工作负载洞察，从而协助识别工作负载问题，以支持根本原因分析和修复。

以下问题主要针对卓越运营的准备阶段。

#### OPS 8: How do you utilize workload observability in your organization?

Ensure optimal workload health by leveraging observability. Utilize relevant metrics, logs, and traces to gain a comprehensive view of your workload's performance and address issues efficiently.

#### OPS 9: How do you understand the health of your operations?

Define, capture, and analyze operations metrics to gain visibility to operations events so that you can take appropriate action.

#### OPS 10: How do you manage workload and operations events?

Prepare and validate procedures for responding to events to minimize their disruption to your workload.

您收集的所有指标都应该与业务需求及其支持的结果相符。为充分理解的事件开发脚本式响应，并自动执行响应以识别事件。

## 演进

学习、分享和不断改进，以保持卓越运营。将工作周期专用于持续进行渐进式改进。对影响客户的所有事件执行事后分析。确定导致这些事件的因素和预防措施，以限制或防止再次发生。根据需要与受影响的团体沟通导致这些事件的因素。定期评估并优先处理改进机会（例如，功能请求、问题修复和合规性要求），包括工作负载和运营程序。

将反馈周期纳入您的流程，以快速确定需要改进的领域，并从运营执行中获取经验教训。

在团队中分享得到的经验教训，并从中受益。分析经验教训中的趋势，并对运营指标进行跨团队回顾性分析，以确定改进的机会和方法。实施改进措施，并评估结果以确定是否成功。

在 AWS 上，您可以将日志数据导出到 Amazon S3 或将日志直接发送到 Amazon S3 以便长期存储。使用 AWS Glue，您可以在 Amazon S3 中发现并准备您的日志数据以供分析，并将相关元数据存储在 AWS Glue Data Catalog 中。然后，通过 Amazon Athena 与 AWS Glue 的原生集成，可以对您的日志数据进行分析，并使用标准 SQL 查询日志数据。使用像 Amazon QuickSight 这样的商业智能工具，您可以直观显示、浏览和分析您的数据。发现可能推动改进的相关趋势和活动。

以下问题主要针对卓越运营方面的注意事项。

### OPS 11: How do you evolve operations?

Dedicate time and resources for nearly continuous incremental improvement to evolve the effectiveness and efficiency of your operations.

运营的成功演进建立在以下基础上：频繁的小规模改进；提供安全的环境和时间来试验、开发和测试改进；以及鼓励人们从失败中获取经验教训的整体氛围。随着运营控制水平的提高，对于沙盒、开发、测试和生产环境的运营支持促进了开发，并提高了对生产环境中部署的变更结果成功与否的可预测性。

## 资源

请参阅以下资源，详细了解卓越运营的最佳实践。

## 文档

- [DevOps 和 AWS](#)

## 白皮书

- [卓越运营支柱](#)

## 视频

- [Amazon DevOps](#)

## 安全性

安全性支柱包括保护数据、系统和资产以利用云技术来改善安全性的能力。

安全性支柱概述了设计原则、最佳实践和问题。如需有关具体实施的说明性指导，请参阅 [《安全性支柱》白皮书](#)。

### 主题

- [设计原则](#)
- [定义](#)
- [最佳实践](#)
- [资源](#)

## 设计原则

在云中实现安全性有七个设计原则：

- 健壮的身份验证体系：实施最小权限原则，并通过对每一次与 AWS 资源之间的交互进行适当授权来强制执行职责分离。集中进行身份管理，并努力消除对长期静态凭证的依赖。
- 实现可追溯性：实时监控和审计对环境执行的操作和更改并发送警报。为系统集成日志和指标收集功能，以自动调查并采取措施。
- 在所有层面应用安全措施：利用多种安全控制措施实现深度防御。应用到所有层面（例如网络边缘、VPC、负载均衡、每个实例和计算服务、操作系统、应用程序和代码）。
- 自动实施安全最佳实践：借助基于软件的自动化安全机制，您能够以更为快速且更具成本效益的方式实现安全扩展。创建安全架构，包括实施可在版本控制模板中以代码形式定义和管理的控制措施。
- 保护动态数据和静态数据：将您的数据按敏感程度进行分类，并采用加密、令牌和访问控制等机制（如适用）。



- 限制对数据的访问：使用相关机制和工具来减少和消除直接访问或人工处理数据的需求。这样可以降低处理敏感数据时数据处理不当、被修改以及人为错误的风险。
- 做好应对安全性事件的准备：制定符合您组织要求的事件管理和调查策略和流程，做好应对事件的准备工作。开展事件响应模拟演练并使用具有自动化功能的工具来提高检测、调查和恢复的速度。

## 定义

在云中实现安全性包括六个方面的最佳实践：

- 安全性
- 身份和权限管理
- 检测
- 基础设施保护
- 数据保护
- 事件响应

在为任何工作负载设计架构之前，您需要确定可能影响安全性的实践。您需要控制谁可以执行什么操作。另外，您希望能够识别安全事件、保护您的系统和服务，并通过数据保护机制来保持数据的机密性和完整性。您应该具备一个定义明确且经过实践的流程来响应安全事件。这些工具和方法非常重要，因为它们有助于实现诸如避免财务损失或遵循法律与合规性要求等一系列目标。

借助 AWS 责任共担模式，组织能够利用云服务实现其安全性和合规性目标。AWS 负责保护用于支持云服务的基础设施，作为 AWS 的客户，您能够专注于使用云上的各种服务来实现您的目标。还可通过 AWS Cloud 更好地访问安全数据，并以自动化方式响应安全性事件。

## 最佳实践

主题

- [安全性](#)
- [身份与权限管控](#)
- [检测](#)
- [基础设施保护](#)
- [数据保护](#)
- [事件响应](#)

## 安全性

为了安全地操作您的工作负载，您必须对安全性的各个方面应用总体最佳实践。采用您在组织和工作负载层面的卓越运营中定义的要求和流程，并将它们应用到各个方面。

及时了解最新的 AWS、行业建议以及威胁情报信息可帮助您改进您的威胁模型和控制目标。实现安全流程、测试和验证的自动化可扩展您的安全运营。

以下问题主要针对安全方面的注意事项。（有关安全性问题的列表和最佳实践，请参阅 [附录](#)）。

### SEC 1：如何安全地操作您的工作负载？

为了安全地操作您的工作负载，您必须对安全性的各个方面应用总体最佳实践。采用您在组织和工作负载层面的卓越运营中定义的要求和流程，并将它们应用到各个方面。及时了解来自 AWS 的建议、行业资源以及威胁情报信息可帮助您改进您的威胁模型和控制目标。实现安全流程、测试和验证的自动化可扩展您的安全运营。

在 AWS 中，建议根据账户的功能和合规性或数据敏感性要求分离不同的工作负载。

## 身份与权限管控

身份识别与访问管理是信息安全计划的关键部分，可以确保只有经过授权和通过身份验证的用户和组件才能访问您的资源，并且只能以您要求的方式进行访问。例如，您需要定义一些主体（即可以在您的账户中执行操作的账户、用户、角色和服务）、创建与这些主体相匹配的策略，并实施严格的凭证管理。这些权限管理元素构成了身份验证和授权的核心。

在 AWS 中，权限管理主要通过 AWS Identity and Access Management (IAM) 服务来实现，您可以使用该服务控制对 AWS 服务和资源的用户和编程访问。您需要应用细粒度的策略向用户、组、角色或资源分配权限。您还可以应用强密码原则（例如复杂程度）来避免重复使用并强制执行多重验证 (MFA)。您可以将联合身份验证与现有的目录服务配合使用。对于需要系统接入 AWS 的工作负载，IAM 可以通过角色、实例配置文件、身份联合验证和临时凭证进行安全访问。

以下问题主要针对安全方面的注意事项。

### SEC 2：如何管理人员和机器的身份？

在访问和运行安全的 AWS 工作负载时，您需要管理两种类型的身份。了解管理和授予访问权限所需的身份类型，这有助于确保正确的身份能够在正确的条件下访问正确的资源。

## SEC 2：如何管理人员和机器的身份？

**人员身份：**您的管理员、开发人员、操作员和最终用户需要确定身份才能访问您的 AWS 环境和应用程序。这些是您的组织成员或您与之协作的外部用户，以及通过 Web 浏览器、客户端应用程序或交互式命令行工具与您的 AWS 资源交互的用户。

**机器身份：**您的服务应用程序、操作工具和工作负载需要一个身份来向 AWS 服务发出请求，例如，读取数据。这些身份包括在 AWS 环境中运行的机器，例如 Amazon EC2 实例或 AWS Lambda 函数。您还可以管理需要访问权限的外部各方的机器身份。此外，您可能还有需要访问您 AWS 环境的 AWS 之外的机器。

## SEC 3：如何管理人员和机器的权限？

管理权限以控制对需要访问 AWS 和您的工作负载的人员和机器身份的访问。权限用于控制哪些人可以在什么条件下访问哪些内容。

凭证不得与任何用户或系统共享。应使用最小权限原则授予用户访问权限，并采用密码规则和强制执行 MFA 等最佳实践。应使用临时凭证和有限权限凭证（例如 AWS Security Token Service 发放的凭证）来执行程序访问（包括对 AWS 服务的 API 调用）。

AWS 提供了能够帮助您使用 Identity and Access Management 的资源。为了帮助学习最佳实践，请探索我们的 [管理凭证和身份验证](#)，[控制人员访问](#)和[控制程序访问的相关动手实验](#)。

## 检测

您可以使用检测控制来识别潜在的安全威胁或事件。检测控制是管理框架的重要组成部分，并且可以用于支持质量流程、法律或合规，还可以用于威胁识别和响应工作。检测控制分为多种不同类型。例如，编制资产清单及其详细属性有助于更有效地做出决策（以及进行生命周期管理），从而有助于建立运营基准。您可以通过内部审计（是指对信息系统相关的控制措施进行的检查）来确保实践符合策略和要求，并确保您已根据定义的条件设置了正确的自动告警通知。这些控制措施都是重要的响应手段，可以帮助您的组织识别和了解异常活动的范围。

在 AWS 中，您可以通过处理可用于审计、自动化分析和触发警报的日志、事件以及监控来实施检测控制。CloudTrail 日志、AWS API 调用和 CloudWatch 可以提供对指标进行监控以及报警的功能，AWS Config 可以提供配置历史记录。Amazon GuardDuty 是一种托管的威胁检测服务，可以持续监控恶意或未经授权的行为，从而帮助您保护您的 AWS 账户和工作负载。您还可以使用服务级别日志，例如，您可以使用 Amazon Simple Storage Service (Amazon S3) 来记录访问请求。

以下问题主要针对安全方面的注意事项。

## SEC 4：您如何检测和调查安全事件？

通过日志和指标来记录和分析事件，以便了解信息。针对安全事件和潜在的威胁采取措施，以便保护您的工作负载。

日志管理对于架构完善的工作负载至关重要，这其中原因众多，包括安全性或取证、法律或法规要求。分析日志并相应地做出响应至关重要，这样您能够识别潜在的安全事件。借助 AWS 提供的功能，您能够定义数据保留生命周期或定义数据保存、存档或最终删除的位置，从而更轻松地管理日志。这样，您就能够以更为简单且更具成本效益的方式进行可预测且可靠的数据处理。

## 基础设施保护

基础设施保护包括满足最佳实践和组织、法律及监管义务所必需的控制方法（例如深度防御）。使用这些方法对于在云中或本地持续成功运营是至关重要的。

在 AWS 中，您可以通过使用 AWS 原生技术或使用 AWS Marketplace 提供的合作伙伴产品和服务来进行有状态和无状态数据包检查。您还可使用 Amazon Virtual Private Cloud (Amazon VPC) 创建一个安全且可扩展的私有环境，您可以在其中定义拓扑结构，包括网关、路由表以及公有子网和私有子网。

以下问题主要针对安全方面的注意事项。

## SEC 5：如何保护您的网络资源？

任何以某种形式连接至网络的工作负载（互联网或私有网络）都需要多层防御，以帮助防御基于外部和内部网络的威胁。

## SEC 6：如何保护计算资源？

工作负载内的计算资源需要采用多层防御，才有助于免受内部和外部威胁。计算资源包括 EC2 实例、容器、AWS Lambda 函数、数据库服务、IoT 设备等。

在任何类型的环境，我们都建议使用多层防御。在基础设施保护方面，许多概念和方法在跨云和本地模型中都有效。实施边界保护、监控入站点和出站点以及建立全面的日志记录、监控和告警机制对于制定有效的信息安全计划至关重要。

AWS 客户能够定制或加强 Amazon Elastic Compute Cloud ( Amazon EC2 )、Amazon Elastic Container Service ( Amazon ECS ) 容器或 AWS Elastic Beanstalk 实例的配置，并将配置保存到不可变的亚马逊云机器镜像 ( AMI , Amazon Machine Image )。之后，无论是由 Auto Scaling 触发还是手动启动，使用此 AMI 启动的所有新虚拟服务器 ( 实例 ) 都会收到上述加强的配置。

## 数据保护

在为任何系统设计架构之前，您应确定可能影响安全性的基本实践。例如，数据分级提供了一种基于敏感程度对组织数据进行分类的方法，加密通过让未经授权的用户无法获知数据的真正内容来保护数据。这些工具和方法非常重要，因为它们有助于实现诸如避免财务损失或遵循法律与合规性要求等一系列目标。

在 AWS 中，以下实践有助于保护数据：

- 作为 AWS 客户，您拥有对自己的数据的完全控制权。
- AWS 可帮助您更轻松地进行加密和管理密钥（包括定期密钥轮换），这些操作可以由 AWS 轻松自动执行，也可由您执行。
- 我们还提供包含文件访问和更改等重要内容的详细日志记录。
- AWS 设计的存储系统具有优异的弹性。例如，Amazon S3 标准、S3 标准 – IA、S3 单区 – IA 和 Amazon Glacier 都设计为可以在一年内实现 99.99999999% 的对象持久性。这一持久性级别相当于平均每年有 0.000000001% 的数据对象丢失。
- 作为较大规模数据生命周期管理流程中的一部分，版本控制可以防止意外覆盖、删除数据和类似损害。
- AWS 永远不会主动在区域之间移动数据。除非您明确启用相关功能或利用提供该功能的服务移动数据，否则放置在某个区域中的内容将保留在该区域中。

以下问题主要针对安全方面的注意事项。

### SEC 7：如何对数据进行分类？

分类提供了一种基于关键性和敏感度对数据进行分类的方法，以帮助您确定适当的保护和保留控制措施。

### SEC 8：如何保护静态数据？

通过实施多个控制措施来保护静态数据，以降低未经授权的访问或处理不当带来的风险。

## SEC 9：如何保护传输中的数据？

通过实施多个控制措施来保护传输中的数据，以降低未经授权的访问或数据丢失所带来的风险。

AWS 提供了多种加密静态数据和传输中数据的方法。我们将这些功能内置在我们的服务中，这样您就可以更轻松地进行加密。例如，我们为 Amazon S3 实施了服务器端加密 (SSE)，这样您就可以更轻松地进行加密。您还可以将整个 HTTPS 加密和解密过程（通常称为 SSL 终端）交给 Elastic Load Balancing (ELB) 来完成。

### 事件响应

即使采用极为成熟的预防和检测控制机制，您的组织仍应制定相关流程来响应安全事件并缓解安全事件可能带来的影响。工作负载的架构会极大地影响团队在事件发生期间采取行动、隔离或约束系统并将运行状态恢复到已知的良好状态的能力。在安全事件发生之前确保相关工具部署到位，而后定期进行响应演练，将有助于确保您的架构有能力及时进行调查和恢复。

在 AWS 中，以下实践有助于做出有效的事故响应：

- 我们提供包含文件访问和更改等重要内容的详细日志记录。
- 事件可以自动处理，并且会触发通过使用 AWS API 自动做出响应的工具。
- 您可以使用 AWS CloudFormation 预先配置工具和一个“清洁屋”。这样您就可以在安全且隔离的环境中进行取证。

以下问题主要针对安全方面的注意事项。

## SEC 10：如何预测、响应事件以及从事件中恢复？

准备工作对于及时有效地调查、响应安全事件以及从安全事件中恢复至关重要，可以尽可能减少对组织的破坏。

确保您能够快速授予安全团队访问权限，而且系统可以自动隔离实例并自动捕捉数据与状态信息用于取证。

### 资源

请参阅以下资源，详细了解安全方面的最佳实践。

## 文档

- [AWS Cloud 安全](#)
- [AWS 合规性](#)
- [AWS 安全性博客](#)

## 白皮书

- [安全性支柱](#)
- [AWS 安全性概述](#)
- [AWS 风险和合规性](#)

## 视频

- [AWS 安全中心](#)
- [责任共担模式概述](#)

## 可靠性

可靠性支柱涵盖相关工作负载按照计划正确而稳定执行其预期功能的能力。它包括在其全部生命周期内运行和测试工作负载的能力。本白皮书深度介绍了有关在 AWS 中实施可靠工作负载的最佳实践指导。

可靠性支柱概述了设计原则、最佳实践和问题。如需有关具体实施的说明性指导，请参阅 [《可靠性支柱》白皮书](#)。

### 主题

- [设计原则](#)
- [定义](#)
- [最佳实践](#)
- [资源](#)

## 设计原则

在云中实现可靠性有五个设计原则：



- **自动从故障中恢复**：通过监控工作负载的关键绩效指标 (KPI)，您可以在指标超过阈值时触发自动化功能。这些 KPI 应该是对商业价值（而不是服务运营的技术方面）的一种度量。这包括自动发送故障通知和跟踪故障，以及启动解决或修复故障的自动恢复流程。借助更高级的自动化功能，您可以在故障发生之前预测和修复故障。
- **测试恢复过程**：在本地环境中，经常会通过执行测试来证明工作负载能够在特定场景中正常运作。通常不会利用测试来验证恢复策略。在云中，您可以测试工作负载的故障情况，并验证您的恢复程序。您可以采用自动化方式来模拟不同的故障，也可以重新建立之前导致故障的场景。此方式可以在实际的故障发生以前揭示您可以测试与修复的故障路径，从而降低风险。
- **横向扩展以提高聚合工作负载的可用性**：使用多个小型资源替换一个大型资源，以降低单个故障对整个工作负载的影响。跨多个较小的资源分配请求，以确保它们不共用常见故障点。
- **无需再预估容量**：本地工作负载出现故障的常见原因是资源饱和，即对工作负载的需求超过该工作负载的容量（这通常是拒绝服务攻击的目标）。在云中，您可以监控需求和工作负载利用率，并自动添加或删除资源，以保持最佳水平来满足需求，而不会出现超额预置或预置不足的问题。虽然还有很多限制，但有些配额是可控的，其他配额也可以管理（请参阅“管理Service Quotas与限制”）。
- **管理自动化变更**：应利用自动化功能对基础设施进行更改。需要管理的变更包括，对自动化的变更，可对其进行跟踪与审查。

## 定义

在云中实现可靠性包括四个方面的最佳实践：

- 基础
- 工作负载架构
- 变更管理
- 故障管理

要实现可靠性，您必须从基础入手，而基础是服务配额和网络拓扑适应工作负载的环境。在设计时，分布式系统的工作负载架构必须能够预防与减少故障。工作负载必须处理需求或要求的变化，而且它的设计必须能够检测故障，并自动加以修复。

## 最佳实践

### 主题

- [基础](#)
- [工作负载架构](#)



- [变更管理](#)
- [故障管理](#)

## 基础

基础要求是指其范围超出单个工作负载或项目的因素。在为任何系统设计架构之前，您应确定影响可靠性的基本要求。例如，您必须为数据中心提供足够的网络带宽。

在您使用 AWS 时，这些基础要求中的大部分已经包含在内，并且可以根据需要进行处理。云环境在设计层面拥有几乎无限的资源，因此 AWS 要负责满足对足够联网和计算容量的需求，让您可以根据需求随意更改资源大小和分配。

以下问题主要针对可靠性的注意事项。（有关可靠性问题的列表和最佳实践，请参阅 [附录](#)）。

### REL 1：如何管理服务配额和限制？

基于云的工作负载架构存在服务配额（也被称作服务限制）。存在这些配额是为了防止意外预置超过您所需的资源，并对 API 操作的请求速率进行限制，以保护服务不会遭到滥用。还存在资源限制，例如，将比特推入光缆的速率，或物理磁盘上的存储量。

### REL 2：如何规划网络拓扑？

工作负载通常存在于多个环境中。其中包括多个云环境（可公开访问云和私有云），可能还包括现有数据中心基础设施。相关计划必须涵盖网络注意事项，如系统内部和系统间连接、公有 IP 地址管理、私有 IP 地址管理，以及域名解析。

基于云的工作负载架构存在服务配额（也被称作服务限制）。这些配额的存在目的在于，防止您意外预置超出必要量的资源，限制 API 操作的请求速率，从而避免服务遭到滥用。工作负载通常存在于多个环境中。您必须为所有工作负载环境监控和管理这些配额。其中包括多个云环境（可公开访问的云和私有云），可能还包括您的现有数据中心基础设施。相关计划必须涵盖网络注意事项，如系统内部和系统间连接、公有 IP 地址管理、私有 IP 地址管理以及域名解析。

## 工作负载架构

可靠的工作负载始于前期的软件和基础设施设计决策。您的架构选择将影响所有 Well-Architected 支柱的工作负载行为。针对可靠性，您必须遵循特定的模式。

使用 AWS 时，工作负载开发人员可以选择要使用的语言和技术。AWS 开发工具包通过为 AWS 服务提供特定于语言的 API，省去了复杂的代码编写过程。通过这些开发工具包，以及语言选择，开发人员可以实现此处列出的可靠性最佳实践。开发人员还可以通过以下资料库阅读并了解 Amazon 构建和运营软件的方法：[Amazon Builders' Library](#)。

以下问题主要针对可靠性的注意事项。

### REL 3：如何设计工作负载服务架构？

使用面向服务的架构 (SOA) 或微服务架构构建高度可扩展的可靠工作负载。面向服务的架构 (SOA) 可通过服务接口使软件组件可重复使用。微服务架构则进一步让组件变得更小、更简单。

### REL 4：您如何在分布式系统中设计交互以预防发生故障？

分布式系统依赖于通信网络实现组件（例如服务器或服务）的互联。尽管这些网络中存在数据丢失或延迟，但是您的工作负载必须可靠运行。分布式系统组件的运行方式不得对其他组件或工作负载产生负面影响。这些最佳实践能够预防故障，并改善平均故障间隔时间（MTBF）。

### REL 5：您如何在分布式系统中进行交互设计，从而缓解或经受住故障影响？

分布式系统依赖于通信网络以便使组件互相连接（如服务器或服务）。尽管这些网络中存在数据丢失或延迟，但是您的工作负载必须可靠运行。分布式系统组件的运行方式不得对其他组件或工作负载产生负面影响。这些最佳实践使工作负载能够承受压力或故障，从中更快地恢复，并且降低此类伤害的影响。其结果是缩短平均恢复时间（MTTR）。

## 变更管理

您必须提前为工作负载或其环境的更改做好准备，从而实现工作负载的可靠操作。此类更改包括，外部因素施加到工作负载上的更改（如，需求高峰），以及内部更改（如功能部署和安全补丁）。

使用 AWS，您可以监控工作负载的行为并自动对 KPI 做出响应。例如，您的工作负载可以在某项工作负载用户增加时，添加更多服务器。您可以控制谁有权进行工作负载变更并审核这些变更的历史记录。

以下问题主要针对可靠性的注意事项。

## REL 6：如何监控工作负载资源？

日志和指标是用于了解工作负载运行状况的强大工具。您可以配置工作负载以监控日志和指标，并在超出阈值或发生重大事件时发送通知。监控让您的工作负载可以发现超出低性能阈值和发生故障的情形，从而在响应中自动恢复。

## REL 7：您如何设计工作负载，以适应不断变化的需求？

可扩展工作负载具有自动添加或移除资源的弹性，因此确保在任何时间点都能准确满足当前的需求。

## REL 8：如何实施更改？

要部署新功能，必须对更改加以控制，以确保工作负载和操作环境正在运行已知的软件，并以可预测的方式进行修补和替换。如果此类更改不受控制，您将难以预测这些更改的影响，或难以处理由它们引发的问题。

当您构建工作负载来根据需求变化自动添加和删除资源时，这不仅可以提高可靠性，还可以确保业务成功不至于带来额外负担。借助既有的监控功能，当 KPI 偏离预期标准时，系统会自动向您的团队发送提醒。通过自动记录环境变更，您可以审核并快速识别可能影响可靠性的操作。对变更管理的控制确保您可以实施可提供所需的可靠性的规则。

## 故障管理

在任何具备一定复杂度的系统中，发生故障在意料之中。可靠性要求您的工作负载知晓故障的发生，并采取相应措施以避免对可用性产生影响。工作负载必须既能承受故障，又能自动解决问题。

使用 AWS，您可以利用自动化机制对监控数据做出响应。例如，当特定指标超过阈值时，您可以触发自动操作来解决问题。此外，与其尝试诊断并修复作为生产环境一部分的失败资源，您可以将其替换为新的资源，并对被替换的旧有资源进行故障排查。由于云使您能够以低成本构建整个系统的临时版本，您可以使用自动化测试来验证完整的恢复流程。

以下问题主要针对可靠性的注意事项。

## REL 9：如何备份数据？

备份数据、应用程序和配置，以满足恢复时间目标（RTO）和恢复点目标（RPO）的要求。

## REL 10：如何使用故障隔离来保护您的工作负载？

故障隔离边界可将一个工作负载内的故障影响限制于有限数量的组件。边界以外的组件不会受到故障的影响。使用多个故障隔离边界，您可以限制作用于您的工作负载的影响。

## REL 11：如何将您的工作负载设计为可承受组件故障的影响？

在设计具有高可用性和较短平均恢复时间（MTTR）要求的工作负载时必须考虑到弹性。

## REL 12：如何测试可靠性？

在为您的工作负载采用弹性设计以应对生产压力以后，测试是确保其按设计预期运行，并且提供您所预期弹性的唯一方式。

## REL 13：如何规划灾难恢复（DR）？

拥有适当的备份和冗余工作负载组件是您的 DR 策略的开始。[RTO 和 RPO 是您恢复工作负载的目标](#)。根据业务需求设置这些目标。通过实施策略来实现这些目标，同时考虑工作负载资源和数据的位置和功能。中断概率和恢复成本也是关键因素，有助于了解为工作负载提供灾难恢复的商业价值。

请定期备份数据并测试备份文件，确保您可以从逻辑和物理错误中恢复。管理故障的关键在于自动且频繁地测试工作负载以致其出现故障，然后观察它们如何恢复。请定期执行此操作，并确保在工作负载发生重大变更后也会触发此测试。主动跟踪 KPI（以及恢复时间目标（RTO，Recovery Time Objective）和恢复点目标（RPO，Recovery Point Objective））以评估工作负载的弹性（特别是在故障测试场景中）。跟踪 KPI 将有助于您识别和减少单点故障。充分测试您的工作负载恢复流程，确保可以恢复所有数据并继续为您的客户提供服务，即使面对持续存在的问题也是如此。您的恢复流程应该与您的标准生产流程一样完备而有效。

## 资源

请参阅以下资源，详细了解可靠性的最佳实践。

### 文档

- [AWS 文档](#)
- [AWS 全球基础设施](#)
- [AWS Auto Scaling：扩展计划的工作原理](#)
- [什么是 AWS Backup？](#)

### 白皮书

- [可靠性支柱：AWS Well-Architected](#)
- [在 AWS 上实施微服务](#)

## 性能效率

性能效率要素包括有效地使用计算资源以满足系统要求的能力以及在需求变化和技术改进时保持此效率的能力。

性能效率支柱概述了设计原则、最佳实践和问题。如需有关具体实施的说明性指导，请参阅 [《性能效率支柱》白皮书](#)。

### 主题

- [设计原则](#)
- [定义](#)
- [最佳实践](#)
- [资源](#)

## 设计原则

在云中实现性能效率有五个设计原则：

- 普及先进技术：通过将复杂的任务委派给云供应商，让您的团队更顺畅地实施高级技术。与要求您的 IT 团队学习有关托管和运行新技术的知识相比，考虑将新技术作为服务使用是一种更好的选择。例

如，NoSQL 数据库、媒体转码和机器学习都是需要专业知识才能使用的技术。在云中，这些技术会转变为团队可以使用的服务，让团队能够专注于产品开发，而不是资源预置和管理。

- 数分钟内实现全球化部署：您可以在全球多个 AWS 区域中部署工作负载，从而以更低的成本为客户提供更低的延迟和更好的体验。
- 使用无服务器架构：借助无服务器架构，您无需运行和维护物理服务器即可执行传统计算活动。例如，无服务器存储服务可以充当静态网站（从而无需再使用 Web 服务器），事件服务则可以实现代码托管。这不仅能够消除管理物理服务器产生的运行负担，还可以借由以云规模运行的托管服务来降低业务成本。
- 更频繁地进行实验：利用虚拟资源和可自动化的资源，您可以使用不同类型的实例、存储或配置来快速进行比较测试。
- 考虑软硬件协同编程：了解如何使用云服务，并始终使用最适合您工作负载目标的技术方法。例如，在选择数据库或存储方法时考虑数据访问模式。

## 定义

在云中实现性能效率包括五个方面的最佳实践：

- 架构选择
- 计算和硬件
- 数据管理
- 联网和内容分发
- 流程和文化

采用数据驱动型方法来构建高性能架构。收集架构各方面的数据，从总体设计到资源类型的选择与配置都包括在内。

定期审核您的选择，确保充分利用不断发展的 AWS 云的优势。监控可以确保您随时发现与预期性能的偏差。您可以对您的架构作出权衡以提高性能，例如使用压缩或缓存，或放宽一致性要求。

## 最佳实践

### 主题

- [架构选择](#)
- [计算和硬件](#)
- [数据管理](#)

- [联网和内容分发](#)
- [流程和文化](#)

## 架构选择

针对特定工作负载的最佳解决方案各不相同，而且解决方案通常会结合多种方法。Well-Architected 工作负载会使用多种解决方案，并且允许使用各种不同的功能来提高性能。

我们提供多种类型和配置的 AWS 资源，可让您更轻松找到最能满足您的需求的方法。此外，我们还提供了无法使用本地基础设施轻松实现的选项。例如，Amazon DynamoDB 之类的托管服务可以提供完全托管的 NoSQL 数据库，确保在任何规模下都只会有几毫秒的延迟。

以下问题主要针对性能效率方面的注意事项。（有关性能效率问题的列表和最佳实践，请参阅 [Appendix](#)）。

PERF 1: How do you select appropriate cloud resources and architecture patterns for your workload?

Often, multiple approaches are required for more effective performance across a workload. Well-Architected systems use multiple solutions and features to improve performance.

## 计算和硬件

适合特定工作负载的最佳计算方案会因应用程序设计、使用模式和配置设置而有所不同。架构可能会使用不同的计算方案来支持各种组件，并允许使用不同的功能来提高性能。为架构选择错误的计算方案可能会降低性能效率。

在 AWS 中，计算资源有三种形式：实例、容器和函数：

- 实例是虚拟化服务器，因此您只需通过一个按钮或一次 API 调用即可对其功能进行调整。因为云中的资源决策不是固定不变的，所以您可以尝试使用不同的服务器类型。在 AWS 中，这些虚拟服务器实例具有不同的系列和大小，并且可以提供各种功能，包括固态硬盘（SSD，Solid-State Drive）和图形处理单元（GPU，Graphics Processing Unit）。
- 容器是一种操作系统虚拟化方法，允许您在资源隔离的进程中运行应用程序及其依赖项。AWS Fargate 是适用于容器的无服务器计算引擎。如果您需要控制计算环境的安装、配置和管理，则可以使用 Amazon EC2。此外，您还可以从多个容器编排平台中进行选择：Amazon Elastic Container Service（ECS）或 Amazon Elastic Kubernetes Service（EKS）。



- 函数从您想要应用的代码中抽象出运行环境。例如，AWS Lambda 允许您在不运行实例的情况下运行代码。

以下问题主要针对性能效率方面的注意事项。

## PERF 2: How do you select and use compute resources in your workload?

The more efficient compute solution for a workload varies based on application design, usage patterns, and configuration settings. Architectures can use different compute solutions for various components and turn on different features to improve performance. Selecting the wrong compute solution for an architecture can lead to lower performance efficiency.

## 数据管理

针对特定系统的最佳数据管理解决方案往往取决于数据类型（数据块、文件或对象）、访问模式（随机或连续）、所需吞吐量、访问频率（在线、离线、归档）、更新频率（WORM、动态）以及可用性与持久性限制等因素。Well-Architected 工作负载使用专门构建的数据存储，这些存储允许使用不同的功能来提高性能。

在 AWS 中，存储有三种形式：对象、数据块和文件：

- 对象存储提供了一个可扩展的耐用平台，允许从任何互联网位置访问数据，适用于用户生成的内容、活跃的存档、无服务器计算、大数据存储或备份，以及恢复。Amazon Simple Storage Service (Amazon S3) 是一种对象存储服务，提供行业领先的可扩展性、数据可用性、安全性和性能。Amazon S3 的耐用性可达到 99.999999999%（11 个 9），为全球各地的公司存储数百万个应用程序的数据。
- 数据块存储可为每个虚拟主机提供具有高可用性、一致性且低延迟的数据块存储，类似于直连式存储（DAS）或存储区域网络（SAN）。Amazon Elastic Block Store (Amazon EBS) 旨在满足需要持久性存储的工作负载的需求，此类持久性存储可通过 EC2 实例访问，可助力您根据适合的存储容量、性能和成本对应用程序进行微调。
- 文件存储可以跨多个系统提供对共享文件系统的访问。Amazon Elastic File System (Amazon EFS) 等文件存储解决方案非常适合大型内容存储库、开发环境、媒体存储或用户主目录等使用案例。Amazon FSx 让您可以经济高效地启动和运行常用文件系统，因此您可以利用应用广泛的开源和商业许可文件系统的丰富功能集和高速性能。

以下问题主要针对性能效率方面的注意事项。

### PERF 3: How do you store, manage, and access data in your workload?

The more efficient storage solution for a system varies based on the kind of access operation (block, file, or object), patterns of access (random or sequential), required throughput, frequency of access (online, offline, archival), frequency of update (WORM, dynamic), and availability and durability constraints. Well-architected systems use multiple storage solutions and turn on different features to improve performance and use resources efficiently.

## 联网和内容分发

适合某个工作负载的最佳网络解决方案会因延迟、吞吐量要求、抖动和带宽而有所不同。物理限制（例如用户资源或本地资源）决定位置选项。这些限制可以通过边缘站点或资源置放来抵消。

在 AWS 中，网络资源以虚拟化形式存在，而且支持多种类型和配置。这让您可以更轻松地找到贴合您需求的网络方案。AWS 提供多种产品功能（例如增强联网、Amazon EC2 网络优化实例、Amazon S3 Transfer Acceleration 和动态 Amazon CloudFront）来优化网络流量。AWS 还可以提供多种联网功能（例如 Amazon Route 53 延迟路由、Amazon VPC 端点、AWS Direct Connect 和 AWS Global Accelerator）来减少网络距离或抖动。

以下问题主要针对性能效率方面的注意事项。

### PERF 4: How do you select and configure networking resources in your workload?

This question includes guidance and best practices to design, configure, and operate efficient networking and content delivery solutions in the cloud.

## 流程和文化

在最初构建工作负载时，您可以采用一些原则和实践，来协助您更好地运行高效、高性能的云工作负载。要采用能提高云工作负载性能效率的文化，请考虑以下关键原则和实践。

要打造这种文化，请考虑以下关键原则：

- **基础设施即代码**：使用 AWS CloudFormation 模板之类的方法定义您的基础设施即代码。使用模板，您可以将您的基础设施与应用程序代码和配置一道放入源代码控制中。这使您能够将用于开发软件的实践应用到基础设施，从而能够快速迭代。

- **部署管道**：使用持续集成/连续部署 ( CI/CD ) 管道 ( 例如，源代码存储库、构建系统、部署和测试自动化 ) 来部署您的基础设施。这使您能够以可重复、一致且低成本的方式进行迭代部署。
- **明确定义的指标**：设置并监控指标以捕获关键性能指标 ( KPI )。我们建议您使用技术和业务指标。网站或移动应用程序的关键指标是首个字节捕获时间或渲染时间。其他常规的适用指标包括线程计数、垃圾收集速率以及等待状态。业务指标，如单次请求累计总成本，可以提醒您留意降低成本的方法。仔细考虑解读指标的方式。例如，您可以选择最大值或第 99 个百分位数，而不是平均值。
- **自动性能测试**：作为部署过程的一部分，在快速运行测试成功通过后自动开始性能测试。自动化应创建新环境、设置初始条件 ( 如测试数据 )，然后执行一系列基准和负载测试。这些测试的结果应回绑到构建中，以便您可以随着时间推移跟踪性能变化。对于长时间运行的测试，您可以使管道的这一部分与构建的剩余部分异步进行。或者，您也可以使用 Amazon EC2 竞价型实例来通宵执行性能测试。
- **负载生成**：您应该创建复制综合或预先记录的用户旅程的一系列测试脚本。这些脚本应该是幂等的，而不是耦合，您可能需要包含“预热”脚本以便产生有效结果。测试脚本应尽可能再现生产中的使用行为。您可以使用软件或软件即服务 ( SaaS ) 解决方案来生成负载。考虑使用 [AWS Marketplace](#) 解决方案和[竞价型实例](#) – 它们是用于生成负载的经济高效的方法。
- **性能可见性**：关键指标应该对您的团队可见，尤其是针对每个构建版本的指标。这让您能够随着时间推移看到所有重大的正面或负面趋势。您还应展示有关错误或异常数量的指标，以确保测试的是正常工作的系统。
- **可视化**：使用可视化技术，清楚了解出现性能问题、热点、等待状态或低利用率的位置。在架构图上叠加性能指标 – 调用图表或代码有助于快速发现问题。
- **定期审查流程**：通常，不存在或不完整的性能审核流程会导致架构性能不佳。如果您的架构性能不佳，请实施性能审核流程，以便推动迭代改进。
- **持续优化**：采用一种文化，不断优化云工作负载的性能效率。

以下问题主要针对性能效率方面的注意事项。

#### PERF 5: What process do you use to support more performance efficiency for your workload?

When architecting workloads, there are principles and practices that you can adopt to help you better run efficient high-performing cloud workloads. To adopt a culture that fosters performance efficiency of cloud workloads, consider these key principles and practices.

## 资源

请参阅以下资源，详细了解提升性能效率的最佳实践。

### 文档

- [Amazon S3 Performance Optimization](#)
- [Amazon EBS Volume Performance](#)

### 白皮书

- [性能效率支柱](#)

### 视频

- [AWS re:Invent 2019: Amazon EC2 foundations \(CMP211-R2\)](#)
- [AWS re:Invent 2019: Leadership session: Storage state of the union \(STG201-L\)](#)
- [AWS re:Invent 2019: Leadership session: AWS purpose-built databases \(DAT209-L\)](#)
- [AWS re:Invent 2019: Connectivity to AWS and hybrid AWS network architectures \(NET317-R1\)](#)
- [AWS re:Invent 2019: Powering next-gen Amazon EC2: Deep dive into the Nitro system \(CMP303-R2\)](#)
- [AWS re:Invent 2019: Scaling up to your first 10 million users \(ARC211-R\)](#)

## 成本优化

成本优化支柱包括以最低价格运行系统来交付商业价值的能力。

成本优化支柱概述了设计原则、最佳实践和问题。如需有关具体实施的说明性指导，请参阅 [部分](#)。

### 主题

- [设计原则](#)
- [定义](#)
- [最佳实践](#)
- [资源](#)

## 设计原则

在云中实现成本优化有五个设计原则：

- **践行云财务管理**：为获得财务上的成功并加速在云中实现商业价值，需要投资云财务管理/成本优化。您的组织需要投入时间和资源增强自身在这个新的技术和使用情况管理领域中的能力。与安全性或卓越运营能力类似，您的组织需要通过知识构建、计划、资源和流程来培养能力，从而成为一家具有成本效益的组织。
- **采用消费模型**：仅为所需计算资源付费，并可根据业务需求而非复杂的预测增加或减少使用量。例如，开发和测试环境通常只需要在每个工作日运行八个小时。您可以在不需要时停用这些资源，从而实现 75% 的潜在成本节约（40 小时对比 168 小时）。
- **衡量整体效率**：衡量工作负载的业务产出及这些产出的实现成本。使用这种衡量方式了解您通过提高产出和降低成本获得的收益。
- **不再将资金投入无差别的繁重任务上**：AWS 会负责繁重的数据中心运维任务，例如服务器的安装、堆叠和供电。它还消除了使用托管服务管理操作系统和应用程序的运营负担。因此，您可以集中精力处理客户和业务项目而非 IT 基础设施。
- **对支出进行分析和归因**：使用云，您可以更轻松确定系统的准确使用量和成本，从而将 IT 成本透明地分摊到各个工作负载拥有者。这有助于衡量投资回报率 (ROI)，并让工作负载拥有者能够据此优化资源和降低成本。

## 定义

在云中实现成本优化包括五个方面的最佳实践：

- 践行云财务管理
- 支出和使用情况意识
- 具有成本效益的资源
- 管理需求和供应资源
- 随着时间的推移不断优化

与良好架构框架中的其他支柱一样，成本优化支柱也需要权衡各种因素，例如，是优化上市速度还是优化成本。在某些情况下，最好优化上市速度以便快速上市、交付新功能或只是为了按时完成任务，而不是优化预付成本。设计决策有时是在仓促中而非是由数据决定的，并且人们总是倾向于过度补偿“以防万一”，而不是花时间进行基准测试以获得成本最优的部署。这可能会导致过度预置和优化不足的部署。但是，当您需要将资源从本地环境“直接迁移”到云，然后再进行优化时，这是一个合理的选择。通

过预先在成本优化策略中投入适量的精力，您可以确保始终如一地遵守最佳实践，避免不必要的过度预置，从而更轻松地实现云的经济优势。以下部分介绍了一些技巧和最佳实践，可帮助您开始并持续实施工作负载的云财务管理和成本优化。

## 最佳实践

### 主题

- [践行云财务管理](#)
- [支出和使用情况意识](#)
- [具有成本效益的资源](#)
- [管理需求和供应资源](#)
- [随着时间的推移不断优化](#)

### 践行云财务管理

采用云后，由于缩短了审批、采购和基础设施部署周期，技术团队的创新速度会更快。要实现商业价值和财务成功，需要实施一种在云中管理财务的新方法。这种方法便是云财务管理，通过实施组织范围的知识构建、计划、资源和流程，在整个组织内培养能力。

许多组织由许多不同的单位构成，而这些单位又具有不同的要务。若能让组织遵循一组商定的财务目标并为组织提供实现这些目标的机制，将会打造一个更高效的组织。一个有能力的组织的创新和构建速度更快，更敏捷，并能够适应任何内部或外部因素。

在 AWS 中，您可以使用 Cost Explorer，也可以选择使用 Amazon Athena 和 Amazon QuickSight 查看成本和使用情况报告（CUR，Cost and Usage Report），从而了解整个组织的成本和使用情况。AWS Budgets 可主动发出成本和使用情况通知。AWS 博客提供有关新服务和新功能的信息，以确保您及时了解新发布的服务。

以下问题主要针对成本优化方面的注意事项。（有关成本优化问题的列表和最佳实践，请参阅 [附录](#)）。

#### COST 1：如何实施云财务管理？

实施云财务管理后，组织可以在 AWS 上优化成本和使用情况并进行扩展，从而实现商业价值和财务成功。

在组建成本优化部门时，需要包括成员并为团队配备 CFM 和成本优化方面的专家。现有的团队成员将了解组织的当前运作方式以及如何快速实施改进。此外，还可以考虑配备拥有辅助或专业技能组合的人员，例如具备分析和项目管理能力的人员。

在组织中树立成本意识时，需要改进现有计划和流程或基于现有计划和流程进行构建。与构建新流程和计划相比，向现有流程和计划增添内容要快得多。这样将更快地取得成果。

## 支出和使用情况意识

通过云，您可以获得更大的灵活性和敏捷性，从而支持创新以及快速的开发和部署。这样便节省了自建本地基础设施所需的人工环节和时间，包括确定硬件规格、协商报价、管理购买订单、安排发货和部署资源。然而，要实现这种易用性并利用近乎无限的按需容量，我们需要以新方式考虑支出。

很多企业有多个由不同团队运行的系统。将资源成本分摊到各个组织或产品拥有者可以推动更高效的资源使用模式，减少浪费。准确的成本分摊能够帮助您了解哪些产品是真正盈利的，让您能够做出更明智的预算分配决策。

在 AWS 中，您可以使用 AWS Organizations 或 AWS Control Tower 创建账户结构，这种方式不仅实现了分离，而且有助于对成本和使用进行分配。此外，也可以通过资源标记在使用情况和成本中标注业务和组织信息。使用 AWS Cost Explorer 查看您的成本和使用情况，或者使用 Amazon Athena 和 Amazon QuickSight 创建自定义控制面板和分析。成本和使用情况控制通过 AWS Budgets 的通知来实现，并使用 AWS Identity and Access Management ( IAM ) 和 Service Quotas 进行控制。

以下问题主要针对成本优化方面的注意事项。

### COST 2：您如何管理使用情况？

制定各种策略和机制，确保花费适当的成本来达到目标。采用制约与平衡方法，您可以在不超支的情况下进行创新。

### COST 3：如何监控使用情况和成本？

建立策略和程序以便监控并适当分配您的成本。这让您能够衡量和改进工作负载的成本效益。



## COST 4：您如何停用资源？

在从项目开始到结束的过程中实施变更控制和资源管理。这可以确保您关闭或终止未使用的资源，以便减少浪费。

您可以使用成本分配标签对 AWS 使用情况和成本进行分类并跟踪。当您为 AWS 资源（例如 EC2 实例或 S3 存储桶）应用标签后，AWS 将通过使用情况和成本标签生成成本和使用情况报告。您可以使用代表组织类别的标签（例如成本中心、工作负载名称或所有者）整理您的多个服务的成本。

确保在成本和使用情况报告和监控中使用正确的详细级别和粒度。要获得大概见解和趋势，请在 AWS Cost Explorer 中使用每日粒度。要更深入地进行分析和检查，请在 AWS Cost Explorer 中使用每小时粒度，或者在 Amazon Athena 和 Amazon QuickSight 中以每小时为粒度查看成本和使用情况报告 (CUR)。

结合标记资源和实体生命周期跟踪（员工、项目），您可以确定无法再为组织创造价值而应停用的孤立资源或项目。您可以设置账单提醒，以在预计超支时通知您。

### 具有成本效益的资源

为工作负载使用合适的实例和资源是节约成本的关键。例如，在小型服务器上运行某个报告需要五个小时，而在另一个两倍成本的大型服务器上运行只需要一个小时。虽然两个服务器提供同样的结果，但小型服务器随着时间推移会产生更多成本。

良好架构的工作负载会使用最具有成本效益的资源，这样可以产生巨大而积极的经济效益。您还可以使用托管服务降低成本。例如，您可以使用按电子邮件收费的服务，而无需自己维护电子邮件服务器。

AWS 提供各种灵活且具有成本效益的定价选项，以最符合您需求的方式从 Amazon EC2 和其他服务获取实例。按需实例允许按小时支付计算容量的费用，且无需最低使用承诺。Savings Plans 和预留实例与按需定价相比最高可节约 75% 的成本。使用 Spot 实例，您可以利用未使用的 Amazon EC2 容量，并且与按需定价相比最高可节约 90% 的成本。Spot 实例适用于以下情况：系统可以容忍使用服务器队列，其中单个服务器可以动态装卸（例如无状态 Web 服务器）、批处理或使用 HPC 和大数据。

选择合适的服务还可以减少使用量和降低成本；例如，使用 CloudFront 可以最大限度地减少数据传输成本；例如，使用 Amazon Aurora on RDS 可以消除昂贵的数据库许可成本。

以下问题主要针对成本优化方面的注意事项。

### COST 5：您在选择服务时如何评估成本？

Amazon EC2、Amazon EBS 和 Amazon S3 属于构建块 AWS 服务。托管服务（如 Amazon RDS 和 Amazon DynamoDB）属于更高级别或应用程序级别的 AWS 服务。通过选择适当的基础服务和托管服务，您可以优化工作负载，从而降低成本。例如，使用托管服务，您可以节省或消除大部分管理和运营开销，从而使您有精力从事应用程序和业务相关活动。

### COST 6：在选择资源类型、规模和数量时，如何实现成本目标？

确保选择适合当前任务的资源规模和资源数量。选择最经济实惠的资源类型、规模和数量可以尽可能减少浪费。

### COST 7：您如何使用定价模式来降低成本？

使用最适合的资源定价模式可以尽可能减少支出。

### COST 8：您如何规划数据传输费用？

务必要监控和规划您的数据传输费用，以便制定架构决策，尽可能降低成本。持续以小步迭代的方式进行架构优化可以实现运营成本的大幅降低。

通过在选择服务时考虑成本因素，并使用 Cost Explorer 和 AWS Trusted Advisor 等工具定期检查 AWS 使用情况，您可以主动监控利用率并相应地调整部署。

## 管理需求和供应资源

在您迁移到云时，您仅为所需内容付费。您可以在需要时供应与工作负载需求匹配的资源，从而消除昂贵且浪费的过度预置需求。还可以通过限流、缓冲区或队列来修改需求，以满足需求并以更少的资源达成目标，从而降低成本，或者在以后使用批处理服务处理需求。

在 AWS 中，您可以自动预置资源来满足工作负载需求。通过使用基于需求或时间的方法进行 Auto Scaling，您可以根据需要添加和删除资源。如果您可以预测需求变化，便可以节省更多资金并确保资源与工作负载需求匹配。您可以使用 Amazon API Gateway 实施限流，也可以使用 Amazon SQS 在工作负载中实施队列。这两种方法都允许您修改工作负载组件的需求。

以下问题主要针对成本优化方面的注意事项。

### COST 9：如何管理需求和供应资源？

为了工作负载的性能与支出实现平衡，请确保您支付过费用的所有资源都得到利用，并避免出现资源利用率过低的情况。无论是从运维成本（由于过度使用导致性能下降）还是从浪费 AWS 支出（由于超额配置）的角度衡量，利用率指标过高或过低都会对您的组织产生负面影响。

当进行修改需求和供应资源的设计时，请主动考虑资源使用模式、预置新资源所需要耗费的时间，以及需求模式的可预测性。当管理需求时，确保您具有大小正确的队列或缓冲区，并在所需的时间内响应工作负载需求。

### 随着时间的推移不断优化

AWS 不断发布新服务和功能，因此您最好不断审视现有架构决策，以便确保其始终最具成本效益。当您的需求发生变化时，请主动停用不再需要的资源、整体服务和系统。

实施新功能或资源类型可以逐步优化您的工作负载，同时最大程度地减少实施变更所需的工作量。这样可不断提高效率，并确保您始终使用最新的技术，从而降低运营成本。您还可以使用新服务替换或向工作负载中添加新组件。这可以显著提高效率，因此必须定期审查您的工作负载，并实施新服务和新功能。

以下问题主要针对成本优化方面的注意事项。

### COST 10：如何评估新服务？

AWS 不断发布新服务和功能，因此您最好不断审视现有架构决策，以便确保其始终最具成本效益。

定期审查部署时，评估更新的服务如何帮助您节省成本。例如，Amazon Aurora on RDS 可以降低关系数据库的成本。使用无服务器（例如 Lambda）服务，无需操作和管理实例即可运行代码。

## 资源

请参阅以下资源，详细了解成本优化的最佳实践。

## 文档

- [AWS 文档](#)

## 白皮书

- [成本优化支柱](#)

## 可持续性

可持续性支柱侧重于环境影响，尤其是能源消耗和效率，因为它们是架构师在直接采取行动以减少资源使用时依据的重要杠杆。如需有关具体实施的说明性指导，请参阅 [《可持续性支柱》白皮书](#)。

### 主题

- [设计原则](#)
- [定义](#)
- [最佳实践](#)

## 设计原则

在云中实现可持续性有六个设计原则：

- **了解您的影响：** 衡量您的云工作负载的影响并为您的工作负载的未来影响建模。包括所有影响来源，例如客户使用您的产品所产生的影响，以及产品最终淘汰和停用所产生的影响。通过查看每个工作单元所需的资源和排放量，将生产性输出与云工作负载的总体影响进行比较。使用这些数据来建立关键绩效指标（KPI），评估在降低影响的同时提高生产力的方法，并估计提议的更改随时间的推移所产生的影响。
- **设定可持续性目标：** 对于每个云工作负载，建立长期可持续性目标，例如减少每个事务所需的计算和存储资源。针对现有工作负载的可持续性改进的投资回报进行建模，并为负责人提供投资于可持续性目标所需的资源。规划增长并构建您的工作负载，以便增长可降低影响强度（以适当的单位衡量，例如每用户或每事务）。目标可帮助您支持您的企业或组织更广泛的可持续发展目标、识别回归并确定潜在改进领域的优先级。
- **实现利用率最大化：** 适当调整工作负载规模并实施高效设计，以确保高利用率并最大限度地提高底层硬件的能源效率。由于每台主机的基准功耗，两台以 30% 利用率运行的主机的效率低于一台以 60% 利用率运行的主机。同时，消除或尽可能减少空闲资源、处理和存储，以减少支持工作负载所需的总能源。
- **预测并采用更高效的新硬件和软件产品/服务：** 支持您的合作伙伴和供应商进行上游改进，以帮助您减少云工作负载的影响。持续监控和评估更高效的新硬件和软件产品。设计灵活性以允许快速采用高效的新技术。

- **使用托管服务：** 在庞大的客户群中共享服务有助于更充分地利用资源，从而减少支持云工作负载所需的基础设施数量。例如，客户可以通过将工作负载迁移到 AWS Cloud 并采用托管服务（例如用于无服务器容器的 AWS Fargate，AWS 在其中大规模运行并负责其高效运行）来分散电力和网络等常见数据中心组件的影响。使用有助于将影响降至最低的托管服务，例如使用 Amazon S3 生命周期配置将不经常访问的数据自动移动到冷存储，或使用 Amazon EC2 Auto Scaling 来调整容量以满足需求。
- **减少云工作负载的下游影响：** 减少使用您的服务所需的能源或资源量。减少或消除客户为了使用您的服务而升级其设备的需求。使用设备场进行测试以了解预期影响，并对客户进行测试以了解使用您服务的实际影响。

## 定义

在云中实现可持续性包括六个方面的最佳实践：

- 区域选择
- 用户行为模式
- 软件和架构模式
- 数据模式
- 硬件模式
- 开发和部署流程

云中的可持续性是一项持续的工作，主要关注工作负载的所有组件的节能和效率，通过从预置的资源中获得最大收益，并最大限度地减少所需的总资源来达成此目标。这项工作范围很广，包括一开始就选择高效的编程语言、采用现代算法、使用高效的数据存储技术、部署到适当规模的高效计算基础设施，以及最大限度地减少对高功耗最终用户硬件的需求。

## 最佳实践

### 主题

- [区域选择](#)
- [用户行为模式](#)
- [软件和架构模式](#)
- [数据模式](#)
- [硬件模式](#)

- [开发和部署模式](#)
- [资源](#)

## 区域选择

根据您的业务需求和可持续发展目标，选择您将在其中实施工作负载的区域。

以下问题主要针对可持续性方面的注意事项。（有关可持续性问题 and 最佳实践的列表，请参阅 [附录](#)。）

### SUS 1：如何选择区域来支持您的可持续发展目标？

选择亚马逊可再生能源项目附近的区域和其电网公布的碳强度低于其他位置（或区域）的区域。

## 用户行为模式

用户使用您的工作负载和其他资源的方式可以帮助您确定改进措施，以实现可持续性目标。扩展基础设施以持续匹配用户负载，并确保仅部署支持用户所需的最少资源。使服务水平与客户需求保持一致。定位资源以限制用户使用它们所需的网络。移除现有的未使用资产。识别已创建但未使用的资产并停止生成它们。为您的团队成员提供满足其需求的设备，同时最大限度地减少对可持续性的影响。

以下问题主要针对可持续性方面的注意事项：

### SUS 2：您如何利用用户行为模式来支持您的可持续发展目标？

用户使用您的工作负载和其他资源的方式可以帮助您确定改进措施，以实现可持续性目标。扩展基础设施以持续匹配用户负载，并确保仅部署支持用户所需的最少资源。使服务水平与客户需求保持一致。定位资源以限制用户使用它们所需的网络。移除现有的未使用资产。识别已创建但未使用的资产并停止生成它们。为您的团队成员提供满足其需求的设备，同时最大限度地减少对可持续性的影响。

扩缩基础设施以匹配用户负载：确定利用率低或利用率为零的时段，缩减资源以消除过剩容量并提高效率。

使 SLA 与可持续发展目标保持一致：定义和更新服务等级协议（SLA，Service Level Agreement），例如可用性 or 数据留存期，以最大限度地减少支持工作负载所需的资源数量，同时继续满足业务需求。



消除创建和维护未使用资产的需求：分析应用程序资产（例如预编制的报告、数据集和静态图像）和资产访问模式，以识别冗余、利用率低下的情况和潜在的淘汰目标。整合具有冗余内容的已生成资产（例如，具有重叠或公用数据集和输出的月度报告），以消除重复输出时消耗的资源。淘汰未使用的资产（例如，已停售产品的图片）以释放消耗的资源，并减少用于支持工作负载的资源数量。

针对用户位置优化工作负载的地理位置：分析网络访问模式以识别您的客户建立连接的地理位置。选择可减少网络流量必须传输的距离的区域和服务，以减少支持您的工作负载所需的总网络资源。

针对执行的活动优化团队成员资源：优化提供给团队成员的资源，在支持其需求的同时最大程度地降低对可持续性的影响。例如，在利用率高的共享云桌面上，而不是在利用率不高的强力单用户系统上，执行渲染和编译等复杂的操作。

## 软件和架构模式

实施用于执行负载平滑和保持已部署资源始终如一的高利用率的模式，以最大限度地减少资源消耗。由于用户行为会随着时间的推移而发生变化，因此组件可能会因缺乏使用而变得空闲。修改模式和架构以整合未充分利用的组件，从而提高整体利用率。停用不再需要的组件。了解工作负载组件的性能，并优化消耗资源最多的组件。注意客户用来访问您服务的设备，并实施相应的模式以最大限度地减少设备升级需要。

以下问题主要针对可持续性的注意事项：

### SUS 3：您如何利用软件和架构模式来支持您的可持续发展目标？

实施用于执行负载平滑和保持已部署资源始终如一的高利用率的模式，以最大限度地减少资源消耗。由于用户行为会随着时间的推移而发生变化，因此组件可能会因缺乏使用而变得空闲。修改模式和架构以整合未充分利用的组件，从而提高整体利用率。停用不再需要的组件。了解工作负载组件的性能，并优化消耗资源最多的组件。注意客户用来访问您服务的设备，并实施相应的模式以最大限度地减少设备升级需要。

针对异步和计划作业优化软件和架构：使用高效的软件设计和架构来尽可能减少每个工作单元所需的平均资源。实施可促成均匀的组件利用率的机制，以减少任务之间的空闲资源并最大限度地减少负载峰值的影响。

删除或重构很少或没有使用的工作负载组件：监控工作负载活动以识别各个组件的利用率随时间的变化。移除未使用且不再需要的组件，并重构利用率低的组件，以限制资源浪费。

优化消耗最多时间或资源的代码区域：监控工作负载活动以识别消耗最多资源的应用程序组件。优化在这些组件中运行的代码，以最大限度地减少资源使用和提高性能。



优化对客户设备的影响：了解客户用来使用您服务的设备、它们的预期生命周期，以及更换这些组件对财务和可持续性的影响。实施软件模式和架构，以最大限度地减少客户更换和升级设备的需求。例如，使用与旧硬件和操作系统版本向后兼容的代码来实施新功能，或管理有效负载的大小，使其不超过目标设备的存储容量。

使用最能支持数据访问和存储模式的软件模式和架构：了解数据在工作负载中的使用方式、用户使用数据的方式，以及数据的传输和存储方式。选择相应的技术以最大限度地减少数据处理和存储要求。

## 数据模式

实施用于执行负载平滑和保持已部署资源始终如一的高利用率的模式，以最大限度地减少资源消耗。由于用户行为会随着时间的推移而发生变化，因此组件可能会因缺乏使用而变得空闲。修改模式和架构以整合未充分利用的组件，从而提高整体利用率。停用不再需要的组件。了解工作负载组件的性能，并优化消耗资源最多的组件。注意客户用来访问您服务的设备，并实施相应的模式以最大限度地减少设备升级需要。

以下问题主要针对可持续性方面的注意事项：

### SUS 4：您如何利用数据访问模式和使用模式来支持您的可持续发展目标？

实施数据管理实践以减少支持工作负载所需的预置存储，以及使用存储所需的资源。了解您的数据，并使用最能支持数据的商业价值及其使用方式的存储技术和配置。当需求减少时，将数据移到更高效、性能更低的存储中，并删除不再需要的数据。

实施数据分类策略：对数据进行分类以了解其对业务成果的重要性。使用此信息来确定何时可以将数据移动到更节能的存储，或者何时可以安全删除数据。

使用支持您的数据访问模式和存储模式的技术：使用最能支持您的数据访问和存储方式的存储，以在支持您的工作负载的同时最大限度地减少预置的资源。例如，固态硬盘（SSD，Solid State Device）比磁性驱动器更耗能，应该仅用于活跃的数据使用场景。对不常访问的数据使用节能的存档级存储。

使用生命周期策略删除不必要的数据：管理所有数据的生命周期并自动执行删除时间表，以最大限度地减少工作负载的总存储需求。

最大限度地减少数据块存储中的过度预置：要尽可能减少总预置存储，请创建大小分配适合工作负载的数据块存储。随着数据的增长，使用弹性卷扩展存储，而无需调整附加到计算资源的存储大小。定期检查弹性卷并缩小过度配置的卷，以适应当前数据大小。

删除不需要或多余的数据：仅在必要时复制数据，以最大程度地减少消耗的总存储空间。使用备份技术在文件和数据块级别进行重复数据删除。限制使用独立驱动器冗余阵列 ( RAID ) 配置，除非需要满足 SLA。

使用共享文件系统或对象存储来访问公用数据：采用共享存储和单一事实来源，以避免重复数据删除并降低工作负载的总存储需求。仅在需要从共享存储中获取数据。分离未使用的卷以释放资源。最大限度地减少跨网络的数据移动：使用共享存储和访问区域数据存储中的数据，以最大限度地减少支持工作负载数据移动所需的总网络资源。

仅在难以重新创建时备份数据：为了最大限度地减少存储消耗，仅备份具有商业价值或满足合规性要求所必需的数据。检查备份策略并在恢复方案中排除没有价值的临时存储。

## 硬件模式

寻找机会，通过更改硬件管理实践来降低工作负载可持续性影响。最大限度地减少预置和部署所需的硬件数量，并为您的各项工作负载选择最高效的硬件。

以下问题主要针对可持续性方面的注意事项：

### SUS 5：您的硬件管理和使用实践如何支持您的可持续发展目标？

寻找机会，通过更改硬件管理实践来降低工作负载可持续性影响。最大限度地减少预置和部署所需的硬件数量，并为您的各项工作负载选择最高效的硬件。

使用最少的硬件来满足您的需求：通过使用云的功能，您可以对工作负载实施进行频繁更改。在需求变化时更新已部署的组件。

使用影响最小的实例类型：持续监控新实例类型的发布并利用能效改进，包括那些旨在支持特定工作负载（例如机器学习训练和推理以及视频转码）的实例类型。

使用托管服务：托管服务将维持已部署硬件的高平均利用率和可持续性优化的责任转移给 AWS。使用托管服务将服务的可持续性影响分散到服务的所有租户，从而减少您的个人份额。

优化您对 GPU 的使用：图形处理单元 ( GPU , Graphics Processing Unit ) 可能是高功耗的来源，许多 GPU 工作负载是高度可变的，例如渲染、转码以及机器学习训练和建模。仅在需要时运行 GPU 实例，并在不需要时自动停用它们，以最大限度地减少资源消耗。

## 开发和部署模式

寻找机会，通过对开发、测试和部署实践进行更改来降低可持续性影响。

以下问题主要针对可持续性方面的注意事项：

### SUS 6：您的开发和部署流程如何支持您的可持续发展目标？

寻找机会，通过对开发、测试和部署实践进行更改来降低可持续性影响。

采用可以快速引入可持续性改进的方法：在将潜在改进部署到生产环境之前，先进行测试和验证。在计算改进的潜在未来收益时，考虑测试成本。开发低成本的测试方法，以实现细微的改进。

让您的工作负载保持最新：最新的操作系统、库和应用程序可以提高工作负载效率，并简化更高效技术的采用。最新的软件可能还包括更准确地衡量工作负载对可持续性的影响的功能，因为供应商提供的功能是为了满足其自身的可持续性目标。

提高构建环境的利用率：使用自动化功能和基础设施即代码，在需要时启动预生产环境，并在不使用时将其关闭。一种常见模式是安排与开发团队成员的工作时间相吻合的可用时段。休眠是一个有用的工具，它可以保存状态，并且只在需要时才快速将实例上线。使用具有突增容量的实例类型、Spot 实例、弹性数据库服务、容器和其他技术，使开发和测试能力与使用相一致。

使用托管 Device Farm 进行测试：托管式设备场将硬件制造和资源使用的可持续性影响分散到多个租户。托管式设备场提供多种设备类型，使您能够支持不太受欢迎的较旧硬件，并避免不必要的设备升级对客户可持续性的影响。

## 资源

请参阅以下资源，详细了解可持续性的最佳实践。

### 白皮书

- [可持续性支柱](#)

### 视频

- [The Climate Pledge](#)

# 审查流程

需要持续不断对架构进行审查，同时要允许试错，建立良好的研究探索氛围。架构审查本身应该是一个简单流程（数小时，而不是几天），是一种对话，而不是审核。审查架构的目的是找出任何需要解决的关键问题或可以改进之处。审查后应采取一些措施，以改善客户体验。

正如在“关于架构”部分讨论的那样，团队中的每位成员都应该对架构质量负责。我们建议负责架构的团队利用架构完善的框架持续检视架构，而不仅仅只是召开一个正式的审查会议。持续的审查使团队成员能够随着架构的演进不断获得知识体系与对架构认识的更新，并在您推出新功能时改进架构。

AWS Well-Architected Framework 高度借鉴了 AWS 在内部审查系统和服务的方式，并与之保持一致。它基于一套可以影响架构方法的设计原则，并确保不会忽略常见于根因分析 (RCA) 中的那些因素。当内部系统、AWS 服务或客户遇到严重问题时，我们会查看 RCA，寻求改进所用审查流程的可能性。

应在产品生命周期内的关键里程碑阶段和设计阶段早期进行多次审查，以避免单向决策，因为它们很难进行更改，然后在正式投入使用之前再次审查。（许多决策都是可逆的，是双向的。这些决策可以采用简单流程。单向决策很难，不可逆，所以在做出决策之前需要更加全面的检查。）进入生产阶段后，您的工作负载会随着您不断添加新功能和更改技术实施而继续演进。架构也会随之演进。您需要遵循良好的架构实践，避免出现架构退化。当架构面临重大变更时，您应遵循一套系统健康规范与流程，包括执行 Well-Architected 审查。

如果您想把审查用作一次性快照或独立的衡量方法，则需要确保让所有相关人员参与对话。我们经常发现，通过审查，团队才第一次真正了解他们实施了什么。在审查其他团队的工作负载时，一种有效的方法是围绕架构展开一系列非正式的对话，在此过程中您可以收集到大多数问题的答案。然后通过一两次会议进行跟进，来帮助您理清思路，或深入了解不明确的方面或已感知的风险。

下面建议了一些召开会议需要准备的事项：

- 配有白板的会议室
- 任何图表或设计说明的打印件
- 需要带外研究来获取答案的问题（例如，“是否已启用加密？”）

完成审查后，您应有一个问题清单，并基于您的业务环境来确定这些问题的优先级。您还需要考虑这些问题对团队日常工作的影响。如能及早解决这些问题，您就可以腾出时间开展创造商业价值的工作，而不是解决重复出现的问题。在解决问题时，您可以反复进行审查，来确认架构的改进效果。

虽然在完成审查后，审查的价值显而易见，但您可能发现新团队在开始时可能会对审查抱有抵触情绪。可以通过与团队沟通审查的益处来解决下列异议：

- “我们实在太忙了！”（一般会在团队准备重大发布时这么说。）
  - 如果您正在为重大发布做准备，您会希望一切进展顺利。审查能够帮助您发现您可能错过的任何问题。
  - 我们建议您在产品生命周期早期执行审查，以及时发现风险，并制定与功能交付路线图一致的规避计划。
- “我们已经没有时间了，结果已成定局！”（一般会在他们面临无法改变的事件 [比如超级碗] 时这么说。）
  - 这些事件是无法改变的。您真的想在不了解架构风险的情况下将它投入使用吗？即使不能解决所有问题，您仍然可以编制一个潜在问题处理手册。
- “我们不希望其他人知道我们实施解决方案的秘诀！”
  - 如果您向团队指出 Well-Architected Framework 中的问题，他们不会在这些问题中发现任何商业或技术专有信息。

在您与团队进行多次审核后，您可能会发现一些问题。例如，您可能发现一些团队在某个支柱或主题方面出现较多问题。建议您以全局眼光看待所有审查，找出能够帮助解决这些问题的任何机制、培训或首席工程师会谈方案。

# 总结

AWS Well-Architected Framework 涵盖了六大支柱，强调了在云中设计和运行可靠、安全、高效、经济实惠且可持续的系统的架构最佳实践。该框架提供了一系列问题清单，来帮助您审查现有或将要实现的架构。它还为每个支柱提供了一组 AWS 最佳实践。在架构中应用该框架将能帮助您打造稳定且高效的系统，从而使您能够将主要精力集中在功能需求上。

# 贡献者

以下是对本文做出贡献的个人和组织：

- Brian Carlson , Amazon Web Services“架构完善”的操作主管
- Ben Potter , Amazon Web Services Well-Architected 安全主管
- Seth Eliot , Amazon Web Services Well-Architected 可靠性主管
- Eric Pullen , Amazon Web Services 高级解决方案架构师
- Rodney Lester , Amazon Web Services 首席解决方案架构师
- Jon Steele , Amazon Web Services 高级技术客户经理
- Max Ramsay , Amazon Web Services 首席安全解决方案架构师
- Callum Hughes , Amazon Web Services 解决方案架构师
- Aden Leirer , Amazon Web Services Well-Architected 内容计划经理



# 延伸阅读

[AWS Architecture Center](#)

[AWS 云合规性](#)

[AWS Well-Architected 合作伙伴计划](#)

[AWS Well-Architected Tool](#)

[AWS Well-Architected 主页](#)

[《卓越运营支柱》白皮书](#)

[《安全性支柱》白皮书](#)

[《可靠性支柱》白皮书](#)

[《性能效率支柱》白皮书](#)

[部分](#)

[《可持续性支柱》白皮书](#)

[Amazon Builders' Library](#)

# 文档修订

要获得有关此白皮书的更新通知，请订阅 RSS 源。

| 变更                         | 说明                                                                                                                                                                                                                                                        | 日期                |
|----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
| <a href="#">主要更新</a>       | 重大 <a href="#">性能支柱</a> 重组，将最佳实践领域的数量增加到五个。对 <a href="#">事件响应 ( SEC 10 )</a> 访问 AWS 资源。对以下卓越运营领域的内容进行重要变更和整合 <a href="#">OPS 04、05、06、08 和 09</a> 访问 AWS 资源。更新了整个 <a href="#">成本优化</a> 和 <a href="#">可靠性</a> 支柱中的指南。对 <a href="#">可持续性支柱</a> 风险等级进行了次要更新。 | October 3, 2023   |
| <a href="#">针对新框架进行了更新</a> | 为最佳实践更新了规范性指南并增加了新的最佳实践。在“安全性”和“成本优化”支柱中增加了新问题。                                                                                                                                                                                                           | April 10, 2023    |
| <a href="#">次要更新</a>       | 在附录中增加了工作量定义和更新的最佳实践。                                                                                                                                                                                                                                     | October 20, 2022  |
| <a href="#">已更新白皮书</a>     | 增加了可持续性支柱并更新了链接。                                                                                                                                                                                                                                          | December 2, 2021  |
| <a href="#">主要更新</a>       | 可持续性支柱已添加到框架。                                                                                                                                                                                                                                             | November 20, 2021 |
| <a href="#">次要更新</a>       | 删除了非包容性用语。                                                                                                                                                                                                                                                | April 22, 2021    |
| <a href="#">次要更新</a>       | 修复了许多链接。                                                                                                                                                                                                                                                  | March 10, 2021    |
| <a href="#">次要更新</a>       | 贯穿全文的次要编辑更改。                                                                                                                                                                                                                                              | July 15, 2020     |
| <a href="#">针对新框架进行了更新</a> | 审核并重写大多数问题和答案。                                                                                                                                                                                                                                            | July 8, 2020      |

|                        |                                                                                                               |                  |
|------------------------|---------------------------------------------------------------------------------------------------------------|------------------|
| <a href="#">已更新白皮书</a> | 增加了 AWS Well-Architected Tool 以及指向 AWS Well-Architected 实验室和 AWS Well-Architected 合作伙伴的链接，进行了小的修复以支持框架的多语言版本。 | July 1, 2019     |
| <a href="#">已更新白皮书</a> | 审核并重写大多数问题和答案，确保问题一次集中在一个主题上。这导致某些之前的问题被拆分成多个问题。向定义中添加了常见术语（工作负载、组件等）。更改了正文中问题的表达以包含描述性文本。                    | November 1, 2018 |
| <a href="#">已更新白皮书</a> | 进行了更新，以简化问题文本、实现答案的标准化和提高可读性。                                                                                 | June 1, 2018     |
| <a href="#">已更新白皮书</a> | 将卓越运营移至第一个支柱，并进行重新编写，以此引出其他支柱。更新了其他支柱以反映 AWS 的发展。                                                             | November 1, 2017 |
| <a href="#">已更新白皮书</a> | 更新了框架，添加了卓越运营支柱，修订并更新了其他支柱，以减少重复并整合从成千上万的客户审查实践中吸取的经验。                                                        | November 1, 2016 |
| <a href="#">次要更新</a>   | 为附录更新了当前的 Amazon CloudWatch Logs 信息。                                                                          | November 1, 2015 |
| <a href="#">原始版本</a>   | 发布了 AWS Well-Architected Framework。                                                                           | October 1, 2015  |

# 附录：问题和最佳实践

此附录汇总了 AWS Well-Architected Framework 中的所有问题和最佳实践。

## 支柱

- [卓越运营](#)
- [安全性](#)
- [可靠性](#)
- [性能效率](#)
- [成本优化](#)
- [可持续性](#)

## 卓越运营

卓越运营支柱能够有效地支持发展和运行工作负载，获取对运营的洞察，以及不断改进支持流程和程序以实现商业价值。如需有关具体实施的说明性指导，请参阅 [《卓越运营支柱》白皮书](#) 访问 AWS 资源。

## 最佳实践领域

- [组织](#)
- [准备](#)
- [运营](#)
- [演进](#)

## 组织

## 问题

- [OPS 1. 您如何确定自己的重点？](#)
- [OPS 2. 如何构建组织结构来为业务成果提供支持？](#)
- [OPS 3. 组织文化如何为业务成果提供支持？](#)

## OPS 1. 您如何确定自己的重点？

每个人都应该了解自己在业务取得成功的过程中扮演的角色。设置共同的目标，以便为资源设定重点。这可以让您的工作效益最大化。

### 最佳实践

- [OPS01-BP01 评估外部客户需求](#)
- [OPS01-BP02 评估内部客户需求](#)
- [OPS01-BP03 评估治理要求](#)
- [OPS01-BP04 评估合规性要求](#)
- [OPS01-BP05 评估威胁形势](#)
- [OPS01-BP06 评估权衡](#)
- [OPS01-BP07 管理收益和风险](#)

### OPS01-BP01 评估外部客户需求

让包括业务、开发和运维团队在内的主要利益相关方参与进来，以便确定将工作重心放在哪里来满足外部客户的需求。这可以确保您充分了解实现您期望的业务成果所需的运营支持。

### 常见反模式：

- 您决定核心业务时间之外不再提供客户支持，但是您还没有查看历史支持请求数据。您不知道这是否会对客户产生影响。
- 您正在开发一项新功能，但尚未与客户沟通，不了解客户是否需要；如果需要，以什么形式提供；并且尚未通过试验来验证交付需求和方法。

建立此最佳实践的好处：需求得到满足的客户流失的可能性更小。评估和了解外部客户需求将为您提供相关信息，告知您如何通过安排工作的优先级来实现商业价值。

未建立此最佳实践暴露的风险等级：高

### 实施指导

- 了解业务需求：包括业务、开发和运营团队在内的利益相关方需要有共同的目标和共同的理解，才能实现业务成功。

- 审查外部客户的业务目标、需求和重点：让包括业务、开发和运营团队在内的主要利益相关方参与进来，讨论外部客户的目标、需求和重点。这可以确保您充分了解实现业务成果和客户成果所需的运营支持。
- 建立共识：建立共识，确定工作负载的业务功能、每个团队在运行工作负载方面的角色，以及这些因素如何支持内部和外部客户共同的业务目标。

## 资源

### 相关文档：

- [AWS Well-Architected Framework 概念 – 反馈循环](#)

## OPS01-BP02 评估内部客户需求

让包括业务、开发和运营团队在内的主要利益相关方参与进来，以便确定怎样将工作重心放在内部客户的需求上。这可以确保您充分了解实现业务成果所需的运营支持。

使用这些已明确的重点，将改进工作集中部署在能发挥最大影响（例如，开发团队技能、提高工作负载性能、降低成本、自动化运行手册或增强监控）的方面。要随着需求的变化更新重点。

### 常见反模式：

- 您决定更改产品团队的 IP 地址分配（没有与他们商议），以便更轻松地管理网络。您不知道这是否会对您的产品团队产生影响。
- 您正在采用一种新的开发工具，但尚未与内部客户沟通，不了解他们是否需要，或者是否与他们的现有实践兼容。
- 您正在实施一个新的监控系统，但尚未与内部客户沟通，不了解他们是否有监控或报告需求需要考虑。

建立此最佳实践的好处：评估和了解内部客户需求将为您提供相关信息，告知您通过安排工作的优先级来实现商业价值。

未建立此最佳实践暴露的风险等级：高

### 实施指导

- 了解业务需求：包括业务、开发和运营团队在内的利益相关方需要有共同的目标和共同的理解，才能实现业务成功。

- 分析内部客户的业务目标、需求和重点：让包括业务、开发和运营团队在内的主要利益相关方参与进来，讨论内部客户的目标、需求和重点。这可以确保您充分了解实现业务成果和客户成果所需的运营支持。
- 建立共识：建立共识，确定工作负载的业务功能、每个团队在运行工作负载方面的角色，以及这些因素如何支持内部和外部客户共同的业务目标。

## 资源

### 相关文档：

- [AWS Well-Architected Framework 概念 – 反馈循环](#)

## OPS01-BP03 评估治理要求

治理是公司用来实现其业务目标的一系列策略、规则或框架。从组织内部生成治理要求。它们会影响您选择的技术类型或影响您运营工作负载的方式。将组织治理要求纳入工作负载中。合规是证明您已实施治理要求的能力。

### 期望结果：

- 将治理要求纳入架构设计和工作负载运营中。
- 您可以提供证据来证明您遵循治理要求。
- 定期审核和更新治理要求。

### 常见反模式：

- 您的组织要求根账户具有多重身份验证。您未能实施此要求，根账户已泄露。
- 在设计工作负载时，您选择了未得到 IT 部门批准的实例类型。您无法启动工作负载，必须重新设计。
- 您需要制定灾难恢复计划。您没有制定灾难恢复计划，且您的工作负载遭受了长时间的中断。
- 您的团队希望使用新实例，但您的治理要求没有更新，不允许使用新实例。

### 建立此最佳实践的好处：

- 遵循治理要求，使您的工作负载与更广泛的组织政策保持一致。
- 治理要求反映了行业标准和您组织的最佳实践。



在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

与利益攸关方和治理组织合作，确定治理要求。将治理要求包括在您的工作负载中。可以提供证据来证明您遵循治理要求。

## 客户示例

在 AnyCompany Retail，云运营团队与整个组织内的利益攸关方一起制定治理要求。例如，他们禁止对 Amazon EC2 实例进行 SSH 访问。如果团队需要系统访问，他们需要使用 AWS Systems Manager Session Manager。随着新服务推出，云运营团队定期更新治理要求。

## 实施步骤

1. 为您的工作负载确定利益攸关方，包括任何集中式团队。
2. 与利益攸关方一起确定治理要求。
3. 生成列表之后，按优先顺序列出改进项，并开始将它们实施到您的工作负载中。
  - a. 使用 [AWS Config](#) 等服务来创建治理即代码，并确认遵循治理要求。
  - b. 如果您使用 [AWS Organizations](#)，则可以利用服务控制策略来实施治理要求。
4. 提供用于验证实施的文档。

实施计划的工作量级别：中等。实施时未满足治理要求可能会导致工作负载返工。

## 资源

相关最佳实践：

- [OPS01-BP04 评估合规性要求](#) - 合规性类似于治理，但它来自组织外部。

相关文档：

- [AWS 管理和治理云环境指南](#)
- [多账户环境中的 AWS Organizations 服务控制策略的最佳实践](#)
- [AWS Cloud 中的治理：敏捷性和安全性之间的适当平衡](#)
- [什么是治理、风险与合规性 \( GRC \) ？](#)

相关视频：

- [AWS 管理和治理：配置、合规性和审计 - AWS 在线技术讲座](#)
- [AWS re:Inforce 2019：云时代的治理 \( DEM12-R1 \)](#)
- [AWS re:Invent 2020：使用 AWS Config 实现合规性即代码](#)
- [AWS re:Invent 2020：AWS GovCloud \(US\) 中的敏捷治理](#)

相关示例：

- [AWS Config 合规包示例](#)

相关服务：

- [AWS Config](#)
- [AWS Organizations - 服务控制策略](#)

#### OPS01-BP04 评估合规性要求

监管、行业和内部合规性要求是定义组织优先级的重要驱动因素。您的合规性框架可能会阻止您使用特定技术或地理位置。如果未确定外部合规框架，则进行尽职调查。生成验证合规性的审计或报告。

如果您宣称自己的产品符合特定的合规性标准，则您必须有内部流程来确保持续合规。合规性标准的示例包括 PCI DSS、FedRAMP 和 HIPAA。适用的合规性标准由各种因素决定，例如解决方案存储或传输的数据类型，以及解决方案支持的地理区域。

期望结果：

- 将监管、行业和内部合规性要求纳入架构选择。
- 您可以验证合规性并生成审计报告。

常见反模式：

- 部分工作负载属于支付卡行业数据安全标准 ( PCI-DSS ) 框架，但您的工作负载以未加密方式存储信用卡数据。
- 软件开发人员和架构师不了解组织必须遵循的合规性框架。
- 年度系统与组织控制 ( SOC2 ) 类型 II 审计即将开始，您无法验证控制措施是否已到位。

建立此最佳实践的好处：

- 评估和了解适用于工作负载的合规性要求将为您提供相关信息，告知您如何通过安排工作的优先级来实现业务价值。
- 选择与合规性框架保持一致的适当位置和技术。
- 设计工作负载以实现可审计性，这样就可以证明您遵守合规性框架。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

实施此最佳实践意味着将合规性要求纳入架构设计过程。您的团队成员了解所需的合规性框架。您依照框架验证合规性。

## 客户示例

AnyCompany Retail 存储客户的信用卡信息。卡存储团队的开发人员明白他们需要遵守 PCI-DSS 框架。他们已采取措施来确认按照 PCI-DSS 框架安全地存储和访问信用卡信息。他们每年都会与安全团队一起验证合规性。

## 实施步骤

1. 与我们的安全性和治理团队合作，确定您的工作负载必须遵守哪些行业、监管或内部合规性框架。将合规性框架纳入您的工作负载。
  - a. 使用 [AWS Compute Optimizer](#) 和 [AWS Security Hub](#) 等服务验证 AWS 资源是否持续合规。
2. 向团队成员介绍合规性要求，以便他们可以根据要求运行和改进工作负载。架构和技术选择中应包括合规性要求。
3. 根据合规性框架，您可能需要生成审计或合规性报告。与组织合作，尽可能使此过程实现自动化。
  - a. 使用 [AWS Audit Manager](#) 等服务验证合规性和生成审计报告。
  - b. 您可以使用 [AWS Artifact](#) 下载 AWS 安全性和合规性文档。

实施计划的工作量级别：中等。实施合规性框架并非易事。生成审计报告或合规性文档带来了额外的复杂性。

## 资源

相关最佳实践：

- [SEC01-BP03 识别并验证控制目标](#) - 安全控制目标是整体合规性的重要组成部分。

- [SEC01-BP06 在管道中自动测试和验证安全控制措施](#) - 作为管道的一部分，验证安全控制。您还可以生成新变更的合规性文档。
- [SEC07-BP02 定义数据保护控制措施](#) - 许多合规性框架都基于数据处理和存储策略。
- [SEC10-BP03 准备取证能力](#) - 有时候审计合规性中会使用取证功能。

相关文档：

- [AWS 合规中心](#)
- [AWS 合规性资源](#)
- [AWS 风险和合规性白皮书](#)
- [AWS 责任共担模式](#)
- [合规性计划范围内的 AWS 服务](#)

相关视频：

- [AWS re:Invent 2020：使用 AWS Compute Optimizer 实现合规性即代码](#)
- [AWS re:Invent 2021 - 云合规性、保证和审计](#)
- [AWS Summit ATL 2022 - 在 AWS 上实施合规性、保证和审计 \( COP202 \)](#)

相关示例：

- [AWS 上的 PCI DSS 和 AWS 基础安全最佳实践](#)

相关服务：

- [AWS Artifact](#)
- [AWS Audit Manager](#)
- [AWS Compute Optimizer](#)
- [AWS Security Hub](#)

## OPS01-BP05 评估威胁形势

评估对业务的威胁（例如竞争、业务风险和负债、运营风险和信息安全威胁），并在风险注册表中维护当前信息。在确定工作重心时，将风险的影响考虑在内。

如示例所示，[Well-Architected Framework](#) 强调学习、衡量和改进。它为您提供了一种一致的方法来评估架构，并实施将随着时间推移而扩展的设计。AWS 提供 [AWS Well-Architected Tool](#)，可帮助您在开发之前查看方法、生产前的工作负载状态以及生产中的工作负载状态。您可以将其与最新的 AWS 架构最佳实践进行比较，监控工作负载的整体状态，并深入了解潜在风险。

AWS 客户可以使用针对任务关键型工作负载的指导式 Well-Architected 审核，以 [根据](#) AWS 最佳实践来衡量其架构。企业支持客户可以使用 [运营审核](#)，该审核旨在帮助他们找出云中的运营方法所存在的漏洞。

这些审核需要跨团队参与，可帮助各团队在工作负载以及他们在实现成功中的角色方面达成一致的理解。通过审查所确定的需求可以帮助确定您的运营重点。

[AWS Trusted Advisor](#) 是一种工具，让您可以访问一组核心检查，这些检查会提出优化建议，帮助确定您的运营重点。[商业和企业支持客户](#) 可以访问其他检查，这些检查重点关注安全性、可靠性、性能和成本优化，可进一步帮助他们帮助确定运营重点。

常见反模式：

- 您在产品中使用的是旧版软件库。对于可能会对工作负载产生意外影响的问题，需要对库进行安全更新，而您忽略了这一点。
- 您的竞争对手刚刚发布了新的产品版本，可以解决许多客户对您产品的投诉。您没有优先解决这些已知问题。
- 监管机构一直在追查像您这样的不符合法律法规要求的公司。您没有优先处理任何未解决的合规性要求。

建立此最佳实践的好处：发现并了解对组织和工作负载的威胁后，您可以确定要解决的威胁、需解决的威胁的优先级以及执行操作所需的资源。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

- 评估威胁形势：评估对业务的威胁（例如竞争、业务风险和负债、运营风险和信息安全威胁），以便您在确定工作重心时可以将其影响考虑在内。
  - [AWS 最新安全公告](#)
  - [AWS Trusted Advisor](#)
- 维护威胁模型：建立并维护威胁模型，确定潜在威胁、计划内和已实施的缓解措施及其优先级。审核威胁酿成意外事件的可能性、从意外事件恢复的成本和预期造成的危害，以及防止这些意外事件发生的成本。根据威胁模型内容的更改修订优先级。

## 资源

相关文档：

- [AWS Cloud 合规性](#)
- [AWS 最新安全公告](#)
- [AWS Trusted Advisor](#)

## OPS01-BP06 评估权衡

在有冲突的利益或替代方法之间做出权衡并评估其影响，以便在确定工作重心或选择行动方案时做出明智的决策。例如，加快新功能上市的速度可能会比成本优化更重要，或者您可以为非关系数据选择关系数据库，以简化迁移系统的工作，而不是迁移到针对您的数据类型优化的数据库和更新您的应用程序。

AWS 可以帮您向团队介绍 AWS 及其服务，让他们深入了解自己的选择会如何影响工作负载。您应该使用由 [AWS 支持](#)（[AWS 知识中心](#)，[AWS 开发论坛](#)和 [AWS 支持 中心](#)）和 [AWS 文档提供的资源](#) 来培训您的团队。请通过 AWS 支持 中心联系 AWS 支持，获取与 AWS 问题相关的帮助。

AWS 还在 Amazon Builders' Library 中分享了我们通过 AWS 运营 [学到的最佳实践和模式](#)。您可以通过 [AWS Blog](#) 和 [AWS 官方播客](#)，获得各种其他有用信息。

常见反模式：

- 您正在使用关系数据库管理时间序列和非关系数据。有一些数据库选项经过优化后可以支持您正在使用的数据类型，但是您没有意识到这些好处，因为您尚未评估解决方案之间的权衡。
- 您的投资者要求您证明符合支付卡行业数据安全标准 (PCI DSS)。您没有在满足他们的要求和继续进行当前的开发工作之间进行权衡，而是在没有证明合规性的情况下继续进行开发工作。投资者出于对平台安全性和投资的担忧停止了对公司的支持。

建立此最佳实践的好处：了解您的选择产生的影响和后果可以帮助您排定选择的优先顺序。

未建立此最佳实践暴露的风险等级：中

## 实施指导

- 评估权衡：在利益互有冲突的目标之间做出权衡并评估其影响，以便在确定工作重心时做出明智的决策。例如，加快新功能上市的速度可能比成本优化更重要。

- AWS 可以帮您向团队介绍 AWS 及其服务，让他们深入了解自己的选择会如何影响工作负载。您应该使用由 AWS 支持（AWS 知识中心、AWS 开发论坛和 AWS 支持中心）和 AWS 文档提供的资源来培训您的团队。请通过 AWS 支持中心联系 AWS 支持，获取与 AWS 问题相关的帮助。
- AWS 还在 Amazon Builders' Library 中分享了我们通过 AWS 运营学到的最佳实践和模式。您可以通过 AWS Blog 和 AWS 官方播客，获得各种其他有用信息。

## 资源

### 相关文档：

- [AWS Blog](#)
- [AWS Cloud 合规性](#)
- [AWS 开发论坛](#)
- [AWS 文档](#)
- [AWS 知识中心](#)
- [AWS 支持](#)
- [AWS 支持中心](#)
- [Amazon Builders' Library](#)
- [AWS 官方播客](#)

## OPS01-BP07 管理收益和风险

管理收益和风险，以便在确定工作重心时做出明智的决策。例如，为了向客户提供重要的新功能，部署仍存在未决问题的的工作负载是可以接受的。这可能会降低相关风险，或者允许风险继续存在可能会令人无法接受，在这种情况下，您将采取措施来化解风险。

您可能会发现，您需要在某个时间点侧重于一小部分运营重点。长期使用平衡的方法来确保所需能力的发展和风险管理。要随着需求的变化更新重点

### 常见反模式：

- 您决定设置一个库，这是您的一位开发人员在互联网上找到的万能库。您尚未评估从未知来源采用此库的风险，也不知道它是否包含漏洞或恶意代码。
- 您决定开发和部署新功能，而不是修复现有问题。您尚未评估在部署功能之前将问题继续留存的风险，也无从知晓会对客户产生哪些影响。
- 由于合规团队提出了未指明的顾虑，您决定不部署客户频繁请求的功能。



建立此最佳实践的好处：确定选择可以带来的收益并了解组织所面临的风险有助于您做出明智的决定。

未建立此最佳实践暴露的风险等级：低

### 实施指导

- 管理收益和风险：在决策的收益与涉及的风险之间取得平衡。
  - 确定收益：根据业务目标、需求和优先事项来确定收益。例如上市时间、安全性、可靠性、性能和成本等。
  - 确定风险：根据业务目标、需求和优先事项来确定风险。例如上市时间、安全性、可靠性、性能和成本等。
  - 对照风险评估收益并做出明智决策：根据包括业务、开发和运营团队在内的主要利益相关方的目标、需求和优先事项，确定收益和风险的影响。对照发生风险的可能性及其影响产生的成本来评估收益的价值。例如，强调上市速度而不是可靠性可能会带来竞争优势。但是如果出现可靠性问题，就可能会导致正常运行时间缩短。

## OPS 2.如何构建组织结构来为业务成果提供支持？

您的团队必须了解他们在实现业务成果方面所发挥的作用。团队应该了解自己在其他团队获得成功的过程中所扮演的角色、其他团队在他们获得成功的过程中所扮演的角色，并设定共同的目标。了解责任分配、所有权归属、决策制定方式以及决策者将有助于集中精力，最大限度地发挥团队的优势。

### 最佳实践

- [OPS02-BP01 确定资源所有者](#)
- [OPS02-BP02 确定流程和程序所有者](#)
- [OPS02-BP03 确定对运营活动绩效负责的所有者](#)
- [OPS02-BP04 团队成员知道自己的责任](#)
- [OPS02-BP05 制定用于确定责任和所有权的机制](#)
- [OPS02-BP06 制定用于请求添加、更改和例外的机制](#)
- [OPS02-BP07 预先定义或协商团队间的职责](#)

## OPS02-BP01 确定资源所有者

工作负载的资源必须具有已确定的所有者，以便实现变更控制、故障排除和其他功能。为工作负载、账户、基础设施、平台 and 应用程序分配所有者。使用集中登记册或附加到资源的元数据等工具记录所有权。组件的商业价值指明应用于它们的流程和程序。

期望结果：

- 使用元数据或集中登记册确定资源所有者。
- 团队成员可以确定谁拥有资源。
- 在可能的情况下，账户只有一个所有者。

常见反模式：

- 未填入 AWS 账户 的备用联系人。
- 资源缺少用于标识哪些团队拥有它们的标记。
- 您的 ITSM 队列没有电子邮件映射。
- 两个团队对一个关键基础设施的所有权出现重叠。

建立此最佳实践的好处：

- 通过分配所有权，资源的变更控制变得非常简单。
- 在排查问题时，您可以让适合的所有者参与进来。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

定义所有权对于环境中的资源使用案例的意义。所有权表示谁监督资源的变更、在排查故障时对资源提供支持或谁在财务上负责。指定并记录资源所有者，包括名称、联系信息、组织和团队。

### 客户示例

AnyCompany Retail 将所有权定义为控制资源变更和支持的团队或个人。他们利用 AWS Organizations 来管理其 AWS 账户。使用组收件箱配置账户备用联系人。每个 ITSM 队列映射到一个电子邮件别名。标签确定谁拥有 AWS 资源。对于其他平台和基础设施，他们有 Wiki 页面，其中确定了所有权和联系信息。

### 实施步骤

1. 首先定义组织的所有权。所有权意味着谁承担资源的风险、谁控制对资源的变更，或在排查故障时谁为资源提供支持。所有权还意味着资源的财务或管理所有权。
2. 使用 [AWS Organizations](#) 来管理账户。您可以集中管理账户的备用联系人。
  - a. 联系信息使用公司拥有的电子邮件地址和电话号码，即使它们所属的个人不再是公司的员工，您仍可以访问它们。例如，为账单、运营和安全性创建单独的电子邮件分发列表，并在各个活跃的 AWS 账户 中将它们配置为账单、安全性和运营联系人。有多个人会收到 AWS 通知，所以即使有人在度假、变更角色或离开公司，也有其他人能够作出回复。
  - b. 如果一个账户未由 [AWS Organizations](#) 管理，则在需要时，备用账户联系人可帮助 AWS 与相关人员联系。将账户的备用联系人配置为指向群组而不是指向个人。
3. 使用标签来识别 AWS 资源的所有者。您可以用单独的标签指定所有者及其联系信息。
  - a. 您可以使用 [AWS Config](#) 规则强制使资源具有所需的所有权标签。
  - b. 有关如何为组织构建标记策略的深入指导，请参阅 [AWS 标记最佳实践白皮书](#)。
4. 对于其他资源、平台和基础设施，创建用于标识所有权的文档。所有团队成员应该都可以访问此文档。

实施计划的工作量级别：低。利用账户联系信息和标签来分配 AWS 资源的所有权。对于其他资源，您可以使用像 Wiki 中的表格这样简单的东西来记录所有权和联系信息，或使用 ITSM 工具来映射所有权。

## 资源

### 相关最佳实践：

- [OPS02-BP02 确定流程和程序所有者](#) - 用于支持资源的流程和程序取决于资源所有权。
- [OPS02-BP04 团队成员知道自己的责任](#) - 团队成员应了解他们是哪些资源的所有者。
- [OPS02-BP05 制定用于确定责任和所有权的机制](#) - 可以使用标签或账户联系人等机制来发现所有权。

### 相关文档：

- [AWS 账户管理 - 更新联系信息](#)
- [AWS Config 规则 - 必需的标签](#)
- [AWS Organizations - 更新组织中的备用联系人](#)
- [AWS 标记最佳实践白皮书](#)

相关示例：

- [AWS Config 规则 - 带有必需标签和有效值的 Amazon EC2](#)

相关服务：

- [AWS Config](#)
- [AWS Organizations](#)

## OPS02-BP02 确定流程和程序所有者

了解谁对各个流程和程序的定义拥有所有权、为何使用这些特定的流程和程序，以及为什么存在这种所有权。了解使用特定流程和程序的原因将有助于发现改进机会。

建立此最佳实践的好处：了解所有权可以确定谁有权批准改进和/或实施改进。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

- 确定负责定义流程和程序的所有者：捕获环境中使用的流程和程序，以及负责其定义的个人或团队。
  - 确定流程和程序：确定为支持工作负载而开展的运营活动。将这些活动记录在易于发现的位置。
  - 确定谁负责定义流程或程序：唯一标识负责活动规范的个人或团队。他们负责确保由技能娴熟且具有正确的权限、访问权限和工具的团队成员来成功执行活动。如果执行活动时遇到问题，那么执行活动的团队成员有责任提供详细反馈，推进活动改进。
  - 在活动构件的元数据中捕获所有权：在 AWS Systems Manager 之类的服务中通过文档和 AWS Lambda 函数自动执行的程序支持以标签形式捕获元数据信息。使用标签或资源组捕获资源所有权，详细说明所有权和联系信息。使用 AWS Organizations 创建标记策略，并确保捕获所有权和联系信息。

## OPS02-BP03 确定对运营活动绩效负责的所有者

了解谁负责针对定义的工作负载执行特定活动，以及为什么负责。了解谁负责执行活动可让我们知晓谁来开展活动、验证结果并向活动所有者提供反馈。

建立此最佳实践的好处：了解谁负责执行活动可让我们知晓需要采取行动时要通知谁以及谁将执行操作、验证结果并向活动所有者提供反馈。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

- 确定对运营活动绩效负责的所有者：了解环境中使用的流程和程序的责任分配
  - 确定流程和程序：确定为支持工作负载而开展的运营活动。将这些活动记录在易于发现的位置。
  - 确定各项活动的执行负责人：确定负责某项活动的团队。确保他们具有活动的详细信息，具备执行活动所需的技能以及正确的权限、访问权限和工具。他们必须了解活动执行条件（例如基于某个事件或计划）。显示这些信息，以便组织成员确定针对特定需求他们需要联系的人员（团队或个人）。

### OPS02-BP04 团队成员知道自己的责任

了解您的角色具有哪些责任以及如何为业务成果做出贡献，这可帮助您确定任务的优先级以及自身角色的重要性。这使团队成员能够了解需求并做出适当响应。

建立此最佳实践的好处：了解您的责任可帮助明确所做的决定、采取的行动以及需要将哪些活动交给适当的所有者。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

- 确保团队成员了解各自的角色和责任：确定团队成员的角色和责任，并确保他们了解对其角色的期望。显示这些信息，以便组织成员确定他们为满足特定需求而需要联系的人员（团队或个人）。

### OPS02-BP05 制定用于确定责任和所有权的机制

在未确定个人或团队时，要为有权分配所有权或计划满足该需求的人定义升级路径。

建立此最佳实践的好处：了解谁负责或拥有所有权可使您与合适的团队或团队成员联系，以提出请求或转换任务。确定谁有权分配责任或所有权或计划满足需求，可以降低不作为的风险并减少无法满足的需求。

未建立此最佳实践暴露的风险等级：高

## 实施指导

- 制定用于确定责任和所有权的机制：为组织成员提供可访问机制，以发现和确定所有权和责任。这些机制将使他们能够根据特定的需求确定相关的联系人（团队或个人）。

## OPS02-BP06 制定用于请求添加、更改和例外的机制

您可以向流程、程序和资源的所有者提出请求。请求包括添加、更改和例外。这些请求都要经过变更管理流程。对收益和风险进行评估之后，做出明智的决定，批准可行和确认合适的请求。

期望结果：

- 您可以根据分配的所有权提出变更流程、程序和资源的请求。
- 以慎重的态度作出变更，权衡益处和风险。

常见反模式：

- 您必须更新部署应用程序的方式，但运营团队无法请求更改部署过程。
- 必须更新灾难恢复计划，但没有可向其请求变更的已确定所有者。

建立此最佳实践的好处：

- 流程、程序和资源会随着需求变化而演进。
- 进行变更时，所有者可以作出明智的决定。
- 以慎重的态度作出变更。

在未建立这种最佳实践的情况下暴露的风险等级：中等

实施指导

为实施这种最佳实践，您需要能够请求对流程、程序和资源作出变更。变更管理流程可以很轻巧。记录变更管理流程。

客户示例

AnyCompany Retail 使用责任分配 ( RACI ) 矩阵来确定谁控制流程、程序和资源的变更。他们有书面的变更管理流程，这些流程轻巧且易于遵循。使用 RACI 矩阵和流程，任何人都可以提交变更请求。

实施步骤

1. 确定工作负载的流程、程序和资源及它们各自的所有者。在知识管理系统中记录它们。
  - a. 如果您未实施 [OPS02-BP01 确定资源所有者](#)、[OPS02-BP02 确定流程和程序所有者](#) 或 [OPS02-BP03 确定对运营活动绩效负责的所有者](#)，请先从实施这些项开始。

2. 与您组织中的利益攸关方合作，制定变更管理流程。该流程应涵盖资源、流程和程序的添加、更改和例外。
  - a. 您可以使用 [AWS Systems Manager Change Manager](#) 作为工作负载资源的变更管理平台。
3. 在知识管理系统中记录变更管理流程。

实施计划的工作量级别：中等。制定变更管理流程需要与整个组织的多个利益攸关方达成一致。

资源

相关最佳实践：

- [OPS02-BP01 确定资源所有者](#) - 在构建变更管理流程之前，需要确定资源的所有者。
- [OPS02-BP02 确定流程和程序所有者](#) - 在构建变更管理流程之前，需要确定流程的所有者。
- [OPS02-BP03 确定对运营活动绩效负责的所有者](#) - 在构建变更管理流程之前，需要确定运营活动的所有者。

相关文档：

- [AWS 规范性指南 - AWS 大型迁移的基础行动手册：创建 RACI 矩阵](#)
- [云中的变更管理白皮书](#)

相关服务：

- [AWS Systems Manager Change Manager](#)

#### OPS02-BP07 预先定义或协商团队间的职责

团队之间具有明确或协商好的协议，规定了团队之间的合作和相互支持方式（例如响应时间、服务级别目标或服务等级协议）。记录团队间沟通渠道。了解团队工作对业务成果以及其他团队和组织的成果的影响，可以确定其任务的优先级，并帮助他们做出适当的响应。

当责任和所有权不确定或未知时，您将面临以下风险：没有及时处理必要的活动，以及在处理这些需求时可能出现工作冗余和潜在冲突。

期望结果：

- 商定并记录团队间工作或支持协议。
- 相互支持或合作的团队有明确的沟通渠道和响应期望。



## 常见反模式：

- 生产中出现问题，两个单独的团队开始彼此独立地排查问题。他们各自为政，这延长了中断时间。
- 运营团队需要开发团队提供帮助，但没有商定好响应时间。请求卡滞在待办事项中。

## 建立此最佳实践的好处：

- 团队知道如何互动和相互支持。
- 知道响应期望。
- 明确定义沟通渠道。

在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

实施这种最佳实践意味着明确团队相互合作的方式。正式协议规定了团队如何协同工作或相互支持。记录团队间沟通渠道。

## 客户示例

AnyCompany Retail 的 SRE 团队与其开发团队达成服务等级协议。开发团队在其工单系统中提出请求后，预计可以在十五分钟内得到答复。如果站点发生中断，则 SRE 团队在开发团队的支持下主导调查。

## 实施步骤

1. 与整个组织的利益攸关方合作，根据流程和程序在团队之间达成一致。
  - a. 如果在两个团队之间分享了流程或程序，则编制有关团队如何协同工作的运行手册。
  - b. 如果团队之间存在依赖关系，请商定请求的响应 SLA。
2. 在知识管理系统中记录责任。

实施计划的工作量级别：中等。如果团队之间还没有达成一致，则需要努力与组织中的利益攸关方达成一致。

## 资源

## 相关最佳实践：

- [OPS02-BP02 确定流程和程序所有者](#) - 在团队之间达成协议之前，必须先确定流程所有权。

- [OPS02-BP03 确定对运营活动绩效负责的所有者](#) - 在团队之间达成协议之前，必须先确定运营活动所有权。

相关文档：

- [AWS Executive Insights - 利用双披萨团队助力创新](#)
- [AWS 上的 DevOps 简介 - 双披萨团队](#)

## OPS 3.组织文化如何为业务成果提供支持？

为您的团队成员提供支持，以便他们可以更有效地采取行动并为您的业务成果提供支持。

最佳实践

- [OPS03-BP01 高管支持](#)
- [OPS03-BP02 赋能团队成员在结果有风险时采取行动](#)
- [OPS03-BP03 鼓励上报](#)
- [OPS03-BP04 沟通及时、清晰、可行](#)
- [OPS03-BP05 鼓励试验](#)
- [OPS03-BP06 支持和鼓励团队成员保持和增强他们的技能组合](#)
- [OPS03-BP07 为团队配置适当的资源](#)
- [OPS03-BP08 鼓励在团队内部和团队之间提出不同的观点](#)

### OPS03-BP01 高管支持

高层领导明确为组织设定期望并评估是否成功。高层领导是采用最佳实践和组织发展的发起人、倡导者和推动者

建立此最佳实践的好处：积极参与的领导层、表达清晰的期望和共同的目标可确保团队成员了解组织对自己的期望。对成功进行评估可以确定阻碍成功的障碍，以便通过发起人、倡导者或他们的代表进行干预，从而消除这些障碍。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

- 高管支持：高级领导层明确为组织设定期望并评估是否成功。高层领导是采用最佳实践和组织发展的发起人、倡导者和推动者

- 设定期望：为您的组织制定和发布目标，包括目标衡量方式。
- 跟踪目标实现情况：定期衡量目标的逐步实现情况并分享结果，以便在结果有风险时可以采取适当的措施。
- 为实现目标提供必要的资源：定期审核资源是否仍然合适，或者根据以下项目决定是否添加资源：新信息、目标变更、责任或您的业务环境。
- 支持您的团队：与团队保持沟通，以便您了解他们的进展情况以及是否受到外部因素的影响。团队受外部因素影响时，需重新评估目标并适当地调整目标。确定阻碍团队进度的障碍。代表团队做出行动，帮助消除障碍，除去不必要的负担。
- 推动最佳实践的采用：认可可量化收益的最佳实践并确定创建者和采用者。鼓励进一步采用，实现更大收益。
- 推动团队发展：打造持续改进的文化。鼓励个人和组织的成长与发展。制定长期目标并为此奋斗，逐步实现成功。根据需求、业务目标和业务环境的变化调整此愿景。

### OPS03-BP02 赋能团队成员在结果有风险时采取行动

工作负载所有者定义了指南和范围，赋能团队成员在结果有风险时做出响应。当事件超出定义的范围时，使用上报机制获取指示。

建立此最佳实践的好处：在早期进行测试和验证更改，可以将解决问题的成本降至最低，并降低对客户的影响。经过测试之后再部署，可以最大限度地减少错误。

未建立此最佳实践暴露的风险等级：高

#### 实施指导

- 赋能团队成员在结果有风险时采取行动：为您的团队成员提供权限、工具和机会，实践有效响应所需的技能。
- 为您的团队成员提供机会，实践响应所需的技能：提供替代的安全环境，以便在其中安全地对流程和程序进行测试和培训。进行实际演练，让团队成员在模拟的安全环境中获得响应现实意外事件的经验。
- 定义并确认团队成员采取行动的权限：通过分配对他们所支持的工作负载和组件的权限和访问权限，来明确定义团队成员采取行动的权限。确认授权他们在结果存在风险时采取行动。

### OPS03-BP03 鼓励上报

团队成员具有相应机制，如果他们认为结果存在风险，鼓励他们向决策者和利益相关者上报问题。应经常尽早上报，以便能够确定风险，并防止造成意外事件。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

- 鼓励经常尽早上报：从组织的角度认可，经常尽早上报是一种最佳实践。组织确认并接受，上报的内容最终可能证明并无依据，但最好要抓住机会预防意外事件的发生，而不要因为没有上报而错失机会。
- 制定上报机制：我们设定明文程序来确定何时应上报以及上报方式。记录权限逐级提升（以采取行动或批准行动）的人员及其联系信息。上报应一直持续，直到团队成员认为他们已将风险移交给能够化解风险的人员，或者他们已联系到对运营该工作负载的风险和责任负责的人员。（即可以最终对工作负载做出决策的人员）。上报内容应包括风险的性质、工作负载的严重性、受影响的人、受到的影响以及紧迫性（即预计影响发生的时间）。
- 保护上报的员工：制定政策保护团队成员，如果他们上报关于决策者或利益相关者未做出响应的问题，保护他们免遭报复。制定适当的机制，确定是否发生了这种情况并做出相应响应。

## OPS03-BP04 沟通及时、清晰、可行

制定相应机制，用于将已知风险和计划内事件及时通知给团队成员。提供必要的相关信息、详细信息和时间（如果可能），为确定是否需要采取措施、需要采取什么措施以及及时采取措施提供支持。例如，提供软件漏洞通知可以加快修补过程；或者，提供计划内促销活动的通知可以实施变更冻结以避免发生服务中断的风险。可以将计划内事件记录在变更日历或维护时间表中，以便团队成员可以确定哪些活动待处理。

### 期望结果：

- 通过沟通提供背景、细节和时间期望。
- 团队成员清楚地知道何时以及如何采取行动回应沟通。
- 利用变更日历提供预期变更的通知。

### 常见反模式：

- 每周会出现几次误报警报。每次发生警报时，您都要关掉通知声音。
- 系统会要求您对安全组进行更改，但未告诉你应何时进行修改。
- 当系统纵向扩展时，您会在聊天中不断收到通知，但无需执行任何操作。您避开聊天频道，而错过了重要通知。
- 在不通知运营团队的情况下对生产作出更改。变更会触发警报，并激活随时待命的团队。

建立此最佳实践的好处：

- 您的组织可避免警报疲劳。
- 团队成员可以在必要的背景和期望下行动。
- 可以在变更时段进行变更，从而降低风险。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

为实施这种最佳实践，您必须与整个组织的利益攸关方合作，商定沟通标准。向您的组织公布这些标准。识别并删除误报或始终打开的警报。使用变更日历，以便团队成员知道何时可以采取行动以及哪些活动待处理。确认附带必要背景的沟通可带来明确的行动。

## 客户示例

AnyCompany Retail 使用聊天作为其主要沟通媒介。警报和其他信息会填入特定渠道。当有人必须采取行动时，会清楚地说明预期的结果，并提供运行手册或行动手册供他们使用。他们使用变更日历来安排生产系统的重大变更。

## 实施步骤

1. 分析警报是否为误报，或警报是否会持续触发。删除或更改这些警报，以便仅在需要人工干预时才触发这些警报。如果触发了警报，则提供运行手册或行动手册。
  - a. 您可以使用 [AWS Systems Manager 文档](#) 为警报编制行动手册和运行手册。
2. 制定合理的机制，以清晰、可操作的方式提供风险或计划内事件的通知，而且要引起足够的注意，以做出适当的响应。使用电子邮件列表或聊天频道在计划事件之前发送通知。
  - a. [Amazon Q Developer in chat applications](#) 可用于在组织的消息传送平台内发送警报和响应事件。
3. 提供可访问的信息源，其中包含计划内事件。通知来自同一系统的计划内事件。
  - a. [AWS Systems Manager 变更日历](#) 可用于创建变更时段，指明何时会发生变更。因而在团队成员可以安全地进行变更时，为他们提供通知。
4. 监控漏洞通知和补丁程序信息，以了解外部漏洞以及与工作负载组件相关的潜在风险。向团队成员发送通知，以便他们可以采取行动。
  - a. 您可以订阅 [AWS 安全公告](#)，以便接收有关 AWS 上漏洞的通知。

## 资源

### 相关最佳实践：

- [OPS07-BP03 使用运行手册执行程序](#) - 在结果已知时提供运行手册，使沟通具有可操作性。
- [OPS07-BP04 根据行动手册调查问题](#) - 如果结果未知，则行动手册使沟通具有可操作性。

### 相关文档：

- [AWS 安全公告](#)
- [开放的 CVE](#)

### 相关示例：

- [Well-Architected 实验室：清单和补丁管理（级别 100）](#)

### 相关服务：

- [Amazon Q Developer in chat applications](#)
- [AWS Systems Manager 变更日历](#)
- [AWS Systems Manager 文档](#)

## OPS03-BP05 鼓励试验

试验是将新想法转化为产品和功能的催化剂。它可加快学习速度，并使团队成员保持兴趣和参与热情。鼓励团队成员经常试验，以便推动创新。即使出现了不希望看到的结果，我们知道什么不该做也是有价值的。团队成员不会因为试验成功但结果不理想而受到惩罚。

### 期望结果：

- 您的组织鼓励试验以促进创新。
- 将试验当作学习的机会。

### 常见反模式：

- 您想要运行 A/B 测试，但没有运行试验的机制。您部署了 UI 更改，但无法对其进行测试。这会造成负面的客户体验。

- 您的公司只有一个模拟和生产环境。没有沙盒环境来试验新功能或产品，因此您必须在生产环境中进行试验。

建立此最佳实践的好处：

- 试验推动创新。
- 通过试验，您可以更快地对用户的反馈作出反应。
- 您的组织发展了一种学习的文化。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

试验应以安全的方式进行。利用多个环境来试验，而不危及生产资源。使用 A/B 测试和功能标记来测试试验。使团队成员能够在沙盒环境中执行试验。

### 客户示例

AnyCompany Retail 鼓励试验。团队成员可以每周使用 20% 的工作时间来试验或学习新技术。他们可以实现创新的沙盒环境。为新功能使用 A/B 测试，用真实的用户反馈来验证它们。

### 实施步骤

1. 与整个组织的领导层合作以支持试验。应鼓励团队成员以安全的方式进行试验。
2. 为团队成员提供可以安全进行试验的环境。他们必须能够访问类似于生产的环境。
  - a. 您可以使用单独的 AWS 账户 来创建用于试验的沙盒环境。[AWS Control Tower](#) 可用于预置这些账户。
3. 使用功能标记和 A/B 测试安全地试验和收集用户反馈。
  - a. [AWS AppConfig Feature Flags](#) 可用于创建功能标记。
  - b. [Amazon CloudWatch Evidently](#) 可用于在有限的部署上运行 A/B 测试。
  - c. 您可以使用 [AWS Lambda 版本](#) 来部署一项功能的新版本以进行 Beta 测试。

实施计划的工作量级别：高。为团队成员提供试验环境和进行试验的安全方法需要大量投资。您可能还需要修改应用程序代码以使用功能标记或支持 A/B 测试。

### 资源

相关最佳实践：



- [OPS11-BP02 在意外事件发生后执行分析](#) - 从事件中学习是创新和试验的重要驱动因素。
- [OPS11-BP03 实施反馈环路](#) - 反馈环路是试验的重要部分。

#### 相关文档：

- [深入了解亚马逊文化：试验、失败和客户至尚](#)
- [在 AWS 中创建和管理沙盒账户的最佳实践](#)
- [营造由云支持的试验文化](#)
- [在 SulAmérica Seguros 实现云中试验和创新](#)
- [试验更多，失败更少](#)
- [使用多个账户组织 AWS 环境 - 沙盒 OU](#)
- [使用 AWS AppConfig Feature Flags](#)

#### 相关视频：

- [AWS On Air，主讲：Amazon CloudWatch Evidently | AWS Events](#)
- [AWS On Air San Fran Summit 2022，主讲：AWS AppConfig Feature Flags 与 Jira 集成](#)
- [AWS re:Invent 2022 - 部署不是发布：使用功能标记控制您的启动 \( BOA305-R \)](#)
- [使用 AWS Control Tower 以编程方式创建 AWS 账户](#)
- [设置使用 AWS Organizations 最佳实践的多账户 AWS 环境](#)

#### 相关示例：

- [AWS 创新沙盒](#)
- [适合电子商务的端到端个性化 101](#)

#### 相关服务：

- [Amazon CloudWatch Evidently](#)
- [AWS AppConfig](#)
- [AWS Control Tower](#)

## OPS03-BP06 支持和鼓励团队成员保持和增强他们的技能组合

团队必须增强自己的技能组合，以采用新技术；并随需求和职责的变化继续提供支持，以支持工作负载。新技术技能的增强通常能提升团队成员满意度并支持创新。支持您的团队成员获取和维护行业认证，以验证和认可他们不断增强的技能。进行交叉培训，以促进知识转移并降低在您失去熟练掌握机构知识、经验丰富的团队成员时产生重大影响的风险。专门安排时间进行学习。

AWS 提供了许多资源，包括 [AWS 入门资源中心](#)，[AWS Blog](#)，[AWS 在线技术讲座](#)，[AWS 活动和网络研讨会](#)，以及 [AWS Well-Architected 实验室](#)，这些资源提供了指导、示例和详细演练，用以培训您的团队。

AWS 还在 Amazon Builders' Library 中分享了我们通过 AWS 运营 [学到的最佳实践和模式](#)；并通过 [AWS Blog](#) 和 [AWS 官方播客](#) 分享了各种实用的教材。

您应该利用 AWS 提供的教育资源，例如 Well-Architected 实验室、[AWS 支持](#)（[AWS 知识中心](#)，[AWS 开发论坛](#) 和 [AWS 支持中心](#)）和 [AWS 文档](#) 来培训您的团队。请通过 AWS 支持中心联系 AWS 支持，获取与 AWS 问题相关的帮助。

[AWS 培训与认证](#) 提供了一些免费培训，可以通过自定进度的数字课程，学习 AWS 的基础知识。您还可以注册讲师指导培训，进一步帮助培养您团队的 AWS 技能。

未建立此最佳实践暴露的风险等级：中

### 实施指导

- 支持和鼓励团队成员保持和增强他们的技能组合：我们必须不断学习，这样才能采用新技术、支持创新，并随需求和责任的变化继续提供支持，以支持工作负载。
- 提供教育资源：专门安排时间，提供培训材料、实验室资源，并支持参加会议和加入专业组织，以便有机会向讲师和同行学习。为初级团队成员提供与高级团队成员接触的机会，可以请高级团队成员作为他们的导师，或允许他们跟随高级团队成员学习并了解方法和技能。鼓励学习与工作没有直接关系的内容，拓展视野。
- 团队教育和跨团队合作：为团队成员的继续教育需求做好规划。为团队成员提供（临时或永久）加入其他团队的机会，以分享技能和最佳实践，惠及整个组织
- 支持获取和维护行业认证：支持团队成员获取和维护行业认证，以验证他们所学到的知识并认可他们的成就。

### 资源

相关文档：

- [AWS 入门资源中心](#)
- [AWS Blog](#)
- [AWS Cloud 合规性](#)
- [AWS 开发论坛](#)
- [AWS 文档](#)
- [AWS 在线技术讲座](#)
- [AWS 活动和网络研讨会](#)
- [AWS 知识中心](#)
- [AWS 支持](#)
- [AWS 培训与认证](#)
- [AWS Well-Architected 实验室](#) ,
- [Amazon Builders' Library](#)
- [AWS 官方播客](#).

## OPS03-BP07 为团队配置适当的资源

培养团队成员的能力，并提供工具和资源来支持工作负载需求。团队成员超负荷工作会增加人为错误导致事故发生的风险。投资于工具和资源（例如，对频繁执行的活动实现自动化）可以提高团队的效率，让他们为其他活动提供支持。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

- 为团队配置适当的资源：确保您了解团队取得的成功，以及推动团队成功或导致团队不成功的因素。采取行动为团队提供适当的资源。
- 了解团队绩效：衡量团队运营成果的实现和资产的开发。跟踪输出和错误率随时间发生的变化。与团队沟通，了解会对他们工作产生影响的挑战（例如责任增加、技术变化、人员流失或支持的客户增加）。
- 了解对团队绩效的影响：与团队保持沟通，以便您了解他们的进展情况以及是否受到外部因素的影响。团队受外部因素影响时，需重新评估目标并适当地调整目标。确定阻碍团队进度的障碍。代表团队做出行动，帮助消除障碍，除去不必要的负担。
- 为团队提供必要资源，助力团队取得成功：定期审核资源是否仍然合适，或者是否需要添加新资源，并做出适当的调整，为团队提供支持。

## OPS03-BP08 鼓励在团队内部和团队之间提出不同的观点

利用跨组织的多样性来寻求多种独特的见解。利用这种见解提高创新能力、对您的假设提出质疑，并降低确认偏差的风险。在团队内部提升包容性、多样性和可达性有助于获取有益的见解。

组织文化会直接影响团队成员的工作满意度和保留率。增强团队成员的参与度和能力，助力业务成功。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

- 寻求不同的观点和视角：鼓励所有人做出贡献。为弱势群体发声。在会议中轮换角色和职责。
- 扩展角色和职责：让团队成员有机会尝试他们可能不会担任的角色。他们将从角色以及与其他团队成员的互动中获得经验和见解，而之前他们可能没有机会与他们互动。他们会将自己的经验和见解赋予新角色，以及就此与新团队成员沟通交流。随着见解的不断增多，可能会出现新的商机，或者可能会发现新的改进机会。让团队成员轮流体验他人日常执行的任务，了解执行这些任务的需求和影响。
- 提供安全舒适的环境：制定相应的政策和控制措施，保护组织内团队成员的身心健康。团队成员应该能够彼此敞开心扉，而不是处在会受到报复的担惊受怕之中。当团队成员处于安全舒适的环境中时，才能有更高的参与热情、更高的工作成效。您的组织越多元化，您就越能更好地理解您所支持的人，包括客户。当您的团队成员感到舒服自在、能够畅所欲言并确信自己的意见会被听到时，他们会更愿意分享有价值的见解（例如营销机会、可访问性需求、尚待开发的细分市场、环境中未被发现的风险）。
- 让团队成员充分参与：为员工提供必要的资源，让他们充分参与到所有与工作相关的活动中。每天都面临各种挑战的团队成员已不断开发应对这些挑战的技能。这些开发的技能独一无二，可以为您的组织带来巨大的效益。根据需要为团队成员提供住所将有助于他们增加对团队的贡献，从而提升您的效益。

## 准备

### 问题

- [OPS 4.如何在工作负载中实现可观测性？](#)
- [OPS 5.如何减少缺陷、简化修复和改进生产流程？](#)
- [OPS 6.您如何缓解部署风险？](#)
- [运营 7.如何知道您已经准备好支持某种工作负载？](#)

## OPS 4.如何在工作负载中实现可观测性？

在工作负载中实现可观测性，以便您可以了解其状态并根据业务需求做出数据驱动型决策。

### 最佳实践

- [OPS04-BP01 识别关键绩效指标](#)
- [OPS04-BP02 实施应用程序遥测](#)
- [OPS04-BP03 实施用户体验遥测](#)
- [OPS04-BP04 实施依赖项遥测](#)
- [OPS04-BP05 实施分布式跟踪](#)

### OPS04-BP01 识别关键绩效指标

要在工作负载中实现可观测性，首先要了解其状态并根据业务需求做出数据驱动型决策。确保监控活动与业务目标相一致的最有效方法之一是，定义和监控关键绩效指标 ( KPI )。

期望的结果：与业务目标紧密协调的高效可观测性实践，确保监控工作始终为切实的业务成果服务。

常见反模式：

- 未定义 KPI：在没有明确 KPI 的情况下工作可能会导致监控过多或过少内容，从而缺少重要信号。
- 静态 KPI：不会随着工作负载或业务目标的变化而重新审视或完善 KPI。
- 不一致：重点关注与业务成果不直接相关或难以与现实问题关联的技术指标。

建立此最佳实践的好处：

- 易于识别问题：业务 KPI 通常比技术指标能够更清楚地揭示问题。与筛查众多技术指标相比，业务 KPI 的下降有助于更有效地查明问题。
- 业务协调：确保监控活动直接支持业务目标。
- 效率：将监控资源和注意力优先放在重要的指标上。
- 积极主动：在问题对业务产生更广泛影响之前识别并解决问题。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

要有效地定义工作负载 KPI，请执行以下操作：

1. 从业务成果着手：在深入研究指标之前，请先了解所需的业务成果。是销售额增加、用户参与度提高还是响应时间更短？
2. 将技术指标与业务目标相关联：并非所有技术指标都会对业务结果产生直接影响。确定那些确实会产生直接影响的指标，但使用业务 KPI 来识别问题通常更为简单。
3. 使用 [Amazon CloudWatch](#)：利用 CloudWatch 定义和监控代表您的 KPI 的指标。
4. 定期审查和更新 KPI：随着工作负载和业务的发展，保持 KPI 的相关性。
5. 让利益相关方参与进来：让技术和业务团队参与定义和审查 KPI。

实施计划的工作量级别：中

资源

相关最佳实践：

- [the section called “OPS04-BP02 实施应用程序遥测”](#)
- [the section called “OPS04-BP03 实施用户体验遥测”](#)
- [the section called “OPS04-BP04 实施依赖项遥测”](#)
- [the section called “OPS04-BP05 实施分布式跟踪”](#)

相关文档：

- [AWS 可观测性最佳实践](#)
- [CloudWatch 用户指南](#)
- [AWS 可观测性 Skill Builder 课程](#)

相关视频：

- [开发可观测性战略](#)

相关示例：

- [One Observability Workshop](#)

## OPS04-BP02 实施应用程序遥测

应用程序遥测是实现工作负载可观测性的基础。发射遥测数据至关重要，它可以提供切实可行的见解，让您了解应用程序的状态以及技术和业务成果的实现情况。从故障排除到衡量新功能的影响或确保与业务关键绩效指标（KPI）保持一致，应用程序遥测可为您构建、操作和演进工作负载的方式提供指导。

指标、日志和跟踪数据构成了可观测性的三个主要支柱。它们用作诊断工具来描述应用程序状态。随着时间的推移，它们会协助创建基线和识别异常情况。但是，为了确保监控活动与业务目标协调一致，定义和监控 KPI 至关重要。与只考虑纯粹的技术指标相比，业务 KPI 通常有助于更轻松地了解问题。

其他遥测类型，例如真实用户监控（RUM）和综合事务，是对这些主要数据源的补充。RUM 让您了解实时用户交互，而综合事务则模拟潜在的用户行为，有助于提前发现瓶颈，以防真实用户遇到瓶颈。

期望的结果：获得有关工作负载性能的可操作见解。这些见解使您能够主动做出性能优化决策，提高工作负载稳定性，简化 CI/CD 流程，并有效地利用资源。

常见反模式：

- 可观测性不完整：忽略将可观测性纳入工作负载的每一层，造成盲点，从而掩盖重要的系统性能和行为洞察。
- 支离破碎的数据视图：当数据分散在多个工具和系统中时，要全面了解工作负载的运行状况和性能，会非常困难。
- 用户报告的问题：这表明缺乏通过遥测和业务 KPI 监控来主动发现问题的功能。

建立此最佳实践的好处：

- 明智的决策：借助从遥测和业务 KPI 中获得的见解，您可以做出以数据为导向的决策。
- 提高运营效率：数据驱动的资源利用率可提高成本效益。
- 增强工作负载稳定性：更快地检测和解决问题，延长正常运行时间。
- 简化 CI/CD 流程：从遥测数据获得的见解有助于完善流程和可靠地交付代码。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

要为您的工作负载实现应用程序遥测，请使用 AWS 服务，例如 [Amazon CloudWatch](#) 和 [AWS X-Ray](#)。Amazon CloudWatch 提供了一套全面的监控工具，让您能够观察 AWS 和本地环境中的资源和



应用程序。它会收集、跟踪和分析指标，整合和监控日志数据，并对资源的变化做出响应，从而增进您对工作负载运行方式的了解。同时，利用 AWS X-Ray，您还可以跟踪、分析和调试应用程序，从而深入了解工作负载的行为。借助服务地图、延迟分布和跟踪时间表等功能，X-Ray 可让您深入了解工作负载的性能和影响工作负载性能的瓶颈。

## 实施步骤

1. 确定要收集哪些数据：确定有助于您深入了解工作负载的运行状况、性能和行为的基本指标、日志和跟踪数据。
2. 部署 [CloudWatch](#) 代理：CloudWatch 代理在从您的工作负载及其底层基础设施中获取系统和应用程序指标和日志方面发挥重要作用。该 CloudWatch 代理还可用于收集 OpenTelemetry 或 X-Ray 跟踪数据，并将其发送到 X-Ray。
3. 定义和监控业务 KPI：建立 [自定义指标](#) 并使其与您的 [业务成果相一致](#) 访问 AWS 资源。
4. 使用 AWS X-Ray 来检测您的应用程序：除了部署 CloudWatch 代理外，还必须 [检测您的应用程序](#) 以发出跟踪数据。此过程可让您进一步了解工作负载行为和性能。
5. 标准化整个应用程序中的数据收集：标准化整个应用程序中的数据收集实践。统一性有助于关联和分析数据，从而全面了解应用程序的行为。
6. 分析数据并据此采取行动：数据收集和规范化完成后，使用 [Amazon CloudWatch](#) 来分析指标和日志，并使用 [AWS X-Ray](#) 来分析跟踪数据。此类分析可得出有关您的工作负载的运行状况、性能和行为的重要见解，从而指导您的决策过程。

实施计划的工作量级别：高

## 资源

相关最佳实践：

- [OPS04-BP01 识别关键绩效指标](#)
- [OPS04-BP03 实施用户体验遥测](#)
- [OPS04-BP04 实施依赖项遥测](#)
- [OPS04-BP05 实施分布式跟踪](#)

相关文档：

- [AWS 可观测性最佳实践](#)
- [CloudWatch 用户指南](#)

- [AWS X-Ray 开发人员指南](#)
- [检测分布式系统的运营可见性](#)
- [AWS 可观测性 Skill Builder 课程](#)
- [Amazon CloudWatch 新增功能](#)
- [AWS X-Ray 新增功能](#)

相关视频：

- [AWS re:Invent 2022 – Amazon 的可观测性最佳实践](#)
- [AWS re:Invent 2022 - 开发可观测性战略](#)

相关示例：

- [One Observability Workshop](#)
- [AWS 解决方案库：使用 Amazon CloudWatch 进行应用程序监控](#)

## OPS04-BP03 实施用户体验遥测

深入了解客户体验以及与应用程序的交互至关重要。真实用户监控 ( RUM ) 和综合事务是实现此目的的强大工具。RUM 提供有关真实用户交互的数据，从未经过滤的视角反映用户满意度，而综合事务可模拟用户交互，有助于在潜在问题影响真实用户之前就发现它们。

期望的结果：全面了解客户体验，主动检测问题，优化用户互动，以提供无缝的数字体验。

常见反模式：

- 应用程序没有真实用户监控 ( RUM ) 功能
  - 问题检测被延误：如果没有 RUM，可能要等到用户抱怨时，您才会意识到性能瓶颈或问题。这种被动应对的方法可能会导致客户不满。
  - 缺乏对客户体验的了解：不使用 RUM 意味着您无法掌握揭示真实用户如何与应用程序交互的关键数据，从而限制您优化用户体验的能力。
- 应用程序缺乏综合事务
  - 错过边缘案例：综合事务有助于您测试普通用户可能不经常使用、但对某些业务职能至关重要的路径和功能。没有它们，这些路径可能会出现故障并被忽视。
  - 在应用程序未使用时检查问题：定期的综合测试可以模拟真实用户未积极与应用程序交互时的情况，确保系统始终正常运行。

## 建立此最佳实践的好处：

- 主动检测问题：在潜在问题影响真实用户之前，识别并解决这些问题。
- 优化用户体验：来自 RUM 的持续反馈有助于完善和增强整体用户体验。
- 获得有关设备和浏览器性能的见解：了解您的应用程序在各种设备和浏览器上的表现，从而实现进一步优化。
- 经过验证的业务工作流程：定期的综合事务可确保核心功能和关键路径始终可以使用且高效。
- 增强应用程序性能：利用从真实用户数据中收集的见解，提高应用程序的响应能力和可靠性。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

为了利用 RUM 和综合事务进行用户活动遥测，AWS 提供多项服务，例如 [Amazon CloudWatch RUM](#) 和 [Amazon CloudWatch Synthetics](#)。指标、日志和跟踪，再加上用户活动数据，可让您全面了解应用程序的运行状态和用户体验。

## 实施步骤

1. 部署 Amazon CloudWatch RUM：将您的应用程序与 CloudWatch RUM 集成，收集、分析和呈现真实的用户数据。
  - a. 使用 [CloudWatch RUM JavaScript 库](#) 将 RUM 与您的应用程序集成。
  - b. 设置控制面板以可视化形式呈现和监控真实的用户数据。
2. 配置 CloudWatch Synthetics：创建金丝雀或脚本化例程，模拟用户与应用程序的交互。
  - a. 定义关键应用程序工作流程和路径。
  - b. 使用 [CloudWatch Synthetics 脚本](#) 设计金丝雀，模拟用户在这些路径上的交互。
  - c. 安排和监控金丝雀按指定的间隔运行，确保一致的性能检查。
3. 分析数据并据此采取行动：利用来自 RUM 和综合事务的数据来获取见解，并在检测到异常时采取纠正措施。使用 CloudWatch 控制面板和警报及时了解情况。

实施计划的工作量级别：中

## 资源

## 相关最佳实践：

- [OPS04-BP01 识别关键绩效指标](#)

- [OPS04-BP02 实施应用程序遥测](#)
- [OPS04-BP04 实施依赖项遥测](#)
- [OPS04-BP05 实施分布式跟踪](#)

相关文档：

- [Amazon CloudWatch RUM 指南](#)
- [Amazon CloudWatch Synthetics 指南](#)

相关视频：

- [使用 Amazon CloudWatch RUM 通过最终用户洞察优化应用程序](#)
- [AWS on Air ft.Real-User Monitoring for Amazon CloudWatch](#)

相关示例：

- [可观测性研讨会](#)
- [适用于 Amazon CloudWatch RUM Web 客户端的 Git 存储库](#)
- [使用 Amazon CloudWatch Synthetics 来测量页面加载时间](#)

## OPS04-BP04 实施依赖项遥测

要想监控您的工作负载所依赖的外部服务和组件的运行状况及性能，依赖项遥测必不可少。它提供有关与 DNS、数据库或第三方 API 等依赖项相关的可访问性、超时及其他关键事件的宝贵见解。通过检测您的应用程序以发出有关这些依赖关系的指标、日志和跟踪，您可以更清楚地了解可能影响您的工作负载的潜在瓶颈、性能问题或故障。

期望的结果：您的工作负载所依赖的依赖项按预期执行，使您能够主动解决问题并确保最佳的工作负载性能。

常见反模式：

- 忽略外部依赖项：仅关注内部应用程序指标，而忽略与外部依赖项相关的指标。
- 缺乏主动监控：等待问题出现，而不是持续监控依赖项运行状况和性能。
- 孤立监控：使用多种不同的监控工具，这可能会导致依赖项运行状况视图支离破碎且不一致。

## 建立此最佳实践的好处：

- 提高工作负载可靠性：通过确保外部依赖项始终可用且性能出色来实现。
- 更快地检测 and 解决问题：在依赖项问题影响工作负载之前，主动识别和解决这些问题。
- 全面视图：全面了解影响工作负载运行状况的内部和外部组件。
- 增强工作负载可扩展性：通过了解外部依赖项的可扩展性限制和性能特征来实现。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

从确定您的工作负载所依赖的服务、基础设施和流程开始，实施依赖项遥测。量化这些依赖项按预期运行时的良好状况，然后确定需要哪些数据来衡量这些状况。利用这些信息，您可以创建控制面板和警报，为运营团队提供有关这些依赖项状态的见解。当依赖项无法按需交付时，使用 AWS 工具来发现和量化影响。不断重新审视您的策略，以考虑优先事项、目标和所获见解的变化。

## 实施步骤

要有效地实现依赖项遥测，请执行以下操作：

1. 确定外部依赖项：与利益相关方合作，查明您的工作负载所依赖的外部依赖项。外部依赖项可包括外部数据库、第三方 API、通往其他环境的网络连接路由以及 DNS 服务等内容。实现有效的依赖项遥测的第一步是全面了解这些依赖项是什么。
2. 制定监控策略：一旦您清楚地了解了外部依赖项，就可以针对它们设计一个量身定制的监控策略。这包括了解每个依赖项的重要程度、其预期行为以及任何相关的服务级别协议或目标（SLA 或 SLT）。设置主动警报，在出现状态变化或性能偏差时通知您。
3. 利用 [Amazon CloudWatch 网络监测仪](#)：它提供对全球互联网的见解，有助于了解可能影响外部依赖项的中断。
4. 随时了解最新信息 - 借助 [AWS Health Dashboard](#)：当 AWS 遇到可能会影响您的服务的事件时，它会提供警报和修正指导。
5. 检测您的应用程序 - 使用 [AWS X-Ray](#)：AWS X-Ray 让您能够深入了解应用程序及其底层依赖项的运行情况。通过从头到尾跟踪请求，您可以找出应用程序所依赖的外部服务或组件中的瓶颈或故障。
6. 使用 [Amazon DevOps Guru](#)：这项服务由机器学习驱动，可识别操作问题，预测何时可能出现严重问题，并建议可采取的具体行动。这对于深入了解依赖项并确定它们是不是造成操作问题的根源非常重要。
7. 定期监控：持续监控与外部依赖项相关的指标和日志。针对意外行为或性能下降设置警报。

8. 更改后验证：每当任何外部依赖项有更新或更改时，都应验证其性能，并检查它们是否符合应用程序的要求。

实施计划的工作量级别：中

资源

相关最佳实践：

- [OPS04-BP01 识别关键绩效指标](#)
- [OPS04-BP02 实施应用程序遥测](#)
- [OPS04-BP03 实施用户体验遥测](#)
- [OPS04-BP05 实施分布式跟踪](#)

相关文档：

- [什么是 AWS Health?](#)
- [Using Amazon CloudWatch Internet Monitor](#)
- [AWS X-Ray 开发人员指南](#)
- [Amazon DevOps Guru 用户指南](#)

相关视频：

- [Visibility into how internet issues impact app performance](#)
- [Introduction to Amazon DevOps Guru](#)

相关示例：

- [Gaining operational insights with AIOps using Amazon DevOps Guru](#)
- [AWS Health Aware](#)

## OPS04-BP05 实施分布式跟踪

分布式跟踪提供了一种方法，可在请求通过分布式系统的各个组件时对其进行监控和可视化。通过从多个来源捕获跟踪数据并在一个统一视图中对其进行分析，团队可以更好地了解请求是如何流动的、哪里存在瓶颈以及优化工作的重点。

期望的结果：全面了解流经分布式系统的请求，从而精确调试、优化性能和改善用户体验。

常见反模式：

- 检测不一致：并非分布式系统中的所有服务都经过跟踪检测。
- 忽略延迟：只关注错误，而不考虑延迟或性能逐渐下降的情况。

建立此最佳实践的好处：

- 全面了解系统：以可视化方式呈现请求从进入到退出的整个路径。
- 增强调试功能：快速识别出现故障或性能问题的地方。
- 改善用户体验：监控并根据实际用户数据进行优化，确保系统满足现实需求。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

首先确定工作负载中所有需要检测的元素。将所有组件都考虑在内后，利用 AWS X-Ray 和 OpenTelemetry 之类的工具收集跟踪数据，以便使用 X-Ray 和 Amazon CloudWatch ServiceLens Map 等工具进行分析。定期与开发人员一起进行审查，并使用 Amazon DevOps Guru、X-Ray Analytics 和 X-Ray Insights 等工具来补充这些讨论，以挖掘更深层次的信息。根据跟踪数据确立警报，以便在结果面临风险时，按照工作负载监控计划中定义的流程发出通知。

## 实施步骤

要有效地实施分布式跟踪，请执行以下操作：

1. 采用 [AWS X-Ray](#)：将 X-Ray 集成到您的应用程序中，以深入了解其行为和性能并查明瓶颈。利用 X-Ray Insights 自动分析跟踪数据。
2. 检测您的服务：确认从 [AWS Lambda](#) 函数到 [EC2 实例](#) 的每项服务都发送跟踪数据。您检测的服务越多，端到端视图就越清晰。
3. 纳入 [CloudWatch 真实用户监控](#) 和 [合成监控](#)：将真实用户监控（RUM）和合成监控与 X-Ray 集成。这允许捕捉现实世界的用户体验并模拟用户交互，以识别潜在问题。
4. 使用 [CloudWatch 代理](#)：代理可以从 X-Ray 或 OpenTelemetry 发送跟踪，从而增强所获得见解的深度。
5. 使用 [Amazon DevOps Guru](#)：DevOps Guru 使用来自 X-Ray、CloudWatch、AWS Config 和 AWS CloudTrail 的数据来提供可行的建议。



6. 分析跟踪数据：定期查看跟踪数据，以识别可能影响应用程序性能的模式、异常或瓶颈。
7. 设置警报：在 [CloudWatch](#) 中针对异常模式或过长的延迟时间配置警报，以便于主动解决问题。
8. 持续改进：在添加或修改服务时，重新审视您的跟踪策略，以捕获所有相关数据点。

实施计划的工作量级别：中

资源

相关最佳实践：

- [OPS04-BP01 识别关键绩效指标](#)
- [OPS04-BP02 实施应用程序遥测](#)
- [OPS04-BP03 实施用户体验遥测](#)
- [OPS04-BP04 实施依赖项遥测](#)

相关文档：

- [AWS X-Ray 开发人员指南](#)
- [Amazon CloudWatch 代理用户指南](#)
- [Amazon DevOps Guru 用户指南](#)

相关视频：

- [Use AWS X-Ray Insights](#)
- [AWS on Air ft. Observability: Amazon CloudWatch and AWS X-Ray](#)

相关示例：

- [Instrumenting your Application with AWS X-Ray](#)

## OPS 5.如何减少缺陷、简化修复和改进生产流程？

采用可改进将更改应用于生产环境的流程的方法，激活重构、快速质量反馈和错误修复。这些方法可以加快有益更改进入生产环境的速度、减少产生的问题，并能够快速识别和修复通过部署活动引入的问题。

## 最佳实践

- [OPS05-BP01 使用版本控制](#)
- [OPS05-BP02 测试并验证变更](#)
- [OPS05-BP03 使用配置管理系统](#)
- [OPS05-BP04 使用构建和部署管理系统](#)
- [OPS05-BP05 执行补丁管理](#)
- [OPS05-BP06 共享设计标准](#)
- [OPS05-BP07 实施提高代码质量的实践](#)
- [OPS05-BP08 使用多个环境](#)
- [OPS05-BP09 频繁进行可逆的小规模更改](#)
- [OPS05-BP10 完全自动化集成和部署](#)

### OPS05-BP01 使用版本控制

使用版本控制来跟踪更改和发布。

许多 AWS 服务都提供版本控制功能。使用修订或源代码控制系统（如 [AWS CodeCommit](#)）管理代码和其他构件，如基础设施的版本控制的 [AWS CloudFormation](#) 模板。

期望的结果：您的团队就代码开展协作。合并后，代码将保持一致，并且不会丢失任何更改。通过正确的版本控制，可以很容易纠正错误。

常见反模式：

- 您一直在工作站上开发和存储代码。工作站上发生了不可恢复的存储故障，您的代码丢失了。
- 用更改内容覆盖现有代码后，您重新启动应用程序，但其无法运行。您无法撤消所做更改。
- 您对报告文件执行了写入锁定，而其他人需要对此文件进行编辑。他们与您联系要求您停止写入锁定，以便他们可以完成自己的任务。
- 您的研究团队一直在进行详细的分析，以便对未来的工作进行规划。有人不小心把购物单保存在最终报告上了。您无法撤消更改，不得不重新创建报告。

建立此最佳实践的好处：借助版本控制功能，您可以轻松地恢复到已知的良好状态和以前的版本，并降低资产丢失的风险。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

在受到版本控制的存储库中维护资产。这让您能够跟踪更改、部署新版本、检测对现有版本的更改，以及恢复到以前的版本（例如在发生故障时回滚到已知的良好状态）。将配置管理系统的版本控制功能集成到程序中。

## 资源

相关最佳实践：

- [OPS05-BP04 使用构建和部署管理系统](#)

相关文档：

- [什么是 AWS CodeCommit？](#)

相关视频：

- [AWS CodeCommit 简介](#)

## OPS05-BP02 测试并验证变更

部署的每一项变更都必须经过测试，以避免在生产中出现错误。此最佳实践的重点是测试从版本控制到构件构建的变更。除应用程序代码变更外，测试还应该包括基础设施、配置、安全控制和操作程序。测试有多种形式，从单元测试到软件组件分析（SCA）等等。在软件集成和交付过程中，尽早进行测试可进一步确保构件质量。

您的组织必须为所有的软件构件制定测试标准。自动化测试可以减少工作量，并避免人工测试的错误。有些情况下，可能必须进行手动测试。开发人员必须能够访问自动化测试结果，以创建反馈循环，提高软件质量。

期望的结果：软件更改须在交付前进行测试。开发人员可以访问测试结果和验证结果。您的组织有一个适用于所有软件更改的测试标准。

常见反模式：

- 您在没有进行任何测试的情况下部署一项新软件更改。它无法在生产环境中运行，从而导致中断。
- 使用 AWS CloudFormation 部署新安全组，而没有在生产前环境中进行测试。这些安全组使客户无法访问您的应用程序。
- 修改了一个方法，但没有进行单元测试。该软件在部署到生产环境中后无法运行。

建立此最佳实践的好处：软件部署的更改失败率降低。软件质量得到改进。开发人员提高了对其代码可行性的认识。可以放心地推出安全策略，以支持组织实现合规性。可以提前测试基础设施更改（如自动扩缩策略的更新），以满足流量需求。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

作为持续集成实践的一部分，对从应用程序代码到基础设施的所有更改都进行测试。将公布测试结果，以便开发人员快速提供反馈。您的组织有一个测试标准，所有更改都必须通过测试。

## 客户示例

作为持续集成管道的一部分，AnyCompany Retail 对所有软件构件进行几种类型的测试。他们实行测试驱动型开发，因此所有软件都有单元测试。构件构建完毕后，他们会立即运行端到端测试。第一轮测试完成后，他们会运行静态应用程序安全扫描，寻找已知漏洞。在每个测试关口通过时，开发人员都会收到消息。所有测试均完成后，软件构件就会存储在构件库中。

## 实施步骤

1. 与您组织中的利益相关方合作，为软件构件制定测试标准。所有构件均应通过哪些标准测试？是否有合规性或治理要求必须包括在测试范围内？您是否需要进行代码质量测试？测试完成后，需要通知谁？
  - a. 此 [AWS 部署管道参考架构](#) 包含一个权威的测试类型列表，可作为集成管道的一部分对软件构件执行这些测试。
2. 根据您的软件测试标准，利用必要的测试来检测您的应用程序。每组测试应在 10 分钟内完成。测试应该作为集成管道的一部分运行。
  - a. [Amazon CodeGuru Reviewer](#) 可以测试您的应用程序代码是否存在缺陷。
  - b. 您可以使用 [AWS CodeBuild](#) 对软件构件执行测试。
  - c. [AWS CodePipeline](#) 可以将您的软件测试编排到管道中。

## 资源

相关最佳实践：

- [OPS05-BP01 使用版本控制](#)
- [OPS05-BP06 共享设计标准](#)
- [OPS05-BP10 完全自动化集成和部署](#)

## 相关文档：

- [采用测试驱动型开发方法](#)
- [使用 TaskCat 和 AWS CloudFormation 的 CodePipeline 自动化测试管道](#)
- [使用开源 SCA、SAST 和 DAST 工具构建端到端 AWS DevSecOps CI/CD 管道](#)
- [开始测试无服务器应用程序](#)
- [我的 CI/CD 管道统领我的发布](#)
- [《在 AWS 上实践持续集成和持续交付》白皮书](#)

## 相关视频：

- [AWS re:Invent 2020：可测试的基础设施：AWS 上的集成测试](#)
- [2021 AWS 峰会（澳大利亚和新西兰）- 用 CDK 和测试驱动型开发推动测试优先战略](#)
- [用 AWS CDK 测试您的基础设施即代码](#)

## 相关资源：

- [AWS 部署管道参考架构 - 应用程序](#)
- [AWS Kubernetes DevSecOps 管道](#)
- [策略即代码研讨会 – 测试驱动型开发](#)
- [通过使用 AWS CodeBuild 从 GitHub 为 Node.js 应用程序运行单元测试](#)
- [使用 Serverspec 对基础设施代码进行测试驱动型开发](#)

## 相关服务：

- [Amazon CodeGuru Reviewer](#)
- [AWS CodeBuild](#)
- [AWS CodePipeline](#)

## OPS05-BP03 使用配置管理系统

使用配置管理系统来实现和跟踪配置更改。这些系统可以减少手动过程引起的错误，并减少部署更改的工作量。

静态配置管理在初始化资源时设置值，这些值在资源的生命周期内预期会保持一致。这样的例子包括为实例上的 Web 或应用程序服务器设置配置，或者定义 AWS 服务的配置（在 [AWS Management Console](#) 内或者通过 [AWS CLI](#)）。

动态配置管理在初始化时设置值，这些值在资源的生命周期内可能或预期会发生变化。例如，您可以设置一个功能切换，通过配置更改在代码中激活功能，或者在意外事件期间更改日志详细级别以捕获更多数据，然后在意外事件完成后更改回来，避免不再必要的日志记录及其相关费用。

在 AWS 上，您可以使用 [AWS Config](#) 持续监控 AWS 资源配置 - [跨账户和区域](#)。这有助于您跟踪其配置历史记录，了解配置更改会如何影响其他资源，并使用 [AWS Config 规则](#) 和 [AWS Config 合规包](#) 根据预期或所需的配置审计它们。

如果您在 Amazon EC2 实例、AWS Lambda、容器、移动应用程序或物联网设备上运行的应用程序具有动态配置，则可以使用 [AWS AppConfig](#) 在您的环境中配置、验证、部署和监控它们。

在 AWS 中，您可以使用像 [AWS 开发人员工具](#)（例如，[AWS CodeCommit](#)、[AWS CodeBuild](#)、[AWS CodePipeline](#)、[AWS CodeDeploy](#) 和 [AWS CodeStar](#)）这样的服务来构建持续集成/持续部署（CI/CD）管道。

期望的结果：可以作为持续集成、持续交付（CI/CD）管道的一部分进行配置、验证和部署。通过监控来验证配置是否正确。这样可以最大限度地减少对最终用户和客户的任何影响。

常见反模式：

- 您手动更新整个队列中的 Web 服务器配置，由于更新错误，许多服务器变得没有响应。
- 手动更新应用程序服务器队列需要花费很长时间。在变更过程中，如果配置不一致会导致意外行为发生。
- 有人更新了您的安全组，您的 Web 服务器无法访问了。如果不知道发生了哪些变更，您需要花费大量时间来调查问题，导致恢复时间延长。
- 未经验证即通过 CI/CD 将预生产配置推送到生产环境中。您让用户和客户接触到不正确的数据和服务。

建立此最佳实践的好处：采用配置管理系统可以减少更改及对其进行跟踪的工作量，还可以降低手动程序导致错误的频率。配置管理系统为治理、合规性和监管要求提供了保障。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

配置管理系统用于跟踪和实施对应用程序和环境配置的更改。配置管理系统还用于减少手动流程引起的错误，使配置更改可重复且可审核，并减少工作量。

## 实施步骤

1. 确定配置负责人。
  - a. 让配置负责人了解任何合规性、治理或监管需求。
2. 确定配置项和可交付成果。
  - a. 配置项是受您的 CI/CD 管道内的部署影响的所有应用程序和环境配置。
  - b. 可交付成果包括成功标准、验证和要监控的内容。
3. 根据您的业务需求和交付管道，选择配置管理工具。
4. 考虑使用加权部署，例如用于重大配置更改的金丝雀部署，以尽可能减少错误配置的影响。
5. 将您的配置管理集成到 CI/CD 管道中。
6. 验证所有推送的更改。

## 资源

相关最佳实践：

- [OPS06-BP01 针对不成功的更改制定计划](#)
- [OPS06-BP02 测试部署](#)
- [OPS06-BP03 采用安全部署策略](#)
- [OPS06-BP04 自动测试和回滚](#)

相关文档：

- [AWS Control Tower](#)
- [AWS 登录区加速器](#)
- [AWS Config](#)
- [什么是 AWS Config？](#)
- [AWS AppConfig](#)
- [什么是 AWS CloudFormation？](#)
- [AWS 开发人员工具](#)



相关视频：

- [AWS re:Invent 2022 - Proactive governance and compliance for AWS workloads](#)
- [AWS re:Invent 2020：使用 AWS Config 实现合规性即代码](#)
- [Manage and Deploy Application Configurations with AWS AppConfig](#)

## OPS05-BP04 使用构建和部署管理系统

使用构建和部署管理系统。这些系统可以减少手动过程引起的错误，并减少部署更改的工作量。

在 AWS 中，您可以使用像 [AWS 开发人员工具](#)（例如，AWS CodeCommit、[AWS CodeBuild](#)、[AWS CodePipeline](#)、[AWS CodeDeploy](#)和 [AWS CodeStar](#)）这样的服务来构建持续集成/持续部署（CI/CD）管道。

期望的结果：您的构建和部署管理系统支持组织的持续集成/持续交付（CI/CD）系统，后者提供使用正确的配置自动进行安全部署的功能。

常见反模式：

- 在开发系统上编译代码后，您将可执行文件复制到生产系统上，但它无法启动。本地日志文件显示这是因为缺少依赖项。
- 您成功地在开发环境中构建了具有新功能的应用程序，并将代码送交质量检查（QA）。由于缺少静态资产，它没有通过质量检查。
- 星期五，经过大量的努力，您成功地在开发环境中手动构建了应用程序，包括新编码的功能。星期一，您无法重复这一成功构建应用程序的步骤。
- 您执行为新版本创建的测试。下周，您将设置测试环境，并执行所有现有的集成测试，然后执行性能测试。新代码产生了难以接受的性能影响，因此必须重新开发并测试。

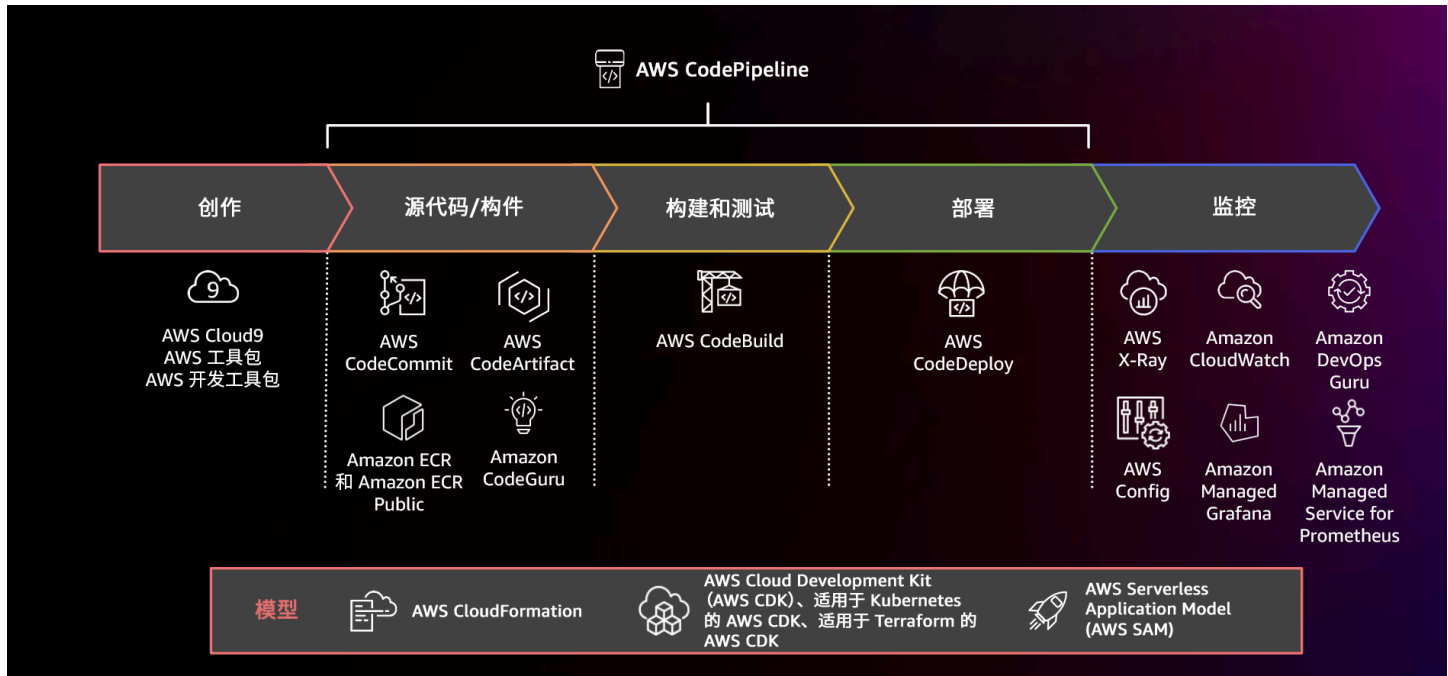
建立此最佳实践的好处：制定相应机制来管理活动的构建和部署。这样，您可以减少执行重复任务的工作量，让团队成员腾出时间专注于高价值的创造性任务，还可以减少手动程序导致的错误。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

构建和部署管理系统用于跟踪和实施变更，减少手动过程引起的错误，并减少安全部署所需的工作量。将集成和部署管道完全自动化，从代码签入到构建、测试、部署和验证都包含在内。这可以缩短准备时间，降低成本，鼓励更频繁地进行变更，减少工作量并增进协作。

## 实施步骤



显示使用 AWS CodePipeline 和相关服务的 CI/CD 管道的示意图

1. 使用 AWS CodeCommit 对资产（例如文档、源代码和二进制文件）进行版本控制、存储和管理。
2. 使用 CodeBuild 编译源代码、运行单元测试和生成可随时部署的构件。
3. 使用 CodeDeploy 作为部署服务，自动将应用程序部署到 [Amazon EC2](#) 实例、本地实例、[无服务器 AWS Lambda 函数](#) 或 [Amazon ECS](#)。
4. 监控您的部署。

## 资源

相关最佳实践：

- [OPS06-BP04 自动测试和回滚](#)

相关文档：

- [AWS 开发人员工具](#)
- [什么是 AWS CodeCommit？](#)
- [什么是 AWS CodeBuild？](#)

- [AWS CodeBuild](#)
- [什么是 AWS CodeDeploy ?](#)

相关视频：

- [AWS re:Invent 2022 - AWS Well-Architected best practices for DevOps on AWS](#)

## OPS05-BP05 执行补丁管理

执行补丁管理以便实现功能、解决问题并保持监管合规性。实现自动补丁管理，以便减少手动过程引起的错误，进行扩展，并减少修补工作量。

补丁和漏洞管理是优势和风险管理活动的一部分。最好是具有不可变的基础设施和已在验证的已知良好状态下部署工作负载。如果该方法不可行，那就只能进行修补。

[Amazon EC2 Image Builder](#) 提供更新计算机映像的管道。作为补丁管理的一部分，请考虑 [亚马逊机器映像](#)（AMI）（使用 [AMI 映像管道](#)）或带 [Docker 映像管道](#) 的容器镜像，同时 AWS Lambda 会提供 [自定义运行时模式和其他库](#) 以消除漏洞。

您应使用以下工具来管理适用于 Linux 或 Windows Server 映像的 [亚马逊机器映像](#) 的更新：[Amazon EC2 Image Builder](#)。您可以将 [Amazon Elastic Container Registry \(Amazon ECR\)](#) 与现有管道配合使用，以管理 Amazon ECS 映像和 Amazon EKS 映像。Lambda 包括 [版本管理功能](#)。

在未事先在安全环境中测试的情况下，不对生产系统执行修补操作。仅当补丁支持操作或业务结果时，才应该应用补丁。在 AWS 上，您可以使用 [AWS Systems Manager Patch Manager](#) 来自动执行修补托管系统的过程和安排修补活动 - 同时使用 [Systems Manager 维护时段](#)。

期望的结果：您的 AMI 和容器映像已修补、处于最新状态，随时可以启动。您可以跟踪所有已部署映像的状态，并了解补丁合规性。您可以报告当前状态，并有一个流程来满足您的合规需求。

常见反模式：

- 您接到任务，需要在两个小时内应用所有新的安全补丁，但由于应用程序与补丁不兼容，导致了多次停机。
- 没有安装补丁的库会引发意外后果，这是因为未知方会利用其中的漏洞来访问您的工作负载。
- 您在未通知开发人员的情况下自动修补开发人员环境。您收到来自开发人员的多起投诉，称他们的环境不能按预期运行。
- 您尚未修补持久性实例上的现有商用软件。当您遇到软件问题并与供应商联系时，他们告知您已不再为该版本提供支持，您必须安装特定级别的补丁才能获得帮助。

- 您使用的加密软件最近发布了新补丁，对性能进行了重大改进。您未安装补丁的系统仍然存在性能问题，恰恰是因为没有安装补丁造成的。
- 您收到通知，告知您存在零日漏洞，需要紧急修复，而您不得不手动为所有环境打补丁。

建立此最佳实践的好处：通过建立补丁管理流程，包括修补标准以及在环境中分发补丁的方法，您可以扩展和报告补丁级别。这为安全补丁提供了保障，并确保清楚地了解已知修复程序的状态。这会促进采用所需特性和功能、快速解决问题并保持监管合规性。实施补丁管理系统和自动化，以减少部署补丁的工作量，并减少手动过程引起的错误。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

修补系统以便纠正问题、获得所需的特性或功能、符合监管政策并满足供应商支持需求。在不可变系统中，使用适当的补丁集进行部署，以便实现所需结果。自动执行补丁管理机制以便缩短修补时间、避免手动过程引起的错误，并减少修补工作量。

## 实施步骤

对于 Amazon EC2 Image Builder：

1. 使用 Amazon EC2 Image Builder，指定管道详细信息：
  - a. 创建映像管道并为其命名
  - b. 定义管道计划和时区
  - c. 配置任何依赖项
2. 选择方案：
  - a. 选择现有方案或创建新方案
  - b. 选择映像类型
  - c. 为您的方案命名并确定其版本
  - d. 选择您的基础映像
  - e. 添加构建组件并添加到目标注册表
3. 可选 - 定义您的基础设施配置。
4. 可选 - 定义配置设置。
5. 查看设置。
6. 定期检查方案，保持方案正常发挥作用。

对于 Systems Manager Patch Manager :

1. 创建补丁基准。
2. 选择路径操作方法。
3. 启用合规性报告和扫描。

资源

相关最佳实践 :

- [OPS06-BP04 自动测试和回滚](#)

相关文档 :

- [What is Amazon EC2 Image Builder?](#)
- [Create an image pipeline using the Amazon EC2 Image Builder](#)
- [Create a container image pipeline](#)
- [AWS Systems Manager Patch Manager](#)
- [Working with Patch Manager](#)
- [使用补丁合规性报告](#)
- [AWS 开发人员工具](#)

相关视频 :

- [AWS 上面向无服务器应用程序的 CI/CD](#)
- [Ops 设计理念](#)

相关示例 :

- [Well-Architected 实验室 - 清单和补丁管理](#)
- [AWS Systems Manager Patch Manager 教程](#)

## OPS05-BP06 共享设计标准

在不同团队间共享最佳实践，以便提高认识并最大程度地实现开发工作的效益。随着架构的发展，记录它们并使它们保持最新。如果在组织中强制实施了共享标准，则必须存在相应的机制来请求对标准进行添加、更改和例外处理。如果没有这样的机制，标准将成为创新的约束。

期望的结果：在组织中的不同团队间共享设计标准。随着最佳实践的发展，记录标准并使它们保持最新。

常见反模式：

- 两个开发团队各自创建了一个用户身份验证服务。对于用户来说，他们想要访问系统的每一部分，都必须使用一套单独的凭据。
- 每个团队管理他们自己的基础设施。新的合规性要求迫使您变更基础设施，各个团队以不同的方式实施变更。

建立此最佳实践的好处：使用共享标准支持最佳实践的采用，并充分发挥开发工作的作用。记录和更新设计标准，让您的组织可以了解最新的最佳实践以及安全和合规性要求。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

在不同团队间共享现有的最佳实践、设计标准、检查清单、操作程序、指南和监管要求。建立针对设计标准的更改、添加和例外请求程序，以便支持改进和创新。让团队了解已发布的内容。随着新最佳实践的出现，使用一种机制以使设计标准保持最新。

## 客户示例

AnyCompany Retail 拥有负责创建软件架构模式的跨职能架构团队。此团队构建具有内置合规性和监管的架构。采用这些共享标准的团队可以从内置合规性和监管中受益。他们可以在设计标准的基础上快速构建。架构团队每季度召开一次会议，评估架构模式，如有必要，更新架构模式。

## 实施步骤

1. 确定一个跨职能团队，负责开发和更新设计标准。此团队应与整个组织的利益相关方合作，制定设计标准、操作程序、检查清单、指南和监管要求。记录设计标准并在组织内共享。
  - a. [AWS Service Catalog](#) 可用于使用基础设施即代码创建代表设计标准的产品组合。您可以在不同账户间共享产品组合。
2. 随着新最佳实践的确定，使用一种机制以使设计标准保持最新。
3. 如果集中执行设计标准，则制定一个流程来请求更改、更新和豁免。

实施计划的工作量级别：中。制定一个流程来创建和共享设计标准，就可以与整个组织的利益攸关方进行协调与合作。

## 资源

### 相关最佳实践：

- [OPS01-BP03 评估治理要求](#) - 监管要求会影响设计标准。
- [OPS01-BP04 评估合规性要求](#) - 在创建设计标准时，合规性是一项至关重要的输入。
- [OPS07-BP02 确保以一致的方式对运维准备情况进行审查](#) - 运营准备检查清单是一种在设计工作负载时实施设计标准的机制。
- [OPS11-BP01 设置持续改进流程](#) - 更新设计标准是持续改进的一部分。
- [OPS11-BP04 执行知识管理](#) - 在知识管理实践过程中，记录和共享设计标准。

### 相关文档：

- [使用 AWS Service Catalog 实现 AWS Backup 自动化](#)
- [AWS Service Catalog Account Factory 增强版](#)
- [Expedia Group 如何使用 AWS Service Catalog 构建数据库即服务 \(DBaaS\) 产品](#)
- [对云架构模式的使用保持可见性](#)
- [简化在 AWS Organizations 设置中共享 AWS Service Catalog 产品组合的操作](#)

### 相关视频：

- [AWS Service Catalog – 入门](#)
- [AWS re:Invent 2020：像专家一样管理您的 AWS Service Catalog 产品组合](#)

### 相关示例：

- [AWS Service Catalog 参考架构](#)
- [AWS Service Catalog 研讨会](#)

### 相关服务：

- [AWS Service Catalog](#)



## OPS05-BP07 实施提高代码质量的实践

实施能够提高代码质量并尽可能减少缺陷的最佳实践。一些示例包括测试驱动型开发、代码审查、标准采用和结对编程。将这些实践合并到您的持续集成和交付流程中。

期望的结果：您的组织使用代码审查或结对编程等最佳实践来提高代码质量。在软件开发生命周期内，开发人员和操作人员采用代码质量最佳实践。

常见反模式：

- 在没有进行代码审查的情况下将代码提交到应用程序的主分支。变更会自动部署到生产环境并导致中断。
- 开发新应用程序，而不进行任何单元测试、端到端测试或集成测试。在部署之前无法测试应用程序。
- 您的团队在生产中进行手动变更以解决缺陷问题。变更没有经过测试或代码审查，也不会通过持续集成和交付流程捕获或记录。

建立此最佳实践的好处：通过采用提高代码质量的实践，能够最大限度地减少引入生产中的问题。使用最佳实践（例如，结对编程和代码审查）提高代码质量。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

实施提高代码质量的实践，以便在部署代码之前尽可能减少缺陷。使用测试驱动型开发、代码审查和结对编程等实践来提高开发的质量。

### 客户示例

AnyCompany Retail 采用几种做法来提高代码质量。他们采用了测试驱动型开发作为编写应用程序的标准。对于一些新功能，他们让开发人员在冲刺阶段结对编程。在集成和部署之前，由高级开发人员对每个拉取请求进行代码审查。

### 实施步骤

1. 在持续集成和交付流程中采用测试驱动型开发、代码审查和结对编程等代码质量实践。使用这些技术来提高软件质量。
  - a. [Amazon CodeGuru Reviewer](#) 可以使用机器学习为 Java 和 Python 代码提供编程建议。
  - b. 您可以使用 [AWS Cloud9](#) 创建共享开发环境，在那里您可以协作开发代码。

实施计划的工作量级别：中。实施此最佳实践有很多方法，但获得组织采用可能并非易事。

## 资源

### 相关最佳实践：

- [OPS05-BP06 共享设计标准](#) - 作为代码质量实践的一部分，您可以共享设计标准。

### 相关文档：

- [敏捷软件指南](#)
- [我的 CI/CD 管道统领我的发布](#)
- [使用 Amazon CodeGuru Reviewer 自动审查代码](#)
- [采用测试驱动型开发方法](#)
- [DevFactory 如何使用 Amazon CodeGuru 构建更好的应用程序](#)
- [关于结对编程](#)
- [RENGA Inc. 使用 Amazon CodeGuru 自动进行代码审查](#)
- [敏捷开发的艺术：测试驱动型开发](#)
- [为什么代码审查很重要（而且实际上节省了时间！）](#)

### 相关视频：

- [AWS re:Invent 2020：使用 Amazon CodeGuru 持续改进代码质量](#)
- [2021 AWS 峰会（澳大利亚和新西兰）- 用 CDK 和测试驱动型开发推动测试优先战略](#)

### 相关服务：

- [Amazon CodeGuru Reviewer](#)
- [Amazon CodeGuru Profiler](#)
- [AWS Cloud9](#)

## OPS05-BP08 使用多个环境

使用多个环境来试验、开发和测试您的工作负载。当环境接近于生产环境时，逐步加强控制，以确保工作负载在部署后能够按预期运行。

期望的结果：您有多个环境，这些环境均反映您的合规性和监管需求。在通往生产的道路上，您可以通过环境来测试和推广代码。

## 常见反模式：

- 您正在共享开发环境中执行开发，另一位开发人员将覆盖您的代码更改。
- 共享开发环境上严苛的安全控制令您无法试验新的服务和功能。
- 您在生产系统上执行负载测试，导致用户停机。
- 生产中发生了严重错误，导致数据丢失。在生产环境中，您尝试重新创建导致数据丢失的条件，以便能够确定它是如何发生的，并防止它再次发生。为了防止在测试期间再次丢失数据，您被迫采取措施，导致用户无法使用应用程序。
- 您正在运行多租户服务，无法支持客户对专用环境的请求。
- 您可能并不总是进行测试，但在需要测试时，您在生产环境中进行。
- 您认为单一环境的简单性比更改在环境中的影响范围更加重要。

建立此最佳实践的好处：您可以为多个同时进行的开发、测试和生产环境提供支持，而不会在开发人员或用户社区间造成冲突。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

使用多个环境，为开发人员提供控制机制最少的沙盒环境，以协助进行试验。提供单独的开发环境以协助并行工作，并提高开发的敏捷性。在接近生产的环境中实施更严格的控制，让开发人员能够创新。使用基础设施即代码和配置管理系统来部署与生产环境中的控制机制配置一致的环境，以便确保系统在部署后按照预期运行。关闭不使用的环境，以免空闲资源（例如晚上和周末的开发系统）产生费用。在负载测试时部署与生产等效的环境，以改善有效结果。

## 资源

### 相关文档：

- [AWS 上的实例计划程序](#)
- [什么是 AWS CloudFormation？](#)

## OPS05-BP09 频繁进行可逆的小规模更改

频繁进行可逆的小规模变更可以减少变更的范围和影响。当与变更管理系统、配置管理系统以及构建和交付系统结合使用时，频繁进行可逆的小规模更改可以减少变更的范围和影响。这样可以提高故障排除工作的效果、加快修复速度，并支持回滚更改。

## 常见反模式：

- 您每季度部署一个新版本的应用程序，这会存在一个变更窗口，意味着核心服务将关闭。
- 您经常更改数据库架构，而不在管理系统中跟踪变更。
- 您执行手动就地更新，覆盖现有安装和配置，并且没有明确的回滚计划。

建立此最佳实践的好处：频繁部署小的更改有助于提高开发速度。更改很小时，更易于确定是否会带来意外后果，并且更容易撤回。更改可逆时，由于简化了恢复，因此实施更改的风险更小。变更过程的风险降低，变更失败的影响也减小。

未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

频繁进行可逆的小规模变更可以减小变更的范围和影响。这可以简化故障排除、加快修复速度，并支持回滚更改。这还可以加快企业实现价值的速度。

## 资源

### 相关最佳实践：

- [OPS05-BP03 使用配置管理系统](#)
- [OPS05-BP04 使用构建和部署管理系统](#)
- [OPS06-BP04 自动测试和回滚](#)

### 相关文档：

- [在 AWS 上实施微服务](#)
- [微服务 - 可观测性](#)

## OPS05-BP10 完全自动化集成和部署

实现自动构建、部署和测试工作负载。这可以减少手动过程引起的错误，并减少部署更改的工作量。

使用 [资源标签](#) 和 [AWS Resource Groups](#)，按照一致的 [标记策略](#) 应用元数据，以协助标识您的资源。标记您的资源，以便进行整理、成本核算、访问控制并有针对性地自动执行操作活动。

期望的结果：开发人员使用工具来交付代码并推广到生产环境。开发人员无需登录 AWS Management Console即可提供更新。对变更和配置进行全面的审计跟踪，可满足监管和合规需求。流程是可重复的，并且跨团队实现标准化。开发人员可以腾出时间专注于开发和代码推送，从而提高工作效率。

常见反模式：

- 星期五，您完成为功能分支编写新代码的工作。星期一，在运行代码质量测试脚本和各单元测试脚本后，您将代码签入计划发行的下一版本中。
- 您接到任务，需要为重要问题编写修复代码，该问题在生产中影响了大量客户。对修复代码进行测试后，您提交代码并通过电子邮件发送变更管理，请求批准，以将其部署到生产环境中。
- 作为开发人员，您可以登录 AWS Management Console，以使用非标准方法和系统创建新的开发环境。

建立此最佳实践的好处：通过自动构建和部署管理系统，可以减少手动流程引起的错误，并减少部署更改的工作量，让您的团队成员能够专注于实现业务价值。可以在推广到生产环境时提高交付速度。

未建立这种最佳实践的情况下暴露的风险等级：低

实施指导

您使用构建和部署管理系统来跟踪并实施更改，以便减少手动流程引起的错误，并减少工作量。将集成和部署管道完全自动化，从代码签入到构建、测试、部署和验证都包含在内。这可以缩短准备时间，鼓励更频繁地进行更改，减少工作量，提高面市速度，提升生产力，并增进代码在推广到生产环境时的安全性。

资源

相关最佳实践：

- [OPS05-BP03 使用配置管理系统](#)
- [OPS05-BP04 使用构建和部署管理系统](#)

相关文档：

- [什么是 AWS CodeBuild？](#)
- [什么是 AWS CodeDeploy？](#)

相关视频：

- [AWS re:Invent 2022 - AWS Well-Architected best practices for DevOps on AWS](#)

## OPS 6.您如何缓解部署风险？

采用可提供快速质量反馈，并在更改没有达到目标成效时支持快速恢复的方法。使用这些实践可以减轻因部署更改而产生的问题的影响。

### 最佳实践

- [OPS06-BP01 针对不成功的更改制定计划](#)
- [OPS06-BP02 测试部署](#)
- [OPS06-BP03 采用安全部署策略](#)
- [OPS06-BP04 自动测试和回滚](#)

### OPS06-BP01 针对不成功的更改制定计划

制定计划，以便在部署没有达到期望结果时，在生产环境中恢复到已知良好状态，或者进行修复。制定一项策略来建立这样的计划，有助于所有团队制定从失败的更改中恢复的策略。这样的策略示例包括部署和回滚步骤、更改策略、功能标记、流量隔离和流量转移。单个发布可能包括多个相关的组件更改。该策略应提供承受任何组件更改的失败或从中恢复过来的能力。

期望的结果：您已经为更改失败准备了详细的恢复计划。此外，您还缩小了发布内容的大小，以最大限度地减少对其他工作负载组件的潜在影响。因此，您通过缩短更改失败可能造成的停机时间，提高恢复时间的灵活性和效率，减少了对业务的影响。

### 常见反模式：

- 执行部署后，应用程序变得不稳定，但是系统上似乎还有活动用户。您必须决定是回滚更改并影响活动用户，还是等到知道用户无论如何都可能受到影响后再回滚更改。
- 执行例行更改后，可以访问新环境，但是其中一个子网无法访问。您必须决定是回滚所有内容还是尝试修复无法访问的子网。在您做决定时，子网仍然无法访问。
- 您的系统的架构不允许使用较小的发布版本进行更新。因此，在部署失败时，您很难撤销这些大批量的更改。
- 您没有使用基础设施即代码（IaC）模式，而且对基础设施进行的手动更新导致了不希望出现的配置。您无法有效地跟踪和撤销手动更改。
- 由于您没有将部署频率的增加作为衡量标准，因此团队没有动力来缩小更改规模，也不愿意改进每次更改的回滚计划，从而导致风险增加和失败率上升。

- 您没有衡量因更改失败而导致的中断的总持续时间。您的团队无法确定部署流程的优先顺序和恢复计划的有效性，也无法进行改进。

建立此最佳实践的好处：制定从失败更改中恢复的计划可以最大限度地缩短平均恢复时间（MTTR，Mean Time To Recover），并减少对业务的影响。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

发布团队采用的一致、有据可查的策略和实践，使组织能够计划在更改失败时应如何处理。该策略应允许在特定情况下向前修复。无论是哪种情况，在部署到实际生产环境之前，都应妥善记录并测试向前修复或回滚计划，以便最大限度地减少从更改中恢复所需的时间。

## 实施步骤

1. 记录要求团队制定有效计划以在指定时间内撤销更改的策略。
  - a. 策略应指定何时允许出现向前修复情况。
  - b. 要求所有相关人员都能查阅记录在案的回滚计划。
  - c. 指定回滚要求（例如，当发现部署了未经授权的更改时）。
2. 分析与工作负载的每个组件相关的所有更改的影响级别。
  - a. 如果可重复的更改遵循的工作流，与执行更改策略的工作流保持一致，则允许对这些更改进行标准化、模板化和预授权。
  - b. 通过缩小更改的规模来减少任何更改的潜在影响，从而减少恢复所需的时间和对业务的影响。
  - c. 确保回滚过程将代码恢复到已知的良好状态，以尽可能避免意外事件。
3. 集成工具和工作流，以编程方式执行策略。
4. 让其他工作负载所有者能够查看有关更改的数据，以提高对无法回滚的任何失败更改的诊断速度。
  - a. 利用可见的更改数据来衡量这一做法是否成功，并确定迭代改进措施。
5. 使用监控工具来验证部署的成败，以加快制定回滚决策的速度。
6. 衡量更改失败时的停机时间，以不断改进恢复计划。

实施计划的工作量级别：中

## 资源

相关最佳实践：



- [OPS06-BP04 自动测试和回滚](#)

相关文档：

- [AWS Builders Library | Ensuring Rollback Safety During Deployments](#)
- [AWS 白皮书 | Change Management in the Cloud](#)

相关视频：

- [re:Invent 2019 | Amazon's approach to high-availability deployment](#)

## OPS06-BP02 测试部署

使用与生产环境相同的部署配置、安全控制、步骤和程序，在预生产环境中测试发布过程。验证所有部署步骤是否按预期完成，如检查文件、配置和服务。通过功能测试、集成测试和负载测试以及运行状况检查等各种监控方法，进一步测试所有更改。通过这些测试，您可以及早发现部署问题，并有机会在进入生产之前规划和缓解问题。

您可以创建临时的并行环境来测试每项更改。使用基础设施即代码 ( IaC ) 自动部署测试环境，有助于减少所涉及的工作量，确保稳定性、一致性和更快的功能交付。

期望的结果：您的组织采用了包含测试部署在内的测试驱动型开发文化。这样可以确保团队专注于提供商业价值，而不是管理发布版本。各团队在发现部署风险后尽早参与进来，以确定适当的缓解方案。

常见反模式：

- 在发布生产版本期间，未经测试的部署会导致问题频发，需要进行故障排除和上报。
- 您的发布版本包含用于更新现有资源的基础设施即代码 ( IaC )。您不确定 IaC 是否会成功运行，还是会对资源造成影响。
- 您在应用程序中部署一项新功能。此功能未按预期运行，并且在受影响的用户报告之前都无从了解问题。
- 您更新了证书。您不小心将证书安装到了错误的组件上，而这却没有被发现，于是因为无法建立与网站的安全连接而影响网站访客。

建立此最佳实践的好处：在生产前对部署程序及其引入的更改进行全面测试，可最大限度地减少部署步骤对生产的潜在影响。这增强了生产版本发布过程中的信心，并最大限度地减少了运营支持，而且不会减慢更改交付速度。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

测试部署过程与测试部署所产生的更改同样重要。要完成这一步骤，可以在尽可能接近生产环境的预生产环境中测试部署步骤。可以在投入生产之前发现一些常见问题，如部署步骤不完整、不正确或配置错误。此外，您还可以测试恢复步骤。

## 客户示例

作为持续集成和持续交付（CI/CD，Continuous Integration/Continuous Delivery）管道的一部分，AnyCompany Retail 在类似生产的环境中执行为客户发布基础设施和软件更新所需的既定步骤。该管道包含预检查过程，用于在部署之前检测资源中的偏差（检测在 IaC 之外对资源执行的更改），以及验证 IaC 在启动时采取的操作。该管道会验证部署步骤，例如在向负载均衡器重新注册之前，验证特定文件和配置是否已准备就绪，服务是否处于正在运行状态，以及是否正确响应本地主机上的运行状况检查。此外，所有更改都要进行一系列自动测试，如功能测试、安全测试、回归测试、集成测试和负载测试。

## 实施步骤

1. 执行预安装检查，模拟生产环境打造预生产环境。
  - a. 使用 [偏差检测](#) 功能，检测是否在 AWS CloudFormation 之外更改了资源。
  - b. 使用 [更改集](#) 功能，验证堆栈更新的意图是否与 AWS CloudFormation 在启动更改集时所采取的操作相匹配。
2. 这会在 [AWS CodePipeline](#) 中触发手动审批步骤，以授权部署到预生产环境。
3. 使用 [AWS CodeDeploy AppSpec](#) 文件等部署配置来定义部署和验证步骤。
4. 在适用的情况下，[可将 AWS CodeDeploy 与其他 AWS 服务集成](#) 或 [将 AWS CodeDeploy 与合作伙伴的产品和服务集成](#)。
5. [监控部署](#) - 使用 Amazon CloudWatch、AWS CloudTrail 和 Amazon SNS 事件通知。
6. 执行部署后的自动化测试，包括功能测试、安全测试、回归测试、集成测试和负载测试。
7. [排查](#) 部署问题。
8. 成功验证上述步骤后应启动手动审批工作流，以授权部署到生产环境。

实施计划的工作量级别：高

## 资源

相关最佳实践：

- [OPS05-BP02 测试并验证变更](#)

相关文档：

- [AWS Builders' Library | Automating safe, hands-off deployments | Test Deployments](#)
- [AWS 白皮书 | Practicing Continuous Integration and Continuous Delivery on AWS](#)
- [The Story of Apollo - Amazon's Deployment Engine](#)
- [How to test and debug AWS CodeDeploy locally before you ship your code](#)
- [Integrating Network Connectivity Testing with Infrastructure Deployment](#)

相关视频：

- [re:Invent 2020 | Testing software and systems at Amazon](#)

相关示例：

- [Tutorial | Deploy and Amazon ECS service with a validation test](#)

## OPS06-BP03 采用安全部署策略

在安全的生产环境滚动部署中，会对有益更改的流程进行控制，目标是尽可能减少这些更改让客户感知到的任何影响。安全控制措施提供检查机制，用于验证是否达成所期望的结果，并针对由于更改或部署失败所引入的任何缺陷，限制这些缺陷的影响范围。安全滚动部署可包括功能标记、单盒、滚动（金丝雀版本）、不可变、流量分割和蓝绿部署等策略。

期望的结果：您的企业使用持续集成/持续交付（CI/CD，Continuous Integration/Continuous Delivery）系统，提供自动进行安全滚动部署的功能。团队必须使用适当的安全滚动部署策略。

常见反模式：

- 您将不成功的更改一次性部署到所有生产环境中。因此，所有客户同时受到影响。
- 在同时部署到所有系统时，引入的一个缺陷需要紧急进行修复。为所有客户修复该缺陷需要几天时间。
- 管理生产版本发布需要多个团队的规划和参与。这限制了您为客户更新功能的频率。
- 您通过修改现有系统来执行可变部署。发现更改不成功时，您被迫再次修改系统，还原旧版本，导致恢复时间延长。

建立此最佳实践的好处：自动化的部署，在快速滚动部署与持续向客户提供有益更改之间取得平衡。限制影响范围可以防止代价高昂的部署失败，并最大限度地提高团队有效应对失败的能力。

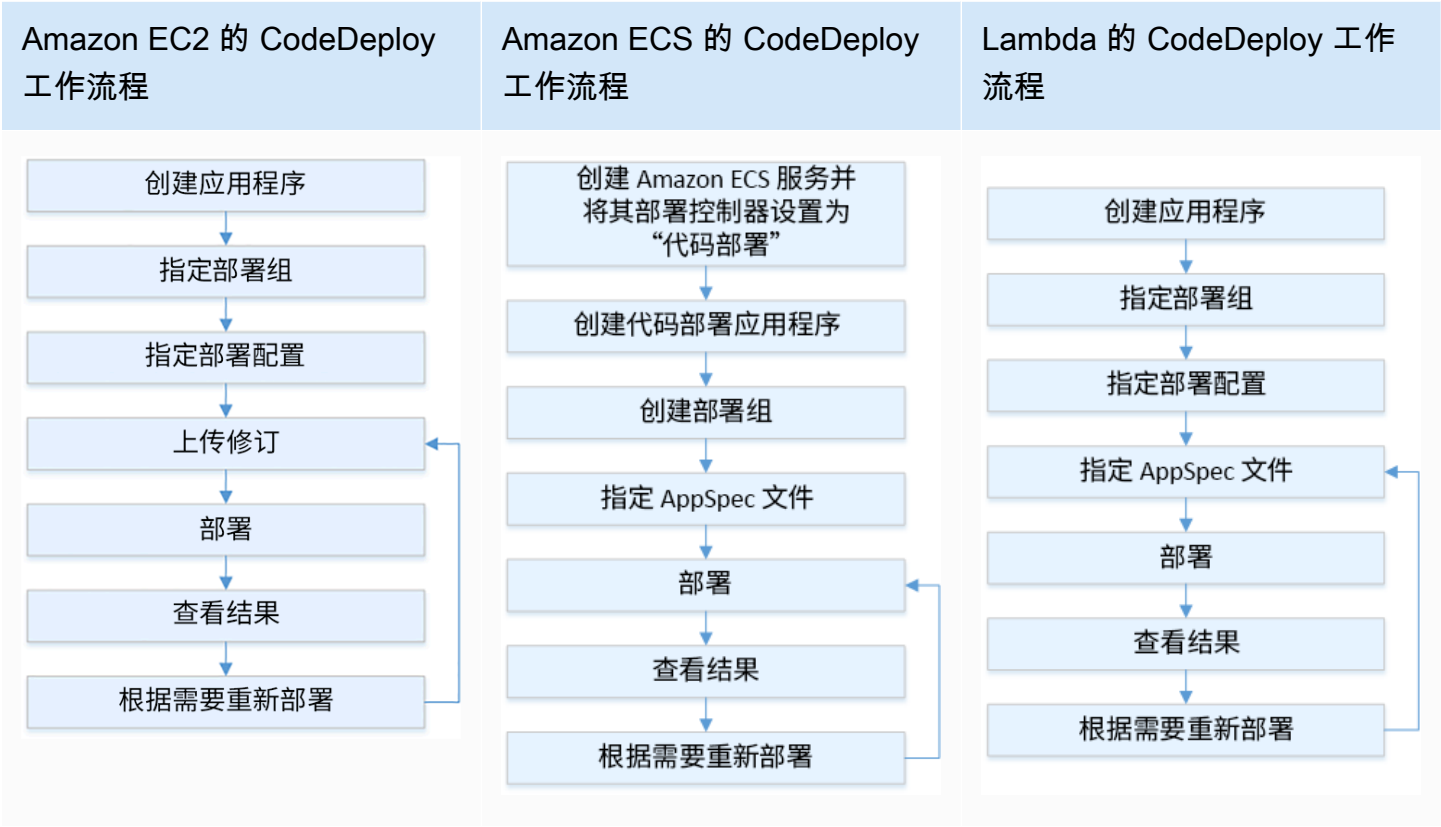
未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

持续交付失败会导致服务可用性降低，带来糟糕的客户体验。为了最大限度地提高部署成功率，请在端到端发布流程中实施安全控制措施，以最大限度地减少部署错误，将达成零部署失败作为目标。

客户示例

AnyCompany Retail 的目标是尽可能减少部署的停机时间，甚至实现零停机，这意味着在部署期间，用户不会感觉到任何影响。为了实现这一目标，公司建立了部署模式（参见以下工作流程图），例如滚动部署和蓝绿部署。所有团队在各自的 CI/CD 管道中都采用了其中一种或多种模式。



实施步骤

1. 使用审批工作流程在提升到生产版本后，启动生产版本滚动部署步骤序列。
2. 使用自动化部署系统，例如 [AWS CodeDeploy](#)。AWS CodeDeploy [部署选项](#) 包括 EC2/本地就地部署及 EC2/本地蓝绿部署、AWS Lambda 和 Amazon ECS（参见前面的工作流程图）。

- a. 在适用的情况下，[可将 AWS CodeDeploy 与其他 AWS 服务集成](#) 或 [将 AWS CodeDeploy 与合作伙伴的产品和服务集成](#)。
3. 对于 [Amazon Aurora](#) 和 [Amazon RDS](#) 等数据库，使用蓝绿部署。
4. [监控部署](#) - 使用 Amazon CloudWatch、AWS CloudTrail 和 Amazon SNS 事件通知。
5. 执行部署后的自动化测试，包括功能测试、安全测试、回归测试、集成测试以及任何负载测试。
6. [排查](#) 部署问题。

实施计划的工作量级别：中

资源

相关最佳实践：

- [OPS05-BP02 测试并验证变更](#)
- [OPS05-BP09 频繁进行可逆的小规模更改](#)
- [OPS05-BP10 完全自动化集成和部署](#)

相关文档：

- [AWS Builders Library | Automating safe, hands-off deployments | Production deployments](#)
- [AWS Builders Library | My CI/CD pipeline is my release captain | Safe, automatic production releases](#)
- [AWS 白皮书 | Practicing Continuous Integration and Continuous Delivery on AWS | Deployment methods](#)
- [AWS CodeDeploy User Guide](#)
- [Working with deployment configurations in AWS CodeDeploy](#)
- [Set up an API Gateway canary release deployment](#)
- [Amazon ECS Deployment Types](#)
- [Fully Managed Blue/Green Deployments in Amazon Aurora and Amazon RDS](#)
- [Blue/Green deployments with AWS Elastic Beanstalk](#)

相关视频：

- [re:Invent 2020 | Hands-off: Automating continuous delivery pipelines at Amazon](#)

- [re:Invent 2019 | Amazon's Approach to high-availability deployment](#)

相关示例：

- [Try a Sample Blue/Green Deployment in AWS CodeDeploy](#)
- [研讨会 | Building CI/CD pipelines for Lambda canary deployments using AWS CDK](#)
- [研讨会 | Blue/Green and Canary Deployment for EKS and ECS](#)
- [研讨会 | Building a Cross-account CI/CD Pipeline](#)

## OPS06-BP04 自动测试和回滚

为了提高部署过程的速度和可靠性以及对该过程的信心，您需要制定一项策略，用于在预生产和生产环境中实现自动化的测试和回滚功能。在部署到生产环境时自动进行测试，模拟人与系统的交互，从而验证已经部署了更改。利用自动回滚功能，可以快速恢复到先前已知的良好状态。回滚应在预先定义的条件自动启动，例如更改未达到期望结果或自动化测试失败时。自动执行这两项活动可以提高部署的成功率，尽可能缩短恢复时间，并减少可能对业务造成的影响。

期望的结果：您的自动化测试和回滚策略已集成到持续集成/持续交付（CI/CD，Continuous Integration/Continuous Delivery）管道中。您的监控功能可以根据成功标准进行验证，并能在失败时启动自动回滚。这样可以最大限度地减少对最终用户和客户的任何影响。例如，当您对所有测试结果都感到满意时，可以将代码提升到生产环境中，该环境启动了使用相同测试案例的自动回归测试。如果回归测试结果与预期不符，则在管道工作流程中启动自动回滚。

常见反模式：

- 您的系统的架构不允许使用较小的发布版本进行更新。因此，在部署失败时，您很难撤销这些大批量的更改。
- 您的部署过程包括一系列人工步骤。将更改部署到工作负载后，您启动了部署后测试。完成测试之后，您发现工作负载不可操作，而且客户断开了连接。然后，您开始回滚到之前的版本。所有这些人工步骤都会延误整个系统的恢复，并对客户造成长时间的影响。
- 您花时间为应用程序中不常用的功能开发了自动化测试案例，这极大地降低了自动化测试功能上的投资回报率。
- 您的发布版本由应用程序、基础设施、补丁和配置更新组成，这些组件相互独立。但是，您只有一个CI/CD管道，只能同时交付所有更改。一个组件中的失败会迫使您撤销所有更改，导致回滚过程复杂且效率低下。

- 您的团队完成了冲刺一的编码工作并开始冲刺二的工作，但是按照您的计划，直到冲刺三才会进行测试。结果，自动化测试发现冲刺一中存在缺陷，需要先解决这些缺陷，然后才能开始测试冲刺二的可交付成果，整个发布都被延误，这大大降低了自动化测试的价值。
- 生产版本的自动化回归测试案例已完成，但您没有监控工作负载运行状况。由于无法监控服务是否已重新启动，因此您不确定是否需要回滚或者是否已经进行了回滚。

建立此最佳实践的好处：自动化测试可提高测试过程的透明度，以及在更短的时间内测试更多功能的能力。通过对生产环境中的更改进行测试和验证，可以让您立即发现问题。利用自动化测试工具改进一致性，可以更好地检测缺陷。通过自动回滚到以前的版本，可以将对客户的影响降至最低。自动回滚可以减少业务影响，最终提升对您部署功能的信心。总体而言，这些功能可在确保质量的同时缩短交付时间。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

自动测试部署的环境，以便更快地确认目标结果。在没有达到预定义的结果时，自动回滚到之前的已知良好状态，尽可能地缩短恢复时间，并减少手动过程引起的错误。将测试工具与管道工作流集成，以便一致地开展测试并尽可能减少手动输入。确定优先执行的自动化测试案例，例如能够降低最大风险的测试案例，以及每次更改都需要频繁测试的测试案例。此外，还可以根据在测试计划中预定义的特定条件自动回滚。

## 实施步骤

1. 为开发生命周期建立测试生命周期，针对测试过程，从需求规划到测试案例开发、工具配置、自动化测试和测试案例关闭，对每个阶段进行定义。
  - a. 根据您的整体测试策略创建特定于工作负载的测试方法。
  - b. 考虑在整个开发生命周期中适宜的阶段实施持续测试策略。
2. 根据您的业务需求和管道投资，选择用于测试和回滚的自动化工具。
3. 确定要自动执行哪些测试案例，以及应手动执行哪些测试案例。这个过程可以根据所测试功能的业务价值优先级来定义。确保所有团队成员都遵守该计划，并核实执行手动测试的责任人。
  - a. 在自动化测试可以实现价值的特定测试案例上使用自动化测试功能，例如可重复或经常运行的案例、需要重复任务的案例或者多种配置中所需的案例。
  - b. 在自动化工具中定义测试自动化脚本和成功标准，以便在特定案例失败时可以启动持续的自动化工作流。
  - c. 为自动回滚定义具体的失败标准。



4. 优先考虑测试自动化，在过于复杂和人工交互会导致更高失败风险的案例中，通过开发全面的测试案例来获得一致的结果。
5. 将您的自动化测试和回滚工具集成到 CI/CD 管道中。
  - a. 为您的更改制定明确的成功标准。
  - b. 监控并观察以检测这些标准，以及在满足特定回滚标准时自动撤销更改。
6. 执行不同类型的自动化生产测试，例如：
  - a. A/B 测试，显示在两个用户测试组之间当前版本的结果对比。
  - b. 金丝雀测试，让您可以先对一部分用户部署更改，然后再向所有用户发布。
  - c. 功能标记测试，让您可以在应用程序外部，每次将新版本的单个功能标记为打开和关闭，以便逐个验证各个新功能。
  - d. 回归测试，用于验证现有相互关联组件的新功能。
7. 监控应用程序的操作情况、事务，以及与其他应用程序和组件的交互。编制报告，用于按工作负载显示更改是否成功，这样您就可以确定对自动化和工作流的哪些部分进一步进行优化。
  - a. 编制测试结果报告，以便您快速决定是否应调用回滚程序。
  - b. 实施策略，以便根据一种或多种测试方法的预定义失败条件进行自动回滚。
8. 开发自动化测试案例，以便在将来的可重复更改中重用。

实施计划的工作量级别：中

资源

相关最佳实践：

- [OPS06-BP01 针对不成功的更改制定计划](#)
- [OPS06-BP02 测试部署](#)

相关文档：

- [AWS Builders Library | Ensuring rollback safety during deployments](#)
- [Redeploy and rollback a deployment with AWS CodeDeploy](#)
- [8 best practices when automating your deployments with AWS CloudFormation](#)

相关示例：

- [使用 Selenium、AWS Lambda、AWS Fargate 和 AWS 开发人员工具进行无服务器 UI 测试](#)

相关视频：

- [re:Invent 2020 | Hands-off: Automating continuous delivery pipelines at Amazon](#)
- [re:Invent 2019 | Amazon's Approach to high-availability deployment](#)

## 运营 7. 如何知道您已经准备好支持某种工作负载？

评估工作负载、流程及程序和工作人员的操作准备情况，以便了解与工作负载相关的操作风险。

最佳实践

- [OPS07-BP01 确保员工能力](#)
- [OPS07-BP02 确保以一致的方式对运维准备情况进行审查](#)
- [OPS07-BP03 使用运行手册执行程序](#)
- [OPS07-BP04 根据行动手册调查问题](#)
- [OPS07-BP05 做出明智的决策来部署系统和变更](#)
- [OPS07-BP06 为生产工作负载启用支持计划](#)

### OPS07-BP01 确保员工能力

通过一种机制来验证您是否有适当数量训练有素的员工来支持工作负载。他们必须接受构成工作负载的平台和服务方面的培训。为他们提供运营工作负载所需的知识。您必须有足够训练有素的员工来支持工作负载的正常运营和排查发生的意外事件。拥有足够的员工轮流值班和休假，避免疲劳。

期望结果：

- 在工作负载可用时，有足够训练有素的员工为工作负载提供支持。
- 为员工提供构成工作负载的软件和服务方面的培训。

常见反模式：

- 部署工作负载，但没有经过培训团队成员来运营所使用的平台和服务。
- 没有足够的员工来支持轮流值班或休假。

建立此最佳实践的好处：

- 拥有技能娴熟的团队成员能够为您的工作负载提供有效支持。
- 有足够的团队成员，可以支持工作负载和实现轮流值班，同时降低疲劳风险。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

确认有足够的训练有素的员工来支持工作负载。确认有足够的团队成员来执行正常运营活动，包括轮流值班。

### 客户示例

AnyCompany Retail 确保支持工作负载的团队得到适当的人员配备和培训。他们有足够的工程师支持轮流值班。他们为员工提供有关构建工作负载的软件和平台方面的培训，并鼓励他们获得认证。他们有足够的员工，所以员工可以休假，同时仍然可以支持工作负载和轮流值班。

### 实施步骤

1. 分配足够数量的员工来运营和支持您的工作负载，并且支持轮流值班。
2. 在构成工作负载的软件和平台方面对员工进行培训。
  - a. [AWS 培训与认证](#) 包括有关 AWS 的课程库。这里提供线上和线下的免费和付费课程。
  - b. [AWS 主持活动和网络研讨会](#)，以便向 AWS 专家学习。
3. 随着运营条件和工作负载发生变化，定期评估团队规模和技能。调整团队规模和技能以满足运营要求。

实施计划的工作量级别：高。雇用和培训团队以支持工作负载需要付出巨大的努力，但可带来可观的长期利益。

### 资源

#### 相关最佳实践：

- [OPS11-BP04 执行知识管理](#) - 团队成员必须具备运营和支持工作负载所需的信息。知识管理是提供这些信息的关键。

#### 相关文档：

- [AWS 活动和网络研讨会](#)

- [AWS 培训和认证](#)

## OPS07-BP02 确保以一致的方式对运维准备情况进行审查

使用运维准备情况审查 ( ORR , Operational Readiness Review ) , 确保可以运营您的工作负载。ORR 是 Amazon 开发的一种机制, 用于验证团队可以安全地运营其工作负载。ORR 是一个使用要求核对清单进行审查和检查的过程。ORR 是一种自助服务体验, 供团队用于验证其工作负载。ORR 中包含的最佳实践源自我们多年构建软件的经验教训。

ORR 核对清单包括架构推荐、运维过程、事件管理和发布质量。我们的更正错误 ( CoE , Correction of Error ) 流程是这些项目的主要推动因素。您的事后分析应该可以推动自己的 ORR 演进。ORR 不仅仅关系到遵循最佳实践, 还关系到预防以前的事件再次发生。最后, ORR 中还可以包括安全性、监管和合规性要求。

在工作负载正式公开发布之前运行 ORR , 然后在整个软件开发生命周期中运行 ORR。在发布之前运行 ORR 可以提升安全运营工作负载的能力。对工作负载定期重新运行 ORR 可以收集任何偏离最佳实践的情况。您可以准备用于新服务发布的 ORR 以及用于定期审查的 ORR。这可以帮助您遵循最新制定的最佳实践, 并吸取从事事后分析中学到的经验教训。随着您对云的使用日趋成熟, 您可以将 ORR 要求作为默认设置整合到自己的架构中。

期望的结果: 您已准备好 ORR 核对清单, 其中包括适合您组织的最佳实践。在工作负载发布之前运行 ORR。在整个工作负载生命周期中定期运行 ORR。

常见反模式:

- 您启动了工作负载, 但不知道谁负责其运维工作。
- 在验证工作负载以便发布时, 没有包括监管和安全性要求。
- 没有定期重新评估工作负载。
- 发布工作负载而没有准备好所需的规程。
- 您在多个工作负载中看到相同的根本原因反复导致出现故障。

建立此最佳实践的好处:

- 您的工作负载包括架构、流程和管理最佳实践。
- 学到的经验教训可合并到 ORR 流程中。
- 在工作负载发布时已准备好所需的规程。
- 在工作负载的整个软件生命周期中运行 ORR。

未建立这种最佳实践的情况下的风险等级：高

## 实施指导

ORR 关系到两点：流程和核对清单。ORR 流程应该由您的组织采用并获得高管支持。至少，ORR 必须在工作负载正式公开发布之前已运行。在整个软件开发生命周期中运行 ORR 可确保软件始终遵循新的最佳实践或新要求。ORR 核对清单应包括配置项目、安全性和监管要求，以及组织的最佳实践。在一段时间后，您可以使用 [AWS Config](#)、[AWS Security Hub](#) 和 [AWS Control Tower 防护机制](#) 等服务，将源自 ORR 的最佳实践整合到防护机制中，以实现自动化的最佳实践检测。

## 客户示例

在经历了多起生产事件之后，AnyCompany Retail 决定实施 ORR 流程。他们构建了核对清单，其中包括最佳实践、监管和合规性要求，以及从中断中学到的经验教训。在发布新工作负载之前，运行 ORR。每个工作负载会每年运行一次 ORR，其中包括一小组最佳实践，用于整合添加到 ORR 核对清单中的新最佳实践和要求。在一段时间后，AnyCompany Retail 使用 [AWS Config](#) 来检测一些最佳实践，以加快 ORR 流程。

## 实施步骤

如需详细了解 ORR，请阅读 [运维准备情况审查 \(ORR\) 白皮书](#)。其中详细介绍了 ORR 流程的历史，如何构建自己的 ORR 实践，以及如何制定自己的 ORR 核对清单。以下步骤是该文档的缩减版本。如需深入了解什么是 ORR 以及如何自行构建，建议您阅读该白皮书。

1. 让关键利益相关方聚在一起讨论，包括来自安全、运维和开发部门的代表。
2. 让每个利益相关方至少提一个要求。对于第一次迭代，请尝试将项目数限制为不超过三十个。
  - [附录 B：ORR 问题示例](#) 源自运维准备情况审查 (ORR) 白皮书，包含您在开始着手时可借鉴的示例问题。
3. 在电子表格中收集您的要求。
  - 您可以使用 [自定义剖析](#)（位于 [AWS Well-Architected Tool](#) 中）开发自己的 ORR，并跨账户以及在 AWS Organization 中分享它们。
4. 确定一个工作负载来运行 ORR。最好选择发布前的工作负载或者内部工作负载。
5. 运行 ORR 核对清单并记录任何发现结果。如果已经有防范措施，那么发现结果可能就不太重要。对于任何没有防范措施的发现结果，请将它们记录到项目的待办事项中，并在发布之前实施它们。
6. 在一段时间后，继续在 ORR 中添加最佳实践和要求。

具有 Enterprise Support 的支持 客户可以向其技术客户经理请求举行 [运维准备情况审查研讨会](#)。该研讨会是一个交互式研讨会，采用 反推式工作方法，可帮助您制定自己的 ORR 核对清单。

实施计划的工作量级别：高。在组织中采用 ORR 实践需要获得高管以及利益相关方的支持。使用整个组织中获得的反馈意见来构建和更新核对清单。

## 资源

### 相关最佳实践：

- [OPS01-BP03 评估治理要求](#) – 监管要求非常适合包括在 ORR 核对清单中。
- [OPS01-BP04 评估合规性要求](#) – 合规性要求有时候包括在 ORR 核对清单中。另一些时候它们可作为单独的流程。
- [OPS03-BP07 为团队配置适当的资源](#) – 团队能力是很适合加入 ORR 要求的候选项。
- [OPS06-BP01 针对不成功的更改制定计划](#) – 在发布工作负载之前，必须建立回滚或前滚计划。
- [OPS07-BP01 确保员工能力](#) – 为了支持工作负载，您必须具备所需的人员。
- [SEC01-BP03 识别并验证控制目标](#) – 安全控制目标会是非常合适的 ORR 要求。
- [REL13-BP01 定义停机和数据丢失的恢复目标](#) – 灾难恢复计划是很好的 ORR 要求。
- [COST02-BP01 根据组织的要求制定各种策略](#) – 成本管理策略非常适合包括在 ORR 核对清单中。

### 相关文档：

- [AWS Control Tower – AWS Control Tower 中的防护机制](#)
- [AWS Well-Architected Tool – 自定义剖析](#)
- [Adrian Hornsby 提供的运维准备情况审查模板](#)
- [运维准备情况审查 \( ORR \) 白皮书](#)

### 相关视频：

- [AWS 支持 为您提供支持 | 构建高效的运维准备情况审查 \( ORR , Operational Readiness Review \)](#)

### 相关示例：

- [运维准备情况审查 \( ORR \) 剖析](#)

### 相关服务：

- [AWS Config](#)

- [AWS Control Tower](#)
- [AWS Security Hub](#)
- [AWS Well-Architected Tool](#)

## OPS07-BP03 使用运行手册执行程序

A 运行手册 是实现特定结果的书面流程。运行手册由某人为完成某件事而遵循的一系列步骤组成。早在航空发展的早期，运行手册便已用于运营。在云运营中，我们使用运行手册来降低风险并实现预期结果。简单而言，运行手册就是完成一项任务的核对清单。

运行手册是运营工作负载的重要组成部分。从新团队成员入职到部署一个主要版本，运行手册都是一个成文的流程，无论谁使用它们，都能获得一致的结果。运行手册应发布在一个中央位置，并随着流程的发展而更新，因为更新运行手册是变更管理流程的一个关键组成部分。它们还应包括关于错误处理、工具、权限、异常和问题发生时上报的指导。

随着贵组织日益成熟，开始自动编写运行手册。从简短且经常使用的运行手册开始。使用脚本语言来实现步骤自动化或使步骤更容易执行。当您自动化前几本运行手册后，您将花时间自动化更复杂的运行手册。随着时间的推移，大多数运行手册应以某种方式实现自动化。

期望结果：您的团队有一系列执行工作负载任务的分步指南。运行手册包含期望结果、必要的工具和权限，以及关于错误处理的说明。它们存储在一个中央位置并经常更新。

常见反模式：

- 依靠记忆完成流程的每个步骤。
- 手动部署更改而不使用核对清单。
- 不同的团队成员执行相同的流程，但执行不同的步骤或取得不同的结果。
- 让运行手册与系统更改和自动化不同步。

建立此最佳实践的好处：

- 降低人工任务的错误率。
- 以一致的方式执行操作。
- 新的团队成员可以更早地开始执行任务。
- 可以自动编写运行手册以减少工作量。

未建立这种最佳实践的情况下暴露的风险等级：中



## 实施指导

根据贵组织的成熟度级别，运行手册可以采用多种形式。它们至少应该包含一个分步文本文档。应明确指出期望结果。清楚地记录必要的特殊权限或工具。提供关于错误处理和出现问题时进行上报的详细指导。列出运行手册负责人，并将运行手册发布在一个中央位置。一旦运行手册编写完成，让您团队中的其他人运行它来进行验证。随着过程的发展，根据变更管理流程更新运行手册。

随着贵组织日益成熟，您的文本运行手册应实现自动化。使用诸如 [AWS Systems Manager 自动化之类的服务](#)，您可以将纯文本转换为可针对您的工作负载运行的自动化功能。这些自动化功能可以根据事件的发生而运行，从而减轻维持工作负载的运营负担。

## 客户示例

AnyCompany Retail 必须在软件部署期间执行数据库模式更新。云运营团队与数据库管理团队合作，构建了一个用于手动部署这些更改的运行手册。运行手册以核对清单的形式列出了流程中的每个步骤。其中有一节是关于出错时的错误处理。他们在内部 Wiki 上发布了该运行手册和其他运行手册。云运营团队计划在未来的冲刺阶段实现运行手册的自动化。

## 实施步骤

如果您没有现有的文档存储库，那么版本控制存储库是开始构建运行手册库的绝佳场所。您可以使用 Markdown 构建运行手册。我们提供了一个示例运行手册模板，您可以用它开始构建运行手册。

```
# Runbook Title ## Runbook Info | Runbook ID | Description | Tools Used
| Special Permissions | Runbook Author | Last Updated | Escalation POC |
|-----|-----|-----|-----|-----|-----|-----| | RUN001 | What is this
runbook for? What is the desired outcome? | Tools | Permissions | Your Name |
2022-09-21 | Escalation Name | ## Steps 1.Step one 2.Step two
```

1. 如果您当前尚没有文档存储库或 Wiki，请在版本控制系统中创建一个新的版本控制存储库。
2. 识别一个没有运行手册的流程。一个理想的流程是半定期执行的流程，步骤少，且故障影响小。
3. 在文档存储库中，使用模板创建新的草稿 Markdown 文档。填写 Runbook Title 以及 Runbook Info #####。
4. 从第一步开始，填写运行手册的 Steps 部分。
5. 将运行手册交给团队成员。让他们使用运行手册来验证这些步骤。如果有遗漏或需要澄清的地方，请更新运行手册。
6. 将运行手册发布到您的内部文档存储区。发布后，告诉您的团队和其他利益相关者。
7. 随着时间的推移，您将构建一个运行手册库。随着该库的增长，开始努力实现运行手册的自动化。

实施计划的工作量级别：低。运行手册的最低标准是一个分步文本指南。实现运行手册自动化可能会增加实施工作量。

## 资源

相关最佳实践：

- [OPS02-BP02 确定流程和程序所有者](#)：运行手册应该有一个负责人负责维护。
- [OPS07-BP04 根据行动手册调查问题](#)：运行手册和行动手册彼此相似，但有一个关键区别：运行手册包含期望结果。在许多情况下，一旦行动手册确定了根本原因，就会触发运行手册。
- [OPS10-BP01 使用流程来管理事件、意外事件和问题](#)：运行手册是良好的事件、意外事件和问题管理实践的一部分。
- [OPS10-BP02 针对每个提醒设置一个流程](#)：应使用运行手册和行动手册来响应警报。随着时间的推移，应自动进行这些响应。
- [OPS11-BP04 执行知识管理](#)：维护运行手册是知识管理的一个关键部分。

相关文档：

- [利用自动化行动手册和运行手册实现卓越运营](#)
- [AWS Systems Manager：使用运行手册](#)
- [适用于 AWS 大型迁移的迁移行动手册 - 任务 4：改进迁移运行手册](#)
- [使用 AWS Systems Manager Automation 运行手册解决运营任务](#)

相关视频：

- [AWS re:Invent 2019：运行手册、事件报告和事件响应 DIY 指南 \( SEC318-R1 \)](#)
- [如何在 AWS | Amazon Web Services 上实现 IT 运营自动化](#)
- [将脚本集成到 AWS Systems Manager](#)

相关示例：

- [AWS Systems Manager：自动化演练](#)
- [AWS Systems Manager：从最新的快照运行手册中还原根卷](#)
- [使用 Jupyter Notebook 和 CloudTrail Lake 构建 AWS 意外事件响应运行手册](#)
- [Gitlab - 运行手册](#)

- [Rubix - 用于在 Jupyter Notebook 中构建运行手册的 Python 库](#)
- [使用 Document Builder 创建自定义运行手册](#)
- [Well-Architected 实验室：使用行动手册和运行手册自动完成操作](#)

相关服务：

- [AWS Systems Manager Automation](#)

## OPS07-BP04 根据行动手册调查问题

行动手册是用于调查事件的分步指南。发生事件时，行动手册用于开展调查，以及确定影响的范围和根本原因。行动手册可用于从失败的部署到安全事件的各种场景。在许多情况下，行动手册可确定根本原因，而运行手册可用来缓解其带来的风险。行动手册是贵组织事件响应计划的必要组成部分。

出色的行动手册有几个主要特点。它逐步指导用户完成事件发现过程。由外而内地思考，用户应执行哪些步骤来诊断事件？如果行动手册中需要特殊工具或提升的权限，请在行动手册中明确地定义。请制定沟通计划，以向利益相关者提供有关调查状态的最新信息，这是事件响应计划的一个重要组成部分。在无法确定根本原因的情况下，行动手册应制定上报计划。如果确定了根本原因，行动手册应指出介绍如何解决根本原因的运行手册。应集中存储并定期维护行动手册。如果行动手册用于特定提醒，请向团队提供关于提醒中的行动手册的提示。

随着组织日趋成熟，可自动实施行动手册。从包含低风险事件的行动手册开始实施。使用脚本自动执行发现步骤。确保有配套的运行手册来缓解常见根本原因带来的风险。

期望的结果：您的组织有针对常见事件的行动手册。行动手册集中存储在一个位置，可供团队成员使用。行动手册经常进行更新。对于任何已知的根本原因，将制定配套的运行手册。

常见反模式：

- 要调查事件，并没有标准方法。
- 团队成员依靠肌肉记忆或对机构的了解，对失败的部署进行排查。
- 新的团队成员将学习如何通过试错法来调查问题。
- 调查问题的最佳实践无法在不同团队之间共享。

建立此最佳实践的好处：

- 行动手册可帮助您减轻事件带来的影响。
- 不同的团队成员可使用同一行动手册，以一致的方式确定根本原因。

- 可以针对已知的根本原因制定运行手册，从而加快恢复速度。
- 团队成员根据行动手册能够更快地开始行动。
- 团队可以使用可重复的行动手册来扩展其流程。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

制定和使用行动手册的方式取决于组织的成熟度。如果您是初次使用云，请在中央文档存储库中以文本形式制定行动手册。随着组织日趋成熟，可以使用 Python 等脚本语言实现行动手册的半自动化。可以在 Jupyter notebook 中运行这些脚本来加快发现速度。先进的组织已针对可通过运行手册自动修正的常见问题，制定完全自动化的行动手册。

通过列出工作负载所发生的常见事件，开始制定行动手册。为风险较低且根本原因范围已缩小到几个问题的事件选择行动手册。在为较简单的场景制定行动手册后，可以着手处理风险较高的场景或根本原因尚不确定的场景。

随着贵组织日趋成熟，您的文本样式的行动手册应实现自动化。通过使用诸如 [AWS Systems Manager Automations](#) 之类的服务，可以将纯文本转换为自动化代码。可以针对工作负载运行这些自动化代码，从而加快调查速度。可以激活这些自动化代码以响应事件，从而减少发现和解决事件所需的平均时间。

客户可以使用 [AWS Systems Manager Incident Manager](#) 来响应事件。此服务提供了单一界面对事件进行分类，在发现和缓解问题期间通知利益相关者，并在整个事件中进行协作。它使用 AWS Systems Manager Automations 加快检测和恢复的速度。

## 客户示例

生产事件影响了 AnyCompany Retail。随时待命的工程师根据行动手册调查了问题。随着他们逐步地解决问题，他们不断为行动手册中确定的关键利益相关者提供最新信息。工程师最终确定，根本原因是后端服务中出现竞态条件。根据运行手册，工程师重新启动了该服务，并使 AnyCompany Retail 重新联机。

## 实施步骤

如果您当前没有文档存储库，建议您为行动手册库创建版本控制存储库。您可以使用 Markdown 制定您的行动手册，该服务兼容大多数行动手册自动化系统。如果您从头开始制定行动手册，请使用以下行动手册示例模板。

```
# Playbook Title ## Playbook Info | Playbook ID | Description
| Tools Used | Special Permissions | Playbook Author | Last
```

```
Updated | Escalation POC | Stakeholders | Communication Plan |
|-----|-----|-----|-----|-----|-----|-----|-----|-----| | RUN001
| What is this playbook for? What incident is it used for? | Tools | Permissions |
Your Name | 2022-09-21 | Escalation Name | Stakeholder Name | How will updates be
communicated during the investigation? | ## Steps 1.Step one 2.Step two
```

1. 如果您当前没有文档存储库或 Wiki，请在版本控制系统中为行动手册创建一个新的版本控制存储库。
2. 确定需要调查的一个常见问题。它应该是根本原因范围限于几个问题且解决方案风险较低的场景。
3. 使用 Markdown 模板填写 Playbook Name##### 部分以及 Playbook Info#####下的字段。
4. 填写问题排查步骤。尽可能清楚地知道要采取哪些行动，或者应调查哪些方面。
5. 将行动手册提供给团队成员，让他们仔细阅读并加以验证。如果发现有遗漏之处或某些内容不清楚，请更新行动手册。
6. 在文档存储库中发布您的行动手册，并告知您的团队和任何利益相关者。
7. 随着您添加更多的行动手册，这个行动手册库将会不断扩大。在您拥有多个行动手册后，可以开始使用 AWS Systems Manager Automations 等工具自动执行它们，从而使自动化操作和行动手册保持同步。

实施计划的工作量级别：低。行动手册应该是集中存储在一个位置的文本文档。对于更加成熟的组织，将转为自动实施行动手册。

## 资源

### 相关最佳实践：

- [OPS02-BP02 确定流程和程序所有者](#)：行动手册应该有一个负责人来负责维护。
- [OPS07-BP03 使用运行手册执行程序](#)：运行手册和行动手册类似，但有一个关键区别：运行手册包含期望的结果。在许多情况下，一旦行动手册确定了根本原因，就会使用运行手册。
- [OPS10-BP01 使用流程来管理事件、意外事件和问题](#)：行动手册是良好的事件、意外事件和问题管理实践的一部分。
- [OPS10-BP02 针对每个提醒设置一个流程](#)：应使用运行手册和行动手册来响应警报。随着时间的推移，应自动进行这些响应。
- [OPS11-BP04 执行知识管理](#)：维护行动手册是知识管理的一个关键部分。

### 相关文档：

- [利用自动化行动手册和运行手册实现卓越运营](#)
- [AWS Systems Manager：使用运行手册](#)
- [使用 AWS Systems Manager Automation 运行手册解决运营任务](#)

相关视频：

- [AWS re:Invent 2019：运行手册、事件报告和事件响应 DIY 指南 \( SEC318-R1 \)](#)
- [AWS Systems Manager Incident Manager – AWS 虚拟研讨会](#)
- [将脚本集成到 AWS Systems Manager](#)

相关示例：

- [AWS 客户行动手册框架](#)
- [AWS Systems Manager：自动化演练](#)
- [使用 Jupyter notebook 和 CloudTrail Lake 构建 AWS 事件响应运行手册](#)
- [Rubix – 用于在 Jupyter notebook 中构建运行手册的 Python 库](#)
- [使用 Document Builder 创建自定义运行手册](#)
- [Well-Architected 实验室：根据行动手册和运行手册自动完成操作](#)
- [Well-Architected 实验室：使用 Jupyter 的事件响应行动手册](#)

相关服务：

- [AWS Systems Manager Automation](#)
- [AWS Systems Manager Incident Manager](#)

## OPS07-BP05 做出明智的决策来部署系统和变更

为工作负载的成功和不成功变更制定恰当的流程。故障演练是一种演习，团队模拟发生故障的情况来制定缓解策略。使用故障演练来预测故障，并在适当的时候创建程序。评估将变更部署到工作负载所获得好处和产生的风险。确认所有变更符合监管要求。

期望结果：

- 将变更部署到工作负载时作出明智的决策。
- 变更符合监管要求。

## 常见反模式：

- 将变更部署到工作负载，而没有处理失败部署的流程。
- 对生产环境作出不符合监管要求的变更。
- 部署新版本的工作负载，而不为资源利用建立基准。

## 建立此最佳实践的好处：

- 为工作负载的不成功变更做好了准备。
- 工作负载的变更符合监管策略。

在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

使用故障演练制定不成功变更的流程。记录不成功变更的流程。确保所有变更符合监管要求。评估将变更部署到工作负载所获得好处和产生的风险。

## 客户示例

AnyCompany Retail 定期执行故障演练以验证他们的不成功变更流程。他们在共享的 Wiki 中记录他们的流程并经常更新。所有变更符合监管要求。

## 实施步骤

1. 将变更部署到工作负载时作出明智的决策。确立并审查成功部署的条件。制定将触发变更回滚的方案或条件。在部署变更带来的好处与不成功变更的风险之间进行权衡。
2. 确认所有变更符合监管政策。
3. 使用故障演练为不成功的变更制定计划，并记录缓解策略。运行桌面练习，为不成功的变更建模，并验证回滚程序。

实施计划的工作量级别：适中。实施故障演练的实践需要整个组织内的利益攸关方进行协调和付出努力。

## 资源

## 相关最佳实践：

- [OPS01-BP03 评估治理要求](#) - 监管要求是确定是否部署变更的关键因素。



- [OPS06-BP01 针对不成功的更改制定计划](#) - 制定计划来缓和失败的部署并使用故障演练来验证它们。
- [OPS06-BP02 测试部署](#) - 在部署之前应适当地测试每项软件变更，以便减少生产中的缺陷。
- [OPS07-BP01 确保员工能力](#) - 有足够训练有素的人员来支持工作负载，这对于作出部署系统变更的明智决定很重要。

相关文档：

- [Amazon Web Services：风险与合规](#)
- [AWS 责任共担模式](#)
- [AWS Cloud 中的监管：敏捷性和安全性之间的适当平衡](#)

### OPS07-BP06 为生产工作负载启用支持计划

为生产工作负载所依赖的所有软件和服务启用支持。选择适当的支持级别以满足您的生产服务级别需求。以防出现服务中断或软件问题，这些依赖项的支持计划是必需的。记录支持计划以及如何要求所有服务和软件供应商提供支持。实施机制以确认主要支持联系人的信息保持最新。

期望结果：

- 为生产工作负载所依赖的软件和服务实施支持计划。
- 根据服务级别需求选择适当的支持计划。
- 记录支持计划、支持级别以及如何请求支持。

常见反模式：

- 您没有制定关键软件供应商的支持计划。您的工作负载受到影响，您无法采取任何措施来加快修复或从供应商获得及时更新。
- 作为软件供应商主要联系人的开发人员离开了公司。您无法直接联系供应商支持人员。您必须花时间研究和浏览通用联系系统，延长在需要时作出反应所需的时间。
- 软件供应商发生生产中断。没有关于如何提交支持案例的文档。

建立此最佳实践的好处：

- 通过适当的支持级别，您可以在满足服务级别需求所需的时间范围内获得响应。

- 作为受支持的客户，如果存在生产问题，您可以上报。
- 发生事件时，软件和服务供应商会协助排除故障。

在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

为生产工作负载所依赖的所有软件和服务供应商启用支持计划。设置适当的支持计划以满足服务级别需求。对于 AWS 客户，这意味着要在具有生产工作负载的任何账户上启用 AWS Business Support 或更高级别的支持。定期与支持供应商会面，获取有关支持产品、流程和联系人的更新。记录如何向软件和服务供应商请求支持，包括在出现中断时如何上报。实施特定机制，使支持联系人的信息保持最新。

## 客户示例

在 AnyCompany Retail，所有商用软件和服务依赖项均有支持计划。例如，他们在有生产工作负载的所有账户上启用 AWS Enterprise Support。如果出现问题，任何开发人员都可以提出支持案例。有一个 Wiki 页面，其中包含有关如何请求支持、向谁发出通知以及加快处理案例最佳实践的信息。

## 实施步骤

1. 与组织内的利益攸关方一起确定您的工作负载所依赖的软件和服务供应商。记录这些依赖项。
2. 确定工作负载的服务级别需求。选择与需求匹配的支持计划。
3. 对于商用软件和服务，与供应商一起制定支持计划。
  - a. 为所有生产账户订阅 AWS Business Support 或更高级别的支持，可以让 AWS 支持 更快响应，并且我们强烈建议订阅此支持。如果您没有 Premium Support，则必须制定行动计划来处理问题，而这需要来自 AWS 支持 的帮助。AWS 支持 提供工具和技术的组合、人员和计划，旨在主动帮助您优化性能、降低成本和加快创新速度。AWS Business Support 可带来额外的好处，包括访问 AWS Trusted Advisor 和 AWS Personal Health Dashboard，并更快响应。
4. 在您的知识管理工具中记录支持计划。包括如何请求支持、在提出支持案例时应通知谁以及在发生事件时如何上报。Wiki 是一种很好的机制，让任何人都可以在发现支持流程或联系人的更改时对文档进行必要的更新。

实施计划的工作量级别：低。大部分软件和服务供应商都提供选择加入的支持计划。在知识管理系统上记录和分享支持最佳实践，可以确认您的团队知道在出现生产问题时该怎么做。

## 资源

相关最佳实践：

- [OPS02-BP02 确定流程和程序所有者](#)

相关文档：

- [AWS 支持 计划](#)

相关服务：

- [AWS Business Support](#)
- [AWS Enterprise Support](#)

## 运营

问题

- [运营 8.如何在组织中利用工作负载可观测性？](#)
- [运营 9.您如何了解自己的运营状况？](#)
- [OPS 10.您如何应对工作负载事件和运营事件？](#)

### 运营 8.如何在组织中利用工作负载可观测性？

利用可观测性确保最佳工作负载运行状况。利用相关的指标、日志和跟踪数据，来全面了解工作负载的性能并有效地解决问题。

最佳实践

- [OPS08-BP01 分析工作负载指标](#)
- [OPS08-BP02 分析工作负载日志](#)
- [OPS08-BP03 分析工作负载跟踪数据](#)
- [OPS08-BP04 创建可操作的警报](#)
- [OPS08-BP05 创建控制面板](#)

#### OPS08-BP01 分析工作负载指标

实施应用程序遥测后，定期分析收集的指标。虽然延迟、请求、错误和容量（或配额）有助于深入了解系统性能，但优先审查业务成果指标至关重要。这样可以确保您做出与业务目标相一致的数据驱动型决策。

期望的结果：准确洞察工作负载性能，推动做出以数据为依据的决策，确保与业务目标相一致。

常见反模式：

- 孤立地分析指标，而不考虑其对业务成果的影响。
- 过度依赖技术指标，而不重视业务指标。
- 很少审查指标，错过了实时决策机会。

建立此最佳实践的好处：

- 进一步了解技术性能与业务成果之间的相互关系。
- 以实时数据为依据改善决策流程。
- 在问题影响业务结果之前主动找出和缓解问题。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

利用 Amazon CloudWatch 之类的工具执行指标分析。AWS Cost Anomaly Detection 和 Amazon DevOps Guru 之类的 AWS 服务可用于检测异常，尤其是在静态阈值未知，或行为模式更适合异常检测时。

实施步骤

1. 分析和审查：定期审查和解释您的工作负载指标。
  - a. 优先考虑业务成果指标，而不是只考虑纯粹的技术指标。
  - b. 了解数据中的高峰、低谷或模式的重要性。
2. 利用 Amazon CloudWatch：使用 Amazon CloudWatch 获得集中式视图并进行深入分析。
  - a. 配置 CloudWatch 控制面板，以可视化形式呈现您的指标，并对一段时间内的指标进行比较。
  - b. 使用 [CloudWatch 中的百分位数](#) 来清楚地了解指标分布，这有助于定义 SLA 和理解异常值。
  - c. 设置 [AWS Cost Anomaly Detection](#) 在不依赖静态阈值的情况下识别异常模式。
  - d. 实施 [CloudWatch 跨账户可观测性](#) 以监控跨区域内多个账户的应用程序并对其进行故障排除。
  - e. 使用 [CloudWatch Metric Insights](#) 来查询和分析跨账户和地区的指标数据，从而识别趋势和异常情况。
  - f. 应用 [CloudWatch Metric Math](#) 对您的指标进行转换、汇总或执行计算，从而获得更深入的见解。

3. 应用 Amazon DevOps Guru：纳入 [Amazon DevOps Guru](#) 以利用其机器学习增强的异常检测，来识别无服务器应用程序操作问题的早期迹象，并在它们影响客户之前将其修复。
4. 根据见解进行优化：根据您的指标分析做出明智的决策，以调整和改进您的工作负载。

实施计划的工作量级别：中

资源

相关最佳实践：

- [OPS04-BP01 识别关键绩效指标](#)
- [OPS04-BP02 实施应用程序遥测](#)

相关文档：

- [The Wheel 博客 - 强调持续审查指标的重要性](#)
- [百分位很重要](#)
- [使用 AWS Cost Anomaly Detection](#)
- [CloudWatch 跨账户可观测性](#)
- [使用 CloudWatch Metrics Insights 查询您的指标](#)

相关视频：

- [Enable Cross-Account Observability in Amazon CloudWatch](#)
- [Introduction to Amazon DevOps Guru](#)
- [Continuously Analyze Metrics using AWS Cost Anomaly Detection](#)

相关示例：

- [One Observability Workshop](#)
- [Gaining operation insights with AIOps using Amazon DevOps Guru](#)

## OPS08-BP02 分析工作负载日志

定期分析工作负载日志对于更深入地了解应用程序的运行至关重要。通过高效地筛选、以可视化方式呈现和解释日志数据，您可以持续优化应用程序性能和安全性。

期望的结果：通过全面的日志分析获得对应用程序行为和操作的丰富见解，确保主动检测和缓解问题。

常见反模式：

- 在出现严重问题之前，忽视对日志的分析。
- 没有使用可分析日志的全套工具，导致错过关键见解。
- 仅依靠人工查看日志，而不利用自动化和查询功能。

建立此最佳实践的好处：

- 主动识别运营瓶颈、安全威胁和其他潜在问题。
- 高效利用日志数据进行持续的应用程序优化。
- 增进对应用程序行为的理解，有助于调试和故障排除。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

[Amazon CloudWatch Logs](#) 是用于日志分析的强大工具。利用 CloudWatch Logs Insights 和 Contributor Insights 等集成功能，可以直观而高效地从日志中获取有意义的信息。

实施步骤

1. 设置 CloudWatch Logs：配置应用程序和服务以将日志发送到 CloudWatch Logs。
2. 设置 CloudWatch Logs Insights：使用 [CloudWatch Logs Insights](#) 以交互方式搜索和分析您的日志数据。
  - a. 创建查询以提取模式、直观呈现日志数据并得出切实可行的见解。
3. 利用 Contributor Insights 使用 [CloudWatch Contributor Insights](#) 在 IP 地址或用户代理等高基数维度中找到主要贡献者。
4. 实现 CloudWatch Logs 指标筛选器：配置 [CloudWatch 日志指标筛选器](#) 将日志数据转换为可操作的指标。这允许您设置警报或进一步分析模式。
5. 定期审查和优化：定期审查您的日志分析策略，以捕获所有相关信息并持续优化应用程序性能。

实施计划的工作量级别：中。

## 资源

相关最佳实践：

- [OPS04-BP01 识别关键绩效指标](#)
- [OPS04-BP02 实施应用程序遥测](#)
- [OPS08-BP01 分析工作负载指标](#)

相关文档：

- [使用 CloudWatch Logs Insights 分析日志数据](#)
- [使用 CloudWatch Contributor Insights](#)
- [创建和管理 CloudWatch Logs 日志指标筛选器](#)

相关视频：

- [使用 CloudWatch Logs Insights 分析日志数据](#)
- [Use CloudWatch Contributor Insights to Analyze High-Cardinality Data](#)

相关示例：

- [CloudWatch Logs 示例查询](#)
- [可观测性研讨会](#)

## OPS08-BP03 分析工作负载跟踪数据

分析跟踪数据对于全面了解应用程序的运营过程至关重要。通过以可视化方式呈现和理解各个组件之间的交互，可以微调性能，识别瓶颈，并增强用户体验。

期望的结果：清晰地了解应用程序的分布式操作，从而更快地解决问题并增强用户体验。

常见反模式：

- 忽略跟踪数据，仅依赖日志和指标。
- 不将跟踪数据与关联日志联系起来。
- 忽略从跟踪数据中得出的指标，例如延迟和故障率。



建立此最佳实践的好处：

- 改善故障排除并缩短平均解决时间（MTTR）。
- 深入了解依赖项及其影响。
- 迅速发现并纠正性能问题。
- 利用从跟踪数据中得出的指标做出明智的决策。
- 通过优化组件交互来改善用户体验。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

[AWS X-Ray](#) 提供了一个完整套件来分析跟踪数据，从而提供服务交互的整体视图、监控用户活动并检测性能问题。ServiceLens、X-Ray Insights、X-Ray Analytics 和 Amazon DevOps Guru 等功能，可增强从跟踪数据中获得的可操作见解的深度。

## 实施步骤

以下步骤提供了一种结构化方法，可使用 AWS 服务有效地实施跟踪数据分析：

1. 集成 AWS X-Ray：确保 X-Ray 已与您的应用程序集成，来捕获跟踪数据。
2. 分析 X-Ray 指标：深入研究从 X-Ray 跟踪数据中得出的指标，例如延迟、请求速率、故障率和响应时间分布，方法是使用 [服务地图](#) 来监控应用程序运行状况。
3. 使用 ServiceLens：利用 [ServiceLens 地图](#) 来增强您的服务和应用程序的可观测性。这允许以集成方式查看跟踪、指标、日志、警报和其他运行状况信息。
4. 启用 X-Ray Insights：
  - a. 开启 [X-Ray Insights](#) 以自动检测跟踪数据中的异常。
  - b. 研究见解以查明模式并确定根本原因，例如故障率或延迟增加。
  - c. 查阅见解时间表，按时间顺序分析检测到的问题。
5. 使用 X-Ray Analytics：[X-Ray Analytics](#) 可用于全面探究跟踪数据、查明模式和挖掘见解。
6. 在 X-Ray 中使用群组：在 X-Ray 中创建群组，以根据高延迟等标准筛选跟踪，从而进行更有针对性的分析。
7. 纳入 Amazon DevOps Guru：利用 [Amazon DevOps Guru](#) 受益于机器学习模型，查明跟踪数据中的操作异常。

8. 使用 CloudWatch Synthetics：使用 [CloudWatch Synthetics](#) 来创建用于持续监控您的端点和 workflows 的金丝雀。这些金丝雀可以与 X-Ray 集成以提供跟踪数据，用于对正在测试的应用程序进行深入分析。
9. 使用真实用户监控 ( RUM )：使用 [AWS X-Ray](#) 和 [CloudWatch RUM](#)，您可以通过下游 AWS 托管服务，从应用程序的最终用户开始分析和调试请求路径。这有助于您识别影响用户的延迟趋势和错误。
10. 与日志关联：将跟踪数据 [与相关日志关联](#)（在 X-Ray 跟踪视图中），可以详细了解应用程序行为。这允许您查看与跟踪的事务直接关联的日志事件。

实施计划的工作量级别：中。

资源

相关最佳实践：

- [OPS08-BP01 分析工作负载指标](#)
- [OPS08-BP02 分析工作负载日志](#)

相关文档：

- [使用 ServiceLens 监控应用程序运行状况](#)
- [使用 X-Ray Analytics 探究跟踪数据](#)
- [使用 X-Ray Insights 检测跟踪数据中的异常](#)
- [使用 CloudWatch Synthetics 进行持续监控](#)

相关视频：

- [Analyze and Debug Applications Using Amazon CloudWatch Synthetics and AWS X-Ray](#)
- [Use AWS X-Ray Insights](#)

相关示例：

- [One Observability Workshop](#)
- [使用 X-Ray 实现 AWS Lambda](#)
- [CloudWatch Synthetics 金丝雀模板](#)

## OPS08-BP04 创建可操作的警报

及时检测和响应应用程序行为的偏差至关重要。尤其重要的是，识别基于关键绩效指标 ( KPI ) 的结果何时处于风险当中，或何时出现意外的异常情况。基于 KPI 的警报可确保您收到的信号与业务或运营影响直接相关。这种可操作警报的方法可促进主动响应，并有助于维护系统性能和可靠性。

期望的结果：接收及时、相关且可操作的警报，以便快速找出和缓解潜在问题，尤其是在 KPI 结果面临风险时。

常见反模式：

- 设置过多非关键警报，导致警报疲劳。
- 不根据 KPI 对警报进行优先级排序，因此很难理解问题对业务的影响。
- 忽视根本原因，导致针对同一问题出现重复警报。

建立此最佳实践的好处：

- 通过关注可操作且相关的警报，减少警报疲劳。
- 通过主动检测和缓解问题，改善系统的正常运行时间和可靠性。
- 通过与常用的警报和通信工具集成，增进团队协作并更快解决问题。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

要创建有效的警报机制，必须使用指标、日志和跟踪数据来标记基于 KPI 的结果何时存在风险，或何时检测到异常情况。

### 实施步骤

1. 定义关键绩效指标 ( KPI )：确定应用程序的 KPI。警报应与这些 KPI 相关联，以准确反映业务影响。
2. 实现异常检测：
  - 使用 AWS Cost Anomaly Detection：设置 [AWS Cost Anomaly Detection](#) 以自动检测异常模式，确保仅针对真正的异常情况生成警报。
  - 使用 X-Ray Insights：
    - a. 设置 [X-Ray Insights](#) 以检测跟踪数据中的异常。

- b. 配置 [X-Ray Insights 通知](#) 以便在检测到问题时收到提醒。
- 与 DevOps Guru 集成：
  - a. 利用 [Amazon DevOps Guru](#) 的机器学习功能，结合现有数据来检测操作异常。
  - b. 导航到 [通知设置](#)（在 DevOps Guru 中）以设置异常警报。
3. 实施可操作的警报：设计能够提供足够信息的警报，以便立即采取行动。
4. 减少警报疲劳：尽量减少非关键警报。如果存在大量无关紧要的警报，团队将不堪重负，可能导致忽视关键问题，并降低警报机制的整体有效性。
5. 设置复合警报：使用 [Amazon CloudWatch 复合警报](#) 来整合多个警报。
6. 与警报工具集成：纳入多个工具，例如 [Ops Genie](#) 和 [PagerDuty](#)。
7. 利用 Amazon Q Developer in chat applications 集成 [Amazon Q Developer in chat applications](#) 以便将警报转发到 Chime、Microsoft Teams 和 Slack。
8. 基于日志的警报：使用 [日志指标筛选器](#)（在 CloudWatch 中）来根据特定的日志事件创建警报。
9. 查看并迭代：定期重新审视和完善警报配置。

实施计划的工作量级别：中。

## 资源

相关最佳实践：

- [OPS04-BP01 识别关键绩效指标](#)
- [OPS04-BP02 实施应用程序遥测](#)
- [OPS04-BP03 实施用户体验遥测](#)
- [OPS04-BP04 实施依赖项遥测](#)
- [OPS04-BP05 实施分布式跟踪](#)
- [OPS08-BP01 分析工作负载指标](#)
- [OPS08-BP02 分析工作负载日志](#)
- [OPS08-BP03 分析工作负载跟踪数据](#)

相关文档：

- [使用 Amazon CloudWatch 警报](#)

- [创建复合警报](#)
- [基于异常检测创建 CloudWatch 警报](#)
- [DevOps Guru 通知](#)
- [X-Ray Insights 通知](#)
- [使用交互式 ChatOps 对 AWS 资源进行监控、操作和故障排除](#)
- [Amazon CloudWatch 集成指南 | PagerDuty](#)
- [将 OpsGenie 与 Amazon CloudWatch 集成](#)

相关视频：

- [Create Composite Alarms in Amazon CloudWatch](#)
- [Amazon Q Developer in chat applications Overview](#)
- [AWS on Air ft. Mutative Commands in Amazon Q Developer in chat applications](#)

相关示例：

- [Alarms, incident management, and remediation in the cloud with Amazon CloudWatch](#)
- [Tutorial: Creating an Amazon EventBridge rule that sends notifications to Amazon Q Developer in chat applications](#)
- [One Observability Workshop](#)

## OPS08-BP05 创建控制面板

控制面板是以人为本的视图，可用于查看工作负载的遥测数据。虽然它们提供了重要的可视化界面，但它们不应取代警报机制，而是补充警报机制。如果精心设计，它们不仅能迅速洞察系统的运行状况和性能，还能为利益相关方提供有关业务成果和问题影响的实时信息。

期望的结果：使用可视化形式，清晰地了解系统和业务运行状况，并可据此采取行动。

常见反模式：

- 指标过多，使得控制面板过于复杂。
- 依靠的控制面板不会对检测到的异常情况发出警报。
- 不会随着工作负载的演进而更新控制面板。

建立此最佳实践的好处：

- 即时了解关键系统指标和 KPI。
- 增进利益相关方的沟通和理解。
- 快速洞察运营问题的影响。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

### 以业务为中心的控制面板

为业务 KPI 量身定制的控制面板可吸引更广泛的利益相关方。尽管这些人可能对系统指标不感兴趣，但他们热衷于了解这些数字对业务的影响。以业务为中心的控制面板可确保所监控和分析的所有技术和运营指标与总体业务目标同步。这种一致性可让每个人了解什么是至关重要的，什么不太重要，并就此达成共识。此外，突出业务 KPI 的控制面板往往更具操作性。利益相关方可以快速了解运营状况、需要关注的领域以及对业务成果的潜在影响。

考虑到这一点，在创建控制面板时，请确保技术指标和业务 KPI 之间保持平衡。两者都至关重要，但它们面向不同的受众。理想情况下，您拥有的控制面板应该有助于您全面了解系统的运行状况和性能，同时还要强调关键业务成果及其影响。

Amazon CloudWatch 控制面板是 CloudWatch 控制台中的可自定义主页，方便您通过单一视图监控您的资源，即使这些资源分布在不同 AWS 区域和账户中。

## 实施步骤

1. 创建基本控制面板：[在 CloudWatch 中创建一个新控制面板](#)，给它起一个描述性名称。
2. 使用 Markdown 小部件：在深入研究指标之前，请使用[Markdown 小部件](#)在控制面板顶部添加文字背景信息。这应该解释控制面板涵盖的内容、所表示的指标的重要性，还可以包含指向其他控制面板和故障排除工具的链接。
3. 创建控制面板变量：[适当时纳入控制面板变量](#)，以创建动态和灵活的控制面板视图。
4. 创建指标小部件：[添加指标小部件](#)以可视化形式呈现应用程序发出的各种指标，定制这些小部件以有效呈现系统运行状况和业务成果。
5. Log Insights 查询：利用[CloudWatch Logs Insights](#)从日志中获取可操作的指标，并在控制面板上显示这些见解。
6. 设置警报：将[CloudWatch 警报](#)集成到您的控制面板，可以快速查看任何超出阈值的指标。

7. 使用 Contributor Insights：纳入 [CloudWatch Contributor Insights](#) 以分析高基数字段，并更清楚地了解您的资源的主要贡献者。
8. 设计自定义小部件：如果标准小部件无法满足特定需求，可以考虑创建 [自定义小部件](#)。它们可以从各种数据源中提取数据，也可以以独特方式表示数据。
9. 迭代和完善：随着应用程序的演进，请定期重新审视控制面板以确保其仍然适用。

## 资源

### 相关最佳实践：

- [OPS04-BP01 识别关键绩效指标](#)
- [OPS08-BP01 分析工作负载指标](#)
- [OPS08-BP02 分析工作负载日志](#)
- [OPS08-BP03 分析工作负载跟踪数据](#)
- [OPS08-BP04 创建可操作的警报](#)

### 相关文档：

- [构建控制面板以获取操作可见性](#)
- [使用 Amazon CloudWatch 控制面板](#)

### 相关视频：

- [创建跨账户和跨区域 CloudWatch 控制面板](#)
- [AWS re:Invent 2021 - Gain enterprise visibility with AWS Cloud operation dashboards](#)

### 相关示例：

- [可观测性研讨会](#)
- [使用 Amazon CloudWatch 进行应用程序监控](#)

## 运营 9.您如何了解自己的运营状况？

定义、记录和分析运营指标以便了解运营事件，从而采取适当的措施。



## 最佳实践

- [OPS09-BP01 使用指标衡量运营目标和 KPI](#)
- [OPS09-BP02 通报状态和趋势，确保了解运营情况](#)
- [OPS09-BP03 审查运营指标并确定改进优先顺序](#)

### OPS09-BP01 使用指标衡量运营目标和 KPI

从您的组织获取定义运营成功的目标和 KPI，并确定指标反映了这些目标和 KPI。将基线设置为参考点，并定期重新评估。制定机制，从团队那里收集这些指标以供评估。

期望的结果：

- 组织运营团队的目标和 KPI 已发布并共享。
- 已建立反映这些 KPI 的指标。示例可能包括：
  - 工单队列深度或平均工单时长
  - 按问题类型分组的工单数量
  - 使用或不使用标准化操作程序 (SOP) 时处理问题所花费的时间
  - 从失败的代码推送中恢复所花费的时间
  - 呼叫量

常见反模式：

- 由于开发人员被拉去执行故障排除任务，因此而错过部署截止日期。开发团队主张增加人手，但由于无法衡量被占用的时间，因此无法量化他们需要多少人手。
- 设置了一级服务台来处理用户呼叫。随着时间的推移，工作负载越来越多，但没有为一级服务台分配人手。随着通话次数的增加以及问题解决时间的延长，客户满意度下降，但管理层看不到此类问题的任何指标，因此未采取任何行动。
- 有问题的 workload 被移交给单独的运营团队进行处理。与其他 workload 不同，这种新的 workload 没有提供适当的文档和运行手册。因此，团队需要花费更长的时间解决和排除故障。但是，没有任何指标记录这一点，这使得问责制变得难以实施。

建立此最佳实践的好处： workload 监控可以显示应用程序和服务的状态，而监控运营团队则可以让所有者深入了解这些 workload 的使用者之间的变化，例如不断变化的业务需求。通过创建能够反映运营状态的指标，衡量这些团队的效率，并根据业务目标对其进行评估。指标可以突出显示支持问题，或确定何时出现偏离服务级别目标的情况。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

安排时间与业务主管和利益相关方会面，以确定服务的总体目标。确定各个运营团队的任务，以及他们可能应对哪些挑战。利用这些信息，针对可能反映这些运营目标的关键绩效指标 ( KPI ) 进行集思广益。这些指标可能是客户满意度、从功能构思到部署所花的时间、平均问题解决时间等。

根据 KPI，确定可能最能反映这些目标的指标和数据来源。客户满意度可能是各种指标的组合，例如呼叫等待或回复时间、满意度得分和提出的问题类型。部署时间可能是测试和部署所需的时间，加上需要添加的所有部署后修复的总和。显示不同类型问题所花费的时间（或这些问题的数量）的统计数据，可以提供一个窗口，让您了解需要在哪些方面开展有针对性的工作。

## 资源

相关文档：

- [QuickSight - 使用 KPI](#)
- [Amazon CloudWatch - 使用指标](#)
- [构建控制面板](#)
- [How to track your cost optimization KPIs with KPI Dashboard](#)

## OPS09-BP02 通报状态和趋势，确保了解运营情况

了解运营状况及其趋势非常有必要，这样才能确定结果何时可能面临风险、是否可以支持新增的工作，或者变更对团队的影响。在运营活动期间，用户和运营团队可通过状态页面获取信息，从而减轻通信渠道的压力并主动传播信息。

期望的结果：

- 运营领导者可以一目了然地了解其团队正在处理的呼叫量，以及可能正在开展的工作（如部署）。
- 当正常运营受到影响时，会向利益相关方和用户群体发出警报。
- 组织领导层和利益相关方可以查看状态页面，以响应警报或影响，并获取与运营事件相关的信息，如联系人、工单信息和预计恢复时间。
- 向领导层和其他利益相关方提供的报告可显示运营统计数据，例如一段时间内的呼叫量、用户满意度分数、未处理工单的数量及其存在时间。

常见反模式：

- 工作负载出现故障，导致服务不可用。用户想知道发生了什么情况，呼叫量激增。管理人员想知道谁在处理问题，从而进一步增加了呼叫量。各运营团队都在努力调查问题，导致工作重复。
- 由于人们渴望获得新功能，数名人员被重新分配到工程工作中。但没有提供后补人员，问题解决时间激增。这些信息没有被记录下来，几周后，在收到用户表达不满的反馈时，领导层才意识到这个问题。

建立此最佳实践的好处：在业务受到影响的运营事件中，为了解情况而向不同团队查询信息可能会浪费大量时间和精力。通过建立广泛传播的状态页面和控制面板，利益相关方可以快速获得相关信息，例如是否发现了问题、谁在负责处理问题，或者预计何时可以恢复正常运行。这样，团队成员就不必花太多时间与他人沟通状态，而是可以将更多时间花在解决问题上。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

构建控制面板，显示运营团队当前的关键指标，并使运营主管和管理层都能随时访问这些指标。

构建可以快速更新的状态页面，以显示事件何时发生、由谁负责以及谁在协调响应。在此页面上分享用户应考虑的任何步骤或解决方法，并广为分发该位置。鼓励用户在遇到未知问题时先查看此位置。

收集并提供显示一段时间内的运营状况的报告，并将其分发给领导者和决策者，以阐述运营工作以及挑战和需求。

在团队之间共享这些指标和报告，这些指标和报告最能反映目标和 KPI，以及它们在推动变革方面的影响力。在这些活动中投入时间，提升运营在团队内部和团队之间的重要性。

## 资源

相关文档：

- [衡量进度](#)
- [构建控制面板以获取操作可见性](#)

相关解决方案：

- [数据操作](#)

## OPS09-BP03 审查运营指标并确定改进优先顺序

留出专门的时间和资源来审查运营状况，可确保为日常业务提供服务始终是优先事项。召集运营主管和利益相关方，定期审查指标，重申或修改长期和短期目标，并确定改进的优先顺序。

期望的结果：

- 运营主管和员工定期开会，审查给定报告期内的指标。交流挑战，庆祝胜利，分享经验教训。
- 定期向利益相关方和业务领导通报运营状况，并征求他们对目标、KPI 和未来举措的意见。结合相关背景，讨论服务交付、运行和维护之间的权衡。

常见反模式：

- 推出了一款新产品，但一级和二级运营团队没有接受过充分的培训，无法为其提供支持，或者没有相应地增加人手。领导者看不到表明工单解决时间缩短和事件量增加的指标。几周后，心怀不满的用户离开平台，订阅数量开始下降，此时才采取行动。
- 对工作负载进行维护的手动流程已经存在很长时间。尽管人们一直渴望实现自动化，但考虑到该系统的重要性较低，并未得到足够的重视。然而，随着时间的推移，该系统的重要性与日俱增，现在这些手动过程耗费了运营团队的大部分时间。没有安排资源为运营团队提供更多工具，这导致随着工作量的增加，员工疲惫不堪。当有人报告员工离职去了其他竞争对手那里时，领导层才意识到这一点。

建立此最佳实践的好处： 在一些组织中，如何将同样的时间和精力用于提供新产品或服务，可能是一项挑战。一旦出现这种情况，预期的服务水平会慢慢降低，业务线就会受到影响。这是因为运营团队没有随着业务的增长而做出改变和发展，很快就被甩在后面。如果不定期审查运营团队收集的见解，等到发现业务面临的风险时，可能为时已晚。通过花时间与运营人员和领导层一起审查指标和程序，运营团队所发挥的关键作用将始终可见，并且能在风险达到临界水平之前及早发现。运营团队可以更好地洞察即将发生的业务变化和举措，从而积极主动地开展工作。领导层对运营指标的了解展示了这些团队在客户满意度（包括内部和外部客户满意度）方面所发挥的作用，使他们能够更好地权衡选择的优先事项，或确保运营团队有足够的时间和资源随着新业务和工作负载计划的变化而做出改变和发展。

未建立这种最佳实践的情况下暴露的风险等级： 中

实施指导

花时间与利益相关方和运营团队一起审查运营指标，并审查报告数据。结合组织的长期和短期目标来审视这些报告，以确定是否实现了这些目标。在目标不明确的地方，或在要求的和给予的东西之间可能存在冲突的地方，找出含糊不清的根源。

确定时间、人员和工具可以在哪些方面推动实现运营成果。确定这将影响哪些 KPI 以及成功的目标应该是什么。定期重新审视，确保运营团队有足够的资源来支持业务线。

## 资源

相关文档：

- [Amazon Athena](#)
- [Amazon CloudWatch 指标和维度参考](#)
- [Amazon QuickSight](#)
- [AWS Glue](#)
- [AWS Glue Data Catalog](#)
- [使用 Amazon CloudWatch 代理从 Amazon EC2 实例和本地服务器收集指标和日志](#)
- [使用 Amazon CloudWatch 指标](#)

## OPS 10.您如何应对工作负载事件和运营事件？

制定和验证用于响应事件的程序，以便尽可能减少其对工作负载的干扰。

### 最佳实践

- [OPS10-BP01 使用流程来管理事件、意外事件和问题](#)
- [OPS10-BP02 针对每个提醒设置一个流程](#)
- [OPS10-BP03 根据业务影响确定运营事件的优先顺序](#)
- [OPS10-BP04 定义上报路径](#)
- [OPS10-BP05 定义中断时的客户沟通计划](#)
- [OPS10-BP06 通过控制面板展现状况信息](#)
- [OPS10-BP07 自动响应事件](#)

### OPS10-BP01 使用流程来管理事件、意外事件和问题

贵组织拥有处理事件、意外事件和问题的流程。事件 是在工作负载中发生但可能不需要干预的事情。意外事件 是需要干预的事件。问题 是需要干预或无法解决的反复发生的事件。您需要一些流程来减轻这些事件对业务的影响，并确保做出适当的响应。

当您的工作负载发生意外事件和问题时，您需要一些流程来处理它们。您将如何与利益相关者沟通事件的状态？谁负责监督领导应对工作？您用什么工具来减轻事件的影响？这些是您建立可靠的响应流程所需回答的一些问题的例子。

这些流程必须记录在一个中央位置，并可供参与您工作负载的任何人使用。如果您没有中央 Wiki 或文档存储区，可以使用版本控制存储库。随着流程的发展，您将不断更新这些计划。

接下来将需要对问题进行自动化。这些事情占用了您的时间，限制了您的创新能力。首先构建一个可重复的流程来缓解问题。随着时间的推移，将重点放在自动化缓解或修复根本问题上。这样就可以腾出时间来改进您的工作负载。

期望结果：贵组织拥有处理事件、意外事件和问题的流程。这些流程被记录下来并存储在一个中央位置。它们随着流程的更改而更新。

常见反模式：

- 周末发生了一起意外事件，值班工程师不知道该怎么办。
- 一位客户向您发送一封电子邮件，说应用程序关闭了。您重新启动服务器以修复该问题。这种情况经常发生。
- 有一起意外事件，多个团队独立工作，试图解决该问题。
- 部署发生在您的工作负载中，而不会被记录下来。

建立此最佳实践的好处：

- 您有一条关于工作负载中事件的审计跟踪。
- 从意外事件中恢复的时间缩短了。
- 团队成员能够一致地解决意外事件和问题。
- 调查意外事件时，大家更加团结一致。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

实施这种最佳实践意味着您正在跟踪工作负载事件。您建立了处理意外事件和问题的流程。这些流程被记录下来、共享并经常更新。发现问题，确定优先级，并加以解决。

客户示例

AnyCompany Retail 的内部 Wiki 中有一部分专门用于事件、意外事件和问题管理的流程。所有事件均发送至 [Amazon EventBridge](#)。问题在 [AWS Systems Manager OpsCenter](#) 中被识别为 OpsItems，并按优先级进行修复，减少了无差别的劳动。当流程发生变化时，它们会在内部 Wiki 中进行更新。他们使用 [AWS Systems Manager Incident Manager](#) 来管理意外事件并协调缓解工作。

## 实施步骤

### 1. 事件

- 跟踪工作负载中发生的事件，即使不需要人工干预。
- 与工作负载利益相关者合作，制定一份应跟踪的事件清单。一些示例包括已完成的部署或成功的修补。
- 您可以使用 [Amazon EventBridge](#) 或 [Amazon Simple Notification Service](#) 之类的服务生成自定义事件以进行跟踪。

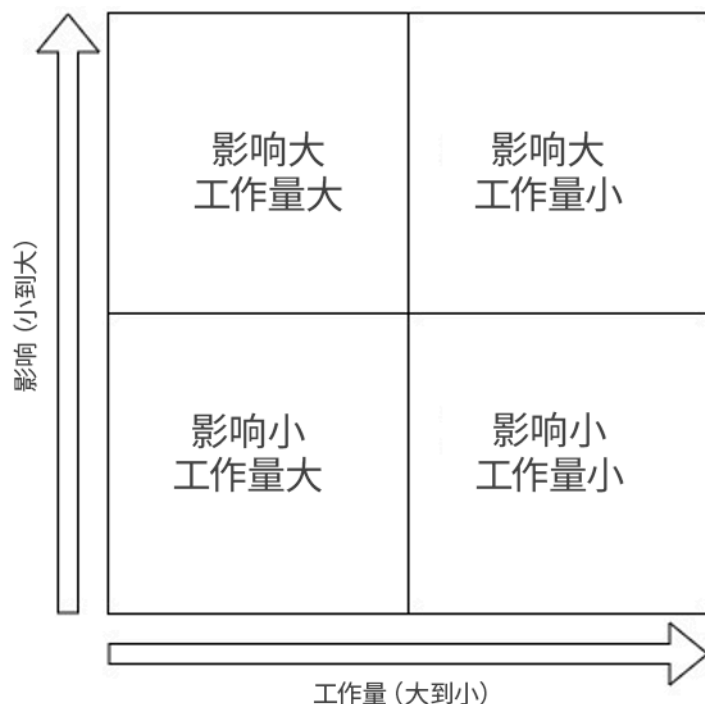
### 2. 意外事件

- 首先要确定意外事件的沟通计划。必须告知哪些利益相关者？您将如何让他们了解情况？谁负责监督协调工作？我们建议建立一个内部聊天渠道进行沟通和协调。
- 为支持您工作负载的团队定义上报路径，特别是在团队没有随时待命的轮换情况下。根据您的支持级别，您还可以向支持提交工单。
- 创建一个调查该意外事件的行动手册。这应该包括沟通计划和详细的调查步骤。在您的调查中包括检查 [AWS Health Dashboard](#)。
- 记录意外事件响应计划。沟通意外事件管理计划，以便内部和外部客户了解参与规则以及对他们的期望。就使用方法对您的团队成员进行培训。
- 客户可以使用 [Incident Manager](#) 来建立和管理他们的意外事件响应计划。
- 企业支持客户可以向他们的技术客户经理请求参加 [意外事件管理研讨会](#)。这场有指导意义的研讨会可测试您现有的意外事件响应计划，并帮助您找出需要改进之处。

### 3. 问题

- 必须在您的 ITSM 系统中识别和跟踪问题。
- 确定所有已知问题，并根据修复工作量和对工作负载的影响来确定它们的优先级。





- 先解决影响大、工作量小的问题。一旦这些问题得到解决，就继续处理那些属于“影响小且工作量小”象限的问题。
- 随着您的工作负载增长和扩展，您可以使用 [Systems Manager OpsCenter](#) 来识别这些问题，为它们附上运行手册，并跟踪它们。

实施计划的工作量级别：中。您需要一个流程和工具来实施这种最佳实践。记录您的流程，让与工作负载相关的任何人都可以使用它们。经常更新它们。您建立了一个管理问题、缓解问题或解决问题的流程。

## 资源

### 相关最佳实践：

- [OPS07-BP03 使用运行手册执行政序](#)：已知问题需要一个相关的运行手册，以使缓解工作保持一致。
- [OPS07-BP04 根据行动手册调查问题](#)：必须使用行动手册对意外事件进行调查。
- [OPS11-BP02 在意外事件发生后执行分析](#)：从意外事件中恢复之后，务必要进行事后分析。

### 相关文档：

- [Atlassian - DevOps 时代的意外事件管理](#)

- [AWS 安全意外事件响应指南](#)
- [DevOps 和 SRE 时代的意外事件管理](#)
- [PagerDuty - 什么是意外事件管理？](#)

相关视频：

- [AWS re:Invent 2020：分布式组织中的意外事件管理](#)
- [AWS re:Invent 2021 - 使用事件驱动型架构构建下一代应用程序](#)
- [AWS 支持您 | 探讨事件管理桌面练习](#)
- [AWS Systems Manager Incident Manager - AWS 虚拟研讨会](#)
- [AWS 后续举措主讲 Incident Manager | AWS 事件](#)

相关示例：

- [AWS 管理和监管工具研讨会 - OpsCenter](#)
- [AWS 主动式服务 – 意外事件管理研讨会](#)
- [使用 Amazon EventBridge 构建事件驱动型应用程序](#)
- [在 AWS 上构建事件驱动型架构](#)

相关服务：

- [Amazon EventBridge](#)
- [Amazon SNS](#)
- [AWS Health Dashboard](#)
- [AWS Systems Manager Incident Manager](#)
- [AWS Systems Manager OpsCenter](#)

OPS10-BP02 针对每个提醒设置一个流程

针对引发提醒的任何事件制定明确的响应措施（运维手册或管理手册），并明确指定负责人。这样可以确保您及时有效地响应运营事件，并防止可以针对其采取措施的事件被不重要的通知所掩盖。

常见反模式：

- 监控系统会向您显示已批准的连接流和其他消息。由于消息量过大，导致您错过了需要干预的周期性错误消息。
- 您收到提醒，指示网站停机。发生这种情况时，没有明确的流程。您被迫采用临时方法来诊断和解决问题。边处理边开发流程会延长恢复时间。

建立此最佳实践的好处：仅在需要采取措施时发出提醒可以防止低价值提醒遮掩高价值提醒。制定一个可随时采取措施的提醒流程，可以对环境中的事件做出一致而迅速的响应。

未建立此最佳实践暴露的风险等级：高

### 实施指导

- 提醒响应流程：对于引发提醒的任何事件，都要制定明确的响应措施（运行手册或管理手册），并明确指定负责其成功完成的负责人（例如个人、团队或角色）。响应的执行可能是自动的，也可能由其他团队完成，但是负责人应负责确保响应流程获得预期的成果。设置这些流程可以确保您及时有效地响应运营事件，并防止可以针对其采取措施的事件被不重要的通知所掩盖。例如，可以实施自动扩展来扩展 Web 前端，但是运营团队应负责确保自动扩展规则和限制符合工作负载需求。

### 资源

相关文档：

- [Amazon CloudWatch 功能](#)
- [什么是 Amazon CloudWatch Events？](#)

相关视频：

- [制定监控计划](#)

### OPS10-BP03 根据业务影响确定运营事件的优先顺序

确保在多个事件需要干预时，优先处理对业务最为重要的事件。人身伤亡、经济损失、名誉或信任损害都是一种影响。

常见反模式：

- 您收到一个支持请求，需为用户添加打印机配置。在处理该问题时，您收到一个支持请求，称您的零售网站停机。为您的用户完成打印机配置后，您着手处理网站问题。

- 您收到通知，指示您的零售网站和薪资系统都发生停机。您不知道应该优先处理哪个问题。

建立此最佳实践的好处： 优先响应对业务影响最大的意外事件，可以帮助您管理这种影响。

未建立此最佳实践暴露的风险等级： 中

#### 实施指导

- 根据业务影响确定运营事件的优先顺序：确保在多个事件需要干预时，优先处理对业务最为重要的事件。影响可能包括人身伤亡、经济损失、违规、名誉或信任损害。

#### OPS10-BP04 定义上报路径

在运维手册和管理手册中定义上报路径，包括触发上报的事件和上报程序。明确指定每项措施的负责人，以便确保有效而及时地响应运营事件。

在采取措施之前，确定何时需要人为决定。与决策者合作，提前做出决策，这样 MTTR 便不会因为等待响应而延长。

常见反模式：

- 您的零售网站停机。您不了解用于网站恢复的运行手册。您开始打电话求助同事。
- 您收到一个关于应用程序无法访问的支持案例。您没有系统管理权限。您不知道谁具有权限。您尝试与创建案例的系统负责人联系，但没有得到响应。您无法联系到系统负责人，而您的同事对此也不太熟悉。

建立此最佳实践的好处： 通过定义上报、上报触发器和上报程序，您可以适当的影响速率系统地向意外事件添加资源。

未建立这种最佳实践的情况下暴露的风险等级： 中

#### 实施指导

- 定义上报路径：在运维手册和管理手册中定义上报路径，包括触发升级的事件和升级程序。例如，当运维手册无法解决问题或者预定义的时间已经过去时，将问题从支持工程师升级给高级支持工程师。当管理手册无法确定修复路径或者预定义的时间已经过去时，将问题从高级工程师升级给开发团队也是一种正确的升级路径。明确指定每项措施的负责人，以便确保有效而及时地响应运营事件。升级可以涉及第三方。例如某个网络连接提供商或软件供应商。升级可以涉及负责受影响的系统并且获得授权的决策者。

## OPS10-BP05 定义中断时的客户沟通计划

定义和测试系统中断时的沟通计划，您可以依赖此计划在中断期间让客户和利益攸关方了解情况。当用户使用的服务受到影响和服务恢复正常时直接传达给用户。

期望结果：

- 为从计划维护到重大意外故障的各种状况（包括调用灾难恢复计划）制定沟通计划。
- 在沟通中，提供有关系统问题的清晰透明信息，帮助客户避免事后猜测其系统的性能。
- 使用自定义错误消息和状态页面减少服务台请求的激增并让用户了解相关情况。
- 定期测试沟通计划，以便确认在真正发生中断时计划会按预期执行。

常见反模式：

- 发生工作负载中断，但您没有沟通计划。因为用户不知晓关于中断的信息，他们不断发出请求，使您的故障单系统不堪重负。
- 您在中断期间向用户发送电子邮件通知。邮件中未包含恢复服务的时间表，因此用户无法针对中断制定计划。
- 您有针对中断的沟通计划，但从未经过测试。发生中断，但因为漏掉了一个本来可以在测试中发现的关键步骤，沟通计划失败。
- 在中断期间，您向用户发送通知，其中包含太多受 AWS 保密协议保护的技术细节和信息。

建立此最佳实践的好处：

- 在中断期间保持沟通可以确保客户可以了解有关问题的进度和预估的解决时间。
- 制定明确的沟通计划，确认您的客户和最终用户可以充分了解情况，以便他们可以采取必要的额外步骤来缓解中断的影响。
- 通过适当的沟通并提高对计划内和计划外中断的认识，您可以提高客户满意度、限制非预期的反应和提高客户保留率。
- 及时和透明的系统中断沟通可以树立信心和建立信任，从而维护您和客户之间的关系。
- 在中断或危机期间，行之有效的沟通策略可以减少猜测和流言蜚语，以免阻碍您恢复运营。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

在中断期间让您的客户了解情况的沟通计划是全面的，涵盖多个接口，包括面向客户的错误页面、自定义 API 错误消息、系统状态横幅和运行状况状态页面。如果您的系统包括注册用户，您可以通过各种消息传递渠道（例如，电子邮件、短信或推送消息）进行沟通，向客户发送个性化的消息内容。

### 客户沟通工具

作为第一道防线，Web 和移动应用程序应在中断期间提供友好且信息丰富的错误消息，并能够将流量重定向到状态页面。[Amazon CloudFront](#) 是完全托管式内容分发网络（CDN），包括定义和提供自定义错误内容的功能。CloudFront 中的自定义错误页面是很好的第一层客户信息传递，适合组件级中断。CloudFront 还可以简化状态页的管理与激活，以便在计划内或计划外中断期间拦截所有请求。

当中断隔离到不相关联的服务时，自定义 API 错误消息可以帮助检测和减少影响。您可以使用 [Amazon API Gateway](#) 为 REST API 配置自定义响应。当 API Gateway 无法访问后端服务时，您可以利用它向 API 使用者提供清晰而有意义的信息。客户消息还可以用于支持中断横幅内容，以及在特定系统功能因服务层中断而降级时发出通知。

直接消息传递是最个性化的客户消息传递类型。[Amazon Pinpoint](#) 是适合可扩展多渠道通信的托管服务。您可以利用 Amazon Pinpoint 来构建活动，通过短信、电子邮件、语音、推送通知或您定义的自定义渠道在受影响的客户群中广泛广播消息。当您使用 Amazon Pinpoint 管理消息传递时，消息活动定义明确、可测试，并且可以智能地应用于目标客户群体。建立活动之后，就可以对其作出安排，或通过事件来触发，然后就可以很容易地进行测试。

### 客户示例

当工作负载受影响时，AnyCompany Retail 向其用户发送电子邮件通知。电子邮件描述了哪些业务功能受到影响，并对何时恢复服务作出合理的预估。此外，他们有一个状态页，其中显示了有关其工作负载运行状况的实时信息。每年在开发环境中对沟通计划测试两次，以便确认计划有效。

### 实施步骤

1. 确定消息传递策略的沟通渠道。考虑应用程序的架构方面，并确定向客户提供反馈的最佳策略。这包括概述的一个或多个指导策略，包括错误和状态页、自定义 API 错误响应或直接消息传递。
2. 设计应用程序的状态页面。如果您已确定状态或自定义错误页面适合您的客户，则需要为这些页面设计内容和信息。错误页面向用户说明为什么应用程序不可用、什么时候可以再次使用以及在此期间他们可以做什么。如果应用程序使用 Amazon CloudFront，您可以提供[自定义错误响应](#)或使用 Lambda at Edge 来[转变错误](#)并重写页面内容。CloudFront 还可以将目的地从应用程序内容转换为包含维护或中断状态页面的静态 [Amazon S3](#) 内容来源。

3. 为服务设计一组正确的 API 错误状态。当 API Gateway 无法访问后端服务时生成的错误消息以及服务层异常可能不包含适合向最终用户显示的友好消息。您无需对后端服务进行代码更改，而可以将 API Gateway [自定义错误响应](#)配置为将 HTTP 响应代码映射到精选的 API 错误消息。
4. 从业务角度设计信息，使其与系统的最终用户相关，并且不包含技术细节。考虑您的受众并调整信息。例如，您可以引导内部用户转向利用替代系统的解决方法或手动流程。外部用户可能需要等到系统恢复，或订阅更新，以便在系统恢复后收到通知。定义多个场景的已批准消息传递，包括意外中断、计划维护以及会导致特定功能可能降级或不可用的部分系统故障。
5. 使您的客户消息传递模板化和自动化。确立消息内容之后，您可以使用 [Amazon Pinpoint](#) 或其他工具使消息传递活动实现自动化。借助 Amazon Pinpoint，您可以为特定的受影响用户创建客户目标细分，并将消息转换为模板。查看 [Amazon Pinpoint 教程](#)，了解如何设置消息传递活动。
6. 避免将消息传递功能与面向客户的系统紧密耦合。消息传递策略不应该严格依赖于系统数据存储或服务，以便确认在遇到中断时您可以成功发送消息。考虑培养从[多个可用区或区域](#)发送消息以实现消息传递可用性的能力。如果您使用 AWS 服务来发送消息，请利用数据面板操作而不是[控制面板操作](#)来调用消息传递。

实施计划的工作量级别：高。制定沟通计划和用于发送计划的机制需要付出巨大的努力。

## 资源

### 相关最佳实践：

- [OPS07-BP03 使用运行手册执行程序](#) - 您的沟通计划应具有与之关联的运行手册，这样您的员工就知道如何应对。
- [OPS11-BP02 在意外事件发生后执行分析](#) - 发生中断之后，执行事后分析以确定可防止再次中断的机制。

### 相关文档：

- [Amazon API Gateway 和 AWS Lambda 中的错误处理模式](#)
- [Amazon API Gateway 响应](#)

### 相关示例：

- [AWS Health 控制面板](#)
- [弗吉尼亚州北部 \( US-EAST-1 \) 区域中的 AWS 服务事件的摘要](#)



## 相关服务：

- [AWS 支持](#)
- [AWS 客户协议](#)
- [Amazon CloudFront](#)
- [Amazon API Gateway](#)
- [Amazon Pinpoint](#)
- [Amazon S3](#)

## OPS10-BP06 通过控制面板展现状况信息

提供为目标受众（例如内部技术团队、领导和客户）专门设计的控制面板，以传达业务当前的运营状况并提供值得关注的指标。

您可以使用 [Amazon CloudWatch 控制面板](#)，在 CloudWatch 控制台中可自定义的主页上创建控制面板。借助像 [QuickSight](#) 这样的商业智能服务，您可以创建和发布工作负载和运营状况（例如，订单达成率、连接的用户和交易时间）的交互式控制面板。您可以创建控制面板，用来显示指标的系统级和业务级视图。

## 常见反模式：

- 您根据请求运行有关管理应用程序当前使用情况的报告。
- 在意外事件发生期间，相关系统负责人每二十分钟与您联系一次，想要知道问题是否已解决。

建立此最佳实践的好处：创建控制面板后，您可以让客户自助访问信息，这样可让他们获知相关信息，并确定是否需要采取措施。

未建立此最佳实践暴露的风险等级：中

## 实施指导

- 通过控制面板展现状况信息：提供为目标受众（例如内部技术团队、领导和客户）专门设计的控制面板，以传达业务当前的运营状况并提供值得关注的指标。提供用于获取状态信息的自助选项，可以减少负责处理状态请求的运营团队的中断。示例包括 Amazon CloudWatch 控制面板和 AWS Health Dashboard。
  - [CloudWatch 控制面板创建并使用自定义指标视图](#)

## 资源

相关文档：

- [QuickSight](#)
- [CloudWatch 控制面板创建并使用自定义指标视图](#)

## OPS10-BP07 自动响应事件

自动响应事件以便减少由手动流程引起的错误，并确保响应及时并且一致。

有多种方法可以在 AWS 上自动执行运行手册和行动手册操作。要响应 AWS 资源中的状态更改事件或您自己的自定义事件，您应创建 [CloudWatch Events 规则](#) 以通过 CloudWatch 目标（例如，Lambda 函数、Amazon Simple Notification Service ( Amazon SNS ) 主题、Amazon ECS 任务和 AWS Systems Manager Automation ) 触发响应。

要响应超过资源阈值的指标（例如，等待时间），您应创建 [CloudWatch 警报](#) 以使用 Amazon EC2 操作或 Auto Scaling 操作执行一个或多个操作，或者向 Amazon SNS 主题发送通知。如果您需要执行自定义操作以响应警报，请通过 Amazon SNS 通知调用 Lambda。使用 Amazon SNS 发布事件通知和升级消息，以便让人们了解情况。

AWS 还通过 AWS 服务 API 和 SDK 支持第三方系统。AWS 合作伙伴和第三方提供了许多用于监控、通知和响应的监控工具。其中一些工具包括 New Relic、Splunk、Loggly、SumoLogic 和 Datadog。

您应该保留关键的手动程序，以备在自动程序出故障时使用。

常见反模式：

- 开发人员检查其代码。发生此事件后，本可开始构建然后执行测试，但您没执行任何操作。
- 在停止运行前，您的应用程序记录了一个特定的错误。重新启动应用程序的流程易于理解，可编写成脚本。您可以使用日志事件来调用脚本并重新启动应用程序。否则的话，如果错误发生在星期天凌晨 3 点，您作为负责修复系统的随叫随到的资源，将不得不起床去处理。

建立此最佳实践的好处：通过自动响应事件，您可以缩短响应时间并减少人工活动中发生的错误。

未建立此最佳实践暴露的风险等级：低

## 实施指导

- 自动响应事件：自动响应事件以便减少由手动流程引起的错误，并确保响应及时并且一致。

- [什么是 Amazon CloudWatch Events ?](#)
- [创建在发生事件时触发的 CloudWatch Events 规则](#)
- [创建在 AWS API 调用上使用 AWS CloudTrail 触发的 CloudWatch Events 规则](#)
- [来自支持的服务的 CloudWatch Events 事件示例](#)

## 资源

### 相关文档：

- [Amazon CloudWatch 功能](#)
- [来自支持的服务的 CloudWatch Events 事件示例](#)
- [创建在 AWS API 调用上使用 AWS CloudTrail 触发的 CloudWatch Events 规则](#)
- [创建在发生事件时触发的 CloudWatch Events 规则](#)
- [什么是 Amazon CloudWatch Events ?](#)

### 相关视频：

- [制定监控计划](#)

### 相关示例：

## 演进

### 问题

- [OPS 11.如何改进运营？](#)

## OPS 11.如何改进运营？

分配专门的时间和资源用于近乎持续的增量改进，以便提高运营的有效性和效率。

### 最佳实践

- [OPS11-BP01 设置持续改进流程](#)
- [OPS11-BP02 在意外事件发生后执行分析](#)
- [OPS11-BP03 实施反馈环路](#)

- [OPS11-BP04 执行知识管理](#)
- [OPS11-BP05 确定推动改进的因素](#)
- [OPS11-BP06 验证分析结果](#)
- [OPS11-BP07 审核运营指标](#)
- [OPS11-BP08 记录和分享经验教训](#)
- [OPS11-BP09 分配时间进行改进](#)

## OPS11-BP01 设置持续改进流程

根据内部和外部架构最佳实践评估您的工作负载。至少每年执行一次工作负载审核。将改进机会优先纳入您的软件开发周期。

期望结果：

- 至少每年都根据架构最佳实践来分析工作负载。
- 在软件开发过程中，为改进机会赋予同等的优先级。

常见反模式：

- 自从几年前部署工作负载以来，您没有对其进行过架构审查。
- 改进机会的优先级较低，并停留在待办事项列表中。
- 不存在对组织的最佳实践实施修改的标准。

建立此最佳实践的好处：

- 您的工作负载符合最新的架构最佳实践。
- 深思熟虑地推进工作负载的演变。
- 您可以利用组织的最佳实践来改善所有工作负载。

在未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

至少每年一次，对工作负载进行架构审查。利用内部和外部最佳实践，评估您的工作负载并确定改进机会。将改进机会优先纳入您的软件开发周期。

## 客户示例

AnyCompany Retail 的所有工作负载都要经过每年一次的架构审查过程。他们开发了自己的最佳实践清单，适用于所有工作负载。使用 AWS Well-Architected Tool 的自定义剖析功能，他们通过该工具和最佳实践自定义剖析进行审查。通过审查发现的改进机会在他们的软件冲刺中被优先考虑。

## 实施步骤

1. 至少每年一次，对您的生产工作负载进行定期架构审查。使用记录在册的架构标准，包括 AWS 特定的最佳实践。
  - a. 建议您使用您自己内部定义的标准来进行这些审查。如果没有内部标准，建议您使用 AWS Well-Architected Framework。
  - b. 您可以使用 AWS Well-Architected Tool 来创建自己的内部最佳实践的自定义剖析，并进行架构审查。
  - c. 客户可以联系他们的 AWS 解决方案架构师，对他们的工作负载进行一次指导式 Well-Architected Framework 审查。
2. 将审查中发现的改进机会优先纳入您的软件开发过程。

实施计划的工作量级别：低。可以使用 AWS Well-Architected Framework 执行年度架构审核。

## 资源

### 相关最佳实践：

- [OPS11-BP02 在意外事件发生后执行分析](#) - 事件后分析是改进项目的另一个来源。将学到的经验教训纳入您自己的内部架构最佳实践清单。
- [OPS11-BP08 记录和分享经验教训](#) - 在开发自己的架构最佳实践时，在组织内分享这些最佳实践。

### 相关文档：

- [AWS Well-Architected Tool – 自定义剖析](#)
- [AWS Well-Architected 白皮书 - 审核流程](#)
- [使用自定义剖析和 AWS Well-Architected Tool 定制 Well-Architected 审查](#)
- [在组织中实施 AWS Well-Architected 自定义剖析生命周期](#)

### 相关视频：

- [Well-Architected 实验室 - 级别 100 : AWS Well-Architected Tool 自定义剖析](#)

相关示例：

- [AWS Well-Architected Tool](#)

## OPS11-BP02 在意外事件发生后执行分析

审核影响客户的事件，确定导致这些事件的因素和预防措施。利用这些信息来制定缓解措施，以限制或防止再次发生同类事件。制定程序以迅速有效地做出响应。根据目标受众，适当传达事件成因和纠正措施。

常见反模式：

- 您管理应用程序服务器。大约每 23 小时 55 分钟，所有活动会话都会终止。您已尝试找出应用程序服务器上出现的问题。您怀疑可能是网络问题，但由于网络团队工作繁忙无法为您提供支持，因此无法与他们合作。由于缺乏可遵循的预定义流程，因此难以获取支持并收集必要的信息来确定发生了什么情况。
- 您的工作负载中出现了数据丢失的情况。这是第一次发生，原因不明。您认为它不重要，因为可以重新创建数据。数据丢失对客户的影响开始变得愈发频繁。还原丢失的数据时，这也会增加您的操作负担。

建立此最佳实践的好处：设置预定义的流程，以确定导致意外事件发生的要素、条件、操作和事件，从而帮助您找到改进机会。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

- 通过流程来确定事件成因：审查所有影响客户的意外事件。设置流程来确定和记录导致意外事件的因素，以便制定缓解措施来限制或防止事件再次发生，并且您还可以据此制定及时有效的应对措施。在适当的情况下向目标受众说明根本原因。

## OPS11-BP03 实施反馈环路

反馈环路提供了可操作的见解，进而推动决策的制定。将反馈环路融入过程和工作负载中。这可帮助您确定问题和需要改进的领域。它们还可以验证在改进方面所做的投入。这些反馈环路为持续改进工作负载奠定了基础。

反馈环路分为两大类：即时反馈 和 回顾性分析。通过审查运营活动的绩效和成果来收集即时反馈。此反馈来自团队成员、客户或活动的自动化输出。通过 A/B 测试和发布新功能等方式接收即时反馈，这对于快速失效机制至关重要。

定期执行回顾性分析，可以获得在运营成果审核和指标审核过程中产生的反馈。这些回顾在冲刺结束时进行、有节奏地进行或者在重大发布或事件之后进行。这种类型的反馈环路验证了在运营或工作负载方面的投入。它有助于衡量成功并验证您的策略。

期望的结果：您可以使用即时反馈和回顾性分析来加快改进。有一种机制可用于捕获用户和团队成员的反馈。回顾性分析用于确定可推动改进的趋势。

常见反模式：

- 您推出了一项新功能，但无法接收客户对此新功能的反馈。
- 在投资进行运营改进后，您无需回顾来验证它们。
- 您可以收集客户反馈，但不用定期进行审查。
- 反馈环路会产生建议的操作项，但它们不包括在软件开发过程中。
- 对于所提出的改进事项，客户不会收到关于它们的反馈意见。

建立此最佳实践的好处：

- 您可以反过来从客户出发，以便推动新的功能。
- 您的组织文化能够更快地对变更做出回应。
- 趋势用于确定改进机会。
- 回顾将验证对工作负载和运营所做的投入。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

实施此最佳实践意味着同时使用即时反馈和回顾性分析。这些反馈环路将推动改进。有许多适用于即时反馈的机制，包括调查、客户投票或反馈表。您的组织还使用回顾来确定改进机会并验证计划。

客户示例

AnyCompany Retail 创建了一个 Web 表单，客户可使用此表单提供反馈或报告问题。在每周 Scrum 期间，软件开发团队将评估用户反馈。反馈定期用于引导相应平台的发展。他们在每个冲刺结束时进行回顾，确定需要改进的项目。



## 实施步骤

### 1. 即时反馈

- 您需要一种机制来接收由客户和团队成员提供的反馈，也可以将您的运营活动配置为交付自动反馈。
- 您的组织需要一个流程，来审查此反馈、确定要改进的方面并安排改进。
- 必须将反馈纳入您的软件开发过程中。
- 在实施改进时，请对反馈提交者进行跟进。
- 您可以使用 [AWS Systems Manager OpsCenter](#) 将这些改进创建为 [OpsItem](#) 并进行跟踪。

### 2. 回顾性分析

- 在开发周期结束时、按设定的节奏或在主要发布后进行回顾。
- 召开回顾性会议，让工作负载中涉及的利益相关者参加。
- 在白板或电子表格上创建三个列：“停止”、“开始”和“继续”。
  - 停止 针对的是您希望团队停止执行的任何工作。
  - 开始 针对的是要开始付诸行动的想法。
  - 继续 针对的是要继续执行的项目。
- 在会议室里四处走动，从利益相关者那里收集反馈。
- 确定反馈的优先级。将操作和利益相关者分配给任何“开始”或“继续”项目。
- 将操作纳入软件开发过程中，并在实施改进时将状态更新传达给利益相关者。

实施计划的工作量级别：中。要实施此最佳实践，您需要一种方法来获取并分析即时反馈。此外，您需要建立一个回顾性分析过程。

## 资源

### 相关最佳实践：

- [OPS01-BP01 评估外部客户需求](#)：反馈环路是一种用于收集外部客户需求的机制。
- [OPS01-BP02 评估内部客户需求](#)：内部利益相关者可以使用反馈环路来传达需求和要求。
- [OPS11-BP02 在意外事件发生后执行分析](#)：事件后分析是发生事件后进行回顾性分析的重要形式。
- [OPS11-BP07 审核运营指标](#)：运营指标审查可确定趋势和需要改进的方面。

## 相关文档：

- [构建 CCOE 时应避免的 7 个陷阱](#)
- [Atlassian 团队行动手册 – 回顾](#)
- [电子邮件定义：反馈环路](#)
- [建立基于 AWS Well-Architected Framework 审查的反馈环路](#)
- [IBM Garage 方法 – 保持回顾](#)
- [Investopedia – PDCA 循环](#)
- [Tim Cochran 所著的《最大限度地提高开发人员效率》](#)
- [运营准备情况审查 \( ORR \) 白皮书 – 迭代](#)
- [TIL CSI – 持续服务改进](#)
- [当丰田转向电子商务：Amazon 的精益方法](#)

相关视频：

- [构建有效的客户反馈环路](#)

相关示例：

- [Astuto – 客户反馈开源工具](#)
- [AWS 解决方案 – AWS 上的 QnABot](#)
- [Fider – 客户反馈整理平台](#)

相关服务：

- [AWS Systems Manager OpsCenter](#)

## OPS11-BP04 执行知识管理

知识管理帮助团队成员找到完成他们的工作所需的信息。在学习型组织中，自由分享信息，从而增强个人的能力。可以发现和搜索信息。信息准确且保持最新。制定有创建新信息、更新现有信息和归档过时信息的机制。知识管理平台的最常见例子是像 Wiki 这样的内容管理系统。

期望结果：

- 团队成员可以及时获取准确的信息。
- 信息可搜索。

- 制定有添加、更新和归档信息的机制。

常见反模式：

- 没有集中式知识存储。团队成员在他们的本地计算机上管理自己的笔记。
- 您有自托管的 Wiki，但没有制定机制来管理信息，导致信息过时。
- 有人识别出缺失的信息，但没有制定流程来请求将其添加到团队 Wiki 中。他们自己添加信息，但他们错过了一个关键步骤，导致发生中断。

建立此最佳实践的好处：

- 因为可以自由分享信息，所以增强了团队成员的能力。
- 因为文档保持最新且可搜索，新团队成员可以更快上手。
- 信息及时、准确和富有实用价值。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

知识管理是学习型组织的一个重要方面。首先，您需要一个中央存储库来存储您的知识（一个常见的例子是自托管 Wiki）。您必须制定用于添加、更新和归档知识的流程。为应该记录的内容制定标准，并让每一个人都能作出贡献。

## 客户示例

AnyCompany Retail 托管一个内部 Wiki，其中存储了他们的所有知识。公司鼓励团队成员在履行日常职责时向知识库中添加内容。跨职能团队每季度评估哪些页面更新最少，并确定这些页面是否需要归档或更新。

## 实施步骤

1. 首先确定用于存储知识的内容管理系统。获得整个组织的利益攸关方的同意。
  - a. 如果您还没有内容管理系统，则在刚开始的时候请考虑运行自托管的 Wiki，或使用版本控制存储库。
2. 编制用于添加、更新和归档信息的运行手册。就这些流程对团队进行培训。
3. 确定应该在内容管理系统中存储哪些知识。从团队成员执行的日常活动（运行手册和行动手册）开始。与利益攸关方一起对增加的知识进行优先级排序。

#### 4. 定期与利益攸关方一起识别过时的信息并将其归档或更新。

实施计划的工作量级别：中等。如果您还没有内容管理系统，则可以设置自托管的 Wiki 或版本控制文档库。

#### 资源

#### 相关最佳实践：

- [OPS11-BP08 记录和分享经验教训](#) - 知识管理促进有关经验教训的信息共享。

#### 相关文档：

- [Atlassian - 知识管理](#)

#### 相关示例：

- [DokuWiki](#)
- [Gollum](#)
- [MediaWiki](#)
- [Wiki.js](#)

### OPS11-BP05 确定推动改进的因素

确定推动改进的因素，以便评估各种机会并确定其优先顺序。

在 AWS 上，您可以聚合所有运营活动、工作负载和基础设施的日志，以创建详细的活动历史记录。然后，您可以根据推动因素，使用 AWS 工具分析您在一段时间内的运营状况和工作负载运行状况（例如，确定趋势、将事件和活动与结果相关联，并在环境之间/跨系统进行比较和对比），以发现改进机会。

您应该使用 CloudTrail 跟踪 API 活动（通过 AWS Management Console、CLI、开发工具包和 API），以了解您账户中发生的情况。您可以使用 CloudTrail 和 CloudWatch 跟踪您的 AWS 开发人员工具部署活动。这将向您的 CloudWatch Logs 日志数据添加部署的详细活动历史记录及其结果。

[将您的日志数据导出到 Amazon S3](#) 以便长期存储。使用 [AWS Glue](#)，您可以在 Amazon S3 中发现并准备您的日志数据以供分析。使用 [Amazon Athena](#)，借助其与 AWS Glue 的原生集成来分析您的日志数据。使用像 [QuickSight](#) 这样的商业智能工具，您可以直观显示、浏览和分析您的数据

## 常见反模式：

- 您有一个脚本，可以正常运行但不完美。您投入时间重新编写。现在，它完美无缺。
- 您的初创公司正试图从风险投资人那里获得另一笔资金。他们希望您证明符合 PCI DSS。您希望让他们满意，因此您记录合规性，但却错过了客户的交付日期，导致客户流失。您没有做错，但是现在您怀疑这样做是否合适。

建立此最佳实践的好处：确定希望用于改进的标准后，您可以最大程度地减小基于事件的动机或情感投入所带来的影响。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

- 了解推动改进的因素：您只应该在能够实现所需成果的情况下更改某个系统。
  - 需要的功能：在评估改进机会时评估需要的特性和功能。
    - [AWS 的新增功能](#)
  - 无法接受的问题：在评估改进机会时评估无法接受的问题、错误和漏洞。
    - [AWS 最新安全公告](#)
    - [AWS Trusted Advisor](#)
  - 合规性要求：在分析改进机会时，评估保持监管和政策合规性或获取第三方支持所需的更新和更改。
    - [AWS 合规性](#)
    - [AWS 合规性计划](#)
    - [AWS 合规性最新新闻](#)

## 资源

### 相关文档：

- [Amazon Athena](#)
- [QuickSight](#)
- [AWS 合规性](#)
- [AWS 合规性最新新闻](#)
- [AWS 合规性计划](#)

- [AWS Glue](#)
- [AWS 最新安全公告](#)
- [AWS Trusted Advisor](#)
- [将您的日志数据导出到 Amazon S3](#)
- [AWS 的新增功能](#)

## OPS11-BP06 验证分析结果

与跨职能团队和业务负责人共同查看分析结果和响应措施。通过这些工作来建立共识、发现其他影响并确定行动方案。适当调整响应措施。

常见反模式：

- 您看到某个系统上的 CPU 利用率达到 95%，因此优先考虑寻找一种方法来减少系统上的负载。您认为最佳行动方案是进行扩展。系统是一个转码器，您可对它进行扩展，让它始终以 95% 的 CPU 利用率运行。如果您先与系统负责人联系，他们会向您做出解释。您浪费了时间。
- 系统负责人坚持认为他们的系统是关键任务型系统。系统未置于高安全性环境中。为了提高安全性，您需要对关键任务型系统实施额外的检测和预防控制措施。您通知系统负责人工作已完成，并向他收取额外资源费用。在收到此通知后的沟通过程中，系统负责人了解到对于关键任务型系统有一个正式定义，而他的系统并不满足。

建立此最佳实践的好处：通过与业务负责人和主题专家一起验证分析结果，您可以建立共识并更有效地指导改进。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

- 验证分析结果：与业务负责人和主题专家沟通，以确保对您收集的数据的价值达成共识和一致。确定其他问题、潜在影响并制定行动方案。

## OPS11-BP07 审核运营指标

定期与来自不同业务领域的跨团队参与者对运营指标进行回顾性分析。通过这些分析来确定改进机会和可能的行动方案，并分享经验教训。

寻找在所有环境（例如，开发、测试和生产环境）中改进的机会。

## 常见反模式：

- 维护时段导致一次重要的零售促销中断。如果存在其他影响业务的事件，可以延迟标准维护时段，而业务部门对此并不知晓。
- 由于使用了组织中常用的错误库，导致了长时间的停机。自此之后，您已经迁移到可靠的库。您组织中的其他团队尚未意识到风险的存在。如果您定期开会并审核此意外事件，他们应该注意到这种风险。
- 转码器的性能一直在不断下降，这对媒体团队产生了影响。但这还不算多严重。真正糟糕的是，除非情况严重到足以引发意外事件，否则您将难以发现。如果您与媒体团队一起审核运营指标，就有机会发现指标的变化，同时认识到他们的经验并利用这些经验将问题解决。
- 您没有审核对客户 SLA 的满足程度。您目前正趋向于无法满足客户 SLA。如果无法满足客户 SLA，将会受到经济处罚。如果您定期开会审核这些 SLA 的指标，您将有机会发现并解决这一问题。

建立此最佳实践的好处：您可以通过会议定期审核运营指标、事件和意外事件，在团队之间保持共识、分享经验教训，以及确定改进的优先级和目标。

未建立此最佳实践暴露的风险等级：中

## 实施指导

- 审核运营指标：定期与来自不同业务领域的跨团队参与者对运营指标进行回顾性分析。与包括业务、开发和运营团队在内的利益相关方共同分析通过即时反馈和回顾性分析得到的发现，并分享经验教训。根据他们的见解来确定改进机会和可能的行动方案。
  - [Amazon CloudWatch](#)
  - [使用 Amazon CloudWatch 指标](#)
  - [发布自定义指标](#)
  - [Amazon CloudWatch 指标和维度参考](#)

## 资源

### 相关文档：

- [Amazon CloudWatch](#)
- [Amazon CloudWatch 指标和维度参考](#)
- [发布自定义指标](#)
- [使用 Amazon CloudWatch 指标](#)



## OPS11-BP08 记录和分享经验教训

记录和分享在运营活动中获得的经验教训，以便在内部和不同团队中利用。

您应该分享团队学到的经验教训，以增加整个组织的效益。您需要分享信息和资源，以防止出现可避免的错误并简化开发工作。这让您可以专注于交付所需的功能。

使用 AWS Identity and Access Management (IAM) 定义权限，以允许对您要在账户内和账户之间共享的资源进行受控访问。然后，您应该使用版本受控的 AWS CodeCommit 存储库来分享应用程序库、脚本程序、程序文档和其他系统文档。您可以共享对 AMI 的访问权限并授权跨账户使用 Lambda 函数，以此来分享您的计算标准。您还应将您的基础设施标准共享为 AWS CloudFormation 模板。

通过 AWS API 和 SDK，您可以集成外部和第三方工具和存储库（例如，GitHub、BitBucket 和 SourceForge）。在分享您学到的和开发的内容时，请注意设定权限以确保共享存储库的完整性。

常见反模式：

- 由于使用了组织中常用的错误库，导致了长时间的停机。自此之后，您已经迁移到可靠的库。您组织中的其他团队尚未意识到风险的存在。如果您记录并分享对于此库的经验，他们就可能会注意到风险。
- 您已经确定了内部共享微服务中导致会话中断的边缘案例。为了避免这一边缘案例的出现，您更新了对服务的调用。您组织中的其他团队尚未意识到风险的存在。如果您记录并分享对于此库的经验，他们就可能会注意到风险。
- 您已找到一种方法，可以显著降低其中一个微服务的 CPU 利用率要求。您不知道其他团队是否可以利用这种技术。如果您记录并分享对于此库的经验，他们将有机会加以利用。

建立此最佳实践的好处：分享经验教训可以为改进提供支持，并最大程度地从经验中获益。

未建立这种最佳实践的情况下暴露的风险等级：低

实施指导

- 记录和分享经验教训：设置程序来记录在运营活动执行和回顾性分析过程中获得的经验教训，供其他团队利用。
- 分享经验教训：设置程序在不同团队中分享经验教训和相关构件。例如，通过可以访问的 Wiki 共享更新后的程序、指南、管理机制和最佳实践。通过公共存储库共享脚本、代码和库。
  - [授权访问 AWS 环境](#)
  - [共享 AWS CodeCommit 存储库](#)

- [AWS Lambda 函数的简单授权](#)
- [将 AMI 与特定 AWS 账户共享](#)
- [利用 AWS CloudFormation Designer URL 快速共享模板](#)
- [将 AWS Lambda 与 Amazon SNS 配合使用](#)

## 资源

### 相关文档：

- [AWS Lambda 函数的简单授权](#)
- [共享 AWS CodeCommit 存储库](#)
- [将 AMI 与特定 AWS 账户共享](#)
- [利用 AWS CloudFormation Designer URL 快速共享模板](#)
- [将 AWS Lambda 与 Amazon SNS 配合使用](#)

### 相关视频：

- [授权访问 AWS 环境](#)

## OPS11-BP09 分配时间进行改进

流程中专用的时间和资源可以实现持续增量改进。

在 AWS 上，您可以创建临时的环境副本，从而降低试验和测试的风险、工作量及成本。这些重复的环境可用于测试分析、试验、开发和测试计划改进时所得出的结论。

### 常见反模式：

- 您的应用程序服务器中存在一个已知性能问题。它被添加到每个计划内功能实施之后的待办事项中。如果计划功能的添加速率保持不变，那么性能问题将永远无法解决。
- 为了支持持续改进，您批准管理员和开发人员利用他们所有的额外时间来选择和实施改进。没有完成任何改进。

建立此最佳实践的好处：通过在流程中投入专用时间和资源，您可以实现持续增量改进。

未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

- 分配时间进行改进：在流程中抽出专门的时间和资源，用于实现持续增量改进。实施更改以便改进，并评估结果以确定是否成功。如果结果不符合目标，并且仍然需要改进，则寻求其他行动方案。

## 安全性

安全性支柱包括保护数据、系统和资产以利用云技术来改善安全性的能力。如需有关具体实施的说明性指导，请参阅 [《安全性支柱》白皮书](#) 访问 AWS 资源。

### 最佳实践领域

- [安全基础知识](#)
- [身份与访问权限管理](#)
- [检测](#)
- [基础设施保护](#)
- [数据保护](#)
- [事件响应](#)
- [应用程序安全性](#)

## 安全基础知识

### 问题

- [SEC 1.如何安全地操作您的工作负载？](#)

### SEC 1.如何安全地操作您的工作负载？

为了安全地操作您的工作负载，您必须对安全性的各个方面应用总体最佳实践。采用您在组织和工作负载层面的卓越运营中定义的要求和流程，并将它们应用到各个方面。及时了解最新的 AWS、行业建议以及威胁情报信息可帮助您改进您的威胁模型和控制目标。实现安全流程、测试和验证的自动化可扩展您的安全运营。

### 最佳实践

- [SEC01-BP01 通过使用账户来分隔工作负载](#)
- [SEC01-BP02 保护账户根用户和属性](#)

- [SEC01-BP03 识别并验证控制目标](#)
- [SEC01-BP04 及时了解最新的安全威胁](#)
- [SEC01-BP05 及时了解最新的安全建议](#)
- [SEC01-BP06 在管道中自动测试和验证安全控制措施](#)
- [SEC01-BP07 使用威胁模型识别威胁并确定缓解措施的优先级](#)
- [SEC01-BP08 定期评估和实施新的安全服务和功能](#)

## SEC01-BP01 通过使用账户来分隔工作负载

通过采取多账户策略，在环境（如生产、开发和测试）和工作负载之间建立共同的防护机制和隔离措施。强烈建议在账户层面进行分离管理，这样可为安全性、账单和访问提供强大的隔离边界。

期望结果：形成一种账户结构，可将云运维、无关工作负载和环境隔离到单独的账户中，从而提高整个云基础设施的安全性。

常见反模式：

- 将多个相互毫无关联，具有不同数据敏感度级别的工作负载放入同一账户中。
- 组织单位（OU）结构界定不清。

建立此最佳实践的好处：

- 即使不该访问的工作负载无意中被访问了，影响范围也会缩小。
- 能够对访问 AWS 服务、资源和区域进行集中治理。
- 可集中管理策略和安全服务，维护云基础设施的安全性。
- 使账户创建和维护流程自动化。
- 集中审计基础设施状况，从而满足法规遵从性和监管要求。

在未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

AWS 账户 提供安全隔离边界，使不同敏感度的工作负载或资源相互分离。AWS 提供相应工具，以多账户策略来大规模管理云工作负载，从而利用此隔离边界。如要获得有关 AWS 多账户策略的概念、模式和实施的指导，请参阅[使用多个账户组织 AWS 环境](#)。

如果您需要集中管理多个 AWS 账户，您的账户应基于组织单位（OU）层，建立层次结构。然后可以建立安全控制机制，并将其应用于 OU 和成员账户，从而为组织内的成员账户建立一致的预防性控制机制。安全控制机制是层层继承的，使您能够筛选位于 OU 层次结构较低层次的成员账户的可用权限。优秀的架构设计将能够利用这种层层继承的特性，减少设置安全策略，降低复杂性，并使每个成员账户的安全控制效果达到预期。

采用 [AWS Organizations](#) 和 [AWS Control Tower](#) 这两种服务，可在您的 AWS 环境中实施和管理多账户结构。AWS Organizations 使得您能够将账户建立成由一个或多个 OU 层定义的层次结构形式，每个 OU 均可包含若干成员账户。[服务控制策略](#)（SCP）使组织管理员能够对成员账户建立精细的预防性控制，而 [AWS Config](#) 可用于建立对成员账户的主动式和检测性控制。许多 AWS 服务与 [AWS Organizations 集成](#)，以提供委派型的管理控制，并在组织内的所有成员账户中执行特定于服务的任务。

[AWS Control Tower](#) 位于 AWS Organizations 之上，为具有[登录区](#)的多账户 AWS 环境提供了一键式最佳实践设置。登录区是 Control Tower 多账户环境的入口处。与 AWS Organizations 相比，采用 Control Tower 具有若干[好处](#)。这三种好处可以改进账户治理状况：

- 将强制安全防护机制集成于系统中，可自动应用于准入组织的账户。
- 有多种防护机制可供选择，还能开启或关闭给定 OU 组的防护机制。
- [AWS Control Tower Account Factory](#) 可在组织内部自动部署账户，设置好预先批准的基准和配置选项。

## 实施步骤

1. 设计组织单位结构：设计良好的组织单位结构减少了创建和维护服务控制策略及其他安全控制机制所需的管理负担。您的组织单位结构应[与您的业务需求、数据敏感度和工作负载结构看齐](#)。
2. 为多账户环境创建登录区：登录区提供了一致的安全性和基础设施基础，您的组织可以从中快速开发、启动和部署工作负载。您可以使用[定制的登录区或 AWS Control Tower](#) 来编排您的环境。
3. 建立防护机制：通过登录区为您的环境实施一致的安全防护机制。AWS Control Tower 提供了可部署的[必选](#)和[可选](#)控制机制的列表。实施 Control Tower 时会自动部署必选控制机制。查看[高度推荐](#)和[可选控制机制的列表](#)，并实施适合您需求的控制机制。
4. 限制访问新添加的区域：对于新的 AWS 区域，诸如用户和角色之类的 IAM 资源将仅传播到您指定的区域。可以在[使用 Control Tower 时通过控制台](#)执行此操作，也可以通过调整 [AWS Organizations 中的 IAM 权限策略](#) 执行此操作。
5. 考虑使用 AWS [CloudFormation StackSets](#)：StackSets 可帮助您通过已批准的模板将资源（包括 IAM 策略、角色和组）部署到不同的 AWS 账户和区域中。

## 资源

相关最佳实践：

- [SEC02-BP04 依赖集中式身份提供程序](#)

相关文档：

- [AWS Control Tower](#)
- [AWS 安全性审计指导原则](#)
- [IAM 最佳实践](#)
- [使用 CloudFormation StackSets 跨多个 AWS 账户 和区域预置资源](#)
- [组织常见问题解答](#)
- [AWS Organizations 术语和概念](#)
- [AWS Organizations 多账户环境中的服务控制策略的最佳实践](#)
- [AWS 账户管理参考指南](#)
- [使用多个账户组织 AWS 环境](#)

相关视频：

- [利用自动化和监管，支持大规模采用 AWS](#)
- [架构完善的安全性最佳实践](#)
- [使用 AWS Control Tower 构建和监管多个账户](#)
- [为现有组织启用 Control Tower](#)

相关研讨会：

- [Control Tower 沉浸日](#)

## SEC01-BP02 保护账户根用户和属性

根用户是 AWS 账户 中权限最高的用户，对账户内的所有资源具有完全管理访问权限，在某些情况下不受安全策略的约束。禁用对根用户的编程访问，为根用户建立适当的控制机制，并避免日常使用根用户。这样有助于降低无意中暴露根凭证以及随后破坏云环境的风险。

期望结果：保护根用户有助于减少因滥用根用户凭证而导致意外或故意损坏的可能性。建立检测性控制机制也可以在有人使用根用户执行操作时向适当人员发出警报。

常见反模式：

- 使用根用户执行各种任务，而非仅在必要时使用根用户凭证。
- 忽略定期测试应急计划，以验证关键基础设施、流程和人员在紧急情况下的运作情况。
- 只考虑典型的账户登录流程，而没有考虑或测试替代的账户恢复方法。
- 因为 DNS、电子邮件服务器和电话提供商要用于账户恢复流程，就不将其作为关键安全边界的一部分进行处理。

建立此最佳实践的好处：保护对根用户的访问可以建立信心，使您账户中的操作受到控制和审计。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

AWS 提供许多有助于保护账户安全的工具。但由于其中一些措施默认情况下未启用，因此您必须采取直接行动来实施它们。请将这些建议视为确保 AWS 账户 安全的基本步骤。实施这些步骤时，务必建立一个持续评估和监控安全控制机制的过程，这非常重要。

当您首次创建 AWS 账户 时，最初使用的是一个对账户中所有 AWS 服务和资源有完全访问权限的身份。该身份称为 AWS 账户 根用户。您可以使用创建账户时使用的电子邮件地址和密码，以根用户的身份登录。由于授予 AWS 根用户的访问权限较高，您必须仅将 AWS 根用户用于执行[特别需要它](#)的任务。必须严格保护根用户登录凭证，并且应始终为 AWS 账户 根用户启用多重身份验证 (MFA)。

除了使用用户名、密码和多重身份验证 (MFA) 设备登录根用户的常规身份验证流程外，还可以使用账户恢复流程登录您的 AWS 账户 根用户，该用户可以访问与您的账户关联的电子邮件地址和电话号码。因此，保护发送恢复电子邮件的根用户电子邮件账户和保护与该账户关联的电话号码同样重要。还应考虑潜在的循环依赖关系，其中与根用户关联的电子邮件地址托管在同一 AWS 账户 的电子邮件服务器或域名服务 (DNS) 资源上。

使用 AWS Organizations 时，有多个 AWS 账户 ( 均有一个根用户 )。将一个账户指定为管理账户，然后可以在管理账户下面添加几层成员账户。优先保护管理账户的根用户，然后解决成员账户根用户问题。保护管理账户根用户的策略可能与保护成员账户根用户的策略不同，您可以对成员账户根用户建立预防性安全控制机制。



## 实施步骤

建议使用以下实施步骤为根用户建立控制机制。在适用情况下，建议与 [CIS AWS Foundations Benchmark 版本 1.4.0](#) 交叉引用。除了这些步骤外，请参阅 [AWS 最佳实践指导](#) 以确保您的 AWS 账户和资源安全。

### 预防性控制机制

1. 为账户设置准确的[联系信息](#)。
  - a. 该信息用于丢失的密码恢复流程、丢失的 MFA 设备账户恢复流程，以及与您的团队进行关键的安全相关通信。
  - b. 使用企业域托管的电子邮件地址（最好是通讯组列表）作为根用户的电子邮件地址。使用通讯组列表而不是个人的电子邮件账户可提供额外的冗余和连续性，以便在很长一段时间内访问根账户。
  - c. 联系信息上所列的电话号码应该是为此目的而设置的专用安全电话的号码。电话号码不应列出或与任何人共享。
2. 不要为根用户创建访问密钥。如果存在访问密钥，请将其删除（CIS 1.4）。
  - a. 消除根用户的任何长期编程凭证（访问密钥和私有密钥）。
  - b. 如果已存在根用户访问密钥，您应将使用这些密钥的进程转换为使用 AWS Identity and Access Management（IAM）角色的临时访问密钥，然后[删除根用户访问密钥](#)。
3. 确定是否需要为根用户存储凭证。
  - a. 如果您使用 AWS Organizations 创建新的成员账户，则新成员账户上根用户的初始密码将设置为一个不向您公开的随机值。如果需要，请考虑使用 AWS 组织管理账户的密码重置流程来[访问成员账户](#)。
  - b. 对于独立 AWS 账户 或管理 AWS 组织账户，请考虑为根用户创建并安全地存储凭证。为根用户启用 MFA。
4. 在 AWS 多账户环境中，为成员账户根用户启用预防性控制机制。
  - a. 考虑为成员账户启用[不允许为根用户创建根访问密钥](#)预防性防护机制。
  - b. 考虑为成员账户启用[不允许以根用户身份执行操作](#)预防性防护机制。
5. 如果需要根用户凭证，请执行以下操作：
  - a. 使用复杂密码。
  - b. 为根用户启用多重身份验证（MFA），特别是 AWS Organizations 管理（付款人）账户（CIS 1.5）。

- c. 考虑使用硬件 MFA 设备以提高弹性和安全性，因为一次性使用的设备可以减少包含 MFA 代码的设备被重复用于其他用途的可能性。验证是否定期更换由电池供电的硬件 MFA 设备。( CIS 1.6 )
    - 要为根用户配置 MFA，请遵循启用[虚拟 MFA](#) 或[硬件 MFA 设备](#) 的说明。
  - d. 考虑注册多个 MFA 设备用于备份。[每个账户最多允许 8 个 MFA 设备](#)。
    - 请注意，为根用户注册多个 MFA 设备将自动禁用[在 MFA 设备丢失情况下恢复账户的流程](#)。
  - e. 安全地存储密码，如果以电子方式存储密码，则考虑循环依赖关系。不要以需要访问同一 AWS 账户 才能获得密码的方式存储密码。
6. 可选：考虑为根用户制定定期密码轮换计划。
- 凭证管理最佳实践取决于您的监管和政策要求。受 MFA 保护的根用户并不依赖密码作为单重身份验证。
  - 定期[更改根用户密码](#)可降低无意中暴露的密码被滥用的风险。

## 检测性控制

- 创建警报以检测根凭证的使用 ( CIS 1.7 )。[启用 Amazon GuardDuty](#) 将通过 [RootCredentialUsage](#) 调查结果对根用户 API 凭证的使用进行监控和发出警报。
- 评估并实施 [AWS Well-Architected 安全性支柱合规包 AWS Config](#) 中包含的检测性控制机制，或者如果使用 AWS Control Tower，则评估并实施 Control Tower 内[强烈建议的控制机制](#)。

## 运营指南

- 确定组织中应该有权访问根用户凭证的人员。
  - 采用双人规则，以便不会出现一个人就能够访问所有必要凭证和 MFA 来获得根用户访问权限的情况。
  - 验证组织 ( 而不是个人 ) 对与账户关联的电话号码和电子邮件别名 ( 用于密码重置和 MFA 重置流程 ) 保持控制。
- 仅在例外情况下使用根用户 ( CIS 1.7 )。
  - 不得使用 AWS 根用户执行日常任务，即使是管理任务也不可以。仅以根用户身份登录，以执行[需要根用户的 AWS 任务](#)。所有其他操作都应由代入适当角色的其他用户执行。
- 定期检查对根用户的访问是否正常，以便在出现需要使用根用户凭证的紧急情况之前对过程进行测试。

- 定期检查与账户关联的电子邮件地址以及[备用联系人](#)下列出的电子邮件地址是否有效。监控这些电子邮件收件箱，查看您可能从中收到的安全通知 <abuse@amazon.com>。还要确保与该账户相关的任何电话号码都有效。
- 准备事件响应程序，以应对根账户滥用情况。请参阅 [AWS 安全事件响应指南](#) 以及 [《安全性支柱》白皮书“事件响应”部分](#) 中的最佳实践，了解有关为 AWS 账户构建事件响应策略的更多信息。

## 资源

### 相关最佳实践：

- [SEC01-BP01 通过使用账户来分隔工作负载](#)
- [SEC02-BP01 使用强大的登录机制](#)
- [SEC03-BP02 授予最低访问权限](#)
- [SEC03-BP03 建立紧急访问流程](#)
- [SEC10-BP05 预置访问权限](#)

### 相关文档：

- [AWS Control Tower](#)
- [AWS 安全性审计指导原则](#)
- [IAM 最佳实践](#)
- [Amazon GuardDuty – 根凭证使用情况警报](#)
- [通过 CloudTrail 监控根凭证使用情况的分步指导](#)
- [获准与 AWS 一起使用的 MFA 令牌](#)
- 在 AWS 上实施 [Break Glass 访问](#)
- [AWS 账户中需要改进的十大安全项目](#)
- [发现我的 AWS 账户 中存在未经授权的活动时该怎么办？](#)

### 相关视频：

- [利用自动化和监管，支持大规模采用 AWS](#)
- [架构完善的安全性最佳实践](#)
- [限制使用来自 AWS re:inforce 2022 的 AWS 根凭证 – AWS IAM 安全最佳实践](#)

相关示例和实验室：

- [实验室：AWS 账户 和根用户](#)

## SEC01-BP03 识别并验证控制目标

根据您的合规性要求以及从威胁模型中发现的风险，获得并验证您需要应用于工作负载的控制目标和控制措施。持续验证控制目标和控制措施可帮助您衡量风险缓解措施的有效性。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

- 确定合规性要求：了解您的工作负载必须符合的组织、法律和合规性要求。
- 确定 AWS 合规性资源：确定 AWS 帮助您实现合规性的资源。
  - <https://aws.amazon.com/compliance/>
  - <https://aws.amazon.com/artifact/>

### 资源

相关文档：

- [AWS 安全性审计指导原则](#)
- [安全公告](#)

相关视频：

- [AWS Security Hub：管理安全警报和自动执行合规性检查](#)
- [架构完善的安全性最佳实践](#)

## SEC01-BP04 及时了解最新的安全威胁

通过及时了解最新的安全威胁，帮助您定义并实施适当的控制措施，识别攻击媒介。使用 AWS Managed Services 可以更轻松地接收 AWS 账户中意外或异常行为的通知。在您的安全信息流程中，使用 AWS 合作伙伴工具或第三方威胁信息源进行调查。此 [通用漏洞披露 \( CVE , Common Vulnerabilities and Exposures \) 列表](#) 包含公开披露的网络安全漏洞，可供您用于掌握最新信息。

未建立此最佳实践暴露的风险等级：高

## 实施指导

- 订阅威胁情报来源：定期查看来自多个来源、与您在工作负载中所用技术相关的威胁情报信息。
  - [通用漏洞披露列表](#)
- 考虑使用 [AWS Shield Advanced](#) 服务：如果您的工作负载可通过互联网访问，则该服务可让您近乎实时地了解情报来源。

## 资源

### 相关文档：

- [AWS 安全性审计指导原则](#)
- [AWS Shield](#)
- [安全公告](#)

### 相关视频：

- [架构完善的安全性最佳实践](#)

## SEC01-BP05 及时了解最新的安全建议

及时了解最新的 AWS 和行业安全建议，以改善您的工作负载安全状况。[AWS 安全公告](#) 包含有关安全性和隐私通知的重要信息。

未建立此最佳实践暴露的风险等级：高

## 实施指导

- 关注 AWS 更新：订阅或定期查看新建议、提示与诀窍。
  - [AWS Well-Architected 实验室](#)
  - [AWS 安全性博客](#)
  - [AWS 服务文档](#)
- 订阅行业新闻：定期查看来自多个来源、与您在工作负载中所用技术相关的新闻动态。
  - [示例：通用漏洞披露列表](#)

## 资源

### 相关文档：

- [安全公告](#)

### 相关视频：

- [架构完善的安全性最佳实践](#)

## SEC01-BP06 在管道中自动测试和验证安全控制措施

为安全机制建立可靠的基准和模板，并将其作为构建、管道和流程的一部分进行测试和验证。利用工具和自动化功能，持续测试并验证所有的安全控制措施。例如，对机器镜像和基础设施即代码模板等项目进行扫描，以发现安全漏洞、异常以及与每个阶段的既定基准的偏差。AWS CloudFormation Guard 可帮助您验证 CloudFormation 模板是否安全，为您节省时间并减少配置错误风险。

减少引入到生产环境中的安全性错误配置的数量至关重要 — 在构建过程中，可以执行的质量控制和可以减少的缺陷越多越好。设计持续集成和持续部署 (CI/CD) 管道，以便尽可能测试安全问题。CI/CD 管道提供了在构建和交付的每个阶段增强安全性的机会。还必须确保 CI/CD 安全工具始终是最新版本，以减轻不断变化的威胁。

跟踪对工作负载配置进行的更改，帮助您进行合规性审计、更改管理以及可能适用于您的调查。您可以使用 AWS Config 记录和评估您的 AWS 和第三方资源。这使您可以依据规则及合规包（合规包是带有补救操作的规则集合），连续审计和评估您的整体合规情况。

更改跟踪应包括计划更改，计划更改可能是组织更改控制流程（有时也称作 MACD，即移动、添加、更改、删除（Move, Add, Change, Delete））、临时更改或意外更改（如意外事件）的一部分。更改可能出现在基础设施中，但也可能涉及其他类别，如代码存储库中的更改、机器镜像和应用程序清单更改、流程和策略更改或文档更改。

未建立此最佳实践暴露的风险等级：中

### 实施指导

- 自动管理配置：使用配置管理服务或工具自动实施安全配置并对其进行验证。
  - [AWS Systems Manager](#)
  - [AWS CloudFormation](#)
  - [在 AWS 上设置 CI/CD 管道](#)

## 资源

### 相关文档：

- [如何使用服务控制策略来设置您在 AWS Organization 中的跨账户权限防护机制](#)

### 相关视频：

- [使用 AWS Organizations 管理多账户 AWS 环境](#)
- [架构完善的安全性最佳实践](#)

## SEC01-BP07 使用威胁模型识别威胁并确定缓解措施的优先级

执行威胁建模，以识别并维护一个针对工作负载的潜在威胁和相关缓解措施的最新登记表。确定威胁优先级并调整安全控制缓解措施，以进行防范、检测和响应。根据您的工作负载以及不断变化的安全环境，重新审视和维护此登记表。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

#### 什么是威胁建模？

“威胁建模可识别、沟通和理解威胁及缓解措施，用于保护重要资产。” – [开放 Web 应用程序安全项目 \(OWASP\) 应用程序威胁建模](#)

#### 为什么应该进行威胁建模？

系统是复杂的，并且随着时间的推移会变得越来越复杂，功能越来越强大，从而提供更多业务价值，提高客户满意度和参与度。这意味着 IT 设计决策需要考虑数量不断增加的使用案例。由于复杂性和使用案例排列组合的数量过多，非结构化方法将通常无法有效地发现和减轻威胁。因此，您需要采用一种系统方法来列举对系统的潜在威胁，制定缓解措施并确定这些缓解措施的优先级，以确保贵组织利用有限资源，在改善系统的整体安全状况方面发挥巨大作用。

威胁建模旨在提供这种系统方法，目的是在设计过程的早期发现和解决问题。此时与生命周期的后期相比，缓解措施的成本和工作量相对较低。这种方法与[左移 安全实践](#)的行业原则相一致。最终，威胁建模将与组织的风险管理流程相互集成，并通过使用威胁驱动的方法，帮助您决定实施哪些控制措施。

#### 应在什么时候进行威胁建模？



应在工作负载的生命周期中尽早开始进行威胁建模，这使您能够更灵活地处理已识别的威胁。就像软件漏洞一样，越早发现威胁，解决它们就越经济高效。威胁模型是一个动态文档，应随着工作负载的变化而不断发展。随着时间的推移（包括当发生重大变更、威胁形势发生变化或您采用新功能或服务时），重新审视您的威胁模型。

## 实施步骤

### 我们如何进行威胁建模？

可以采用许多不同的方式来进行威胁建模。就像编程语言一样，每种方式都有优点和缺点，您应该选择最适合自己的方式。一种方法是从 [Shostack 的威胁建模 4 问题框架](#) 开始，该框架提出开放式问题，为您的威胁建模工作提供结构：

#### 1. 正在做什么？

该问题旨在帮助您了解所构建的系统并达成一致意见，以及了解与安全性相关的系统细节。创建模型或图表是回答该问题的常用方法，因为这可帮助您对所构建的内容进行可视化，例如使用 [数据流图](#)。写下关于系统的假设和重要细节也有助于定义范围内的内容。这使每个参与威胁建模的人员都能专注于同一件事，避免因在超出范围的主题（包括过时的系统版本）上走弯路而浪费时间。例如，如果您要构建一个 Web 应用程序，那么可能不值得花时间为浏览器客户端操作系统可信引导顺序进行威胁建模，因为您无法在设计中影响这一点。

#### 2. 会出现什么问题？

该问题可帮助您识别系统存在的威胁。威胁是指具有不必要影响并可能影响系统安全的意外或故意行为或事件。如果不清楚哪里可能出现问题，您就无从应对。

对于可能出现的问题，并没有一个规范的列表。创建此列表需要团队中的所有个人和参与威胁建模工作的 [相关角色](#) 集思广益并展开协作。您可以通过使用识别威胁的模型（如 [STRIDE](#)）来帮助集思广益，该模型建议了不同的评估类别：欺骗、篡改、抵赖、信息披露、拒绝服务和权限提升。此外，您可能希望通过回顾现有的列表和研究来帮助集思广益，寻找灵感，其中包括 [OWASP Top 10](#)、[HiTrust 威胁目录](#) 和贵组织自己的威胁目录。

#### 3. 我们要怎么做？

与前面的问题一样，我们不可能得到包含所有缓解措施的规范清单。这一步骤需要考虑的是上一步中确定的威胁、威胁行动者和要改进的领域。

安全性和合规性是 [您与 AWS 的共同责任](#)。重要的是要明白，当您问“我们要怎么做？”时，您也在问“谁负责做这件事？”。了解您和 AWS 之间的责任平衡有助于将威胁建模工作的范围限定在您控

制的缓解措施范围内，这些缓解措施通常是 AWS 服务配置选项和您自己的系统特定缓解措施的组合。

对于共担责任中 AWS 应承担的部分，您会发现 [AWS 服务在许多合规计划的范围内](#)。这些计划可帮助您理解 AWS 用以维持云安全性与合规性的可靠控制机制。AWS 客户可以从 [AWS Artifact](#) 下载这些计划的审计报告。

无论您使用哪项 AWS 服务，客户始终要承担一部分责任，并且与这些责任相一致的缓解措施应包含在威胁模型中。对于 AWS 服务自身的安全控制缓解措施，您需要考虑跨域实施安全控制措施，包括身份和访问管理（身份验证和授权）、数据保护（静态和传输中）、基础设施安全性、日志记录和监控等域。每项 AWS 服务的文档都包含一个[专门的安全章节](#)，里面提供了关于安全控制机制的指导，可视为缓解措施。重要的是，需要考虑您正在编写的代码及其代码依赖关系，并思考可以设置哪些控制机制来应对这些威胁。这些控制机制可以是[输入验证](#)、[会话处理](#)和[边界处理](#)等内容。通常，大多数漏洞都是在自定义代码中引入，因此请重点关注这一领域。

#### 4. 我们做得好吗？

该问题旨在随着时间的推移，让您的团队和组织提高威胁模型的质量并加快执行威胁建模的速度。通过将实践、学习、教学和回顾相结合可以取得这些改进。要想深入了解并亲身体验，建议您和您的团队完成[“适合构建者的威胁建模方式”培训课程](#)或[研讨会](#)。此外，如果您正在寻找如何将威胁建模集成到组织的应用程序开发生命周期中的指导，请参阅 AWS 安全博客上的[如何处理威胁建模](#)一文。

### Threat Composer

为了协助和指导您执行威胁建模，可以考虑使用 [Threat Composer](#) 工具，该工具旨在缩短威胁建模时实现价值的时间。该工具有便于您执行以下操作：

- 撰写符合[威胁语法](#)、能够在自然非线性工作流程中发挥作用的有用威胁语句
- 生成人类可读的威胁模型
- 生成机器可读的威胁模型，允许您将威胁模型视为代码
- 使用 Insights 控制面板协助您快速确定质量和覆盖范围有待改进的方面

如需更多参考，请访问 Threat Composer 并切换到系统定义的示例工作区。

### 资源

相关最佳实践：

- [SEC01-BP03 识别并验证控制目标](#)
- [SEC01-BP04 及时了解最新的安全威胁](#)
- [SEC01-BP05 及时了解最新的安全建议](#)
- [SEC01-BP08 定期评估和实施新的安全服务和功能](#)

相关文档：

- [如何处理威胁建模](#) ( AWS 安全性博客 )
- [NIST：以数据为中心的系统威胁建模指南](#)

相关视频：

- [2021 AWS ANZ 峰会 - 如何处理威胁建模](#)
- [2022 AWS ANZ 峰会 - 扩展安全性 – 优化以实现快速而安全的交付](#)

相关培训：

- [适合构建者的威胁建模方式 – AWS Skill Builder 自控进度的虚拟培训](#)
- [适合构建者的威胁建模方式 – AWS 研讨会](#)

相关工具：

- [Threat Composer](#)

## SEC01-BP08 定期评估和实施新的安全服务和功能

评估并实施 AWS 和 AWS 合作伙伴提供的安全服务和功能，以改善您的工作负载安全状况。AWS 安全博客重点介绍新的 AWS 服务和功能、实施指导和常规安全指南。[AWS 的最新内容](#) 是一个很好的工具，可帮助您随时了解所有新的 AWS 功能、服务和公告。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

- 规划定期审核：创建审核活动日历，包括遵守合规性要求、评估新的 AWS 安全功能和服务，以及及时了解行业最新动态。

- 发现 AWS 服务和功能：发现适用于您使用的服务的安全功能，并在新功能发布时查看这些功能。
  - [AWS 安全性博客](#)
  - [AWS 安全公告](#)
  - [AWS 服务文档](#)
- 定义 AWS 服务上线流程：定义用于上线新 AWS 服务的流程。包括您如何评估新 AWS 服务的功能，以及针对工作负载的合规性要求。
- 测试新的服务和功能：当有新的服务和功能发布时，在与生产环境非常相似的非生产环境中对其进行测试。
- 实施其他防御机制：实施自动化机制来保护您的工作负载，并探索可用选项。
  - [按照 AWS Config 规则 修正不合规的 AWS 资源](#)

## 资源

相关视频：

- [架构完善的安全性最佳实践](#)

## 身份与访问权限管理

### 问题

- [SEC 2.如何管理人员和计算机的身份验证？](#)
- [SEC 3.如何管理人员和机器的权限？](#)

### SEC 2.如何管理人员和计算机的身份验证？

在访问和运行安全的 AWS 工作负载时，您必须管理两种类型的身份。了解您必须管理和授予访问权限的身份类型，这有助于确保正确的身份能够在正确的条件下访问正确的资源。

**人员身份：**您的管理员、开发人员、操作员和最终用户需要确定身份才能访问您的 AWS 环境和应用程序。这些是您的组织成员或您与之协作的外部用户，以及通过 Web 浏览器、客户端应用程序或交互式命令行工具与您的 AWS 资源交互的用户。

**机器身份：**您的服务应用程序、操作工具和工作负载需要一个身份来向 AWS 服务发出请求，例如，读取数据。这些身份包括在 AWS 环境中运行的机器，例如 Amazon EC2 实例或 AWS Lambda 函数。您还可以管理需要访问权限的外部各方的机器身份。此外，您可能还有需要访问您 AWS 环境的 AWS 之外的机器。

## 最佳实践

- [SEC02-BP01 使用强大的登录机制](#)
- [SEC02-BP02 使用临时凭证](#)
- [SEC02-BP03 安全地存储和使用密钥](#)
- [SEC02-BP04 依赖集中式身份提供程序](#)
- [SEC02-BP05 定期审计和轮换凭证](#)
- [SEC02-BP06 利用用户组和属性](#)

### SEC02-BP01 使用强大的登录机制

当不使用多重身份验证 ( MFA ) 等机制时，登录 ( 使用登录凭证的身份验证 ) 可能会带来风险，特别是在登录凭证被无意泄露或很容易猜到的情况下。使用强大的登录机制，通过要求使用 MFA 和强密码策略来降低这些风险。

期望结果：通过为 [AWS Identity and Access Management \( IAM \)](#) 用户、[AWS 账户根用户](#)、[AWS IAM Identity Center](#) ( AWS Single Sign-On 的后继者 ) 和第三方身份提供者使用强大的登录机制，降低意外访问 AWS 中凭证的风险。这意味着需要 MFA，强制执行强密码策略，并检测异常登录行为。

常见反模式：

- 没有为您的身份执行强密码策略，包括复杂密码和 MFA。
- 在不同的用户之间共享相同的凭证。
- 不对可疑的登录使用检测性控制。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

人类身份可通过多种方式登录 AWS。在向 AWS 进行身份验证时，AWS 最佳做法是寻找使用联合身份验证 ( 直接联合或使用 AWS IAM Identity Center ) 的集中式身份提供者进行验证。在这种情况下，您应与您的身份提供者或 Microsoft Active Directory 建立安全登录过程。

第一次开设 AWS 账户时，您会从 AWS 账户根用户开始。您应仅使用账户根用户为您的用户 ( 以及为[需要根用户的任务](#) ) 设置访问权限。开设 AWS 账户后立即为账户根用户启用 MFA，并使用 AWS [最佳实践指南](#)保护根用户，这一点非常重要。

如果您在 AWS IAM Identity Center 中创建用户，请确保该服务中的登录过程安全。对于消费者身份，您可以使用 [Amazon Cognito user pools](#) 并保护该服务中的登录过程，或者使用 Amazon Cognito user pools 支持的身份提供者之一。

如果您使用 [AWS Identity and Access Management \( IAM \)](#) 用户，则将使用 IAM 保护登录过程。

无论采用何种登录方法，执行强登录策略都非常关键。

### 实施步骤

以下是一般的强登录建议。应根据您的公司策略或使用 [NIST 800-63](#) 等标准，对您配置的实际设置进行设置。

- 要求使用 MFA。对于人类身份和工作负载，[要求使用 MFA 是 IAM 最佳实践](#)。启用 MFA 提供了一层额外的安全保障，这会要求用户提供登录凭证和一次性密码（OTP），或从硬件设备加密验证和生成的字符串。
- 强制执行最小密码长度，这是密码强度的主要因素。
- 强制执行密码复杂性，使密码更难以猜到。
- 允许用户更改自己的密码。
- 创建个人身份而不是共享凭证。通过创建个人身份，您可以为每个用户提供一组唯一的安全凭证。个人用户可以审计每个用户的活动。

### IAM Identity Center 建议：

- IAM Identity Center 在使用默认目录时提供了预定义的[密码策略](#)，该策略确定了密码长度、复杂性和重用要求。
- 当身份源为默认目录、AWS Managed Microsoft AD 或 AD Connector 时，[启用 MFA](#) 并为 MFA 配置“背景认知”或“始终开启”设置。
- 允许用户[注册自己的 MFA 设备](#)。

### Amazon Cognito user pools 目录建议：

- 配置[密码长度](#)设置。
- 对于用户，[要求使用 MFA](#)。
- 使用 Amazon Cognito user pools [高级安全设置](#)实现[自适应身份验证](#)（可阻止可疑登录）等功能。

### IAM 用户建议：

- 最好是使用 IAM Identity Center 或直接联合。然而，您可能需要 IAM 用户。在这种情况下，为 IAM 用户[设置密码策略](#)。您可以使用密码策略来定义诸如最小长度、密码是否需要非字母字符之类的要求。
- 创建 IAM 策略来[强制执行 MFA 登录](#)，以允许用户管理其自己的密码和 MFA 设备。

## 资源

### 相关最佳实践：

- [SEC02-BP03 安全地存储和使用密钥](#)
- [SEC02-BP04 依赖集中式身份提供程序](#)
- [SEC03-BP08 在组织内安全地共享资源](#)

### 相关文档：

- [AWS IAM Identity Center \( AWS Single Sign-On 的后继者 \) 密码策略](#)
- [IAM 用户密码策略](#)
- [设置 AWS 账户 根用户密码](#)
- [Amazon Cognito 密码策略](#)
- [AWS 凭证](#)
- [IAM 安全最佳实践](#)

### 相关视频：

- [使用 AWS IAM Identity Center 大规模管理用户权限](#)
- [在每个层面掌握身份](#)

## SEC02-BP02 使用临时凭证

进行任何类型的身份验证时，最好使用临时凭证而不是长期凭证，以降低或消除诸如凭证被无意泄露、共享或被盗之类的风险。

期望结果：为了降低长期凭证的风险，请尽可能对人类和机器身份使用临时凭证。长期凭证会带来许多风险，例如，它们能够以代码的形式上传到公有 GitHub 存储库。使用临时凭证可以显著降低凭证被泄露的几率。



## 常见反模式：

- 开发人员使用 IAM users 的长期访问密钥，而不是使用联合身份验证从 CLI 获得临时凭证。
- 开发人员在他们的代码中嵌入长期访问密钥，并将该代码上传到公有 Git 存储库。
- 开发人员在移动应用程序中嵌入长期访问密钥，然后在应用商店中提供这些密钥。
- 用户与其他用户共享长期访问密钥，或员工离开公司时仍持有长期访问密钥。
- 当可以使用临时凭证时，对机器身份使用长期访问密钥。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

对所有 AWS API 和 CLI 请求使用临时安全凭证，而不是长期凭证。在几乎所有情况下，对 AWS 服务的 API 和 CLI 请求都必须使用 [AWS 访问密钥](#) 进行签名。这些请求可以使用临时凭证或长期凭证进行签名。只有在使用 [IAM 用户](#) 或 [AWS 账户根用户](#) 时，才应该使用长期凭证（也称为长期访问密钥）。当您联合到 AWS 或通过其他方法代入 [IAM 角色](#) 时，系统将生成临时凭证。即使您使用登录凭证访问 AWS Management Console，系统也会生成临时凭证供您调用 AWS 服务。需要用到长期凭证的情况很少，您可以使用临时凭证完成几乎所有任务。

尽量不要使用长期凭证，多用临时凭证，并且还应该减少采取 IAM 用户形式，多用联合身份验证和 IAM 角色形式进行访问。虽然 IAM 过去常使用用户来访问人类和机器身份，但我们现在建议不要使用它们，以避免使用长期访问密钥所带来的风险。

## 实施步骤

对于员工、管理员、开发人员、操作员和客户等人类身份：

- 您应该使用[集中式身份提供者的服务](#)，并要求人类用户配合使用[联合身份验证与身份提供者两种方法](#)，以使用临时凭证访问 AWS。可以通过[直接联合到每个 AWS 账户](#)，或使用 [AWS IAM Identity Center \(AWS IAM Identity Center 的后继者\)](#) 和您选择的身份提供者，对用户进行联合身份验证。与使用 IAM 用户相比，联合身份验证除了消除使用长期凭证的情况之外，还具有许多优势。您的用户也可以从[直接联合](#)的命令行或通过使用 [IAM Identity Center](#) 来请求获得临时凭证。这意味着能够大幅减少需要使用 IAM 用户或用户长期凭证的情况。
- 授予软件即服务 (SaaS) 提供商等第三方访问 AWS 账户中资源的权限时，您可以使用[跨账户角色](#)和[基于资源的策略](#)。
- 如果您需要批准消费者或客户申请访问您的 AWS 资源，可以使用 [Amazon Cognito 身份池](#) 或 [Amazon Cognito user pools](#) 提供临时凭证。通过 IAM 角色配置凭证权限。对于访客用户未通过身份验证的情况，您还可以定义一个拥有有限权限的单独 IAM 角色。

对于机器身份，您就可能需要使用长期凭证了。在这些情况下，您应该[要求工作负载使用具有 IAM 角色的临时凭证来访问 AWS](#)。

- 对于 [Amazon Elastic Compute Cloud](#) ( Amazon EC2 )，您可以使用[适用于 Amazon EC2 的角色](#)。
- [AWS Lambda](#) 使您能够配置 [Lambda 执行角色，以授权此服务](#) 利用临时凭证执行 AWS 操作。AWS 服务还有许多其他类似的模型，可以使用 IAM 角色授予临时凭证。
- 对于 IoT 设备，您可以使用 [AWS IoT Core 凭证提供程序](#) 来请求临时凭证。
- 对于需要访问 AWS 资源的本地系统或在 AWS 之外运行的系统，您可以使用 [IAM Roles Anywhere](#)。

某些情况下不能选择临时凭证，此时您可能需要使用长期凭证。在这些情况下，可以[定期审计和轮换凭证](#)，并[定期轮换访问密钥](#)以解决需要使用长期凭证的情况。诸如使用 WordPress 插件和第三方 AWS 客户端等情况，都可能需要用到长期凭证。在必须使用长期凭证的情况下，或者对于并非 AWS 访问密钥的凭证（如数据库登录），您可以使用一种专门用于处理密钥管理的服务，比如 [AWS Secrets Manager](#)。借助 Secrets Manager，您可以使用[支持的服务](#) 轻松管理、轮换和安全地存储加密密钥。有关轮换长期凭证的更多信息，请参阅[轮换访问密钥](#)。

## 资源

相关最佳实践：

- [SEC02-BP03 安全地存储和使用密钥](#)
- [SEC02-BP04 依赖集中式身份提供程序](#)
- [SEC03-BP08 在组织内安全地共享资源](#)

相关文档：

- [临时安全凭证](#)
- [AWS 凭证](#)
- [IAM 安全最佳实践](#)
- [IAM 角色](#)
- [IAM Identity Center](#)
- [身份提供者和联合身份验证](#)
- [轮换访问密钥](#)

- [安全合作伙伴解决方案：访问和访问控制](#)
- [AWS 账户根用户](#)

相关视频：

- [使用 AWS IAM Identity Center \( AWS IAM Identity Center 的后继者 \) 大规模管理用户权限](#)
- [在每个层面掌握身份](#)

## SEC02-BP03 安全地存储和使用密钥

工作负载需要能够自动向数据库、资源和第三方服务证明其身份。这是使用秘密访问凭证（如 API 访问密钥、密码和 OAuth 令牌）完成的。使用专门构建的服务来存储、管理和轮换这些凭证有助于降低这些凭证泄露的可能性。

期望结果：实施安全管理应用程序凭证的机制，以实现以下目标：

- 确定工作负载需要哪些密钥。
- 尽可能用短期凭证代替长期凭证，从而减少所需的长期凭证的数量。
- 建立安全存储并自动轮换剩余的长期凭证。
- 审计对工作负载中存在的密钥的访问。
- 持续监控，以验证开发期间没有在源代码中嵌入任何密钥。
- 降低凭证被无意中泄露的可能性。

常见反模式：

- 不轮换凭证。
- 将长期凭证存储在源代码或配置文件中。
- 在未加密状态下静态存储凭证。

建立此最佳实践的好处：

- 对存储的凭证进行静态和传输中加密。
- 通过 API 来把关对凭证的访问（可将 API 看作凭证自动售货机）。
- 审计和记录对凭证的访问（包括读和写）。
- 关注点分离：凭证轮换由一个单独的组件执行，该组件可与架构的其余部分隔离开来。

- 密钥自动按需分发给软件组件，并在中心位置进行轮换。
- 可以精细地控制对凭证的访问。

在未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

过去，用于对数据库、第三方 API、令牌和其他密钥进行身份验证的凭证可能已嵌入到源代码或环境文件中。AWS 提供了几种机制来安全地存储这些凭证，自动轮换它们，并审计它们的使用情况。

进行密钥管理的妥善方法是遵循相关指导，以正确地删除、替换和轮换密钥。最安全的凭证是您不必存储、管理或处理的凭证。某些凭证可能不再是正常运行工作负载所必需的，可以安全地删除。

对于仍然是正常运行工作负载所必需的凭证，可能有机会用临时或短期凭证替换长期凭证。例如，相比起对 AWS 秘密访问密钥进行硬编码，不妨考虑使用 IAM 角色，将长期凭证替换为临时凭证。

可能无法删除或替换某些长期密钥。这些密钥可以存储在 [AWS Secrets Manager](#) 等服务中，在那里可以集中存储、管理和定期轮换密钥。

对工作负载的源代码和配置文件进行审计，可以发现许多类型的凭证。下表总结了处理常见凭证类型的策略：

| Credential type | Description                                                                  | Suggested strategy                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|-----------------|------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| IAM access keys | AWS IAM access and secret keys used to assume IAM roles inside of a workload | Replace: Use <a href="#">IAM 角色</a> assigned to the compute instances (such as <a href="#">Amazon EC2</a> or <a href="#">AWS Lambda</a> ) instead. For interoperability with third parties that require access to resources in your AWS 账户, ask if they support <a href="#">AWS 跨账户访问</a> . For mobile apps, consider using temporary credentials through <a href="#">Amazon Cognito 身份池 (联合身份)</a> . For workloads running outside of AWS, consider <a href="#">IAM Roles</a> |

| Credential type                                  | Description                                                                                                 | Suggested strategy                                                                                                                                                                                                                                      |
|--------------------------------------------------|-------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                  |                                                                                                             | <a href="#">Anywhere</a> or <a href="#">AWS Systems Manager</a> 混合激活.                                                                                                                                                                                   |
| SSH keys                                         | Secure Shell private keys used to log into Linux EC2 instances, manually or as part of an automated process | Replace: Use <a href="#">AWS Systems Manager</a> or <a href="#">EC2 Instance Connect</a> to provide programmatic and human access to EC2 instances using IAM roles.                                                                                     |
| Application and database credentials             | Passwords – plain text string                                                                               | Rotate: Store credentials in <a href="#">AWS Secrets Manager</a> and establish automated rotation if possible.                                                                                                                                          |
| Amazon RDS and Aurora Admin Database credentials | Passwords – plain text string                                                                               | Replace: Use the <a href="#">Secrets Manager 与 Amazon RDS 集成</a> or <a href="#">Amazon Aurora</a> . In addition, some RDS database types can use IAM roles instead of passwords for some use cases (for more detail, see <a href="#">IAM 数据库身份验证</a> ). |
| OAuth tokens                                     | Secret tokens – plain text string                                                                           | Rotate: Store tokens in <a href="#">AWS Secrets Manager</a> and configure automated rotation.                                                                                                                                                           |
| API tokens and keys                              | Secret tokens – plain text string                                                                           | Rotate: Store in <a href="#">AWS Secrets Manager</a> and establish automated rotation if possible.                                                                                                                                                      |

一种常见的反模式是在源代码、配置文件或移动应用程序中嵌入 IAM 访问密钥。当需要 IAM 访问密钥与 AWS 服务通信时，请使用[临时（短期）安全凭证](#)。可以通过[EC2 实例的 IAM 角色](#)、Lambda 功能的[执行角色](#)、移动用户访问的[Cognito IAM 角色](#)和 IoT 设备的[IoT Core 策略](#)提供这些短期凭证。与第

三方进行交互时，最好[将访问权限委托给 IAM 角色](#)，以获得对您账户资源的必要访问权限，而不是配置 IAM 用户并向第三方发送该用户的秘密访问密钥。

在许多情况下，工作负载需要存储与其他服务和资源进行互操作所必需的密钥。[AWS Secrets Manager](#) 旨在安全地管理这些凭证，以及 API 令牌、密码和其他凭证的存储、使用和轮换。

AWS Secrets Manager 提供五个关键功能，以确保敏感凭证的安全存储和处理：[静态加密](#)、[传输中加密](#)、[全面审计](#)、[精细访问控制](#)和[可扩展凭证轮换](#)。AWS 合作伙伴提供的其他密钥管理服务或提供类似功能和保证的本地开发的解决方案也可以接受。

## 实施步骤

1. 使用自动化工具（如 [Amazon CodeGuru](#)）识别包含硬编码凭证的代码路径。
  - 使用 Amazon CodeGuru 扫描您的代码存储库。审查完成后，在 CodeGuru 中筛选 Type=Secrets 以查找有问题的代码行。
2. 识别可以删除或替换的凭证。
  - a. 识别不再需要的凭证并标明要删除。
  - b. 对于嵌入到源代码的 AWS 私有密钥，将其替换为与必要资源相关的 IAM 角色。如果您的部分工作负载在 AWS 之外，但需要 IAM 凭证才能访问 AWS 资源，请考虑 [IAM Roles Anywhere](#) 或 [AWS Systems Manager 混合激活](#)。
3. 对于其他需要使用轮换策略的第三方、长期密钥，请将 Secrets Manager 集成到代码中，以便在运行时检索第三方密钥。
  - a. CodeGuru 控制台可以使用发现的凭证[在 Secrets Manager 中自动创建密钥](#)。
  - b. 将 Secrets Manager 的密钥检索集成到应用程序代码中。
    - 无服务器 Lambda 功能可以使用与语言无关的 [Lambda 扩展](#)。
    - 对于 EC2 实例或容器，AWS 用几种流行的编程语言提供了示例[客户端代码，用于从 Secrets Manager 检索密钥](#)。
4. 定期检查代码库并重新扫描，以验证代码中没有添加新的密钥。
  - 考虑使用诸如 [git-secrets](#) 之类的工具来防止向源代码存储库提交新的密钥。
5. [监控 Secrets Manager 活动](#)，以发现意外使用、不适当的密钥访问或试图删除密钥的迹象。
6. 减少人类接触凭证的机会。将读取、写入和修改凭证的权限仅授予专用于此目的的 IAM 角色，并仅向一小部分操作用户提供代入该角色的权限。

## 资源

相关最佳实践：

- [SEC02-BP02 使用临时凭证](#)
- [SEC02-BP05 定期审计和轮换凭证](#)

相关文档：

- [开始使用 AWS Secrets Manager](#)
- [身份提供者和联合身份验证](#)
- [Amazon CodeGuru 推出 Secrets Detector](#)
- [AWS Secrets Manager 如何使用 AWS Key Management Service](#)
- [Secrets Manager 中的密钥加密和解密](#)
- [Secrets Manager 博客条目](#)
- [Amazon RDS 宣布与 AWS Secrets Manager 集成](#)

相关视频：

- [有关大规模管理、检索和轮换密钥的最佳实践](#)
- [使用 Amazon CodeGuru Secrets Detector 查找硬编码密钥](#)
- [使用 AWS Secrets Manager 保护混合工作负载的密钥](#)

相关研讨会：

- [在 AWS Secrets Manager 中存储、检索和管理敏感凭证](#)
- [AWS Systems Manager 混合激活](#)

## SEC02-BP04 依赖集中式身份提供程序

对于员工身份（员工和合同工），请依赖允许您在集中位置管理身份的身份提供程序。这样，您就可以更轻松地跨多个应用程序和系统管理访问权限，因为您在从单一位置创建、分配、管理、撤销和审核访问权限。

期望的结果：您有一个集中式身份提供程序，可以在其中集中管理员工用户、身份验证策略 [例如要求多重身份验证（MFA）]，以及对系统和应用程序的授权（例如根据用户的群组成员资格或属性分配访问权限）。您的员工用户登录到中央身份提供程序并联合身份验证（单点登录）到内部和外部应用程序，这样用户就无需记住多个凭证。您的身份提供程序已与您的人力资源（HR）系统集成，因此人事



变动会自动与身份提供程序同步。例如，如果有人离开您的组织，您可以自动撤消对联合应用程序和系统（包括 AWS）的访问权限。您已在身份提供程序中启用了详细的审核日志记录，并且正在监控这些日志以发现异常用户行为。

常见反模式：

- 您不使用联合身份验证和单点登录。您的员工用户在多个应用程序和系统中创建单独的用户账户和凭证。
- 您尚未实现员工用户身份生命周期的自动化，例如将您的身份提供程序与 HR 系统集成。当用户离开您的组织或变换角色时，您使用手动流程来删除或更新他们在多个应用程序和系统中的记录。

建立此最佳实践的好处：通过使用集中式身份提供程序，您可以在一个位置管理员工用户身份和策略，可以向用户和群组分配应用程序的访问权限，还可以监控用户登录活动。通过与您的人力资源（HR）系统集成，当用户的角色发生更改时，这些更改会同步到身份提供程序，并自动更新为他们分配的应用程序和权限。当用户离职时，其身份将在身份提供程序中自动被禁用，从而撤消他们对联合应用程序和系统的访问权限。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

### 员工用户访问 AWS 的指南

员工用户（如组织中的员工和合同工）可能需要使用 AWS Management Console 或 AWS Command Line Interface（AWS CLI）来访问 AWS，以履行其工作职能。您可以通过从集中式身份提供程序联合到 AWS，在两个层面上向员工用户授予 AWS 访问权：直接联合到每个 AWS 账户，或联合到 [AWS 组织](#) 中的多个账户。

- 要将员工用户与每个 AWS 账户直接联合，可以使用集中式身份提供程序来联合到该账户中的 [AWS Identity and Access Management](#)。IAM 的灵活性允许您启用单独的 [SAML 2.0](#) 或 [Open ID Connect \(OIDC\)](#) 身份提供程序（针对每个 AWS 账户），并使用联合用户属性进行访问控制。您的员工用户将使用网络浏览器，通过提供凭证（如密码和 MFA 令牌码）来登录身份提供程序。身份提供程序向其浏览器发出 SAML 断言，该断言将提交到 AWS Management Console 登录 URL，以允许用户 [通过代入 IAM 角色单点登录到 AWS Management Console](#)。用户还可以获取临时 AWS API 凭证以用于 [AWS CLI](#) 或 [AWS SDK](#) - 从 [AWS STS](#)，方法是 [使用身份提供程序的 SAML 断言代入 IAM 角色](#)。
- 要将员工用户与 AWS 组织中的多个账户联合起来，可以使用 [AWS IAM Identity Center](#) 来集中管理员工用户对 AWS 账户和应用程序的访问权限。您为组织启用 Identity Center 并配置身份源。IAM

Identity Center 提供一个默认身份源目录，您可以用它来管理用户和组。或者，您也可以选择外部身份源，方法是使用 SAML 2.0 [连接外部身份提供程序](#) 并使用 SCIM [自动预置](#) 用户和组，或 [连接到您的 Microsoft AD 目录](#)（使用 [AWS Directory Service](#)）。配置身份源后，您可以通过以下方法为用户和组分配 AWS 账户访问权限：在 [权限集](#) 中定义最低权限策略。您的员工用户可以通过您的中央身份提供程序进行身份验证，以登录 [AWS 访问门户](#) 并单点登录到 AWS 账户以及分配给他们的云应用程序。您的用户可以配置 [AWS CLI v2](#) 以使用 Identity Center 进行身份验证，并获取用于运行 AWS CLI 命令的凭证。Identity Center 还允许通过单点登录访问 AWS 应用程序，例如 [Amazon SageMaker AI Studio](#) 和 [AWS IoT Sitewise Monitor 门户](#)。

遵循上述指导后，您的员工用户在 AWS 上管理工作负载时，将不再需要使用 IAM users 和组来进行通用的操作。相反，您的用户和组将在 AWS 外部进行管理，并且用户可以作为联合身份访问 AWS 资源。联合身份使用您的集中式身份提供程序定义的组。您应该识别并删除 AWS 账户中不再需要的 IAM 组、IAM users 和长期用户凭证（密码和访问密钥）。您可以 [查找未使用的凭证](#)（使用 [IAM 凭证报告](#)）、[删除相应的 IAM users](#) 和 [删除 IAM 组](#)。您可以将 [服务控制策略（SCP）](#) 应用于您的组织，它有助于防止创建新的 IAM users 和组，并强制通过联合身份访问 AWS。

## 应用程序用户指南

您可以通过将 [Amazon Cognito](#) 用作您的集中式身份提供程序，来管理应用程序（例如移动应用程序）用户的身份。Amazon Cognito 为您的 Web 和移动应用程序启用身份验证、授权和用户管理。Amazon Cognito 提供可扩展到数百万用户的身份存储，支持社交网络和企业身份联合验证，并提供高级安全功能来协助保护您的用户和业务。您可以将自定义 Web 或移动应用程序与 Amazon Cognito 集成，以便在几分钟内为您的应用程序添加用户身份验证和访问控制。Amazon Cognito 以 SAML 和 Open ID Connect（OIDC）等开放式身份标准为基础构建，支持各种合规性法规，并与前端和后端开发资源集成。

## 实施步骤

### 员工用户访问 AWS 的步骤

- 通过以下方法之一，使用集中式身份提供程序将您的员工用户联合身份验证到 AWS：
  - 通过与您的身份提供程序联合，使用 IAM Identity Center 来允许单点登录到您的 AWS 组织中的多个 AWS 账户。
  - 使用 IAM 将您的身份提供程序直接连接到每个 AWS 账户，从而实现精细的联合访问。
- 识别并移除被联合身份取代的 IAM users 和群组。

### 适用于您的应用程序用户的步骤

- 将 Amazon Cognito 用作应用程序的集中式身份提供程序。
- 使用 OpenID Connect 和 OAuth 将您的自定义应用程序与 Amazon Cognito 集成。您可以使用 Amplify 库开发自定义应用程序，这些库提供了与各种 AWS 服务（例如用于身份验证的 Amazon Cognito）集成的简单接口。

## 资源

相关的 Well-Architected 最佳实践：

- [SEC02-BP06 利用用户组和属性](#)
- [SEC03-BP02 授予最低访问权限](#)
- [SEC03-BP06 基于生命周期管理访问权限](#)

相关文档：

- [AWS 中的身份联合验证](#)
- [IAM 中的安全最佳实践](#)
- [AWS Identity and Access Management 最佳实践](#)
- [IAM Identity Center 委托管理入门](#)
- [如何针对高级用例在 IAM Identity Center 中使用客户托管策略](#)
- [AWS CLI v2：IAM Identity Center 凭证提供程序](#)

相关视频：

- [AWS re:Inforce 2022 - AWS Identity and Access Management \( IAM \) 深入探讨](#)
- [AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center](#)
- [AWS re:Invent 2018: Mastering Identity at Every Layer of the Cake](#)

相关示例：

- [Workshop: Using AWS IAM Identity Center to achieve strong identity management](#)
- [Workshop: Serverless identity](#)

相关工具：

- [AWS 安全能力合作伙伴：身份和访问管理](#)
- [saml2aws](#)

## SEC02-BP05 定期审计和轮换凭证

定期审计和轮换凭证，以限制凭证可用于访问资源的时间。使用长期凭证会产生许多风险，可通过定期轮换来降低这些风险。

期望结果：实施凭证轮换，以帮助降低长期凭证相关风险。定期审计并纠正不符合凭证轮换策略的情况。

常见反模式：

- 不审计凭证的使用情况。
- 不必要地使用长期凭证。
- 使用长期凭证，不定期轮换。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

当您无法依赖临时凭证并需要长期凭证时，请审计凭证，以验证实施了定义的控制措施（例如多重身份验证（MFA）），并且定期轮换凭证，使凭证具有适当的访问级别。

（最好通过自动化工具）定期验证，以确保实施正确的控制措施。对于人员身份，您应要求用户定期更改他们的密码并弃用访问密钥，以支持临时凭证。从 AWS Identity and Access Management（IAM）用户转向集中式身份时，您可以[生成凭证报告](#)来审计您的用户。

我们还建议您在身份提供者中实施并监控 MFA。您可以设置 [AWS Config 规则](#) 或使用 [AWS Security Hub 安全标准](#) 来监控用户是否启用了 MFA。考虑使用 IAM Roles Anywhere 为机器身份提供临时凭证。在无法使用 IAM 角色和临时凭证的情况下，需要经常审计和轮换访问密钥。

### 实施步骤

- 定期审计凭证：对您的身份提供者和 IAM 中配置的身份进行审计，这有助于验证只有经过授权的身份才能访问您的工作负载。此类身份可能包括但不限于 IAM 用户、AWS IAM Identity Center 用户、Active Directory 用户或不同上游身份提供者中的用户。例如，删除离开组织的人员，并删除不再需要的跨账户角色。制定流程，以定期审计 IAM 实体所访问服务的权限。这有助于您确定需要

修改的策略，以删除任何未使用的权限。使用凭证报告和 [AWS Identity and Access Management Access Analyzer](#) 来审计 IAM 凭证和权限。您可以使用 [Amazon CloudWatch](#) 为 AWS 环境中调用的特定 API 调用设置警报。[Amazon GuardDuty](#) 还可以提醒您注意意外活动，出现这种提醒，可表明对 IAM 凭证的访问过于宽松，或出现了意外访问情况。

- 定期轮换凭证：当您无法使用临时凭证时，请定期轮换长期 IAM 访问密钥（最多每 90 天一次）。如果在您不知情的情况下无意中泄露了访问密钥，这将限制凭证用于访问资源的时间。有关轮换 IAM 用户的访问密钥的信息，请参阅[轮换访问密钥](#)。
- 审核 IAM 权限：为了提高您的 AWS 账户的安全性，请定期审核和监控每个 IAM 策略。验证这些策略是否遵循最低权限原则。
- 考虑自动创建和更新 IAM 资源：IAM Identity Center 会自动执行许多 IAM 任务，比如角色和策略管理。或者，AWS CloudFormation 可用于自动部署 IAM 资源（包括角色和策略），以减少人为错误的机会，因为可以验证模板和控制版本。
- 对于机器身份，使用 IAM Roles Anywhere 替换 IAM 用户：IAM Roles Anywhere 将使您能够在传统上无法使用角色的领域（例如本地服务器）使用角色。IAM Roles Anywhere 使用可信的 X.509 证书向 AWS 进行身份验证并接收临时凭证。使用 IAM Roles Anywhere 便无需轮换这些凭证，因为长期凭证不再存储在本地环境中。请注意，您需要监控 X.509 证书，并在该证书即将到期时轮换它。

## 资源

### 相关最佳实践：

- [SEC02-BP02 使用临时凭证](#)
- [SEC02-BP03 安全地存储和使用密钥](#)

### 相关文档：

- [开始使用 AWS Secrets Manager](#)
- [IAM 最佳实践](#)
- [身份提供者和联合身份验证](#)
- [安全合作伙伴解决方案：访问和访问控制](#)
- [临时安全凭证](#)
- [获取 AWS 账户 的凭证报告](#)

### 相关视频：

- [有关大规模管理、检索和轮换密钥的最佳实践](#)
- [使用 AWS IAM Identity Center 大规模管理用户权限](#)
- [在每个层面掌握身份](#)

相关示例：

- [Well-Architected 实验室 - 自动化 IAM 用户清理](#)
- [Well-Architected 实验室 - 自动部署 IAM 组和角色](#)

## SEC02-BP06 利用用户组和属性

随着您管理的用户数量不断增加，您需要确定如何组织这些用户，以便能够实现规模管理。将具有常见安全要求的用户置于由您的身份提供程序定义的组中，并建立机制以确保用于访问控制的用户属性（例如部门或位置）正确无误且已更新。使用这些组和属性（而不是单个用户）来控制访问权限。这样，您就可以通过使用 [权限集](#) 一次性更改用户的组成员身份或属性来集中管理访问，而不是在需要更改用户的访问权限时更新多个单独策略。您可以使用 AWS IAM Identity Center（IAM Identity Center）来管理用户组和属性。IAM Identity Center 支持最常用的属性，无论是在创建用户时手动输入的属性还是使用同步引擎自动预置的属性，例如跨域身份管理系统（SCIM，Cross-Domain Identity Management）规范中定义的那些属性。

将具有常见安全要求的用户置于由您的身份提供程序定义的组中，并建立机制以确保用于访问控制的用户属性（例如部门或位置）正确无误且已更新。使用这些组和属性（而不是单个用户）来控制访问。这使您可以通过一次性更改用户的组成员身份或属性来集中管理访问，而不是在用户的访问需要更改时更新多个单独策略。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

- 如果您在使用 AWS IAM Identity Center（IAM Identity Center）配置组：IAM Identity Center 使您能够配置用户组，并为组分配所需的权限级别。
  - [AWS Single Sign-On – 管理身份](#)
- 了解基于属性的访问控制（ABAC，Attribute-Based Access Control）：ABAC 是一种基于属性定义权限的授权策略。
  - [什么是适用于 AWS 的 ABAC？](#)
  - [实验室：基于 IAM 标签的 EC2 访问控制](#)

## 资源

### 相关文档：

- [开始使用 AWS Secrets Manager](#)
- [IAM 最佳实践](#)
- [身份提供程序和联合](#)
- [AWS 账户根用户](#)

### 相关视频：

- [有关大规模管理、检索和轮换密钥的最佳实践](#)
- [使用 AWS IAM Identity Center 大规模管理用户权限](#)
- [在每个层面掌握身份](#)

### 相关示例：

- [实验室：基于 IAM 标签的 EC2 访问控制](#)

## SEC 3.如何管理人员和机器的权限？

管理权限以控制对需要访问 AWS 和您的工作负载的人员和机器身份的访问。权限用于控制哪些人可以在什么条件下访问哪些内容。

### 最佳实践

- [SEC03-BP01 定义访问要求](#)
- [SEC03-BP02 授予最低访问权限](#)
- [SEC03-BP03 建立紧急访问流程](#)
- [SEC03-BP04 持续减少权限](#)
- [SEC03-BP05 为您的组织定义权限防护机制](#)
- [SEC03-BP06 基于生命周期管理访问权限](#)
- [SEC03-BP07 分析公共和跨账户访问](#)
- [SEC03-BP08 在组织内安全地共享资源](#)
- [SEC03-BP09 与第三方安全地共享资源](#)



## SEC03-BP01 定义访问要求

管理员、最终用户或其他组件都需要访问您工作负载的每个组件或资源。明确定义哪些人员或事物应当有权访问每个组件，选择用于进行身份验证和授权的适当身份类型和方法。

常见反模式：

- 在应用程序中进行硬编码或存储密码。
- 向每个用户授予自定义权限。
- 使用长期有效的凭证。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

管理员、最终用户或其他组件都需要访问您工作负载的每个组件或资源。明确定义哪些人员或事物应当有权访问每个组件，选择用于进行身份验证和授权的适当身份类型和方法。

应提供对组织内 AWS 账户 的常规访问，方法是使用 [联合身份访问](#) 或集中式身份提供者。您还应将身份管理集中处理，确保对于 AWS 将访问集成到员工访问生命周期中已建立了既定做法。例如，当员工转岗到具有不同访问级别的职位时，该员工的小组成员资格也应进行更改以反映新的访问要求。

在定义非人类身份的访问要求时，请确定哪些应用程序和组件需要访问权限以及如何向其授予权限。建议使用通过最低权限访问模型构建的 IAM 角色。[AWS 托管策略](#) 提供了预定义的 IAM 策略，这些策略涵盖了大多数常见使用案例。

AWS 服务（例如 [AWS Secrets Manager](#) 和 [AWS Systems Manager Parameter Store](#)）可以帮助在无法使用 IAM 角色的情况下，安全地将密码与应用程序或工作负载分离。在 Secrets Manager 中，您可以为凭证建立自动轮换。您可以通过 Systems Manager 使用您在创建参数时指定的唯一名称，来引用脚本、命令、SSM 文档、配置和自动化工作流中的参数。

您可以使用 AWS Identity and Access Management Roles Anywhere [获取 IAM 中的临时安全凭证](#)，这种凭证适用于在 AWS 外部运行的工作负载。您的工作负载可以使用 [IAM 策略](#) 和 [IAM 角色](#)，也就是您为访问 AWS 资源在 AWS 应用程序中所用的策略和角色。

如果可能，请优先选择短期临时凭证而不是长期静态凭证。在一些场景中，需要具有编程访问权限和长期凭证的 IAM 用户，此时请使用 [访问密钥上次使用的信息](#) 来轮换和删除访问密钥。

### 资源

相关文档：

- [基于属性的访问控制 \( ABAC \)](#)
- [AWS IAM Identity Center](#)
- [IAM Roles Anywhere](#)
- [适用于 IAM Identity Center 的 AWS 托管策略](#)
- [AWS IAM 策略条件](#)
- [IAM 使用案例](#)
- [删除不必要的凭证](#)
- [策略的使用](#)
- [如何根据 AWS 账户、OU 或组织来控制对 AWS 资源的访问](#)
- [使用 AWS Secrets Manager 中的增强搜索来轻松标识、安排和管理密钥](#)

相关视频：

- [在 60 分钟以内成为 IAM 策略高手](#)
- [职责分离、最低权限、委托和 CI/CD](#)
- [简化身份和访问管理以实施创新](#)

## SEC03-BP02 授予最低访问权限

最佳实践是向身份授予的访问权限只能是在特定条件下对特定资源执行特定操作所必需的。使用组和身份属性来大规模动态设置权限，而不是为单个用户定义权限。例如，您可以允许一组开发人员访问，以便仅管理其项目的资源。使用这种方法，如果某个开发人员离开项目，则可以自动撤销该开发人员的访问权限，而无需更改底层访问策略。

期望结果：用户应仅具有完成工作所必需的权限。应该仅向用户授予访问生产环境以在有限的时间段内执行特定任务的权限，并且在任务完成后，应撤销访问权限。当不再需要权限时（包括当用户转到其他项目或工作职能时），应撤销权限。管理员权限应该只授予一小部分受信任的管理员。应定期审查权限，以避免权限蔓延。应该为机器或系统账户授予一组完全其任务所需的最低权限。

常见反模式：

- 默认为向用户授予管理员权限。
- 使用根用户进行日常活动。
- 创建过于宽松但没有完全管理员权限的策略。

- 不审查权限以了解是否为它们授予了最低访问权限。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

[最低权限](#)原则指出，应仅允许身份执行完成特定任务所需的一组最少操作。这样使可用性、效率 and 安全性达到平衡。根据此原则运营，有助于限制意外访问，还有助于跟踪谁有权访问哪些资源。默认情况下，IAM 用户和角色没有任何权限。根用户默认具有完全访问权限，所以应该严格控制和监控根用户，并且只用于[需要根访问权限的任务](#)。

IAM 策略用于显式地为 IAM 角色或特定资源授予权限。例如，基于身份的策略可以附加到 IAM 组，而 S3 存储桶可由基于资源的策略控制。

创建 IAM 策略时，您可以指定服务操作、资源以及为使 AWS 允许或拒绝访问而必须满足的条件。AWS 支持多种条件以帮助您缩小访问权限范围。例如，通过使用 PrincipalOrgID [条件键](#)，如果请求者不属于 AWS Organization，则您可以拒绝操作。

您还可以控制 AWS 服务代表您发出的请求，例如要求 AWS CloudFormation 使用 CalledVia 条件键创建 AWS Lambda 函数。您应该对不同的策略类型进行分层，以建立深度防御并限制用户的总体权限。您还可以限制可以授予哪些权限以及在什么条件下授予权限。例如，您可以允许应用程序团队为他们构建的系统创建他们自己的 IAM 策略，但还必须应用[权限边界](#)来限制系统可以接收的最大权限。

## 实施步骤

- 实施最低权限策略：向 IAM 组和角色分配具有最低权限的访问策略，以反映所定义的用户角色或职能。
  - 将策略基于 API 使用情况：确定所需权限的一种方法是查看 AWS CloudTrail 日志。这样就使您可以根据用户在 AWS 内实际执行的操作来创建权限。[IAM Access Analyzer 会自动基于活动生成 IAM 策略](#)。您可以在组织或账户级别使用 IAM Access Advisor 来[跟踪特定策略的上次访问信息](#)。
- 考虑使用[针对工作职能的 AWS 托管策略](#)。开始创建细粒度权限策略时，很难知道从哪里开始。AWS 拥有针对常见工作角色（例如计费、数据库管理员和数据科学家）的托管策略。这些策略可以帮助缩减用户具备的访问权限，同时确定如何实施最低权限策略。
- 删除不必要的权限：删除不需要的权限，并削减过于宽松的策略。[IAM Access Analyzer 策略生成](#)可帮助微调权限策略。
- 确保用户对生产环境仅具有有限的访问权限：用户应该只能访问具有有效使用案例的生产环境。在用户执行需要生产访问权限的特定任务后，应撤销访问权限。限制对生产环境的访问可帮助防止发生影响生产的意外事件，并缩小意外访问的影响范围。

- 考虑使用权限边界：权限边界是一项功能，它使用托管策略设置基于身份的策略可向 IAM 实体授予的最高权限。实体的权限边界仅允许实体执行其基于身份的策略及其权限边界都允许的操作。
- 考虑权限的[资源标签](#)：借助使用资源标签且基于属性的访问控制模型，您可以根据资源用途、所有者、环境或其他条件授予访问权限。例如，您可以使用资源标签来区分开发环境和生产环境。您可以使用这些标签将开发人员限制在开发环境中。通过将标记与权限策略结合在一起，您可以实现细粒度的资源访问，而无需为每个工作职能定义复杂的自定义策略。
- 为 AWS Organizations 使用[服务控制策略](#)。服务控制策略集中控制组织中成员账户的最大可用权限。重要的是，您可以使用服务控制策略来限制成员账户中的根用户权限。还要考虑使用 AWS Control Tower，它提供可充实 AWS Organizations 的规范性托管控制。您还可以在 Control Tower 内定义自己的控制。
- 为组织建立用户生命周期策略：用户生命周期策略定义了当用户加入 AWS、更改工作角色或范围，或不再需要访问 AWS 时需要执行的任务。应在用户生命周期的每个步骤中执行权限审查，以确认权限受到适当限制并避免权限蔓延。
- 确立定期的时间表来审查权限并删除任何不需要的权限：您应定期审查用户访问权限，以确认用户不具有过于宽松的访问权限。在审核用户权限时 [AWS Config](#) 和 IAM Access Analyzer 可以提供帮助。
- 确立工作角色矩阵：工作角色矩阵形象地展示 AWS 业务覆盖范围内所需的各种角色和访问级别。使用工作角色矩阵，您可以根据用户在组织内的职责来定义和分离权限。使用组而不是将权限直接应用于单个用户或角色。

## 资源

### 相关文档：

- [授予最低权限](#)
- [IAM 实体的权限边界](#)
- [用于编写最低权限 IAM 策略的方法](#)
- [通过基于访问活动生成 IAM 策略，IAM Access Analyzer 可让您更轻松实施最低权限](#)
- [使用 IAM 权限边界将权限管理委派给开发人员](#)
- [使用上次访问的信息来细化权限](#)
- [IAM 策略类型及其使用时间](#)
- [使用 IAM 策略模拟器测试 IAM 策略](#)
- [AWS Control Tower 中的防护机制](#)
- [零信任架构：AWS 视角](#)
- [如何使用 CloudFormation StackSets 实施最低权限原则](#)

- [基于属性的访问控制 \( ABAC \)](#)
- [通过查看用户活动缩小策略范围](#)
- [查看角色访问权限](#)
- [使用标记来整理环境和促进责任的履行](#)
- [AWS 标记策略](#)
- [标记 AWS 资源](#)

相关视频：

- [新一代权限管理](#)
- [零信任：AWS 视角](#)
- [如何使用权限边界限制用户和角色以防止权限升级？](#)

相关示例：

- [实验室：创建 IAM 权限边界委派角色](#)
- [实验室：基于 IAM 标签的 EC2 访问控制](#)

## SEC03-BP03 建立紧急访问流程

创建一个流程，便于在集中式身份提供程序偶尔出现问题时紧急访问您的工作负载。

必须针对可能导致紧急事件的不同故障模式设计流程。例如，在正常情况下，您的员工用户使用集中式身份提供程序联合到云端 ( [SEC02-BP04](#) ) 来管理他们的工作负载。但是，如果您的集中式身份提供程序出现故障，或者云中联合身份验证的配置被修改，则您的员工用户可能无法联合到云中。紧急访问流程允许授权管理员通过其他方式 ( 例如其他联合形式或直接用户访问 ) 访问云资源，以解决联合配置或工作负载的问题。在恢复正常的联合机制之前，将使用紧急访问流程。

期望的结果：

- 您已经定义并记录了算是紧急情况的故障模式：考虑您的正常情况以及用户管理其工作负载所依赖的系统。考虑这些依赖项中的每一个在哪些情形下会发生故障并导致紧急情况。您可能会发现，[可靠性支柱](#)中的问题和最佳实践有助于识别故障模式和构建更具韧性的系统，从而最大限度地降低发生故障的可能性。
- 您已记录了将故障确认为紧急情况所必须遵循的步骤。例如，您可以要求身份管理员检查主身份提供程序和备用身份提供程序的状态，如果两者均不可用，则宣布身份提供程序故障为紧急事件。

- 您已针对每种紧急情况或故障模式定义了紧急访问流程。应尽可能明确具体，这样可减少用户针对所有类型的紧急情况过度使用通用流程的倾向。紧急访问流程描述了每个流程的使用情形，以及哪些情况下不应使用该流程，并指出了可能适用的替代流程。
- 您的流程有详细的说明和行动手册，便于快速有效地遵循。请记住，对用户来说，紧急事件可能让人很煎熬，他们可能面临极大的时间压力，因此流程设计应尽可能简单。

#### 常见反模式：

- 您没有详细记录并经过充分测试的紧急访问流程。您的用户没有为紧急情况做好准备，在出现紧急事件时遵循临时流程。
- 您的紧急访问流程依赖于与普通访问机制相同的系统（例如集中式身份提供程序）。这意味着，此类系统的故障可能会同时影响您的正常访问和紧急访问机制，并削弱您从故障中恢复的能力。
- 您的紧急访问流程被用于非紧急情况。例如，您的用户经常滥用紧急访问流程，因为他们发现直接进行更改比通过管道提交更改更容易。
- 您的紧急访问流程未生成足够的日志用于审核这些流程，或者没有监控日志以提醒可能存在的流程滥用。

#### 建立此最佳实践的好处：

- 通过拥有记录详实且经过充分测试的紧急访问流程，您可以减少用户响应和解决紧急事件所花费的时间。这可以缩短停机时间，提高您向客户提供的服务的可用性。
- 您可以跟踪每个紧急访问请求，检测未经授权企图对非紧急事件滥用该过程的行为，并发出警报。

#### 未建立这种最佳实践的情况下暴露的风险等级：中

#### 实施指导

本节针对与部署在 AWS 上的工作负载相关的几种故障模式，提供创建紧急访问流程的指导，首先是适用于所有故障模式的通用指导，然后是不同故障模式类型的特定指导。

#### 适用于所有故障模式的通用指南

在针对故障模式设计紧急访问流程时，请考虑以下几点：

- 记录流程的先决条件和假设：何时应使用该流程、何时不应使用该流程。它有助于详细说明故障模式并记录假设，例如其他相关系统的状态。例如，故障模式 2 的流程假设身份提供程序可用，但 AWS 上的配置已修改或已过期。



- 预先创建紧急访问流程所需的资源 ( [SEC10-BP05](#) )。例如，预先创建带有 IAM users 和角色的紧急访问 AWS 账户，并在所有工作负载账户中创建跨账户 IAM 角色。这可以验证在发生紧急事件时这些资源是否已准备就绪并且可用。通过预先创建资源，您就不必依赖 AWS [控制平面](#) API ( 用于创建和修改 AWS 资源 )，在紧急情况下，这些 API 可能不可用。此外，通过预先创建 IAM 资源，您也无需考虑 [由于最终的一致性问题的延迟](#)。
- 将紧急访问流程作为事件管理计划的一部分 ( [SEC10-BP02](#) )。记录如何跟踪紧急事件并将其传达给组织中的其他人，例如同级团队、您的领导层，如果适用，还包括外部的客户和业务合作伙伴。
- 在现有的服务请求工作流程系统 ( 如果有 ) 中定义紧急访问请求流程。通常，此类工作流程系统允许您创建受理表来收集有关请求的信息，在工作流程的每个阶段跟踪请求，并添加自动和手动审批步骤。将每个请求与事件管理系统中所跟踪的相应紧急事件关联起来。通过拥有统一的紧急访问系统，您可以在单个系统中跟踪这些请求，分析使用趋势并改进流程。
- 确保您的紧急访问流程只能由授权用户启动，并且需要用户的同事或管理层 ( 视情况而定 ) 的批准。审批流程在工作时间内和工作时间之外应该能够有效运作。明确在主审批人抽不开身的情况下，如何允许辅助审批人审批请求，并沿管理链条上报，直至获得批准。
- 验证流程是否会针对成功和失败的紧急访问尝试，生成详细的审计日志和事件。监控请求流程和紧急访问机制，以检测滥用或未经授权的访问。将活动与事件管理系统中正在发生的紧急事件关联起来，并在预期时间段之外执行操作时发出警报。例如，您应监控紧急访问 AWS 账户中的活动并发出警报，因为在正常操作过程中绝不应使用该账户。
- 定期测试紧急访问流程，以确保步骤清楚明了，并快速高效地授予正确的访问级别。您的紧急访问流程应作为事件响应模拟 ( [SEC10-BP07](#) ) 和灾难恢复测试 ( [REL13-BP03](#) ) 的一部分进行测试。

#### 故障模式 1：用于联合到 AWS 的身份提供程序不可用

如 [SEC02-BP04 依赖集中式身份提供程序](#) 中所述，我们建议依靠集中式身份提供程序，来联合您的员工用户以授予对 AWS 账户的访问权限。您可以使用 IAM Identity Center 联合到 AWS 组织中的多个 AWS 账户，也可以使用 IAM 将联合到单个 AWS 账户。在这两种情况下，员工用户都要先通过集中式身份提供程序进行身份验证，然后才会被重定向到 AWS 登录端点进行单点登录。

万一集中式身份提供程序不可用，员工用户就无法联合到 AWS 账户或管理其工作负载。在这种紧急情况下，您可以为一些小部分管理员提供紧急访问流程，让他们访问 AWS 账户，来执行等不及集中式身份提供程序恢复正常的任务。例如，您的身份提供程序在 4 小时内不可用，在此期间，您需要修改生产账户中 Amazon EC2 Auto Scaling 组的上限，以应对客户流量意外激增的情况。您的紧急状况管理员应遵循紧急访问流程，以获得对特定生产 AWS 账户的访问权限并进行必要的更改。

紧急访问流程依赖于预先创建的紧急访问 AWS 账户，该账户仅用于紧急访问，并拥有 AWS 资源 ( 如 IAM 角色和 IAM users ) 以支持紧急访问流程。在正常运营期间，任何人都不得访问紧急访问账户，而且您必须对滥用该账户的行为进行监控并发出警报 ( 有关更多详情，请参阅前面的“通用指南”部分 )。



紧急访问账户具有紧急访问 IAM 角色，有权在需要紧急访问的 AWS 账户中代入跨账户角色。这些 IAM 角色是预先创建的，并配置有信任策略，可信任应急账户的 IAM 角色。

紧急访问过程可以使用以下方法之一：

- 您可以在紧急访问账户中为紧急状况管理员预先创建一组 [IAM users](#)，并使用相关的强密码和 MFA 令牌。这些 IAM users 有权代入 IAM 角色，然后在需要紧急访问时，允许跨账户访问 AWS 账户。我们建议尽可能少地创建此类用户，并将每个用户分配给一个紧急状况管理员。在紧急情况下，紧急状况管理员用户使用其密码和 MFA 令牌码登录紧急访问账户，切换到紧急账户中的紧急访问 IAM 角色，最后切换到工作负载账户中的紧急访问 IAM 角色，以执行紧急更改操作。这种方法的优点是，每个 IAM user 都分配给一个紧急状况管理员，您可以通过查看 CloudTrail 事件来了解哪个用户已登录。缺点是，您必须维护多个 IAM users 及其关联的长寿命密码和 MFA 令牌。
- 您可以使用紧急访问 [AWS 账户根用户](#) 来登录紧急访问账户，代入用于紧急访问的 IAM 角色，并代入工作负载账户中的跨账户角色。建议为根用户设置一个强密码和多个 MFA 令牌。我们还建议将密码和 MFA 令牌存储在安全的企业凭证保管库中，该保管库可执行强身份验证和授权。您应确保密码和 MFA 令牌重置因素的安全：将账户的电子邮件地址设置为由云安全管理员监控的电子邮件分发列表，将账户的电话号码设置为同样由安全管理员监控的共享电话号码。这种方法的优点是只需管理一组根用户凭证。缺点是，由于这是共享用户，多个管理员都能以根用户身份登录。您必须审计企业保管库日志事件，以确定是哪位管理员查看了根用户密码。

## 故障模式 2：AWS 上的身份提供程序配置已修改或已过期

要允许您的员工用户联合到 AWS 账户，您可以使用外部身份提供程序配置 IAM Identity Center，或创建 IAM 身份提供程序（[SEC02-BP04](#)）。通常，您需要通过导入身份提供程序提供的 SAML 元数据 XML 文档，来配置这些服务。元数据 XML 文档包含一个 X.509 证书，该证书对应于身份提供程序用来签署其 SAML 断言的私钥。

管理员可能会错误地修改或删除 AWS 端的这些配置。在另一种情形下，导入到 AWS 的 X.509 证书可能会过期，而带有新证书的新元数据 XML 尚未导入到 AWS。这两种情形都可能使您的员工用户无法联合到 AWS，从而出现紧急情况。

在这种紧急情况下，您可以向您的身份管理员提供对 AWS 的访问权限以修复联合问题。例如，身份管理员使用紧急访问流程登录紧急访问 AWS 账户，切换到 Identity Center 管理员账户中的角色，并通过从身份提供程序导入最新的 SAML 元数据 XML 文档来更新外部身份提供程序配置，从而重新启用联合。修复联合后，您的员工用户将继续使用正常操作流程联合到其工作负载账户。

您可以按照前面的故障模式 1 中详述的方法来创建紧急访问流程。您可以向您的身份管理员授予最低访问权限，使其只能访问 Identity Center 管理员账户，并使用该账户对 Identity Center 执行操作。

## 故障模式 3：Identity Center 中断

如果发生 IAM Identity Center 或 AWS 区域中断这样的小概率事件，我们建议您设置一个可用于临时访问 AWS Management Console 的配置。

紧急访问流程使用从身份提供程序到您的紧急账户中的 IAM 的直接联合。有关流程和设计注意事项的详细信息，请参阅 [Set up emergency access to the AWS Management Console](#) 访问 AWS 资源。

### 实施步骤

#### 针对所有故障模式的通用步骤

- 创建专门用于紧急访问流程的 AWS 账户。预先创建账户中所需的 IAM 资源，例如 IAM 角色或 IAM users，以及可选的 IAM 身份提供程序。此外，在工作负载 AWS 账户中预先创建跨账户 IAM 角色，并与紧急访问账户中的相应 IAM 角色建立信任关系。您可以将 [AWS CloudFormation StackSets](#) 与 [AWS Organizations](#) 结合使用，在您组织的成员账户中创建此类资源。
- 创建 AWS Organizations [服务控制策略](#)（SCP），以拒绝删除和修改成员 AWS 账户中的跨账户 IAM 角色。
- 对紧急访问 AWS 账户启用 CloudTrail，并将跟踪事件发送到日志收集 AWS 账户中的中央 S3 存储桶。如果您使用 AWS Control Tower 来设置和管理您的 AWS 多账户环境，则您使用 AWS Control Tower 创建或在 AWS Control Tower 中注册的每个账户默认情况下都已启用 CloudTrail，并发送到专用日志存档 AWS 账户中的 S3 存储桶。
- 通过创建 EventBridge 规则来匹配紧急 IAM 角色所执行的控制台登录和 API 活动，监控紧急访问账户的活动。当事件管理系统中所跟踪的正在发生的紧急事件之外出现活动时，向安全运营中心发送通知。

针对“故障模式 1：用于联合到 AWS 的身份提供程序不可用”和“故障模式 2：AWS 上的身份提供程序配置已修改或已过期”的其他步骤

- 根据您的选择的紧急访问机制，预先创建资源：
  - 使用 IAM users：使用强密码和关联的 MFA 设备预先创建 IAM users。
  - 使用紧急账户的根用户：为根用户配置一个强密码，并将该密码存储在您的企业凭证库中。将多个物理 MFA 设备与根用户关联，并将设备存放在紧急状况管理员团队成员可以快速访问的位置。

针对“故障模式 3：Identity Center 中断”的其他步骤

- 如 [Set up emergency access to the AWS Management Console](#) 中所详述的那样，在紧急访问 AWS 账户中，创建 IAM 身份提供程序，以启用从身份提供程序的直接 SAML 联合。
- 在 IdP 中创建没有成员的紧急行动组。
- 在紧急访问账户中创建与紧急行动组相对应的 IAM 角色。

## 资源

相关的 Well-Architected 最佳实践：

- [SEC02-BP04 依赖集中式身份提供程序](#)
- [SEC03-BP02 授予最低访问权限](#)
- [SEC10-BP02 制定事件管理计划](#)
- [SEC10-BP07 执行实际演练](#)

相关文档：

- [Set up emergency access to the AWS Management Console](#)
- [Enabling SAML 2.0 federated users to access the AWS Management Console](#)
- [Break glass access](#)

相关视频：

- [AWS re:Invent 2022 - Simplify your existing workforce access with IAM Identity Center](#)
- [AWS re:Inforce 2022 - AWS Identity and Access Management \( IAM \) 深入探讨](#)

相关示例：

- [AWS Break Glass 角色](#)
- [AWS 客户行动手册框架](#)
- [AWS 事件响应行动手册样本](#)

## SEC03-BP04 持续减少权限

当您的团队确定好所需的访问权限时，删除不需要的权限，并建立审核流程以实现最低权限。持续监控并删除供人类和机器访问的未使用的身份和权限。

期望结果：权限策略应遵循最低权限原则。随着工作职责和角色变得更加明确，需要审查您的权限策略以删除不必要的权限。如果无意中泄露或未经授权访问凭证，这种方法会缩小影响范围。

常见反模式：

- 默认为向用户授予管理员权限。
- 创建过于宽松但没有完全管理员权限的策略。
- 保留不再需要的权限策略。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

当团队和项目刚刚起步时，可以使用宽松的权限策略来激发创新并提高敏捷性。例如，在开发或测试环境中，开发人员可以获得广泛的访问权限以使用各种 AWS 服务。我们建议您持续评估访问权限，并仅限于访问完成当前作业所必需的服务和服务操作。对于人类和机器身份，均建议进行此项评估。机器身份有时称为系统或服务账户，是让 AWS 访问应用程序或服务器的身份。这种访问权限在生产环境中尤其重要，因为在该环境中，过于宽松的权限会产生广泛的影响，并可能暴露客户数据。

AWS 提供多种方法来帮助识别未使用的用户、角色、权限和凭证。AWS 还可帮助分析 IAM 用户和角色（包括关联的访问密钥）的访问活动，以及对 AWS 资源（如 Amazon S3 存储桶中的对象）的访问。AWS Identity and Access Management Access Analyzer 策略生成可帮助您根据主体与之交互的实际服务和操作来创建限制性权限策略。[基于属性的访问控制 \(ABAC\)](#) 可帮助简化权限管理，因为您可以使用用户的属性为用户提供权限，而不是将权限策略直接附加到每个用户。

## 实施步骤

- 使用 [AWS Identity and Access Management Access Analyzer](#)：IAM Access Analyzer 可帮助识别您组织和账户中[与外部实体共享](#)的资源，例如 Amazon Simple Storage Service (Amazon S3) 存储桶或 IAM 角色。
- 使用 [IAM Access Analyzer 策略生成](#)：IAM Access Analyzer 策略生成可帮助您[基于 IAM 用户或角色的访问活动创建精细的权限策略](#)。
- 为 IAM 用户和角色确定可接受的时间框架和使用策略：使用[上次访问时间戳](#)来[识别未使用的用户和角色](#)并将它们移除。查看关于服务和操作的上次访问情况的信息，并[确定特定用户和角色的权限范围](#)。例如，您可以使用关于上次访问情况的信息，确定您的应用程序角色需要执行的特定 Amazon S3 操作，并只允许该角色访问这些操作。AWS Management Console 中提供了上次获取的信息，您也可以对这些功能进行编程，以便将它们整合到您的基础设施工作流程和自动化工具中。

- 考虑将数据事件录入 [AWS CloudTrail](#)：默认情况下，CloudTrail 不会记录数据事件，例如 Amazon S3 对象级活动（如 GetObject 和 DeleteObject）或 Amazon DynamoDB 表活动（如 PutItem 和 DeleteItem）。考虑为这些事件启用日志记录，以确定哪些用户和角色需要访问特定的 Amazon S3 对象或 DynamoDB 表项目。

## 资源

### 相关文档：

- [授予最低特权](#)
- [删除不必要的凭证](#)
- [什么是 AWS CloudTrail？](#)
- [策略的使用](#)
- [日志记录和监控 DynamoDB](#)
- [为 Amazon S3 存储桶和对象启用 CloudTrail 事件日志记录](#)
- [获取 AWS 账户的凭证报告](#)

### 相关视频：

- [在 60 分钟以内成为 IAM 策略高手](#)
- [职责分离、最低权限、委托和 CI/CD](#)
- [AWS re:Inforce 2022 - AWS Identity and Access Management \( IAM \) 深入探讨](#)

## SEC03-BP05 为您的组织定义权限防护机制

建立通用控件以限制对组织中所有身份的访问。例如，您可以限制对特定 AWS 区域的访问，或防止操作员删除通用资源，例如用于您的核心安全团队的 IAM 角色。

### 常见反模式：

- 在组织管理员账户中运行工作负载。
- 在同一账户中运行生产工作负载和非生产工作负载。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

当您在 AWS 中的工作负载增多并管理这些额外的工作负载时，您应使用账户分离这些工作负载，并使用 AWS Organizations 管理这些账户。我们建议您建立常用权限防护机制，以限制您所在组织中的所有身份的访问权限。例如，您可以限制对特定 AWS 区域的访问，或防止您的团队删除常见资源，例如您的核心安全团队使用的 IAM 角色。

您可以首先实施示例服务控制策略，例如禁止用户禁用密钥服务。SCP 使用 IAM 策略语言，并允许您建立所有 IAM 主体（用户和角色）都要遵循的控制机制。您可以限制对特定服务操作和资源的访问，并根据特定的条件限制访问，以满足您所在组织的访问控制需求。如有必要，您可以为您的防护机制定义异常情况。例如，您可以为账户中除特定管理员角色以外的所有 IAM 实体限制服务操作。

我们建议您避免在管理账户中运行工作负载。应该使用管理账户来管理和部署将影响成员账户的安全防护机制。一些 AWS 服务支持使用委派管理员账户。在可能的情况下，您应使用此委派账户，而不是使用管理账户。您应严格限制对组织管理员账户的访问。

通过使用多账户策略，您可以更灵活地将防护机制应用于工作负载。AWS Security Reference Architecture 提供了有关如何设计账户结构的规范性指南。AWS Control Tower 等 AWS 服务提供了一些功能，可集中管理整个组织内的预防性和检测性控制机制。为组织中的每个账户或 OU 定义明确的用途，并根据该用途限制控制机制。

## 资源

相关文档：

- [AWS Organizations](#)
- [服务控制策略 \( SCP , Service Control Policy \)](#)
- [在多账户环境中充分利用服务控制策略](#)
- [AWS Security Reference Architecture \( AWS SRA \)](#)

相关视频：

- [使用服务控制策略实施预防性防护机制](#)
- [使用 AWS Control Tower 实施大规模管理](#)
- [AWS Identity and Access Management 深入探讨](#)



## SEC03-BP06 基于生命周期管理访问权限

将访问控制措施与操作员和应用程序生命周期以及您的集中联合身份提供者集成。例如，在用户离开组织或角色发生变化时删除用户的访问权限。

当您使用不同的账户管理工作负载时，您有时需要在这些账户之间共享资源。我们建议您使用 [AWS Resource Access Manager \(AWS RAM\) 来共享资源](#)。使用此服务，您可以轻松、安全地在您的 AWS Organizations 和组织部门内共享 AWS 资源。使用 AWS RAM，当账户移进和移出与之共享资源的组织或组织部门时，会自动授予或撤销对共享资源的访问权限。这样有助于您确保只与您的目标账户共享资源。

未建立此最佳实践暴露的风险等级：低

### 实施指导

用户访问生命周期：针对加入的人员、工作职能变更和离开的人员实施用户访问生命周期策略，以确保只有在职用户具有访问权限。

### 资源

相关文档：

- [基于属性的访问控制 \(ABAC\)](#)
- [授予最小特权](#)
- [IAM Access Analyzer](#)
- [删除不必要的凭证](#)
- [策略的使用](#)

相关视频：

- [在最多 60 分钟的时间内成为 IAM 策略高手](#)
- [职责分离、最低权限、委托和 CI/CD](#)

## SEC03-BP07 分析公共和跨账户访问

对于标识出存在公共访问和跨账户访问情况的调查结果，应持续监控。减少公共访问和跨账户访问，使访问仅能触达特定资源。

期望结果：了解您的 AWS 资源中哪些是共享的，以及与谁共享。持续监控和审计您的共享资源，以验证它们仅与授权的主体共享。



## 常见反模式：

- 不保留共享资源的清单。
- 跨账户访问或公开访问资源时，没有遵循流程。

在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

如果您的账户在 AWS Organizations 中，您可以向整个组织、特定组织单位或个人账户授予资源访问权限。如果您的账户不是某个组织的成员，您可以与个人账户共享资源。您可以使用基于资源的策略（例如 [Amazon Simple Storage Service \( Amazon S3 \) 存储桶策略](#)）授予直接跨账户访问权限，也可以允许另一账户中的主体代入您账户中的 IAM 角色来授予该权限。使用资源策略时，请验证访问权限是否仅授予给经过授权的主体。建立一个流程来审批所有需要可公开访问的资源。

[AWS Identity and Access Management Access Analyzer](#) 使用 [可证明的安全性](#) 来标识从账户的外部访问某个资源时的所有访问路径。它持续审核资源策略，并报告公开访问和跨账户访问的调查结果，以使您能够轻松分析可能非常宽泛的访问权限。不妨考虑配置 IAM Access Analyzer 与 AWS Organizations，来验证您是否能够查看所有账户。IAM Access Analyzer 也使得您能够先[预览调查结果](#)，然后再部署资源权限。这样，您便可以证实更改策略之后，只有您所希望的对象能够授权通过公共和跨账户访问方式触达您的资源。在设计多账户访问权限时，您可以使用[信任策略](#)来控制何种情况下允许代入某个角色。例如，您可以使用 [PrincipalOrgId 条件键来拒绝从 AWS Organizations 之外代入角色的尝试](#)。

[AWS Config 可以报告](#) 资源配置错误的情况，并且通过 AWS Config 策略检查，可以检测有何资源配置了公共访问权限。[AWS Control Tower](#) 和 [AWS Security Hub](#) 等服务简化了跨 AWS Organizations 部署检测性控制和防护机制的流程，可以识别并修复资源公开暴露的情况。例如，AWS Control Tower 具有托管防护机制，可以检测是否有任何 [Amazon EBS 快照可由 AWS 账户恢复](#)。

## 实施步骤

- 考虑为 AWS Organizations 启用 [AWS Config](#)：AWS Config 使得您能够将 AWS Organizations 内多个账户的调查结果聚合到一个委派的管理员账户。这将给您提供全局视角，进行[跨账户部署 AWS Config 规则](#)，以检测可公开访问的资源。
- 配置 AWS Identity and Access Management Access Analyzer IAM Access Analyzer 可帮助您识别组织和账户中[与外部实体共享](#)的资源，例如 Amazon S3 存储桶或 IAM 角色。
- 在 AWS Config 中使用自动修复来响应 Amazon S3 存储桶公共访问配置的变更情况：[您可以自动重新启用 Amazon S3 存储桶阻止公共访问的设置](#)。

- 实施监控和警报，以确定 Amazon S3 存储桶是否已变得能够公开访问：您必须设置[监控和警报](#)，以确定何时禁用 Amazon S3 屏蔽公共访问权限，以及 Amazon S3 存储桶是否已变得能够公开访问。此外，如果您使用 AWS Organizations，则可以创建一个[服务控制策略](#)来防止更改 Amazon S3 公共访问策略。AWS Trusted Advisor 检查是否存在具有开放访问权限的 Amazon S3 存储桶。如果向每个人授予“上传/删除”权限，那么任何人都可以向存储桶添加项目或者修改或删除存储桶中的项目，这样会产生潜在的安全问题。Trusted Advisor 可以检查存储桶明确拥有哪些权限，以及是否存在可能能够覆写这些权限的相关存储桶策略。您也可以使用 AWS Config 来监控 Amazon S3 存储桶是否具有公共访问权限。有关更多信息，请参阅[如何使用 AWS Config 监控 Amazon S3 存储桶允许公共访问的情况](#)并作出响应。检查访问权限时，重要的是要考虑 Amazon S3 存储桶中包含哪些类型的数据。[Amazon Macie](#) 有助发现和保护敏感数据，比如 PII、PHI 和凭证（如私有密钥或 AWS 密钥）。

## 资源

### 相关文档：

- [使用 AWS Identity and Access Management Access Analyzer](#)
- [AWS Control Tower 控制机制库](#)
- [AWS 基础安全最佳实践标准](#)
- [AWS Config 托管规则](#)
- [AWS Trusted Advisor 检查参考](#)
- [用 Amazon EventBridge 监控 AWS Trusted Advisor 检查结果](#)
- 对横跨组织内部所有账户的规则[进行管理 AWS Config](#)
- [AWS Config 和 AWS Organizations](#)

### 相关视频：

- [保护多账户环境的最佳实践](#)
- [深入探究 IAM Access Analyzer](#)

## SEC03-BP08 在组织内安全地共享资源

随着工作负载数量的增长，您可能需要将这些工作负载中资源的访问权限进行共享，或者跨多个账户多次预置资源。您可能需要进行构造来划分环境，例如划分成开发、测试和生产环境。但是，采取相互

分离的构造并不会限制您安全共享权限。通过共享重叠的组件，您可以降低运维开销，并提供一致的体验，而不必猜测在多次创建同一资源时可能遗漏了什么。

期望结果：通过使用安全的方法在组织内共享资源，尽可能地减少意外访问，并帮助实施数据丢失防护计划。与管理单个组件相比，降低了运维开销，减少了多次手动创建同一组件时引起的错误，并提高了工作负载的可扩展性。您可以在多点故障场景中缩短问题解决时间，并在确定何时不再需要某个组件时更有信心。有关分析外部共享资源的规范性指南，请参阅[SEC03-BP07 分析公共和跨账户访问](#)。

常见反模式：

- 缺少对意外的外部共享进行持续监控和自动发出警报的流程。
- 缺乏关于应分享什么和不应分享什么的基准。
- 默认采用广泛的开放政策，而不是在需要时明确地分享。
- 手动创建在需要时重叠的基础资源。

在未建立这种最佳实践的情况下暴露的风险等级：中等

实施指导

设计您的访问控制和模式，以安全地管理共享资源的使用，并且仅与可信实体共享。监控共享资源，持续检查共享资源访问权限，并在不适当或意外共享时发出警报。查看[分析公共和跨账户访问](#)来帮助您设置监管机制，以减少外部访问，只对需要的资源进行访问，并建立一个持续监控和自动警报的流程。

[许多 AWS 服务](#)（如 [AWS Security Hub](#)、[Amazon GuardDuty](#) 和 [AWS Backup](#)）支持 AWS Organizations 内的跨账户共享。这些服务允许将数据共享到中心账户，可从中心账户访问，或从中心账户管理资源和数据。例如，AWS Security Hub 可将调查结果从个人账户转移到中心账户，在那里您可以查看所有调查结果。AWS Backup 可以对资源进行备份并在多个账户之间共享。您可以使用 [AWS Resource Access Manager](#)（AWS RAM）来分享其他共用资源，例如 [VPC 子网](#)和 [Transit Gateway 附件](#)、[AWS Network Firewall](#) 或 [Amazon SageMaker AI 管道](#)。

要限制您的账户仅在组织内共享资源，请使用[服务控制策略（SCP）](#)阻止访问外部主体。共享资源时，请将基于身份的控制措施和网络控制措施相结合，[为您的组织创建数据边界](#)，以帮助防止意外访问。数据边界是一组预防性防护机制，用于帮助验证是否只有您的可信身份才能访问预期网络中的可信资源。这些控制措施会施加适当的限制，确定哪些资源可以共享，并防止共享或暴露不应被外泄的资源。例如，作为数据边界的一部分，您可以使用 VPC 端点策略和 AWS:PrincipalOrgId 条件来确保访问 Amazon S3 存储桶的身份属于您的组织。务必要注意，[SCP 并不适用于服务关联角色（LSR）或 AWS 服务主体](#)。

使用 Amazon S3 时，请[对您的 Amazon S3 存储桶禁用 ACL](#)，并使用 IAM 策略来定义访问控制。要[限制从 Amazon CloudFront 访问 Amazon S3 源](#)，请从来源访问身份（OAI）转为采用来源访问控制（OAC），后者支持其他功能，包括使用 [AWS Key Management Service](#) 进行服务器端加密。

在某些情况下，您可能希望允许在组织外部共享资源，或授予第三方访问您资源的权限。有关管理外部共享资源的权限的规范性指南，请参阅[权限管理](#)。

## 实施步骤

### 1. 使用 AWS Organizations。

AWS Organizations 是一项账户管理服务，可让您将多个 AWS 账户整合到您创建并集中管理的组织中。您可以将账户分组为组织单位（OU），并将不同的策略附加到每个 OU，以帮助满足预算、安全性和合规性需求。您还可以控制 AWS 人工智能（AI）和机器学习（ML）服务收集和存储数据的方式，并使用与 Organizations 集成的 AWS 服务的多账户管理。

### 2. 将 AWS Organizations 与 AWS 服务集成。

当您启用 AWS 服务以在组织的成员账户中代表您执行任务时，AWS Organizations 会在每个成员账户中为该服务创建 IAM 服务关联角色。您应使用 AWS Management Console、AWS API 或 AWS CLI 管理可信访问。有关启用可信访问的规范性指南，请参阅[将 AWS Organizations 与其他 AWS 服务结合使用](#)和[可与 Organizations 一起使用的 AWS 服务](#)。

### 3. 建立数据边界。

AWS 边界通常表示为由 AWS Organizations 管理的组织。与本地网络和系统一样，访问 AWS 资源被许多人视为 My AWS 的边界。建立边界的目标，是验证如果身份可信、资源可信并且网络符合预期，则允许访问。

#### a. 定义和实施边界。

对于每个授权条件，请遵循《在 AWS 上构建边界》白皮书的[边界实施](#)中描述的步骤。有关保护网络层的规范性指南，请参阅[保护网络](#)。

#### b. 持续监控并发出警报。

[AWS Identity and Access Management Access Analyzer](#) 可帮助识别您组织和账户中与外部实体共享的资源。您可以将 [IAM Access Analyzer 与 AWS Security Hub 集成](#)，以将资源的调查结果从 IAM Access Analyzer 发送并聚合到 Security Hub，从而帮助分析环境的安全状况。要实现集成，请在每个账户的每个“区域”中同时启用 IAM Access Analyzer 和 Security Hub。您也可以使用 AWS Config 规则来审计配置，并使用 [Amazon Q Developer in chat applications](#) 和 [AWS Security Hub](#) 向相关方发出警报。然后，您可以使用 [AWS Systems Manager Automation 文档](#) 来修复不合规的资源。

- c. 有关对外部共享资源进行持续监控并发出警报的规范性指南，请参阅[分析公共和跨账户访问](#)。
4. 在 AWS 服务中使用资源共享并进行相应限制。

许多 AWS 服务允许您与另一账户共享资源，或以另一账户中的资源为目标，比如[亚马逊云机器镜像 \(AMI\)](#) 和 [AWS Resource Access Manager \(AWS RAM\)](#)。限制 `ModifyImageAttribute` API 以指定可信账户，从而与之共享 AMI。当需要使用 AWS RAM 来将共享限制于您的组织内部时，请指定 `ram:RequestedAllowsExternalPrincipals` 条件，以帮助防止来自不可信身份的访问。有关规范性指南和注意事项，请参阅[资源共享和外部目标](#)。

5. 使用 AWS RAM 在账户中或与其他 AWS 账户 安全共享。

[AWS RAM](#) 可帮助您与账户中的角色和用户以及与其他 AWS 账户安全地共享已创建的资源。在多账户环境中，AWS RAM 使您能够一次性创建资源并与其他账户共享。这种方法有助于降低运维开销，同时通过与 Amazon CloudWatch 和 AWS CloudTrail 的集成提供一致性、可见性和可审计性，使用跨账户访问时无法获得这些好处。

如果您拥有以前使用基于资源的策略共享的资源，则可以使用 [PromoteResourceShareCreatedFromPolicy API](#) 或等效 API 将资源共享升级为完全 AWS RAM 资源共享。

在某些情况下，您可能需要采取其他步骤来共享资源。例如，要共享加密快照，您需要[共享 AWS KMS 密钥](#)。

## 资源

相关最佳实践：

- [SEC03-BP07 分析公共和跨账户访问](#)
- [SEC03-BP09 与第三方安全地共享资源](#)
- [SEC05-BP01 创建网络层](#)

相关文档：

- [存储桶所有者向并非其拥有的对象授予跨账户权限](#)
- [如何将信任策略与 IAM 结合使用](#)
- [在 AWS 上构建数据边界](#)
- [如何在向第三方授予对 AWS 资源的访问权限时使用外部 ID](#)

- [可与 AWS Organizations 一起使用的 AWS 服务](#)
- [在 AWS 上建立数据边界：仅允许可信身份获取公司数据](#)

相关视频：

- [使用 AWS Resource Access Manager 实现精细访问](#)
- [使用 VPC 端点保护您的数据边界](#)
- [在 AWS 上建立数据边界](#)

相关工具：

- [数据边界策略示例](#)

## SEC03-BP09 与第三方安全地共享资源

确保云环境安全，不能仅仅局限于保护您的组织。您的组织有一部分数据可能要依赖第三方来管理。管理第三方托管系统的权限，应遵循及时访问的做法，使用最低权限原则和临时凭证。通过与第三方密切合作，您既可以缩小影响范围，又可以降低意外访问的风险。

期望结果：只要凭证有效且处于激活状态，任何人都可以使用与用户关联的长期 AWS Identity and Access Management ( IAM ) 凭证、IAM 访问密钥和私有密钥。使用 IAM 角色和临时凭证可以减少维护长期凭证的工作量，包括这些敏感细节的管理和运维开销，从而帮助您改善总体安全状况。通过在 IAM 信任策略中对外部 ID 使用全局唯一标识符 ( UUID )，并将附加到 IAM 角色的 IAM 策略置于您的控制之下，您可以审计授予第三方的访问权限，并验证该权限不会过于宽松。有关分析外部共享资源的规范性指南，请参阅[SEC03-BP07 分析公共和跨账户访问](#)。

常见反模式：

- 采用默认的 IAM 信任策略，不附加任何条件。
- 使用长期 IAM 凭证和访问密钥。
- 重用外部 ID。

在未建立这种最佳实践的情况下暴露的风险等级：中等



## 实施指导

您可能会希望允许在 AWS Organizations 外部共享资源，或授予第三方访问您账户的权限。例如，第三方提供的监控解决方案可能会需要访问您账户内部的资源。在这些情况下，请创建 IAM 跨账户角色，并仅向该角色提供第三方所需的权限。此外，使用[外部 ID 条件](#)定义信任策略。使用外部 ID 时，您或第三方可以为每个客户、第三方或租赁生成唯一 ID。唯一 ID 创建后，不应由除您之外的任何人控制。第三方必须实施具体流程，以一种安全、可审计且可复制的方式将外部 ID 与客户关联起来。

您也可以使用 [IAM Roles Anywhere](#) 来管理 AWS 之外使用 AWS API 的应用程序的 IAM 角色。

如果第三方不再需要访问您的环境，则删除该角色。应避免向第三方提供长期凭证。保持对其他支持共享的 AWS 服务的关注。例如，AWS Well-Architected Tool 允许与其他 AWS 账户[共享工作负载](#)，[AWS Resource Access Manager](#) 可帮助您与其他账户安全共享您拥有的 AWS 资源。

## 实施步骤

### 1. 使用跨账户角色提供对外部账户的访问。

[跨账户角色](#)可减少外部账户和第三方为服务客户而存储的敏感信息量。跨账户角色允许您将账户中 AWS 资源的访问权限安全地授予第三方（如 AWS Partner 或组织内的其他账户），同时保持管理和审计该访问权限的能力。

第三方可能从混合基础设施向您提供服务，或者将数据提取到一个异地位置。[IAM Roles Anywhere](#) 可帮助您使第三方工作负载能够安全地与 AWS 工作负载交互，并进一步减少对长期凭证的需求。

不应使用长期凭证或与用户关联的访问密钥来提供外部账户访问，而应使用跨账户角色来提供跨账户访问。

### 2. 对第三方使用外部 ID。

使用[外部 ID](#)，您就可以指定谁可以在 IAM 信任策略中代入角色。信任策略可能要求代入角色的用户声明他们所处的条件和追求的目标。它还为账户所有者提供了一种方法，允许仅在特定情况下代入角色。外部 ID 的主要功能是解决和防止[混淆代理](#)问题。

如果您是 AWS 账户所有者，并且您为第三方配置了一个角色（该角色可以访问您的和其他 AWS 账户），或者当您可以代表不同的客户代入角色时，请使用外部 ID。与第三方或 AWS Partner 合作，建立一个包括在 IAM 信任策略中的外部 ID 条件。

### 3. 使用全局唯一外部 ID。

实施一个为外部 ID（例如全局唯一标识符（UUID））生成随机唯一值的流程。第三方在不同客户之间重用外部 ID 并不能解决混淆代理问题，因为客户 A 可以通过使用客户 B 的角色 ARN 以及重复的



外部 ID 来查看客户 B 的数据。在多租户环境中，第三方支持多个具有不同 AWS 账户 的客户，此时第三方必须使用不同的唯一 ID 作为每个 AWS 账户 的外部 ID。第三方负责检测重复的外部 ID，并将每个客户安全地映射到各自的外部 ID。第三方应进行测试，以验证他们只能在指定外部 ID 时代入该角色。在需要外部 ID 之前，第三方应避免存储客户角色 ARN 和外部 ID。

外部 ID 不视为密钥，但外部 ID 不能是容易猜测的值，例如电话号码、姓名或账户 ID。将外部 ID 设置为只读字段，这样就无法为了冒充设置而更改外部 ID。

您或第三方可以生成外部 ID。定义一个流程，确定谁负责生成 ID。无论创建外部 ID 的实体是什么，第三方都必须确保客户之间的唯一性和格式一致。

#### 4. 弃用客户提供的长期凭证。

弃用长期凭证，使用跨账户角色或 IAM Roles Anywhere。如果必须使用长期凭证，请制定相应计划，逐渐转变成基于角色进行访问。有关管理密钥的详细信息，请参阅[身份管理](#)。同时与 AWS 账户 团队和第三方合作，建立风险缓解运行手册。有关应对和减轻安全事件潜在影响的规范性指南，请参阅[事件响应](#)。

#### 5. 验证设置是否具有规范性指导，或是否实现了自动化。

为您账户中的跨账户访问创建的策略必须遵循[最低权限原则](#)。第三方必须为您提供使用 AWS CloudFormation 模板或等效模板的角色策略文档或自动化设置机制。这减少了手动创建策略时出错的机会，并提供了可审计的跟踪。有关使用 AWS CloudFormation 模板创建跨账户角色的更多信息，请参阅[跨账户角色](#)。

第三方应提供一个自动化的、可审计的设置机制。但是，通过使用角色策略文档（此文档大致列出了所需的访问权限），角色设置的自动化应该由您来完成。使用 AWS CloudFormation 模板或等效模板，您应将偏差检测纳入审计流程以监控变更。

#### 6. 对变更做出解释。

您的账户结构、您对第三方的需求或他们提供的服务可能会发生变更。您应预料到可能会发生变动和失败，并进行相应的规划：请安排合适的人员，建立适当的流程并采用正确的技术进行应对。应定期审计您提供的访问级别，并实施检测方法，以便在发生意外变更时向您发出警报。监控并审计角色的使用情况，以及外部 ID 的数据存储状态。若发生意外变更或存在不当访问模式，您应准备暂时或永久撤销第三方访问权限。此外，还要衡量撤销操作造成的影响，包括执行该操作所需的时间、涉及的人员、成本以及对其他资源的影响。

有关检测方法的规范性指南，请参阅[检测最佳实践](#)。

## 资源

### 相关最佳实践：

- [SEC02-BP02 使用临时凭证](#)
- [SEC03-BP05 为您的组织定义权限防护机制](#)
- [SEC03-BP06 基于生命周期管理访问权限](#)
- [SEC03-BP07 分析公共和跨账户访问](#)
- [SEC04 检测](#)

### 相关文档：

- [存储桶所有者向并非其拥有的对象授予跨账户权限](#)
- [如何将信任策略与 IAM 结合使用](#)
- [使用 IAM 角色委派跨 AWS 账户 的访问权限](#)
- [如何使用 IAM 访问其他 AWS 账户 中的资源？](#)
- [IAM 中的安全最佳实践](#)
- [跨账户策略评估逻辑](#)
- [如何在向第三方授予对 AWS 资源的访问权限时使用外部 ID](#)
- [使用自定义资源从外部账户中创建的 AWS CloudFormation 资源中收集信息](#)
- [安全地使用外部 ID 访问他人拥有的 AWS 账户](#)
- [使用 IAM Roles Anywhere 将 IAM 角色扩展到 IAM 外部工作负载](#)

### 相关视频：

- [如何允许单独 AWS 账户 中的用户或角色访问我的 AWS 账户？](#)
- [AWS re:Invent 2018：在 60 分钟以内成为 IAM 策略高手](#)
- [AWS 知识中心实况：IAM 最佳实践和设计决策](#)

### 相关示例：

- [Well-Architected 实验室 - Lambda 跨账户 IAM 角色代入（第 300 级）](#)
- [配置对 Amazon DynamoDB 的跨账户访问](#)
- [AWS STS 网络查询工具](#)

# 检测

## 问题

- [SEC 4.您如何检测和调查安全事件？](#)

## SEC 4.您如何检测和调查安全事件？

通过日志和指标来记录和分析事件，以便了解信息。针对安全事件和潜在的威胁采取措施，以便保护您的工作负载。

## 最佳实践

- [SEC04-BP01 配置服务和应用程序日志记录](#)
- [SEC04-BP02 集中分析日志、结果和指标](#)
- [SEC04-BP03 自动响应事件](#)
- [SEC04-BP04 实施可操作的安全事件](#)

### SEC04-BP01 配置服务和应用程序日志记录

保留服务和应用程序的安全事件日志。这是审计、调查和运营使用案例的基本安全原则，也是由监管、风险与合规性（GRC）标准、政策和程序驱动的共同安全要求。

期望结果：当需要履行内部流程或义务（如安全事件响应）时，组织应能够及时、可靠且一致地从AWS服务和应用程序中检索安全事件日志。考虑将日志集中起来，以取得更好的运营成果。

常见反模式：

- 日志被永久存储或过早删除。
- 每个人都可以访问日志。
- 完全依赖手动流程进行日志治理和使用。
- 存储每一种类型的日志，以备不时之需。
- 仅在必要时检查日志完整性。

建立此最佳实践的好处：为安全事件实施根本原因分析（RCA）机制，并为您的监管、风险与合规性义务提供证据来源。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

在安全调查或基于要求的其他使用案例期间，您需要能够查看相关日志，以记录并了解事件的来龙去脉和时间线。警报生成也需要日志，以指示发生了某些感兴趣的操作。选择、启用、存储和设置查询、检索机制以及警报至关重要。

## 实施步骤

- 选择并启用日志源。进行安全调查之前，您需要捕获相关日志，以便以回溯方式重建 AWS 账户中的活动。选择并启用与工作负载相关的日志源。

日志源的选择标准应基于业务所需的使用案例。使用 AWS CloudTrail 或 AWS Organizations 跟踪为每个 AWS 账户 建立跟踪，并为其配置 Amazon S3 存储桶。

AWS CloudTrail 是一项日志记录服务，可跟踪针对 AWS 账户 捕获 AWS 服务活动所进行的 API 调用。它默认情况下启用，管理事件保留 90 天，可以使用 AWS Management Console、AWS CLI 或 AWS SDK [通过 CloudTrail 事件历史记录](#)检索这些事件。为了更长久地保留和了解数据事件，请[创建 CloudTrail 跟踪](#)并将其与 Amazon S3 存储桶关联，也可以选择与 Amazon CloudWatch 日志组关联。或者，您可以创建 [CloudTrail Lake](#)，这可保留 CloudTrail 日志长达七年之久，并提供基于 SQL 的查询工具

AWS 建议使用 VPC 的客户分别使用 [VPC 流日志](#)和 [Amazon Route 53 解析器查询日志](#)启用网络流量和 DNS 日志，并将其流式传输到 Amazon S3 存储桶或 CloudWatch 日志组。您可以为 VPC、子网或网络接口创建 VPC 流日志。对于 VPC 流日志，您可以选择使用流日志的方式和位置，以降低成本。

AWS CloudTrail 日志、VPC 流日志和 Route 53 解析器查询日志是支持 AWS 中安全调查的基本日志记录源。您还可以使用[亚马逊安全数据湖](#)以 Apache Parquet 格式和开放网络安全架构框架（OCSF）收集、标准化和存储这些日志数据，以便于查询。安全数据湖还支持其他 AWS 日志和来自第三方的日志。

AWS 服务可以生成基本日志源未捕获到的日志，如 Elastic Load Balancing 日志、AWS WAF 日志、AWS Config 记录器日志、Amazon GuardDuty 调查结果、Amazon Elastic Kubernetes Service（Amazon EKS）审计日志，以及 Amazon EC2 实例操作系统和应用程序日志。有关日志记录和监控选项的完整列表，请参阅 [AWS 安全事件响应指南](#)的[附录 A：云功能定义 – 日志记录和事件](#)。

- 研究每项 AWS 服务和应用程序的日志记录功能：每项 AWS 服务和应用程序都为您提供了日志存储选项，每个选项都有自己的保留和生命周期功能。两种很常见的日志存储服务是 Amazon Simple Storage Service（Amazon S3）和 Amazon CloudWatch。如果保留期较长，建议使用 Amazon

S3，因为它具有成本效益和灵活的生命周期功能。如果主要日志记录选项是 Amazon CloudWatch Logs，作为一种选择，您应该考虑将不太经常访问的日志存档到 Amazon S3。

- 选择日志存储：日志存储的选择通常与您使用的查询工具、保留能力、熟悉程度和成本有关。日志存储的主要选项是 Amazon S3 存储桶或 CloudWatch 日志组。

Amazon S3 存储桶提供持久且经济高效的存储，并具有可选的生命周期策略。可以使用 Amazon Athena 等服务查询存储在 Amazon S3 存储桶中的日志。

CloudWatch 日志组通过 CloudWatch Logs Insights 提供持久存储和内置查询工具。

- 确定适当的日志保留时长：使用 Amazon S3 存储桶或 CloudWatch 日志组存储日志时，必须为每个日志源建立足够的生命周期，以优化存储和检索成本。客户通常可以查询三个月到一年的日志，日志保留期长达七年。可用性和保留时长的选择应与您的安全要求以及法律法规和业务授权的综合因素相一致。
- 使用适当的保留时长和生命周期策略为每个 AWS 服务和应用程序启用日志记录：对于组织内的每个 AWS 服务或应用程序，请查找特定的日志记录配置指南：
  - [配置 AWS CloudTrail 跟踪](#)
  - [配置 VPC 流日志](#)
  - [配置 Amazon GuardDuty 调查结果导出](#)
  - [配置 AWS Config 记录](#)
  - [配置 AWS WAF Web ACL 流量](#)
  - [配置 AWS Network Firewall 网络流量日志](#)
  - [配置 Elastic Load Balancing 访问日志](#)
  - [配置 Amazon Route 53 解析器查询日志](#)
  - [配置 Amazon RDS 日志](#)
  - [配置 Amazon EKS 控制面板日志](#)
  - [为 Amazon EC2 实例和本地服务器配置 Amazon CloudWatch 代理](#)
- 选择和实施日志查询机制：对于日志查询，可以使用 [CloudWatch Logs Insights](#) 对存储在 CloudWatch 日志组中的数据进行查询，使用 [Amazon Athena](#) 和 [Amazon OpenSearch Service](#) 对存储在 Amazon S3 中的数据进行查询。您还可以使用第三方查询工具，如安全信息和事件管理 (SIEM) 服务。

选择日志查询工具的过程中，应考虑安全运营的人员、流程和技术方面。选择一款能够满足运营、业务和安全要求并可长期使用和维护的工具。请记住，当要扫描的日志数量保持在工具的限制范围内时，日志查询工具的工作状态最佳。由于成本或技术限制，拥有多款查询工具的情况并不罕见。

例如，您可能使用第三方安全信息和事件管理 ( SIEM ) 工具对过去 90 天的数据执行查询，但由于 SIEM 的日志提取成本较高，使用 Athena 来执行 90 天以上的查询。无论采用何种实施方式，都要验证您的方法能够尽可能地减少充分提高运营效率所需的工具数量，尤其在安全事件调查期间。

- 使用日志发出警报：AWS 通过多项安全服务提供警报功能：
  - [AWS Config](#) 监控和记录您的 AWS 资源配置，并允许您对照所需的配置自动执行评估和修复。
  - [Amazon GuardDuty](#) 是一项威胁检测服务，可持续监控恶意活动和未经授权的行为，以保护您的 AWS 账户 和工作负载。GuardDuty 可从 AWS CloudTrail 管理和数据事件、DNS 日志、VPC 流日志和 Amazon EKS 审计日志等来源提取、聚合和分析信息。GuardDuty 可直接从 CloudTrail、VPC 流日志、DNS 查询日志和 Amazon EKS 提取独立的数据流。您无需管理 Amazon S3 存储桶策略，也无需修改日志的收集和存储方式。仍建议保留这些日志，以便您自己进行调查和遵守法规。
  - [AWS Security Hub](#) 集中聚合、组织和优先处理来自多个 AWS 服务和可选第三方产品的安全警报或调查结果，以使您全面了解安全警报和合规性状态。

您也可以使用自定义警报生成引擎来处理这些服务未涵盖的安全警报或与您的环境相关的特定警报。有关构建这些警报和检测的信息，请参阅 [AWS 安全事件响应指南中的“检测”](#)。

## 资源

相关最佳实践：

- [SEC04-BP02 集中分析日志、结果和指标](#)
- [SEC07-BP04 定义数据生命周期管理](#)
- [SEC10-BP06 预部署工具](#)

相关文档：

- [AWS 安全事件响应指南](#)
- [亚马逊安全数据湖入门](#)
- [入门：Amazon CloudWatch Logs](#)
- [安全合作伙伴解决方案：日志记录和监控](#)

相关视频：

- [AWS re:Invent 2022 - 介绍亚马逊安全数据湖](#)



相关示例：

- [专为 AWS 提供的 Assisted Log Enabler](#)
- [AWS Security Hub 调查结果历史导出](#)

相关工具：

- [Snowflake 增强网络安全](#)

## SEC04-BP02 集中分析日志、结果和指标

安全运营团队依靠收集日志和使用搜索工具来发现需要关注的潜在事件，这些事件可能代表未经授权的活动或无意的更改。但是，仅仅分析收集的数据和手动处理信息不足以应对从复杂架构流出的大量信息。单凭分析和报告无法及时分配合适的资源来处理事件。

建立成熟的安全运维团队的最佳实践是，将安全事件和调查结果的流程深度集成到通知和工作流系统中，例如票证系统、错误或问题系统或者其他安全信息和事件管理（SIEM，Security Information and Event Management）系统。这样，工作流可以摆脱电子邮件和静态报告，让您能够路由、上报和管理事件或调查结果。许多组织也在逐步将安全警报集成到他们的聊天或协作以及开发人员工作效率平台中。对于正在踏上自动化之旅的组织，在规划首要自动化任务时，一个由 API 驱动的低延迟票证系统能够提供极高的灵活性。

这种最佳实践不仅适用于从描述用户活动或网络事件的日志消息生成的安全事件，还适用于在基础设施本身检测到的更改生成的安全事件。当面对一些更改，而且这些更改的不受欢迎程度足够微妙，以致于目前无法使用 AWS Identity and Access Management（IAM）和 AWS Organizations 配置的组合来防止这些更改发生时，为了保持和验证安全架构，必须能够检测更改、确定更改是否适当，然后将这些信息路由到正确的修复工作流程。

Amazon GuardDuty 和 AWS Security Hub 为日志记录提供了聚合、重复数据删除和分析机制，您也可以通过其他 AWS 服务提供这些机制。GuardDuty 可从 AWS CloudTrail 管理和数据事件、VPC DNS 日志以及 VPC 流日志等来源提取、聚合和分析信息。Security Hub 能够提取、聚合和分析来自 GuardDuty、AWS Config、Amazon Inspector、Amazon Macie、AWS Firewall Manager 以及 AWS Marketplace 中提供的大量第三方安全产品的输出，如果您相应构建了自己的代码，还将包括这些代码。GuardDuty 和 Security Hub 都有一个管理员-成员模型，此模型可以跨多个账户聚合调查结果和见解，拥有本地 SIEM 的客户通常将 Security Hub 用作 AWS 端日志和警报预处理器和聚合器，随后即可通过基于 AWS Lambda 的处理器和转发服务器提取 Amazon EventBridge。

未建立这种最佳实践的情况下暴露的风险等级：高



## 实施指导

- 评估日志处理功能：评估用于处理日志的选项
  - [使用 Amazon OpenSearch Service 来记录和监控 \( 几乎 \) 所有内容](#)
  - [寻找专门提供日志记录和监控解决方案的合作伙伴](#)
- 作为分析 CloudTrail 日志的开始，请测试 Amazon Athena。
  - [配置 Athena 分析 CloudTrail 日志](#)
- 在 AWS 中实施集中式日志记录：请参阅以下 AWS 示例解决方案来集中处理多个来源的日志记录。
  - [集中日志记录解决方案](#)
- 通过合作伙伴集中处理日志记录：APN 合作伙伴拥有可以帮助您集中分析日志的解决方案。
  - [日志记录和监控](#)

## 资源

### 相关文档：

- [AWS Answers：集中式日志记录](#)
- [AWS Security Hub](#)
- [Amazon CloudWatch](#)
- [Amazon EventBridge](#)
- [开始使用：Amazon CloudWatch Logs](#)
- [安全合作伙伴解决方案：日志记录和监控](#)

### 相关视频：

- [集中监控资源配置和合规性](#)
- [修正 Amazon GuardDuty 和 AWS Security Hub 调查结果](#)
- [云中的威胁管理：Amazon GuardDuty 和 AWS Security Hub](#)

## SEC04-BP03 自动响应事件

使用自动化流程调查和修复事件可减少人工处理工作量和人为错误，从而扩展调查功能。定期审核将帮助您优化自动化工具，并实现持续迭代。

在 AWS 中，可以使用 Amazon EventBridge，调查感兴趣的事件以及自动化工作流程可能发生的意外变化的相关信息。此服务提供可扩展的规则引擎，可代理原生 AWS 事件格式（例如 AWS CloudTrail 事件）以及您可以从应用程序中生成的自定义事件。Amazon GuardDuty 还允许您将这些事件路由到构建意外事件响应系统（AWS Step Functions）的工作流程系统中，或者路由到中央安全账户或存储桶中以执行进一步分析。

检测更改并将此信息路由到正确的工作流的操作也可以使用 AWS Config 规则 和 [合规包](#) 完成。AWS Config 会检测对范围内服务的更改（虽然延迟会比 EventBridge 更高），并生成可使用 AWS Config 规则 进行解析的事件，以便进行回滚、强制实施合规性策略以及将信息转发到相关系统（如变更管理平台 and 运营票证系统）。除了编写您自己的 Lambda 函数以响应 AWS Config 事件，您还可以充分利用 [AWS Config 规则 开发工具包](#) 以及 [一组开源](#) AWS Config 规则。合规包是 AWS Config 规则 和修复操作的集合，您可将其作为以 YAML 模板格式创作的单个实体进行部署。一个 [示例合规包模板](#)，面向 Well-Architected 安全性支柱提供。

未建立这种最佳实践的情况下暴露的风险等级：中

#### 实施指导

- 使用 GuardDuty 实施自动化警报：GuardDuty 是一种威胁检测服务，可持续监控恶意活动和未经授权的行为，从而保护您的 AWS 账户和工作负载。启用 GuardDuty 并配置自动化警报。
- 自动执行调查流程：制定自动化流程来调查事件并向管理员报告信息，以便节省时间。
  - [实验室：Amazon GuardDuty 动手实践](#)

#### 资源

##### 相关文档：

- [AWS Answers：集中式日志记录](#)
- [AWS Security Hub](#)
- [Amazon CloudWatch](#)
- [Amazon EventBridge](#)
- [开始使用：Amazon CloudWatch Logs](#)
- [安全合作伙伴解决方案：日志记录和监控](#)
- [设置 Amazon GuardDuty](#)

##### 相关视频：

- [集中监控资源配置和合规性](#)
- [修正 Amazon GuardDuty 和 AWS Security Hub 调查结果](#)
- [云中的威胁管理：Amazon GuardDuty 和 AWS Security Hub](#)

相关示例：

- [实验室：自动部署检测性控制](#)

## SEC04-BP04 实施可操作的安全事件

创建发送给团队并将由团队处理的警报。确保警报包含团队采取措施所需的相关信息。对于您的每个检测性机制，您还应调查一个以 [运行手册](#) 或者 [行动手册](#) 形式存在的流程。例如，当您启用 [Amazon GuardDuty](#) 时，它会生成不同的 [调查结果](#)。您的每个调查结果类型都应具有一个运行手册条目，例如，如果发现了 [特洛伊木马程序](#)，您的运行手册的简单说明可以指示某个人员进行调查和修复。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

- 发现可用于 AWS 服务的指标：发现可通过 Amazon CloudWatch 用于您正在使用的服务的指标。
  - [AWS 服务文档](#)
  - [使用 Amazon CloudWatch 指标](#)
- 配置 Amazon CloudWatch 告警。
  - [使用 Amazon CloudWatch 告警](#)

### 资源

相关文档：

- [Amazon CloudWatch](#)
- [Amazon EventBridge](#)
- [安全合作伙伴解决方案：日志记录和监控](#)

相关视频：

- [集中监控资源配置和合规性](#)

- [修正 Amazon GuardDuty 和 AWS Security Hub 调查结果](#)
- [云中的威胁管理：Amazon GuardDuty 和 AWS Security Hub](#)

## 基础设施保护

### 问题

- [SEC 5.如何保护您的网络资源？](#)
- [SEC 6.如何保护计算资源？](#)

### SEC 5.如何保护您的网络资源？

任何以某种形式连接至网络（互联网或私有网络）的工作负载都需要多层防御，以帮助防御基于外部和内部网络的威胁。

### 最佳实践

- [SEC05-BP01 创建网络层](#)
- [SEC05-BP02 控制所有层的流量](#)
- [SEC05-BP03 自动执行网络防护](#)
- [SEC05-BP04 实施检查和保护](#)

### SEC05-BP01 创建网络层

将具有共同敏感度要求的组件分成若干层，以尽量缩小未经授权访问的潜在影响范围。例如，应将虚拟私有云（VPC）中无需进行互联网访问的数据库集群，放在无法向/从互联网路由的子网中。流量应仅从相邻的下一个最不敏感的资源流出。应考虑设置一个位于负载均衡器后面的 Web 应用程序。不应直接从负载均衡器访问数据库。只有业务逻辑或 Web 服务器才能直接访问数据库。

期望结果：创建分层网络。分层网络有助于对类似的网络组件进行逻辑分组。它们还缩小了未经授权网络访问的潜在影响范围。适当分层的网络使未经授权的用户更难转向 AWS 环境中的其他资源。除了保护内部网络路径之外，还应保护网络边缘，如 Web 应用程序和 API 端点。

### 常见反模式：

- 在单个 VPC 或子网中创建所有资源。
- 使用过于宽松的安全组。

- 未能使用子网。
- 允许直接访问数据库等数据存储。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

具有相同可访问性要求的组件（例如 Amazon Elastic Compute Cloud（Amazon EC2）实例、Amazon Relational Database Service（Amazon RDS）数据库集群和 AWS Lambda 函数）可细分为由子网构成的层。不妨考虑在 VPC 内或 [Amazon API Gateway](#) 后部署无服务器工作负载，如 [Lambda](#) 函数。应将无需进行互联网访问的 [AWS Fargate](#) 任务放在无法向/从互联网路由的子网中。此分层方法可减轻单层错误配置的影响，这种错误可能导致能够发生意外访问。对于 AWS Lambda，您可以在 VPC 中运行您的函数，以充分利用基于 VPC 的控制。

对于可能包括数千个 VPC、AWS 账户和本地网络的网络连接，您应使用 [AWS Transit Gateway](#)。Transit Gateway 充当一个枢纽，以控制如何在类似于辐条的所有互连网络之间路由流量。Amazon Virtual Private Cloud（Amazon VPC）和 Transit Gateway 之间的流量仍在 AWS 专用网络上，这减少了对未经授权用户的外部暴露和潜在的安全问题。Transit Gateway 区域间对等也会对区域间流量加密，而且不会出现任何单点故障或带宽瓶颈。

## 实施步骤

- 根据配置，使用 [Reachability Analyzer](#) 分析源和目标之间的路径：Reachability Analyzer 使得您能够自动验证与 VPC 所连资源的连接性。请注意，此分析是通过检查配置完成的（在进行分析时不发送网络数据包）。
- 使用 [Amazon VPC 网络访问分析器](#) 识别资源意外受到网络访问的情况：Amazon VPC 网络访问分析器使您能够指定网络访问要求，并识别潜在的网络访问路径。
- 考虑资源是否需要在公有子网中：不要将资源放在您的 VPC 的公有子网中，除非它们绝对必须要接收来自公共来源的入站网络流量。
- 在 [VPC 中创建子网](#)：为每个网络层创建子网（在包含多个可用区的组中），以增强微分段。还要验证您已将正确的[路由表](#)与子网关联，以控制路由和互联网连接。
- 使用 [AWS Firewall Manager](#) 管理 VPC 安全组：AWS Firewall Manager 有助于减轻使用多个安全组的管理负担。
- 使用 [AWS WAF](#) 防范常见的 Web 漏洞：AWS WAF 可通过检查流量中是否存在常见的 Web 漏洞（如 SQL 注入）来帮助增强边缘安全性。它还使您能够限制来自特定国家/地区或地理位置的 IP 地址的流量。

- 使用 [Amazon CloudFront](#) 作为内容分发网络 ( CDN ) : Amazon CloudFront 可通过将数据存储在更靠近用户的位置来帮助加快 Web 应用程序的速度。它还可以实施 HTTPS , 限制对地理区域的访问 , 并确保网络流量只能在通过 CloudFront 路由时访问资源 , 从而提高边缘安全性。
- 创建应用程序编程接口 ( API ) 时使用 [Amazon API Gateway](#) : Amazon API Gateway 可帮助发布、监控和保护 REST、HTTPS 和 WebSocket API。

## 资源

### 相关文档 :

- [AWS Firewall Manager](#)
- [Amazon Inspector](#)
- [Amazon VPC 安全性](#)
- [Reachability Analyzer](#)
- [Amazon VPC 网络访问分析器](#)

### 相关视频 :

- [用于各种 VPC 的 AWS Transit Gateway 参考架构](#)
- [使用 Amazon CloudFront、AWS WAF 和 AWS Shield 提供应用程序加速和保护](#)
- [AWS re:Inforce 2022 - 验证 AWS 上的网络访问控制措施的有效性](#)
- [AWS re:Inforce 2022 - 使用 AWS WAF 针对机器人进行高级防护](#)

### 相关示例 :

- [Well-Architected 实验室 - 自动部署 VPC](#)
- [研讨会 : Amazon VPC 网络访问分析器](#)

## SEC05-BP02 控制所有层的流量

当构建您的网络拓扑时 , 您应检查每个组件的连接要求。例如 , 某个组件是否需要互联网可访问性 ( 入站和出站 ) 、连接到 VPC 的能力、边缘服务和外部数据中心。

使用 VPC , 您可以使用您设置的私有 IPv4 地址范围或者 AWS 选择的 IPv6 地址范围来定义跨 AWS 区域的网络拓扑。对于入站和出站流量 , 您应采用深度防御方法应用多种控制 , 包括使用安全组 ( 状态检测防火墙 ) 、网络 ACL、子网和路由表。在 VPC 中 , 您可以在可用区中创建子网。每个子网都可以

拥有一个关联的路由表，此表定义了用于管理流量在子网内所采用路径的路由规则。您可以将要连接到互联网或 NAT 网关的路由连接到 VPC 或使其经过另一个 VPC，以定义互联网可路由子网。

当在 VPC 内启动某个实例、Amazon Relational Database Service ( Amazon RDS ) 数据库或其他服务时，它的每个网络接口都有自己的安全组。此防火墙位于操作系统层之外，可用于定义允许入站和出站流量的规则。您还可以定义安全组之间的关系。例如，通过参考对相关的实例应用的安全组，数据库层安全组中的实例仅接受来自应用程序层内实例的流量。除非您在使用非 TCP 协议，否则不必在以下情况下允许互联网直接访问 Amazon Elastic Compute Cloud ( Amazon EC2 ) 实例（甚至使用安全组禁止使用的端口）：没有负载均衡器或 [CloudFront](#)。这样有助于防止通过操作系统或应用程序问题进行意外访问。您还可以为子网附加网络 ACL，它将用作无状态防火墙。您应配置网络 ACL 以缩小各层之间允许的流量范围，但请注意，您需要定义入站和出站规则。

一些 AWS 服务要求组件访问互联网进行 API 调用，其目标是 [AWS API 端点](#) 所在的位置。另外一些 AWS 服务使用 [VPC 端点](#)，这些端点位于您的 Amazon VPC 中。很多 AWS 服务（包括 Amazon S3 和 Amazon DynamoDB）都支持 VPC 端点，并且已在 [AWS PrivateLink](#) 中广泛使用此技术。我们建议您使用此方法来访问 AWS 服务、第三方服务以及安全地托管在其他 VPC 中您自己的服务。AWS PrivateLink 上的所有网络流量保持在 AWS 骨干网中，永远不会通过互联网。连接只能由服务的使用方启动，不能由服务的提供方启动。为外部服务访问使用 AWS PrivateLink 让您可以创建没有互联网访问的气隙 VPC，帮助您保护 VPC 免受外部威胁因素的影响。第三方服务可以使用 AWS PrivateLink 允许其客户通过私有 IP 地址，从其 VPC 连接到服务。对于需要出站连接到互联网的 VPC 资产，可以让它们通过 AWS 托管的 NAT 网关、仅出站的互联网网关或者您创建并管理的 Web 代理进行仅出站（单向）连接。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

- 控制 VPC 中的网络流量：实施 VPC 最佳实践来控制流量。
  - [Amazon VPC 安全性](#)
  - [VPC 端点](#)
  - [Amazon VPC 安全组](#)
  - [网络 ACL](#)
- 控制边缘站点的流量：实施边缘服务（例如 Amazon CloudFront），以提供一层额外的保护和其他功能。
  - [Amazon CloudFront 使用案例](#)
  - [AWS Global Accelerator](#)
  - [AWS Web Application Firewall \( AWS WAF \)](#)



- [Amazon Route 53](#)
- [Amazon VPC 传入路由](#)
- 控制私有网络流量：实施保护工作负载专有流量的服务。
  - [Amazon VPC 对等连接](#)
  - [Amazon VPC 端点服务 \( AWS PrivateLink \)](#)
  - [Amazon VPC Transit Gateway](#)
  - [AWS Direct Connect](#)
  - [AWS Site-to-Site VPN](#)
  - [AWS Client VPN](#)
  - [Amazon S3 接入点](#)

## 资源

### 相关文档：

- [AWS Firewall Manager](#)
- [Amazon Inspector](#)
- [开始使用 AWS WAF](#)

### 相关视频：

- [用于各种 VPC 的 AWS Transit Gateway 参考架构](#)
- [使用 Amazon CloudFront、AWS WAF 和 AWS Shield 提供应用程序加速和保护](#)

### 相关示例：

- [实验室：自动部署 VPC](#)

## SEC05-BP03 自动执行网络防护

自动运行保护机制，以提供基于威胁情报和异常检测的自我防御网络。例如可应对最新的威胁并减轻它们的影响的那些入侵检测和预防工具。您可以通过实施 Web 应用程序防火墙来实现自动化的网络保护，例如使用 AWS WAF Security Automations 解决方案 ( <https://github.com/aws-labs/aws-waf-security-automations> ) 来自动拦截来自已知威胁媒介相关 IP 地址的请求。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

- 自动执行基于 Web 流量的保护：AWS 提供了使用 AWS CloudFormation 自动部署一组 AWS WAF 规则的解决方案，旨在筛选常见的基于 Web 的攻击。用户可以从预配置的保护功能中进行选择，这些功能定义 AWS WAF Web 访问控制列表（Web ACL）中包含的规则。
  - [AWS WAF 安全自动化](#)
- 考虑使用 AWS Partner 解决方案：AWS 合作伙伴提供数百种业界领先的产品，这些产品与您的本地环境中的现有控制措施等效、相同或与之集成。这些产品对现有 AWS 服务起到补充作用，使您能够在云和本地部署环境中部署全面的安全架构，进而实现更无缝的体验。
  - [基础设施安全性](#)

## 资源

相关文档：

- [AWS Firewall Manager](#)
- [Amazon Inspector](#)
- [Amazon VPC 安全性](#)
- [开始使用 AWS WAF](#)

相关视频：

- [用于各种 VPC 的 AWS Transit Gateway 参考架构](#)
- [使用 Amazon CloudFront、AWS WAF 和 AWS Shield 提供应用程序加速和保护](#)

相关示例：

- [实验室：自动部署 VPC](#)

## SEC05-BP04 实施检查和保护

检查和筛选每层的流量。您可以使用 [VPC Network Access Analyzer](#) 检测 VPC 配置中可能存在的意外访问。您可以指定网络访问需求，然后确定不能满足这些要求的潜在网络路径。对于通过基于 HTTP 的协议处理的组件，Web 应用程序防火墙可帮助防止常见的攻击。[AWS WAF](#) 是一个 Web 应用程序

防火墙，可监控和拦截与转发到 Amazon API Gateway API、Amazon CloudFront 或 Application Load Balancer 的可配置规则匹配的 HTTP(s) 请求。要开始使用 AWS WAF，您可以将 [AWS 托管式规则](#) 与您自己的规则结合使用，也可以使用现有的 [合作伙伴集成](#)。

要管理 AWS WAF、AWS Shield Advanced 保护以及跨 AWS Organizations 的 Amazon VPC 安全组，您可以使用 AWS Firewall Manager。它允许您跨账户和应用程序集中配置和管理防火墙规则，从而更轻松地扩展常见规则的实施。通过使用 [AWS Shield Advanced](#) 或 [能够自动拦截向](#) 您的 Web 应用程序发送非必要请求的解决方案，它还使您能够快速响应攻击。Firewall Manager 也可以与 [AWS Network Firewall](#) 结合使用。AWS Network Firewall 是一种托管服务，使用规则引擎为您提供对有状态和无状态网络流量的精细控制。它支持 [与 Suricata 兼容的](#) 开源入侵防御系统 (IPS, Intrusion Prevention System) 规范，以便使用规则来保护您的工作负载。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

- 配置 Amazon GuardDuty：GuardDuty 是一种威胁检测服务，可持续监控恶意活动和未经授权的行为，从而保护您的 AWS 账户 和工作负载。启用 GuardDuty 并配置自动化警报。
  - [Amazon GuardDuty](#)
  - [实验室：自动部署检测性控制](#)
- 配置虚拟私有云 (VPC) 流日志：VPC 流日志功能使您能够进一步捕获有关传入和传出 VPC 中网络接口的 IP 流量信息。流日志数据可以发布到 Amazon CloudWatch Logs 和 Amazon Simple Storage Service (Amazon S3)。创建流日志后，您可以在选定目标中检索和查看其数据。
- 考虑使用 VPC 流量镜像：流量镜像是一项 Amazon VPC 功能，您可以用它从 Amazon Elastic Compute Cloud (Amazon EC2) 实例的弹性网络接口复制网络流量，然后将其发送到带外安全和监控设备，以进行内容检查、威胁监控和故障排除。
  - [VPC 流量镜像](#)

### 资源

#### 相关文档：

- [AWS Firewall Manager](#)
- [Amazon Inspector](#)
- [Amazon VPC 安全性](#)
- [开始使用 AWS WAF](#)

相关视频：

- [用于各种 VPC 的 AWS Transit Gateway 参考架构](#)
- [使用 Amazon CloudFront、AWS WAF 和 AWS Shield 提供应用程序加速和保护](#)

相关示例：

- [实验室：自动部署 VPC](#)

## SEC 6.如何保护计算资源？

工作负载内的计算资源需要采用多层防御，才有助于免受内部和外部威胁。计算资源包括 EC2 实例、容器、AWS Lambda 函数、数据库服务、IoT 设备等。

最佳实践

- [SEC06-BP01 执行漏洞管理](#)
- [SEC06-BP02 缩小攻击面](#)
- [SEC06-BP03 实施托管服务](#)
- [SEC06-BP04 自动保护计算](#)
- [SEC06-BP05 帮助人员远程执行操作](#)
- [SEC06-BP06 验证软件完整性](#)

### SEC06-BP01 执行漏洞管理

频繁扫描和修补您的代码、依赖项和基础设施中的漏洞，以帮助防御新的威胁。

期望结果：制定并维护漏洞管理计划。定期扫描和修补资源，例如 Amazon EC2 实例、Amazon Elastic Container Service ( Amazon ECS ) 容器和 Amazon Elastic Kubernetes Service ( Amazon EKS ) 工作负载。为 AWS 托管的资源 ( 如 Amazon Relational Database Service ( Amazon RDS ) 数据库 ) 配置维护时段。使用静态代码扫描检查应用程序源代码的常见问题。如果贵组织具备必要的技能或可以聘请外部人员协助，则不妨考虑执行 Web 应用程序渗透测试。

常见反模式：

- 未制定漏洞管理计划。
- 在不考虑严重性或风险规避的情况下执行系统修补。

- 使用已超过供应商提供的生命周期结束 ( EOL ) 日期的软件。
- 在分析安全问题之前，将代码部署到生产环境中。

建立此最佳实践的好处：

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

漏洞管理计划包括安全评估、识别问题、确定优先级，以及执行修补操作（问题解决流程的一部分）。自动化是持续扫描工作负载（以查找问题和意外网络暴露）并执行补救措施的关键。自动创建和更新资源可以节省时间，并降低配置错误产生进一步问题的风险。完善的漏洞管理计划还应考虑在软件生命周期的开发和部署阶段进行漏洞测试。在开发和部署期间实施漏洞管理有助于减少漏洞进入生产环境的机会。

实施漏洞管理计划需要很好地理解 [AWS 责任共担模式](#)，以及它与特定工作负载的关系。在责任共担模式下，AWS 负责保护 AWS Cloud 的基础设施。此基础设施由运行 AWS Cloud 服务的硬件、软件、网络和设施组成。您负责云中的安全（例如，Amazon EC2 实例的实际数据、安全配置和管理任务），并验证 Amazon S3 对象已正确分类和配置。漏洞管理方法也可能因您使用的服务而异。例如，AWS 管理我们的托管关系数据库服务 Amazon RDS 的修补，但您将负责修补自托管数据库。

AWS 提供一系列服务来帮助您制定漏洞管理计划。[Amazon Inspector](#) 持续扫描 AWS 工作负载是否存在软件问题和意外网络访问。[AWS Systems Manager 补丁管理器](#) 可帮助跨 Amazon EC2 实例管理修补。[AWS Security Hub](#) 是一项云安全状况管理服务，有助于自动执行 AWS 安全检查并将安全警报集中起来，您可以通过该服务查看 Amazon Inspector 和 Systems Manager。

[Amazon CodeGuru](#) 可以使用静态代码分析，帮助识别 Java 和 Python 应用程序中的潜在问题。

## 实施步骤

- 配置 [Amazon Inspector](#)：Amazon Inspector 会自动检测新发布的 Amazon EC2 实例、Lambda 函数和推送到 Amazon ECR 的符合条件的容器映像，并立即扫描它们以查找软件问题、潜在缺陷和意外网络暴露。
- 扫描源代码：扫描库和依赖项，以确定是否有问题和缺陷。[Amazon CodeGuru](#) 会进行扫描，并提供建议以修复 Java 和 Python 应用程序的[常见安全问题](#)。[OWASP Foundation](#) 发布了一系列源代码分析工具（也称为 SAST 工具）。
- 实施一种机制来扫描和修补现有环境，以及作为 CI/CD 管道构建流程的一部分进行扫描：实施一种机制来扫描和修补依赖项和操作系统中的问题，以帮助防范新的威胁。定期运行这种机制。软件漏洞

管理对于了解需要应用补丁或解决软件问题的位置至关重要。在持续集成/持续交付 ( CI/CD ) 管道中尽早嵌入漏洞评估，对潜在的安全问题进行优先修复。您的方法可能会因您使用的 AWS 服务而异。要检查 Amazon EC2 实例中运行的软件中的潜在问题，请在管道中添加 [Amazon Inspector](#)，以便在检测到问题或潜在缺陷时发出警报并停止构建过程。Amazon Inspector 会持续监控资源。您还可以使用开源产品 ( 如 [OWASP Dependency-Check](#)、[Snyk](#)、[OpenVAS](#)、程序包管理器和 AWS Partner 工具 ) 进行漏洞管理。

- 使用 [AWS Systems Manager](#)：您负责自己 AWS 资源的补丁管理，包括 Amazon Elastic Compute Cloud ( Amazon EC2 ) 实例、亚马逊云机器镜像 ( AMI ) 以及其他计算资源。[AWS Systems Manager 补丁管理器](#)使用安全相关的更新和其他类型的更新自动执行修补托管实例的流程。补丁管理器可用于在 Amazon EC2 实例上为操作系统和应用程序 ( 包括 Microsoft 应用程序、Windows Service Pack 和基于 Linux 实例的次要版本升级 ) 应用补丁。除了 Amazon EC2 之外，补丁管理器还可用于对本地服务器进行修补。

有关支持的操作系统的列表，请参阅 Systems Manager 用户指南中的[支持的操作系统](#)。您可以扫描实例以单独查看缺失补丁的报告，也可以扫描并自动安装所有缺失的补丁。

- 使用 [AWS Security Hub](#)：Security Hub 可提供 AWS 中安全状况的全面视图。它跨[多项 AWS 服务](#)收集安全性数据，并以标准化格式提供这些调查结果，使您能够对 AWS 服务中的安全性调查结果进行优先级排序。
- 使用 [AWS CloudFormation](#)：[AWS CloudFormation](#) 是一项基础设施即代码 ( IaC ) 服务，可通过跨多个账户和环境实现资源部署自动化和资源架构标准化来帮助管理漏洞。

## 资源

### 相关文档：

- [AWS Systems Manager](#)
- [AWS Lambda 安全性概述](#)
- [Amazon CodeGuru](#)
- [使用新的 Amazon Inspector 改进了云工作负载的自动化漏洞管理](#)
- [使用 Amazon Inspector 和 AWS Systems Manager 自动执行 AWS 中的漏洞管理和修复 – 第 1 部分](#)

### 相关视频：

- [保护无服务器和容器服务](#)
- [有关 Amazon EC2 实例元数据服务的安全最佳实践](#)



## SEC06-BP02 缩小攻击面

通过强化操作系统，并尽量减少所使用的组件、库和外部可用的服务，缩小暴露在意外访问下的危险。首先减少未使用的组件，无论它们是操作系统程序包、应用程序（适用于基于 Amazon Elastic Compute Cloud ( Amazon EC2 ) 的工作负载）还是您代码中的外部软件模块（适用于所有工作负载）。您可以找到许多面向常见的操作系统和服务器软件的强化和安全配置指南。例如，您可以从 [互联网安全中心](#) 开始并进行迭代。

在 Amazon EC2 中，您可以创建自己的亚马逊云机器镜像 ( AMI ) 并进行修补和强化，以帮助满足企业的特定安全要求。您应用到 AMI 上的补丁和其他安全控制措施在其创建时生效，它们并非动态的，除非您在启动之后进行了修改，例如，使用 AWS Systems Manager 进行修改。

您可以使用 EC2 Image Builder 简化构建安全 AMI 的过程。EC2 Image Builder 可大幅减少创建和维护黄金镜像所需的工作，无需编写和维护自动化过程。在有软件更新可用时，Image Builder 自动生成新的镜像，无需用户手动迭代镜像工作版本。通过 EC2 Image Builder，您可以使用 AWS 提供的测试和自己的测试，在将镜像部署到生产环境中之前轻松地验证镜像的功能和安全性。您还可以应用 AWS 提供的安全设置来进一步保护自己的镜像，满足内部安全标准。例如，您可以使用 AWS 提供的模板，生成符合安全技术实施指南 ( STIG , Security Technical Implementation Guide ) 标准的镜像。

使用第三方静态代码分析工具，您可以确定常见的安全问题，例如未检查的函数输入边界，以及适用的通用漏洞披露 ( CVE , Common Vulnerabilities and Exposures )。您可以对所支持的语言使用 [Amazon CodeGuru](#)。您还可以使用第三方依赖关系检查工具，确定代码链接的库是否是最新版本、它们是否不含 CVE，并确保您拥有符合您软件政策要求的许可条件。

使用 Amazon Inspector，您可以针对 CVE，对您的实例执行配置评估、根据安全基准执行评估以及实现缺陷通知自动化。Amazon Inspector 在生产实例或构建管道中运行，它会在发现结果时通知开发人员和工程师。您可以通过编程方式访问调查结果，并将您的团队引导至待办事项和错误跟踪系统。[EC2 Image Builder](#) 可通过自动化修补、AWS 提供的安全策略实施和其他自定义来维护服务器映像 (AMI)。当使用容器时，在您的构建管道中对您的映像存储库定期实施 [ECR 映像扫描](#)，以便在您的容器中查找 CVE。

尽管 Amazon Inspector 和其他工具能够有效地确定配置和存在的任何 CVE，但也需要使用其他方法在应用程序级别测试您的工作负载。[模糊](#) 是一种众所周知的查错方法，可自动将格式不正确的数据注入到您应用程序的输入字段和其他区域来查错。

未建立此最佳实践暴露的风险等级：高

### 实施指导

- 强化操作系统：配置操作系统以符合最佳实践。



- [保护 Amazon Linux](#)
- [保护 Microsoft Windows Server](#)
- 强化容器化资源：配置容器化资源以符合安全最佳实践。
- 实施 AWS Lambda 最佳实践。
- [AWS Lambda 最佳实践](#)

## 资源

### 相关文档：

- [AWS Systems Manager](#)
- [使用 Amazon EC2 Systems Manager 替换堡垒主机](#)
- [AWS Lambda 安全性概述](#)

### 相关视频：

- [在 Amazon EKS 上运行高安全性工作负载](#)
- [保护无服务器和容器服务](#)
- [有关 Amazon EC2 实例元数据服务的安全最佳实践](#)

### 相关示例：

- [实验室：自动部署 Web 应用程序防火墙](#)

## SEC06-BP03 实施托管服务

实施用于托管资源的服务，例如 Amazon Relational Database Service ( Amazon RDS )、AWS Lambda 和 Amazon Elastic Container Service ( Amazon ECS )，以便在责任共担模式中减少安全维护任务。例如，Amazon RDS 可帮助您设置、操作和扩展关系数据库，并自动执行管理任务，例如硬件预置、数据库设置、修补和备份。这意味着您将有更多的空闲时间，因此可以专注于通过 AWS Well-Architected Framework 中所述的其他方法来保护您的应用程序。使用 Lambda，无需使用预置或托管服务器即可运行代码，因此您只需在代码级别专注于连接、调用和安全性，而不是基础设施或操作系统级别。

未建立此最佳实践暴露的风险等级：中

## 实施指导

- 探索可用的服务：探索、测试和实施管理资源的服务，例如 Amazon RDS、AWS Lambda 和 Amazon ECS。

## 资源

### 相关文档：

- [AWS 网站](#)
- [AWS Systems Manager](#)
- [使用 Amazon EC2 Systems Manager 替换堡垒主机](#)
- [AWS Lambda 安全性概述](#)

### 相关视频：

- [在 Amazon EKS 上运行高安全性工作负载](#)
- [保护无服务器和容器服务](#)
- [有关 Amazon EC2 实例元数据服务的安全最佳实践](#)

### 相关示例：

- [实验室：AWS Certificate Manager 请求公有证书](#)

## SEC06-BP04 自动保护计算

自动执行计算保护机制，包括管理漏洞、缩小攻击面和管理资源。此自动化将帮助您投入时间以保护工作负载的其他方面，并降低人为犯错的风险。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

- 自动管理配置：使用配置管理服务或工具自动实施安全配置并对其进行验证。
  - [AWS Systems Manager](#)
  - [AWS CloudFormation](#)
  - [实验室：自动部署 VPC](#)

- [实验室：自动部署 EC2 Web 应用程序](#)
- 自动修补 Amazon Elastic Compute Cloud ( Amazon EC2 ) 实例：AWS Systems Manager Patch Manager 使用安全相关的更新和其他类型的更新来自动执行修补托管实例的流程。您可以使用 Patch Manager 为操作系统和应用程序应用修补程序。
- [AWS Systems Manager 补丁管理器](#)
- [使用 AWS Systems Manager Automation 集中完成多账户和多区域的修补](#)
- 实施入侵检测和预防：实施入侵检测和预防工具，以监控并停止实例上的恶意活动。
- 考虑使用 AWS Partner 解决方案：AWS 合作伙伴提供数百种业界领先的产品，这些产品与您的本地环境中的现有控制措施等效、相同或与之集成。这些产品对现有 AWS 服务起到补充作用，使您能够在云和本地部署环境中部署全面的安全架构，进而实现更无缝的体验。
- [基础设施安全性](#)

## 资源

### 相关文档：

- [AWS CloudFormation](#)
- [AWS Systems Manager](#)
- [AWS Systems Manager 补丁管理器](#)
- [使用 AWS Systems Manager Automation 集中完成多账户和多区域的修补](#)
- [基础设施安全性](#)
- [使用 Amazon EC2 Systems Manager 替换堡垒主机](#)
- [AWS Lambda 安全性概述](#)

### 相关视频：

- [在 Amazon EKS 上运行高安全性工作负载](#)
- [保护无服务器和容器服务](#)
- [有关 Amazon EC2 实例元数据服务的安全最佳实践](#)

### 相关示例：

- [实验室：自动部署 Web 应用程序防火墙](#)
- [实验室：自动部署 EC2 Web 应用程序](#)

## SEC06-BP05 帮助人员远程执行操作

移除交互式访问功能可降低人为错误的风险以及手动配置或管理的可能性。例如，通过更改管理工作流，使用基础设施即代码部署 Amazon Elastic Compute Cloud ( Amazon EC2 ) 实例，然后使用 AWS Systems Manager 等工具管理 Amazon EC2 实例，而不是允许直接访问或通过堡垒主机进行访问。AWS Systems Manager 可以使用 [自动化 工作流](#)、[文档](#)（行动手册）和 [Run Command](#) 等功能自动执行多种维护和部署任务。AWS CloudFormation 堆栈从管道进行构建，并能够自动执行您的基础设施部署和管理任务，而无需直接使用 AWS Management Console 或 API。

未建立此最佳实践暴露的风险等级：低

### 实施指导

- 替换控制台访问：用 AWS Systems Manager Run Command 替换实例的控制台访问（SSH 或 RDP），以自动管理任务。
- [AWS Systems Manager Run Command](#)

### 资源

#### 相关文档：

- [AWS Systems Manager](#)
- [AWS Systems Manager Run Command](#)
- [使用 Amazon EC2 Systems Manager 替换堡垒主机](#)
- [AWS Lambda 安全性概述](#)

#### 相关视频：

- [在 Amazon EKS 上运行高安全性工作负载](#)
- [保护无服务器和容器服务](#)
- [有关 Amazon EC2 实例元数据服务的安全最佳实践](#)

相关示例：

- [实验室：自动部署 Web 应用程序防火墙](#)

## SEC06-BP06 验证软件完整性

实施一些机制（例如代码签名），以确保工作负载中使用的软件、代码和库来自可信的来源且未被篡改。例如，您应验证二进制文件和脚本的代码签名证书以确认作者，并确保证书自作者创建以来未被篡改。[AWS Signer](#) 通过集中管理代码签名生命周期，包括签名证书以及公有和私有密钥，帮助确保代码的可信度和完整性。您可以了解如何对 [AWS Lambda](#) 使用代码签名的高级模式和最佳实践。此外，通过将您下载的软件在校验和与提供商提供的校验和进行对比，可帮助确保它未被篡改。

未建立这种最佳实践的情况下暴露的风险等级：低

实施指导

- 调查机制：代码签名是一种可用来验证软件完整性的机制。
  - [NIST：代码签名的安全注意事项](#)

资源

相关文档：

- [AWS Signer](#)
- [新增 – 代码签名，用于 AWS Lambda 的可信度和完整性控制措施](#)

## 数据保护

问题

- [SEC 7.如何对数据进行分类？](#)
- [SEC 8.如何保护静态数据？](#)
- [SEC 9.如何保护传输中的数据？](#)

### SEC 7.如何对数据进行分类？

分类提供了一种基于关键性和敏感度对数据进行分类的方法，以帮助确定适当的保护和保留控制措施。

## 最佳实践

- [SEC07-BP01 识别工作负载内的数据](#)
- [SEC07-BP02 定义数据保护控制措施](#)
- [SEC07-BP03 自动识别和分类](#)
- [SEC07-BP04 定义数据生命周期管理](#)

### SEC07-BP01 识别工作负载内的数据

了解工作负载所处理数据的类型和分类、关联的业务流程、数据存储位置以及数据所有者至关重要。您还应了解工作负载的适用法律和合规性要求，以及需要执行的数据控制措施。识别数据是数据分类过程的第一步。

建立此最佳实践的好处：

通过数据分类，工作负载所有者可以识别存储敏感数据的位置，并确定访问和共享这些数据的方式。

数据分类旨在回答以下问题：

- 您拥有什么类型的数据？

可能包括以下数据：

- 知识产权（IP），例如商业秘密、专利或合同协议。
- 受保护健康信息（PHI），例如包含与个人相关的病史信息的医疗记录。
- 个人身份信息（PII），例如姓名、地址、出生日期和国民身份证或登记号码。
- 信用卡数据，例如主账号（PAN）、持卡人姓名、到期日期和服务代码编号。
- 敏感数据存储在哪里？
- 谁可以访问、修改和删除数据？
- 了解用户权限对于防范潜在的数据误操作至关重要。
- 谁可以执行创建、读取、更新和删除（CRUD）操作？
- 通过了解谁可以管理数据权限，对潜在的权限升级进行说明。
- 如果数据被无意中披露、更改或删除，可能会产生什么业务影响？
- 了解数据被修改、删除或无意中披露的风险后果。

通过了解这些问题的答案，您可以采取以下行动：

- 缩小敏感数据的范围（如敏感数据位置的数量），并限制敏感数据的访问权限，仅限经批准的用户进行访问。
- 了解不同的数据类型，以便实施适当的数据保护机制和技术，如加密、数据丢失防护以及身份和权限管理。
- 通过为数据提供正确的控制目标来优化成本。
- 自信地回答监管机构和审计人员提出的关于数据类型和数量，以及不同敏感度的数据如何相互隔离的问题。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

数据分类是确定数据敏感度的行为。该行为可能涉及标记，以使数据易于搜索和跟踪。数据分类还可减少数据的重复，这有助于降低存储和备份成本，同时加快搜索过程。

使用 Amazon Macie 等服务大规模将敏感数据的发现和分类自动化。其他服务（如 Amazon EventBridge 和 AWS Config）可用于自动修复数据安全问题，如未加密的 Amazon Simple Storage Service（Amazon S3）存储桶和 Amazon EC2 EBS 卷或未标记的数据资源。有关 AWS 服务集成的完整列表，请参阅 [EventBridge 文档](#)。

通过[使用 Amazon Comprehend](#)（一种自然语言处理（NLP）服务，该服务使用机器学习（ML）在非结构化文本中查找人物、地点、情感和主题等洞察和关系），可以在非结构化数据（如客户电子邮件、支持工单、产品评论和社交媒体）中[检测 PII](#)。有关可协助进行数据识别的 AWS 服务的列表，请参阅[使用 AWS 服务检测 PHI 和 PII 数据的常见技术](#)。

另一种支持数据分类和保护的方法是[AWS 资源标记](#)。通过标记，可以为 AWS 资源分配元数据，您可以使用这些元数据来管理、识别、组织、搜索和筛选资源。

在某些情况下，您可能会选择标记整个资源（例如 S3 存储桶），尤其当特定工作负载或服务预计将存储已知数据分类的进程或传输时。

在适当情况下，您可以标记 S3 存储桶而不是单个对象，以便于管理和安全维护。

## 实施步骤

检测 Amazon S3 内的敏感数据：

1. 开始之前，请确保您拥有访问 Amazon Macie 控制台和执行 API 操作的适当权限。有关其他详细信息，请参阅[开始使用 Amazon Macie](#)。



2. 当敏感数据驻留在 [Amazon S3](#) 中时，使用 Amazon Macie 执行自动化数据发现。

- 使用[开始使用 Amazon Macie](#) 指南为敏感数据发现结果配置存储库，并为敏感数据创建发现作业。
- [如何使用 Amazon Macie 预览 S3 存储桶中的敏感数据。](#)

默认情况下，Macie 通过使用我们建议用于自动化敏感数据发现的一组托管数据标识符来分析对象。您可以定制分析，方法是将 Macie 配置为在为您的账户或组织执行自动化敏感数据发现时，使用特定的托管数据标识符、自定义数据标识符和允许列表。您可以通过排除特定的存储桶（例如，通常存储 AWS 日志记录数据的 S3 存储桶）来调整分析范围。

3. 要配置和使用自动化敏感数据发现，请参阅[使用 Amazon Macie 执行自动化敏感数据发现](#)。

4. 您也可以考虑[针对 Amazon Macie 的自动化数据发现](#)。

检测 Amazon RDS 内的敏感数据：

有关 [Amazon Relational Database Service \( Amazon RDS \)](#) 数据库中数据发现的更多信息，请参阅[使用 Macie 为 Amazon RDS 数据库启用数据分类](#)。

检测 DynamoDB 内的敏感数据：

- [使用 Macie 检测 DynamoDB 内的敏感数据](#)介绍了如何使用 Amazon Macie 通过将数据导出到 Amazon S3 进行扫描来检测 [Amazon DynamoDB](#) 表中的敏感数据。

AWS 合作伙伴解决方案：

- 考虑使用我们广泛的 AWS Partner Network。AWS 合作伙伴拥有广泛的工具和合规性框架，可直接与 AWS 服务集成。合作伙伴可以为您提供量身定制的治理和合规性解决方案，以帮助满足组织需求。
- 有关数据分类中的定制解决方案，请参阅[监管和合规性要求时代的数据治理](#)。

通过使用 AWS Organizations 创建和部署策略，您可以自动实施贵组织所采用的标记标准。您可以通过标签策略来指定规则，规则中定义有效的密钥名称以及对每个密钥有效的值。您可以选择仅监控，这使您有机会评估和清理现有标签。标签符合所选标准后，您可以在标签策略中启用强制执行，以防止创建不合规的标签。有关更多详细信息，请参阅[使用 AWS Organizations 中的服务控制策略保护用于授权的资源标签](#)，以及关于[防止标签被授权主体以外的人员修改](#)的示例策略。

- 要开始在 [AWS Organizations](#) 中使用标签策略，强烈建议您先遵循[开始使用标签策略](#)中的工作流程，然后再使用更高级的标签策略。在扩展到整个组织单位（OU）或组织之前，了解将简单的标签

策略附加到单个账户的效果，可以在强制遵守某项标签策略之前看到该标签策略的效果。[开始使用标签策略](#)提供了指向更高级策略相关任务的说明的链接。

- 不妨考虑评估一下[数据分类](#)白皮书中列出的支持数据分类的其他 [AWS 服务和功能](#)。

## 资源

### 相关文档：

- [开始使用 Amazon Macie](#)
- [使用 Amazon Macie 执行自动化数据发现](#)
- [开始使用标签策略](#)
- [检测 PII 实体](#)

### 相关博客：

- [如何使用 Amazon Macie 预览 S3 存储桶中的敏感数据。](#)
- [使用 Amazon Macie 执行自动化敏感数据发现。](#)
- [使用 AWS 服务检测 PHI 和 PII 数据的常见技术](#)
- [使用 Amazon Comprehend 检测和编辑 PII](#)
- [使用 AWS Organizations 中的服务控制策略保护用于授权的资源标签](#)
- [使用 Macie 为 Amazon RDS 数据库启用数据分类](#)
- [使用 Macie 检测 DynamoDB 中的敏感数据](#)
- 

### 相关视频：

- [使用 Amazon Macie 的事件驱动型数据安全性](#)
- [用于数据保护和治理的 Amazon Macie](#)
- [利用允许列表对敏感数据调查结果进行微调](#)

## SEC07-BP02 定义数据保护控制措施

根据数据分类级别保护数据。例如，使用相关建议保护分类为公共的数据，同时使用其他控制措施保护敏感数据。

通过使用资源标签、根据敏感度（可能还包括限制性条款、飞地或感兴趣的社区）划分 AWS 账户、IAM 策略、AWS Organizations SCP、AWS Key Management Service（AWS KMS）和 AWS CloudHSM，您可以定义并实施您的数据分类和加密保护策略。例如，如果您的项目具有包含极关键数据的 S3 存储桶或者处理机密数据的 Amazon Elastic Compute Cloud（Amazon EC2）实例，则可以使用 Project=ABC 标签对其进行标记。只有您的直属团队知道项目代码的含义，它提供了一种使用基于属性的访问控制措施的方法。您可以通过关键策略和授权定义对 AWS KMS 加密密钥的访问级别，以确保只有适当的服务可以通过安全机制访问敏感内容。如果您正在根据标签做出授权决定，您应确保在 AWS Organizations 中使用标签策略适当定义对于标签的权限。

未建立此最佳实践暴露的风险等级：高

### 实施指导

- 定义您的数据识别和分类架构：对数据执行标识和分类，用于评估您要存储的数据的潜在影响和类型，并确定谁可以访问数据。
  - [AWS 文档](#)
- 发现可用的 AWS 控制措施：对于您正在使用或计划使用的 AWS 服务，发现安全控制措施。许多服务在其文档中都会提供一个安全部分。
  - [AWS 文档](#)
- 确定 AWS 合规性资源：确定 AWS 为您提供帮助的资源。
  - <https://aws.amazon.com/compliance/>

### 资源

#### 相关文档：

- [AWS 文档](#)
- [数据分类白皮书](#)
- [开始使用 Amazon Macie](#)
- [缺少文本](#)

#### 相关视频：

- [新 Amazon Macie 简介](#)

## SEC07-BP03 自动识别和分类

自动识别和分类数据可帮助您实施正确的控制措施。在这方面实现自动化而不是允许人员直接访问，可以降低人为犯错和漏洞的风险。您应使用 [Amazon Macie](#) 等工具执行评估，这些工具使用机器学习来自动发现、分类和保护 Amazon Macie 中的敏感数据。AWS 可以识别个人信息（PII，Personally Identifiable Information）或知识产权之类的敏感数据，并为您提供控制面板和警报，让您了解此类数据的访问或移动情况。

未建立此最佳实践暴露的风险等级：中

### 实施指导

- 使用 Amazon Simple Storage Service（Amazon S3）清单：Amazon S3 清单是可以用来审核和报告对象的复制和加密状态的工具之一。
  - [Amazon S3 清单](#)
- 考虑使用 Amazon Macie：Amazon Macie 使用机器学习来自动发现存储在 Amazon S3 中的数据，并对其进行分类。
  - [Amazon Macie](#)

### 资源

相关文档：

- [Amazon Macie](#)
- [Amazon S3 清单](#)
- [数据分类白皮书](#)
- [开始使用 Amazon Macie](#)

相关视频：

- [新 Amazon Macie 简介](#)

## SEC07-BP04 定义数据生命周期管理

您定义的生命周期策略应基于敏感度级别以及法律和组织要求。应考虑您的数据保留期限、数据销毁流程、数据访问管理、数据转换和数据共享等方面。当选择数据分类方法时，请平衡可用性与访问权限。您还应考虑多种访问级别及其细微差别，以便针对每个级别实施安全且有效的方法。始终采用深度防御方法并减少人工访问数据次数以及数据转换、删除或复制机制。例如，要求用户对应用程序执行严格身

份验证，并为应用程序而不是用户授予执行远程操作的必要访问权限。此外，确保用户来自可信网络路径并要求其获取解密密钥。使用控制面板和自动报告等工具为用户提供数据信息，而不是让他们直接访问数据。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

- 识别数据类型：确定您正在工作负载中存储或处理的数据类型。这些数据可以是文本、图像、二进制数据库等。

### 资源

相关文档：

- [数据分类白皮书](#)
- [开始使用 Amazon Macie](#)

相关视频：

- [新 Amazon Macie 简介](#)

## SEC 8.如何保护静态数据？

通过实施多个控制措施来保护静态数据，以降低未经授权的访问或处理不当带来的风险。

### 最佳实践

- [SEC08-BP01 实施安全密钥管理](#)
- [SEC08-BP02 强制实施静态加密](#)
- [SEC08-BP03 自动执行静态数据保护](#)
- [SEC08-BP04 强制实施访问控制](#)
- [SEC08-BP05 使用机制限制对数据的访问](#)

### SEC08-BP01 实施安全密钥管理

安全密钥管理包括密钥材料的存储、轮换、访问控制和监控，这些都是保护工作负载的静态数据安全所必需的。

期望的结果：一种可扩展、可重复且自动化的密钥管理机制。该机制应能够强制对密钥材料实施最低权限访问，在密钥可用性、机密性和完整性之间取得适当的平衡。应监控密钥的使用，并通过自动化流程轮换密钥材料。密钥材料永远不应能够通过人员的身份来访问。

常见反模式：

- 由人类访问未加密的密钥材料。
- 创建自定义加密算法。
- 访问密钥材料的权限过于宽泛。

建立此最佳实践的好处：通过为您的工作负载建立安全的密钥管理机制，您可以帮助保护您的内容免遭未经授权的访问。此外，您可能需要遵守对数据进行加密的监管要求。有效的密钥管理解决方案可以提供符合这些法规的技术机制，进而保护密钥材料。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

许多监管要求和最佳实践都将静态数据加密作为一项基本的安全控制措施。为了遵守这种控制措施，您的工作负载需要一种机制，来安全地存储和管理用于加密静态数据的密钥材料。

AWS 的 AWS Key Management Service ( AWS KMS ) 可为 AWS KMS 密钥提供持久、安全和冗余的存储空间。[许多 AWS 服务都与 AWS KMS 集成](#) 来支持对您的数据进行加密。AWS KMS 使用经 FIPS 140-2 Level 3 验证的硬件安全模块来保护您的密钥。不存在以纯文本格式导出 AWS KMS 密钥的机制。

使用多账户策略部署工作负载时，[最佳实践](#) 是将 AWS KMS 密钥与使用密钥的工作负载保存在同一个账户中。在这种分布式模型中，由应用程序团队负责管理 AWS KMS 密钥。在其他用例中，组织可以选择将 AWS KMS 密钥存储到集中式账户中。这种集中式结构需要额外的策略，以实现工作负载账户访问集中式账户中存储的密钥所需的跨账户访问，但可能更适用于多个 AWS 账户 账户共享单个密钥的用例。

无论密钥材料存放在哪里，都应通过使用 [密钥策略](#) 和 IAM 策略来严格控制对密钥的访问。密钥策略是控制对 AWS KMS 密钥的访问的主要方式。此外，AWS KMS 密钥授权可以提供对 AWS 服务的访问权限，从而代表您加密和解密数据。花点时间回顾 [AWS KMS 密钥访问控制的最佳实践](#) 访问 AWS 资源。

最佳实践是监控加密密钥的使用情况，以检测异常的访问模式。使用 AWS 托管密钥和 AWS KMS 中存储的客户自主管理型密钥执行的操作可以记录在 AWS CloudTrail 中，并应定期进行审查。应特别注意监控密钥销毁事件。为了减少意外或恶意破坏密钥材料的情况，密钥销毁事件不会立即删除密钥材

料。尝试删除 AWS KMS 中的密钥时会经历一个 [等待期](#)，默认为 30 天，这让管理员有时间查看这些操作并在必要时回滚请求。

大多数 AWS 服务使用 AWS KMS 的方式对您来说都是透明的，您只需要决定是使用 AWS 托管密钥还是客户自主管理型密钥。如果您的工作负载需要直接使用 AWS KMS 来加密或解密数据，则最佳实践是使用 [信封加密](#) 来保护您的数据。此 [AWS 加密开发工具包](#) 可为您的应用程序提供客户端加密原语，来实施信封加密并与 AWS KMS 集成。

## 实施步骤

1. 为密钥确定合适的 [密钥管理选项](#)（AWS 托管或客户自主管理）。
  - 为便于使用，AWS 为大多数服务提供 AWS 自有和 AWS 托管密钥，这样，无需管理密钥材料或密钥策略，即可提供静态加密功能。
  - 使用客户自主管理型密钥时，请考虑使用默认密钥存储，以便在敏捷性、安全性、数据主权和可用性之间取得最佳平衡。其他用例可能需要使用附带 [AWS CloudHSM](#) 的自定义密钥存储或使用 [外部密钥存储](#) 访问 AWS 资源。
2. 查看您用于工作负载的服务列表，以了解 AWS KMS 如何与该服务集成。例如，EC2 实例可以使用加密的 EBS 卷，验证从这些卷创建的 Amazon EBS 快照是否也使用客户自主管理型密钥进行加密，并减少未加密快照数据的意外泄露。
  - [AWS 服务如何使用 AWS KMS](#)
  - 有关 AWS 服务提供的加密选项的详细信息，请参阅该服务的用户指南或开发人员指南中的“静态加密”主题。
3. 实施 AWS KMS：AWS KMS 使您可以轻松创建和管理密钥，并控制各种 AWS 服务和应用程序中的加密使用情况。
  - [入门：AWS Key Management Service \(AWS KMS\)](#)
  - 请查看 [AWS KMS 密钥访问控制的最佳实践](#) 访问 AWS 资源。
4. 考虑 AWS Encryption SDK：当您的应用程序需要加密客户端数据时，使用包含 AWS KMS 集成的 AWS Encryption SDK。
  - [AWS Encryption SDK](#)
5. 启用 [IAM Access Analyzer](#) 以自动审查是否存在过于宽泛的 AWS KMS 密钥策略并相应地发出通知。
6. 启用 [Security Hub](#) 以便在密钥策略配置错误、计划删除密钥或存在未启用自动轮换的密钥时，接收通知。
7. 确定适合您的 AWS KMS 密钥的日志记录级别。由于对 AWS KMS 的调用（包括只读事件）会被记录下来，因此与 AWS KMS 关联的 CloudTrail 日志可能会变得非常庞大。



- 一些组织倾向于将 AWS KMS 日志活动分成单独的跟踪。有关更多详细信息，请参阅 [使用 CloudTrail 记录 AWS KMS API 调用](#) 部分（位于 AWS KMS 开发者指南中）。

## 资源

### 相关文档：

- [AWS Key Management Service](#)
- [AWS 加密服务和工具](#)
- [利用加密保护 Amazon S3 数据](#)
- [信封加密](#)
- [数字主权承诺](#)
- [揭开 AWS KMS 密钥操作的神秘面纱、自带密钥、自定义密钥库和密文可移植性](#)
- [AWS Key Management Service 加密详情](#)

### 相关视频：

- [AWS 中的加密原理](#)
- [在 AWS 上保护您的数据块存储](#)
- [AWS data protection: Using locks, keys, signatures, and certificates](#)

### 相关示例：

- [使用 AWS KMS 实施高级访问控制机制](#)

## SEC08-BP02 强制实施静态加密

您应该强制对静态数据使用加密。加密后，即使遭到未经授权访问或意外泄露，也能保持敏感数据的机密性。

期望结果：私有数据在静态时应默认加密。加密有助于保持数据的机密性，并提供额外一层保护，防止有意或无意的数据泄露或外流。如果不首先对数据进行解密，则无法读取或访问加密的数据。任何未加密便存储的数据都应进行清点和控制。

### 常见反模式：

- 不使用默认加密配置。
- 提供对解密密钥过于宽松的访问权限。
- 不监控加密和解密密钥的使用。
- 未加密便存储数据。
- 对所有数据使用相同的解密密钥，而不考虑数据用途、类型和分类。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

将解密密钥映射到工作负载中的数据分类。当对数据使用单个或非常少量的解密密钥时，这种方法有助于防止过于宽松的访问（请参阅[SEC07-BP01 识别工作负载内的数据](#)）。

AWS Key Management Service ( AWS KMS ) 与许多 AWS 服务集成，使加密静态数据更加轻松。例如，在 Amazon Simple Storage Service ( Amazon S3 ) 中，您可以对存储桶设置[默认加密](#)，以自动加密新对象。使用 AWS KMS 时，请考虑需要对数据进行多严格的限制。默认和服务控制的 AWS KMS 密钥由 AWS 代表您进行管理和使用。对于需要对底层解密密钥进行精细访问的敏感数据，请考虑使用客户自主管理型密钥 ( CMK )。您可以完全控制 CMK，包括通过使用密钥策略进行轮换和访问管理。

此外，[Amazon Elastic Compute Cloud \( Amazon EC2 \)](#) 和 [Amazon S3](#) 支持通过设置默认加密来强制加密。您可以使用 [AWS Config 规则](#) 自动检查您已使用了加密，例如针对 [Amazon Elastic Block Store \( Amazon EBS \) 卷](#)、[Amazon Relational Database Service \( Amazon RDS \) 实例](#) 和 [Amazon S3 存储桶](#)。

AWS 还提供客户端加密选项，使您能够在将数据上传到云之前对数据进行加密。AWS Encryption SDK 提供了一种使用[信封加密](#)对数据进行加密的方法。您提供包装密钥，AWS Encryption SDK 为它加密的每个数据对象生成一个唯一数据密钥。如果需要托管的单租户硬件安全模块 ( HSM )，请考虑 AWS CloudHSM。AWS CloudHSM 使您可在通过 FIPS 140-2 Level 3 验证的 HSM 上生成、导入和管理解密密钥。AWS CloudHSM 的一些使用案例包括保护用于签发证书颁发机构 ( CA ) 的私有密钥，以及为 Oracle 数据库启用透明数据加密 ( TDE )。AWS CloudHSM 客户端开发工具包提供的软件使您可在将数据上传到 AWS 之前，使用存储在 AWS CloudHSM 中的密钥对客户端数据进行加密。Amazon DynamoDB Encryption Client 还使您可在将项目上传到 DynamoDB 表之前，对项目进行加密和签名。

## 实施步骤

- 对 Amazon S3 强制实施静态加密：实施 [Amazon S3 存储桶默认加密](#)。

为新的 Amazon EBS 卷配置[默认加密](#)：指定所有新创建的 Amazon EBS 卷要以加密形式创建，并选择使用 AWS 提供的默认密钥或您创建的密钥。

配置加密亚马逊云机器镜像（AMI）：通过复制启用加密功能的现有 AMI，可自动加密根卷和快照。

配置 [Amazon RDS 加密](#)：通过使用加密选项，配置对您的 Amazon RDS 数据库集群和静态快照的加密。

使用策略创建和配置 AWS KMS 密钥，以限制对每个数据分类的相应主体的访问：例如，创建一个 AWS KMS 密钥用于加密生产数据，创建一个不同的密钥用于加密开发或测试数据。您还可以提供对其他 AWS 账户的密钥访问权限。不妨考虑分开设立开发环境和生产环境的账户。如果您的生产环境需要解密开发账户中的构件，您可以编辑用于加密开发构件的 CMK 策略，使生产账户有能力解密这些构件。然后，生产环境可以摄取解密后的数据以用于生产。

在其他 AWS 服务中配置加密：对于您使用的其他 AWS 服务，请查看该服务的[安全文档](#)，以确定该服务的加密选项。

## 资源

相关文档：

- [AWS 加密工具](#)
- [AWS 文档](#)
- [AWS Encryption SDK](#)
- [AWS KMS 加密详情白皮书](#)
- [AWS Key Management Service](#)
- [AWS 加密服务和工具](#)
- [Amazon EBS 加密](#)
- [Amazon EBS 卷的默认加密](#)
- [加密 Amazon RDS 资源](#)
- [如何为 Amazon S3 存储桶启用默认加密？](#)
- [利用加密保护 Amazon S3 数据](#)

相关视频：

- [AWS 中的加密原理](#)

- [在 AWS 上保护您的数据块存储](#)

## SEC08-BP03 自动执行静态数据保护

利用自动化工具持续验证和实施静态数据控制措施，例如，确保只存在经过加密的存储资源。您可以[自动确认所有 EBS 卷都已经过加密](#)，方法是使用 [AWS Config 规则](#)。[AWS Security Hub](#) 还可以按照安全标准执行自动化检查，以验证多种不同的控制措施。此外，您的 AWS Config 规则可以自动[修复不合规的资源](#)。

未建立此最佳实践暴露的风险等级：中

### 实施指导

静态数据 代表您在工作负载期间的任意时间段内保留在非易失性存储器中的任何数据。其中包括数据块存储、对象存储、数据库、存档、IoT 设备和用来保留数据的任何其他存储介质。在实施了加密和适当的访问控制时，保护静态数据可以降低未经授权访问的风险。

强制实施静态加密：您应确保只以加密的方式存储数据。AWS KMS 与很多 AWS 服务无缝集成，使您能够更轻松地加密所有静态数据。例如，在 Amazon Simple Storage Service (Amazon S3) 中，您可以对存储桶设置[默认加密](#)，以自动加密所有的新对象。此外，[Amazon EC2](#) 和 [Amazon S3](#) 支持通过设置默认加密来强制加密。您可以使用 [AWS 托管 Config 规则](#) 自动检查您已使用了加密，例如针对 [EBS 卷](#)、[Amazon Relational Database Service \(Amazon RDS\) 实例](#) 和 [Amazon S3 存储桶](#)。

### 资源

相关文档：

- [AWS 加密工具](#)
- [AWS 加密开发工具包](#)

相关视频：

- [AWS 中的加密原理](#)
- [在 AWS 上保护您的数据块存储](#)

## SEC08-BP04 强制实施访问控制

为了帮助保护静态数据，请使用隔离和版本控制等机制强制实施访问控制，并应用最低权限原则。防止允许公众访问您的数据。

期望结果：验证只有获得授权的用户才能按照“需要知晓”的原则访问数据。通过定期备份和版本控制来保护您的数据，防止数据被有意或无意地修改或删除。将关键数据与其他数据隔离，以保护其机密性和数据完整性。

常见反模式：

- 将具有不同敏感度要求或分类的数据存储在一起。
- 解密密钥的权限过于宽松。
- 数据分类不当。
- 不保留重要数据的详细备份。
- 提供对生产数据的持久访问。
- 未审计数据访问，也未定期检查权限

在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

多项控制措施可以帮助保护静态数据，包括访问（使用最低权限）、隔离和版本控制。您应使用检测性机制（例如 AWS CloudTrail）和服务级别日志（例如 Amazon Simple Storage Service（Amazon S3）访问日志），审计对您的数据进行的访问。您应清点可公开访问的数据，并制定一份计划，以便随着时间的推移减少公开可用的数据量。

Amazon S3 Glacier 文件库锁定和 Amazon S3 对象锁定可为 Amazon S3 中的对象提供强制访问控制 – 利用合规性选项锁定文件库策略之后，在锁定过期之前，即使根用户也无法对其进行更改。

## 实施步骤

- 强制实施访问控制：强制实施最低权限访问控制，包括对加密密钥的访问。
- 根据不同分类级别隔离数据：针对数据分类级别使用不同的 AWS 账户，并使用 [AWS Organizations](#) 管理这些账户。
- 查看 AWS Key Management Service（AWS KMS）策略：[查看 AWS KMS 策略中授予的访问级别](#)。
- 查看 Amazon S3 存储桶和对象权限：定期查看 S3 存储桶策略中授予的访问级别。最佳做法是避免使用可公开读取或写入的存储桶。考虑使用 [AWS Config](#) 检测可公开访问的存储桶，并使用 Amazon CloudFront 提供 Amazon S3 中的内容。验证正确配置了不允许公开访问的存储桶，以防止公开访问。默认情况下，所有 S3 存储桶都是私有的，只有被明确授予访问权限的用户才可以访问。

- 启用 [AWS IAM Access Analyzer](#) : IAM Access Analyzer 可对 Amazon S3 存储桶进行分析，并在 [S3 策略授予外部实体访问权限](#) 时生成结果。
- 在适当情况下启用 [Amazon S3 版本控制](#) 和 [对象锁定](#)。
- 使用 [Amazon S3 清单](#) : Amazon S3 清单可用来审计和报告 S3 对象的复制和加密状态。
- 查看 [Amazon EBS](#) 和 [AMI 共享](#) 权限 : 共享权限将使得镜像和卷能够与您的工作负载外部的 AWS 账户共享。
- 查看 [AWS Resource Access Manager](#) : 定期共享，以确定是否应继续共享资源。您可以通过 Resource Access Manager 在您的 Amazon VPC 内共享资源，如 AWS Network Firewall 策略、Amazon Route 53 解析器规则和子网。定期审计共享资源，停止共享不再需要共享的资源。

## 资源

### 相关最佳实践：

- [SEC03-BP01 定义访问要求](#)
- [SEC03-BP02 授予最低访问权限](#)

### 相关文档：

- [AWS KMS 加密详情白皮书](#)
- [管理对 Amazon S3 资源的访问权限简介](#)
- [管理对 AWS KMS 资源的访问权限概览](#)
- [AWS Config 规则](#)
- [Amazon S3 + Amazon CloudFront：云中的天作之合](#)
- [使用版本控制](#)
- [使用 Amazon S3 对象锁定来锁定对象](#)
- [共享 Amazon EBS 快照](#)
- [已共享的 AMI](#)
- [在 Amazon S3 上托管单页应用程序](#)

### 相关视频：

- [在 AWS 上保护您的数据块存储](#)

## SEC08-BP05 使用机制限制对数据的访问

禁止所有用户直接访问正常运行环境中的敏感数据和系统。例如，利用变更管理工作流程，借助工具管理 Amazon Elastic Compute Cloud ( Amazon EC2 ) 实例，而不是允许直接访问或通过堡垒主机进行访问。这可以使用 [AWS Systems Manager Automation](#) 来实现，此功能将使用 [包含您的任务执行步骤的](#) 自动化文档。这些文档可以存储在源代码控制中、在运行之前接受对等审核，并接受全面测试以便最大程度降低风险（与 shell 访问相比）。企业用户可以使用一个仪表板而不是通过直接访问数据存储库来执行查询。当未使用 CI/CD 管道时，确定需要利用哪些控制措施和流程来充分提供通常禁用的 Break Glass 访问机制。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

- 实施可限制对数据的访问的机制：这些机制包括使用控制面板（例如 QuickSight），以向用户显示数据，而不是直接查询。
  - [QuickSight](#)
- 自动管理配置：远程执行操作，使用配置管理服务或工具自动实施安全配置并对其进行验证。避免使用堡垒主机或直接访问 EC2 实例。
  - [AWS Systems Manager](#)
  - [AWS CloudFormation](#)
  - [AWS 上适用于 AWS CloudFormation 模板的 CI/CD 管道](#)

### 资源

相关文档：

- [AWS KMS 加密详情白皮书](#)

相关视频：

- [AWS 中的加密原理](#)
- [在 AWS 上保护您的数据块存储](#)

## SEC 9.如何保护传输中的数据？

通过实施多个控制措施来保护传输中的数据，以降低未经授权的访问或数据丢失所带来的风险。



## 最佳实践

- [SEC09-BP01 实施安全密钥和证书管理](#)
- [SEC09-BP02 在传输中执行加密](#)
- [SEC09-BP03 自动检测意外数据访问](#)
- [SEC09-BP04 对网络通信进行身份验证](#)

### SEC09-BP01 实施安全密钥和证书管理

传输层安全性协议 ( TLS ) 证书用于保障网络通信的安全，确立网站、资源和工作负载在互联网上以及专用网络上的身份。

期望的结果：一个安全的证书管理系统，可以在公钥基础设施 ( PKI , Public Key Infrastructure ) 中预置、部署、存储和续订证书。安全密钥和证书管理机制可防止证书私钥材料泄露，并定期自动续订证书。它还与其他服务集成，为工作负载内的计算机资源提供安全的网络通信和标识。密钥材料永远不应能够通过人员的身份来访问。

常见反模式：

- 在证书部署或续订流程中执行人工步骤。
- 在设计私有证书颁发机构 ( CA , Certificate Authority ) 时，对 CA 层次结构的关注不够。
- 对公共资源使用自签名证书。

建立此最佳实践的好处：

- 通过自动化的部署和续订流程简化证书管理
- 鼓励使用 TLS 证书对传输中数据进行加密
- 提高了证书颁发机构执行的证书操作的安全性和可审计性
- 在 CA 层次结构的不同层级上组织管理职责

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

现代化工作负载广泛使用通过 PKI 协议 ( 如 TLS ) 进行加密的网络通信。PKI 证书管理可能很复杂，但是，通过自动化的证书预置、部署和续订机制，可以减少与证书管理相关的麻烦。

AWS 提供了两种服务用于管理通用 PKI 证书：[AWS Certificate Manager](#) 和 [AWS Private Certificate Authority \( AWS Private CA \)](#)。ACM 是客户用于预置、管理和部署证书的主要服务，适用于面向公众的工作负载以及私有 AWS 工作负载。ACM 使用 AWS Private CA 颁发证书，并 [集成](#) 许多其他 AWS 托管服务，用于向工作负载提供安全的 TLS 证书。

利用 AWS Private CA，您可以建立自己的根证书颁发机构或从属证书颁发机构，并通过 API 颁发 TLS 证书。在 TLS 连接的客户端一侧控制和管理信任链的场景中，您可以使用这些类型的证书。除了 TLS 使用场景外，还可以使用 AWS Private CA 通过 [自定义模板](#) 向 Kubernetes 容器组 ( pod )、Matter 设备产品认证、代码签名和其他使用场景颁发证书。您还可以使用 [IAM Roles Anywhere](#)，向已经为其颁发了 X.509 证书 ( 使用您的私有 CA 签名 ) 的本地工作负载，提供临时 IAM 凭证。

除了 ACM 和 AWS Private CA 之外，[AWS IoT Core](#) 针对为物联网设备预置、管理和部署 PKI 证书提供专业化支持。AWS IoT Core 提供专门的机制，用于大规模 [将物联网设备载入](#) 到您的公钥基础设施中。

### 建立私有 CA 层次结构的注意事项

当您需要建立私有 CA 时，请务必重视预先正确设计 CA 层次结构。在创建私有 CA 层次结构时，最佳实践是将 CA 层次结构的每个级别部署到单独的 AWS 账户中。这个有意而为的步骤可减少 CA 层次结构中每个级别的暴露范围，使得发现 CloudTrail 日志数据中的异常变得更加简单，并可在某个账户遭到未经授权的访问时，缩小访问或影响的范围。根 CA 应位于自己的独立账户中，并且只能用于发布一个或多个中间 CA 证书。

然后，在不同于根 CA 账户的账户中创建一个或多个中间 CA，为最终用户、设备或其他工作负载发布证书。最后，从您的根 CA 向中间 CA 颁发证书，后者随之向您的最终用户或设备颁发证书。有关规划 CA 部署和设计 CA 层次结构 ( 包括弹性规划、跨区域复制、在组织中共享 CA 等 ) 的更多信息，请参阅 [规划 AWS Private CA 部署](#)。

### 实施步骤

#### 1. 确定您的使用场景所需的相关 AWS 服务：

- 许多使用场景都可以利用现有的 AWS 公钥基础设施并使用 [AWS Certificate Manager](#)。ACM 可用于为 Web 服务器、负载均衡器或公共可信证书的其他用途部署 TLS 证书。
- 在您需要建立自己的私有证书颁发机构层次结构或需要使用可导出证书时，请考虑 [AWS Private CA](#)。然后，可以使用 ACM 颁发 [多种类型的终端实体证书](#) ( 使用 AWS Private CA )。
- 对于必须为嵌入式物联网 ( IoT ) 设备大规模预置证书的使用场景，请考虑使用 [AWS IoT Core](#)。

#### 2. 尽可能实施自动证书续订：

- 将 [ACM 托管续订](#) 用于 ACM 颁发的证书以及集成的 AWS 托管服务。

### 3. 建立日志记录和审计跟踪：

- 启用 [CloudTrail 日志](#)，以便跟踪对具有证书颁发机构的账户的访问。请考虑在 CloudTrail 中配置日志文件完整性验证，用于验证日志数据的真实性。
- 定期生成和审查 [审计报告](#)，列出您的私有 CA 已颁发或撤销的证书。这些报告可以导出到 S3 存储桶。
- 部署私有 CA 时，您还需要创建一个 S3 存储桶，用于存储证书撤销列表 (CRL, Certificate Revocation List)。有关根据工作负载要求配置此 S3 存储桶的指南，请参阅 [计划证书撤销列表 \(CRL\)](#)。

#### 资源

##### 相关最佳实践：

- [SEC02-BP02 使用临时凭证](#)
- [SEC08-BP01 实施安全密钥管理](#)
- [SEC09-BP04 对网络通信进行身份验证](#)

##### 相关文档：

- [How to host and manage an entire private certificate infrastructure in AWS](#)
- [How to secure an enterprise scale ACM Private CA hierarchy for automotive and manufacturing](#)
- [Private CA best practices](#)
- [How to use AWS RAM to share your ACM Private CA cross-account](#)

##### 相关视频：

- [Activating AWS Certificate Manager Private CA \(研讨会\)](#)

##### 相关示例：

- [Private CA workshop](#)
- [物联网设备管理研讨会](#) (包括设备预置)

##### 相关工具：

- [使用 AWS Private CA 的 Kubernetes 证书管理器插件](#)

## SEC09-BP02 在传输中执行加密

根据贵组织的政策、监管义务和标准定义，实施加密要求，以帮助满足组织、法律和合规性要求。如要在虚拟私有云 ( VPC ) 外部传输敏感数据，务必仅使用具有加密功能的协议。即使数据在不可信的网络中传输，加密也有助于保持数据的机密性。

期望结果：所有数据在传输过程中都应使用安全的 TLS 协议和密码套件进行加密。必须对资源和互联网之间的网络流量进行加密，以减少对数据的未经授权访问。对于仅在您内部 AWS 环境中的网络流量，应尽可能使用 TLS 进行加密。默认情况下，AWS 内部网络进行了加密，除非未经授权的一方能够访问正在生成流量的任何资源（如 Amazon EC2 实例和 Amazon ECS 容器），否则无法假冒或嗅探 VPC 内的网络流量。不妨考虑使用 IPsec 虚拟专用网络 ( VPN ) 保护网络到网络流量。

常见反模式：

- 使用已弃用的 SSL、TLS 和密码套件组件版本（例如，SSL v3.0、1024 位 RSA 密钥和 RC4 密码）。
- 允许未加密的 ( HTTP ) 流量进出面向公众的资源。
- 未在 X.509 证书到期前监控和替换证书。
- 对 TLS 使用自签名 X.509 证书。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

AWS 服务提供使用 TLS 的 HTTPS 端点进行通信，从而可以在与 AWS API 通信时提供传输中加密。可以使用安全组在 VPC 中审计和拦截不安全的协议，例如 HTTP。也可以在 Amazon CloudFront 中或 [Application Load Balancer](#) 上，将 HTTP 请求 [自动重定向到 HTTPS](#)。您可以完全控制计算资源，以便在整个服务中实施加密。您也可以利用 VPN 连接从外部网络或 [AWS Direct Connect](#) 连接到您的 VPC 中，以便于对流量进行加密。请验证您的客户端至少使用 TLS 1.2 调用 AWS API，因为 [AWS 将在 2023 年 6 月弃用 TLS 1.0 和 1.1](#)。如果您有特殊要求，可以使用 AWS Marketplace 中提供的第三方解决方案。

### 实施步骤

- 实施传输中加密：您定义的加密要求应基于最新的标准和最佳实践，且仅允许使用安全协议。例如，配置一个安全组，仅允许通过 HTTPS 协议访问应用程序负载均衡器或 Amazon EC2 实例。

- 在边缘服务中配置安全协议：[使用 Amazon CloudFront 配置 HTTPS](#)，并使用[适合您的安全状况和使用案例的安全配置文件](#)。
- 将[VPN 用于外部连接](#)：考虑使用 IPsec VPN 来保护点对点或网络对网络连接，以帮助实现数据隐私和完整性。
- 在负载均衡器中配置安全协议：选择一个安全策略，该策略提供受客户端支持且将要连接到侦听器的强大密码套件。[为 Application Load Balancer 创建 HTTPS 侦听器](#)。
- 在 Amazon Redshift 中配置安全协议：将集群配置为要求[安全套接字层 \( SSL \) 或传输层安全性协议 \( TLS \) 连接](#)。
- 配置安全协议：查看 AWS 服务文档，以确定传输中加密功能。
- 上传到 Amazon S3 存储桶时配置安全访问：使用 Amazon S3 存储桶策略控制措施[执行对数据的安全访问](#)。
- 不妨考虑使用 [AWS Certificate Manager](#)：ACM 允许您预置、管理和部署用于 AWS 服务的公有 TLS 证书。
- 不妨考虑使用 [AWS Private Certificate Authority](#) 满足私有 PKI 需求：AWS Private CA 允许您创建私有证书颁发机构 ( CA ) 层次结构，以签发可用于创建加密 TLS 通道的终端实体 X.509 证书。

## 资源

### 相关文档：

- [AWS 文档](#)
- [将 HTTPS 与 CloudFront 搭配使用](#)
- [使用 AWS Virtual Private Network 将 VPC 连接到远程网络](#)
- [为 Application Load Balancer 创建 HTTPS 侦听器](#)
- [教程：在 Amazon Linux 2 上配置 SSL/TLS](#)
- [使用 SSL/TLS 加密与数据库实例的连接](#)
- [针对连接配置安全选项](#)

## SEC09-BP03 自动检测意外数据访问

使用 Amazon GuardDuty 等工具自动检测可疑活动或尝试将数据移动到定义的边界之外。例如，GuardDuty 可以通过以下方法，检测异常的 Amazon Simple Storage Service ( Amazon S3 ) 读取活动：[Exfiltration:S3/AnomalousBehavior finding](#)。除了 GuardDuty 以外，还可以将[Amazon VPC 流日志](#) ( 用于捕获网络流量信息 ) 与 Amazon EventBridge 配合使用，以触发对已成功和被拒绝的异常连

接的检测。[Amazon S3 Access Analyzer](#) 可以帮助评估您的 Amazon S3 存储桶中的哪些数据可供哪些人访问。

未建立此最佳实践暴露的风险等级：中

### 实施指导

- 自动检测意外数据访问：使用工具或检测机制自动检测试图将数据移出定义边界的行为；例如，检测正在将数据复制到无法识别的主机的数据库系统。
- [VPC 流日志](#)
- 考虑 Amazon Macie：Amazon Macie 是一项完全托管式数据安全和数据隐私服务，该服务使用机器学习和模式匹配发现和保护 AWS 中的敏感数据。
- [Amazon Macie](#)

### 资源

相关文档：

- [VPC 流日志](#)
- [Amazon Macie](#)

## SEC09-BP04 对网络通信进行身份验证

使用传输层安全性 (TLS) 或 IPsec 等支持身份验证的协议来验证通信的身份。

将您的工作负载设计为在服务 and 应用程序之间通信或与用户通信时，使用经过身份验证的安全网络协议。使用支持身份验证和授权的网络协议，可以加强对网络流量的控制，并减少未经授权访问的影响。

期望结果：工作负载具有明确定义的数据面板和控制面板，它们控制流量在服务之间的流动。在技术上可行的情况下，流量将使用经过身份验证和加密的网络协议。

常见反面模式：

- 工作负载中存在未加密或未经身份验证的流量。
- 在多个用户或实体之间重用身份验证凭证。
- 仅依赖网络控制作为访问控制机制。
- 创建自定义身份验证机制，而不是依赖行业通用身份验证机制。

- VPC 中的服务组件或其他资源之间有过于宽松的流量流动。

建立此最佳实践的好处：

- 限制未经授权访问工作负载某一部分的影响范围。
- 提供更高级别的保障，即操作只能由经过身份验证的实体执行。
- 通过明确定义和强制执行预期的数据传输接口，改善服务的解耦。
- 通过请求归因和明确定义的通信界面，增强监控、日志记录和事件响应。
- 通过将网络控制与身份验证和授权控制相结合，为您的工作负载提供深度防御。

在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

您的工作负载的网络流量模式可分为两类：

- 东西向流量 代表构成工作负载的服务之间的流量。
- 南北向流量 代表您的工作负载和使用器之间的流量。

对南北向流量进行加密是常见做法，而使用经过身份验证的协议保护东西向流量则不太常见。现代安全实践建议，仅靠网络设计并不足以在两个实体之间建立可信关系。当两项服务可能位于公共网络边界内时，最佳做法仍然是对这些服务之间的通信进行加密、身份验证和授权。

例如，无论请求来自哪个网络，AWS 服务 API 都使用 [AWS 签名版本 4 \( SigV4 \)](#) 签名协议对调用方进行身份验证。这种身份验证可确保 AWS API 可以验证请求操作的身份，然后将该身份与策略结合起来，作出授权决策，以确定是否应该允许该操作。

[Amazon VPC Lattice](#) 和 [Amazon API Gateway](#) 等服务允许您使用相同的 SigV4 签名协议，为自己的工作负载中的东西向流量添加身份验证和授权。如果您的 AWS 环境之外的资源要与服务进行通信，而服务需要基于 Sigv4 的身份验证和授权，则您可以对非 AWS 资源使用 [AWS Identity and Access Management \( IAM \) Roles Anywhere](#) 来获取临时 AWS 凭证。这些凭证可用于对使用 SigV4 的服务请求进行签名，以授权访问权限。

验证东西向流量的另一种常见机制是 TLS 双向身份验证 ( mTLS )。许多物联网 ( IoT )、企业对企业应用程序和微服务都使用 mTLS，通过使用客户端和服务端 X.509 证书来验证 TLS 通信两端的身份。这些证书可以由 AWS Private Certificate Authority ( AWS Private CA ) 颁发。您可以使用



[Amazon API Gateway](#) 和 [AWS App Mesh](#) 等服务，针对工作负载间或工作负载内的通信提供 mTLS 身份验证。虽然 mTLS 为 TLS 通信的两端提供身份验证信息，但它不提供授权机制。

最后，OAuth 2.0 和 OpenID Connect ( OIDC ) 这两种协议通常用于控制用户对服务的访问，但如今在服务间流量中也变得越来越流行。API Gateway 提供了 [JSON Web 令牌 \( JWT \) 授权方](#)，允许工作负载使用 OIDC 或 OAuth 2.0 身份提供商颁发的 JWT 来限制对 API 路由的访问。可依据 OAuth2 范围来作出基本授权决策，但授权检查仍需要在应用层实现，仅靠 OAuth2 范围无法支持更复杂的授权需求。

## 实施步骤

- 定义并记录您的工作负载网络流：实施深度防御策略的第一步是定义工作负载的流量。
  - 创建数据流示意图，明确定义构成工作负载的不同服务之间如何传输数据。此示意图是强制这些流量通过经身份验证的网络渠道传输的第一步。
  - 在开发和测试阶段对您的工作负载进行检测，以验证数据流示意图是否准确反映了工作负载在运行时的行为。
  - 在执行威胁建模练习时，数据流示意图可能也很有用，如 [SEC01-BP07 使用威胁模型识别威胁并确定缓解措施的优先级](#) 中所述。
- 建立网络控制：考虑使用 AWS 功能建立起与数据流相符的网络控制。虽然网络边界不应该是唯一的安全控制措施，但它们在深度防御策略中提供了一个安全层来保护您的工作负载。
  - 使用 [安全组](#) 来建立、定义和限制资源之间的数据流。
  - 考虑使用 [AWS PrivateLink](#) 与 AWS 以及支持 AWS PrivateLink 的第三方服务通信。通过 AWS PrivateLink 接口端点发送的数据保留在 AWS 网络主干内，而不通过公共互联网传输。
- 在工作负载中实施跨服务的身份验证和授权：选择最适合在工作负载中提供经过身份验证的加密流量的一组 AWS 服务。
  - 考虑使用 [Amazon VPC Lattice](#) 来保护服务间的通信。VPC Lattice 可以结合使用 [Sigv4 身份验证与身份验证策略](#) 来控制服务间的访问。
  - 对于使用 mTLS 进行的服务间通信，请考虑使用 [API Gateway](#) 或 [App Mesh](#)。[AWS Private CA](#) 可用于建立能够颁发与 mTLS 结合使用的证书的私有 CA 层次结构。
  - 与使用 OAuth 2.0 或 OIDC 的服务集成时，可以考虑 [使用 JWT 授权方的 API Gateway](#)。
  - 对于工作负载和物联网设备之间的通信，可以考虑使用 [AWS IoT Core](#)，它提供多种网络流量加密和身份验证选项。
- 监控未经授权的访问：持续监控非预期的通信渠道、试图访问受保护资源的未授权主体以及其他不当访问模式。
  - 如果使用 VPC Lattice 来管理对服务的访问，请考虑启用和监控 [VPC Lattice 访问日志](#)。这些访问日志包括有关请求实体的信息、源和目的地 VPC 等网络信息以及请求元数据。

- 考虑启用 [VPC 流日志](#)，以捕获网络流量的元数据并定期检查是否存在异常。
- 有关规划、模拟和响应安全事件的更多指导，请参阅 [AWS 安全事件响应指南](#) 和 AWS Well-Architected Framework 安全性支柱的[“事件响应”部分](#)。

## 资源

相关最佳实践：

- [SEC03-BP07 分析公共和跨账户访问](#)
- [SEC02-BP02 使用临时凭证](#)
- [SEC01-BP07 使用威胁模型识别威胁并确定缓解措施的优先级](#)

相关文档：

- [评估用于保护 Amazon API Gateway API 的访问控制方法](#)
- [为 REST API 配置双向 TLS 身份验证](#)
- [如何使用 JWT 授权方保护 API Gateway HTTP 端点](#)
- [使用 AWS IoT Core 凭证提供程序授权直接调用 AWS 服务](#)
- [AWS 安全事件响应指南](#)

相关视频：

- [AWS re: invent 2022：VPC Lattice 简介](#)
- [AWS re: invent 2020：AWS 上 HTTP API 的无服务器 API 身份验证](#)

相关示例：

- [Amazon VPC Lattice 研讨会](#)
- [零信任第 1 集 — 幻影服务边界研讨会](#)

## 事件响应

### 问题

- [SEC 10.如何预测、响应事件以及从事件中恢复？](#)

## SEC 10.如何预测、响应事件以及从事件中恢复？

即使采用成熟的预防和检测性控制措施，您的组织也应实施机制来响应安全事件并缓解安全事件可能带来的影响。您的准备工作会极大地影响团队在事件发生期间采取有效行动、对问题进行隔离、遏制和取证并将运行状态恢复到已知良好状态的能力。在安全事件发生之前确保相关工具和访问权限部署到位，然后通过实际试用定期进行事件响应演练，这样有助于确保您有能力恢复并最大限度避免业务中断。

### 最佳实践

- [SEC10-BP01 确定关键人员和外部资源](#)
- [SEC10-BP02 制定事件管理计划](#)
- [SEC10-BP03 准备取证能力](#)
- [SEC10-BP04 制定和测试安全事件响应行动手册](#)
- [SEC10-BP05 预置访问权限](#)
- [SEC10-BP06 预部署工具](#)
- [SEC10-BP07 运行模拟](#)
- [SEC10-BP08 建立从事件中吸取经验教训的框架](#)

### SEC10-BP01 确定关键人员和外部资源

确定能够帮助您的组织响应事件的内部和外部人员、资源、以及法律义务。

当您与其他团队（例如法律顾问、领导、业务利益相关者、AWS Support 服务等）一起在云中定义您的事件响应方法时，您必须确定关键人员、利益相关者和相关联系人。为了降低依赖性并缩短响应时间，请确保为您的团队、专家安全团队和响应者开展培训，以使它们了解您使用的服务并有机会练习动手实践。

我们鼓励您寻找外部 AWS 安全合作伙伴，他们应当能够为您带来外部专业知识和不同的视角，以增强您的响应能力。您的可靠安全合作伙伴可以帮助您发现您可能并不熟悉的潜在风险或威胁。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

- 确定组织中的关键人员：制定联系人列表，列出组织内需要参与事件响应和事件恢复的人员。
- 确定外部合作伙伴：根据需要联系能够帮助您响应事件并从事件中恢复的外部合作伙伴。

## 资源

### 相关文档：

- [AWS 事件响应指南](#)

### 相关视频：

- [准备和响应 AWS 环境中的安全事件](#)

### 相关示例：

#### SEC10-BP02 制定事件管理计划

为事件响应制定的第一个文档是事件响应计划。事件响应计划旨在为您的事件响应计划和战略奠定基础。

建立此最佳实践的好处：要想成功实现可扩展的事件响应计划，制定全面且明确定义的事件响应流程是关键。在发生安全事件时，明确的步骤和工作流有助于您及时做出响应。您可能已经有事故响应流程。无论您当前的状态如何，定期更新、迭代和测试事件响应流程都很重要。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

对于响应、缓解安全事件的潜在影响并从中恢复来说，事件管理计划是至关重要的。事件管理计划是一个结构化的过程，用于及时地确定、补救和响应安全事件。

云的许多操作角色和要求都与本地环境中的相同。在创建事件管理计划时，应考虑最符合业务成果和合规性要求的响应和恢复策略，这一点非常重要。例如，如果您在 AWS 中运行在美国符合 FedRAMP 标准的工作负载，则遵守 [《NIST SP 800-61 计算机安全处理指南》](#) 会很有帮助。同样，在使用欧洲个人信息（PII）数据运行工作负载时，请考虑具体的场景，例如如何根据以下条例的强制要求，保护和应对与数据驻留相关的问题：[欧盟通用数据保护条例（GDPR）](#)。

在为 AWS 中的工作负载制定事件管理计划时，请首先使用 [AWS 责任共担模式](#)，以便构建针对事件响应的深度防御方法。在此模式中，AWS 负责管理云本身的安全，云内部的安全则由您负责。这意味着您将保留控制权，并对选择实施的安全控制机制负责。此 [AWS 安全事件响应指南](#) 详细介绍了构建以云为中心的事件管理计划的关键概念和基本指南。

必须不断地迭代有效的事件管理计划，使其与您的云运营目标保持一致。在创建和改进事件管理计划时，请考虑使用下面详述的实施计划。

## 实施步骤

### 定义角色和职责

处理安全事件需要在整个组织中落实纪律要求和行动意愿。在您的组织结构中，发生事件时，负责、追责、咨询或者告知信息等各个环节会涉及到不同的人员，例如人力资源（HR）、高管团队和法律部门的代表。请考虑这些角色和职责，以及是否必须有第三方参与。请注意，许多地区的当地法律都规定了，哪些事情能做，哪些事情不能做。尽管为安全响应计划建立一个负责、问责、咨询和知情（RACI）的图表可能显得过于繁文缛节，但这样做有利于快速直接地进行沟通，并清楚地概述在事件不同阶段负责的领导层。

在事件发生期间，让受影响应用程序和资源的负责人和开发人员参与进来非常关键，因为这些人员是主题专家（SME），可以提供信息和背景情况来协助衡量影响。您应该先练习并与开发人员和应用程序负责人建立关系，然后才能依靠他们的专业知识进行事件响应。应用程序负责人或 SME，如云管理员或工程师，可能需要在不熟悉环境、面临复杂情况或响应人员没有访问权限的情况下采取行动。

最后，值得信赖的合作伙伴可以参与到调查或响应中，因为他们可以提供额外的专业知识和宝贵的审查工作。当您自己的团队缺乏具备这些技能的人员时，您可能需要聘请外部人员寻求帮助。

### 了解 AWS 响应团队和支持

- AWS 支持
  - [支持](#) 提供了一系列计划，可供您利用各种工具和专业知识和专业知识，为您的 AWS 解决方案的成功和正常运行提供支持。如果您需要技术支持及更多资源来规划、部署和优化 AWS 环境，则可以选择最符合您 AWS 使用场景的支持计划。
  - 考虑将 [支持中心](#)（在 AWS Management Console 中，需要登录）作为中心联系点，为影响您 AWS 资源的问题获取支持。对 [支持](#) 的访问由 AWS Identity and Access Management 控制。有关获取对 [支持](#) 功能的访问的更多信息，请参阅 [开始使用 支持](#)。
- AWS 客户事件响应团队（CIRT）
  - AWS 客户事件响应团队（CIRT）是一支专业的 AWS 全球团队，全天候向客户提供支持，协助客户解决根据 [AWS 责任共担模式](#) 应由客户一方负责的安全事件。
  - 当 AWS CIRT 为您提供支持时，他们会为 AWS 上出现的安全事件提供分类和恢复方面的协助。他们可以使用 AWS 服务日志来协助分析根本原因，并为您提供恢复建议。他们还可以提供安全建议和最佳实践，从而让您以后能够避免出现安全事件。
  - AWS 客户要与 AWS CIRT 交流，可以开立 [支持 案例](#)。
- DDoS 响应支持
  - AWS 提供 [AWS Shield](#)，它提供了托管的分布式拒绝服务（DDoS）攻击保护服务，可保护在 AWS 上运行的 Web 应用程序。Shield 提供不间断检测和自动化内嵌缓解措施，可以最大限度地

减少应用程序停机时间和延迟，因此无需与支持交流即可从 DDoS 保护中受益。Shield 分为两个级别：AWS Shield Standard 和 AWS Shield Advanced。要了解这两个级别之间的区别，请参阅 [Shield 功能文档](#)。

- AWS Managed Services ( AMS )
  - [AWS Managed Services \( AMS \)](#) 可持续管理您的 AWS 基础设施，让您可以专注于应用程序。AMS 实施最佳实践来维护您的基础设施，让您能够降低运营开销和风险。AMS 可以自动执行常见活动（例如更改请求、监控、补丁管理、安全性和备份服务），并可以提供全生命周期服务来预置、运行和支持您的基础设施。
  - AMS 负责部署一套安全检测控制措施，并全天候提供对警报的第一线响应。启动警报后，AMS 遵循一组标准的自动和手动行动手册，验证是否有一致的响应。这些行动手册在功能部署期间与 AMS 客户共享，这样客户就能够开发并与 AMS 协调响应措施。

## 制定事件响应计划

事件响应计划旨在为您的事件响应计划和战略奠定基础。事件响应计划应包含在正式文档中。事件响应计划通常包括以下部分：

- 事件响应团队概述：概述事件响应团队的目标和职能。
- 角色和职责：列出事件响应利益相关者，并详细说明他们在发生事件时的角色。
- 沟通计划：详细的联系信息，以及在事件发生期间如何进行沟通。
- 后备沟通方法：此时的最佳实践是采用带外通信，作为事件沟通的后备。AWS Wickr 就是一个提供安全的带外通信渠道的应用程序示例。
- 事件响应阶段和应采取的行动：列举事件响应的各个阶段（例如，检测、分析、消除、遏制和恢复），包括在这些阶段中要采取的高级别操作。
- 事件严重性和优先级定义：详细说明如何对事件的严重性进行分类，如何确定事件的优先级，然后详细说明严重性定义对上报程序有何影响。

尽管这些内容部分在各种规模和行业的公司中很常见，但每个组织的事件响应计划都是独一无二的。您需要制定最适合贵组织的事件响应计划。

## 资源

相关最佳实践：

- [SEC04 \( “您如何检测和调查安全事件？” \)](#)



相关文档：

- [AWS 安全事件响应指南](#)
- [NIST：计算机安全事件处理指南](#)

## SEC10-BP03 准备取证能力

在发生安全事件之前，可以考虑构建取证能力来支持安全事件调查工作。

未建立这种最佳实践的情况下暴露的风险等级：中

传统本地取证的概念适用于 AWS。有关开始在 AWS Cloud 中构建取证功能的关键信息，请参阅 [AWS Cloud 中的取证调查环境策略](#)。

设置好取证的环境和 AWS 账户结构后，确定在以下四个阶段有效执行可靠取证方法所需的技术：

- 收集：收集相关的 AWS 日志，例如 AWS CloudTrail、AWS Config、VPC 流日志和主机级日志。收集受影响的 AWS 资源的快照、备份和内存转储（如果有）。
- 检查：通过提取和评测相关信息来检查收集到的数据。
- 分析：分析收集到的数据，以便了解事件并从中得出结论。
- 报告：提供分析阶段得出的信息。

## 实施步骤

### 准备取证环境

[AWS Organizations](#) 有助于您随着 AWS 资源的增长和扩展，集中管理和监管 AWS 环境。AWS 组织会整合您的 AWS 账户，这样您就能够将这些账户作为一个单元进行管理。可以使用组织单位（OU）将账户分组到一起，以便作为一个单元进行管理。

对于事件响应，拥有一个支持事件响应功能的 AWS 账户结构（包括安全 OU 和取证 OU）会很有帮助。在安全 OU 中，您应该拥有以下账户：

- 日志存档：将日志聚合到权限有限的日志存档 AWS 账户中。
- 安全工具：将安全服务集中在安全工具 AWS 账户中。此账户以安全服务的委托管理员身份运行。

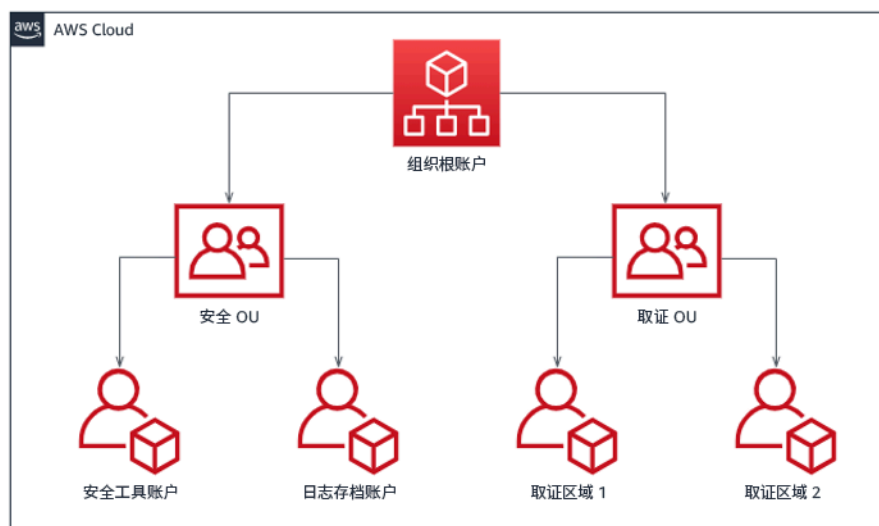
在取证 OU 中，您可以选择实施单一取证账户，也可以为您运营的每个区域实施账户，具体取决于哪种账户最适合您的业务和运营模式。如果为每个区域创建一个取证账户，就可以阻止在该区域之外创建 AWS 资源，降低资源被复制到非预期区域的风险。例如，如果您只在 US East (N. Virginia)



Region ( us-east-1 ) 和 US West (Oregon) ( us-west-2 ) 运营，那么您将在取证 OU 中拥有两个账户：一个用于 us-east-1，另一个用于 us-west-2。

可以为多个区域创建取证 AWS 账户。在将 AWS 资源复制到该账户时应小心谨慎，确保符合数据主权要求。由于预置新账户需要时间，因此必须在事件发生前创建和分析取证账户，以便响应人员能够做好准备，有效地使用这些账户进行响应。

下图显示了一个账户结构示例，其中包括一个取证 OU，涵盖了根据每个区域创建的取证账户：



用于响应事件而根据区域创建的账户结构

## 捕获备份和快照

为关键系统和数据库建立备份对于从安全事件中恢复和取证至关重要。有了备份，您就能够将系统恢复到以前的安全状态。在 AWS 上，您可以创建各种资源的快照。快照为您提供这些资源的时间点备份。有许多 AWS 服务能够在备份和恢复方面为您提供支持。有关这些服务以及备份和恢复方法的详细信息，请参阅 [备份和恢复规范性指南](#) 以及 [使用备份从安全事件中恢复](#)。

特别是遇到勒索软件等情况时，妥善保管备份至关重要。有关保护备份的指导，请参阅 [AWS 中保护备份的 10 大安全最佳实践](#)。除了确保备份安全外，您还应当定期测试备份和还原流程，从而确保现有的技术和流程按预期运行。

## 自动取证

在安全事件期间，您的事件响应团队必须能够快速收集和分析证据，同时保持事件相关时间段的准确性（例如捕获与特定事件或资源相关的日志，或收集 Amazon EC2 实例的内存转储）。对于事件响应团队来说，手动收集相关证据既具有挑战性又很耗时，尤其是在存在大量实例和账户的情况下。此外，手动收集容易出现人为错误。出于这些原因，您应该尽可能开发和实现取证自动化功能。

AWS 提供了大量用于取证的自动化资源，这些资源在下面的“资源”部分中列出。这些资源是我们开发并由客户实施的取证模式示例。虽然这些资源可能是有用的参考架构，但可以考虑根据您的环境、要求、工具和取证流程对资源进行修改，或者创建新的取证自动化模式。

## 资源

### 相关文档：

- [AWS 安全事件响应指南 – 构建取证能力](#)
- [AWS 安全事件响应指南 – 取证资源](#)
- [AWS Cloud中的取证调查环境策略](#)
- [如何在 AWS 中自动实施取证磁盘收集](#)
- [AWS Prescriptive Guidance – 自动化事件响应和取证](#)

### 相关视频：

- [自动化事件响应和取证](#)

### 相关示例：

- [自动事件响应和取证框架](#)
- [Amazon EC2 的自动取证编排工具](#)

## SEC10-BP04 制定和测试安全事件响应行动手册

准备事件响应流程的关键环节是制定行动手册。事件响应行动手册提供了一系列规范性指导和步骤，供发生安全事件时遵循。清晰的结构和步骤可简化响应，减少发生人为错误的可能性。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

应针对以下事件场景创建行动手册：

- 预期事件：应针对预期的事件创建行动手册。这包括拒绝服务 ( DoS )、勒索软件和凭证泄露等威胁。
- 已知的安全调查发现或警报：应针对已知的安全调查发现和警报 ( 如 GuardDuty 调查发现 ) 创建行动手册。您可能会收到一个 GuardDuty 调查发现，然后想：“现在怎么办？”为防止错误处理或忽略

GuardDuty 调查发现，应针对每个可能的 GuardDuty 调查发现创建行动手册。有关补救细节和指导的信息可在 [GuardDuty 文档](#) 中找到。需要注意的是，默认情况下并不会启用 GuardDuty，而且需要付费。有关 GuardDuty 的详细信息，请参阅 [附录 A：云功能定义 – 可见性和警报](#)。

行动手册应包含安全分析师需要完成的技术步骤，以便充分调查和应对潜在的安全事件。

## 实施步骤

行动手册中应包括的项目有：

- 行动手册概述：本行动手册针对哪些风险或事件场景？本行动手册的目标是什么？
- 先决条件：此事件场景需要哪些日志、检测机制和自动化工具？预期的通知是什么？
- 沟通和上报信息：谁参与其中，他们的联系信息是什么？每个利益相关者的责任是什么？
- 响应步骤：在事件响应的各个阶段，应采取哪些战术性措施？分析师应该进行哪些查询？应该运行什么代码才能达到预期的结果？
  - 检测：如何检测事件？
  - 分析：如何确定影响范围？
  - 控制：如何隔离事件来限制其影响范围？
  - 消除：如何从环境中消除威胁？
  - 恢复：受影响的系统或资源将如何恢复生产？
- 预期结果：运行查询和代码后，行动手册的预期结果是什么？

## 资源

相关的 Well-Architected 最佳实践：

- [SEC10-BP02 – 制定事件管理计划](#)

相关文档：

- [事件响应行动手册框架](#)
- [制定自己的事件响应行动手册](#)
- [事件响应行动手册样本](#)
- [使用 Jupyter 行动手册和 CloudTrail Lake 构建 AWS 事件响应运行手册](#)

## SEC10-BP05 预置访问权限

确保事件响应者将正确的访问权限预置到 AWS 中，以缩短调查到恢复所需的时间。

常见反模式：

- 使用根账户进行事件响应。
- 变更现有用户账户。
- 在提供实时权限提升时直接操作 IAM 权限。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

AWS 建议尽可能减少或消除对长期有效凭证的依赖，转而使用临时凭证和实时权限提升机制。长期有效的凭证容易带来安全风险，并且会增加运营开销。对于大多数管理任务以及事件响应任务，建议您对管理访问实施 [身份联合验证](#) 以及 [临时上报](#)。在此模型中，用户请求提升到更高级别的权限（例如事件响应角色），如果用户符合提升条件，则会向审批者发送请求。如果请求获得批准，用户将收到一组临时的 [AWS 凭证](#)，可用于完成用户任务。在这些凭证过期后，用户必须提交新的提升请求。

在大多数事件响应场景中，建议使用临时权限提升。执行此操作的正确方法是使用 [AWS Security Token Service](#) 和 [会话策略](#) 来限定访问范围。

在一些场景中，联合身份不可用，例如：

- 与被盗用的身份提供者（IdP）相关的中断。
- 导致联合访问管理系统损坏的错误配置或人为错误。
- 恶意活动，例如分布式拒绝服务（DDoS，Distributed Denial of Service）事件或导致系统不可用的活动。

在上述情况下，应配置紧急 Break Glass 访问，以允许调查事件并及时给予补救。我们建议您使用 [具有适当权限的 IAM 用户](#)，来执行任务和访问 AWS 资源。请仅将根凭证用于 [需要根用户访问权限的任务](#)。要确认事件响应者对 AWS 和其他相关系统是否具有正确的访问权限级别，建议预置专用的用户账户。用户账户需要特许的访问权限，并且必须受到严格的控制和监视。在构建账户时，必须使用执行必要任务所需的最少权限，并且访问级别应基于作为事件管理计划的一部分创建的行动手册。

最好使用专门构建的专用用户和角色。通过添加 IAM 策略来临时提升用户或角色的访问权限，既会导致无法清楚地了解用户在事件期间拥有哪些访问权限，又会带来无法撤销提升的权限的风险。

请务必删除尽可能多的依赖项，以确保能在尽可能多的故障场景中获得访问权限。为了支持此操作，可创建一个行动手册，验证是否在专用的安全账户中创建事件响应用户作为 AWS Identity and Access Management 用户，而不是通过任何现有的联合身份验证或单点登录 (SSO) 解决方案管理他们。每个响应者都必须有自己的指定账户。账户配置必须实施 [强密码策略](#) 和多重身份验证 (MFA)。如果事件响应行动手册仅需要对 AWS Management Console 的访问权限，则用户不应配置访问密钥，并且应明确禁止用户创建访问密钥。可以使用 IAM 策略或服务控制策略 (SCP, Service Control Policy) 进行此配置，如 AWS 安全最佳实践 (适用于 [AWS Organizations SCP](#)) 中所述。用户仅能够在其他账户中代入事件响应角色，而不应具有其他任何权限。

在事件处理期间，可能需要向其他内部或外部人员授予访问权限，以支持调查、补救或恢复活动。在这种情况下，可以使用前面提到的行动手册机制，并且必须创建一个流程，确保在事件结束后立即撤销其他任何访问权限。

要确保能正确地监控和审计对事件响应角色的使用，至关重要的一点是，为此目的创建的 IAM 用户账户不会在人员之间共享，并且不会使用 AWS 账户根用户，除非 [特定任务要求这样做](#)。如果需要根用户 (例如，对特定账户的 IAM 访问权限不可用)，请使用单独的流程和可用的行动手册来验证根用户密码和 MFA 令牌的可用性。

要为事件响应角色配置 IAM 策略，请考虑使用 [IAM Access Analyzer](#) 来生成基于 AWS CloudTrail 日志的策略。为此，请在非生产账户中向事件响应角色授予管理员访问权限，并运行行动手册。完成后，会创建一个策略，仅允许已执行的操作。之后，可以跨所有账户将此策略应用于所有事件响应角色。您可能希望为每个行动手册创建一个单独的 IAM 策略，以便更轻松地进行管理和审计。示例行动手册可能包括针对勒索软件、数据泄露、丢失生产访问权限和其他场景的响应计划。

使用事件响应用户账户可在 [其他 AWS 账户中代入专用的事件响应 IAM 角色](#)。必须将这些角色配置为仅可由安全账户中的用户代入，并且信任关系必须要求调用主体已使用 MFA 进行身份验证。角色必须使用严格界定的 IAM 策略来控制访问。确保这些角色的所有 AssumeRole 请求都记录在 CloudTrail 中并发出提醒，并确保记录使用这些角色执行的任何操作。

强烈建议清楚地命名 IAM 用户账户和 IAM 角色，以便在 CloudTrail 日志中轻松找到他们。例如，将 IAM 账户命名为 `<USER_ID>-BREAK-GLASS#` 并将 IAM 角色命名为 `BREAK-GLASS-ROLE`。

[CloudTrail](#) 用于记录 AWS 账户中的 API 活动，并且应该用于 [配置关于使用事件响应角色的提醒](#)。请参阅博文，了解有关配置使用根密钥时的提醒。可以修改说明以配置 [Amazon CloudWatch](#) 指标筛选条件，从而筛选 AssumeRole 事件 (与事件响应 IAM 角色相关)：

```
{ $.eventName = "AssumeRole" && $.requestParameters.roleArn =  
  "<INCIDENT_RESPONSE_ROLE_ARN>" && $.userIdentity.invokedBy NOT EXISTS && $.eventType !=  
  "AwsServiceEvent" }
```

由于事件响应角色可能具有高级别的访问权限，因此，请务必将这些提醒转至广泛的群体，并及时采取适当的行动。

在事件处理期间，响应者可能需要访问不受 IAM 直接保护的系统。它们可能包括 Amazon Elastic Compute Cloud 实例、Amazon Relational Database Service 数据库或软件即服务 ( SaaS ) 平台。强烈建议不要使用 SSH 或 RDP 等本机协议，而是使用[AWS Systems Manager Session Manager](#) 对 Amazon EC2 实例进行所有管理访问。可以使用安全且经过审计的 IAM 控制此访问。此外，还可以使用 [AWS Systems Manager Run Command 文档自动实施行动手册的部分内容](#)，这可以减少用户出错的机会并缩短恢复时间。对于访问数据库和第三方工具，我们建议将访问凭证存储在 AWS Secrets Manager 中，并向事件响应者角色授予访问权限。

最后，事件响应 IAM 用户账户的管理应该添加到您的 [合并人员、移动人员和离开人员流程中](#)，并定期进行检查和测试，以确认只允许预期访问。

## 资源

### 相关文档：

- [管理对 AWS 环境的临时提升的访问权限](#)
- [AWS 安全事件响应指南](#)
- [AWS Elastic Disaster Recovery](#)
- [AWS Systems Manager Incident Manager](#)
- [为 IAM 用户设置账户密码策略](#)
- [在 AWS 中使用多重身份验证 \( MFA \)](#)
- [使用 MFA 配置跨账户存取](#)
- [使用 IAM Access Analyzer 生成 IAM 策略](#)
- [多账户环境中的 AWS Organizations 服务控制策略的最佳实践](#)
- [如何在使用 AWS 账户的根访问密钥时接收通知](#)
- [使用 IAM 托管策略创建精细会话权限](#)

### 相关视频：

- [在 AWS 中自动化事件响应和取证](#)
- [运行手册、事件报告和事件响应 DIY 指南](#)
- [准备和响应 AWS 环境中的安全事件](#)



相关示例：

- [实验室：AWS 账户设置和根用户](#)
- [实验：使用 AWS 控制台和 CLI 的事件响应](#)

## SEC10-BP06 预部署工具

确保安全人员预部署了适当的工具，来缩短从调查到恢复的时间。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

要自动执行安全响应和操作功能，您可以使用 AWS 提供的一整套 API 和工具。您可以完全自动执行身份管理、网络安全、数据保护和监控功能，并使用您已采用的常见软件开发方法交付这些功能。当构建安全自动化时，您的系统可以监控、审核和启动响应，您不必安排人员监控您的安全位置并对事件做出人为响应。

如果您的事件响应团队继续以同样的方式响应警报，警报可能会让他们应接不暇。久而久之，团队对警报的敏感性可能会下降，并可能在处理正常情况时犯错或者错过异常警报。利用一些功能自动处理重复和正常的警报，并将敏感、特殊的事件交由人员来处理，这样有助于避免疲于应对警报。集成异常检测系统（例如 Amazon GuardDuty、AWS CloudTrail Insights 和 Amazon CloudWatch Anomaly Detection）可以减轻常见阈值警报的负担。

您可以通过编程方式自动执行此流程中的步骤，从而改进手动流程。为事件定义修复模式之后，您可以将此模式分解为可执行的逻辑，并编写代码以执行此逻辑。然后，响应人员可以运行该代码来修复问题。久而久之，您就可以自动化越来越多的步骤，并最终自动处理各类常见事件。

在安全调查期间，您需要能够查看相关日志，以便记录并了解事件的来龙去脉和时间线。生成警报时也需要日志，因为日志可以指示某些相关操作已经发生。选择、启用、存储、设置查询和检索机制以及设置警报至关重要。此外，提供工具来搜索日志数据的有效方法是 [Amazon Detective](#)。

AWS 提供 200 多种云服务和数千种功能。我们建议您检查可支持和简化事件响应策略的服务。

除日志记录外，还应当制定并实施 [标记策略](#)。标记有助于提供有关 AWS 资源用途的背景信息。标记也可用于实现自动化。

### 实施步骤

选择并设置用于分析和报警的日志

请参阅以下关于配置事件响应日志记录的文档：



- [安全事件响应的日志记录策略](#)
- [SEC04-BP01 配置服务和应用程序日志记录](#)

## 启用安全服务来支持检测和响应

AWS 提供了本机检测、预防和响应功能，而其他服务可用于构建自定义安全解决方案。有关与安全事件响应最相关的服务列表，请参阅 [云功能定义](#)。

## 制定和实施标记策略

要获取围绕 AWS 资源的业务场景和相关内部利益相关者的背景信息，可能很困难。要做到这一点，可以采用标签的形式，标签为 AWS 资源分配元数据，并由用户定义的键和值组成。您可以创建标签，按照用途、所有者、环境、处理的数据类型以及您选择的其他标准对资源进行分类。

采用一致的标记策略可以加快响应速度，并通过快速识别和辨别 AWS 资源的背景信息，最大限度地减少在组织背景方面所花费的时间。标签还可以充当启动自动响应的机制。有关要标记的内容的详细信息，请参阅 [标记 AWS 资源](#)。首先，您需要定义要在组织内实施的标签。之后，实施并强制执行这一标记策略。有关实施和强制执行的详细信息，请参阅 [使用 AWS 标签策略和服务控制策略 \(SCP\) 实施 AWS 资源标记策略](#)。

## 资源

相关的 Well-Architected 最佳实践：

- [SEC04-BP01 配置服务和应用程序日志记录](#)
- [SEC04-BP02 集中分析日志、结果和指标](#)

相关文档：

- [安全事件响应的日志记录策略](#)
- [事件响应云功能定义](#)

相关示例：

- [使用 Amazon GuardDuty 和 Amazon Detective 进行威胁检测和响应](#)
- [Security Hub 研讨会](#)
- [使用 Amazon Inspector 进行漏洞管理](#)

## SEC10-BP07 运行模拟

随着组织不断发展壮大，威胁形势也会不断变化，因此务必要持续评估组织的事件响应能力。运行模拟（也称为实际演练）是可用于执行这种评估的一种方法。模拟过程使用现实世界中的安全事件场景，旨在模仿威胁主体采取的战术、技术和程序（TTP），让组织通过响应现实中可能发生的模拟网络事件，来练习和评估自己的事件响应能力。

建立这种最佳实践的好处：模拟有多种好处：

- 检验网络准备情况，有助于事件响应人员树立信心。
- 测试工具和工作流程的准确性和有效性。
- 完善沟通和上报环节，使之与您的事件响应计划相吻合。
- 提供机会来应对不太常见的攻击载体。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

模拟主要分为三种类型：

- **桌面演练**：桌面演练模拟方法是一种基于讨论的会议，让各个事件响应利益相关者参与进来，练习角色和职责，以及练习使用既定的沟通工具和行动手册。通常是用一整天的时间在虚拟场地和/或实地中协调完成演练。由于桌面演练以讨论为基础，因此侧重于流程、人员和协作。在讨论中，技术是必不可少的一部分，但事件响应工具或脚本的实际使用通常不包括在桌面演练中。
- **紫队演练**：紫队演练可提高事件响应人员（蓝队）和模拟威胁主体（红队）之间的协作能力。蓝队由安全运营中心（SOC）的成员组成，但也可以包括在实际网络事件中会参与进来的其他利益相关者。红队由渗透测试团队或接受过攻击安全培训的关键利益相关者组成。在设计场景时，红队会与演练协调员相互协作，以确保场景的准确性与可行性。在紫队演练中，主要的关注点是支持事件响应工作的检测机制、工具和标准操作程序（SOP）。
- **红队演练**：在红队演练中，进攻方（红队）模拟进行攻击，从而在预定范围内实现特定目标或一系列目标。防御方（蓝队）不一定知道演练的范围和持续时间，如此，可以更真实地评估他们应对真实事件的能力。由于红队的演练可能是侵入性测试，因此务必谨慎行事，并实施控制措施，以确保该演练不会对环境造成实际破坏。

请考虑定期协调开展网络模拟。对于参与者和整个组织而言，每种演练类型都可以带来独特的好处，因此您可以选择从不太复杂的模拟类型（例如桌面演练）入手，然后再慢慢过渡到较为复杂的模拟类型

(红队演练)。您应根据自身的安全成熟度、资源和预期结果选择模拟类型。由于红队演练的复杂性和成本，一些客户可能不会选择进行红队演练。

## 实施步骤

无论您选择哪种模拟类型，模拟通常都遵循以下实施步骤：

1. 确定核心演练要素：确定模拟场景和模拟要达成的目标。这两者都应该得到领导层的认同。
2. 确定关键利益相关者：演练至少需要演练协调员和参与者。根据具体的场景，可能会涉及其他利益相关者，例如法务、通信或行政等领域的领导层。
3. 构建和测试场景：如果有特定要素不可行，则可能需要在构建时重新定义该场景。本阶段的预期结果是最终确定的场景。
4. 协调开展模拟：采用的模拟类型决定了所需的协调工作（书面讨论场景对比高技术含量的模拟场景）。协调员应根据演练目标调整其协调战术，并应尽可能让所有演练参与者都参与进来，以实现最大利益。
5. 撰写事后报告（AAR）：确定哪些方面进展较为顺利、哪些方面需要改进以及可能存在的差距。AAR 应衡量模拟的有效性，并记录团队对模拟事件的响应情况，以便在将来的模拟中可以不断跟踪进度。

## 资源

相关文档：

- [AWS 事件响应指南](#)

相关视频：

- [AWS 实际演练 – 安全版](#)

## SEC10-BP08 建立从事件中吸取经验教训的框架

实现 经验教训总结 框架和根本原因分析能力不仅有助于提高事件响应能力，还有助于防止事件再次发生。通过从每次事件中吸取教训，您可以避免重复同样的错误、泄露或错误配置，这不仅可以改善您的安全态势，还可以最大限度地减少因可预防的情况而损失的时间。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

重要的是要实现一个 经验教训总结 框架，大体上确立并实现以下几点：

- 何时总结经验教训？
- 总结经验教训的过程涉及什么？
- 如何总结经验教训？
- 谁参与了这个过程，具体情况如何？
- 如何确定需要改进的领域？
- 如何确保有效跟踪和实施改进措施？

该框架不应关注或指责个人，而应侧重于改进工具和流程。

## 实施步骤

除了前面列出的大体上的成果外，重要的是要确保提出正确的问题，以便从流程中获得最大价值（可以带来切实可行的改进的信息）。请考虑以下问题，以便于您启动经验教训讨论：

- 发生了什么事件？
- 何时首次发现该事件？
- 是如何发现的？
- 哪些系统针对该活动发出了警报？
- 涉及哪些系统、服务和数据？
- 具体发生了什么？
- 哪些地方做得好？
- 哪些地方做得不好？
- 哪些流程或程序出现问题或未能扩展以应对事件？
- 以下方面有哪些地方有待改进：
  - 人员
    - 需要联系的人是否真的可以联系上，联系名单是否是最新名单？
    - 相应人员是否缺少有效应对和调查事件所需的培训或能力？
    - 相应的资源是否已就绪并随时可用？
  - 流程

- 是否遵循了流程和程序？
- 是否针对这种事件记录并提供了流程和程序？
- 是否缺少必要的流程和程序？
- 响应人员是否能够及时获得所需的信息来处理问题？
- 技术
  - 现有警报系统是否能有效识别活动并发出警报？
  - 我们如何将检测时间缩短 50%？
  - 现有警报是否需要改进，或者是否需要针对这种事件设置新的警报？
  - 现有工具是否允许对事件进行有效调查（搜索/分析）？
  - 怎样才能更快地识别这种事件？
  - 如何防止这种事件再次发生？
  - 谁是改进计划的负责人，如何检验改进计划的执行情况？
  - 实施和测试额外监控或预防性控制和流程的时间表是怎样的？

此列表并非详尽无遗，但旨在作为一个起点，确定组织和业务需求是什么，以及如何分析这些需求，以便最有效地从事件中吸取经验教训，并不断改进您的安全态势。最重要的是，该列表开始将经验教训作为事件响应流程、文档和利益相关者期望的标准组成部分。

## 资源

相关文档：

- [AWS 安全事件响应指南：建立从事件中吸取经验教训的框架](#)
- [NCSC CAF 指南：总结经验教训](#)

## 应用程序安全性

### 问题

- [SEC 11.如何在整个设计、开发和部署生命周期中纳入并验证应用程序的安全属性？](#)

### SEC 11.如何在整个设计、开发和部署生命周期中纳入并验证应用程序的安全属性？

开展员工培训、执行自动化测试、了解依赖项，并验证各种工具和应用程序的安全属性，有助于降低生产工作负载中出现安全问题的可能性。

## 最佳实践

- [SEC11-BP01 应用程序安全性培训](#)
- [SEC11-BP02 在整个开发和发布生命周期中执行自动化测试](#)
- [SEC11-BP03 定期执行渗透测试](#)
- [SEC11-BP04 人工代码审查](#)
- [SEC11-BP05 集中管理服务，方便获取软件包和依赖项](#)
- [SEC11-BP06 以编程方式部署软件](#)
- [SEC11-BP07 定期评测管道的安全属性](#)
- [SEC11-BP08 建立规程，让工作负载团队负责安全领域](#)

### SEC11-BP01 应用程序安全性培训

向贵组织的构建者提供培训，使其了解有关安全开发和操作应用程序的常见做法。在开发时采取安全至上的做法，有助于降低到安全审查阶段才会检测出问题的可能性。

期望结果：依照安全的理念设计和构建软件。如果组织中的构建者接受始于威胁建模的安全开发实践培训，就可以提高所开发软件的整体质量和安全性。因为经过安全审查阶段之后所需的返工较少，所以此方法可以减少交付软件或功能的时间。

就本最佳实践而言，安全开发是指正在编写的软件以及支持软件开发生命周期（SDLC）的工具或系统。

常见反模式：

- 等到安全审查阶段才考虑系统的安全属性。
- 将所有安全决策交给安全团队。
- 未能传达在 SDLC 中作出的决策如何与组织的总体安全期望或策略相关联。
- 太迟参与安全审查过程。

建立此最佳实践的好处：

- 在开发周期的早期更好地了解组织对安全性的要求。
- 能够更快地识别和修复潜在的安全问题，从而更快地交付功能。
- 提高软件和质量。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

为组织中的构建者提供培训。从[威胁建模](#)课程开始，为帮助进行安全培训打下了良好的基础。理想情况下，构建者应该能够自助访问与其工作负载相关的信息。这种访问有助于他们就所构建系统的安全属性作出明智的决策，而不需要询问另一个团队。应该明确地定义让安全团队参与审查的流程，且这个流程应该简单易行。安全培训中应包括审查流程中的步骤。如果提供了已知的实施模式或模板，应该很容易找到它们并与总体安全要求关联起来。考虑使用 [AWS CloudFormation](#)、[AWS Cloud Development Kit \(AWS CDK\) Constructs](#)、[Service Catalog](#) 或其他模板工具减少对自定义配置的需求。

## 实施步骤

- 让构建者从有关[威胁建模](#)的课程开始，打下很好的基础，并帮助培训他们如何思考安全性。
- 让他们可以访问 [AWS 培训与认证](#)、行业或 AWS 合作伙伴培训。
- 提供有关组织安全审查流程的培训，明确安全团队、工作负载团队和其他利益相关方之间的职责分工。
- 发布有关如何满足安全要求的自助服务指南，包括代码示例和模板（如可用）。
- 定期从构建者团队那里获得有关他们在安全审查流程和培训方面的体验反馈，并利用这些反馈来作出改进。
- 使用实际试用或错误大扫除活动来帮助减少问题数量，以及增强构建者的技能。

## 资源

相关最佳实践：

- [SEC11-BP08 建立规程，让工作负载团队负责安全领域](#)

相关文档：

- [AWS 培训和认证](#)
- [如何看待云安全治理](#)
- [如何处理威胁建模](#)
- [加速培训 – AWS Skills Guild](#)

相关视频：



- [主动式安全性：注意事项和方法](#)

相关示例：

- [有关威胁建模的研讨会](#)
- [开发人员的行业意识](#)

相关服务：

- [AWS CloudFormation](#)
- [AWS Cloud Development Kit \(AWS CDK\) \( AWS CDK \) Constructs](#)
- [Service Catalog](#)
- [AWS BugBust](#)

## SEC11-BP02 在整个开发和发布生命周期中执行自动化测试

在整个开发和发布生命周期中自动测试安全属性。自动化使得在发布软件前，更加容易始终如一反复识别软件中可能存在的问题，进而减少所提供的软件中存在安全风险的风险。

期望结果：自动测试的目标是提供一种程序化方式，在整个开发生命周期中尽早常常检测潜在问题。自动执行回归测试时，您可以重新运行功能测试和非功能测试，确认以前测试过的软件在更改后仍按预期执行。定义安全性单元测试以检查常见的错误配置（如身份验证中断或缺失）时，可以在开发过程的早期识别并修复这些问题。

测试自动化根据应用程序的要求和期望的功能，使用专门构建的测试用例进行应用程序验证。自动测试的结果基于将生成的测试输出与其各自的预期输出进行比较，从而加快整个测试生命周期。回归测试和单元测试套件等测试方法最适合自动化。自动执行安全属性的测试使构建者无需等待安全审查即可自动接收反馈。静态或动态代码分析形式的自动化测试可以提高代码质量，并帮助在开发生命周期的早期检测潜在的软件问题。

常见反模式：

- 不传达自动化测试的测试用例和测试结果。
- 就在发布之前执行自动化测试。
- 使用自动化测试用例来应对经常变化的需求。
- 未能就如何处理安全测试的结果提供指导。

建立此最佳实践的好处：

- 减少对评估系统安全属性的人员的依赖。
- 在多个工作流程中得到一致的结果可提高一致性。
- 降低在生产软件中引入安全问题的可能性。
- 由于及早发现软件问题，可以缩短检测和修复之间的时间。
- 增加多个工作流中的系统或重复行为的可见性，可用于促进组织范围内的改进。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

在构建软件时，采用各种软件测试机制，以确保根据应用程序的业务逻辑测试应用程序的功能要求和非功能要求（重点关注应用程序可靠性、性能和安全性）。

静态应用程序安全性测试（SAST）分析源代码是否存在异常安全模式，并指出容易出现缺陷的代码。SAST 依赖于文档（需求规范、设计文档和设计规范）和应用程序源代码等静态输入来测试一系列已知的安全问题。静态代码分析器可以帮助加快大量代码的分析。[NIST Quality Group](#) 对[源代码安全性分析器](#)进行了比较，包括针对[字节码扫描器](#)和[二进制码扫描器](#)的开源工具。

动态分析安全测试（DAST）方法针对正在运行的应用程序执行测试，以识别潜在的意外行为，能够对静态测试作出补充。动态测试可用于检测通过静态分析无法检测到的潜在问题。通过在代码存储库、构建和管道阶段进行测试，您可以检查出不同类型的潜在问题，防止这些问题进入到代码中。[Amazon CodeWhisperer](#) 在构建器的 IDE 中提供代码建议，包括安全扫描。[Amazon CodeGuru Reviewer](#) 可以识别应用程序开发过程中的关键问题、安全问题以及难以发现的错误，并提供可提高代码质量的建议。

通过[开发人员安全性研讨会](#)，学会使用 AWS 开发人员工具（例如，[AWS CodeBuild](#)、[AWS CodeCommit](#) 和 [AWS CodePipeline](#)）来自动执行发布管道，包括 SAST 和 DAST 测试方法。

在 SDLC 中，建立一个迭代过程，其中包括与安全团队一起定期审查应用程序。在发布准备情况审查过程中，应该处理和验证从这些安全审查中收集到的反馈。这些审查建立了健壮的应用程序安全态势，并为构建者提供切实可行的反馈，以解决潜在问题。

### 实施步骤

- 实施一致的 IDE、代码审查和 CI/CD 工具，其中包括安全测试。
- 考虑在 SDLC 中的哪个阶段适合阻塞管道，而不仅仅是通知构建者需要修复问题。
- [开发人员安全性研讨会](#)提供了将静态和动态测试集成到发布管道的示例。

- 使用自动化工具（例如，与开发人员 IDE 集成的 [Amazon CodeWhisperer](#)，以及用于在提交时扫描代码的 [Amazon CodeGuru Reviewer](#)）执行测试或代码分析，帮助构建者适时获得反馈。
- 使用 AWS Lambda 构建应用程序时，您可以使用 [Amazon Inspector](#) 来扫描函数中的应用程序代码。
- 利用 [AWS CI/CD 研讨会](#) 作为在 AWS 上构建 CI/CD 管道的起点。
- 当 CI/CD 管道中包括自动化测试时，您应该使用工单系统来跟踪软件问题的通知和修正。
- 对于可能生成结果的安全测试，链接到补救指南可帮助构建者提高代码质量。
- 定期分析使用自动化工具获得的结果，以确定下一个自动化、构建者培训或认知宣传活动的优先级。

## 资源

### 相关文档：

- [持续交付和持续部署](#)
- [AWS DevOps 能力合作伙伴](#)
- [适用于应用程序安全性的 AWS 安全能力合作伙伴](#)
- [选择 Well-Architected CI/CD 方法](#)
- [使用 Amazon EventBridge 和 Amazon CloudWatch Events 监控 CodeCommit 事件](#)
- [Amazon CodeGuru 审查中的密钥检测](#)
- [通过有效的治理加快 AWS 上的部署](#)
- [AWS 方法如何自动执行安全、不需要人工介入的部署](#)

### 相关视频：

- [不需要人工介入：在亚马逊自动实现持续交付管道](#)
- [自动执行跨账户 CI/CD 管道](#)

### 相关示例：

- [开发人员的行业意识](#)
- [AWS CodePipeline 治理](#) ( GitHub )
- [开发人员安全性研讨会](#)
- [AWS CI/CD 研讨会](#)

## SEC11-BP03 定期执行渗透测试

定期对软件执行渗透测试。此机制有助于识别无法通过自动化测试或人工代码审查检测到的潜在软件问题。它还有助于了解检测控制的有效性。渗透测试应设法确定软件是否会以意外方式执行，例如公开应受保护的数据，或者授予比预期更广泛的权限。

期望结果：使用渗透测试来检测、修复和验证应用程序的安全属性。在软件开发生命周期 ( SDLC ) 中应定期执行计划的渗透测试。在发布软件之前应处理渗透测试的结果。您应该分析渗透测试的结果，以确定是否存在使用自动化可以发现的问题。拥有包括主动反馈机制的定期且可重复渗透测试流程，有助于为构建者提供指导并提高软件质量。

常见反模式：

- 仅对已知或普遍存在的安全问题进行渗透测试。
- 未使用相关的第三方工具和库对应用程序执行渗透测试。
- 仅对软件包安全问题进行渗透测试，而不评估已实施的业务逻辑。

建立此最佳实践的好处：

- 在发布之前增强对软件安全属性的信心。
- 有机会确定首选的应用程序模式，从而提高软件质量。
- 获得一个反馈环路，在开发周期早期确定自动化或额外培训可以在哪些方面改进软件的安全属性。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

渗透测试是一项结构化安全测试练习，让您可以运行计划的安全漏洞方案，以便检测、修复和验证安全控制机制。渗透测试从侦察开始，在这个过程中，根据应用程序的当前设计及其依赖项收集数据。生成并运行特定于安全方面的测试方案的精选列表。这些测试的主要目的是发现应用程序中的安全问题，有人会利用这些安全问题来获得对环境的非预期访问，或未经授权访问数据。当推出新功能时，或者应用程序的功能或技术实施方面发生重大变更时，您应该执行渗透测试。

您应该确定在开发生命周期的哪个阶段执行渗透测试最为合适。应当尽量早些时候执行此测试，以便系统功能接近预期的发布状态，但也要留有足够的时间来修复任何问题。

### 实施步骤

- 采用结构化流程来确定渗透测试的范围，让这个流程基于[威胁模型](#)是保留场景相关性的好方法。

- 确定在开发周期的什么阶段执行渗透测试较为合适。这个阶段应该是在应用程序预期改动很细微，但仍留有足够时间进行修复的时候。
- 为构建者提供以下方面的培训：从渗透测试结果中可以期待获得什么，以及如何获得有关修复的信息。
- 使用工具自动执行常见或可重复的测试，从而加快渗透测试的速度。
- 分析渗透测试结果，以便确定系统性安全问题，并使用此数据为额外的自动化测试和正在进行的构建者培训提供信息。

## 资源

### 相关最佳实践：

- [SEC11-BP01 应用程序安全性培训](#)
- [SEC11-BP02 在整个开发和发布生命周期中执行自动化测试](#)

### 相关文档：

- [AWS 渗透测试](#)提供有关 AWS 上的渗透测试的详细指导
- [通过有效的治理加快 AWS 上的部署](#)
- [AWS 安全能力合作伙伴](#)
- [使 AWS Fargate 上的渗透测试架构实现现代化改造](#)
- [AWS Fault Injection Simulator](#)

### 相关示例：

- [使用 AWS CodePipeline 自动执行 API 测试](#) ( GitHub )
- [自动安全助手](#) ( GitHub )

## SEC11-BP04 人工代码审查

对您制作的软件执行人工代码审查。此流程有助于确保编写代码的人不是唯一检查代码质量的人。

期望结果：在开发过程中纳入人工代码审查步骤可以提高所编写的软件的质量，帮助团队中缺乏经验的成员提升自身技能，并且有助于确定哪些情况可以使用自动化。自动化工具和测试可以支持人工代码审查。

## 常见反模式：

- 在部署前不执行代码审查。
- 让同一个人编写和审查代码。
- 不使用自动化来协助或编排代码审查。
- 在审查代码之前没有对构建者进行应用程序安全方面的培训。

## 建立此最佳实践的好处：

- 提高代码质量。
- 通过重复利用通用方法提高代码开发的一致性。
- 减少在渗透测试和后续阶段发现的问题。
- 改进团队内部的知识传授。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

在整个代码管理流程中应实施审查步骤。具体细节取决于分支、拉取请求和合并请求所用的方法。您可以使用 AWS CodeCommit 或第三方解决方案，例如 GitHub、GitLab 或 Bitbucket。无论使用哪种方法，务必要确认在将代码部署到生产环境之前，您的流程需要进行代码审查。使用 [Amazon CodeGuru Reviewer](#) 等工具可以更轻松地编排代码审查过程。

## 实施步骤

- 在代码管理流程中实施人工审查步骤，并在继续之前执行此审查。
- 考虑使用 [Amazon CodeGuru Reviewer](#) 来管理和协助代码审查工作。
- 实施审批流程，要求在代码需要完成代码审查后方可进入下一阶段。
- 验证是否存在这样一个流程：识别在人工代码审查期间发现并可以自动检测到的问题。
- 根据您的代码开发实践集成人工代码审查步骤。

## 资源

### 相关最佳实践：

- [SEC11-BP02 在整个开发和发布生命周期中执行自动化测试](#)

## 相关文档：

- [在 AWS CodeCommit 存储库中使用拉取请求](#)
- [在 AWS CodeCommit 中使用审批规则模板](#)
- [关于 GitHub 中的拉取请求](#)
- [使用 Amazon CodeGuru Reviewer 自动审查代码](#)
- [使用 Amazon CodeGuru Reviewer CLI 自动检测 CI/CD 管道中的安全漏洞和错误](#)

## 相关视频：

- [使用 Amazon CodeGuru 持续改进代码质量](#)

## 相关示例：

- [开发人员安全性研讨会](#)

## SEC11-BP05 集中管理服务，方便获取软件包和依赖项

提供集中管理的服务，方便构建者团队获取软件包和其他依赖项。通过采取这种做法，可以在将软件包纳入所编写的软件之前，对软件包进行验证；另外，还可以为分析贵组织所使用的软件提供数据来源。

期望结果：除了正在编写的代码之外，软件还包含一组其他软件包。这样就可以轻松实施可重复使用的功能，例如 JSON 解析器或加密库。从逻辑上将这些软件包和依赖项的来源集中在一起，从而为安全团队提供了一种机制，可以在使用软件包之前对其属性进行验证。这种方法还降低了由于现有软件包中的更改或构建者团队（包括直接来自互联网的任意软件包）所引起的意外问题的风险。使用此方法与手动和自动测试流程相结合，增加对所开发软件质量的信心。

## 常见反模式：

- 从互联网上的任意存储库中提取软件包。
- 在将新软件包提供给构建者之前不进行测试。

## 建立此最佳实践的好处：

- 更好地了解正在构建的软件中使用了哪些软件包。
- 了解谁使用了哪些软件包后，在需要更新软件包时，能够向工作负载团队发出通知。



- 降低软件中存在问题软件包的风险。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

以构建者易于使用的方式为软件包和依赖项提供集中管理服务。集中管理服务可以在逻辑上集中，而不用作为一个整体系统来实施。利用此方法，您可以通过满足构建者需求的方法来提供服务。您应该实施一种有效的方法：在发生更新或出现新需求时将软件包添加到存储库。[AWS CodeArtifact](#) 等 AWS 服务或类似的 AWS 合作伙伴解决方案提供了一种实现此功能的方法。

### 实施步骤：

- 实施可在用于开发软件的所有环境中使用的逻辑集中式存储库服务。
- 在 AWS 账户 账户分配过程中包括对存储库的访问权限。
- 构建自动化以在存储库中发布软件包之前对其进行测试。
- 维护最常用软件包、语言和更改量最大的团队的指标。
- 为构建者团队提供一种自动化机制来请求新软件包和提供反馈。
- 定期扫描存储库中的软件包，以确定新发现的问题的潜在影响。

### 资源

#### 相关最佳实践：

- [SEC11-BP02 在整个开发和发布生命周期中执行自动化测试](#)

#### 相关文档：

- [通过有效的治理加快 AWS 上的部署](#)
- [使用 CodeArtifact Package Origin Control 工具包加强软件包的安全性](#)
- [使用 Amazon CodeGuru Reviewer 检测日志记录中的安全问题](#)
- [软件构件的供应链级别 \( SLSA \)](#)

#### 相关视频：

- [主动式安全性：注意事项和方法](#)

- [AWS 安全理念 \( re:Invent 2017 \)](#)
- [当安全、保障和紧迫性都很重要时：处理 Log4Shell](#)

相关示例：

- [多区域软件包发布管道 \( GitHub \)](#)
- [使用 AWS CodePipeline 在 AWS CodeArtifact 上发布 Node.js 模块 \( GitHub \)](#)
- [AWS CDK Java CodeArtifact 管道示例 \( GitHub \)](#)
- [使用 AWS CodeArtifact 分发专用 .NET NuGet 包 \( GitHub \)](#)

## SEC11-BP06 以编程方式部署软件

尽可能以编程方式部署软件。通过采取这种做法，可以降低由于人为错误导致部署失败或引入意外问题的可能性。

期望结果：让人们远离数据是在 AWS Cloud 中安全构建的一项关键原则。此原则包括如何部署软件。

不依赖人来部署软件的好处是，您可以更加确信，您测试的内容就是部署的内容，并且确信每次都一致地执行部署。无需更改软件即可在不同的环境中运行。使用十二要素应用程序开发原则，特别是配置的外部化，使您无需更改即可将相同的代码部署到多个环境。对软件包进行加密签名可以很好地确认不同环境之间什么也没有改变。这种方法的总体结果是降低更改过程中的风险以及提升软件版本的一致性。

常见反模式：

- 手动将软件部署到生产环境中。
- 手动对软件进行更改，以适应不同的环境。

建立此最佳实践的好处：

- 增强对软件发布过程的信心。
- 降低了失败的更改对业务功能造成影响的风险。
- 由于更改风险降低，从而加快了发布节奏。
- 针对部署过程中的意外事件的自动回滚功能。
- 能够以加密方式证明所测试的软件是部署的软件。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

构建 AWS 账户 结构时减少持续的人类访问环境的情况，并使用 CI/CD 工具来进行部署。适当地设计应用程序，以便从外部源（例如，[AWS Systems Manager Parameter Store](#)）获得特定于环境的配置数据。在测试软件包后对其进行签名，并在部署期间验证这些签名。配置 CI/CD 管道以推送应用程序代码，并使用金丝雀来确认已成功部署。使用 [AWS CloudFormation](#) 或 [AWS CDK](#) 等工具来定义基础设施，然后使用 [AWS CodeBuild](#) 和 [AWS CodePipeline](#) 来执行 CI/CD 操作。

## 实施步骤

- 构建明确定义的 CI/CD 管道，以便简化部署过程。
- 使用 [AWS CodeBuild](#) 和 [AWS Code Pipeline](#) 来提供 CI/CD 功能，从而更容易将安全测试集成到管道中。
- 遵循[使用多个账户组织 AWS 环境](#)白皮书中有关环境分离的指导。
- 确认在运行生产工作负载的环境中沒有持续的人类访问。
- 设计应用程序以支持配置数据外部化。
- 考虑使用蓝绿部署模式进行部署。
- 实施金丝雀以验证软件是否成功部署。
- 使用 [AWS Signer](#) 或 [AWS Key Management Service \( AWS KMS \)](#) 等加密工具为部署的软件包签名和进行验证。

## 资源

### 相关最佳实践：

- [SEC11-BP02 在整个开发和发布生命周期中执行自动化测试](#)

### 相关文档：

- [AWS CI/CD 研讨会](#)
- [通过有效的治理加快 AWS 上的部署](#)
- [自动执行安全、不需要人工介入的部署](#)
- [使用 AWS Certificate Manager Private CA 和 AWS Key Management Service 非对称密钥进行代码签名](#)
- [代码签名，用于 AWS Lambda 的可信度和完整性控制措施](#)

相关视频：

- [不需要人工介入：在亚马逊自动实现持续交付管道](#)

相关示例：

- [使用 AWS Fargate 进行蓝绿部署](#)

## SEC11-BP07 定期评测管道的安全属性

对您的管道运用 Well-Architected 安全性支柱原则，尤其注意权限分离。定期评测管道基础设施的安全属性。通过有效管理管道的安全性，可以确保通过管道的软件的安全性。

期望结果：用于构建和部署软件的管道应遵循与环境中任何其他工作负载相同的推荐做法。正在使用测试的构建者应该不能编辑在管道中实施的测试。管道应该只拥有它们正在执行的部署所需的权限，并应实施保护措施以避免部署到错误的环境。管道不应该依赖长期凭证，且应配置为发出状态，以便可以验证构建环境的完整性。

常见反模式：

- 构建者可以绕过安全测试。
- 用于部署管道的权限过于宽松。
- 未将管道配置为验证输入。
- 不定期审查与 CI/CD 基础设施关联的权限。
- 使用长期或硬编码凭证。

建立此最佳实践的好处：

- 对通过管道构建和部署的软件的安全性有了更大的信心。
- 在出现可疑活动时可以停止部署。

在未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

从支持 IAM 角色的托管 CI/CD 服务开始，可以降低凭证泄露的风险。将安全性支柱原则应用到 CI/CD 管道基础设施有助于确定可以在哪些方面作出安全改进。遵循 [AWS 部署管道参考架构](#) 是构建 CI/CD

环境的一个很好的起点。定期审查管道实施和分析日志以了解意外行为，这样有助于您了解用于部署软件的管道的使用模式。

## 实施步骤

- 从 [AWS 部署管道参考架构](#) 开始。
- 考虑使用 [AWS IAM Access Analyzer](#) 以编程方式生成管道的最低权限 IAM 策略。
- 将管道与监控和警报集成在一起，以便在发生意外或异常活动时您会得到通知，对于 AWS 托管服务，[Amazon EventBridge](#) 允许您将数据路由到目标，例如 [AWS Lambda](#) 或 [Amazon Simple Notification Service](#) ( Amazon SNS ) 。

## 资源

相关文档：

- [AWS 部署管道参考架构](#)
- [监控 AWS CodePipeline](#)
- [AWS CodePipeline 的安全最佳实践](#)

相关示例：

- [DevOps 监控控制面板](#) ( GitHub )

## SEC11-BP08 建立规程，让工作负载团队负责安全领域

建立规程或机制，使构建者团队能够针对创建的软件作出安全决策。这些决策仍然需要由安全团队通过审查加以验证，但让构建者团队负责安全领域可以构建速度更快、安全性更高的工作负载。此机制还可促进负责任文化，进而对所构建系统的运营产生积极影响。

期望结果：为了让构建者团队负责安全性和决策制定，您可以向构建者团队提供安全思维方式方面的培训，或者让安全人员加入构建者团队或与构建者团队联系在一起，从而增强他们的培训。这两种方法都有效，并且可以让团队在开发周期的早期作出更高质量的安全决策。这种负责任模式基于应用程序安全性培训。从特定工作负载的威胁建模开始，有助于将设计思维集中在合适的场景中。拥有一个专注于安全的构建者社区，或者拥有一组与构建者团队合作的安全工程师带来的另一项好处是，您可以更深入地了解如何编写软件。这种了解帮助您确定自动化能力的下一个改进领域。

常见反模式：

- 将所有安全设计决策交给安全团队。
- 在开发过程中没有及早满足安全要求。
- 没有从构建者和安全人员那里获得关于计划运营的反馈。

建立此最佳实践的好处：

- 缩短完成安全审查的时间。
- 减少等到安全审查阶段才检测到安全问题的情况。
- 提高所编写软件的整体质量。
- 有机会识别和了解系统性问题或高价值改进领域。
- 进行安全审查后，发现的问题可以在早期进行修复，从而减少所需的返工量。
- 提升对安全功能的认知。

在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

从[SEC11-BP01 应用程序安全性培训](#)中的指导开始。然后确定您认为可能最适合您组织的计划的运营模式。两个主要模式是对构建者进行培训，或在构建者团队中加入安全人员。确定初始方法后，应使用单个或一小组工作负载团队进行试点，以证明该模式适用于您的组织。来自组织的构建者和安全团队的领导层支持有助于计划的成功交付。在构建此计划时，重要的是选择可以用来显示项目价值的指标。了解 AWS 如何解决这个问题是一个很好的学习经验。这个最佳实践非常注重组织变革和文化。您使用的工具应支持构建者和安全社区之间的协作。

## 实施步骤

- 首先对构建者进行应用程序安全性培训。
- 创建一个社区和入门培训计划来对构建者进行培训。
- 为计划选择一个名称。通常使用守护者、拥护者或倡导者。
- 确定要使用的模式：培训构建者、加入安全工程师或具有相关性安全角色。
- 从安全性、构建者和可能的其他相关团体中确定项目发起人。
- 跟踪参与计划的人数、审查所花时间以及来自构建者和安全人员的反馈等指标。使用这些指标来作出改进。

## 资源

相关最佳实践：

- [SEC11-BP01 应用程序安全性培训](#)
- [SEC11-BP02 在整个开发和发布生命周期中执行自动化测试](#)

相关文档：

- [如何处理威胁建模](#)
- [如何看待云安全治理](#)

相关视频：

- [主动式安全性：注意事项和方法](#)

## 可靠性

可靠性支柱涵盖相关工作负载按照计划正确而稳定执行其预期功能的能力。如需有关具体实施的说明性指导，请参阅 [《可靠性支柱》白皮书](#) 访问 AWS 资源。

最佳实践领域

- [基础](#)
- [工作负载架构](#)
- [变更管理](#)
- [故障管理](#)

## 基础

问题

- [REL 1. 如何管理服务限额和限制？](#)
- [REL 2. 如何规划网络拓扑？](#)



## REL 1. 如何管理服务限额和限制？

基于云的工作负载架构存在服务限额（也称作服务限制）。存在这些限额是为了防止意外预置超过您所需的资源，并对 API 操作的请求速率进行限制，以保护服务不会遭到滥用。还存在资源限制，例如，将比特推入光缆的速率，或物理磁盘上的存储量。

### 最佳实践

- [REL01-BP01 了解服务限额和约束](#)
- [REL01-BP02 跨多个账户和区域管理服务限额](#)
- [REL01-BP03 通过架构适应固定服务限额和限制](#)
- [REL01-BP04 监控和管理限额](#)
- [REL01-BP05 自动管理限额](#)
- [REL01-BP06 确保在当前限额与最大使用量之间存在足够的差距，以便应对失效转移](#)

### REL01-BP01 了解服务限额和约束

了解您的工作负载架构的默认限额并管理限额提高请求。了解哪些云资源约束（如磁盘或网络）可能会对您产生影响。

期望结果：客户可以通过实施正确的准则，来监视关键指标、基础设施审查和自动化补救步骤，以确认没有达到服务限额和约束（这可能导致服务降级或中断），从而防止其 AWS 账户中的服务降级或中断。

### 常见反模式：

- 在不了解所用服务的硬限额或软限额及其限制的情况下部署工作负载。
- 在未分析和重新配置必要限额或未事先联系支持部门的情况下，部署替代工作负载。
- 假设云服务没有限制，并且认为可以在不考虑费率、限制、计数、数量的情况下使用服务。
- 假设限额会自动增加。
- 不了解限额请求的流程和时间表。
- 假设每个服务的默认云服务限额在不同区域都是相同的。
- 假设可以突破服务约束，并且系统会自动扩展或提高限制以超出资源约束。
- 没有在流量高峰期测试应用程序，以便对资源的利用率进行压力测试。
- 在没有分析所需资源规模的情况下配置资源。

- 通过选择远远超出实际需求或预期峰值的资源类型来过量配置容量。
- 在新的客户事件或部署新技术之前，不评估新流量水平的容量需求。

建立此最佳实践的好处：对服务限额和资源约束的监视和自动化管理可以主动减少故障。如果不遵循最佳实践，客户服务的流量模式变化可能会导致中断或降级。通过监视和管理所有区域和所有账户的这些值，应用程序可以在出现不利或意外事件时具有更好的复原能力。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

Service Quotas 是一项 AWS 服务，可帮助您在同一位置管理 250 多项 AWS 服务的限额。除了查找限额值，您还可以在 Service Quotas 控制台或使用 AWS SDK 请求增加限额并跟踪。AWS Trusted Advisor 提供服务限额检查，显示您的服务使用情况，以及某些服务在某些方面的限额。有关每个服务的默认服务限额，请查看相应服务的 AWS 文档，例如，请参阅 [Amazon VPC 限额](#)。

通过配置使用计划，可在 Amazon API Gateway 内设置某些服务限制，例如限流 API 的速率限制。可通过配置对应的服务进行设置的一些限制包括预置 IOPS、已分配的 Amazon RDS 存储，以及 Amazon EBS 卷分配等。Amazon Elastic Compute Cloud 有自身的服务限制控制面板，可帮助您管理您的实例、Amazon Elastic Block Store 和弹性 IP 地址限制。如果在某用例中，服务限额会对您应用程序的性能造成影响，而且无法按您的需求进行调整，请联系 支持 了解是否有解决的办法。

服务限额可以是区域特定的，也可以是全局性的。使用达到其限额的 AWS 服务将不会具有正常使用中预期的行为，并且可能会导致服务中断或降级。例如，服务限额可限制一个区域中使用的 DL Amazon EC2 数量，在使用 Auto Scaling 组 ( ASG ) 的流量扩展事件中，可能会达到该限制。

应定期评估每个账户的服务限额的使用情况，以确定适合该账户的适当服务限制。这些服务限额是作为操作防护机制存在的，以防止意外地配置超出所需数量的资源。它们还用于限制 API 操作的请求率，以保护服务不被滥用。

服务约束与服务限额不同。服务约束代表由特定资源类型定义的该资源的限制，可能是存储容量（例如，gp2 的大小限制为 1GB - 16TB）或磁盘吞吐量（10,000 iops）。必须对资源类型的约束进行设计，并不断评估可能达到其限制的使用量。如果意外地达到约束条件，账户的应用程序或服务可能会降级或中断。

如果在某用例中，服务限额对应用程序的性能造成影响，而且无法根据所要求进行调整，请联系 支持 了解是否有解决办法。有关调整固定限额的更多详情，请参阅 [REL01-BP03 通过架构适应固定服务限额和限制](#)。

有一些 AWS 服务和工具可以帮助监视和管理 Service Quotas。应该利用这些服务和工具来自动或手动检查限额水平。

- AWS Trusted Advisor 提供服务限额检查，显示您的服务使用情况，以及某些服务在某些方面的限额。它可以帮助识别接近限额的服务。
- AWS Management Console 提供了显示服务限额值，管理、请求新限额，监视限额请求状态以及显示限额历史记录的方法。
- AWS CLI 和 CDK 提供了通过编程方式自动管理和监视服务限额水平和使用情况的方法。

## 实施步骤

对于 Service Quotas：

- [审核 AWS Service Quotas。](#)
- 为了了解您现有的服务限额，请确定使用的服务（如 IAM Access Analyzer）。大约有 250 个 AWS 服务由服务限额控制。然后，确定每个账户和区域内可能使用的具体服务限额名称。每个区域大约有 3000 个服务限额名称。
- 使用 AWS Config 来增强这种限额分析，以查找您的 AWS 账户中使用的所有 [AWS 资源](#)。
- 使用 [AWS CloudFormation 数据](#) 来确定使用的 AWS 资源。查看 AWS Management Console 中创建的资源或通过 [list-stack-resources](#) AWS CLI 命令创建的资源。您还可以查看配置为要在模板自身部署的资源。
- 通过查看部署代码来确定工作负载所需的所有服务。
- 确定适用的服务限额。通过 Trusted Advisor 和 Service Quotas 使用能够以编程方式访问的信息。
- 建立一个自动监视方法（请参阅[REL01-BP02 跨多个账户和区域管理服务限额](#)和[REL01-BP04 监控和管理限额](#)），以便在服务限额接近或已达到其限制时发出提醒和通知。
- 建立一个自动化编程方法，以检查服务限额是否在一个区域已更改，但在同一账户的其他区域没有更改（请参阅[REL01-BP02 跨多个账户和区域管理服务限额](#)和[REL01-BP04 监控和管理限额](#)）。
- 自动扫描应用程序日志和指标，以确定是否存在任何限额或服务约束错误。如果存在这些错误，则向监视系统发送警报。
- 一旦确定特定服务需要更高限额，则制定工程设计步骤来计算所需的限额变化（请参阅[REL01-BP05 自动管理限额](#)）。
- 创建一个配置和批准工作流，以请求更改服务限额。这应该包括在请求被拒绝或部分批准情况下的例外工作流。
- 创建一个工程设计方法，在配置和使用新的 AWS 服务之前，以及在推出到生产环境或加载的环境之前（例如，负载测试账户），审查服务限额。

对于服务约束：

- 建立监视和度量方法，以提醒资源接近其资源约束。适当地利用 CloudWatch 进行指标或日志监视。
- 为每个具有对应用程序或系统有意义的约束的资源建立警报阈值。
- 创建工作流和基础设施管理过程，以在约束条件接近利用率时更改资源类型。该工作流应包括负载测试作为最佳实践，以验证新类型是新约束条件下的正确资源类型。
- 使用现有的过程和流程，将确定的资源迁移到推荐的新资源类型。

资源

相关最佳实践：

- [REL01-BP02 跨多个账户和区域管理服务限额](#)
- [REL01-BP03 通过架构适应固定服务限额和限制](#)
- [REL01-BP04 监控和管理限额](#)
- [REL01-BP05 自动管理限额](#)
- [REL01-BP06 确保在当前限额与最大使用量之间存在足够的差距，以便应对失效转移](#)
- [REL03-BP01 选择如何划分工作负载](#)
- [REL10-BP01 将工作负载部署到多个位置](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)
- [REL11-BP03 自动修复所有层](#)
- [REL12-BP05 使用混沌工程测试弹性](#)

相关文档：

- [AWS Well-Architected Framework 的可靠性支柱：可用性](#)
- [AWS Service Quotas \( 以前称为服务限制 \)](#)
- [AWS Trusted Advisor 最佳实践检查 \( 请参阅“服务限制”部分 \)](#)
- [AWS Answers 上的 AWS Limit Monitor](#)
- [Amazon EC2 服务限制](#)
- [什么是 Service Quotas ?](#)
- [如何请求增加限额](#)

- [服务终端节点和限额](#)
- [Service Quotas 用户指南](#)
- [AWS 的限额监控](#)
- [AWS 故障隔离界限](#)
- [通过冗余实现可用性](#)
- [AWS 数据解决方案](#)
- [什么是持续集成？](#)
- [什么是持续交付？](#)
- [APN 合作伙伴：可帮助进行配置管理的合作伙伴](#)
- [在 AWS 上的每个租户一个账户的 SaaS 环境中管理账户生命周期](#)
- [管理和监控工作负载中的 API 节流](#)
- [使用 AWS Organizations 大规模查看 AWS Trusted Advisor 建议](#)
- [使用 AWS Control Tower 自动提升服务限制并实现企业支持](#)

相关视频：

- [AWS Live re:Inforce 2019 - Service Quotas](#)
- [使用 Service Quotas 查看和管理 AWS 服务的限额](#)
- [AWS IAM 限额演示](#)

相关工具：

- [Amazon CodeGuru Reviewer](#)
- [AWS CodeDeploy](#)
- [AWS CloudTrail](#)
- [Amazon CloudWatch](#)
- [Amazon EventBridge](#)
- [Amazon DevOps Guru](#)
- [AWS Config](#)
- [AWS Trusted Advisor](#)
- [AWS CDK](#)

- [AWS Systems Manager](#)
- [AWS Marketplace](#)

## REL01-BP02 跨多个账户和区域管理服务限额

如果您目前使用多个账户或区域，请确保在运行生产工作负载的所有环境中都请求适当的限额。

期望结果：对于跨账户或区域的配置，或具有使用扩展区、区域或账户失效转移的弹性设计的配置，服务和应用程序不应受到服务限额耗尽的影响。

常见反模式：

- 允许一个隔离区域内的资源利用率增加，但没有相关机制保持其他隔离区域中的容量。
- 手动单独设置隔离区域中的所有限额。
- 没有考虑到在非主要区域出现降级期间，弹性架构（如主动或被动）对未来限额需求的影响。
- 不定期评估限额，并且不在工作负载运行的每个区域和账户中进行必要更改。
- 不利用[限额请求模板](#)来请求提高多个区域和账户的限额。
- 不更新服务限额，因为错误地认为提高限额会像计算预留请求一样产生影响成本。

建立此最佳实践的好处：验证在区域服务不可用时，您能否在辅助区域或账户中处理您当前的负载。这可以帮助降低在区域丢失期间发生的错误数量或降级水平。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

每个账户的服务限额都可被跟踪。除非另有说明，否则每个限额都针对的是特定的 AWS 区域。除生产环境以外，还要管理所有适用的非生产环境中的限额，以避免妨碍测试与开发。保持高度复原能力需要持续评估服务限额（无论是自动还是手动）。

由于实施的设计采用主动/主动、主动/被动 – 常用、主动/被动 – 非常用和主动/被动 – 指示灯方法，会有更多工作负载跨越区域，因此了解所有区域和账户限额水平至关重要。如果服务限额设置正确，过去的流量模式并不总是一个好的指标。

同样重要的是，服务限额名称限制并非对每个区域都始终相同。在一个区域，该值可能是 5，而在另一个区域，该值可能是 10。必须跨越所有相同的服务、账户和区域来管理这些限额，以便在负载状态下提供一致的复原能力。

协调不同区域（主动区域或被动区域）之间的所有服务限额差异，并创建流程来持续协调这些差异。被动区域失效转移的测试计划很少扩展到能够满足峰值主动容量，这意味着实际试用或桌面演练可能无法发现区域之间的服务限额差异，因此也无法保持正确的限制。

服务限额偏移对于跟踪和评估非常重要，它是指特定命名限额在一个区域而非所有区域发生更改的情况。应考虑更改在有流量或可能有流量的区域的限额。

- 根据您的服务要求、延迟、法规和灾难恢复（DR）要求选择相关账户和区域。
- 确定跨所有相关账户、区域和可用区的服务限额。限制的范围具体到账户和区域。应比较这些值以了解差异。

## 实施步骤

- 审查可能已经超出风险使用水平的 Service Quotas 值。AWS Trusted Advisor 会在超出 80% 和 90% 阈值时提供警报。
- 审查任何被动区域（在主动/被动设计中）的服务限额值。验证在主区域发生故障时，负载能否在辅助区域成功运行。
- 自动评估同一账户的不同区域之间是否发生了任何服务限额偏移，并采取相应的行动来更改限制。
- 如果客户的组织单位（OU）以受支持的方式构建，则应更新服务限额模板，以反映任何限额的变化，这些变化应该应用于多个区域和账户。
  - 创建模板并将区域与限额变化关联起来。
  - 审查所有现有的服务限额模板，看看是否需要任何更改（区域、限制和账户）。

## 资源

相关最佳实践：

- [REL01-BP01 了解服务限额和约束](#)
- [REL01-BP03 通过架构适应固定服务限额和限制](#)
- [REL01-BP04 监控和管理限额](#)
- [REL01-BP05 自动管理限额](#)
- [REL01-BP06 确保在当前限额与最大使用量之间存在足够的差距，以便应对失效转移](#)
- [REL03-BP01 选择如何划分工作负载](#)
- [REL10-BP01 将工作负载部署到多个位置](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)



- [REL11-BP03 自动修复所有层](#)
- [REL12-BP05 使用混沌工程测试弹性](#)

#### 相关文档：

- [AWS Well-Architected Framework 的可靠性支柱：可用性](#)
- [AWS Service Quotas \( 以前称为服务限制 \)](#)
- [AWS Trusted Advisor 最佳实践检查 \( 请参阅“服务限制”部分 \)](#)
- [AWS Answers 上的 AWS Limit Monitor](#)
- [Amazon EC2 服务限制](#)
- [什么是 Service Quotas ?](#)
- [如何请求增加限额](#)
- [服务终端节点和限额](#)
- [Service Quotas 用户指南](#)
- [AWS 的限额监控](#)
- [AWS 故障隔离界限](#)
- [通过冗余实现可用性](#)
- [AWS 数据解决方案](#)
- [什么是持续集成？](#)
- [什么是持续交付？](#)
- [APN 合作伙伴：可帮助进行配置管理的合作伙伴](#)
- [在 AWS 上的每个租户一个账户的 SaaS 环境中管理账户生命周期](#)
- [管理和监控工作负载中的 API 节流](#)
- [使用 AWS Organizations 大规模查看 AWS Trusted Advisor 建议](#)
- [使用 AWS Control Tower 自动提升服务限制并实现企业支持](#)

#### 相关视频：

- [AWS Live re:Inforce 2019 - Service Quotas](#)
- [使用 Service Quotas 查看和管理 AWS 服务的限额](#)
- [AWS IAM 限额演示](#)

## 相关服务：

- [Amazon CodeGuru Reviewer](#)
- [AWS CodeDeploy](#)
- [AWS CloudTrail](#)
- [Amazon CloudWatch](#)
- [Amazon EventBridge](#)
- [Amazon DevOps Guru](#)
- [AWS Config](#)
- [AWS Trusted Advisor](#)
- [AWS CDK](#)
- [AWS Systems Manager](#)
- [AWS Marketplace](#)

## REL01-BP03 通过架构适应固定服务限额和限制

了解不可更改的服务限额、服务限制和物理资源限制。为应用程序和服务设计架构，以防止这些限制影响可靠性。

其中的示例包括网络带宽、无服务器功能调用有效负载大小、API Gateway 的节流突发速率，以及并发用户连接至数据库。

期望结果：应用程序或服务在正常条件下和高流量条件下按预期执行。它们设计为在该资源的固定约束或服务限额的限制范围内工作。

## 常见反模式：

- 选择使用一项服务资源的设计时，没有意识到设计存在限制，这些限制将导致扩展时设计失败。
- 执行不现实的基准测试，并且在测试期间将达到服务固定限额。例如，以突发限制运行测试，但运行时间较长。
- 选择在超过固定服务限额时无法扩展或修改的设计。例如，SQS 有效负载大小为 256KB。
- 没有设计和实施可观测性，以便监控在高流量事件期间可能面临风险的服务限额阈值并发出警报。

建立此最佳实践的好处：确认应用程序将在所有预计的服务负载水平下运行，而不会中断或降级。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

与软服务限额或替换为更高容量单位的资源不同，无法更改 AWS 服务的固定限额。这意味着，在应用程序设计中使用所有这些类型的 AWS 服务时，必须评估是否存在潜在的硬容量限制。

Service Quotas 控制台中显示了硬限制。如果列中显示 ADJUSTABLE = No，则服务具有硬限制。一些资源配置页中也显示了硬限制。例如，Lambda 具有无法调整的特定硬限制。

例如，在设计 Python 应用程序以在 Lambda 函数中运行时，应评估应用程序，以确定 Lambda 是否有可能运行超过 15 分钟。如果代码可能运行超过此服务限额限制，则必须考虑替代技术或设计。如果在生产部署后达到此限制，则应用程序将遭受降级和中断，直到可以补救为止。与软限额不同，即使在紧急的严重性 1 事件下，也无法更改这些限制。

应用程序部署到测试环境之后，应使用策略来查明是否会达到任何硬限制。引入测试计划中应包括压力测试、负载测试和混沌测试。

## 实施步骤

- 查看可在应用程序设计阶段使用的 AWS 服务的完整列表。
- 查看所有这些服务的软限额限制和硬限额限制。Service Quotas 控制台中并没有显示所有限制。有些服务[描述了备用位置的这些限制](#)。
- 在设计应用程序时，检查工作负载的业务和技术驱动因素，例如业务成果、使用案例、相依系统、可用性目标和灾难恢复对象。让业务和技术驱动因素指导流程，以确定适合工作负载的分布式系统。
- 分析各个区域和账户的服务负载。服务的许多硬限制基于区域。但有些限制基于账户。
- 分析弹性架构在可用区故障和区域故障期间的资源使用情况。在使用“主动/主动”、“主动/被动 - 热”、“主动/被动 - 冷”和“主动/被动 - 指示灯”方法的多区域设计过程中，这些故障情况会导致使用率更高。这会形成达到硬限制的潜在使用案例。

## 资源

相关最佳实践：

- [REL01-BP01 了解服务限额和约束](#)
- [REL01-BP02 跨多个账户和区域管理服务限额](#)
- [REL01-BP04 监控和管理限额](#)
- [REL01-BP05 自动管理限额](#)

- [REL01-BP06 确保在当前限额与最大使用量之间存在足够的差距，以便应对失效转移](#)
- [REL03-BP01 选择如何划分工作负载](#)
- [REL10-BP01 将工作负载部署到多个位置](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)
- [REL11-BP03 自动修复所有层](#)
- [REL12-BP05 使用混沌工程测试弹性](#)

相关文档：

- [AWS Well-Architected Framework 的可靠性支柱：可用性](#)
- [AWS Service Quotas \( 以前称为服务限制 \)](#)
- [AWS Trusted Advisor 最佳实践检查 \( 请参阅“服务限制”部分 \)](#)
- [AWS Answers 上的 AWS Limit Monitor](#)
- [Amazon EC2 服务限制](#)
- [什么是 Service Quotas ?](#)
- [如何请求增加限额](#)
- [服务终端节点和限额](#)
- [Service Quotas 用户指南](#)
- [AWS 的限额监控](#)
- [AWS 故障隔离界限](#)
- [通过冗余实现可用性](#)
- [AWS 数据解决方案](#)
- [什么是持续集成？](#)
- [什么是持续交付？](#)
- [APN 合作伙伴：可帮助进行配置管理的合作伙伴](#)
- [在 AWS 上的每个租户一个账户的 SaaS 环境中管理账户生命周期](#)
- [管理和监控工作负载中的 API 节流](#)
- [使用 AWS Organizations 大规模查看 AWS Trusted Advisor 建议](#)
- [使用 AWS Control Tower 自动提升服务限制并实现企业支持](#)
- [Service Quotas 的操作、资源和条件键](#)

## 相关视频：

- [AWS Live re:Inforce 2019 - Service Quotas](#)
- [使用 Service Quotas 查看和管理 AWS 服务的限额](#)
- [AWS IAM 限额演示](#)
- [AWS re:Invent 2018：闭环系统和开放思维：如何掌控不同规模的系统](#)

## 相关工具：

- [AWS CodeDeploy](#)
- [AWS CloudTrail](#)
- [Amazon CloudWatch](#)
- [Amazon EventBridge](#)
- [Amazon DevOps Guru](#)
- [AWS Config](#)
- [AWS Trusted Advisor](#)
- [AWS CDK](#)
- [AWS Systems Manager](#)
- [AWS Marketplace](#)

## REL01-BP04 监控和管理限额

评估您的可能使用情况，并适当提高您的限额，支持使用量按计划增长。

期望结果：部署了可进行管理和监控的主动和自动化系统。这些操作解决方案可确保接近达到限额使用阈值。根据请求的限额更改主动修复这些问题。

## 常见反模式：

- 未配置监控以检查服务限额阈值
- 没有为硬限制配置监控，即使这些值不能更改。
- 假定请求和确立软限额变化所需的时间是即时或短时间。
- 配置警报，以在快达到服务限额时发出警报，但没有关于如何对提醒做出响应的流程。
- 只为 AWS Service Quotas 支持的服务配置警报，不监控其他 AWS 服务。

- 不考虑多区域弹性设计 ( 如“主动/主动”、“主动/被动 – 热”、“主动/被动 – 冷”和“主动/被动 – 指示灯”方法 ) 的限额管理。
- 不评估区域之间的限额差异。
- 不评估每个区域对特定限额增加请求的需求。
- 不利用[模板进行多区域限额管理](#)。

建立此最佳实践的好处：自动跟踪 AWS Service Quotas，并根据这些限额监控您的使用情况，使您可以了解何时达到限额。您还可以使用此监控数据帮助限制由于限额耗尽而导致的降级。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

对于支持的服务，您可以通过配置各种可以进行评测的不同服务，然后发送警报，从而监控限额。这有助于监控使用情况，并可在接近限额时提醒您。这些警报可以从 AWS Config、Lambda 函数、Amazon CloudWatch 或从 AWS Trusted Advisor 触发。您还可以使用 CloudWatch Logs 上的指标筛选条件来搜索与提取日志中的模式，确定使用量是否快达到限额阈值。

### 实施步骤

对于监控：

- 获取当前资源使用量 ( 例如存储桶或实例 )。使用服务 API 操作 ( 例如 Amazon EC2 DescribeInstances API ) 来收集当前资源使用量。
- 使用以下项获得必要且适用于服务的当前限额：
  - AWS Service Quotas
  - AWS Trusted Advisor
  - AWS 文档
  - AWS 服务特定页面
  - AWS Command Line Interface ( AWS CLI )
  - AWS Cloud Development Kit (AWS CDK)
- AWS Service Quotas 是一项 AWS 服务，使用该服务可帮助您在一个地方管理超过 250 项 AWS 服务的限额。
- 使用 Trusted Advisor 服务限制来监控在各种阈值下的当前服务限制。
- 使用服务限额历史记录 ( 控制台或 AWS CLI ) 来检查区域增长情况。

- 如果需要，比较每个区域和每个账户中的服务限额变化，以形成等效性。

对于管理：

- 自动：设置 AWS Config 自定义规则以扫描各个区域的服务限额，并比较它们之间的差异。
- 自动：设置计划好的 Lambda 函数以扫描各个区域的服务限额，并比较它们之间的差异。
- 手动：通过 AWS CLI、API 或 AWS 控制台来扫描各个区域的服务限额，并比较它们之间的差异。报告差异。
- 如果在不同区域之间发现限额差异，如有需要，请求限额更改。
- 检查所有请求的结果。

资源

相关最佳实践：

- [REL01-BP01 了解服务限额和约束](#)
- [REL01-BP02 跨多个账户和区域管理服务限额](#)
- [REL01-BP03 通过架构适应固定服务限额和限制](#)
- [REL01-BP05 自动管理限额](#)
- [REL01-BP06 确保在当前限额与最大使用量之间存在足够的差距，以便应对失效转移](#)
- [REL03-BP01 选择如何划分工作负载](#)
- [REL10-BP01 将工作负载部署到多个位置](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)
- [REL11-BP03 自动修复所有层](#)
- [REL12-BP05 使用混沌工程测试弹性](#)

相关文档：

- [AWS Well-Architected Framework 的可靠性支柱：可用性](#)
- [AWS Service Quotas \( 以前称为服务限制 \)](#)
- [AWS Trusted Advisor 最佳实践检查 \( 请参阅“服务限制”部分 \)](#)
- [AWS Answers 上的 AWS Limit Monitor](#)
- [Amazon EC2 服务限制](#)



- [什么是 Service Quotas ?](#)
- [如何请求增加限额](#)
- [服务终端节点和限额](#)
- [Service Quotas 用户指南](#)
- [AWS 的限额监控](#)
- [AWS 故障隔离界限](#)
- [通过冗余实现可用性](#)
- [AWS 数据解决方案](#)
- [什么是持续集成 ?](#)
- [什么是持续交付 ?](#)
- [APN 合作伙伴：可帮助进行配置管理的合作伙伴](#)
- [在 AWS 上的每个租户一个账户的 SaaS 环境中管理账户生命周期](#)
- [管理和监控工作负载中的 API 节流](#)
- [使用 AWS Organizations 大规模查看 AWS Trusted Advisor 建议](#)
- [使用 AWS Control Tower 自动提升服务限制并实现企业支持](#)
- [Service Quotas 的操作、资源和条件键](#)

#### 相关视频：

- [AWS Live re:Inforce 2019 - Service Quotas](#)
- [使用 Service Quotas 查看和管理 AWS 服务的限额](#)
- [AWS IAM 限额演示](#)
- [AWS re:Invent 2018：闭环系统和开放思维：如何掌控不同规模的系统](#)

#### 相关工具：

- [AWS CodeDeploy](#)
- [AWS CloudTrail](#)
- [Amazon CloudWatch](#)
- [Amazon EventBridge](#)
- [Amazon DevOps Guru](#)
- [AWS Config](#)

- [AWS Trusted Advisor](#)
- [AWS CDK](#)
- [AWS Systems Manager](#)
- [AWS Marketplace](#)

## REL01-BP05 自动管理限额

实施工具以便在接近阈值时向您发送提醒。您可以自动发出限额提高请求：通过使用 AWS Service Quotas API，您可以自动发出限额提高请求。

如果将您的配置管理数据库 (CMDB) 或票证系统与 Service Quotas 集成，您可以自动跟踪配额提高请求和当前配额。除了 AWS 开发工具包之外，Service Quotas 还使用 AWS Command Line Interface (AWS CLI) 提供自动化。

常见反模式：

- 以电子表格的形式跟踪配额和使用情况。
- 每天、每周或每月运行使用情况报告，然后将使用情况与配额进行比较。

建立此最佳实践的好处：自动跟踪 AWS 服务限额，并根据这些限额监控您的使用情况，从而让您了解何时接近限额。您可以设置自动流程，帮助您在需要时提出配额提高请求。当使用情况趋向于相反的方向时，您可能需要考虑降低一些限额，以实现降低风险（如果凭据被盗）和节省成本的效果。

未建立此最佳实践暴露的风险等级：中

## 实施指导

- 设置自动监控：通过开发工具包实施各种工具，以便在接近阈值时向您发出提醒。
  - 利用 Service Quotas，通过自动限额监控解决方案（例如 AWS Limit Monitor 或从 AWS Marketplace 获得的产品）来增强服务。
  - [什么是 Service Quotas？](#)
  - [AWS 上的限额监控 – AWS 解决方案](#)
- 使用 Amazon SNS 和 AWS Service Quotas API 来根据限额阈值来设置触发响应。
- 测试自动化。
  - 配置限制阈值。
  - 与来自 AWS Config、部署管道、Amazon EventBridge 或第三方的更改事件集成。

- 人工设置低限额阈值来测试响应。
- 设置触发器以根据通知采取适当措施，并在必要时联系 AWS 支持。
- 人工触发更改事件。
- 运行实际测试以测试限额提高更改流程。

## 资源

### 相关文档：

- [APN 合作伙伴：可帮助进行配置管理的合作伙伴](#)
- [AWS Marketplace：可以帮助跟踪限制的 CMDB 产品](#)
- [AWS Service Quotas \( 以前称为服务限制 \)](#)
- [AWS Trusted Advisor 最佳实践检查 \( 见“服务限制”部分 \)](#)
- [AWS 上的限额监控 – AWS 解决方案](#)
- [Amazon EC2 服务限制](#)
- [什么是 Service Quotas ?](#)

### 相关视频：

- [AWS Live re:Inforce 2019 – Service Quotas](#)

REL01-BP06 确保在当前限额与最大使用量之间存在足够的差距，以便应对失效转移

当资源出现故障或无法访问时，该资源可能仍会被计入限额，直到成功终止资源。确认您的限额涵盖出现故障或无法访问的资源及其替换资源的重叠部分。在计算此差距时，应考虑网络故障、可用区故障或区域故障等使用案例。

期望结果：在当前服务阈值内可以覆盖资源或资源可访问性方面的或大或小的故障。在资源规划中已考虑到可用区故障、网络故障或甚至是区域故障。

### 常见反模式：

- 根据当前需求设置服务限额，而不考虑失效转移场景。
- 在计算服务的峰值限额时不考虑静态稳定性原则。
- 在计算每个区域所需的总限额时，不考虑可能无法访问资源的情况。

- 不考虑某些服务的 AWS 服务故障隔离边界及其可能的异常使用模式。

建立此最佳实践的好处：当服务中断事件影响应用程序可用性时，您可以借助云实施策略来缓解影响或从这些事件中恢复。此类策略通常包括创建额外资源来替换出现故障或无法访问的资源。您的限额策略会适应这些失效转移条件，并且不会由于服务限制耗尽而导致额外的降级。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

在评估限额限制时，请考虑由于某些降级而可能发生的失效转移情况。应考虑以下类型的失效转移情况：

- VPC 中断或无法访问。
- 子网无法访问。
- 可用区已显著降级，导致影响许多资源的可访问性。
- 系统阻止或更改了各种网络路由或入口点和出口点。
- 区域已显著降级，导致影响许多资源的可访问性。
- 有多个资源，但并非所有资源都受到区域或可用区故障的影响。

如上所列的故障可能是启动失效转移事件的触发器。因为业务影响可能会有很大差异，每种情况和每个客户的失效转移决策都是独特的。但是，当从操作上决定失效转移应用程序或服务时，必须在事件发生之前解决失效转移位置中资源的容量规划及其相关限额。

检查每个服务的服务限额，要考虑到可能会发生高于正常峰值的情况。这些峰值可能与由于网络或权限而可以访问但仍处于活动状态的资源有关。未终止的活动资源仍将计入服务限额限制。

### 实施步骤

- 确认您的服务限额与最高使用量之间有足够的差距，以便适应失效转移或失去可访问性。
- 根据您的部署模式、可用性要求和使用量增长情况确定服务限额。
- 根据需要请求增加限额。预计完成限额提高请求所需的时间。
- 确定可靠性要求（也称为“X 个 9”）。
- 构建故障场景（例如组件、可用区或区域缺失）。
- 确定部署方法（例如金丝雀部署、蓝/绿部署、红/黑部署或滚动部署）。
- 在当前限制中包含适当的缓冲区（例如 15%）。

- 在适当情况下包括静态稳定性（可用区和区域）的计算。
- 预计使用量增长（例如监控使用量趋势）。
- 考虑静态稳定性对最关键工作负载的影响。评估所有区域和可用区中适应静态稳定系统的资源。
- 考虑使用按需容量预留，以便在发生任何失效转移之前安排容量。在最关键的业务计划期间，这是一种有用的策略，可以在失效转移期间获得正确数量和类型的资源时降低潜在的风险。

## 资源

### 相关最佳实践：

- [REL01-BP01 了解服务限额和约束](#)
- [REL01-BP02 跨多个账户和区域管理服务限额](#)
- [REL01-BP03 通过架构适应固定服务限额和限制](#)
- [REL01-BP04 监控和管理限额](#)
- [REL01-BP05 自动管理限额](#)
- [REL03-BP01 选择如何划分工作负载](#)
- [REL10-BP01 将工作负载部署到多个位置](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)
- [REL11-BP03 自动修复所有层](#)
- [REL12-BP05 使用混沌工程测试弹性](#)

### 相关文档：

- [AWS Well-Architected Framework 的可靠性支柱：可用性](#)
- [AWS Service Quotas（以前称为服务限制）](#)
- [AWS Trusted Advisor 最佳实践检查（请参阅“服务限制”部分）](#)
- [AWS Answers 上的 AWS Limit Monitor](#)
- [Amazon EC2 服务限制](#)
- [什么是 Service Quotas？](#)
- [如何请求增加限额](#)
- [服务终端节点和限额](#)
- [Service Quotas 用户指南](#)

- [AWS 的限额监控](#)
- [AWS 故障隔离界限](#)
- [通过冗余实现可用性](#)
- [AWS 数据解决方案](#)
- [什么是持续集成？](#)
- [什么是持续交付？](#)
- [APN 合作伙伴：可帮助进行配置管理的合作伙伴](#)
- [在 AWS 上的每个租户一个账户的 SaaS 环境中管理账户生命周期](#)
- [管理和监控工作负载中的 API 节流](#)
- [使用 AWS Organizations 大规模查看 AWS Trusted Advisor 建议](#)
- [使用 AWS Control Tower 自动提升服务限制并实现企业支持](#)
- [Service Quotas 的操作、资源和条件键](#)

#### 相关视频：

- [AWS Live re:Inforce 2019 - Service Quotas](#)
- [使用 Service Quotas 查看和管理 AWS 服务的限额](#)
- [AWS IAM 限额演示](#)
- [AWS re:Invent 2018：闭环系统和开放思维：如何掌控不同规模的系统](#)

#### 相关工具：

- [AWS CodeDeploy](#)
- [AWS CloudTrail](#)
- [Amazon CloudWatch](#)
- [Amazon EventBridge](#)
- [Amazon DevOps Guru](#)
- [AWS Config](#)
- [AWS Trusted Advisor](#)
- [AWS CDK](#)
- [AWS Systems Manager](#)
- [AWS Marketplace](#)

## REL 2.如何规划网络拓扑？

工作负载通常存在于多个环境中。其中包括多个云环境（可公开访问云和私有云），可能还包括现有数据中心基础设施。相关计划必须涵盖网络注意事项，如系统内部和系统间连接、公有 IP 地址管理、私有 IP 地址管理，以及域名解析。

### 最佳实践

- [REL02-BP01 为工作负载公共端点使用高度可用的网络连接](#)
- [REL02-BP02 为云环境和本地部署环境之间的私有网络预置冗余连接](#)
- [REL02-BP03 确保 IP 子网分配考虑扩展和可用性](#)
- [REL02-BP04 轴辐式拓扑优先于多对多网格](#)
- [REL02-BP05 在互相连接的所有私有地址空间中强制实施非重叠的私有 IP 地址范围](#)

### REL02-BP01 为工作负载公共端点使用高度可用的网络连接

构建与工作负载的公共端点的高可用性网络连接有助于减少因连接丢失而导致的停机，并提高工作负载的可用性和改进 SLA。为此，需使用高度可用的 DNS、内容分发网络（CDN）、API Gateway、负载均衡或反向代理。

期望结果：为公共端点规划、构建和实施高可用性网络连接至关重要。如果您的工作负载由于连接中断而无法访问，即使您的工作负载正在运行且可用，您的客户也会看到您的系统处于关闭状态。通过将工作负载的公共端点的高可用性和弹性网络连接与工作负载本身的弹性架构结合起来，您可以为客户提供最佳的可用性和服务级别。

AWS Global Accelerator、Amazon CloudFront、Amazon API Gateway、AWS Lambda 函数 URL、AWS AppSync API 和 Elastic Load Balancing（ELB）均提供高可用性公共端点。Amazon Route 53 为域名解析提供高可用性 DNS 服务，以便确认可以解析公共端点地址。

您还可以评估 AWS Marketplace 软件设备是否适用于负载均衡和代理。

### 常见反模式：

- 在没有规划 DNS 和网络连接以实现高可用性的情况下设计高可用性工作负载。
- 在各个实例或容器中使用公有互联网地址并使用 DNS 管理与它们的连接。
- 使用 IP 地址而非域名来查找服务。
- 没有测试与公共端点的连接丢失的场景。
- 没有分析网络吞吐量需求和分发模式。



- 没有测试和计划与工作负载的公共端点的互联网连接可能中断的情况。
- 为较大地理区域提供内容（如网页、静态资产或媒体文件），而不使用内容分发网络。
- 没有为分布式拒绝服务（DDoS）攻击制定计划。DDoS 攻击会引发关闭您的用户的合法流量并降低可用性的风险。

建立此最佳实践的好处：设计高可用性和弹性网络连接，确保用户可以访问和使用您的工作负载。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

构建与公共端点的高可用性网络连接的核心是流量的路由。为了验证您的流量是否能够到达端点，DNS 必须能够将域名解析为其相应的 IP 地址。使用 Amazon Route 53 等高可用性和可扩展[域名系统 \(DNS\)](#) 来管理域的 DNS 记录。您还可以使用 Amazon Route 53 提供的运行状况检查。运行状况检查确认您的应用程序可访问、可用且正常运行，并且通过特殊的方式设置它们，使它们可以模仿用户的行为，例如请求网页或特定 URL。如果发生故障，Amazon Route 53 会响应 DNS 解析请求，并仅将流量定向到运行正常的端点。您还可以考虑使用 Amazon Route 53 提供的 Geo DNS 和基于延迟的路由功能。

为了确认您的工作负载本身具有高可用性，请使用 Elastic Load Balancing (ELB)。Amazon Route 53 可用于将流量定向到 ELB，然后它将流量分发到目标计算实例。您还可以将 Amazon API Gateway 与 AWS Lambda 结合使用，以实现无服务器解决方案。客户也可以在多个 AWS 区域中运行工作负载。借助[多站点主动/主动模式](#)，工作负载可以从多个区域提供流量。借助多站点主动/被动模式，工作负载可以从活动区域提供流量，而数据复制到辅助区域，在主区域发生故障时它变为活动。然后可使用 Route 53 运行状况检查来控制 DNS 从主区域中的任何端点失效转移到辅助区域中的端点，确认用户可以访问和使用您的工作负载。

Amazon CloudFront 提供一个简单的 API，通过使用世界各地的边缘站点网络提供请求，从而分发具有低延迟和高数据传输速率的内容。内容分发网络 (CDN) 提供位于或缓存在用户附近位置的内容，从而为客户提供服务。因为内容的加载从您的服务器转移到 CloudFront 的[边缘站点](#)，所以这也可以提高应用程序的可用性。边缘站点和区域边缘高速缓存在靠近查看者的位置保存内容的缓存副本，以便可以快速检索并提高工作负载的可访问性和可用性。

对于用户在地理上分散的工作负载，AWS Global Accelerator 可帮助您提高应用程序的可用性和性能。AWS Global Accelerator 提供任播静态 IP 地址，它充当在一个或多个 AWS 区域中托管的应用程序的固定入口点。这样就可以让流量在尽可能靠近用户的位置进入 AWS 全球网络，从而提高工作负载的可访问性和可用性。AWS Global Accelerator 还使用 TCP、HTTP 和 HTTPS 运行状况检查来监控应用程序端点的运行状况。端点的运行状况或配置发生任何更改，都会触发用户流量重定向到正常运行

的端点，为用户提供最佳性能和可用性。此外，AWS Global Accelerator 采用故障隔离设计，使用由独立网络区域提供的两个静态 IPv4 地址，提高了应用程序的可用性。

为帮助客户防范 DDoS 攻击，AWS 提供了 AWS Shield Standard。Shield Standard 会自动启用并防范 SYN/UDP 泛洪和反射攻击等常见的基础设施（第 3 层和第 4 层）攻击，以便在 AWS 上支持应用程序的高可用性。要对更复杂和更大规模的攻击（如 UDP 泛洪）、状态耗尽攻击（如 TCP SYN 泛洪）提供额外保护，以及为了帮助保护 Amazon Elastic Compute Cloud（Amazon EC2）、Elastic Load Balancing（ELB）、Amazon CloudFront、AWS Global Accelerator 和 Route 53 上运行的应用程序，您可以考虑使用 AWS Shield Advanced。为了防范 HTTP POST 或 GET 泛洪等应用程序层攻击，请使用 AWS WAF。AWS WAF 可使用 IP 地址、HTTP 标头、HTTP 主体、URI 字符串、SQL 注入和跨站点脚本条件来确定是应该阻止还是允许请求。

## 实施步骤

1. 设置高可用性 DNS：Amazon Route 53 是高可用性和可扩展的[域名系统（DNS）](#) Web 服务。Route 53 将用户请求连接到在 AWS 上或在本地运行的互联网应用程序。有关更多信息，请参阅[将 Amazon Route 53 配置为 DNS 服务](#)。
2. 设置运行状况检查：当使用 Route 53 时，确认只有正常运行的目标才可解析。首先[创建 Route 53 运行状况检查和配置 DNS 故障转移](#)。在设置运行状况检查时，必须考虑以下方面：
  - a. [Amazon Route 53 如何确定运行状况检查是否正常](#)
  - b. [创建、更新和删除运行状况检查](#)
  - c. [监控运行状况检查状态和获得通知](#)
  - d. [Amazon Route 53 DNS 的最佳实践](#)
3. [将 DNS 服务连接到端点](#)。
  - a. 当使用 Elastic Load Balancing 作为流量的目标时，使用指向负载均衡器的区域端点的 Amazon Route 53 来创建[别名记录](#)。在创建别名记录期间，将评估目标运行状况选项设置为“是”。
  - b. 对于无服务器工作负载或私有 API，当使用 API Gateway 时，使用[Route 53 将流量引向 API Gateway](#)。
4. 决定内容分发网络。
  - a. 要使用更靠近用户的边缘站点传输内容，首先了解[CloudFront 如何传输内容](#)。
  - b. 开始使用[简单 CloudFront 分发](#)。然后，CloudFront 知道您希望从何处传输内容，还知道有关如何跟踪和管理内容传输的详细信息。在设置 CloudFront 分发时，必须了解和考虑以下方面：
    - i. [缓存如何与 CloudFront 边缘站点配合使用](#)
    - ii. [提高直接从 CloudFront 缓存提供的请求的比例（缓存命中率）](#)
    - iii. [使用 Amazon CloudFront Origin Shield](#)

#### iv. [使用 CloudFront 来源失效转移优化高可用性](#)

5. 设置应用程序层保护：AWS WAF 帮助您免受可能会影响可用性、危及安全性或消耗过多资源的常见 Web 漏洞和机器人的攻击。若要更深入了解，请查看 [AWS WAF 的工作原理](#)，并且在您准备好实施保护措施来防范应用程序层 HTTP POST AND GET 泛洪时，请查看[开始使用 AWS WAF](#)。您还可以使用 AWS WAF 和 CloudFront，请参阅有关 [AWS WAF 如何与 Amazon CloudFront 功能配合使用](#) 的文档。
6. 设置额外的 DDoS 防护：默认情况下，所有 AWS 客户都可以免费使用 AWS Shield Standard 获得保护，防范针对您的网站或应用程序的常见、最常发生的网络和传输层 DDoS 攻击。对于在 Amazon EC2、Elastic Load Balancing、Amazon CloudFront、AWS Global Accelerator 和 Amazon Route 53 上运行，面向互联网的应用程序，要获得额外保护，您可以考虑 [AWS Shield Advanced](#) 并查看 [DDoS 弹性架构的示例](#)。要保护您的工作负载和公共端点，防止受到 DDoS 攻击，请参阅[开始使用 AWS Shield Advanced](#)。

## 资源

### 相关最佳实践：

- [REL10-BP01 将工作负载部署到多个位置](#)
- [REL10-BP02 为您的多位置部署选择合适的位置](#)
- [REL11-BP04 恢复期间依赖于数据平面而不是控制平面](#)
- [REL11-BP06 当事件影响可用性时发出通知](#)

### 相关文档：

- [AWS 合作伙伴：可帮助您规划联网的合作伙伴](#)
- [适用于网络基础设施的 AWS Marketplace](#)
- [什么是 AWS Global Accelerator？](#)
- [什么是 Amazon CloudFront？](#)
- [什么是 Amazon Route 53？](#)
- [什么是 Elastic Load Balancing？](#)
- [网络连接能力 - 建立您的云基础](#)
- [什么是 Amazon API Gateway？](#)
- [什么是 AWS WAF、AWS Shield 和 AWS Firewall Manager？](#)
- [什么是 Amazon Route 53 Application Recovery Controller？](#)

- [配置自定义运行状况检查来进行 DNS 故障转移](#)

相关视频：

- [AWS re:Invent 2022 - 使用 AWS Global Accelerator 提高性能和可用性](#)
- [AWS re:Invent 2020：使用 Amazon Route 53 进行全球流量管理](#)
- [AWS re:Invent 2022 - 运行高可用性多可用区应用程序](#)
- [AWS re:Invent 2022 - 深入了解 AWS 网络基础设施](#)
- [AWS re:Invent 2022 - 建立弹性网络](#)

相关示例：

- [使用 Amazon Route 53 Application Recovery Controller \(ARC\) 进行灾难恢复](#)
- [可靠性研讨会](#)
- [AWS Global Accelerator 研讨会](#)

REL02-BP02 为云环境和本地部署环境之间的私有网络预置冗余连接

在单独部署的私有网络之间使用多个 AWS Direct Connect 连接或 VPN 隧道。使用多个 Direct Connect 位置实现高可用性。如果使用多个 AWS 区域，请确保其中至少有两个区域存在冗余。您可能想要评估终止 VPN 的 AWS Marketplace 设备。如果您使用 AWS Marketplace 设备，请在不同的可用区中部署冗余实例以实现高可用性。

AWS Direct Connect 是一项云服务，可以在本地环境和 AWS 之间轻松建立专用网络连接。使用 Direct Connect Gateway，您的本地数据中心可以连接至跨多个 AWS 区域的多个 AWS VPC。

此类冗余可解决会对连接弹性造成影响的潜在故障：

- 您将如何灵活应对拓扑中的故障？
- 如果您配置错了某些内容并删除了连接，会发生什么情况？
- 您是否能应对服务流量或使用量的意外增加？
- 您是否能够吸收未遂的分布式拒绝服务 (DDoS) 攻击？

若通过 VPN 将您的 VPC 连接至本地数据中心，您应该考虑在选择运行该设备所需的供应商和实例大小时所需要的弹性和带宽要求。如果您使用的 VPN 设备在其实施中没有弹性，则您应该通过第二个设

备获取冗余连接。对于所有这些场景，您需要定义可接受的恢复时间并进行测试，以确保您可以满足这些要求。

如果选择通过 Direct Connect 连接将您的 VPC 连接至数据中心，而且您要求连接具有高可用性，请从每个数据中心获得冗余 Direct Connect 连接。冗余连接应使用来自与第一个不同位置的其他 Direct Connect 连接。如果您有多个数据中心，则确保连接在不同的位置终止。使用 [Direct Connect 弹性工具包](#) 以帮助您完成设置。

如果您选择使用 AWS VPN 并通过互联网失效转移到 VPN，一定要了解，它支持每个 VPN 隧道高达 1.25-Gbps 的吞吐量，但在多个 AWS 托管 VPN 隧道终止于同一 VGW 的情况下，不支持将等价多路径 (ECMP, Equal Cost Multi Path) 用于出站流量。我们不建议您使用 AWS 托管 VPN 作为 Direct Connect 连接的备份，除非您可以接受失效转移期间的速度低于 1 Gbps。

您还可以使用 VPC 端点将您的 VPC 私下连接至受支持的 AWS 服务和 VPC 端点服务，它们得到 AWS PrivateLink 的支持而无需通过公有互联网传输。终端节点为虚拟设备。它们是水平扩展、冗余，而且高度可用的 VPC 组件。它们支持在您的 VPC 和服务实例之间进行通信，而不会对您的网络流量施加可用性风险或带宽限制。

常见反模式：

- 在本地网络和 AWS 之间只有一个连接供应方。
- 使用 AWS Direct Connect 连接的功能，但只有一个连接。
- 您的 VPN 连接只有一条路径。

建立此最佳实践的好处：通过在云环境和公司或本地部署环境之间实施冗余连接，您可以确保两个环境之间的依赖服务能够可靠通信。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

- 确保您在 AWS 和本地部署环境之间具有高度可用的连接。在单独部署的私有网络之间使用多个 AWS Direct Connect 连接或 VPN 隧道。使用多个 Direct Connect 位置实现高可用性。如果使用多个 AWS 区域，请确保其中至少有两个区域存在冗余。您可能想要评估终止 VPN 的 AWS Marketplace 设备。如果您使用 AWS Marketplace 设备，请在不同的可用区中部署冗余实例以实现高可用性。
- 确保您拥有面向本地部署环境的冗余连接。您可能需要面向多个 AWS 区域的冗余连接，以满足可用性需求。
- [AWS Direct Connect 弹性建议](#)

- [使用冗余 Site-to-Site VPN 连接以提供故障转移](#)
  - 使用服务 API 操作确定正确使用了 Direct Connect 线路。
    - [DescribeConnections](#)
    - [DescribeConnectionsOnInterconnect](#)
    - [DescribeDirectConnectGatewayAssociations](#)
    - [DescribeDirectConnectGatewayAttachments](#)
    - [DescribeDirectConnectGateways](#)
    - [DescribeHostedConnections](#)
    - [DescribeInterconnects](#)
  - 如果您仅有一个 Direct Connect 连接或没有此连接，请设置连接虚拟私有网关的冗余 VPN 隧道。
    - [什么是 AWS Site-to-Site VPN ?](#)
- 捕获您的当前连接（例如，Direct Connect、虚拟私有网关、AWS Marketplace 设备）。
  - 使用服务 API 操作查询 Direct Connect 连接的配置。
    - [DescribeConnections](#)
    - [DescribeConnectionsOnInterconnect](#)
    - [DescribeDirectConnectGatewayAssociations](#)
    - [DescribeDirectConnectGatewayAttachments](#)
    - [DescribeDirectConnectGateways](#)
    - [DescribeHostedConnections](#)
    - [DescribeInterconnects](#)
  - 使用服务 API 操作收集路由表使用的虚拟私有网关。
    - [DescribeVpnGateways](#)
    - [DescribeRouteTables](#)
  - 使用服务 API 操作收集路由表使用的 AWS Marketplace 应用程序。
    - [DescribeRouteTables](#)

## 资源

### 相关文档：

基础

- [AWS 合作伙伴：可帮助您规划联网的合作伙伴](#)



- [AWS Direct Connect 弹性建议](#)
- [适用于网络基础设施的 AWS Marketplace](#)
- [Amazon Virtual Private Cloud 连接选项白皮书](#)
- [多数据中心 HA 网络连接](#)
- [使用冗余 Site-to-Site VPN 连接以提供故障转移](#)
- [使用 Direct Connect 弹性工具包开始操作](#)
- [VPC 终端节点和 VPC 终端节点服务 \(AWS PrivateLink\)](#)
- [什么是 Amazon VPC ?](#)
- [什么是 Transit Gateway ?](#)
- [什么是 AWS Site-to-Site VPN ?](#)
- [使用 Direct Connect 网关](#)

相关视频：

- [AWS re:Invent 2018 : Amazon VPC 的高级 VPC 设计和新功能 \( NET303 \)](#)
- [AWS re:Invent 2019 : 适用于众多 VPC 的 AWS Transit Gateway 参考架构 \( NET406-R1 \)](#)

## REL02-BP03 确保 IP 子网分配考虑扩展和可用性

Amazon VPC IP 地址范围必须足够大，以满足工作负载的要求，包括考虑未来的扩展以及跨可用区为子网分配 IP 地址。这包括负载均衡器、EC2 实例和基于容器的应用程序。

当您规划网络拓扑时，第一步是定义 IP 地址空间本身。应（按照 RFC 1918 准则）为每个 VPC 分配私有 IP 地址范围。作为此流程的一部分，要满足以下要求：

- 在每个区域中为多个 VPC 留出 IP 地址空间。
- 在 VPC 内，为跨多个可用区的多个子网留出空间。
- 始终在 VPC 内保留未使用的 CIDR 块空间以用于未来扩展。
- 确保有 IP 地址空间以满足您可能使用的任何 EC2 实例临时性队列的需求，如适用于机器学习的 Spot 队列、Amazon EMR 集群或 Amazon Redshift 集群。
- 注意，每个子网 CIDR 块中的前四个 IP 地址和最后一个 IP 地址将被预留而无法供您使用。
- 您应计划部署大型 VPC CIDR 块。注意，最初被分配到您的 VPC 的 VPC CIDR 块无法被更改或删除，但您可以向 VPC 添加额外的非重叠的 CIDR 块。虽然无法更改 IPv4 CIDR，但可以对



IPv6 CIDR 进行更改。请记住，部署最大的 VPC (/16) 会产生超过 65000 个 IP 地址。单单在基础 10.x.x.x IP 地址空间内，您可以预置 255 个这样的 VPC。因此，您可以趋向于过大数量而不是过小数量，这样更容易管理 VPC。

常见反模式：

- 创建小型 VPC。
- 创建小型子网，然后在业务增长过程中向配置添加子网。
- 错误估计 Elastic Load Balancer 可以使用的 IP 地址数量。
- 在相同子网中部署多个高流量负载均衡器。

建立此最佳实践的好处：这可确保您能适应工作负载增长要求，并在扩展过程中继续提供可用性。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

- 规划您的网络以适应增长、符合监管合规性以及实现与其他服务的集成。如果没有合理的规划，则增长可能会被低估、监管合规性可能会发生变化并且收购或私有网络连接可能难以实施。
- 根据您的服务要求、延迟、法规和灾难恢复 (DR, Disaster Recovery) 要求选择相关 AWS 账户和区域。
- 确定您的区域 VPC 部署需求。
- 确定 VPC 的大小。
  - 确定您是否要部署多 VPC 连接。
    - [什么是 Transit Gateway ?](#)
    - [单区域多 VPC 连接](#)
  - 确定您是否需要隔离网络以满足法规要求。
  - 使 VPC 尽可能大。最初为 VPC 分配的 VPC CIDR 块无法更改或删除，但您可以向 VPC 添加额外的非重叠 CIDR 块。但是，这样可能会分割您的地址范围。
  - 使 VPC 尽可能大。最初为 VPC 分配的 VPC CIDR 块无法更改或删除，但您可以向 VPC 添加额外的非重叠 CIDR 块。但是，这样可能会分割您的地址范围。

资源

相关文档：

- [AWS 合作伙伴：可帮助您规划联网的合作伙伴](#)
- [适用于网络基础设施的 AWS Marketplace](#)
- [Amazon Virtual Private Cloud 连接选项白皮书](#)
- [多数据中心 HA 网络连接](#)
- [单区域多 VPC 连接](#)
- [什么是 Amazon VPC？](#)

相关视频：

- [AWS re:Invent 2018：Amazon VPC 的高级 VPC 设计和新功能 \( NET303 \)](#)
- [AWS re:Invent 2019：适用于众多 VPC 的 AWS Transit Gateway 参考架构 \( NET406-R1 \)](#)

#### REL02-BP04 轴辐式拓扑优先于多对多网络

如果通过 VPC 对等连接、AWS Direct Connect 或 VPN 连接的网络地址空间超过两个（例如，VPC 和本地网络），则使用与 AWS Transit Gateway 所提供的模型类似的轴辐式模型。

如果只有两个此类网络，您可以简单地使其相互连接，但随着网络数量的增加，这种网络连接的复杂性将变得无法接受。AWS Transit Gateway 提供易于维护的轴辐式模型，允许在您的多个网络中对流量进行路由。

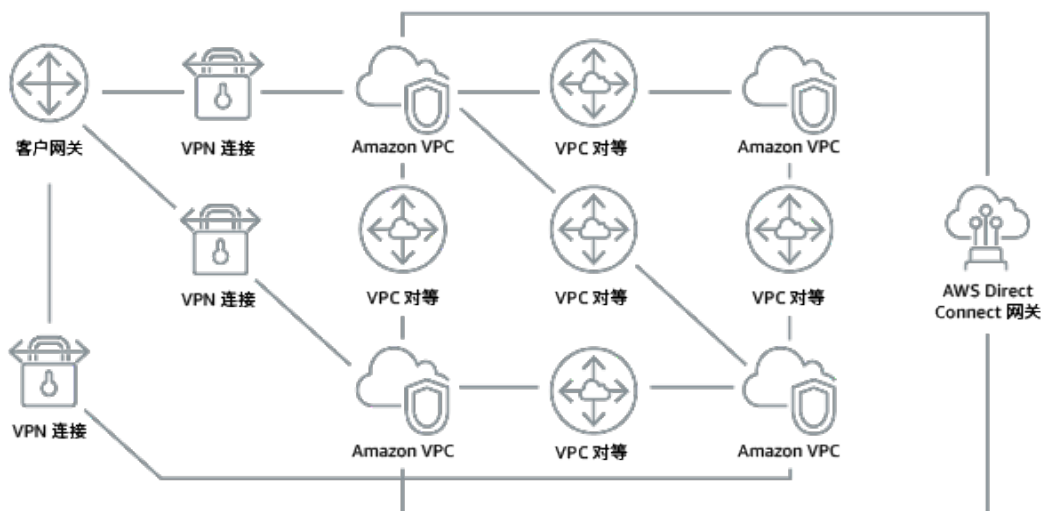


图 1：没有 AWS Transit Gateway：您需要将各个 Amazon VPC 对等连接并使用 VPN 连接来连接到各个现场位置，其复杂程度会随着它的扩展而递增。

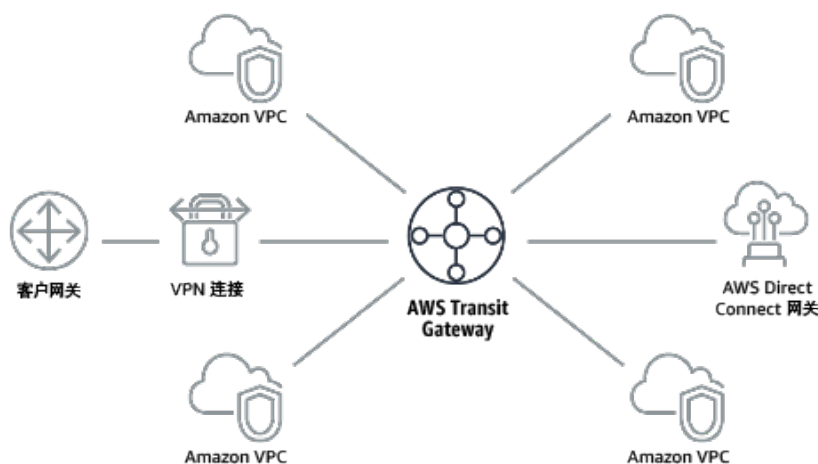


图 2：使用 AWS Transit Gateway：您只需将每个 Amazon VPC 或 VPN 连接至 AWS Transit Gateway，而且它会在每个 VPC 或 VPN 之间往返路由流量。

常见反模式：

- 使用 VPC 对等连接来连接两个以上的 VPC。
- 为每个 VPC 建立多个 BGP 会话，从而建立跨多个 AWS 区域的虚拟私有云 ( VPC，Virtual Private Cloud ) 连接。

建立此最佳实践的好处：随着网络数量增加，这种复杂的网格连接将变得无法维持。AWS Transit Gateway 提供易于维护的轴辐式模型，允许在您的多个网络中路由流量。

未建立这种最佳实践的情况下暴露的风险等级：中

#### 实施指导

- 轴辐式拓扑优先于多对多网格。如果通过 VPC 对等连接、AWS Direct Connect 或 VPN 连接的网络地址空间超过两个 ( VPC，本地网络 )，则使用与 AWS Transit Gateway 所提供的模型类似的轴辐式模型。
- 如果只有两个此类网络，您只需将其相互连接即可，但随着网络数量的增长，这种复杂的网格连接将变得无法维持。AWS Transit Gateway 提供易于维护的轴辐式模型，允许在您的多个网络中路由流量。
- [什么是 Transit Gateway？](#)

## 资源

### 相关文档：

- [AWS 合作伙伴：可帮助您规划联网的合作伙伴](#)
- [适用于网络基础设施的 AWS Marketplace](#)
- [多数据中心 HA 网络连接](#)
- [VPC 终端节点和 VPC 终端节点服务 \(AWS PrivateLink\)](#)
- [什么是 Amazon VPC？](#)
- [什么是 Transit Gateway？](#)

### 相关视频：

- [AWS re:Invent 2018：Amazon VPC 的高级 VPC 设计和新功能 \( NET303 \)](#)
- [AWS re:Invent 2019：适用于众多 VPC 的 AWS Transit Gateway 参考架构 \( NET406-R1 \)](#)

REL02-BP05 在互相连接的所有私有地址空间中强制实施非重叠的私有 IP 地址范围

多个 VPC 通过对等连接或 VPN 连接时，各个 VPC 的 IP 地址范围不得重叠。与之类似，您必须避免 VPC 和本地部署环境或与其他您使用的云提供商之间出现 IP 地址冲突。您还必须能够在需要时分配私有 IP 地址范围。

IP 地址管理 (IPAM) 系统可以帮助解决这个问题。AWS Marketplace 提供多个 IPAM。

### 常见反模式：

- 在 VPC 中使用与本地部署或企业网络相同的 IP 范围。
- 不必追踪用于部署工作负载的 VPC 的 IP 范围。

建立此最佳实践的好处：主动规划网络可确保您不会遇到互连网络中多次出现相同 IP 地址的情况。这可防止使用不同应用程序的工作负载部分出现路由问题。

未建立此最佳实践暴露的风险等级：中

## 实施指导

- 监控和管理您的 CIDR 使用。评估您在 AWS 上的可能使用量、将 CIDR 范围添加到现有 VPC 并创建 VPC 以便使用量实现计划增长。

- 捕获当前的 CIDR 使用量 ( 例如 , VPC、子网 )
  - 使用服务 API 操作收集当前的 CIDR 使用量数据。
- 捕获您当前的子网使用量。
  - 使用服务 API 操作在每个区域中按 VPC 收集子网。
    - [DescribeSubnets](#)
  - 记录当前使用量。
  - 确定您是否创建了任何重叠 IP 范围。
  - 计算备用容量。
  - 记录重叠的 IP 范围。您可以迁移到新地址范围 , 或使用 AWS Marketplace 的网络和端口转换 ( NAT ) 设备 ( 如果需要连接重叠范围 ) 。

## 资源

### 相关文档 :

- [AWS 合作伙伴 : 可帮助您规划联网的合作伙伴](#)
- [适用于网络基础设施的 AWS Marketplace](#)
- [Amazon Virtual Private Cloud 连接选项白皮书](#)
- [多数据中心 HA 网络连接](#)
- [什么是 Amazon VPC ?](#)
- [什么是 IPAM ?](#)

### 相关视频 :

- [AWS re:Invent 2018 : Amazon VPC 的高级 VPC 设计和新功能 \( NET303 \)](#)
- [AWS re:Invent 2019 : 适用于众多 VPC 的 AWS Transit Gateway 参考架构 \( NET406-R1 \)](#)

## 工作负载架构

### 问题

- [REL 3.如何设计工作负载服务架构 ?](#)
- [REL 4.您如何在分布式系统中设计交互以预防发生故障 ?](#)
- [REL 5.您如何在分布式系统中进行交互设计 , 从而缓解或经受住故障影响 ?](#)

## REL 3. 如何设计工作负载服务架构？

使用服务导向型架构 ( SOA ) 或微服务架构构建高度可扩展的可靠工作负载。服务导向型架构 ( SOA ) 可通过服务接口使软件组件可重复使用。微服务架构则进一步让组件变得更小、更简单。

### 最佳实践

- [REL03-BP01 选择如何划分工作负载](#)
- [REL03-BP02 构建专注于特定业务领域和功能的服务](#)
- [REL03-BP03 根据 API 提供服务合同](#)

### REL03-BP01 选择如何划分工作负载

在确定应用程序的弹性要求时，工作负载划分很重要。应尽可能避免使用整体式架构。相反，应仔细考虑哪些应用程序组件可以分解为多项微服务。根据您的应用程序要求，最终应尽可能采用服务导向型架构 ( SOA ) 与微服务组合的形式。能够实现无状态的工作负载更容易部署为微服务。

期望结果：工作负载应该可支持、可扩展，并尽可能地松散耦合。

在选择如何划分工作负载时，要权衡其优点和复杂性。适用于即将首次发布的新产品的功能有别于从一开始就构建用于扩展的工作负载的需求。重构一个现有的整体架构时，您需要考虑应用程序对无状态分解的支持程度。通过将服务分解为较小的部分，可以让职责明确的小型团队来开发和管理它们。然而，较小的服务会带来复杂性，包括可能会增加延迟，调试变得更复杂，而且加重运营负担。

### 常见反模式：

- 如示例所示，[微服务 Death Star](#) 是这样一种情况：原子组件变得高度相互依赖，牵一发而动全身，使组件像一块整体一样死板而又脆弱。

### 建立此实践的好处：

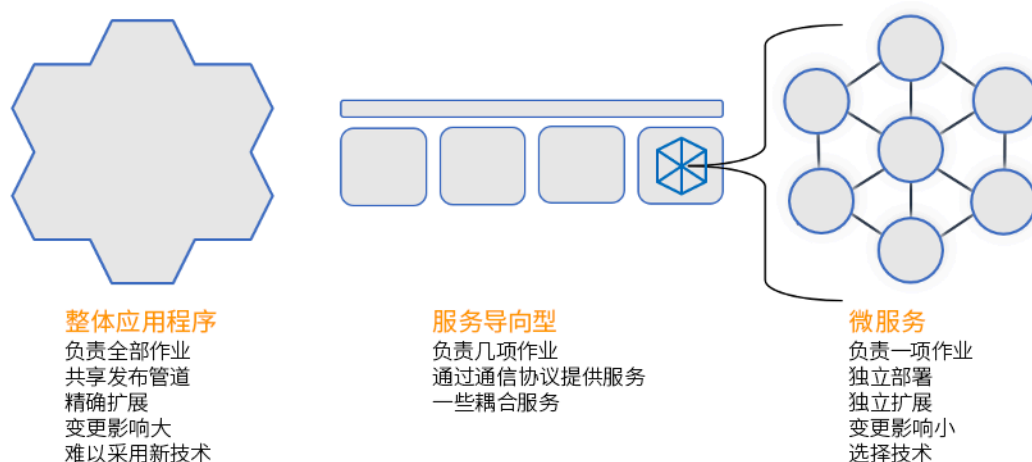
- 更多特定分段可以提高敏捷性、组织灵活性和可扩展性。
- 减小了服务中断的影响。
- 应用程序组件可能有不同的可用性要求，可通过更加原子化的分段来满足这些要求。
- 支持工作负载的团队职责分明。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

请根据工作负载的分段方式选择架构类型。选择 SOA 或微服务架构（极少数情况下是整体架构）。即便在刚开始选择整体架构，您必须确保它是模块化的，而且由于您的产品随着采用的用户增加而扩展，它最终也可转变成为 SOA 或微服务架构。SOA 和微服务各自提供较小的区段，它们是现代可扩展和可靠架构的首选，但您需要认真权衡利弊，尤其在部署微服务架构时。

一项主要的权衡是您现在使用的是分布式计算架构，可能更难实现用户延迟要求，还增加了调试和跟踪用户交互的复杂性。您可以使用 AWS X-Ray 来帮助解决此问题。需要考虑的另一个影响是，您管理的应用程序数量增加时，运营复杂性也会随之增加，这要求部署多个相互独立组件。



## 整体架构、服务导向型架构和微服务架构

### 实施步骤

- 确定构建或重构应用程序所需的适当架构。SOA 和微服务分别提供较小分段，这是现代可扩展的可靠架构的首选。要在实现较小分段的同时避免一些微服务复杂性，SOA 是很好的折中方案。有关更多详细信息，请参阅 [微服务利弊权衡](#)。
- 如果您的工作负载适合，并且您的组织可以支持，则应使用微服务架构来实现最佳敏捷性和可靠性。有关更多详细信息，请参阅 [在 AWS 上实施微服务](#)。
- 考虑遵循 [Strangler Fig 模式](#)，将整体架构重构为较小的组件。这包括逐步用新的应用程序和服务替换特定的应用程序组件。[AWS Migration Hub Refactor Spaces](#) 作为增量重构的起点。有关更多详细信息，请参阅 [使用绞杀者模式无缝迁移本地传统工作负载](#)。
- 实施微服务可能需要采用一种服务发现机制，让这些分布式服务能够相互通信。[AWS App Mesh](#) 可用于服务导向型架构，以实现可靠的发现和服务访问。[AWS Cloud Map](#) 也可用于动态的基于 DNS 的服务发现。



- 如果您要从整体架构迁移到 SOA，[Amazon MQ](#) 可作为服务总线来帮助弥合这一差距（在云中重新设计传统应用程序时）。
- 对于具有单个共享数据库的现有整体架构，请选择将数据重组为较小分段的方式。可以按业务部门、访问模式或数据结构来划分。在重构过程的这一阶段，应选择是使用关系型还是非关系型（NoSQL）数据库来继续操作。有关更多详细信息，请参阅 [从 SQL 到 NoSQL](#)。

实施计划的工作量级别：高

资源

相关最佳实践：

- [REL03-BP02 构建专注于特定业务领域和功能的服务](#)

相关文档：

- [Amazon API Gateway：使用 OpenAPI 配置 REST API](#)
- [什么是服务导向型架构？](#)
- [边界上下文（领域驱动设计的中心模式）](#)
- [在 AWS 上实施微服务](#)
- [微服务利弊权衡](#)
- [微服务 – 一个全新架构术语的定义](#)
- [AWS 上的微服务](#)
- [什么是 AWS App Mesh？](#)

相关示例：

- [迭代应用程序现代化研讨会](#)

相关视频：

- [利用 AWS 上的微服务实现卓越](#)

## REL03-BP02 构建专注于特定业务领域和功能的服务

服务导向型架构 ( SOA ) 采用按业务需求定义的、划分明确的功能来定义服务。微服务使用域模型和限界上下文，沿业务环境边界划定服务边界。通过将重点放在业务领域和功能上，有助于团队为其服务定义独立的可靠性要求。限界上下文隔离和封装业务逻辑，让团队能够更好地解释如何处理故障。

期望的结果：工程师和业务利益相关者共同定义限界上下文，并使用它们将系统设计为实现特定业务功能的服务。这些团队使用事件风暴等既定的实践来定义需求。新的应用程序被设计为服务，具有明确定义的边界并采用松耦合。现有的整体式架构被分解为 [限界上下文](#)，而系统设计转向 SOA 或微服务架构。重构整体式架构时，会应用气泡上下文和整体式架构分解模式等既定方法。

面向领域的服务作为一个或多个进程执行，彼此之间不分享状态。这些进程独立应对需求的波动，并根据特定领域的要求处理故障情景。

常见反模式：

- 围绕特定技术领域（例如 UI 和 UX、中间件或数据库）组建团队，而不是根据特定的业务领域组建。
- 应用程序进行了领域职责划分。跨越限界上下文的服务可能更难于维护，需要更多测试工作，并且需要多个领域团队参与软件更新。
- 领域依赖关系（例如领域实体库）在服务之间共享，因此对一个服务领域进行更改也需要更改其他服务领域。
- 服务合同和业务逻辑没有使用通用且一致的领域语言来描述实体，这会导致翻译层的存在，使得系统更加复杂并增加调试工作量。

建立此最佳实践的好处：应用程序被设计为独立的服务，按照业务领域确定界限，并使用通用的业务语言。服务可独立测试和部署。对于所实施的领域，服务满足特定于领域的弹性要求。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

领域驱动型决策 ( DDD ) 是围绕业务领域设计和构建软件的基本方法。围绕业务领域构建服务时，使用现有框架会很有帮助。在使用现有的整体式应用程序时，您可以利用分解模式提供成熟的技术，将应用程序现代化改造为服务。



## 领域驱动型决策

### 实施步骤

- 团队可以举办 [事件风暴](#) 研讨会，以简便的便签格式，快速确定事件、命令、聚合和领域。
- 在领域上下文中构建领域实体和功能后，您可以使用 [限界上下文](#) 将领域划分为服务，具有相似功能和属性的实体分组在一起。通过将模型细分为不同的上下文，即可得到如何确定微服务边界的模板。
  - 以 Amazon.com 网站为例，实体可能包括包装、配送、时间表、价格、折扣和货币。
  - 包装、配送和时间表分组到运输上下文中，而价格、折扣和货币分组到定价上下文中。
- [将整体式架构分解为微服务](#) 概述了重构微服务的模式。使用按照业务功能、子领域或事务进行分解的模式，与领域驱动型方法非常吻合。
- 利用 [气泡上下文](#) 等战术性技巧，您可以在现有或旧版应用程序中引入 DDD，无需预先重新编写和承诺完全转向 DDD。在气泡上下文方法中，使用服务映射和协调来建立小型限界上下文，或者建立 [防护层](#)，该层保护新定义的领域模型免受外部影响。

在团队进行了领域分析并定义实体和服务合同后，他们可以利用 AWS 服务，将自己的领域驱动型设计作为云端服务来实施。

- 通过定义执行领域业务规则的测试来开始开发。测试驱动型开发（TDD）和行为驱动型开发（BDD），有助于团队将服务重点放在解决业务问题上。
- 选择最能满足您的业务领域要求的 [AWS 服务](#) 以及 [微服务架构](#)：
  - [AWS 无服务器](#) 让您的团队可以将精力集中于具体的领域逻辑上，而不是管理服务器和基础设施。
  - [AWS 上的容器](#) 可简化基础设施的管理，这样您便可以将精力集中于领域需求方面。
  - [专用数据库](#) 有助于您将领域需求与最适合的数据库类型相匹配。
- 在 [AWS 上构建六边形架构](#) 中概述了一个框架，从业务领域出发，采用逆向工作方法，将业务逻辑构建到服务中，以满足功能需求，然后连接集成适配器。采用将接口详细信息从业务逻辑与 AWS 服务之间分离的模式，让团队能够专注于研究领域功能并提高软件质量。

## 资源

### 相关最佳实践：

- [REL03-BP01 选择如何划分工作负载](#)
- [REL03-BP03 根据 API 提供服务合同](#)

### 相关文档：

- [AWS 微服务](#)
- [在 AWS 上实施微服务](#)
- [如何将整体式架构分解为多项微服务](#)
- [在被遗留系统包围时通过 DDD 开始着手](#)
- [领域驱动型设计：解决软件核心的复杂性](#)
- [在 AWS 上构建六边形架构](#)
- [将整体式架构分解为微服务](#)
- [事件风暴](#)
- [限界上下文之间的消息](#)
- [微服务](#)
- [测试驱动型开发](#)
- [行为驱动型开发](#)

### 相关示例：

- [企业云原生研讨会](#)
- [在 AWS 上设计云原生微服务 \( 来源：DDD/EventStormingWorkshop \)](#)

### 相关工具：

- [AWS Cloud 数据库](#)
- [AWS 上的无服务器](#)
- [AWS 上的容器](#)

## REL03-BP03 根据 API 提供服务合同

服务合同是 API 生产者与使用者之间的书面协议，采用机器可读的 API 定义形式进行定义。合同版本控制策略让使用者能够继续使用现有的 API，并在更新的 API 准备就绪时，将其应用程序迁移到更新的 API。只要遵守合同，生产者部署可随时进行。服务团队可以使用自己选择的技术堆栈来满足 API 合同要求。

期望的结果：

常见反模式：使用服务导向型架构或微服务架构所构建的应用程序能够独立运行，但具有集成的运行时系统依赖项。当双方都遵循共同的 API 合同时，向 API 使用者或生产者部署更改并不会影响整个系统的稳定性。通过服务 API 进行通信的组件能够独立执行功能发布、升级到运行时系统依赖项或者失效转移到灾难恢复（DR）站点，而彼此之间的影响很小，或者根本没有影响。此外，离散服务能够独立扩展来满足资源需求，无需统一扩展其他服务。

- 创建不使用强类型架构的服务 API。这样，API 不能用来生成无法通过编程方式验证的 API 绑定和有效负载。
- 不采用版本控制策略，会迫使 API 使用者在服务合同变化时进行更新和发布，否则会出现故障。
- 错误消息会泄露基础服务的实施细节，而不是按照域环境和语言描述集成故障。
- 不使用 API 合同开发测试用例和模拟 API 实施，以便对服务组件进行独立测试。

建立此最佳实践的好处：分布式系统由通过 API 服务合同进行通信的组件组成，可以提高可靠性。开发人员可以在开发过程的早期发现潜在问题，在编译期间进行类型检查，来验证请求和响应是否符合 API 合同以及是否存在必填字段。API 合同为 API 提供了清晰的自描述接口，并在不同的系统和编程语言之间提供了更好的互操作性。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

确定业务领域并确定工作负载划分后，即可开发服务 API。首先，为 API 定义机器可读的服务合同，然后实施 API 版本控制策略。在准备好通过 REST、GraphQL 等常见协议或异步事件来集成服务时，您可以将 AWS 服务整合到架构中，从而将您的组件与强类型的 API 合同集成。

### 面向服务 API 合同的 AWS 服务

将 AWS 服务（包括 [Amazon API Gateway](#)、[AWS AppSync](#) 和 [Amazon EventBridge](#)）等整合到架构中，以便在您的应用程序中使用 API 服务合同。使用 Amazon API Gateway 有助于直接与原生 AWS

服务及其他 Web 服务集成。API Gateway 支持 [OpenAPI 规范](#) 和版本控制。AWS AppSync 是托管的 [GraphQL](#) 端点，在配置该端点时，您需要定义 GraphQL 架构，进而为查询、突变和订阅定义服务接口。Amazon EventBridge 使用事件架构来定义事件并为您的事件生成代码绑定。

## 实施步骤

- 首先，为您的 API 定义一个合同。合同将说明 API 的功能，并为 API 的输入和输出定义强类型的数据对象和字段。
- 在 API Gateway 中配置 API 时，您可以导入和导出端点的 OpenAPI 规范。
  - [导入 OpenAPI 定义](#) 可简化 API 的创建过程，并可与 AWS 基础设施即代码工具（例如 [AWS Serverless Application Model](#) 和 [AWS Cloud Development Kit \(AWS CDK\)](#)）集成。
  - [导出 API 定义](#) 可简化与 API 测试工具的集成，并为服务使用者提供集成规范。
- 要使用 AWS AppSync 定义和管理 GraphQL API，您可以 [定义 GraphQL 架构](#) 文件来生成合同接口，并简化与复杂的 REST 模型、众多数据库表或传统服务的交互。
- [AWS Amplify](#) 项目与 AWS AppSync 集成后，会生成强类型的 JavaScript 查询文件供应用程序使用，还会生成 AWS AppSync GraphQL 客户端库供 [Amazon DynamoDB](#) 表使用。
- 当您使用来自 Amazon EventBridge 的服务事件时，事件遵循架构注册表中已有的架构或您使用 OpenAPI 规范定义的架构。通过在注册表中定义架构，您还可以从架构合同生成客户端绑定，以将代码与事件集成。
- 扩展 API 或者实施 API 版本控制。在添加可以配置为可选字段的字段时，或者为必填字段添加默认值时，扩展 API 是一种相对比较简单的事项。
  - 对于 REST 和 GraphQL 等协议，基于 JSON 的合同可能非常适合合同扩展。
  - 对于 SOAP 等协议，基于 XML 的合同应与服务使用者一起进行测试，来确定合同扩展的可行性。
- 对 API 进行版本控制时，可以考虑实施代理版本控制，其中使用 Facade 模式来支持版本，这样就能够维护在单个代码库中维护逻辑。
  - 通过 API Gateway，您能够使用 [请求和响应映射](#)，通过建立 Facade 模式为新字段提供默认值，或者从请求或响应中除去已删除的字段，从而简化接受合同变更的过程。通过这种方法，底层服务可以维护单个代码库。

## 资源

相关最佳实践：

- [REL03-BP01 选择如何划分工作负载](#)
- [REL03-BP02 构建专注于特定业务领域和功能的服务](#)

- [REL04-BP02 实施松耦合的依赖关系](#)
- [REL05-BP03 控制与限制重试调用](#)
- [REL05-BP05 设置客户端超时](#)

相关文档：

- [什么是 API \( 应用程序编程接口 \) ？](#)
- [在 AWS 上实施微服务](#)
- [微服务利弊权衡](#)
- [微服务 – 一个全新架构术语的定义](#)
- [AWS 上的微服务](#)
- [使用 OpenAPI 的 API Gateway 扩展](#)
- [OpenAPI 规范](#)
- [GraphQL：架构和类型](#)
- [Amazon EventBridge 代码绑定](#)

相关示例：

- [Amazon API Gateway：使用 OpenAPI 配置 REST API](#)
- [从 Amazon API Gateway 到使用 OpenAPI 的 Amazon DynamoDB CRUD 应用程序](#)
- [无服务器时代的现代化应用程序集成模式：API Gateway 服务集成](#)
- [通过 Amazon CloudFront 实施基于标头的 API Gateway 版本控制](#)
- [AWS AppSync：构建客户端应用程序](#)

相关视频：

- [在 AWS SAM 中使用 OpenAPI 来管理 API Gateway](#)

相关工具：

- [Amazon API Gateway](#)
- [AWS AppSync](#)
- [Amazon EventBridge](#)



## REL 4.您如何在分布式系统中设计交互以预防发生故障？

分布式系统依赖于通信网络实现组件（例如服务器或服务）的互联。尽管这些网络中存在数据丢失或延迟，但是您的工作负载必须可靠运行。分布式系统组件的运行方式不得对其他组件或工作负载产生负面影响。这些最佳实践能够预防故障，并改善平均故障间隔时间（MTBF）。

### 最佳实践

- [REL04-BP01 确定需要哪种类型的分布式系统](#)
- [REL04-BP02 实施松耦合的依赖关系](#)
- [REL04-BP03 持续工作](#)
- [REL04-BP04 使所有响应幂等](#)

### REL04-BP01 确定需要哪种类型的分布式系统

硬性实时分布式系统需要同步并快速地做出响应，而软性实时系统有更宽松的以分钟计的时间窗口，或更多响应。离线系统会对响应进行批处理或异步处理。硬性实时分布式系统具有最严格的可靠性要求。

硬性实时分布式系统 [要面对分布式系统中的](#) 最艰巨的挑战，又被称作请求/回复服务。因为收到请求的时间不可预测，而响应必须迅速（例如，客户正急切地等待响应）。此类示例包括，前端 Web 服务器、指令管道、信用卡交易，以及每个 AWS API 和语音通话。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

- 确定需要哪种类型的分布式系统。分布式系统要应对的挑战包括延迟、扩展、了解联网 API、对数据进行集结与解集，以及 Paxos 等算法的复杂性。随着系统规模扩大并呈现出更明显的分布态势，我们现在需要在日常生活中面对过去只存在于理论中的边缘情况。
- [Amazon Builders' Library：分布式系统相关挑战](#)
  - 硬性实时分布式系统需要快速的同步响应。
  - 软性实时系统则有更宽松的以分钟计的时间窗口，或更好响应。
  - 离线系统会对响应进行批处理或异步处理。
  - 硬性实时分布式系统具有最严格的可靠性要求。

### 资源

相关文档：

- [Amazon EC2：确保幂等性](#)
- [Amazon Builders' Library：分布式系统相关挑战](#)
- [Amazon Builders' Library：可靠性、持续工作和安然无忧](#)
- [什么是 Amazon EventBridge？](#)
- [什么是 Amazon Simple Queue Service？](#)

相关视频：

- [2019 年 AWS 纽约峰会：介绍事件驱动型架构和 Amazon EventBridge \( MAD205 \)](#)
- [AWS re:Invent 2018：闭环系统和开放思维：如何掌控不同规模的系统 \( ARC337 \) \( 包括松耦合、持续工作和静态稳定性 \)](#)
- [AWS re:Invent 2019：迁移到事件驱动型架构 \( SVS308 \)](#)

## REL04-BP02 实施松耦合的依赖关系

队列系统、流系统、工作流和负载均衡器等依赖关系是松耦合的。松耦合有助于隔离某个组件的行为与依赖于它的其他组件的行为，从而提升韧性和敏捷性。

在紧耦合的系统中，更改一个组件可能导致需要更改依赖该组件的其他组件，从而使所有组件的性能均降低。松耦合会打破这种依赖关系，使存在依赖关系的组件只需了解经过版本控制而且已发布的接口。在依赖项之间实施松耦合将隔离一个组件中的故障，防止对其他组件造成影响。

松耦合允许您修改代码或向组件添加功能，同时最大限度地降低依赖于该组件的其他组件的风险。它还允许在组件级别实现精细的韧性，您可以横向扩展或甚至改变依赖关系的底层实现。

要通过松耦合进一步提升韧性，在可能的情况下采用异步组件交互。若确定对请求进行注册已足够，则此模型适用于无需立即响应的任何交互。它包含一个生成事件的组件和另外一个使用事件的组件。两个组件不会通过直接点对点交互，但通常经由中间持久存储层集成，例如，Amazon SQS 队列或诸如 Amazon Kinesis 或 AWS Step Functions 流数据平台。

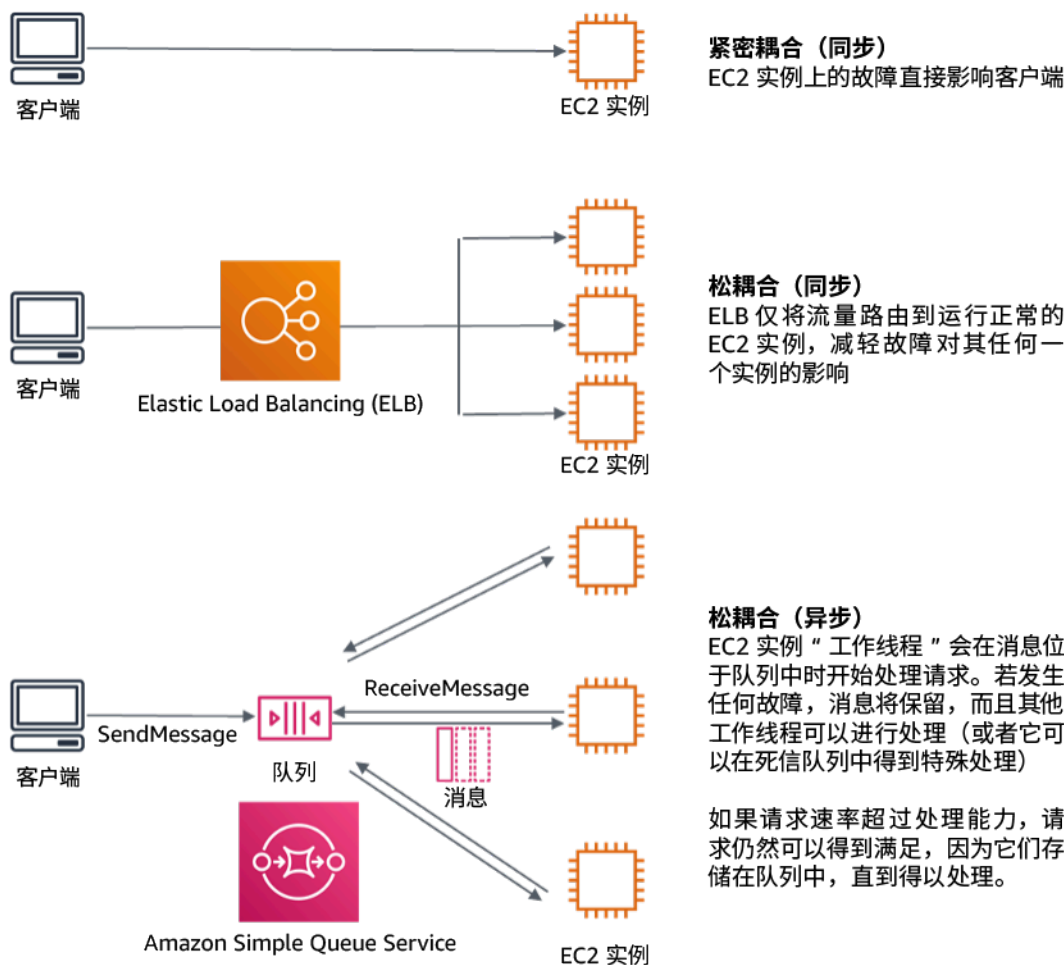


图 4：队列系统和负载均衡器等依赖关系是松耦合的。

Amazon SQS 队列和 Elastic Load Balancer 只是为松耦合增加中间层的两种方式。您还可以使用 Amazon EventBridge 在 AWS Cloud 中构建事件驱动型架构，而前者可从其依赖的服务（事件使用者）中提取客户端（事件产生器）。当您需要进行高吞吐量、基于推送的多对多消息收发时，Amazon Simple Notification Service（Amazon SNS）是可供选择的高效解决方案。通过 Amazon SNS 主题，您的发布者系统可以呈扇形将消息分发到大量订阅者端点以便进行并行处理。

虽然队列具有多项优点，但在大多数硬性实时系统中，早于阈值时间（通常为秒）的请求应被视为过时（客户端已放弃而且不再等待响应）而不被处理。因此，较新（而且可能依然有效）的请求会被处理。

**预期结果：**实现松耦合的依赖关系可以最大限度地减少组件级别的故障范围，这有助于诊断和解决问题。它还简化了开发周期，允许团队在模块级别实施更改，而不会影响依赖它的其他组件的性能。这种方法能够根据资源需求以及有助于提高成本效益的组件利用率，在组件层面进行扩展。

常见的反面模式：

- 部署整体式工作负载。
- 直接在工作负载层之间调用 API，不具备失效转移或异步处理请求的功能。
- 使用共享数据进行紧密耦合。松耦合的系统应避免通过共享数据库或其他形式的紧密耦合数据存储共享数据，这可能会重新引入紧耦合并阻碍可扩展性。
- 忽略背压。当组件无法以相同速度处理传入数据时，您的工作负载应该能够减慢或停止传入数据。

建立此最佳实践的好处：松耦合有助于隔离来自其他依赖于它的组件的行为，从而提升韧性和敏捷性。组件中的故障相互隔离。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

实施松耦合的依赖关系。可利用多种解决方案来构建松耦合的应用程序。其中包括用于实施全面托管的队列、自动化 workflows、对事件的反应以及 API 等服务，这些服务有助于将组件的行为相互隔离，从而提高韧性和敏捷性。

- 构建事件驱动型架构：[Amazon EventBridge](#) 有助于您构建松耦合的分布式事件驱动型架构。
- 在分布式系统中实现队列：可以使用 [Amazon Simple Queue Service \( Amazon SQS \)](#) 来集成和解耦分布式系统。
- 将组件容器化为微服务：利用[微服务](#)，团队可以构建由小型独立组件组成的应用程序，这些组件通过明确定义的 API 进行通信。[Amazon Elastic Container Service \( Amazon ECS \)](#) 和 [Amazon Elastic Kubernetes Service \( Amazon EKS \)](#) 可以帮助您更快地开始使用容器。
- 使用 Step Functions 管理工作流程：[Step Functions](#) 有助于您将多种 AWS 服务协调成灵活的工作流程。
- 利用发布-订阅 (pub/sub) 消息架构：[Amazon Simple Notification Service \( Amazon SNS \)](#) 提供从发布者到订阅者（也称为产生器和使用器）的消息传输。

### 实施步骤

- 事件驱动型架构中的组件由事件启动。事件是系统中发生的操作，例如用户将商品添加到购物车。操作成功后，将生成一个激活系统的下一个组件的事件。
  - [使用 Amazon EventBridge 构建事件驱动型应用程序](#)
  - [AWS re: Invent 2022 - 使用 Amazon EventBridge 设计事件驱动型集成](#)
- 分布式消息系统主要有三个部分，需要为基于队列的架构实现这些部分。它们包括分布式系统的组件、用于解耦的队列（分布在 Amazon SQS 服务器上）以及队列中的消息。典型的系统有将消息发

送到队列中的产生器和从队列接收消息的使用器。该队列将消息存储在多台 Amazon SQS 服务器上以实现冗余。

- [基本 Amazon SQS 架构](#)
- [使用 Amazon Simple Queue Service 在分布式应用程序之间发送消息](#)
- 由于松耦合的组件由独立团队管理，因此微服务如果得到充分利用，可以增强可维护性并提高可扩展性。它还允许在发生变化时将行为隔离到单个组件。
- [在 AWS 上实施微服务](#)
- [让我们来构造！使用容器构造微服务](#)
- 借助 AWS Step Functions，您可以构建分布式应用程序、实现流程自动化、编排微服务等。将多个组件编排到一个自动化工作流程中，让您可以解耦应用程序中的依赖关系。
- [使用 AWS Step Functions 和 AWS Lambda 创建无服务器工作流程](#)
- [开始使用 AWS Step Functions](#)

## 资源

相关文档：

- [Amazon EC2：确保幂等性](#)
- [Amazon Builders' Library：分布式系统相关挑战](#)
- [Amazon Builders' Library：可靠性、持续工作和安然无忧](#)
- [什么是 Amazon EventBridge？](#)
- [什么是 Amazon Simple Queue Service？](#)
- [拆分整体式应用程序](#)
- [使用 AWS Step Functions 和 Amazon SQS 协调基于队列的微服务](#)
- [基本 Amazon SQS 架构](#)
- [基于队列的架构](#)

相关视频：

- [2019 年 AWS 纽约峰会：介绍事件驱动型架构和 Amazon EventBridge \( MAD205 \)](#)
- [AWS re:Invent 2018：闭环系统和开放思维：如何掌控不同规模的系统 \( ARC337 \) \( 包括松耦合、持续工作和静态稳定性 \)](#)
- [AWS re:Invent 2019：迁移到事件驱动型架构 \( SVS308 \)](#)

- [AWS re:Invent 2019：使用 Amazon SQS 和 Lambda 的可扩展无服务器事件驱动型应用程序 \( API304 \)](#)
- [AWS re:Invent 2019：使用 Amazon SQS 和 Lambda 的可扩展无服务器事件驱动型应用程序](#)
- [AWS re: Invent 2022 - 使用 Amazon EventBridge 设计事件驱动型集成](#)
- [AWS re: Invent 2017：Elastic Load Balancing 深入探讨和最佳实践](#)

## REL04-BP03 持续工作

系统会在负载中存在剧烈快速更改时失败。例如，如果您的工作负载执行的一项运行状况检查监控着数千个服务器的运行状况，每次都应发送相同大小的有效负载（当前状态的完整快照）。无论是否有服务器或有多少服务器发生故障，运行状况检查系统都会持续工作，而不会有剧烈、快速的变动。

例如，如果运行状况检查系统正在监控 100000 个服务器，它的标称负载低于通常而言较低的服务器故障率。但如果发生重大事件使一半的服务器运行不正常，则运行状况检查系统会因为尝试更新通知系统以及向其客户端传送状态而变得不堪重负。因此，运行状况检查系统每次应发送当前状态的完整快照。100000 个服务器的运行状况，若每个以一个比特代表，则仅需要 12.5-KB 有效负载。无论是没有服务器发生故障还是所有服务器都发生故障，运行状况检查系统都会持续工作，而大幅度骤变也不会威胁到系统的稳定性。这实际上是 Amazon Route 53 处理端点（例如 IP 地址）的运行状况检查来确定最终用户如何路由到这些端点的方式。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

- 持续工作，使系统不会在负载出现骤变时失败。
- 实施松耦合的依赖关系。队列系统、流系统、工作流和负载均衡器等依赖关系是松耦合的。松耦合有助于隔离某个组件的行为与依赖于它的其他组件的行为，从而提升弹性和敏捷性。
- [Amazon Builders' Library：可靠性、持续工作和安然无忧](#)
- [AWS re:Invent 2018：闭环系统和开放思维：如何掌控不同规模的系统 \( ARC337 \) \( 包括持续工作 \)](#)
  - 例如，如果运行状况检查系统正在监控 10 万台服务器工程设计工作负载，不论成功或失败的次数，有效负载大小均能保持稳定。

## 资源

相关文档：



- [Amazon EC2：确保幂等性](#)
- [Amazon Builders' Library：分布式系统相关挑战](#)
- [Amazon Builders' Library：可靠性、持续工作和安然无忧](#)

相关视频：

- [2019 年 AWS 纽约峰会：介绍事件驱动型架构和 Amazon EventBridge \( MAD205 \)](#)
- [AWS re:Invent 2018：闭环系统和开放思维：如何掌控不同规模的系统 \( ARC337 \) \( 包括持续工作 \)](#)
- [AWS re:Invent 2018：闭环系统和开放思维：如何掌控不同规模的系统 \( ARC337 \) \( 包括松耦合、持续工作和静态稳定性 \)](#)
- [AWS re:Invent 2019：迁移到事件驱动型架构 \( SVS308 \)](#)

#### REL04-BP04 使所有响应幂等

幂等服务承诺每个请求只完成一次，因此发起多个相同请求与进行单个请求的效果相同。幂等服务使客户端可以轻松进行重试，而不必担心错误地处理多次。要执行此操作，客户端可以发出具有幂等性令牌的 API 请求，每当重复请求时都会使用同一令牌。幂等服务 API 使用令牌返回响应，该响应与首次完成请求时返回的响应相同。

在分布式系统中，至多（客户端仅发起一个请求）或至少（持续发起请求直到客户端收到成功确认）执行某项操作一次并不难。难就难在要保证某项操作具有幂等性，亦即它被恰好执行一次，从而使发起多个相同的请求与发起一个请求的效果相同。在 API 中使用幂等性令牌，服务可以一次或多次收到变异请求，而不会创建重复记录或产生副作用。

未建立这种最佳实践的情况下暴露的风险等级：中

#### 实施指导

- 使所有响应幂等。幂等服务承诺每个请求只完成一次，因此发起多个相同请求与进行单个请求的效果相同。
  - 客户端可以发出具有幂等性令牌的 API 请求，每当重复请求时都会使用同一令牌。幂等服务 API 使用令牌返回响应，该响应与首次完成请求时返回的响应相同。
  - [Amazon EC2：确保幂等性](#)



## 资源

### 相关文档：

- [Amazon EC2：确保幂等性](#)
- [Amazon Builders' Library：分布式系统相关挑战](#)
- [Amazon Builders' Library：可靠性、持续工作和安然无忧](#)

### 相关视频：

- [2019 年 AWS 纽约峰会：介绍事件驱动型架构和 Amazon EventBridge \( MAD205 \)](#)
- [AWS re:Invent 2018：闭环系统和开放思维：如何掌控不同规模的系统 \( ARC337 \) \( 包括松耦合、持续工作和静态稳定性 \)](#)
- [AWS re:Invent 2019：迁移到事件驱动型架构 \( SVS308 \)](#)

## REL 5.您如何在分布式系统中进行交互设计，从而缓解或经受住故障影响？

分布式系统依赖于通信网络实现组件（例如服务器或服务）的互联。尽管这些网络中存在数据丢失或延迟，但是您的工作负载必须可靠运行。分布式系统组件的运行方式不得对其他组件或工作负载产生负面影响。这些最佳实践使工作负载能够承受压力或故障，从中更快地恢复，并且降低此类损坏的影响。其结果是缩短平均恢复时间（MTTR）。

### 最佳实践

- [REL05-BP01 实施轻松降级以将适用的硬依赖关系转换为软依赖关系](#)
- [REL05-BP02 限制请求](#)
- [REL05-BP03 控制与限制重试调用](#)
- [REL05-BP04 快速失效机制和限制队列](#)
- [REL05-BP05 设置客户端超时](#)
- [REL05-BP06 尽可能使服务为无状态](#)
- [REL05-BP07 实施紧急杠杆](#)

### REL05-BP01 实施轻松降级以将适用的硬依赖关系转换为软依赖关系

即使依赖项不可用，应用程序组件也应继续执行其核心功能。应用程序组件可以提供稍微陈旧的数据、替代数据，甚至没有数据。这可确保在提供核心业务价值的同时，将局部故障对整体系统功能造成的障碍减至最少。

期望的结果：某个组件的依赖项运行不正常时，该组件仍可在性能降低的条件下运行。组件的故障模式应视为正常运行。工作流在设计时，应确保此类故障不会导致完全失败，或者至少实现可预测和可恢复的状态。

常见反模式：

- 未确定所需的核心业务功能。即使在依赖项故障期间也不测试组件是否正常运行。
- 不论是出错时，还是当多个依赖项中只有一个不可用且仍可以返回部分结果时，不提供任何数据。
- 在事务部分失败时造成不一致的状态。
- 没有替代方法用于访问中央 Parameter Store。
- 在刷新失败时，使本地状态失效或清空，而没有考虑这样做的后果。

建立此最佳实践的好处：轻松降级可以提高整个系统的可用性，即使在故障期间也能保持最重要功能的功能。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

实施轻松降级有助于最大限度地减少依赖项故障对组件功能的影响。理想情况下，组件检测依赖项故障，并以对其他组件或客户影响最小的方式解决这些故障。

为轻松降级设计架构意味着在依赖项设计期间，需要考虑潜在的故障模式。对于每种故障模式，都要有办法向调用方或客户提供组件的大部分功能，或者至少提供最关键的功能。这些注意事项可以作为额外的要求进行测试和验证。理想情况下，即使一个或多个依赖项出现故障，一个组件也能够以可接受的方式执行其核心功能。

这既是商业议题，也是技术议题。所有业务要求都很重要，应尽可能满足。但是，确定在无法满足所有要求时会出现什么情况，这种做法同样很有意义。系统可以设计为具备可用性和一致性，但是如果必须放弃一个要求，那么哪个要求更重要？对于付款处理而言，可能一致性更重要。对于实时应用程序，可能可用性更重要。对于面向客户的网站，这个回答可能取决于客户的期望值。

其具体的意义取决于组件的要求，以及将什么功能视为其核心功能。例如：

- 在登录页面上，电子商务网站可能会显示来自多个不同系统的数据，例如个性化推荐、最热销的产品和客户订单状态。当一个上游系统出现故障时，合理的做法是显示其余的所有信息，而不是向客户显示一个错误页面。

- 对于执行批量写入的组件，如果某个单独的操作失败，它仍应继续执行批处理。实施重试机制应该很简单。要实施重试机制，可以向调用方返回哪些操作成功、哪些操作失败的信息，或者将失败的请求放入死信队列中来实施异步重试。同时还应记录有关失败操作的信息。
- 处理事务的系统必须确保，要么执行了所有更新，要么未执行任何更新。对于分布式事务，在同一个事务后面的操作失败时，可以使用 Segal 模式来回滚先前的操作。这里的核心功能是保持一致性。
- 时间关键型系统应能够处理未及时响应的依赖项。在这些情况下，可以使用断路器模式。当来自依赖项的响应开始超时，系统可以切换为关闭状态，不进行额外的调用。
- 应用程序可以从 Parameter Store 中读取参数。创建具有默认参数集的容器镜像，并在 Parameter Store 不可用时使用这些镜像，这种做法会很有用。

请注意，需要对组件出现故障时所采取的途径进行测试，而且这一途径应该比主要途径简单得多。通常，[应避免使用回退策略](#)。

## 实施步骤

确定外部和内部依赖项。考虑它们可能出现什么样的故障。思考在故障期间，尽可能减少对上游和下游系统以及客户的负面影响的方法。

以下是依赖项列表以及在依赖项故障时如何轻松降级：

1. 依赖项部分故障：一个组件可以向下游系统发出多个请求，这可以是向一个系统发出多个请求，也可以是向多个系统发出一个请求。根据具体的业务环境，可能需要采用不同的处理方式（有关更多详细信息，请参阅实施指南前文中的示例）。
2. 由于高负载，下游系统无法处理请求：如果发送到下游系统的请求持续失败，则继续重试就毫无意义。这可能会对已经过载的系统造成额外负载，使得系统更难于恢复。这时可以使用断路器模式，该模式监视对下游系统的失败调用。如果大量调用失败，它会停止向下游系统发送更多请求，仅不定期让调用通过，以测试下游系统是否再次可用。
3. Parameter Store 不可用：要转换 Parameter Store，可以使用容器或机器镜像中包含的软依赖项缓存或合理默认值。请注意，这些默认值需要保持为最新并包含在测试套件中。
4. 监控服务或其他非功能性依赖项不可用：如果某个组件间歇性地无法将日志、指标或跟踪发送到中央监控服务，通常最好还是正常执行业务功能。长时间静默地不进行日志记录或不推送指标通常不可接受。此外，某些使用场景可能需要完整的审计条目才能满足合规性要求。
5. 关系数据库的主实例可能不可用：与几乎所有关系数据库一样，Amazon Relational Database Service 只能有一个主写入器实例。对于写入工作负载，这会造成单点故障，并增加扩缩的难度。使用多可用区配置来实现高可用性，或者使用 Amazon Aurora 无服务器架构来实现更好的扩展能力，可以部分缓解这种情况。对于非常高的可用性要求，完全不依赖于主写入器是有意义的。对于只读

查询，可以使用只读副本，这提供了冗余和横向扩展能力，而不仅仅是纵向扩展。对写入操作可以进行缓冲，例如缓冲在 Amazon Simple Queue Service 队列中，这样即使主写入器暂时不可用，仍然可以接受来自客户的写入请求。

## 资源

### 相关文档：

- [Amazon API Gateway：限制 API 请求以提高吞吐量](#)
- [CircuitBreaker \(对《发布它！》一书中的“断路器”部分进行的总结\)](#)
- [AWS 中的错误重试和指数回退](#)
- [Michael Nygard 《发布它！设计和部署生产就绪的软件》 \(Release It! Design and Deploy Production-Ready Software\)](#)
- [Amazon Builders' Library：避免在分布式系统中回退](#)
- [Amazon Builders' Library：避免无法克服的队列积压](#)
- [Amazon Builders' Library：缓存挑战和策略](#)
- [Amazon Builders' Library：为超时、重试和回退引入抖动](#)

### 相关视频：

- [重试、回退和抖动：AWS re:Invent 2019：介绍 Amazon Builders' Library \(DOP328\)](#)

### 相关示例：

- [Well-Architected 实验室：第 300 级：实施运行状况检查和管理依赖项以提高可靠性](#)

## REL05-BP02 限制请求

限制请求，以防需求意外增加而导致的资源耗尽的情况。系统将处理未超过限制速率的请求，而超过所定义限制的请求将被拒绝，并返回一条消息，指出请求已受限制。

期望的结果：使用请求限制可以缓解客户流量突增、泛洪攻击或重试风暴所造成的大量容量峰值情况，让工作负载能够继续正常处理支持的请求量。

### 常见反模式：

- 未实施 API 端点限制，或者未考虑预期容量即保留默认值。

- API 端点未经过负载测试，也未测试节流限制。
- 限制请求速率而未考虑请求大小或复杂性。
- 测试最大请求速率或最大请求大小，但未同时测试两者。
- 资源预置的限制与测试中确定的限制不同。
- 尚未为应用程序到应用程序的 ( A2A ) API 使用者配置或考虑使用量计划。
- 横向扩展的队列使用者没有配置最大并发设置。
- 没有基于每个 IP 地址实施速率限制。

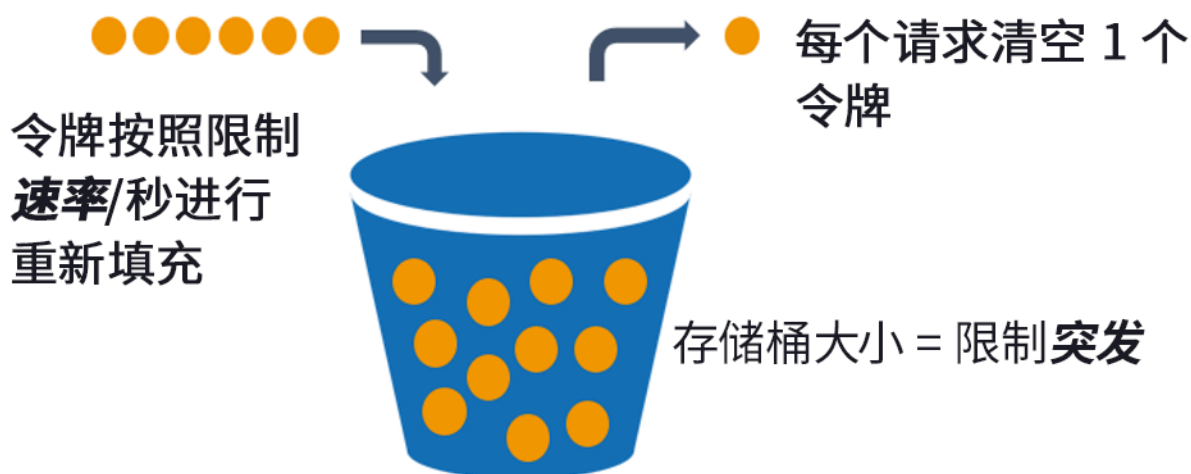
建立此最佳实践的好处：在遇到意外的容量峰值时，设置了节流限制的工作负载能够正常运行，并成功处理已接受的请求负载。API 和队列上突然或持续出现的请求峰值会受到限制，不会耗尽请求处理资源。速率限制会限制单独的请求者，这样来自单个 IP 地址或 API 使用者的大量流量就不会耗尽资源，从而不会影响到其他使用者。

未建立这种最佳实践的情况下暴露的风险等级：高

#### 实施指导

服务应设计为处理已知的请求容量；这种容量可以通过负载测试来确立。如果请求到达速率超过限制，则会发出相应的响应，表示请求已被限制。这让使用者可以处理错误并稍后重试。

当您的服务需要实施节流时，可以考虑实施令牌存储桶算法，每个令牌对应于一个请求。令牌按照每秒的限制速率重新填充，并按照每个请求一个令牌的模式异步清空。



令牌存储桶算法。

[Amazon API Gateway](#) 根据账户和区域限制实施令牌存储桶算法，可通过使用量计划为每个客户端配置。此外，[Amazon Simple Queue Service \( Amazon SQS \)](#) 和 [Amazon Kinesis](#) 可以缓冲请求以稳定请求速率，并在有足够的请求处理能力时实现更高的限制速率。最后，您可以通过 [AWS WAF](#) 实施速率限制，以限制产生过高负载的特定 API 使用者。

## 实施步骤

您可以为 API 配置具有节流限制的 API Gateway，并在超出限制时返回“429 #####”错误。您可以将 AWS WAF 与 AWS AppSync 和 API Gateway 端点结合使用，根据各个 IP 地址来启用速率限制。此外，如果您的系统能够接受异步处理，则可以将消息放入队列或流中，以加快对服务客户端的响应，这样您便可以突增到更高的限制速率。

使用异步处理，当您配置 Amazon SQS 作为 AWS Lambda 的事件源时，您可以 [配置最大并发度](#)，以避免高速率的事件消耗账户中可用的并发执行配额，因为您的工作负载或账户中的其他服务会需要这些配额。

虽然 API Gateway 提供了令牌存储桶的托管实施，但在无法使用 API Gateway 的情况下，您可以针对服务利用具体语言的令牌存储桶开源实施（参见“资源”中的相关示例）。

- 了解和配置 [API Gateway 节流限制](#)：对于区域在账户级别，对于阶段在 API 级别，而对于使用量计划级别则为 API 密钥。
- 应用 [AWS WAF 速率限制规则](#) 到 API Gateway 和 AWS AppSync 端点，以阻止泛洪攻击并屏蔽恶意 IP。对于 A2A 使用者，也可以在 AWS AppSync API 密钥上配置速率限制规则。
- 对于 AWS AppSync API，请考虑您需要的节流控制是否超过了速率限制，如果超过，请在 AWS AppSync 端点前面配置 API Gateway。
- 将 Amazon SQS 队列设置为 Lambda 队列使用者的触发器时，请将 [最大并发度](#) 设置为一个值，使其处理量足以满足服务等级目标，同时使用量又不会超过并发度限制而影响其他 Lambda 函数。当您通过 Lambda 使用队列时，请考虑为相同账户和区域中的其他 Lambda 函数设置预留并发度。
- 将 API Gateway 与 Amazon SQS 或 Kinesis 的原生服务集成结合使用，以缓冲请求。
- 如果您无法使用 API Gateway，请查看具体语言的库，以便为工作负载实施令牌存储桶算法。查看示例部分，然后自行研究以查找合适的库。
- 对您计划设置的限制或者您计划允许增加的限制进行测试，并记录测试后的限制值。
- 不要将限制值提高到超出您在测试中确立的限制值。增加限制时，请先确认预置的资源是否已经等于或大于测试场景中预置的资源，然后再进行增加。

## 资源

相关最佳实践：

- [REL04-BP03 持续工作](#)
- [REL05-BP03 控制与限制重试调用](#)

相关文档：

- [Amazon API Gateway：限制 API 请求以提高吞吐量](#)
- [AWS WAF：基于速率的规则语句](#)
- [引入了使用 Amazon SQS 作为事件源时 AWS Lambda 的最大并发度](#)
- [AWS Lambda：最大并发度](#)

相关示例：

- [三条最重要的基于 AWS WAF 速率的规则](#)
- [Java Bucket4j](#)
- [Python token-bucket](#)
- [Node token-bucket](#)
- [.NET 系统线程速率限制](#)

相关视频：

- [使用 AWS AppSync 实施 GraphQL API 安全最佳实践](#)

相关工具：

- [Amazon API Gateway](#)
- [AWS AppSync](#)
- [Amazon SQS](#)
- [Amazon Kinesis](#)
- [AWS WAF](#)



## REL05-BP03 控制与限制重试调用

使用指数回退来重试请求，每次重试之间的间隔会逐渐延长。在两次重试之间引入抖动，以随机调整重试间隔。限制最大重试次数。

期望的结果：分布式软件系统中的常见组件包括服务器、负载均衡器、数据库和 DNS 服务器。在正常运行期间，这些组件对请求的响应可能是临时错误或者受限制错误，也能是无论如何重试都会持续存在的错误。当客户端向服务发出请求时，请求会消耗资源，包括内存、线程、连接、端口或任何其他有限的资源。控制和限制重试策略用于释放资源并最大限度地减少资源消耗，这样就可以避免承受压力的系统组件不堪重负。

当客户端请求超时或收到错误响应时，客户端应决定是否重试。如果进行重试，则会按照最大重试次数值，采用指数回退和抖动方法进行重试。因此，后端服务和进程可以缓解负载并缩短自我修复时间，从而加快恢复速度并成功处理服务请求。

常见反模式：

- 实施重试，但没有添加指数回退、抖动和最大重试次数值。指数回退和抖动有助于避免因无意间按照相同间隔协调进行重试，从而导致的人为流量峰值。
- 实施重试但没有测试重试的效果，或者假设 SDK 中已经内置了重试而不测试重试场景。
- 无法理解依赖项发布的错误代码，导致重试所有错误，包括那些有明确原因的错误，这些错误指出缺乏权限、配置错误或其他预计需要手动干预才能解决的情况。
- 没有解决可观测性实践，包括监控反复出现的服务故障并发出警报，以便了解和解决潜在问题。
- 在内置或第三方重试功能便已足够时，开发自定义重试机制。
- 在应用程序堆栈的多层进行重试，而重试方法导致重试尝试复杂化，进一步加剧了重试风暴中的资源消耗。一定要了解这些错误对您的应用程序以及所依赖的依赖项有何影响，然后仅在一个级别实施重试。
- 重试非幂等的服务调用，导致意外的副作用，例如重复的结果。

建立此最佳实践的好处：重试有助于客户端在请求失败时获得预期的结果，但也会消耗更多的服务器时间来获取所需的成功响应。当故障比率很低或者是临时性故障时，重试效果很好。当故障是由资源过载导致时，重试会导致情况进一步恶化。通过在客户端重试中添加指数回退和抖动，服务器可以从因资源过载导致的故障中恢复。抖动可避免请求同时出现造成峰值，指数回退可以减少因在正常请求负载中增添重试而导致的负载上升。最后，务必要配置最大重试次数或用时，以避免由于造成积压而导致亚健康故障。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

控制与限制重试调用。在逐渐延长的间隔以后使用指数回退进行重试。引入抖动以使重试间隔随机化，并限制重试的最大次数。

默认情况下，一些 AWS SDK 实施重试和指数回退。在适用于您的工作负载的情况下，使用这些内置 AWS 实施。在调用幂等性服务时，以及重试可以提高客户端可用性时，在工作负载中实施类似的逻辑。根据您的使用场景确定超时以及何时停止重试。为重试使用场景构建和演练测试场景。

## 实施步骤

- 对于您的应用程序所依赖的服务，确定应用程序堆栈中最适合实施重试的层。
- 请注意，现有的 SDK 会针对您选择的语言，实施采用了指数回退和抖动方法的成熟重试策略，相比您自己编写重试实施，使用这些实施方法会更好。
- 确认 [服务具有幂等性](#)，然后再实施重试。实施重试后，请确保在生产环境中进行测试，并定期进行演练。
- 调用 AWS 服务 API 时，使用 [AWS SDK](#) 和 [AWS CLI](#) 并了解重试配置选项。确定默认值是否适用于您的使用场景，进行测试，并根据需要进行调整。

## 资源

相关最佳实践：

- [REL04-BP04 使所有响应幂等](#)
- [REL05-BP02 限制请求](#)
- [REL05-BP04 快速失效机制和限制队列](#)
- [REL05-BP05 设置客户端超时](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)

相关文档：

- [AWS 中的错误重试和指数回退](#)
- [Amazon Builders' Library：为超时、重试和回退引入抖动](#)
- [指数回退和抖动](#)
- [使用幂等 API 确保重试安全](#)

相关示例：

- [Spring 重试](#)
- [Resilience4j 重试](#)

相关视频：

- [重试、回退和抖动：AWS re:Invent 2019：介绍 Amazon Builders' Library \( DOP328 \)](#)

相关工具：

- [AWS SDK 和工具：重试行为](#)
- [AWS Command Line Interface：AWS CLI 重试](#)

## REL05-BP04 快速失效机制和限制队列

当服务无法成功响应请求时，可采用快速失效机制。这样可释放与请求关联的资源，并允许该服务在资源不足的情况下进行恢复。快速失效机制一种成熟的软件设计模式，可用于在云端构建高度可靠的工作负载。队列也是一种成熟的企业集成模式，其能够实现平稳的负载，并在能够容忍异步处理的情况下，使得客户端能够释放资源。如果某个服务在正常条件下能够成功响应，但在请求速率过高时失败，请使用队列来缓冲请求。不过，不要允许出现较长的队列积压，否则可能导致处理已被客户端放弃的过时请求。

期望的结果：当系统遇到资源争用、超时、异常或灰色故障等情况，导致无法实现服务等级目标时，快速失效机制策略可以加快系统恢复速度。如果系统必须承受流量峰值并能够适应异步处理，就可以使用队列来缓冲对后端服务的请求，让客户端可以快速释放请求，从而提高可靠性。在将请求缓冲到队列中时，系统将实施队列管理策略，来避免出现无法克服的积压。

常见反模式：

- 实施消息队列，但不配置死信队列 ( DLQ ) 或 DLQ 数量警报来检测系统出现故障的时间。
- 不测量队列中消息的时限 ( 关于延迟的度量 )，来了解队列使用者何时落后或者出错并导致重试。
- 当业务不需要再存在时，处理积压的消息没有任何价值，但不从队列中清除这些消息。
- 当后进先出 ( LIFO ) 队列可以更好地满足客户端需求时，配置先进先出 ( FIFO ) 队列，例如，在不需要严格排序并且处理积压内容会延误所有新的和注重时效性的请求时，先进先出队列会导致所有客户端出现违反服务等级协议的情况。
- 向客户端公开内部队列，而不是公开那些管理工作摄入并将请求放入内部队列的 API。

- 将过多的工作请求类型合并到一个队列中，这会导致在一个队列中分配对多种请求类型的资源需求，进而会加剧积压情况。
- 在同一个队列中处理复杂请求和简单请求，但这些请求具有不同的监控、超时和资源分配需求。
- 不验证输入，也不使用断言在软件中实施快速失效机制，这些机制可将异常上报到更高级别的组件来轻松处理错误。
- 不从请求路由中移除出现故障的资源，尤其是在由于崩溃和重启、间歇性依赖项故障、容量减少或网络数据包丢失，导致同时出现成功和失败的灰色故障时。

建立此最佳实践的好处：采用快速失效机制的系统更容易调试和修复，通常在将版本发布到生产环境之前，在编码和配置阶段就会暴露问题。采用有效排队策略的系统在面对流量高峰和间歇性系统故障的情况时，能够提供更出色的韧性和可靠性。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

快速失效机制策略能够通过编码形式构建到软件解决方案中，也能够基础设施中配置。除了快速失效机制之外，还可通过队列这种简单而强大的架构技术，来解耦系统组件，实现平稳的负载。[Amazon CloudWatch](#) 提供监控故障和发出警报的功能。在确定系统出现故障时，可以调用缓解策略，包括从受损资源进行故障转移。当系统使用 [Amazon SQS](#) 和其他队列技术实施队列来实现平稳的负载时，须考虑如何管理队列积压以及消息使用故障。

### 实施步骤

- 在软件中实施编程式断言或特定指标，并将其用于明确发出关于系统问题的警报。[Amazon CloudWatch](#) 有助于根据应用程序日志模式和 SDK 工具创建指标和警报。
- 使用 [CloudWatch](#) 指标和警报从受损资源进行故障转移，这些受损资源会增加处理延迟或者在处理请求时反复失败。
- 要使用异步处理，您可以设计 API 来接受请求，并使用 [Amazon SQS](#) 将请求附加到内部队列，然后向生成消息的客户端发送成功消息，这样客户端就可以释放资源，并继续处理其他工作，同时后端队列使用者可以处理请求。
- 在每次从队列中删除消息时，通过将现在的时间戳与消息时间戳进行比较来生成 [CloudWatch](#) 指标，从而测量和监控队列处理延迟。
- 如果故障导致无法成功处理消息，或者有大量的流量高峰无法按照服务等级协议的要求处理，则将较旧或过多的流量转到溢出队列。这样便可以优先处理新的工作，在有可用容量时再处理较早的工作。这种技术与 LIFO 处理有相似之处，使得可以对所有新工作进行正常的系统处理。
- 使用死信或再驱动队列，将无法处理的消息从积压中移至另一个位置，供以后研究和解决

- 您可以重试，或者在允许的情况下，通过将现在的时间戳与消息时间戳进行比较，丢弃与发出请求的客户端不再相关的消息，以此来删除旧消息。

## 资源

### 相关最佳实践：

- [REL04-BP02 实施松耦合的依赖关系](#)
- [REL05-BP02 限制请求](#)
- [REL05-BP03 控制与限制重试调用](#)
- [REL06-BP02 定义与计算指标（聚合）](#)
- [REL06-BP07 对系统中的请求进行端到端跟踪监控](#)

### 相关文档：

- [避免无法克服的队列积压](#)
- [快速失效机制](#)
- [如何防止我的 Amazon SQS 队列中积压的消息不断增加？](#)
- [Elastic Load Balancing：可用区转移](#)
- [Amazon Route 53 应用程序恢复控制器：用于流量失效转移的路由控制](#)

### 相关示例：

- [企业集成模式：死信通道](#)

### 相关视频：

- [AWS re:Invent 2022 – 运行高可用性多可用区应用程序](#)

### 相关工具：

- [Amazon SQS](#)
- [Amazon MQ](#)
- [AWS IoT Core](#)
- [Amazon CloudWatch](#)

## REL05-BP05 设置客户端超时

您应适当设置连接和请求的超时，对其进行系统性验证，不要依赖默认值，因为默认值并不了解具体的工作负载情况。

期望的结果：客户端超时应考虑当完成请求需要超长时间时，与等待请求相关的客户端、服务器和工作负载成本。由于无法知晓任何超时的确切原因，因此客户端必须运用对服务的了解，预测可能的原因和相应的超时时间

根据配置的值，客户端连接超时。遇到超时后，客户端决定回退并重试，或者打开 [断路器](#)。这些模式可避免发出会加剧底层错误状况的请求。

常见反模式：

- 不了解系统超时或默认超时。
- 不了解正常的请求完成时间。
- 不了解完成请求需要超长时间的可能原因，也不了解与等待完成这些请求相关的客户端、服务器或工作负载性能成本。
- 不了解网络受损只要在达到超时后就可能会导致请求失败，也不了解未采用更短的超时时间而招致的客户端和工作负载性能成本。
- 未针对连接和请求来测试超时场景。
- 将超时设置得过高，这会导致等待时间过长并增加资源使用。
- 将超时设置得过低，会导致人为故障。
- 忽略处理远程调用超时错误的模式，例如断路器和重试。
- 不考虑监控服务调用错误率、延迟的服务等级目标和延迟异常值。这些指标能够提供关于过长或不合理超时的洞察信息

建立此最佳实践的好处：配置了远程调用超时，且系统在设计上可以轻松处理超时，这样在远程调用响应异常缓慢时能够节省资源，且服务客户端可以轻松处理超时错误。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

针对所有服务依赖项调用以及一般情况下的所有跨流程调用，设置连接超时和请求超时。许多框架具有内置超时功能，但仍需谨慎，因为一些超时的默认值为无限值，或者高于您的服务目标可以接受的值。过高的值会降低超时的实用性，因为客户端等待超时发生时，系统会继续消耗资源。过低的值可能会重

试请求过多次，因而导致后端流量增加以及延迟变长。在有些情况下，由于要对全部请求进行重试，从而可能导致完全中断。

在确定超时策略时，请考虑以下几点：

- 由于请求的内容、目标服务受损或网络分区故障，处理请求所需的时间可能比正常时间要长。
- 请求如果具有成本异常高的内容，就可能会消耗不必要的服务器和客户端资源。在这种情况下，让这些请求超时而不进行重试可以节省资源。服务还应利用限制和服务器端超时，来保护自身免受成本异常高的内容的侵害。
- 如果由于服务受损而导致请求用时超长，则可以使请求超时并进行重试。请求和重试的服务成本需要考虑在内，但如果原因是局部受损，则重试的成本可能不会太高，而且会减少客户端资源消耗。根据损害的性质，超时还可能会释放服务器资源。
- 如果由于网络未能传输请求或响应，导致请求完成用时很长，则可以使请求超时并进行重试。由于未能传输请求或响应，因此无论超时时间多长，结果都是失败。在这种情况下，超时不会释放服务器资源，但可以释放客户端资源并提高工作负载性能。

利用重试和断路器等成熟的设计模式，来轻松地处理超时并支持快速失效机制方法。[AWS SDK](#) 和 [AWS CLI](#) 可用于配置连接和请求超时，以及使用指数回退和抖动进行重试。[AWS Lambda](#) 函数支持配置超时，通过 [AWS Step Functions](#)，您可以利用与 AWS 服务和 SDK 的预先构建集成，以低代码方式构建断路器。[AWS App Mesh](#) Envoy 提供超时和断路器功能。

## 实施步骤

- 配置远程服务调用的超时，并利用内置语言超时功能或开源超时库。
- 当您的工作负载使用 AWS SDK 进行调用时，请查看文档以了解具体语言的超时配置。
  - [Python](#)
  - [PHP](#)
  - [.NET](#)
  - [Ruby](#)
  - [Java](#)
  - [Go](#)
  - [Node.js](#)
  - [C++](#)
- 在工作负载中使用 AWS SDK 或 AWS CLI 命令时，可通过设置 AWS [配置默认值](#)，为 `connectTimeoutInMillis` 和 `tlsNegotiationTimeoutInMillis` 配置默认超时值。



- 应用 [命令行选项](#) `cli-connect-timeout` 和 `cli-read-timeout`，控制向 AWS 服务发出一次性 AWS CLI 命令。
- 监控远程服务调用的超时，并针对持续性错误设置警报，这样您就能够主动处理错误情况。
- 对调用错误率、延迟的服务等级目标和延迟异常值实施 [CloudWatch Metrics](#) 和 [CloudWatch 异常检测](#)，有助于深入了解如何管理过长或不合理的超时。
- 在 [Lambda 函数](#) 上配置超时。
- 处理超时的时候，API Gateway 客户端必须实施自己的重试。对于下游集成，API Gateway 支持 [50 毫秒到 29 秒的集成超时](#)，而且在集成请求超时不会重试。
- 实施 [断路器](#) 模式，以避免在超时进行远程调用。打开断路器以避免失败调用，当调用响应正常时关闭断路器。
- 对于基于容器的工作负载，请查看 [App Mesh Envoy](#) 功能，以利用内置的超时和断路器。
- 使用 AWS Step Functions，以低代码方式为远程服务调用构建断路器，尤其是在调用 AWS 原生 SDK 和所支持的 Step Functions 集成时，以简化工作负载。

## 资源

### 相关最佳实践：

- [REL05-BP03 控制与限制重试调用](#)
- [REL05-BP04 快速失效机制和限制队列](#)
- [REL06-BP07 对系统中的请求进行端到端跟踪监控](#)

### 相关文档：

- [AWS SDK：重试和超时](#)
- [Amazon Builders' Library：为超时、重试和回退引入抖动](#)
- [Amazon API Gateway 配额和重要注意事项](#)
- [AWS Command Line Interface：命令行选项](#)
- [AWS SDK for Java 2.x：配置 API 超时](#)
- [AWS Botocore 使用配置对象和配置引用](#)
- [适用于 .NET 的 AWS SDK：重试和超时](#)
- [AWS Lambda：配置 Lambda 函数选项](#)

相关示例：

- [将断路器模式与 AWS Step Functions 和 Amazon DynamoDB 配合使用](#)
- [Martin Fowler：断路器](#)

相关工具：

- [AWS SDK](#)
- [AWS Lambda](#)
- [Amazon SQS](#)
- [AWS Step Functions](#)
- [AWS Command Line Interface](#)

REL05-BP06 尽可能使服务为无状态

服务应该不需要状态，或者在不同的客户端请求之间卸载状态，磁盘上和内存中本地存储的数据不存在依赖关系。这使任意替换服务器成为可能，而且不会对可用性产生影响。Amazon ElastiCache 或 Amazon DynamoDB 是卸载状态的理想目标位置。

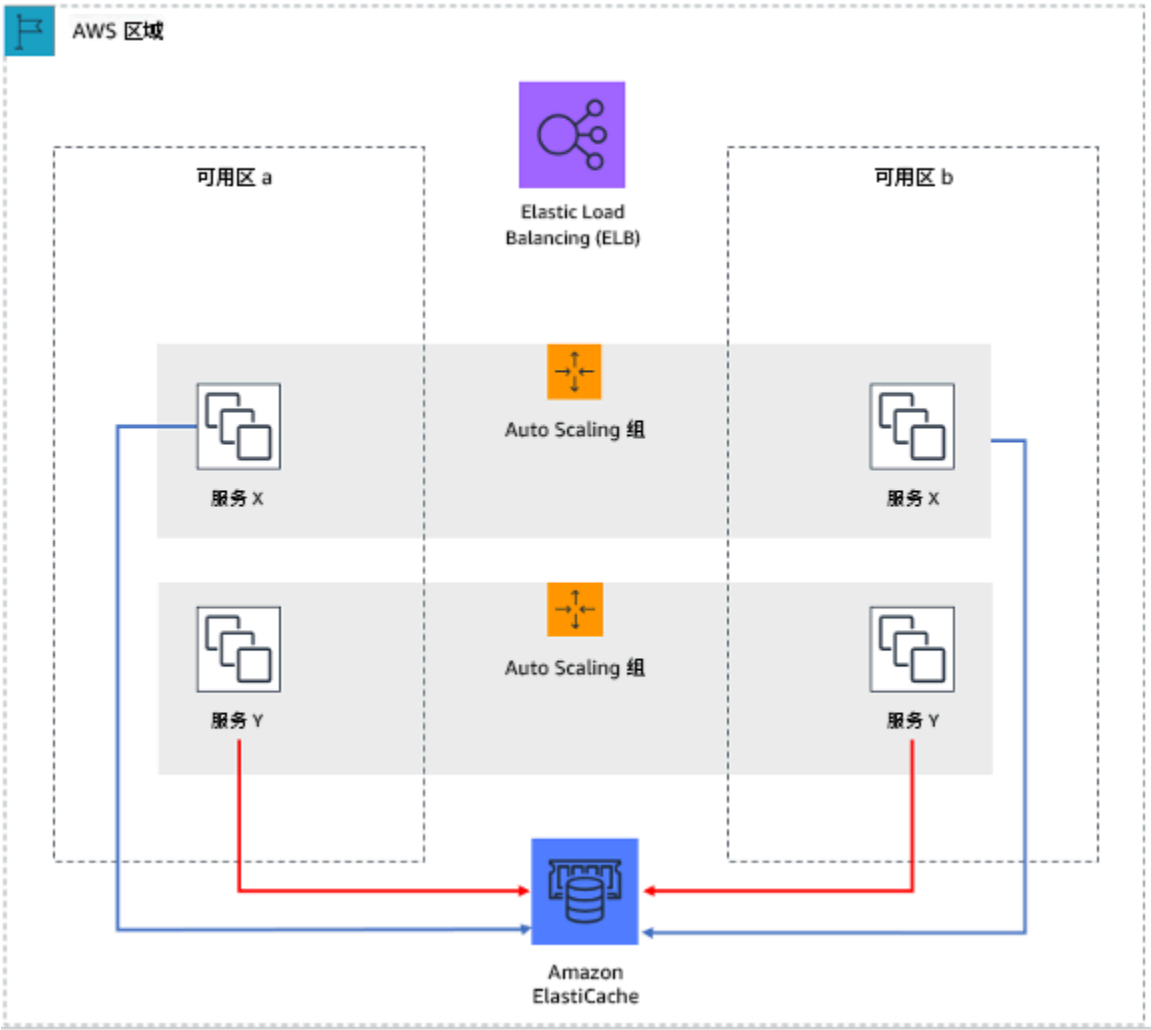


图 7：在此无状态 Web 应用程序中，会话状态会被卸载到 Amazon ElastiCache。

当用户或服务与应用程序进行交互，它们通常会执行一系列交互并构成一次会话。对于用户来说，会话是他们在使用应用程序时持续存在于请求之间的特殊数据。无状态应用程序是无需掌握之前交互而且不会存储会话信息的应用程序。

若采用无状态设计，则您可以使用无服务器计算服务，如 AWS Lambda 或 AWS Fargate。

除了服务器替换，无状态应用程序的另一项优点是，由于任何可用的计算资源（如 EC2 实例和 AWS Lambda 函数）都可以处理任何请求，因此它们可以进行横向扩展。

未建立此最佳实践暴露的风险等级：中

实施指导

- 让应用程序无状态。无状态应用程序支持水平扩展，并且可以承受单个节点故障。

- 删除可能存储在请求参数中的状态。
- 在检查是否需要状态之后，将任何状态追踪移动到具有弹性的多区域缓存或数据存储（如 Amazon ElastiCache、Amazon RDS、Amazon DynamoDB 或第三方分布式数据解决方案）。存储无法移动到弹性数据存储的状态。
- 某些数据（例如 cookie）可能在标头或查询参数中传递。
- 进行重构，从而删除可能在请求中快速传递的状态。
- 提交请求时实际上并不需要某些数据，这些数据可以按需检索。
- 删除可以异步检索的数据。
- 确定满足所需状态要求的数据存储。
- 考虑针对非关系型数据使用 NoSQL 数据库。

## 资源

### 相关文档：

- [Amazon Builders' Library：避免在分布式系统中回退](#)
- [Amazon Builders' Library：避免无法克服的队列积压](#)
- [Amazon Builders' Library：缓存挑战和策略](#)

## REL05-BP07 实施紧急杠杆

紧急杠杆是可帮助您在工作负载减轻可用性影响的快速流程。

紧急杠杆的工作原理是使用已知且经过测试的机制，禁用、节流或更改组件或依赖项的行为。这可以缓解因需求意外增加导致资源耗尽而造成的工作负载损失，并减少工作负载中非关键组件故障的影响。

期望结果：通过实施紧急杠杆，您可以建立已知良好的流程，以保持工作负载中关键组件的可用性。在激活紧急杠杆期间，工作负载应适当降级，并继续执行其关键业务功能。有关适当降级的更多详细信息，请参阅 [REL05-BP01 实施适当降级以将适用的硬依赖关系转换为软依赖关系](#)。

### 常见的反面模式：

- 非关键依赖关系的故障会影响核心工作负载的可用性。
- 在非关键组件受损时，不测试或验证关键组件的行为。
- 没有为紧急杠杆的激活或停用定义明确的标准。

建立此最佳实践的好处：实施紧急杠杆可以为解决问题的人或程序提供既定流程，以应对意外的需求高峰或非关键依赖关系的故障，从而提高工作负载中关键组件的可用性。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

- 识别工作负载中的关键组件。
- 设计和构造工作负载中的关键组件，使其能够承受非关键组件的故障。
- 进行测试以验证关键组件在非关键组件出现故障期间的行为。
- 定义和监控相关指标或触发器，以启动紧急杠杆程序。
- 定义构成紧急杠杆的（手动或自动）程序。

## 实施步骤

- 识别工作负载中的关键业务组件。
  - 工作负载中的每个技术组件都应与其相关业务职能相对应，并评定为关键组件或非关键组件。有关亚马逊的一些关键和非关键功能的示例，请参阅[任何一天都可能成为 Prime Day : Amazon.com 搜索如何使用混沌工程每秒处理超过 8.4 万次请求](#)。
  - 这既是技术决策又是业务决策，而且因组织和工作负载而异。
- 设计和构造工作负载中的关键组件，使其能够承受非关键组件的故障。
  - 在分析依赖关系期间，考虑所有潜在的故障模式，并验证您的紧急杠杆机制是否为下游组件提供了关键功能。
- 进行测试以验证关键组件在紧急杠杆激活期间的行为。
  - 避免双模态行为。有关更多详细信息，请参阅[REL11-BP05 使用静态稳定性来防止双模态行为](#)。
- 定义、监控相关指标并发出警报，以启动紧急杠杆程序。
  - 根据工作负载，找到要监控的正确指标。例如，这些指标可以是延迟，或者是对依赖关系请求失败的次数。
- 定义构成紧急杠杆的手动或自动程序。
  - 这可能包括[负载卸除](#)、[节流请求](#)或实施[适当降级](#)等机制。

## 资源

相关最佳实践：

- [REL05-BP01 实施适当降级以将适用的硬依赖关系转换为软依赖关系](#)
- [REL05-BP02 限制请求](#)
- [REL11-BP05 使用静态稳定性来防止双模态行为](#)

相关文档：

- [自动执行安全、不需要人工介入的部署](#)
- [任何一天都可能成为 Prime Day：Amazon.com 搜索如何使用混沌工程每秒处理超过 8.4 万次请求](#)

相关视频：

- [AWS re:Invent 2020：通过不可变性带来的可靠性、一致性和信心](#)

## 变更管理

问题

- [REL 6.如何监控工作负载资源？](#)
- [REL 7.您如何设计工作负载，以适应不断变化的需求？](#)
- [REL 8.如何实施更改？](#)

### REL 6.如何监控工作负载资源？

日志和指标是用于了解工作负载运行状况的强大工具。您可以配置工作负载以监控日志和指标，并在超出阈值或发生重大事件时发送通知。通过监控，您的工作负载可以发现超出低性能阈值和发生故障的情形，从而自动恢复作为回应。

最佳实践

- [REL06-BP01 为工作负载监控全部组件（生成）](#)
- [REL06-BP02 定义与计算指标（聚合）](#)
- [REL06-BP03 发送通知（实时处理和报警）](#)
- [REL06-BP04 自动响应（实时处理和告警）](#)
- [REL06-BP05 分析](#)
- [REL06-BP06 定期进行审核](#)

- [REL06-BP07 对系统中的请求进行端到端跟踪监控](#)

## REL06-BP01 为工作负载监控全部组件 (生成)

使用 Amazon CloudWatch 或第三方工具监控工作负载组件。使用 AWS Health 控制面板监控 AWS 服务。

应监控您的工作负载的全部组件，包括前端、业务逻辑和存储层。定义关键指标，描述如何将其从日志中提取出来（如有必要），并且设置用于触发对应警报事件的阈值。确保这些指标与您工作负载的关键性能指标（KPI，Key Performance Indicator）相关，并使用指标和日志来确定服务性能下降的早期警告信号。例如，每分钟成功处理的订单数等与业务成果相关的指标，相比 CPU 利用率等技术指标，可以更快地指示工作负载问题。使用 AWS Health 控制面板提供 AWS 资源底层的 AWS 服务的性能和可用性的个性化视图。

云中监控创造新的机会。大多数云提供商都已经开发出可自定义的挂钩，可以提供分析洞察来帮助您监控工作负载的多个层。Amazon CloudWatch 等 AWS 服务应用统计和机器学习算法，集中分析系统与应用程序的指标，确定正常基准，并发现异常，同时最大程度地减少用户干预。异常检测算法考虑了指标的季节性和趋势变动。

AWS 提供了丰富的监控和日志信息以供使用，这些信息可用于定义特定于工作负载的指标、需求变化流程并且采用机器学习技术而无需机器学习专业知识。

此外还会监控您的所有外部端点，确保它们独立于基本实施。这种主动监控可通过合成事务（有时被称为用户金丝雀，但不要与金丝雀部署相混淆）实现，它们会按照工作负载的客户端所执行的操作，定期执行许多常见任务。请确保这些任务的持续时间较短，并且在测试期间不要使您的工作流过载。Amazon CloudWatch Synthetics 使您能够 [创建合成金丝雀](#) 以便监控您的终端节点和 API。您还可以整合 Synthetic Canary 客户端节点和 AWS X-Ray 控制台，精确定位哪些 Synthetic Canary 遇到错误、故障，或对指定时段的速率进行限制的问题。

期望结果：

从工作负载的所有组件收集并使用关键指标，用于确保工作负载的可靠性和提供最佳用户体验。通过检测未能实现业务成果的工作负载，您可以快速发现灾难并从意外事件中恢复。

常见反模式：

- 仅监控连接到工作负载的外部接口。
- 未生成任何特定于工作负载的指标，并且只依靠工作负载所用的 AWS 服务提供给您的指标。
- 仅使用工作负载中的技术指标，不监控与工作负载所带来的非技术 KPI 相关的任何指标。



- 依靠生产流量和简单的运行状况检查来监控并评估工作负载状态。

建立此最佳实践的好处：在工作负载的所有层级进行监控，方便您更快地预测并解决组成工作负载的组件中的问题。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

1. 启用日志记录功能（如适用）。应从工作负载的所有组件获取监控数据。启用额外的日志记录，例如 S3 访问日志，并使工作负载记录特定于工作负载的数据。从 Amazon ECS、Amazon EKS、Amazon EC2、Elastic Load Balancing、AWS Auto Scaling 和 Amazon EMR 等服务收集 CPU、网络 I/O 和磁盘 I/O 平均值等指标。请参阅 [发布 CloudWatch 指标的 AWS 服务](#) 以了解将指标发布到 CloudWatch 的 AWS 服务列表。
2. 审查所有默认指标并探究任何数据收集欠缺。每项服务都生成默认指标。通过收集默认指标，您可以更好地了解工作负载组件之间的依赖关系，以及组件的可靠性和性能如何影响工作负载。您还可以 [使用](#) AWS CLI 或 API 创建自己的指标并发布到 CloudWatch。这将
3. 评估所有指标，以确定对于工作负载中的每个 AWS 服务，需要针对哪些指标发布警报。您还可以选择对工作负载可靠性有重大影响的指标的子集。专注于关键指标和阈值让您可以精调 [警报](#) 的数量，并可帮助尽可能减少误报。
4. 定义警报，以及在触发警报之后工作负载的恢复流程。通过定义警报，您可以快速通知、上报意外事件，按照必要的步骤从意外事件中恢复，并满足规定的恢复时间目标（RTO，Recovery Time Objective）。您可以使用 [Amazon CloudWatch 告警](#)，根据定义的阈值来调用自动工作流并启动恢复程序。
5. 探索使用合成事务来收集有关工作负载状态的相关数据。合成监控遵循与客户相同的路线并执行相同的操作，这使得您可以持续验证客户体验，甚至在您的工作负载上没有任何客户流量时也可以。通过使用 [合成事务](#)，您可以先于客户发现问题。

## 资源

相关最佳实践：

- [REL11-BP03 自动修复所有层](#)

相关文档：

- [开始使用 AWS Health 控制面板 – 账户的运行状况](#)

- [发布 CloudWatch 指标的 AWS 服务](#)
- [Network Load Balancer 的访问日志](#)
- [应用程序负载均衡器的访问日志](#)
- [访问 AWS Lambda 的 Amazon CloudWatch Logs](#)
- [Amazon S3 服务器访问日志记录](#)
- [启用 Classic Load Balancer 的访问日志](#)
- [将日志数据导出到 Amazon S3](#)
- [在 Amazon EC2 实例上安装 CloudWatch 代理](#)
- [发布自定义指标](#)
- [使用 Amazon CloudWatch 控制面板](#)
- [使用 Amazon CloudWatch 指标](#)
- [使用金丝雀 \( Amazon CloudWatch Synthetics \)](#)
- [什么是 Amazon CloudWatch Logs ?](#)

用户指南：

- [创建跟踪记录](#)
- [监控 Amazon EC2 Linux 实例的内存和磁盘指标](#)
- [结合使用 CloudWatch Logs 与容器实例](#)
- [VPC 流日志](#)
- [什么是 Amazon DevOps Guru ?](#)
- [什么是 AWS X-Ray ?](#)

相关博客：

- [使用 Amazon CloudWatch Synthetics 和 AWS X-Ray 进行调试](#)

相关示例和研讨会：

- [AWS Well-Architected 实验室：卓越运营 – 依赖项监控](#)
- [Amazon Builders' Library：检测分布式系统的运营可见性](#)
- [可观测性研讨会](#)

## REL06-BP02 定义与计算指标 (聚合)

存储日志数据并在必要时应用筛选条件以计算指标，例如，特定日志事件的数量，或从日志事件时间戳计算得到的延迟。

Amazon CloudWatch 和 Amazon S3 充当主要聚合层和存储层。某些服务（如 AWS Auto Scaling 和 Elastic Load Balancing）针对整个集群或实例，默认情况下为 CPU 负载或平均请求延迟提供了一些默认指标。对于流式处理服务（如 VPC 流日志和 AWS CloudTrail），事件数据将被转发给 CloudWatch Logs，您需要定义和应用指标筛选条件，才能从事件数据中提取指标。这为您提供了时间序列数据，可被输入到您定义的触发提醒的 CloudWatch 警报。

未建立此最佳实践暴露的风险等级：高

### 实施指导

- 定义与计算指标（聚合）。存储日志数据并在必要时应用筛选条件以计算指标，例如，特定日志事件的数量，或从日志事件时间戳计算得到的延迟
  - 指标筛选条件定义在将日志数据发送到 CloudWatch Logs 中所查找的术语和模式。CloudWatch Logs 使用这些指标筛选条件将日志数据转换为 CloudWatch 数字指标，您可以对这些指标绘制图形或设置警报。
    - [搜索和筛选日志数据](#)
  - 使用受信任第三方来聚合日志。
    - 遵循第三方的说明。大多数第三方产品可以与 CloudWatch 和 Amazon S3 集成。
  - 某些 AWS 服务可以直接向 Amazon S3 发布日志。如果您的主要需求是将日志存储在 Amazon S3 中，则可以让生成日志的服务轻松将其直接发送至 Amazon S3，无需设置额外的基础设施。
    - [将日志直接发送到 Amazon S3](#)

### 资源

相关文档：

- [Amazon CloudWatch Logs Insights 查询示例](#)
- [使用 Amazon CloudWatch Synthetics 和 AWS X-Ray 进行调试](#)
- [可观测性研讨会](#)
- [搜索和筛选日志数据](#)
- [将日志直接发送到 Amazon S3](#)

- [Amazon Builders' Library：检测分布式系统的运营可见性](#)

## REL06-BP03 发送通知（实时处理和报警）

当组织检测到潜在问题时，会向相应的人员和系统发送实时通知和警报，以便快速有效地处理这些问题。

期望的结果：通过根据服务和应用程序指标配置相关警报，可对运维事件做出快速响应。当超出警报阈值时，相应的人员和系统会收到通知，以便解决潜在的问题。

常见反模式：

- 配置的警报阈值过高，导致无法发送重要通知。
- 配置的警报阈值过低，导致通知过多，而重要警报无法得到处理。
- 使用情况发生变化时不更新警报及其阈值。
- 对于最好通过自动操作来处理的警报，向人员发送通知而不是生成自动操作，从而导致发送的通知过多。

建立此最佳实践的好处：通过向相应的人员和系统发送实时通知和警报，可以及早发现问题并快速处理运维方面的意外事件。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

工作负载应配备实时处理和报警功能，从而更及时地检测到可能影响应用程序可用性的问题，并充当自动响应的触发器。组织可以通过使用定义的指标创建警报来执行实时处理和报警，以便在发生重大事件或指标超过阈值时收到通知。

[Amazon CloudWatch](#) 让您可以基于静态阈值、异常检测和其他标准创建 [指标](#) 和复合警报（使用 CloudWatch 警报）。有关您可以使用 CloudWatch 配置的警报类型的详细信息，请参阅 [CloudWatch 文档的警报部分](#)。

您可以为团队自定义包含 AWS 资源指标和警报的视图 – 使用 [CloudWatch 控制面板](#)。通过 CloudWatch 控制台中的可自定义主页，您可以在单一视图中监控多个区域的资源。

警报可以执行一项或多项操作，例如，向 [Amazon SNS 主题](#) 发送通知，执行 [Amazon EC2](#) 操作或 [Amazon EC2 Auto Scaling](#) 操作，或者 [创建 OpsItem](#) 或 [事件](#)（在 AWS Systems Manager 中）。

Amazon CloudWatch 使用 [Amazon SNS](#) 在警报状态发生变化时发送通知，提供从发布者（生产者）到订阅用户（消费者）的信息传递。有关设置 Amazon SNS 通知的详细信息，请参阅 [配置 Amazon SNS](#)。

CloudWatch 会在以下情况下发送 [EventBridge 事件](#)：每当创建、更新、删除 CloudWatch 警报或者其状态发生变化时。您可以结合使用 EventBridge 与这些事件来创建执行操作的规则，例如，每当警报状态发生变化时通知您，或者使用 [Systems Manager 自动化](#) 功能自动触发账户中的事件。

何时应使用 EventBridge？何时应使用 Amazon SNS？

EventBridge 和 Amazon SNS 都可用于开发事件驱动型应用程序，您可以根据自己的具体需求进行选择。

如果您想构建一个能对来自您自己的应用程序、SaaS 应用程序和 AWS 服务的事件做出反应的应用程序，建议使用 Amazon EventBridge。EventBridge 是唯一一项直接与第三方 SaaS 合作伙伴集成的基于事件的服务。EventBridge 还可以自动从 200 多个 AWS 服务中摄取事件，而无需开发人员在自己的账户中创建任何资源。

EventBridge 使用已定义的基于 JSON 的事件结构，有助于您创建应用于整个事件主体的规则，以便选择要转发到 [目标](#) 的事件。EventBridge 目前支持将 20 多种 AWS 服务作为目标，包括 [AWS Lambda](#)、[Amazon SQS](#)、Amazon SNS、[Amazon Kinesis Data Streams](#) 和 [Amazon Data Firehose](#)。

对于需要高扇出（数千或数百万个端点）的应用程序，建议使用 Amazon SNS。我们常见的一种模式是，客户将 Amazon SNS 用作规则的目标，来筛选所需的事件并扇出到多个端点。

消息是非结构化的，可以是任何格式。Amazon SNS 支持将消息转发到六种目标，包括 Lambda、Amazon SQS、HTTP/S 端点、SMS、移动推送和电子邮件。Amazon SNS [典型延迟不超过 30 毫秒](#)。许多 AWS 服务配置为发送 Amazon SNS 消息（超过 30 个服务，包括 Amazon EC2、[Amazon S3](#) 和 [Amazon RDS](#)）。

## 实施步骤

### 1. 使用 [Amazon CloudWatch 警报](#) 来创建警报。

- a. 指标警报可监控单个 CloudWatch 指标或依赖于 CloudWatch 指标的表达式。这种警报会根据在若干时间间隔内，指标或表达式的值与阈值的比较结果，启动一项或多项操作。这些操作可能包括：向 [Amazon SNS 主题](#) 发送通知，执行 [Amazon EC2](#) 操作或 [Amazon EC2 Auto Scaling](#) 操作，或者 [创建 OpsItem](#) 或 [事件](#)（在 AWS Systems Manager 中）。
- b. 复合警报由一个规则表达式组成，该规则表达式考虑了您创建的其他警报的警报条件。只有满足所有规则条件，复合警报才会进入警报状态。复合警报的规则表达式中指定的警报可以包括指标警报和其他复合警报。复合警报可以在状态发生变化时发送 Amazon SNS 通知，并且可以在进入

警报状态时创建 Systems Manager [OpsItem](#) 或 [事件](#)，但无法执行 Amazon EC2 或 Auto Scaling 操作。

2. 设置 [Amazon SNS 通知](#)。创建 CloudWatch 警报时，可以包括 Amazon SNS 主题，以便在警报状态发生变化时发送通知。
3. [在 EventBridge 中创建规则](#) 以便与指定的 CloudWatch 警报匹配。每条规则都支持多个目标，包括 Lambda 函数。例如，您可以定义一个警报，该警报在可用磁盘空间不足时启动，它会通过 EventBridge 规则触发 Lambda 函数来清理空间。有关 EventBridge 目标的详细信息，请参阅 [EventBridge 目标](#)。

## 资源

相关的 Well-Architected 最佳实践：

- [REL06-BP01 为工作负载监控全部组件（生成）](#)
- [REL06-BP02 定义与计算指标（聚合）](#)
- [REL12-BP01 使用行动手册调查故障](#)

相关文档：

- [Amazon CloudWatch](#)
- [CloudWatch Logs Insights](#)
- [使用 Amazon CloudWatch 警报](#)
- [使用 Amazon CloudWatch 控制面板](#)
- [使用 Amazon CloudWatch 指标](#)
- [设置 Amazon SNS 通知](#)
- [CloudWatch 异常检测](#)
- [CloudWatch Logs 数据保护](#)
- [Amazon EventBridge](#)
- [Amazon Simple Notification Service](#)

相关视频：

- [re:Invent 2022 可观测性视频](#)
- [AWS re:Invent 2022 – Amazon 的可观测性最佳实践](#)

相关示例：

- [One Observability Workshop](#)
- [Amazon EventBridge 通过 Amazon CloudWatch 警报对 AWS Lambda 进行反馈控制](#)

## REL06-BP04 自动响应 ( 实时处理和告警 )

检测到事件后，利用自动化功能执行操作；例如，更换故障组件。

实施警报的自动实时处理，以便系统可以快速采取纠正措施，并在触发警报时尝试防止故障或服务降级。警报的自动响应可能包括更换故障组件，调整计算容量，将流量重定向到运行状况良好的主机、可用区或其他区域，以及通知操作员。

期望结果：识别实时警报，并设置警报的自动处理，以便调用适当的措施来维护服务级别目标和服务级别协议 ( SLA )。自动处理的范围可以是单个组件的自我修复活动，也可以是全站点的失效转移。

常见的反面模式：

- 没有明确的关键实时警报的清单或目录。
- 关键警报没有自动响应 ( 例如，当计算资源即将耗尽时自动进行扩展 )。
- 警报响应操作相互矛盾。
- 操作员在收到警报通知时没有任何标准操作程序 ( SOP ) 可以遵循。
- 不监控配置更改，因为未检测到的配置更改可能会导致工作负载停机。
- 没有撤消意外配置更改的策略。

建立此最佳实践的好处：自动警报处理可以提高系统的韧性。系统会自动采取纠正措施，从而减少手动操作，而手动操作往往是容易出错的人工干预。工作负载的运行符合可用性目标，并减少服务中断。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

为了有效地管理警报并自动进行响应，请根据警报的严重程度和影响对警报进行分类，记录响应程序，并在对任务进行评级之前制定好响应计划。

确定需要特定操作的任务 ( 通常在运行手册中详细说明 )，并检查所有运行手册和行动手册，以确定哪些任务可以自动执行。如果操作可以定义，它们通常可以自动化。如果操作无法自动化，请在 SOP 中



记录手动步骤并对操作员进行培训。不断挑战手动流程，寻找自动化机会，以便制定和维护自动响应警报的计划。

## 实施步骤

1. 创建警报清单：要获取所有警报的列表，您可以通过结合使用 [AWS CLI](#) 和 [Amazon CloudWatch](#) 命令 [describe-alarms](#)。根据您的警报数量，您可能必须使用分页来检索每次调用的警报子集，或者也可以使用 AWS SDK [通过 API 调用](#) 来获取警报。
2. 记录所有警报操作：更新关于所有警报及其操作的运行手册，无论它们是手动还是自动。[AWS Systems Manager](#) 提供了预定义的运行手册。有关运行手册的更多信息，请参阅[使用运行手册](#)。有关如何查看运行手册内容的详细信息，请参阅[查看运行手册内容](#)。
3. 设置和管理警报操作：对于任何需要操作的警报，请[使用 CloudWatch SDK 指定自动操作](#)。例如，您可以通过创建和启用警报操作或禁用警报操作，根据 CloudWatch 警报自动更改 Amazon EC2 实例的状态。

还可以使用 [Amazon EventBridge](#) 自动响应系统事件，例如应用程序可用性問題或资源变化。您可以创建规则来指明您对哪些事件感兴趣，以及当事件与规则匹配时要采取的操作。可以自动启动的操作包括调用 [AWS Lambda](#) 函数、调用 [Amazon EC2](#) Run Command、将事件中继到 [Amazon Kinesis Data Streams](#)，以及查看[使用 EventBridge 实现 Amazon EC2 自动化](#)。

4. 标准操作程序 (SOP)：根据您的应用程序组件，[AWS Resilience Hub](#) 会推荐多个 [SOP 模板](#)。您可以使用这些 SOP 来记录在出现警报时操作员应遵循的所有流程。您还可以根据 Resilience Hub 的建议[构建 SOP](#)，其中您需要一个具有相关韧性策略的 Resilience Hub 应用程序，以及针对该应用程序的历史韧性评估。针对您的 SOP 的建议由韧性评估生成。

Resilience Hub 与 Systems Manager 结合，通过提供许多可用作 SOP 基础的 [SSM 文档](#)，自动执行 SOP 的步骤。例如，Resilience Hub 可以基于现有 SSM 自动化文档，推荐用于添加磁盘空间的 SOP。

5. 使用 Amazon DevOps Guru 执行自动化操作：可以使用 [Amazon DevOps Guru](#) 自动监控应用程序资源的异常行为并提供针对性的建议，以缩短识别问题和进行修复所需的时间。借助 DevOps Guru，您可以近乎实时地监控来自多个来源的运营数据流，包括 Amazon CloudWatch 指标、[AWS Config](#)、[AWS CloudFormation](#) 和 [AWS X-Ray](#)。您还可以使用 DevOps Guru 在 OpsCenter 中自动创建 [OpsItems](#)，并将事件发送到 [EventBridge 以实现更多自动化操作](#)。

## 资源

相关最佳实践：

- [REL06-BP01 为工作负载监控全部组件 \(生成\)](#)

- [REL06-BP02 定义与计算指标 \( 聚合 \)](#)
- [REL06-BP03 发送通知 \( 实时处理和报警 \)](#)
- [REL08-BP01 对部署等标准活动使用运行手册](#)

#### 相关文档：

- [AWS Systems Manager 自动化](#)
- [创建通过 AWS 资源中的事件触发的 EventBridge 规则](#)
- [可观测性研讨会](#)
- [Amazon Builders' Library：检测分布式系统的运营可见性](#)
- [什么是 Amazon DevOps Guru？](#)
- [使用自动化文档 \( 行动手册 \)](#)

#### 相关视频：

- [AWS re:Invent 2022 – Amazon 的可观测性最佳实践](#)
- [AWS re: Invent 2020：使用 AWS Systems Manager 实现全面自动化](#)
- [AWS Resilience Hub 简介](#)
- [为 Amazon DevOps Guru 通知创建自定义工单系统](#)
- [使用 Amazon DevOps Guru 启用多账户洞察聚合](#)

#### 相关示例：

- [可靠性研讨会](#)
- [Amazon CloudWatch 和 Systems Manager 研讨会](#)

#### REL06-BP05 分析

收集日志文件和指标历史，并对其进行分析以获得更广泛的趋势和工作负载见解。

Amazon CloudWatch Logs Insights 支持 [简单但强大的查询语言](#)，您可以用它分析日志数据。Amazon CloudWatch Logs 还支持订阅，允许数据无缝流动到 Amazon S3 ( 您可以在其中使用此类数据 ) 或 Amazon Athena 以便对数据进行查询。它还支持查询多种格式。请参阅 [支持的 SerDes 和数据格式](#) ( 参见 Amazon Athena 用户指南 )。针对大型日志文件集的分析，您可以运行 Amazon EMR 集群以执行 PB 级分析。

AWS 合作伙伴和第三方提供了许多用于聚合、处理、存储和分析的工具。这些工具包括 New Relic、Splunk、Loggly、Logstash、CloudHealth 和 Nagios。但是，系统和应用程序日志之外的生成对于每个云提供商，甚至每个服务来说都是独一无二的。

监控过程中常常被忽视的部分是数据管理。您需要确定数据监控的保留要求，然后相应地应用生命周期策略。Amazon S3 支持 S3 存储桶级别的生命周期管理。此生命周期管理可以通过不同的方式应用到存储桶中的不同路径。您可以在生命周期临近结束时，将数据转移到 Amazon S3 Glacier 进行长期存储，然后在保留期结束后让它们过期。S3 智能分层存储类旨在通过将数据自动移动到最具成本效益的访问层，而不会对性能或运营开销产生影响，从而实现优化成本的目的。

未建立此最佳实践暴露的风险等级：中

### 实施指导

- 借助 CloudWatch Logs Insights，您可对 Amazon CloudWatch Logs 中的日志数据进行交互搜索和分析。
  - [使用 CloudWatch Logs Insights 分析日志数据](#)
  - [Amazon CloudWatch Logs Insights 查询示例](#)
- 使用 Amazon CloudWatch Logs 将日志发送到 Amazon S3，您可以在此处使用这些日志或者使用 Amazon Athena 来查询数据。
  - [如何使用 Athena 分析我的 Amazon S3 服务器访问日志？](#)
    - 为服务器访问日志存储桶创建 S3 生命周期策略。配置生命周期策略以定期删除日志文件。这样做可以减少 Athena 针对每次查询分析的数据量。
    - [如何为 S3 存储桶创建生命周期策略？](#)

### 资源

#### 相关文档：

- [Amazon CloudWatch Logs Insights 查询示例](#)
- [使用 CloudWatch Logs Insights 分析日志数据](#)
- [使用 Amazon CloudWatch Synthetics 和 AWS X-Ray 进行调试](#)
- [如何为 S3 存储桶创建生命周期策略？](#)
- [如何使用 Athena 分析我的 Amazon S3 服务器访问日志？](#)
- [可观测性研讨会](#)
- [Amazon Builders' Library：检测分布式系统的运营可见性](#)

## REL06-BP06 定期进行审核

经常审核工作负载监控的实施情况，并根据重大事件和变更加以更新。

关键业务指标可促进有效监控。确保随着业务优先事项的变化在您的工作负载中对这些指标进行调整。

审计监控有助于确保您了解应用程序何时达到其可用性目标。根本原因分析需要具备在出现故障时发现具体情况的能力。AWS 提供的服务让您能够在意外事件发生期间跟踪服务的状态：

- Amazon CloudWatch Logs：您可以将日志存储在此服务中并检查日志内容。
- Amazon CloudWatch Logs Insights：是一项完全托管式服务，让您可以在数秒内分析大量日志。它为您提供快速、交互式的查询和可视化。
- AWS Config：您可以查看在不同的时间点使用了哪些 AWS 基础设施。
- AWS CloudTrail：您可以查看哪些委托人在什么时候调用了哪些 AWS API。

AWS 每周召开一次会议，[以审查运营性能](#)并在团队之间分享经验。因为 AWS 有很多团队，我们设置了 [The Wheel](#) 以随机挑选一个工作负载进行审查。定期开展运营性能审查和知识共享，有助于您增强帮助运营团队提高绩效的能力。

常见反模式：

- 仅收集默认指标。
- 设置监控策略后不再过问。
- 部署重大更改后不讨论监控问题。

建立此最佳实践的好处：定期审核监控可主动预测潜在问题，而不是当预测问题真实发生后被动应对通知。

未建立此最佳实践暴露的风险等级：中

实施指导

- 为工作负载创建多个控制面板。您必须具有顶级控制面板，其中包含关键业务指标，以及已确定与使用情况发生变化时工作负载的预期运行状况最相关的技术指标。您还应该具有可以检查各种应用程序层和依赖项的控制面板。
  - [使用 Amazon CloudWatch 控制面板](#)
- 计划和执行工作负载控制面板常规检查。执行控制面板常规检查。您可能对检查深度具有不同的安排。

- 检查指标中的趋势。对比指标值与历史值，了解是否有趋势表明需要调查某些情况。这种情况的示例包括：延迟增加、主要业务功能减少以及故障响应增加。
- 检查指标中的离群值/异常值。平均值或中值会掩盖离群值和异常值。查看时间范围内的最高值和最低值，调查出现这些极值的原因。当您继续消除这些原因时，降低对极值的定义可以使您继续提高工作负载性能的一致性。
- 查找清晰的行为变化。指标数量或方向的立即更改可能表示应用程序出现更改，或者出现了您需要添加额外指标进行跟踪外部因素。

## 资源

### 相关文档：

- [Amazon CloudWatch Logs Insights 查询示例](#)
- [使用 Amazon CloudWatch Synthetics 和 AWS X-Ray 进行调试](#)
- [可观测性研讨会](#)
- [Amazon Builders' Library：检测分布式系统的运营可见性](#)
- [使用 Amazon CloudWatch 控制面板](#)

## REL06-BP07 对系统中的请求进行端到端跟踪监控

跟踪各个服务组件的请求处理情况，这样产品团队便能够更轻松地对分析和调试问题并提高性能。

期望的结果：针对所有组件全面跟踪工作负载，实现轻松调试，进而通过简化发现错误根本原因的过程，缩短错误的 [平均解决时间](#)（MTTR）和延迟。采用端到端的跟踪方式，有助于更快地发现受影响的组件，并详细深入地了解造成错误或延迟的根本原因。

### 常见反模式：

- 只针对部分组件而不是全部组件进行跟踪。例如，如果不跟踪 AWS Lambda，团队可能无法清楚地了解高峰工作负载中冷启动所造成的延迟。
- Synthetics Canary 或真实用户监控（RUM）未配置跟踪功能。没有 Canary 或 RUM，跟踪分析中会忽略客户端交互遥测数据，这样得出的性能概况就不够完整。
- 混合工作负载包括云原生跟踪工具和第三方跟踪工具，但尚未采取措施来选择并完全集成单个跟踪解决方案。根据所选跟踪解决方案，应使用云原生跟踪 SDK 来检测非云原生组件，或者应将第三方工具配置为摄取云原生跟踪遥测数据。

建立此最佳实践的好处：当开发团队收到问题提醒时，能够查看系统组件交互情况的全貌，包括各个组件在日志记录、性能和故障方面的相关性。由于跟踪有助于直观且轻松地找出根本原因，因此调查根本原因所花费的时间得以减少。在解决问题时，团队如果能详细了解组件的交互情况，就可以更快地做出更好的决策。分析系统跟踪数据有助于改进多种决策，例如何时调用灾难恢复（DR）失效转移，或者在何处实施自我修复策略最合适等，最终势必能够提高客户对服务的满意度。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

团队在运行分布式应用程序时，能够借助跟踪工具来建立关联标识符、收集请求跟踪数据，以及构建互联组件的服务地图。请求跟踪中应该涵盖所有应用程序组件，包括服务客户端、中间件网关和事件总线、计算组件以及存储（包括键/值存储和数据库）。在端到端跟踪配置中纳入 Synthetics Canary 和真实用户监控，来衡量远程客户端交互情况和延迟，这样您就能够根据服务等级协议和目标准确地评估系统性能。

您可以使用 [AWS X-Ray](#) 和 [Amazon CloudWatch 应用程序监控](#) 检测服务，全面了解应用程序的请求处理情况。X-Ray 可收集应用程序遥测数据，让您能够跨有效负载、函数、跟踪、服务、API 对数据进行可视化和筛选，并且能够通过无代码或低代码方式，为系统组件开启此功能。CloudWatch 应用程序监控包括 ServiceLens，可将您的跟踪数据与指标、日志和警报整合在一起。CloudWatch 应用程序监控还包括用于监控端点和 API 的 Synthetics，以及用于检测 Web 应用程序客户端的真实用户监控。

## 实施步骤

- 对所有受支持的原生服务使用 AWS X-Ray，例如 [Amazon S3](#)、[AWS Lambda](#) 和 [Amazon API Gateway](#)。这些 AWS 服务可使用基础设施即代码、AWS SDK 或 AWS Management Console，通过切换配置来启用 X-Ray。
- 检测应用程序 [适用于 OpenTelemetry 的 AWS Distro 以及 X-Ray](#) 或第三方数据收集代理。
- 请查看 [AWS X-Ray 开发人员指南](#)，了解实施所用的具体编程语言。这些文档部分详细介绍了如何检测 HTTP 请求、SQL 查询和其他特定于应用程序编程语言的进程。
- 使用 X-Ray 来跟踪 [Amazon CloudWatch Synthetics Canary](#) 和 [Amazon CloudWatch RUM](#)，以便分析从最终用户客户端到下游 AWS 基础设施的请求路径。
- 根据资源运行状况和 Canary 遥测数据来配置 CloudWatch 指标和警报，这样团队就能够快速收到问题提醒，然后使用 ServiceLens 深入探究跟踪数据和服务地图。
- 如果您使用第三方工具作为主要跟踪解决方案，则将 X-Ray 与下列第三方跟踪工具进行集成：[Datadog](#)、[New Relic](#)或 [Dynatrace](#)。



## 资源

### 相关最佳实践：

- [REL06-BP01 为工作负载监控全部组件（生成）](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)

### 相关文档：

- [什么是 AWS X-Ray？](#)
- [Amazon CloudWatch：应用程序监控](#)
- [使用 Amazon CloudWatch Synthetics 和 AWS X-Ray 进行调试](#)
- [Amazon Builders' Library：检测分布式系统的运营可见性](#)
- [将 AWS X-Ray 与其他 AWS 服务集成](#)
- [适用于 OpenTelemetry 的 AWS Distro 以及 AWS X-Ray](#)
- [Amazon CloudWatch：使用 Synthetics 监控](#)
- [Amazon CloudWatch：使用 CloudWatch RUM](#)
- [设置 Amazon CloudWatch Synthetics Canary 和 Amazon CloudWatch 警报](#)
- [可用性及其他内容：了解和提高 AWS 上分布式系统的韧性](#)

### 相关示例：

- [可观测性研讨会](#)

### 相关视频：

- [AWS re:Invent 2022 – 如何跨多个账户监控应用程序](#)
- [如何监控 AWS 应用程序](#)

### 相关工具：

- [AWS X-Ray](#)
- [Amazon CloudWatch](#)
- [Amazon Route 53](#)



## REL 7.您如何设计工作负载，以适应不断变化的需求？

可扩展工作负载具有自动添加或移除资源的弹性，因此确保在任何时间点都能准确满足当前的需求。

### 最佳实践

- [REL07-BP01 在获取或扩展资源时利用自动化](#)
- [REL07-BP02 在检测到对工作负载的破坏时获取资源](#)
- [REL07-BP03 当检测到某个工作负载需要更多资源时，就会获取资源](#)
- [REL07-BP04 对工作负载进行负载测试](#)

### REL07-BP01 在获取或扩展资源时利用自动化

在替换被损坏的资源或扩展您的工作负载时，通过采用托管 AWS 服务（如 Amazon S3 和 AWS Auto Scaling）对流程进行自动处理。您还可以使用第三方工具和 AWS 开发工具包自动扩展。

托管 AWS 服务包括 Amazon S3、Amazon CloudFront、AWS Auto Scaling、AWS Lambda、Amazon DynamoDB、AWS Fargate 和 Amazon Route 53。

AWS Auto Scaling 让您检测与替换被破坏的实例。它还可以帮助您为资源制定扩展计划，包括 [Amazon EC2](#) 实例和 Spot 队列、[Amazon ECS](#) 任务、[Amazon DynamoDB](#) 表和索引，以及 [Amazon Aurora](#) 副本。

在扩展 EC2 实例时，请确保您使用多个可用区（最好至少三个）并增加或减少容量以保持这些可用区之间的平衡。ECS 任务或 Kubernetes 容器组（pod）（使用 Amazon Elastic Kubernetes Service 时）也应分布在多个可用区中。

如果使用 AWS Lambda，实例会自动扩展。每次收到关于您的函数的事件通知时，AWS Lambda 会快速找到其计算队列中的可用容量，然后运行您的代码至分配的并发值。您需要确保在特定的 Lambda 上，以及在您的 Service Quotas 中配置必要的并发值。

Amazon S3 会自动扩展以处理较高的请求速率。例如，您的应用程序可以在存储桶中为每个前缀每秒至少发送 3500 个 PUT/COPY/POST/DELETE 或 5500 个 GET/HEAD 请求。存储桶中的前缀数量没有限制。您可以通过并行化读取提高您的读取或写入性能。例如，如果在 Amazon S3 存储桶中创建 10 个前缀以便对读取进行并行化，您可以将读取性能扩展至每秒 55000 个读取请求。

配置和使用 Amazon CloudFront 或受信任的内容分发网络（CDN，Content Delivery Network）。CDN 可以缩短最终用户的响应时间，并从缓存中为请求提供内容，从而减少扩展工作负载的请求。

常见反模式：

- 实施 Auto Scaling 组进行自动修复，但无法实施弹性。
- 使用 Auto Scaling 响应流量激增。
- 部署高状态应用程序，消除了部署弹性选项。

建立此最佳实践的好处：自动化可以避免部署和淘汰资源时的潜在手动错误。自动化可以避免由于缓慢响应部署或淘汰需求而导致的服务超支和拒绝服务风险。

未建立此最佳实践暴露的风险等级：高

## 实施指导

- 配置和使用 AWS Auto Scaling。它会监控您的应用程序，并自动调整容量来维持稳定、可预测的性能，并且成本最低。使用 AWS Auto Scaling，您可以跨多个服务为多个资源轻松设置应用程序扩展。
  - [什么是 AWS Auto Scaling？](#)
    - 在您的 Amazon EC2 实例和竞价型实例集、Amazon ECS 任务、Amazon DynamoDB 表和索引、Amazon Aurora 副本以及 AWS Marketplace 设备上配置自动扩展（如果适用）。
    - [使用 DynamoDB Auto Scaling 自动管理吞吐能力](#)
      - 使用服务 API 操作来指定警报、扩展策略、预热时间和冷却时间。
  - 使用 Elastic Load Balancing。负载均衡器可以按路径或网络连接分配负载。
    - [什么是 Elastic Load Balancing？](#)
      - Application Load Balancers 可以按路径分配负载。
        - [什么是 Application Load Balancer？](#)
          - 配置 Application Load Balancer，根据域名下的路径将流量分配给不同的工作负载。
          - Application Load Balancers 可以与 AWS Auto Scaling 集成来分配负载，以便管理需求。
            - [将负载均衡器与自动扩缩组配合使用](#)
      - 网络负载均衡器可以按连接分配负载。
        - [什么是网络负载均衡器？](#)
          - 配置网络负载均衡器，以便使用 TCP 将流量分配给不同的工作负载，或者为工作负载指定一组恒定的 IP 地址。
          - 网络负载均衡器可以与 AWS Auto Scaling 集成来分配负载，以便管理需求。
    - 使用高度可用的 DNS 提供商。使用 DNS 名称，用户可以输入名称而不是 IP 地址来访问您的工作负载，并将该信息分发到指定的范围内，通常面向全局范围内工作负载的所有用户。
      - 使用 Amazon Route 53 或可信 DNS 提供商。

- [什么是 Amazon Route 53 ?](#)
- 使用 Route 53 管理 CloudFront 分配和负载均衡器。
  - 确定要管理的域和子域。
  - 使用 ALIAS 或 CNAME 记录来创建适当的记录集。
  - [使用记录](#)
- 使用 AWS 全球网络可优化用户与应用程序之间的路径。AWS Global Accelerator 持续监控应用程序端点的运行状况，可在 30 秒内将流量重定向到运行状况良好的端点。
  - AWS Global Accelerator 是一项可帮助本地或全球用户提高应用程序可用性和性能的服务。它提供的静态 IP 地址可用作从单个或多个 AWS 区域 区域 ( 例如 Application Load Balancers、网络负载均衡器或 Amazon EC2 实例 ) 访问应用程序端点的固定入口点。
  - [什么是 AWS Global Accelerator ?](#)
- 配置和使用 Amazon CloudFront 或受信任的内容分发网络 ( CDN , Content Delivery Network ) 。内容分发网络可以缩短最终用户的响应时间，还可以对可能导致工作负载进行不必要扩展的内容请求做出响应。
  - [什么是 Amazon CloudFront ?](#)
    - 针对您的工作负载配置 Amazon CloudFront 分配，或者使用第三方 CDN。
      - 您可以通过在端点安全组或访问策略中使用 CloudFront 的 IP 范围，将对工作负载的访问限制为只能从 CloudFront 访问。

## 资源

### 相关文档：

- [APN 合作伙伴：可以帮您制定自动计算解决方案的合作伙伴](#)
- [AWS Auto Scaling：扩展计划的工作原理](#)
- [AWS Marketplace：可以与 Auto Scaling 一起使用的产品](#)
- [使用 DynamoDB Auto Scaling 自动管理吞吐能力](#)
- [将负载均衡器与自动扩缩组配合使用](#)
- [什么是 AWS Global Accelerator ?](#)
- [什么是 Amazon EC2 Auto Scaling ?](#)
- [什么是 AWS Auto Scaling ?](#)
- [什么是 Amazon CloudFront ?](#)
- [什么是 Amazon Route 53 ?](#)

- [什么是 Elastic Load Balancing ?](#)
- [什么是网络负载均衡器 ?](#)
- [什么是 Application Load Balancer ?](#)
- [使用记录](#)

## REL07-BP02 在检测到对工作负载的破坏时获取资源

如果可用性受到影响，在必要时被动扩展资源，从而还原工作负载的可用性。

首先，您必须配置运行状况检查和关于此类检查的标准，表示在什么时候可用性会因缺少资源而受到影响。然后，通知适当的人员手动扩展资源，或启动自动化以对其进行自动扩展。

可以为您的工作负载手动调整规模（例如，通过 AWS Management Console 或 AWS CLI 更改 Auto Scaling 组中 EC2 实例的数量，或者修改 DynamoDB 表的吞吐量来实现）。但是，应尽可能使用自动化（请参阅[获取或扩展资源时使用自动化](#)）。

期望结果：在检测到故障或客户体验下降时，启动扩展活动（自动或手动）以恢复可用性。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

对工作负载中的所有组件实施可观测性和监控，以监控客户体验并检测故障。定义扩展所需资源的手动或自动程序。有关更多信息，请参阅[REL11-BP01 监控工作负载的所有组件以检测故障](#)。

### 实施步骤

- 定义扩展所需资源的手动或自动程序。
  - 扩展程序取决于工作负载中不同组件的设计方式。
  - 扩展程序还因所使用的底层技术而异。
    - 使用 AWS Auto Scaling 的组件可以使用扩展计划来配置一组用于扩展资源的指令。如果使用 AWS CloudFormation 或为 AWS 资源添加标签，您可以根据应用程序为不同的资源集设置扩展计划。Auto Scaling 提供了针对每个资源自定义的扩展策略建议。创建扩展计划后，Auto Scaling 结合了动态扩展和预测式扩缩方法来支持扩展策略。有关更多详细信息，请参阅[扩展计划的工作原理](#)。
  - Amazon EC2 Auto Scaling 会验证您是否拥有适量的 Amazon EC2 实例，可处理您的应用程序负载。您可创建 EC2 实例的集合，称为 Auto Scaling 组。您可以指定每个 Auto Scaling 组中的

最小和最大实例数，Amazon EC2 Auto Scaling 会确保您的组永远不会低于或超过这些限制。有关更多详细信息，请参阅[什么是 Amazon EC2 Auto Scaling ?](#)

- Amazon DynamoDB 自动扩缩使用 Application Auto Scaling 服务，代表您动态调整预置的吞吐能力，以响应实际的流量模式。这将使表或全局二级索引提高预置读取和写入容量，从而在不节流的情况下应对流量突增。有关更多详细信息，请参阅[使用 DynamoDB 自动扩缩自动管理吞吐容量](#)。

## 资源

### 相关最佳实践：

- [REL07-BP01 在获取或扩展资源时利用自动化](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)

### 相关文档：

- [AWS Auto Scaling：扩展计划的工作原理](#)
- [使用 DynamoDB 自动扩缩自动管理吞吐容量](#)
- [什么是 Amazon EC2 Auto Scaling ?](#)

REL07-BP03 当检测到某个工作负载需要更多资源时，就会获取资源

主动扩展资源以满足需求并避免影响可用性。

很多 AWS 服务会自动扩展以满足需求。如果使用 Amazon EC2 实例或 Amazon ECS 集群，您可以根据与您的工作负载的需求对应的使用指标，配置它们会在何时自动扩展。针对 Amazon EC2，平均 CPU 利用率、负载均衡器请求数量，或网络带宽可被用于扩展（或缩减）EC2 实例。而对于 Amazon ECS，可使用平均 CPU 利用率、负载均衡器请求数量和内存利用率横向扩展（或横向缩减）ECS 任务。在 AWS 上使用 Target Auto Scaling，Autoscaler 将扮演“家用恒温器”的角色，增加或减少资源以保持您所指定的目标值（例如，70% CPU 利用率）。

AWS Auto Scaling 还可以执行 [Predictive Auto Scaling](#)，该操作利用机器学习来分析每个资源的历史工作负载，并且定期预测未来两天的负载。

利特尔法则可帮助计算您需要多少计算实例（EC2 实例、并发 Lambda 函数，等等）。

$$L = \lambda W$$

$L$  = 实例数量 ( 或系统中的平均并发值 )

$\lambda$  = 收到请求的平均速率 ( 请求数量/秒 )

$W$  = 每个请求在系统中所花的平均时间 ( 秒 )

例如，假设每秒请求数为 100，若每个请求所需的处理时间为 0.5 秒，您将需要 50 个实例才能满足需求。

未建立此最佳实践暴露的风险等级：中

### 实施指导

- 当检测到某个工作负载需要更多资源时，就会获取资源。主动扩展资源以满足需求并避免影响可用性。
  - 计算处理给定请求速率需要多少计算资源 ( 计算并发 ) 。
    - [讲述与利特尔法则有关的故事](#)
  - 当您具有历史使用模式时，请为 Amazon EC2 Auto Scaling 设置计划扩展。
    - [Amazon EC2 Auto Scaling 的计划扩缩](#)
  - 使用 AWS 预测式扩缩。
    - [由机器学习提供支持的 EC2 预测式扩缩](#)

### 资源

相关文档：

- [AWS Auto Scaling：扩展计划的工作原理](#)
- [AWS Marketplace：可以与 Auto Scaling 一起使用的产品](#)
- [使用 DynamoDB Auto Scaling 自动管理吞吐能力](#)
- [由机器学习提供支持的 EC2 预测式扩缩](#)
- [Amazon EC2 Auto Scaling 的计划扩缩](#)
- [讲述与利特尔法则有关的故事](#)
- [什么是 Amazon EC2 Auto Scaling？](#)

### REL07-BP04 对工作负载进行负载测试

采用负载测试方法来衡量扩展活动能否满足工作负载要求。

持续开展负载测试，这一点很重要。负载测试用于发现工作负载的断点并测试工作负载的性能。利用 AWS，您可以轻松设置能够模拟生产工作负载规模的临时测试环境。在云中，您可以根据需要创建一套生产规模等级的测试环境，完成测试，然后停用资源。由于测试环境只需在运行时付费，您模拟真实环境的成本仅为本地测试成本的一小部分。

生产中的负载测试还应该被视为实际试用活动的一部分，因为在客户使用量降低的那几个小时内，在场的员工都忙于解读结果与处理任何出现的问题，生产系统承受着很大的压力。

常见反模式：

- 对与您的生产采用不同配置的部署执行负载测试。
- 仅对单个工作负载分段（而非整个工作负载）执行负载测试。
- 使用请求子集，而不是具有代表性的实际请求集执行负载测试。
- 对超出预期负载的较小安全系数执行负载测试。

建立此最佳实践的好处：您知道架构中哪些组件会在负载下失败，而且能够确定要监控哪些可指示您即将达到该负载的指标，从而及时解决问题，防止故障影响。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

- 执行负载测试，确定工作负载的哪些方面表明您必须添加或移除容量。负载测试应具有您在生产中接收的流量类似的代表性流量。增加负载，同时监视所有已检测指标，以便确定哪种指标指示何时必须添加或移除资源。
  - [AWS 上的分布式负载测试：模拟数千个连接的用户](#)
    - 确定请求组合。您可能拥有不同的请求组合，因此应当在确定流量组合时查看不同的时间范围。
    - 实施负载驱动程序。您可以使用自定义代码、开源或商用软件来实施负载驱动程序。
    - 最初使用小容量进行负载测试。通过将负载降低到较小容量（可能小到一个实例或容器），可能会有立竿见影的效果。
    - 针对更大的容量进行负载测试。分布式负载的效果会有所不同，因此您必须对尽量接近生产环境的目标进行测试。

资源

相关文档：



- [AWS 上的分布式负载测试：模拟数千个连接的用户](#)

## REL 8. 如何实施更改？

要部署新功能，必须对更改加以控制，以确保工作负载和操作环境正在运行已知的软件，并以可预测的方式进行修补和替换。如果此类更改不受控制，您将难以预测这些更改的影响，或难以处理由它们引发的问题。

### 最佳实践

- [REL08-BP01 对部署等标准活动使用运行手册](#)
- [REL08-BP02 将功能测试作为部署的一部分进行集成](#)
- [REL08-BP03 将弹性测试作为部署的一部分进行集成](#)
- [REL08-BP04 使用不可变基础设施进行部署](#)
- [REL08-BP05 使用自动化功能部署更改](#)

### REL08-BP01 对部署等标准活动使用运行手册

运行手册是用来实现特定结果的预定义程序。使用运行手册执行标准活动，无论这些活动是手动还是自动执行。其中的示例包括部署工作负载，修补工作负载，或修改 DNS。

例如，实施流程以 [确保部署期间安全回滚](#)。确保您可以为客户进行部署回滚而不会出现中断，这是保证服务可靠的关键。

针对运行手册程序，从一个有效的手动流程开始，用代码进行实施，并在适当的情况下触发其自动运行。

即使是高度自动化的复杂工作负载，运行手册同样适用于 [运行实际试用](#) 或用于满足严格的报告和审计要求。

请注意，行动手册可用于对特定事件做出响应，运行手册则用来达成特定的结果。通常，运行手册适用于例行活动，而行动手册则被用于对非例行事件做出响应。

常见反模式：

- 对生产中的配置执行计划外更改。
- 跳过计划中的步骤以加快部署速度，导致部署失败。
- 在未测试反向更改的情况下做出更改。

建立此最佳实践的好处：有效更改计划有助于您成功执行更改，因为您知道所有受影响的系统。在测试环境中验证更改能够增强您的信心。

未建立此最佳实践暴露的风险等级：高

### 实施指导

- 通过在运行手册中记录程序，实现对为人熟知的事件的一致且及时的响应。
  - [AWS Well-Architected Framework：概念：运行手册](#)
- 使用基础设施即代码的原则定义您的基础设施。通过使用 AWS CloudFormation（或受信任的第三方）来定义您的基础设施，您可以使用版本控制软件对更改实施版本控制并进行跟踪。
  - 使用 AWS CloudFormation（或受信任的第三方提供商）定义您的基础设施。
    - [什么是 AWS CloudFormation？](#)
  - 使用良好的软件设计原则创建单个解耦模板。
    - 确定实施的权限、模板和责任方。
      - [使用 AWS Identity and Access Management 控制访问权限](#)
    - 使用源代码控制（例如 AWS CodeCommit 或受信任的第三方工具）进行版本控制。
      - [什么是 AWS CodeCommit？](#)

### 资源

#### 相关文档：

- [AWS 合作伙伴：可以帮助您创建自动化部署解决方案的合作伙伴](#)
- [AWS Marketplace：可用于自动实施部署的产品](#)
- [AWS Well-Architected Framework：概念：运行手册](#)
- [什么是 AWS CloudFormation？](#)
- [什么是 AWS CodeCommit？](#)

#### 相关示例：

- [使用行动手册和运行手册自动完成操作](#)

REL08-BP02 将功能测试作为部署的一部分进行集成

功能测试作为自动化部署的一部分运行。若未满足成功条件，则相关管道会中止或回滚。

这些测试在预生产环境中运行，该环境会在管道中的生产开始前被暂存。在理想情况下，此操作是部署管道的一部分。

未建立这种最佳实践的情况下暴露的风险等级：高

#### 实施指导

- 将功能测试作为部署的一部分进行集成。功能测试作为自动化部署的一部分运行。若未满足成功条件，则相关管道会中止或回滚。
- 当在 AWS CodePipeline 中建模的软件发布管道执行“测试操作”时，调用 AWS CodeBuild。此功能使您能够对代码轻松运行各种测试，例如单元测试、静态代码分析和集成测试。
  - [AWS CodePipeline 增加了对通过 AWS CodeBuild 进行单位和自定义集成测试的支持](#)
- 使用 AWS Marketplace 解决方案，将自动化测试作为软件交付管道的一部分执行。
  - [软件测试自动化](#)

#### 资源

相关文档：

- [AWS CodePipeline 增加了对通过 AWS CodeBuild 进行单位和自定义集成测试的支持](#)
- [软件测试自动化](#)
- [什么是 AWS CodePipeline？](#)

REL08-BP03 将弹性测试作为部署的一部分进行集成

将弹性测试（使用 [混沌工程的原则](#)）作为预生产环境中自动化部署管道的一部分执行。

这些测试会在预生产环境的管道中暂存并运行。它们应在生产中运行，作为 [实际试用](#) 的一部分。

未建立这种最佳实践的情况下暴露的风险等级：中

#### 实施指导

- 将弹性测试作为部署的一部分进行集成。混沌工程是对工作负载进行试验的规范，用于建立人们对工作负载能够在生产中经受住混乱情形的信心。
  - 弹性测试会注入故障或资源降级，以此评估您的工作负载能否以预期弹性做出响应。
  - [Well-Architected 实验室：第 300 级：测试 EC2 RDS 和 S3 的弹性](#)
- 这些测试可以在自动部署管道的预生产环境中定期执行。

- 它们还应作为计划实际演练的一部分在生产环境中运行。
- 使用混沌工程原则，提出有关工作负载在各种破坏情况下如何表现的假设，然后使用弹性测试验证您的假设。
- [混沌工程的原则](#)

## 资源

### 相关文档：

- [混沌工程的原则](#)
- [什么是 AWS Fault Injection Simulator?](#)

### 相关示例：

- [Well-Architected 实验室：第 300 级：测试 EC2 RDS 和 S3 的弹性](#)

## REL08-BP04 使用不可变基础设施进行部署

不可变基础设施模式要求在生产工作负载上不会出现就地更新、安全补丁或配置更改。需要更改时，会在新的基础设施上构建架构，并将其部署到生产环境中。

遵循不可变基础设施部署策略，以提高工作负载部署的可靠性、一致性和可重复性。

期望结果：使用不可变基础设施，不允许为了运行工作负载中的基础设施资源而[就地修改](#)。相反，当需要更改时，会将一组包含所有必要更改的新基础设施资源与现有资源并行部署。会自动验证此部署，如果成功，流量将逐渐转移到新的资源集。

此部署策略适用于软件更新、安全补丁、基础设施更改、配置更新和应用程序更新等。

### 常见的反面模式：

- 对正在运行的基础设施资源实施就地更改。

### 建立此最佳实践的好处：

- 提高跨环境的一致性：由于不同环境的基础设施资源没有差异，因此提高了一致性并简化了测试。
- 减少配置偏差：通过使用已知且受版本控制的配置替换基础设施资源，可以将基础设施设置为已知、经过测试和可信的状态，从而避免配置偏差。

- 可靠的原子部署：部署要么成功完成，要么没有任何变化，从而提高部署过程的一致性和可靠性。
- 简化部署：由于无需支持升级，部署得到简化。升级即意味着新的部署。
- 采用快速回滚和恢复流程的更安全部署：由于之前运行的版本未发生更改，因此部署变得更安全。您可以在检测到错误时进行回滚。
- 增强安全状况：通过不允许更改基础设施，可以禁用远程访问机制（例如 SSH）。这样做可以减少攻击向量，改善您的组织的安全状况。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

### 自动化

在定义不可变基础设施部署策略时，建议尽可能使用[自动化](#)，来提高可重复性并最大限度地减少出现人为错误的可能性。有关更多详细信息，请参阅[REL08-BP05 使用自动化功能部署更改](#)和[自动执行安全、不需要人工介入的部署](#)。

使用[基础设施即代码 \(IaC\)](#)，基础设施预置、编排和部署步骤以编程、描述性和声明性方式定义，并存储在源代码管理系统中。利用基础设施即代码可以更轻松地自动化基础设施部署，并有助于实现基础设施的不可变性。

### 部署模式

当需要更改工作负载时，不可变基础设施部署策略要求部署一组新的基础设施资源，包括所有必要的更改。这组新资源必须遵循可最大限度地减少对用户影响的推出模式，这一点非常重要。此部署有两种主要策略：

[金丝雀部署](#)：将您的少量客户引导到新版本的做法，它通常在单个服务实例（金丝雀）上运行。然后，您可以深入检查生成的任何行为更改或错误。如果遇到了严重问题，您可以将 Canary 中的流量删除，并将用户发回到以前的版本。如果部署成功，您可以继续以期望的速度进行部署，同时监控更改以便发现错误，直到所有部署完成。AWS CodeDeploy 的[部署配置](#)可以配置为允许金丝雀部署。

[蓝绿部署](#)：与金丝雀部署类似，只是会并行部署一整套应用程序。您可以在两个堆栈（蓝和绿）之间轮流部署。同样，您可以将流量发送到新版本中，如果发现部署中存在问题，可以对其进行故障恢复，然后送回旧版本中。通常来说，所有流量会被一次性切换，但您也可以通过 Amazon Route 53 的加权 DNS 路由功能向每个版本发送部分流量，以加快采用新版本的速度。AWS CodeDeploy 和[AWS Elastic Beanstalk](#) 的部署配置可以配置为允许蓝绿部署。

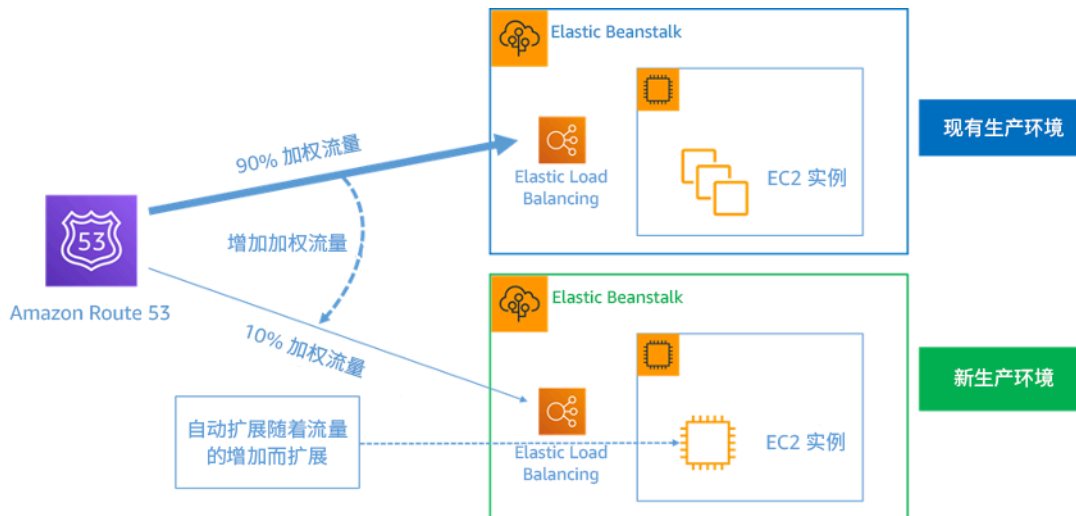


图 8：使用 AWS Elastic Beanstalk 和 Amazon Route 53 进行蓝绿部署

## 偏差检测

偏差是指导致基础设施资源的状态或配置不同于预期的任何更改。任何类型非受管配置更改都与不可变基础设施的概念背道而驰，因此应加以检测和修复，以便成功实施不可变基础设施。

## 实施步骤

- 禁止就地修改正在运行的基础设施资源。
  - 您可以使用 [AWS Identity and Access Management \(IAM\)](#) 来指定谁或什么可以访问 AWS 中的服务和资源，集中管理精细权限，并分析访问权限以细化 AWS 中的权限。
- 自动部署基础设施资源，以提高可重复性并最大限度地减少出现人为错误的可能性。
  - 正如《[AWS 上的 DevOps 简介](#)》白皮书中所述，自动化是 AWS 服务的基石，在所有服务、功能和产品中都受到内部支持。
  - [预先制作](#)您的亚马逊机器映像 (AMI) 可以加快启动它们的时间。[EC2 Image Builder](#) 是一项完全托管式 AWS 服务，可协助您自动创建、维护、验证、共享和部署自定义、安全且最新的 Linux 或 Windows 自定义 AMI。
- 一些支持自动化的服务包括：
  - [AWS Elastic Beanstalk](#) 这项服务可用于在 Apache、NGINX、Passenger 和 IIS 等熟悉的服务器上快速部署和扩展使用 Java、.NET、PHP、Node.js、Python、Ruby、Go 和 Docker 开发的 Web 应用程序。
  - [AWS Proton](#) 让平台团队能够连接和协调开发团队进行基础设施预置、代码部署、监控和更新所需的所有不同工具。AWS Proton 可实现自动化基础设施即代码预置，以及无服务器和基于容器的应用程序的部署。

- 利用基础设施即代码可以轻松实现基础设施部署的自动化，并有助于实现基础设施的不可变性。AWS 提供的服务可用于以编程、描述和声明的方式创建、部署和维护基础设施。
- [AWS CloudFormation](#) 让开发人员能够以有序和可预测的方式创建 AWS 资源。资源使用 JSON 或 YAML 格式写入文本文件。模板需要特定的语法和结构，具体取决于创建和管理的资源类型。您可以使用任何代码编辑器（例如 AWS Cloud9）以 JSON 或 YAML 格式编写资源，将其签入版本控制系统，然后 CloudFormation 会以安全、可重复的方式构建指定的服务。
- [AWS Serverless Application Model \( AWS SAM \)](#) 是一个开源框架，可用于在 AWS 上构建无服务器应用程序。AWS SAM 与其他 AWS 服务集成，是 AWS CloudFormation 的扩展。
- [AWS Cloud Development Kit \(AWS CDK\)](#) 是一个开源软件开发框架，用于使用熟悉的编程语言对您的云应用程序资源进行建模和预置。您可以使用 AWS CDK 通过 TypeScript、Python、Java 和 .NET 对应用程序基础设施进行建模。AWS CDK 在后台使用 AWS CloudFormation，以安全、可重复的方式提供资源。
- [AWS 云端控制 API](#) 引入了一组通用的创建、读取、更新、删除和列出（CRUDL）API，以协助开发人员以简单一致的方式管理其云基础设施。Cloud Control API 通用 API 允许开发人员统一管理 AWS 和第三方服务的生命周期。
- 实施能够最大限度减少对用户的影响的部署模式。
  - 金丝雀部署：
    - [设置 API Gateway 金丝雀版本部署](#)
    - [使用 AWS App Mesh 为 Amazon ECS 的金丝雀部署创建管道](#)
  - 蓝绿部署：[《AWS 上的蓝绿部署》白皮书](#)描述了实施蓝绿部署策略的[示例技术](#)。
- 检测配置或状态偏差。有关更多详细信息，请参阅[检测堆栈和资源的非受管配置更改](#)。

## 资源

### 相关最佳实践：

- [REL08-BP05 使用自动化功能部署更改](#)

### 相关文档：

- [自动执行安全、不需要人工介入的部署](#)
- [利用 AWS CloudFormation 在 Nubank 创建不可变的基础设施](#)
- [基础设施即代码](#)
- [实现警报以自动检测 AWS CloudFormation 堆栈中的偏差](#)



## 相关视频：

- [AWS re:Invent 2020：通过不可变性带来的可靠性、一致性和信心](#)

## REL08-BP05 使用自动化功能部署更改

自动部署与修补以消除负面影响。

对许多组织来说，对生产系统进行变更是风险最大的工作之一。除了软件解决的业务问题外，我们认为部署也是亟待解决的首要问题。如今，这意味着根据实际情况在操作中使用自动化，包括测试和部署更改、添加或删除容量以及迁移数据。AWS CodePipeline 让您可以管理释放您的工作负载所需的步骤。其中包括，采用 AWS CodeDeploy 将应用程序代码自动部署到 Amazon EC2 实例、本地实例、无服务器 Lambda 函数或 Amazon ECS 服务的部署状态。

### 推荐

虽然传统智慧告诉我们，循环中最困难的操作程序应该由人来负责，但出于相同的原因，我们建议您将最困难的程序自动化。

## 常见反模式：

- 手动执行更改。
- 通过紧急工作流程跳过自动化中的步骤。
- 未遵守您的计划。

建立此最佳实践的好处：通过自动化功能部署所有更改，可消除引入人为错误的可能性，还能在更改生产之前进行测试，从而确保计划完成。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

- 实现部署管道的自动化。借助部署管道，您可以调用自动化测试和异常检测，并且能够在生产部署前的某个步骤停止管道，或自动回滚更改。
  - [Amazon Builders' Library：确保部署期间安全回滚](#)
  - [Amazon Builders' Library：采用持续交付，加速交付进度](#)
    - 使用 AWS CodePipeline（或受信任的第三方产品）定义和运行您的管道。

- 将管道配置为在将更改实施到代码存储库后开始。
- [什么是 AWS CodePipeline ?](#)
- 使用 Amazon Simple Notification Service ( Amazon SNS ) 和 Amazon Simple Email Service ( Amazon SES ) 在管道中发送有关问题的通知，或者与 Amazon Chime 等团队聊天工具集成。
- [什么是 Amazon Simple Notification Service ?](#)
- [什么是 Amazon SES ?](#)
- [什么是 Amazon Chime ?](#)
- [使用 Webhook 自动发送聊天消息。](#)

## 资源

### 相关文档：

- [AWS 合作伙伴：可以帮助您创建自动化部署解决方案的合作伙伴](#)
- [AWS Marketplace：可用于自动实施部署的产品](#)
- [使用 Webhook 自动发送聊天消息。](#)
- [Amazon Builders' Library：确保部署期间安全回滚](#)
- [Amazon Builders' Library：采用持续交付，加速交付进度](#)
- [什么是 AWS CodePipeline ?](#)
- [什么是 CodeDeploy ?](#)
- [AWS Systems Manager 补丁管理器](#)
- [什么是 Amazon SES ?](#)
- [什么是 Amazon Simple Notification Service ?](#)

### 相关视频：

- [2019 年 AWS 峰会：AWS 上的 CI/CD](#)

## 故障管理

### 问题

- [REL 9.如何备份数据 ?](#)

- [REL 10.如何使用故障隔离来保护您的工作负载？](#)
- [REL 11.如何将您的工作负载设计为可承受组件故障的影响？](#)
- [REL 12.如何测试可靠性？](#)
- [REL 13.如何规划灾难恢复（DR）？](#)

## REL 9.如何备份数据？

备份数据、应用程序和配置，以满足恢复时间目标（RTO）和恢复点目标（RPO）的要求。

### 最佳实践

- [REL09-BP01 识别和备份需要备份的所有数据，或从源复制数据](#)
- [REL09-BP02 保护并加密备份](#)
- [REL09-BP03 自动执行数据备份](#)
- [REL09-BP04 定期执行数据恢复以验证备份完整性和流程](#)

### REL09-BP01 识别和备份需要备份的所有数据，或从源复制数据

了解并使用工作负载所用的数据服务和资源的备份功能。大多数服务提供了备份工作负载数据的功能。

期望结果：数据来源已确定，并根据重要性进行了分类。然后，根据 RPO 为数据恢复建立了策略。此策略涉及到备份这些数据来源，或者能够从其他来源复制数据。在出现数据丢失的情况下，所实施的策略可以在定义的 RPO 和 RTO 内实现数据的恢复或复制。

云成熟度阶段：基础

常见反模式：

- 不了解工作负载的所有数据来源及其重要性。
- 没有对关键数据来源进行备份。
- 仅对部分数据来源进行备份，但没有考虑重要性标准。
- 没有定义 RPO，或者备份频率无法满足 RPO。
- 没有评估备份是否必需或者是否可以从其他来源复制数据。

建立此最佳实践的好处：确定需要备份的位置并实施某种机制来创建备份，或者具备从外部来源复制数据的能力，这样可以提高在停机期间还原和恢复数据的能力。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

所有 AWS 数据存储均提供备份功能。Amazon RDS 和 Amazon DynamoDB 等服务还额外地支持可实现时间点故障恢复 ( PITR ) 的自动备份，这使您可以将备份恢复到距当前时间不超过五分钟的任意时间点。许多 AWS 服务提供了将备份复制到其他 AWS 区域的功能。AWS Backup 工具向您提供了在不同 AWS 服务中集中实现自动化数据保护的能力。[AWS Elastic Disaster Recovery](#) 使您可以从本地、跨可用区或跨区域复制完整的服务器工作负载并保持连续数据保护，恢复点目标 ( RPO ) 以秒为单位。

Amazon S3 可用作自行管理数据来源和 AWS 托管数据来源的备份目标。Amazon EBS、Amazon RDS 和 Amazon DynamoDB 等 AWS 服务具有可用于创建备份的内置功能。此外，也可使用第三方备份软件。

可以使用 [AWS Storage Gateway](#) 或 [AWS DataSync](#) 将本地数据备份到 AWS Cloud。Amazon S3 存储桶可用于在 AWS 中存储此数据。Amazon S3 提供多个存储层 ( 例如 [Amazon S3 Glacier](#) 或 [S3 Glacier Deep Archive](#) )，可用于降低数据存储的成本。

您可以从其他来源复制数据，以此来满足数据恢复需求。例如，[Amazon ElastiCache 副本节点](#) 或 [Amazon RDS 只读副本](#) 可用于在主来源丢失时复制数据。如果像这样的来源可用于满足 [恢复点目标 \( RPO \)](#) 和 [恢复时间目标 \( RTO \)](#) 要求，您可能不需要备份。在另一个例子中，如果使用 Amazon EMR，只要可以 [将数据从 Amazon S3 复制到 Amazon EMR 中](#)，则可能不需要备份 HDFS 数据存储。

在选择备份策略时，请考虑恢复数据所用的时间。恢复数据所需的时间取决于备份的类型 ( 在采用备份策略时 ) 或数据复制机制的复杂性。此时间应该符合工作负载的 RTO。

## 实施步骤

1. 确定工作负载的所有数据来源。数据可以存储在多种资源中，例如[数据库](#)、[卷](#)、[文件系统](#)、[日志记录系统](#)和[对象存储](#)。请参阅资源部分，查找有关存储数据的不同 AWS 服务的相关文档，以及这些服务提供的备份功能。
2. 根据重要性对数据来源进行分类。对于工作负载，不同数据集具有不同的重要程度，因此对弹性具有不同的要求。例如，一些数据可能会非常重要，要求接近于零的 RPO，而另一些数据则不那么重要，可以承受较高的 RPO 和某种程度的数据丢失。与此类似，不同数据集也可能会有不同的 RTO 要求。
3. 使用 AWS 或第三方服务来创建数据的备份。[AWS Backup](#) 是一项托管服务，支持在 AWS 上创建各种数据源的备份。[AWS Elastic Disaster Recovery](#) 处理到 AWS 区域的自动亚秒级数据复制。大多数 AWS 服务还具有原生的创建备份功能。AWS Marketplace 有许多解决方案同样提供了这些功能。请参阅下面所列的资源，了解如何从不同 AWS 服务创建数据备份的信息。

4. 对于没有备份的数据，请建立数据复制机制。您可能会出于各种原因，不对可从其他来源复制的数据进行备份。您可能会遇到一种情况，在需要时从来源复制数据的成本相比创建备份更低，因为可能会有与存储备份相关的成本。另一个例子是从备份进行还原的时间比从来源复制数据用时更长，使得备份不符合 RTO 要求。在此类情况下请做出权衡，并建立明确定义的流程，确定在需要进行恢复时如何从这些来源复制数据。例如，如果您从 Amazon S3 将数据加载到数据仓库（如 Amazon Redshift）或 MapReduce 集群（如 Amazon EMR），以便对此类数据进行分析，这就是可从其他来源复制数据的例子。只要此类分析的结果被存储在某位置，或者可重现，您不会因为数据仓库或 MapReduce 集群故障而遭遇数据丢失的情况。其他可从数据源复制数据的例子包括，缓存（如 Amazon ElastiCache）或 RDS 只读副本。
5. 建立备份数据的频率。创建数据来源的备份是一个定期执行的流程，其频率取决于 RPO。

实施计划的工作量级别：适中

资源

相关最佳实践：

[REL13-BP01 定义停机和数据丢失的恢复目标](#)

[REL13-BP02 使用定义的恢复策略来实现恢复目标](#)

相关文档：

- [什么是 AWS Backup？](#)
- [什么是 AWS DataSync？](#)
- [什么是卷网关？](#)
- [AWS 合作伙伴：可以帮助进行备份的合作伙伴](#)
- [AWS Marketplace：可以用于备份的产品](#)
- [Amazon EBS 快照](#)
- [备份 Amazon EFS](#)
- [备份 Amazon FSx for Windows File Server](#)
- [ElastiCache for Redis 备份和还原](#)
- [在 Neptune 中创建数据库集群快照](#)
- [创建数据库快照](#)
- [创建按计划触发的 EventBridge 规则](#)

- [使用 Amazon S3 进行跨区域复制](#)
- [EFS 到 EFS AWS Backup](#)
- [将日志数据导出到 Amazon S3](#)
- [对象生命周期管理](#)
- [DynamoDB 的按需备份和还原](#)
- [DynamoDB 的时间点恢复](#)
- [使用 Amazon OpenSearch Service 索引快照](#)
- [什么是 AWS Elastic Disaster Recovery ?](#)

相关视频：

- [AWS re:Invent 2021 – 使用 AWS 进行备份、灾难恢复和勒索软件防护](#)
- [AWS Backup 演示：跨账户和跨区域备份](#)
- [AWS re:Invent 2019：深入了解 AWS Backup，主讲：Rackspace\(STG341\)](#)

相关示例：

- [Well-Architected 实验室 - 为 Amazon S3 实施双向跨区域复制 \( CRR \)](#)
- [Well-Architected 实验室 - 测试数据的备份与还原](#)
- [Well-Architected 实验室 - 面向分析工作负载的备份和还原 \( 具备失效自动恢复功能 \)](#)
- [Well-Architected 实验室 - 灾难恢复 - 备份与还原](#)

## REL09-BP02 保护并加密备份

使用身份验证和授权来控制并检测对备份的访问。使用加密功能，防止并检测备份的数据完整性是否受到损坏。

常见反模式：

- 对备份和还原自动化的访问权限与对数据的访问权限相同。
- 未加密您的备份。

建立此最佳实践的好处：保护备份安全可防止篡改数据，而加密数据可防止数据意外暴露时对其访问。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

使用 AWS Identity and Access Management ( IAM ) 等身份验证和授权服务，控制并检测对备份的访问。使用加密功能，防止并检测备份的数据完整性是否受到损坏。

Amazon S3 支持多种对您的静态数据进行加密的方式。借助服务器端加密功能，Amazon S3 以未加密数据的形式接受您的对象，然后在存储此类数据时进行加密。若采用客户端加密，您的工作负载应用程序需要负责在将其发送到 Amazon S3 之前加密数据。这两种方式都让您可以使用 AWS Key Management Service ( AWS KMS ) 创建并存储数据密钥，或者您也可以提供自己的密钥并自行对其负责。使用 AWS KMS，您可以通过 IAM 设置策略，决定谁可以以及谁不可以访问您的数据密钥与解密数据。

针对 Amazon RDS，如果您已选择对数据库进行加密，那么您的备份也会被加密。DynamoDB 备份则始终被加密。使用 AWS Elastic Disaster Recovery 时，加密所有传输中数据和静态数据。借助 Elastic Disaster Recovery，可以使用默认 Amazon EBS 加密 Volume Encryption Key 或自定义客户管理的密钥来加密静态数据。

## 实施步骤

1. 对每个数据存储使用加密。如果源数据已加密，则备份也将被加密。
  - [在 Amazon RDS 中使用加密](#)。当您创建 RDS 实例时，可以使用 AWS Key Management Service 配置静态加密。
  - [在 Amazon EBS 卷上使用加密](#)。您可以配置默认加密或在创建卷时指定唯一密钥。
  - 使用所需的 [Amazon DynamoDB 加密](#)。DynamoDB 会加密所有静态数据。您可以使用 AWS 拥有的 AWS KMS 密钥或者 AWS 托管 KMS 密钥，指定存储在您账户中的密钥。
  - [加密 Amazon EFS 中存储的数据](#)。在创建文件系统时配置加密。
  - 在源和目标区域中配置加密。您可以使用 KMS 中存储的密钥配置 Amazon S3 中的静态加密，但这些密钥是特定于区域的。您在配置复制时可以指定目标密钥。
  - 选择是为 Elastic Disaster Recovery 使用默认还是自定义 [Amazon EBS 加密](#)。使用此选项会加密暂存区域子网磁盘和复制磁盘上的已复制静态数据。
2. 实施用于访问您的备份的最低权限。请遵循最佳实践，根据[安全最佳实践](#)来限制对备份、快照和副本的访问。

## 资源

### 相关文档：

- [AWS Marketplace：可以用于备份的产品](#)



- [Amazon EBS 加密](#)
- [Amazon S3：利用加密保护数据](#)
- [CRR 附加配置：复制通过存储在 AWS KMS 中的加密密钥、使用服务器端加密 \( SSE \) 创建的对象](#)
- [DynamoDB 静态加密](#)
- [加密 Amazon RDS 资源](#)
- [在 Amazon EFS 中加密数据和元数据](#)
- [AWS 中的备份的加密](#)
- [管理加密的表](#)
- [安全性支柱 – AWS Well-Architected Framework](#)
- [什么是 AWS Elastic Disaster Recovery？](#)

相关示例：

- [Well-Architected 实验室 - 为 Amazon S3 实施双向跨区域复制 \( CRR \)](#)

## REL09-BP03 自动执行数据备份

将备份配置为根据遵循恢复点目标 ( RPO ) 的定期计划自动备份，或者在数据集发生更改时自动备份。具有低数据丢失需求的关键数据资产需要频繁地自动备份，而可以接受某些丢失的较不重要数据的备份频率可以更低。

期望结果：按照确定的节奏创建数据来源备份的自动流程。

常见反模式：

- 手动执行备份。
- 使用具有备份功能的资源，但不包括自动化中的备份。

建立此最佳实践的好处：自动化备份可以确保按照 RPO 执行备份，并在未备份时发出警报。

在未建立这种最佳实践的情况下暴露的风险等级：中等

实施指导

AWS Backup 可用于创建各种 AWS 数据来源的自动数据备份。Amazon RDS 实例可以按照五分钟频率进行几乎连续的备份，Amazon S3 对象可以按照十五分钟频率进行几乎连续的备份，提供可

恢复到备份历史记录中的特定时间点的时间点故障恢复 ( PITR ) 功能。对于其他 AWS 数据来源 ( 如 Amazon EBS 卷、Amazon DynamoDB 表或 Amazon FSx 文件系统 ) , AWS Backup 最快可以按每小时的频率运行自动备份。这些服务还提供了原生备份功能。以下 AWS 服务提供了具备时间点故障恢复的自动备份功能 : [Amazon DynamoDB](#)、[Amazon RDS](#) 和 [Amazon Keyspaces \( Apache Cassandra 兼容 \)](#) - 这些备份可以恢复到备份历史记录中的特定时间点。大部分其他 AWS 数据存储服务提供了计划定期备份的功能 , 频率最快为每小时一次。

Amazon RDS 和 Amazon DynamoDB 提供支持时间点恢复的持续备份。一旦启用 , Amazon S3 版本控制即是自动的。[Amazon Data Lifecycle Manager](#) 可用于自动拍摄、复制和删除 Amazon EBS 快照。它还可以自动创建、复制、弃用和取消注册 Amazon EBS 支持的亚马逊云机器镜像 ( AMI ) 及其底层 Amazon EBS 快照。

AWS Elastic Disaster Recovery 提供从源环境 ( 本地或 AWS ) 到目标恢复区域的持续、块级复制。服务会自动创建和管理时间点 Amazon EBS 快照。

针对您的备份自动化和历史的集中式视图 , AWS Backup 提供完全托管的基于策略的备份解决方案。它会使用 AWS Storage Gateway 将云端和本地的多项 AWS 服务的数据备份集中在一起并自动处理。

除了版本控制 , Amazon S3 还具有复制功能。整个 S3 存储桶都可自动复制到相同或不同 AWS 区域中的其他存储桶。

## 实施步骤

1. 确定当前在手动备份的数据来源。有关更多详细信息 , 请参阅 [REL09-BP01 识别和备份需要备份的所有数据 , 或从源复制数据](#)。
2. 确定工作负载的 RPO。有关更多详细信息 , 请参阅 [REL13-BP01 定义停机和数据丢失的恢复目标](#)。
3. 使用自动化备份解决方案或托管服务。AWS Backup 是一项完全托管式服务 , 可让您在云端和本地[对不同 AWS 服务中的数据保护实现集中化和自动化](#)。使用 AWS Backup 中的备份计划 , 创建规则来定义要备份的资源 , 以及创建这些备份的频率。此频率应遵循在第 2 步中确定的 RPO。有关如何使用 AWS Backup 创建自动备份的动手实践指导 , 请参阅[测试数据的备份和恢复](#)。用于存储数据的大多数 AWS 服务提供了原生备份功能。例如 , 可以利用 RDS 来实现支持时间点故障恢复 ( PITR ) 的自动备份。
4. 对于自动备份解决方案或托管服务不支持的数据来源 ( 如本地数据来源或消息队列 ) , 请考虑使用受信任的第三方解决方案来创建自动备份。或者 , 您可以使用 AWS CLI 或开发工具包创建自动化过程来完成此操作。您可以使用 AWS Lambda 函数或 AWS Step Functions 来定义创建数据备份中涉及的逻辑 , 并使用 Amazon EventBridge 按照基于 RPO 确定的频率来执行它。

实施计划的工作量级别 : 低

## 资源

### 相关文档：

- [AWS 合作伙伴：可以帮助进行备份的合作伙伴](#)
- [AWS Marketplace：可以用于备份的产品](#)
- [创建按计划触发的 EventBridge 规则](#)
- [什么是 AWS Backup？](#)
- [什么是 AWS Step Functions？](#)
- [什么是 AWS Elastic Disaster Recovery？](#)

### 相关视频：

- [AWS re:Invent 2019：深入了解 AWS Backup，主讲：Rackspace \(STG341\)](#)

### 相关示例：

- [Well-Architected 实验室 - 测试数据的备份与还原](#)

## REL09-BP04 定期执行数据恢复以验证备份完整性和流程

通过执行恢复测试，验证您的备份流程实施是否满足恢复时间目标（RTO）和恢复点目标（RPO）要求。

期望结果：使用明确定义的机制定期从备份恢复数据，确认可以按照为工作负载确定的恢复时间目标（RTO）来恢复数据。验证从备份进行还原可以得到包含原始数据的资源，而不会造成数据损坏或无法访问数据，并且数据丢失在恢复点目标（RPO）之内。

### 常见反模式：

- 还原备份，但未查询或检索任何数据以确认还原操作可用。
- 假定备份存在。
- 假定系统的备份完全正常运行，并且可从中恢复数据。
- 假定从备份还原或恢复数据的时间满足工作负载的 RTO。
- 假定备份中包含的数据符合工作负载的 RPO
- 需要时进行还原，没有使用运行手册或者没有按照确定的自动程序执行。

建立此最佳实践的好处：测试备份的恢复过程可以确认在需要时能够将数据还原，不必担心数据可能丢失或损坏，可以按照工作负载要求的 RTO 还原和恢复，并且任何数据丢失都符合工作负载的 RPO。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

测试备份和还原功能可树立信心，确信能够在出现中断时执行这些操作。定期将备份还原到新的位置，并运行测试以验证数据的完整性。应该执行一些常见的测试，以核实所有数据是否均可用、未损坏、可访问且任何数据丢失都符合工作负载的 RPO。此类测试还可以帮助确定恢复机制是否足够快以满足工作负载的 RTO 要求。

使用 AWS，您可以构建一个测试环境，还原您的备份以评估 RTO 和 RPO 功能，并且对数据的内容和完整性执行测试。

此外，Amazon RDS 和 Amazon DynamoDB 还允许时间点恢复（PITR）。您可以使用持续备份将您的数据集还原到其在指定日期与时间所处的状态。

数据是否可用、没有损坏、是否可以访问并且任意数据丢失都符合工作负载的 RPO。此类测试还可以帮助确定恢复机制是否足够快以满足工作负载的 RTO 要求。

AWS Elastic Disaster Recovery 提供 Amazon EBS 卷的持续时间点恢复快照。复制源服务器时，根据配置的策略记录一段时间内的时间点状态。Elastic Disaster Recovery 可以启动实例用于测试和演练，而不重定向流量，从而帮助您验证这些快照的完整性。

## 实施步骤

1. 确定当前备份的数据来源以及存储这些备份的位置。有关实施指导，请参阅 [REL09-BP01 识别和备份需要备份的所有数据，或从源复制数据](#)。
2. 为每个数据来源建立数据验证标准。不同类型的数据具有不同的属性，这可能需要不同的验证机制。在您确信将此数据用于生产之前，请考虑可以如何验证此数据。一些验证数据的常见方法包括使用数据和备份属性，例如数据类型、格式、校验和、大小，或者将这些属性与自定义的验证逻辑结合使用。例如，可以将所恢复资源的校验和值，与创建备份时数据来源的校验和值进行比较。
3. 建立 RTO 和 RPO，以根据数据重要性来还原数据。有关实施指导，请参阅 [REL13-BP01 定义停机和数据丢失的恢复目标](#)。
4. 评估恢复能力。检查备份和还原策略，了解它是否可以满足您的 RTO 和 RPO，并根据需要调整策略。使用 [AWS 韧性监测中心](#)，您可以对工作负载运行评估。该评估根据弹性策略评估您的应用程序配置，并报告是否能够满足 RTO 和 RPO 目标。
5. 使用当前为生产环境中数据还原所确立的流程执行测试还原。这些流程依赖于对原始数据来源进行备份的方法，备份本身的格式和存储位置，或者数据是否从其他来源复制。例如，如果您使用 [AWS](#)

[Backup](#) 等托管服务，则此过程可能就是简单地将备份还原到新的资源。如果您使用 AWS Elastic Disaster Recovery，则可以[启动恢复演练](#)。

6. 根据您之前为数据验证确立的标准，从还原后的资源验证数据恢复。还原和恢复的数据是否包含备份时的最新记录或项目？此数据是否在工作负载的 RPO 之内？
7. 测量还原和恢复所需的时间，并与确立的 RTO 进行比较。此流程是否符合工作负载的 RTO？例如，比较还原过程开始时的时间戳以及恢复验证完成时的时间戳，以计算此过程的用时。所有 AWS API 调用均有时间戳，此信息在 [AWS CloudTrail](#) 中提供。虽然此信息可以提供还原过程何时开始的详细信息，但验证完成时的结束时间戳应该由您的验证逻辑来记录。如果使用自动化过程，则 [Amazon DynamoDB](#) 等服务可用于存储此信息。此外，许多 AWS 服务提供了事件历史记录，其中可提供发生特定操作时的时间戳信息。在 AWS Backup 中，备份和还原操作称为作业，这些作业在其元数据中包含时间戳信息，可用于测量还原和恢复所需的时间。
8. 如果数据验证失败，或者如果还原和恢复所需的时间超过了为工作负载设定的 RTO，则通知利益攸关方。在实施自动化以完成此操作时（[例如在本实验中](#)），可以使用 Amazon Simple Notification Service（Amazon SNS）等服务将推送通知（例如电子邮件或短信）发送给利益攸关方。[这些消息还可以发布到消息传递应用程序，例如 Amazon Chime、Slack 或 Microsoft Teams](#)，或用于[使用 AWS Systems Manager OpsCenter 来创建 OpsItems 等任务](#)。
9. 自动执行此流程以便定期运行。例如，AWS Lambda 等服务或 AWS Step Functions 中的状态机可用于自动完成还原和恢复流程，Amazon EventBridge 可用于定期触发此自动工作流，如以下架构图中所示。了解如何[使用 AWS Backup 自动完成数据恢复验证](#)。此外，[这个 Well-Architected 实验室](#)提供动手实践体验，可用于练习针对此处的多个步骤实现自动化的方法。

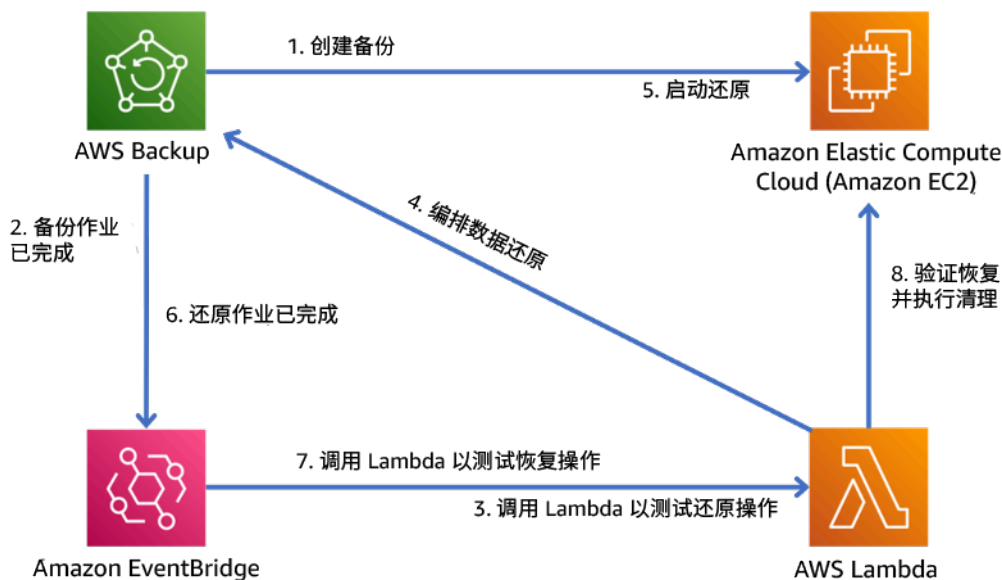


图 9.自动化的备份和还原流程

实施计划的工作量级别：中到高，具体取决于验证条件的复杂性。

## 资源

相关文档：

- [使用 AWS Backup 自动完成数据恢复验证](#)
- [AWS 合作伙伴：可以帮助进行备份的合作伙伴](#)
- [AWS Marketplace：可以用于备份的产品](#)
- [创建按计划触发的 EventBridge 规则](#)
- [DynamoDB 的按需备份和还原](#)
- [什么是 AWS Backup？](#)
- [什么是 AWS Step Functions？](#)
- [什么是 AWS Elastic Disaster Recovery](#)
- [AWS Elastic Disaster Recovery](#)

相关示例：

- [Well-Architected 实验室：测试数据的备份与还原](#)

## REL 10.如何使用故障隔离来保护您的工作负载？

故障隔离边界可将一个工作负载内的故障影响限制于有限数量的组件。边界以外的组件不会受到故障的影响。使用多个故障隔离边界，您可以限制作用于您的工作负载的影响。

### 最佳实践

- [REL10-BP01 将工作负载部署到多个位置](#)
- [REL10-BP02 为您的多位置部署选择合适的位置](#)
- [REL10-BP03 组件的自动恢复受限于单个位置](#)
- [REL10-BP04 采用隔板架构来限制影响范围](#)

### REL10-BP01 将工作负载部署到多个位置

将工作负载数据和资源分布到多个可用区，或在必要时分布到多个 AWS 区域。可通过选择不同位置满足各种需求。



在 AWS，服务设计的其中一个基本原则是避免底层物理基础设施中存在单点故障。这促使我们构建使用多个可用区并能灵活应对单个可用区故障的软件和系统。同样，系统也被构建可灵活应对单个计算节点、单个存储卷或单个数据库实例故障。构建依赖冗余组件的系统时，务必要确保组件独立运行；如果是 AWS 区域，组件应自主运行。只有实现了这一点，采用冗余组件的理论可用性计算的优点才能发挥作用。

## 可用区 ( AZ , Availability Zone )

AWS 区域由在设计上彼此相互独立的多个可用区组成。每个可用区之间都间隔相当的物理距离，以避免因环境公害（如火灾、洪水和龙卷风）导致的相互关联的故障情况。每个可用区还拥有独立的物理基础设施：专用的公用电源连接、独立备份电源、独立机械服务以及可用区内外的独立网络连接。此设计将任意这些系统的故障限制在受影响的那一个 AZ 中。尽管可用区在地理位置上相互分离，但它们位于相同的区域中，从而实现高吞吐量、低延迟的联网。整个 AWS 区域（跨多个可用区，由多个物理上独立的设计中心组成）可以视为工作负载的单个逻辑部署目标，包括同步复制数据（例如，在两个数据库之间）的能力。这样一来，您便能在主动/主动或主动/备用配置中使用可用区。

可用区是独立的，因此当工作负载采用了使用多个可用区的架构时，可以提高工作负载的可用性。一些 AWS 服务（包括 Amazon EC2 实例数据面板）作为严格的区级别服务部署，与其所在的可用区共存亡。其他 AZ 中的 Amazon EC2 实例不受影响，可以继续正常工作。与此类似，如果某个可用区中的故障导致 Amazon Aurora 数据库失败，则不受影响的 AZ 中的只读副本 Aurora 实例可以自动提升为主实例。另一方面，区域性 AWS 服务（例如 Amazon DynamoDB）在内部以主动/主动配置的形式使用多个可用区，以实现为该服务设定的可用性设计目标，而且无需您配置 AZ 置放。

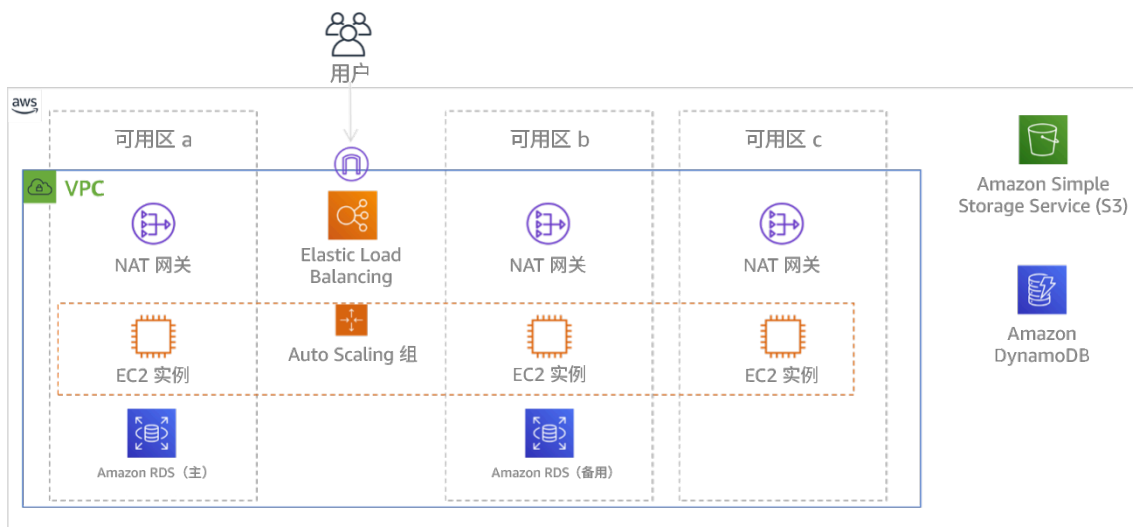


图 9：跨三个可用区部署多层架构。请注意，Amazon S3 和 Amazon DynamoDB 始终会自动部署到多个可用区。而 ELB 也会被部署到所有三个区。



虽然 AWS 控制面板通常提供在整个区域（多个可用区）内管理资源的功能，但某些控制面板（包括 Amazon EC2 和 Amazon EBS）能够将结果筛选到单个可用区。完成筛选后，请求仅在指定可用区中进行处理，从而降低其他可用区的中断风险。此 AWS CLI 示例演示仅从 us-east-2c 可用区中获取 Amazon EC2 实例信息：

```
AWS ec2 describe-instances --filters Name=availability-zone,Values=us-east-2c
```

## AWS Local Zones

AWS Local Zones 在其各自的 AWS 区域内的作用与可用区相似，它们可被选择作为区级别 AWS 资源（如子网和 EC2 实例）的置放位置。特别之处在于，它们并不位于相关的 AWS 区域内，而是靠近目前还未设置 AWS 区域的人口密集的工业和 IT 中心。但是，您还是可以享受高带宽，并且能够在本地区域的本地工作负载与在 AWS 区域内运行的工作负载之间进行安全连接。您应该利用 AWS Local Zones 将工作负载尽量部署在接近用户的地方，以满足低延迟的要求。

## Amazon 全球边缘网络

Amazon 全球边缘网络由全球各大城市的边缘站点组成。Amazon CloudFront 使用此网络以较低的延迟向最终用户分发内容。您可以通过 AWS Global Accelerator 在这些边缘站点创建您的工作负载端点，以便在靠近您的用户的 AWS 全球网络进行接入。利用 Amazon API Gateway，使用 CloudFront 分配的边缘优化 API 端点可以通过最近的边缘站点方便客户端访问。

## AWS 区域

AWS 区域采用自主设计，因此通过多区域方法，您可以将服务的专用副本部署到每个区域。

多区域方法对于灾难恢复策略很常见，用于在偶发的大规模事件中满足恢复目标。请参阅 [灾难恢复 \(DR\) 计划](#) 以了解有关这些策略的更多信息。不过，这里我们的重点是可用性，旨在达成长期使用中的平均正常运行时间目标。对于高可用性目标，通常可以将多区域架构设计为主动/主动模式，各个服务副本（在其各自的区域中）处于活动状态（为请求提供服务）。

### 推荐

大多数工作负载的可用性目标都可通过在单个 AWS 区域内采用多 AZ 策略来实现。只有当工作负载具有极高的可用性要求或者其他业务目标时，才考虑多区域架构，在这些情况下需要使用多区域架构。

AWS 提供了跨区域运行服务的功能。例如，AWS 使用 Amazon Simple Storage Service ( Amazon S3 ) 复制、Amazon RDS 只读副本 ( 包括 Aurora 只读副本 ) 和 Amazon DynamoDB 全局表，提供了连续异步数据复制功能。通过连续复制，您的数据版本可近乎实时地供各个活动区域使用。

使用 AWS CloudFormation，您可以跨 AWS 账户和 AWS 区域定义基础设施并一致地进行部署。AWS CloudFormation StackSets 扩展了此功能，允许您通过单个操作，跨多个账户和区域创建、更新或删除 AWS CloudFormation 堆栈。对于 Amazon EC2 实例部署，亚马逊云机器镜像 ( AMI，Amazon Machine Image ) 可用于提供诸如硬件配置和已安装软件等信息。您可以实施 Amazon EC2 Image Builder 管道来创建所需的 AMI，并将这些 AMI 复制到您的活动区域。这可以确保这些 Golden AMI 具有您需要部署的所有内容，并可在各个新区域中扩展您的工作负载。

对于路由流量，Amazon Route 53 和 AWS Global Accelerator 均可定义策略来确定哪些用户转向哪个活动的区域端点。使用 Global Accelerator，您可以设置流量转盘，控制导向各个应用程序端点的流量的百分比。Route 53 支持这种百分比方法，还有多个其他策略可用，包括基于地理位置距离和延迟的策略。Global Accelerator 自动利用 AWS 边缘服务器广泛的网络，尽可能快地将流量载入到 AWS 主干网，从而得到较低的请求延迟。

所有这些功能在执行时，都保留了各个区域的自主性。这种方法有极少的例外，包括我们提供全球边缘交付的服务 ( 例如 Amazon CloudFront 和 Amazon Route 53 )，以及 AWS Identity and Access Management ( IAM ) 服务的控制面板。大多数服务都完全在单个区域中运行。

## 本地数据中心

对于在本地数据中心运行的工作负载来说，尽可能打造混合体验。AWS Direct Connect 提供从您的本地到 AWS 的专用网络连接，使您可以同时在两者中运行。

另一个选项是，通过 AWS Outposts 在本地运行 AWS 基础设施和服务。AWS Outposts 是一项完全托管式服务，可将 AWS 基础设施、AWS 服务、API 和工具延伸到您的数据中心。在设计中心会安装与 AWS Cloud 中使用的相同硬件基础设施。然后，AWS Outposts 会连接到最近的 AWS 区域。您可以使用 AWS Outposts 支持您的低延迟工作负载，或满足本地数据处理要求。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

- 使用多个可用区和 AWS 区域。将工作负载数据和资源分布到多个可用区，或在必要时分布到多个 AWS 区域。可通过选择不同位置满足各种需求。
- 区域性服务本质上是跨多个可用区部署的。
  - 这包括 Amazon S3、Amazon DynamoDB 和 AWS Lambda ( 未连接到 VPC 时 )

- 将容器、实例和基于功能的工作负载部署到多个可用区中。使用包括缓存在内的多可用区数据存储。使用 EC2 Auto Scaling、ECS 任务置放、AWS Lambda 函数配置（在 VPC 中运行时）和 ElastiCache 集群的功能。
- 部署 Auto Scaling 组时，请使用单独可用区中的子网。
  - [示例：跨多个可用区分布实例](#)
  - [Amazon ECS 任务置放策略](#)
  - [配置 AWS Lambda 函数以访问 Amazon VPC 中的资源](#)
  - [选择区域和可用区](#)
- 部署 Auto Scaling 组时，请使用单独可用区中的子网。
  - [示例：跨多个可用区分布实例](#)
- 使用 ECS 任务置放参数，并指定数据库子网组。
  - [Amazon ECS 任务置放策略](#)
- 配置要在 VPC 中运行的函数时，请使用多个可用区中的子网。
  - [配置 AWS Lambda 函数以访问 Amazon VPC 中的资源](#)
- 将多个可用区与 ElastiCache 集群配合使用。
  - [选择区域和可用区](#)
- 如果您的工作负载必须部署到多个区域，请选择一个多区域策略。大多数可靠性需求可通过在单个 AWS 区域中使用多可用区策略来满足。可在必要时使用多区域策略来满足您的业务需求。
  - [AWS re:Invent 2018：适用于多区域主动-主动应用程序的架构模式 \(ARC209-R2\)](#)
    - 备份到另一个 AWS 区域可以让您更加确信，数据在需要时可用。
    - 有些工作负载具有法规要求，需要使用多区域策略。
- 评估您工作负载的 AWS Outposts。如果您的工作负载需要到本地部署数据中心的较低延迟，或具有本地数据处理要求，然后使用 AWS Outposts 在本地运行 AWS 基础设施和服务
  - [什么是 AWS Outposts？](#)
- 确定 AWS Local Zones 是否可以帮助您为用户提供服务。如果您有低延迟要求，请查看 AWS Local Zones 是否距离您的用户较近。如果是，则使用它将工作负载部署到离这些用户较近的位置。
  - [AWS Local Zones 常见问题](#)

## 资源

### 相关文档：

- [AWS 全球基础设施](#)

- [AWS Local Zones 常见问题](#)
- [Amazon ECS 任务置放策略](#)
- [选择区域和可用区](#)
- [示例：跨多个可用区分布实例](#)
- [全局表：使用 DynamoDB 的多区域复制](#)
- [使用 Amazon Aurora 全局数据库](#)
- [使用 AWS 服务创建多区域应用程序博客系列](#)
- [什么是 AWS Outposts？](#)

相关视频：

- [AWS re:Invent 2018：适用于多区域主动-主动应用程序的架构模式（ARC209-R2）](#)
- [AWS re:Invent 2019：AWS 全球网络基础设施的创新与运营（NET339）](#)

REL10-BP02 为您的多位置部署选择合适的位置

期望结果

要实现高可用性，请始终（在可能时）将您的工作负载组件部署到多个可用区（AZ，Availability Zone），如图 10 中所示。对于具有极高弹性要求的工作负载，请谨慎评估用于多区域架构的选项。

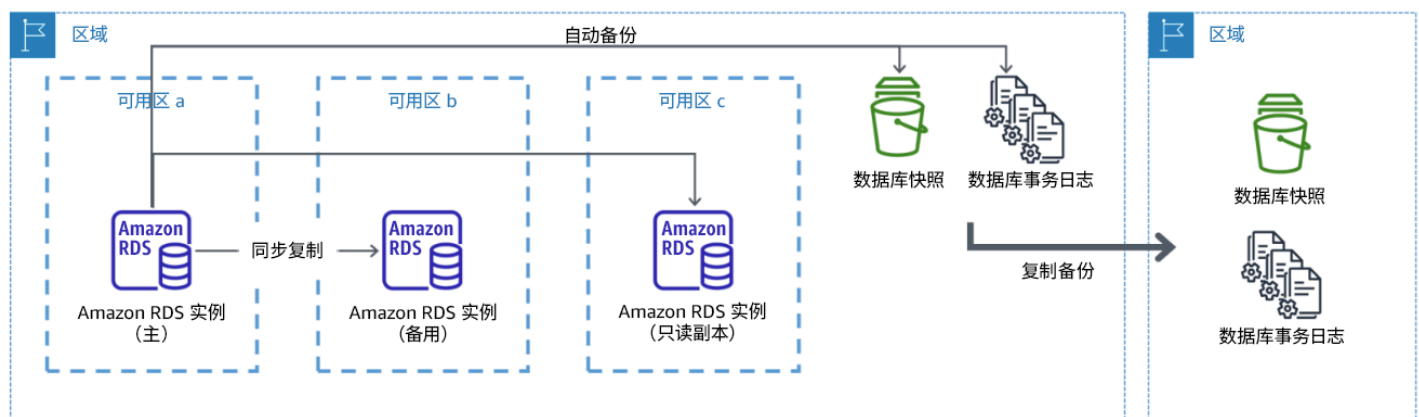


图 10：一个弹性多 AZ 数据库部署，该部署备份到另一个 AWS 区域

常见反模式

- 在多 AZ 架构可以满足需求时选择设计多区域架构。

- 当应用程序部件之间的弹性和多位置需求不同时，没有考虑它们之间的依赖关系。

## 建立此最佳实践的好处

要实现弹性，您应使用构建防御层的方法。其中一层使用多 AZ，通过构建高度可用的架构，防护较小规模的、更常见的中断。另一个防御层用于防御很少发生的事件，例如大范围的自然灾害和区域级别的中断。这个第二层涉及到设计应用程序的架构来跨越多个 AWS 区域。

- 99.5% 的可用性与 99.99% 的可用性相比，每个月的正常运行时间之差超过 3.5 小时。采用多 AZ 的工作负载的可用性，预期只能达到“四个九”。
- 通过在多个 AZ 中运行工作负载，您可以隔离电力、冷却和网络中的故障，以及火灾和洪水之类的大多数自然灾害。
- 为工作负载实施多区域策略，有助于防御影响到某个国家/地区中较大地理面积的大范围自然灾害，或者区域范围的技术故障。请注意，实施多区域架构会有很高的复杂性，对于大部分工作负载通常来说都是不必要的。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

对于一个可用区的中断或部分丢失而导致的灾难事件，在单个 AWS 区域内的多个可用区中实施高可用工作负载，有助于防范自然灾害和技术灾难。每个 AWS 区域由多个可用区组成，各个可用区之间实现了故障隔离并且间隔相当的物理距离。不过，对于可能造成间隔相当距离的多个可用区组件丢失风险的灾难事件，您应该实施灾难恢复选项，以防范整个区域的自然灾难和技术故障。对于需要极高弹性的工作负载（关键基础设施、与生命健康相关的应用程序、财务系统基础设施等），需要使用多区域策略。

## 实施步骤

1. 评估您的工作负载并确定需要使用多 AZ 方法（单个 AWS 区域）还是多区域方法才能够满足弹性需求。实施多区域架构来满足这些需求会引入额外的复杂性，因此请谨慎考虑您的使用场景及其需求。弹性需求几乎总是可以使用单个 AWS 区域来满足。在确定您是否需要使用多区域时，请考虑以下可能的需求：
  - a. 灾难恢复（DR，Disaster Recovery）：对于一个可用区的中断或部分丢失而导致的灾难事件，在单个 AWS 区域内的多个可用区中实施高可用工作负载，有助于防范自然灾害和技术灾难。对于可能造成间隔相当距离的多个可用区组件丢失风险的灾难事件，您应该实施跨多个区域的灾难恢复，以防范整个区域的自然灾难和技术故障。
  - b. 高可用性（HA，High Availability）：多区域架构（在每个区域中使用多个 AZ）可用于实现四个 9 以上（> 99.99%）的可用性。

- c. 堆栈本地化：面向全球受众部署工作负载时，您可以将本地化的堆栈部署在不同的 AWS 区域中，以便服务于这些区域中的受众。本地化可以包括语言、货币和所存储数据的类型。
  - d. 靠近用户：面向全球受众部署工作负载时，您可以通过在靠近最终用户所在位置的 AWS 区域中部署堆栈，从而减少延迟。
  - e. 数据驻留：一些工作负载面临着数据驻留要求，来自特定用户的数据必须保留在特定国家/地区的边界内。根据相关的法规，您可以选择将整个堆栈或者仅仅将数据部署到这些边界内的 AWS 区域中。
2. 以下是 AWS 服务提供的一些多 AZ 功能的示例：
- a. 为了使用 EC2 或 ECS 保护工作负载，请在计算资源前端部署 Elastic Load Balancer。然后，Elastic Load Balancing 提供解决方案来检测未正常运行的区中的实例，并将流量路由到正常运行的区中。
    - i. [开始使用 Application Load Balancers](#)
    - ii. [开始使用网络负载均衡器](#)
  - b. 当 EC2 实例运行不支持负载均衡的现成商用软件时，您可以通过实施多 AZ 灾难恢复方法来实现某种形式的容错能力。
    - i. [the section called “REL13-BP02 使用定义的恢复策略来实现恢复目标”](#)
  - c. 对于 Amazon ECS 任务，将您的服务均匀地部署在三个 AZ 上以实现可用性与成本的平衡。
    - i. [Amazon ECS 可用性最佳实践 | 容器](#)
  - d. 对于非 Aurora Amazon RDS，您可以选择多 AZ 作为配置选项。在主数据库实例出现故障时，Amazon RDS 会自动提升备用数据库，用于接收其他可用区中的流量。还可以创建多区域只读副本来改进弹性。
    - i. [Amazon RDS 多可用区部署](#)
    - ii. [在不同 AWS 区域中创建只读副本](#)
3. 以下是 AWS 服务提供的一些多区域功能的示例：
- a. 对于 Amazon S3 工作负载，当服务自动提供了多 AZ 可用性时，如果需要多区域部署，请考虑多区域接入点。
    - i. [Amazon S3 中的多区域接入点](#)
  - b. 对于 DynamoDB 表，此时服务自动提供了多 AZ 可用性，您可以轻松地将现有表转换为全局表来利用多区域的优势。
    - i. [将单区域 Amazon DynamoDB 表转换为全局表](#)
  - c. 如果您的工作负载采用 Application Load Balancers 或网络负载均衡器作为前端，请将流量引导到包含正常运行端点的多个区域，从而使用 AWS Global Accelerator 来改进应用程序的可用性。



- i. [AWS Global Accelerator 中标准加速器的端点 – AWS Global Accelerator \( amazon.com \)](#)
- d. 对于利用 AWS EventBridge 的应用程序而言，请考虑使用跨区域总线来将事件转发到您选择的其他区域。
  - i. [在 AWS 区域之间发送和接收 Amazon EventBridge 事件](#)
- e. 对于 Amazon Aurora 数据库，请考虑使用跨多个 AWS 区域的 Aurora 全局数据库。可以对现有集群进行修改来添加新的区域。
  - i. [开始使用 Amazon Aurora 全局数据库](#)
- f. 如果您的工作负载包括 AWS Key Management Service ( AWS KMS ) 加密密钥，请考虑多区域密钥是否适合您的应用程序。
  - i. [AWS KMS 中的多区域密钥](#)
- g. 对于其他 AWS 服务功能，请参阅此博客系列中的以下内容：[使用 AWS 服务创建多区域应用程序系列](#)

实施计划的工作量级别：中到高

资源

相关文档：

- [使用 AWS 服务创建多区域应用程序系列](#)
- [AWS 上的灾难恢复 \( DR , Disaster Recovery \) 架构，第 IV 部分：多站点主动/主动](#)
- [AWS 全球基础设施](#)
- [AWS Local Zones 常见问题](#)
- [AWS 上的灾难恢复 \( DR , Disaster Recovery \) 架构，第 I 部分：云中的恢复策略](#)
- [云中的灾难恢复不相同](#)
- [全局表：使用 DynamoDB 的多区域复制](#)

相关视频：

- [AWS re:Invent 2018：适用于多区域主动-主动应用程序的架构模式 \( ARC209-R2 \)](#)
- [Auth0：多区域高可用性架构，可扩展至每月 15 亿+ 次登录，并具有自动故障转移功能](#)

相关示例：



- [AWS 上的灾难恢复 \( DR , Disaster Recovery \) 架构 , 第 I 部分 : 云中的恢复策略](#)
- [DTCC 实现了本地部署无法企及的弹性](#)
- [Expedia Group 使用具有专有 DNS 服务的多区域、多可用区架构来增加应用程序的弹性](#)
- [Uber : 用于多区域 Kafka 的灾难恢复](#)
- [Netflix : 实现多区域弹性的主动-主动架构](#)
- [我们如何为 Atlassian Cloud 构建数据驻留](#)
- [Intuit TurboTax 跨两个区域运行](#)

## REL10-BP03 组件的自动恢复受限于单个位置

如果工作负载的组件只能在单个可用区或本地部署数据中心内运行，您必须利用相关功能在定义的恢复目标内彻底重建工作负载。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

如果由于技术限制无法使用将工作负载部署到多个位置的最佳实践，您必须实施其他的弹性路径。在这种情况下，您必须让重建必要基础设施、重新部署应用程序和重建必要数据的操作实现自动化。

例如，Amazon EMR 会为相同可用区内的特定集群启动全部节点，因为在相同区内运行集群可以改善作业流的性能，提高数据访问速率。如果这是工作负载弹性所需的必要组件，则您必须设法重新部署集群及其数据。同样对于 Amazon EMR，您还应该通过除多可用区以外的方式对冗余进行预置。您可以预置[多个节点](#)。使用 [EMR 文件系统 \( EMRFS \)](#)，EMR 中的数据可以存储在 Amazon S3 中，进而可以实现跨多个可用区或 AWS 区域复制。

同样，对于 Amazon Redshift 来说，它默认会在您选择的 AWS 区域内随机选择可用区，然后对其中的集群进行预置。相同区内的全部集群节点都会被预置。

对于部署到本地数据中心基于服务器的有状态工作负载，您可以使用 AWS Elastic Disaster Recovery 来保护 AWS 中的工作负载。如果您已在 AWS 中托管，则可以使用 Elastic Disaster Recovery 将工作负载保护到备用可用区或区域。Elastic Disaster Recovery 使用持续块级复制将数据复制到轻量级暂存区域，以便为本地和基于云的应用程序提供快速、可靠的恢复。

### 实施步骤

1. 实施自我修复。尽可能使用弹性伸缩部署实例或容器。如果不能使用弹性伸缩，则使用 EC2 实例的自动恢复功能，或者基于 Amazon EC2 或 ECS 容器生命周期事件实施自我修复自动化。

- 将 [Amazon EC2 Auto Scaling 组](#) 用于对单个实例 IP 地址、私有 IP 地址、弹性 IP 地址和实例元数据没有要求的实例和容器工作负载。
  - 启动模板用户数据可以用于实现自动化，从而让大多数工作负载可以自我修复。
- 将 [Amazon EC2 实例的自动恢复](#) 功能用于需要单个实例 IP 地址、私有 IP 地址、弹性 IP 地址和实例元数据的工作负载。
  - 自动恢复功能会在检测到实例故障时，向 SNS 主题发送恢复状态提醒。
- 在无法使用弹性伸缩或 EC2 恢复的情况下，使用 [Amazon EC2 实例生命周期事件](#) 或 [Amazon ECS 事件](#) 实现自动化的自我修复。
  - 使用这些事件调用自动化，该自动化将根据您需要的流程逻辑来修复组件。
- 使用 [AWS Elastic Disaster Recovery](#) 保护仅限于单个位置的有状态工作负载。

## 资源

### 相关文档：

- [Amazon ECS 事件](#)
- [Amazon EC2 Auto Scaling 生命周期挂钩](#)
- [恢复您的实例。](#)
- [服务弹性伸缩](#)
- [什么是 Amazon EC2 Auto Scaling ?](#)
- [AWS Elastic Disaster Recovery](#)

## REL10-BP04 采用隔板架构来限制影响范围

实施隔板架构（也称为基于 cell 的架构），将工作负载中故障的影响限制到有限数量的组件上。

期望结果：基于 cell 的架构使用工作负载的多个独立实例，每个实例称为一个 cell。每个 cell 是独立的，不与其他 cell 分享状态，并处理整个工作负载请求的一个子集。这样减少了故障（例如，错误的软件更新）对单个 cell 及其正在处理的请求的潜在影响。如果工作负载使用 10 个 cell 来提供 100 个请求，则在发生故障时，总请求中有 90% 不受故障的影响。

### 常见反模式：

- 让 cell 无限制地发展。
- 同时将代码更新或部署应用到所有 cell。

- 在不同 cell 之间分享状态或组件（路由器层除外）。
- 向路由器层添加复杂的业务或路由逻辑。
- 没有尽量减少 cell 间的相互作用。

建立此最佳实践的好处：借助基于 cell 的架构，许多常见类型的故障控制在 cell 本身，从而实现了额外的故障隔离。在出现难以控制的故障类型（例如不成功的代码部署，或者是出错或触发特定故障模式的请求，也称为“毒药”请求）时，这些故障边界可以提供弹性。

实施指导

在船舶上，隔板确保船体裂口控制在船体的一个区段内。在复杂的系统中，通常会复制此模式以实现故障隔离。故障隔离边界可将一个工作负载内的故障影响限制于有限数量的组件。边界以外的组件不会受到故障的影响。使用多个故障隔离边界，您可以限制作用于您的工作负载的影响。在 AWS 中，客户可以使用多个可用区和区域来实现故障隔离，但故障隔离的概念可以也可以扩展到工作负载的架构。

整个工作负载是分区的 cell，按分区键进行划分。这个键需要与服务的粒度，或者与使用最小的跨 cell 交互来细分服务工作负载的自然方式保持一致。分区键的示例包括客户 ID、资源 ID 或可以在大多数 API 调用中轻松访问的任何其他参数。cell 路由层根据分区键将请求分发到单个 cell，并向客户端提供单个端点。

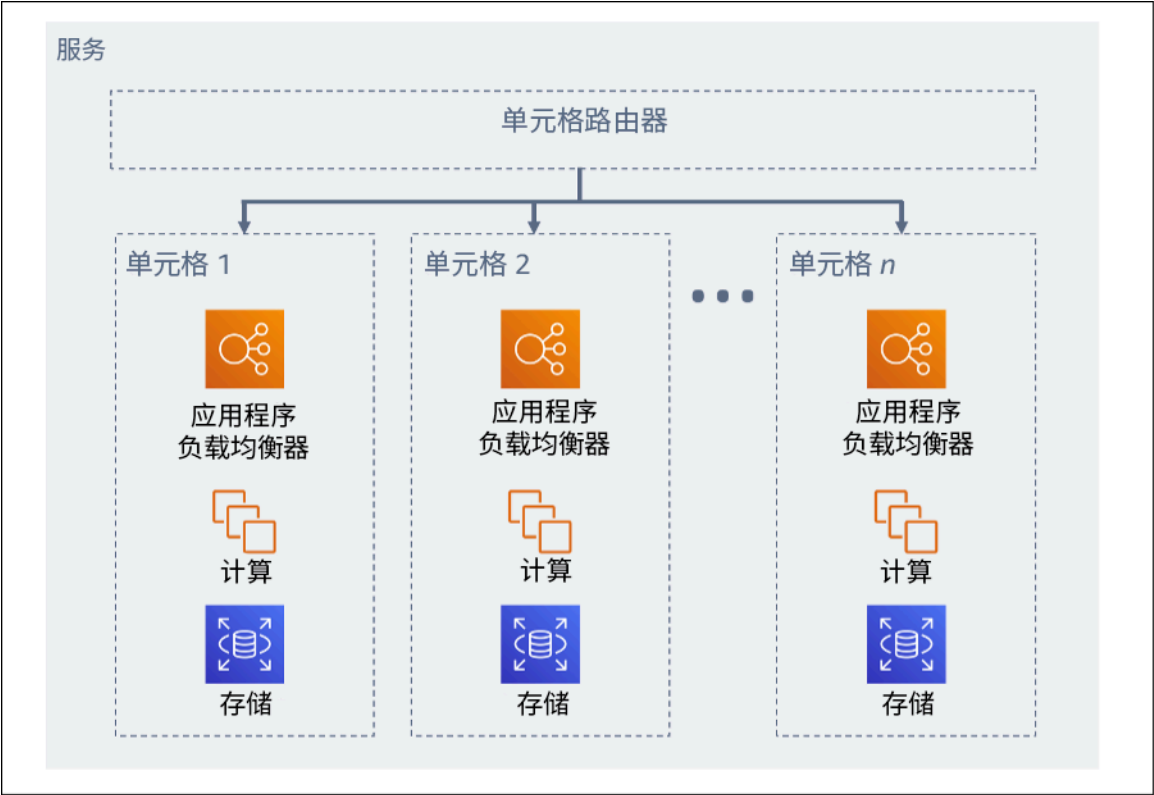


图 11：基于 Cell 的架构

## 实施步骤

在设计基于 cell 的架构时，需要考虑几个设计注意事项：

1. 分区键：选择分区键时应特别考虑。
  - 它应与服务的粒度，或者与使用最小的跨 cell 交互来细分服务工作负载的自然方式保持一致。示例包括 ## ID 或 ## ID。
  - 在所有请求中都必须提供分区键，要么直接提供，要么通过很容易由其他参数确定地推断出来的方式提供。
2. 持久 cell 映射：上游服务在其资源的生命周期内应只与单个 cell 交互。
  - 根据工作负载，可能需要使用 cell 迁移策略将数据从一个 cell 迁移到另一个 cell。可能需要进行 cell 迁移的一种情况是：工作负载中的一个特定用户或资源变得太大，要求它具有专用的 cell。
  - Cell 之间不应该共享状态或组件。
  - 因此，应避免或尽量减少跨 cell 的交互，因为这些交互会在 cell 之间产生依赖关系，从而削弱故障隔离的改进。
3. 路由器层：路由器层是 cell 之间的一个共享组件，因此无法遵循与 cell 相同的分隔策略。
  - 建议路由器层使用分区映射算法以一种计算效率高的方式将请求分发到各个 cell，例如结合加密哈希函数和模块化算法，将分区键映射到 cell。
  - 为避免产生多 cell 影响，路由层必须尽可能保持简单和可横向扩展，这就需要在此层中避免出现复杂的业务逻辑。这还有一个额外的好处，即始终可以轻松地理解其预期行为，从而实现彻底的可测试性。正如 Colm MacCárthaigh 在[可靠性、持续工作和安然无忧](#)中所说，简单的设计和持续工作模式可产生可靠的系统并降低抗脆弱性。
4. Cell 大小：Cell 应具有最大大小，不得发展到超过这个大小。
  - 应通过执行彻底的测试来确定最大大小，直至达到临界点并建立安全的运营边际。有关如何实施测试实践的更多详细信息，请参阅[REL07-BP04 对工作负载进行负载测试](#)
  - 总体工作负载增长时应增加额外的 cell，使得工作负载能够随着需求的增加而扩展。
5. 多可用区或多区域策略：应该利用多层弹性来防止出现不同的故障域。
  - 要实现弹性，您应使用构建防御层的方法。其中一层使用多可用区，通过构建高度可用的架构，防止出现较小规模、更常见的中断。另一个防御层用于防御很少发生的事件，例如大范围的自然灾害和区域级别的中断。这个第二层涉及到设计应用程序的架构来跨越多个 AWS 区域。为工作负载实施多区域策略，有助于防御影响到某个国家/地区中较大地理面积的大范围自然灾害，或者区域范围的技术故障。请注意，实施多区域架构会有很高的复杂性，对于大部分工作负载通常来说都是不必要的。有关更多详细信息，请参阅[REL10-BP02 为您的多位置部署选择合适的位置](#)。
6. 代码部署：交错的代码部署策略应该优于同时将代码更改部署到所有 cell。

- 这将有助于最大限度地减少由于部署不当或人为错误而导致多个 cell 可能出现故障。有关更多详细信息，请参阅[自动执行安全、不需要人工介入的部署](#)。

在未建立这种最佳实践的情况下暴露的风险等级：高

资源

相关最佳实践：

- [REL07-BP04 对工作负载进行负载测试](#)
- [REL10-BP02 为您的多位置部署选择合适的位置](#)

相关文档：

- [可靠性、持续工作和安然无忧](#)
- [AWS 和分隔](#)
- [采用随机分区进行工作负载隔离](#)
- [自动执行安全、不需要人工介入的部署](#)

相关视频：

- [AWS re:Invent 2018：闭环系统和开放思维：如何掌控不同规模的系统](#)
- [AWS re:Invent 2018：AWS 如何将故障的影响范围最小化 \( ARC338 \)](#)
- [随机分片：AWS re:Invent 2019：Amazon Builders' Library 简介 \( DOP328 \)](#)
- [2021 AWS 峰会 \( 澳大利亚和新西兰 \) - 故障在所难免：针对弹性而设计](#)

相关示例：

- [Well-Architected 实验室 - 使用随机分片进行故障隔离](#)

## REL 11. 如何将您的工作负载设计为可承受组件故障的影响？

在设计具有高可用性和较短平均恢复时间 ( MTTR ) 要求的工作负载时必须考虑到弹性。

最佳实践

- [REL11-BP01 监控工作负载的所有组件以检测故障](#)
- [REL11-BP02 失效转移到运行状况良好的资源](#)
- [REL11-BP03 自动修复所有层](#)
- [REL11-BP04 恢复期间依赖于数据平面而不是控制平面](#)
- [REL11-BP05 使用静态稳定性来防止双模态行为](#)
- [REL11-BP06 当事件影响可用性时发出通知](#)
- [REL11-BP07 构造您的产品以满足可用性目标和正常运行时间服务协议 \( SLA \)](#)

## REL11-BP01 监控工作负载的所有组件以检测故障

持续监控您的工作负载的运行状况，以便您和您的自动化系统立即发现任何故障或性能下降情况。监控基于商业价值的关键性能指标 ( KPI ) 。

所有恢复和修复机制必须从快速检测问题的能力入手。首先，应该检测技术故障并加以解决。不过，可用性基于您的工作负载创造商业价值的能力，因此衡量它的关键性能指标 ( KPI ) 需要成为您的检测和补救策略的一部分。

期望的结果：独立监控工作负载的重要组成部分，以检测故障发生的时间和位置，并发出警报。

常见反模式：

- 由于未配置警报，因此在发生中断时不会进行通知。
- 虽然存在警报，但只有在达到阈值时才会发出警报，导致没有足够的响应时间。
- 收集指标的频率不够高，无法满足恢复时间目标 ( RTO ) 。
- 工作负载中只有面向客户的接口才会被主动监控。
- 只收集技术指标，不收集业务功能指标。
- 没有衡量工作负载用户体验的指标。
- 创建的监控太多。

建立此最佳实践的好处：如果您在所有层面都设置了适当的监控，则可以通过减少检测时间来缩短恢复时间。

未建立这种最佳实践的情况下暴露的风险等级：高



## 实施指导

确定将接受审查以决定是否监控的所有工作负载。确定工作负载中所有需要监控的组成部分后，就需要确定监控间隔。根据检测故障所花费的时间，监控间隔将直接影响启动恢复的速度。平均检测时间（MTTD）是从故障发生到修复操作开始之间的时间。服务清单应广泛而完整。

监控必须覆盖应用程序堆栈的所有层，包括应用、平台、基础设施和网络。

您的监控策略应考虑以下因素的影响：灰色故障。有关灰色故障的更多详细信息，请参阅《Advanced Multi-AZ Resilience Patterns》白皮书中的 [Gray failures](#)。

## 实施步骤

- 您的监控间隔取决于您必须以多快的速度恢复。您的恢复时间取决于恢复所需的时间，因此您在确定收集频率时，必须考虑此时间和恢复时间目标（RTO）。
- 为组件和托管服务配置详细监控。
  - 确定 [详细监控对于 EC2 实例](#) 和 [Auto Scaling](#) 来说是否有必要。详细监控以 1 分钟为间隔提供指标，默认监控以 5 分钟为间隔提供指标。
  - 确定 [增强监控](#) 对于 RDS 来说是否有必要。增强监控使用 RDS 实例上的代理，来获取关于不同进程或线程的有用信息。
  - 确定以下各项的关键无服务器组件的监控要求：[Lambda](#)、[API Gateway](#)、[Amazon EKS](#)、[Amazon ECS](#)，以及所有类型的 [负载均衡器](#)。
  - 确定以下各项的存储组件的监控要求：[Amazon S3](#)、[Amazon FSx](#)、[Amazon EFS](#) 和 [Amazon EBS](#)。
- 创建 [自定义指标](#) 来衡量业务关键绩效指标（KPI）。工作负载会实现关键业务功能，这些功能应用作 KPI 来协助在发生间接问题时予以识别。
- 使用用户金丝雀来监控用户的故障体验。[可运行和模拟客户行为的综合事务测试](#)（又称为“金丝雀测试”，但不要和金丝雀部署相混淆）是最重要的测试流程之一。从不同的远程位置针对您的工作负载端点持续地运行此类测试。
- 创建 [自定义指标](#) 来跟踪用户体验。如果您可以衡量客户体验，就可以确定发生了客户体验下降。
- [设置警报](#)，以在检测到工作负载的任何部分未正常运行时发出警报，并指示什么时候自动扩展资源。警报可以直观地显示在控制面板上，通过 Amazon SNS 或电子邮件发送，并与 Auto Scaling 结合使用来扩展或缩减工作负载资源。
- 创建 [控制面板](#) 以可视化形式呈现指标。可以使用控制面板直观地查看趋势、离群值和表示其他潜在问题的指标，或者提供您可能需要进行调查的问题的指示。



- 为您的服务创建 [分布式跟踪监控](#)。使用分布式监控，您可以了解应用程序及其底层服务的运行情况，以便确定和诊断性能问题及错误的根本原因。
- 在单独的区域和账户中创建监控系统（使用 [CloudWatch](#) 或 [X-Ray](#)）控制面板和数据收集。
- 为 [Amazon Health Aware](#) 集成监控功能，以便监控可能出现性能下降的 AWS 资源。对于关键业务工作负载，此解决方案可提供对 AWS 服务的主动实时警报的访问。

## 资源

### 相关最佳实践：

- [可用性定义](#)
- [REL11-BP06 当事件影响可用性时发出通知](#)

### 相关文档：

- [Amazon CloudWatch Synthetics 使您能够创建用户金丝雀](#)
- [为您的实例启用或禁用详细监控](#)
- [增强监控](#)
- [Monitoring Your Auto Scaling Groups and Instances Using Amazon CloudWatch](#)
- [发布自定义指标](#)
- [使用 Amazon CloudWatch 警报](#)
- [使用 CloudWatch 控制面板](#)
- [跨区域跨账户使用 CloudWatch 控制面板](#)
- [跨区域跨账户使用 X-Ray 跟踪](#)
- [了解可用性](#)
- [实施 Amazon Health Aware \( AHA \)](#)

### 相关视频：

- [Mitigating gray failures](#)

### 相关示例：

- [Well-Architected 实验室：第 300 级：实施运行状况检查和管理依赖项以提高可靠性](#)

- [可观测性研讨会：探索 X-Ray](#)

相关工具：

- [CloudWatch](#)
- [CloudWatch X-Ray](#)

## REL11-BP02 失效转移到运行状况良好的资源

如果资源发生故障，运行状况良好的资源应继续为请求提供服务。对于位置受损（如可用区或 AWS 区域），确保您拥有适当的系统，以失效转移到未受损位置内运行状况良好的资源。

设计服务时，应在资源、可用区或区域之间分配负载。因此，可以通过将流量转移到运行状况良好的剩余资源，缓解单个资源的故障或损坏。考虑在出现故障时如何发现服务并将流量路由到相应服务。

在设计服务时要考虑故障恢复。在 AWS，我们设计服务时会尽量缩短从故障恢复的时间并降低对数据的影响。我们的服务主要使用的数据存储，只有在数据持久存储在一个区域中的多个副本之后，才会确认请求。它们被构建为使用基于单元格的隔离，并使用可用区提供的故障隔离功能。我们在自己的运营过程中广泛使用自动化。我们还将替换和重新启动功能优化为可从中断快速恢复。

允许失效转移的模式和设计因各项 AWS 平台服务而异。许多 AWS 原生托管服务本质上跨多个可用区（如 Lambda 或 API Gateway）。其他 AWS 服务（如 EC2 和 EKS）需要特定的最佳实践设计，来支持跨可用区的资源或数据存储的失效转移。

应设置监控功能，以检查失效转移资源是否正常，跟踪资源失效转移的进度，并监控业务流程的恢复情况。

期望的结果：系统能够自动使用新资源从降级中恢复，或手动恢复。

常见反模式：

- 规划和设计阶段未考虑如何应对失败。
- 未确立 RTO 和 RPO。
- 监控不足，无法检测出故障的资源。
- 适当隔离故障域。
- 不考虑多区域失效转移。
- 在决定是否进行失效转移时，故障检测过于敏感或过于激进。

- 未测试或验证失效转移设计。
- 执行自动修复，但不通知需要进行该修复。
- 缺少缓冲期，无法避免太快进行失效自动恢复。

建立此最佳实践的好处：您可以构建更具韧性的系统，在遇到故障时，通过正常降级和快速恢复，保持可靠性。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

AWS 服务（例如 [Elastic Load Balancing](#) 和 [Amazon EC2 Auto Scaling](#)）有助于跨资源和可用区分配负载。因此，可以通过将流量转移到运行状况良好的剩余资源，缓解单个资源（例如 EC2 实例）的故障或可用区的损坏。

对于多区域工作负载，设计就比较复杂。例如，跨区域只读副本允许您将数据部署到多个 AWS 区域。但是，仍需要进行失效转移才能将只读副本提升为主副本，然后将流量指向新的终端节点。Amazon Route 53、Route 53、ARC、CloudFront 和 AWS Global Accelerator 可以协助跨 AWS 区域路由流量。

Amazon S3、Lambda、API Gateway、Amazon SQS、Amazon SNS、Amazon SES、Amazon Pinpoint、Amazon ECR、AWS Certificate Manager、EventBridge 或 Amazon DynamoDB 等 AWS 服务将通过 AWS 自动部署到多个可用区。如果出现故障，这些 AWS 服务会自动将流量路由到状态正常的位置。数据在多个可用区中进行冗余存储，并保持可用。

对于 Amazon RDS、Amazon Aurora、Amazon Redshift、Amazon EKS 或 Amazon ECS，多可用区是一个配置选项。如果启动了失效转移，AWS 可以将流量引导到运行正常的实例。此失效转移操作可由 AWS 执行，或根据客户要求执行

对于 Amazon EC2 实例、Amazon Redshift、Amazon ECS 任务或 Amazon EKS Pod，您要选择部署到哪些可用区。对于某些设计，Elastic Load Balancing 会提供解决方案以检测运行不正常的可用区内的实例，并将流量路由至运行正常的可用区。Elastic Load Balancing 还可以将流量路由至本地数据中心内的组件。

对于多区域流量失效转移，重新路由可以利用 Amazon Route 53、ARC、AWS Global Accelerator、Route 53 Private DNS for VPCs 或 CloudFront 来提供一种方法，以定义互联网域并分配路由策略（包括运行状况检查），从而将流量路由到运行正常的区域。AWS Global Accelerator 提供静态 IP 地址，这些地址充当应用程序的固定入口点，然后使用 AWS 全球网络而不是互联网来路由到您选择的 AWS 区域中的端点，以获得更高的性能和可靠性。

## 实施步骤

- 为所有相应的应用程序和服务创建失效转移设计。隔离每个架构组件，为每个组件创建符合 RTO 和 RPO 的失效转移设计。
- 使用失效转移计划所需的所有服务配置底层环境（例如开发或测试）。使用基础设施即代码（IaC）部署解决方案，以确保可重复性。
- 配置恢复站点（例如第二个区域）以实施和测试失效转移设计。如有必要，可以临时配置测试资源以限制额外成本。
- 确定哪些失效转移计划由 AWS 自动执行，哪些可以通过 DevOps 流程自动执行，哪些可能需要手动执行。记录并测量每项服务的 RTO 和 RPO。
- 创建失效转移行动手册，包括对每个资源、应用程序和服务进行失效转移的所有步骤。
- 创建失效自动恢复行动手册，包括对每个资源、应用程序和服务进行失效自动恢复的所有步骤（包含时机）
- 制定计划以启动和演练行动手册。使用模拟和混沌测试来测试行动手册步骤和自动化功能。
- 对于位置受损（如可用区或 AWS 区域），确保您拥有适当的系统，以失效转移到未受损位置内运行状况良好的资源。在测试失效转移之前，请检查配额、自动扩展级别和正在运行的资源。

## 资源

相关的 Well-Architected 最佳实践：

- [REL13 - 制定灾难恢复计划](#)
- [REL10 - 使用故障隔离来保护您的工作负载](#)

相关文档：

- [设置 RTO 和 RPO 目标](#)
- [使用应用程序负载均衡器设置 ARC](#)
- [使用 Route 53 加权路由进行失效转移](#)
- [使用 ARC 进行灾难恢复](#)
- [带自动扩缩功能的 EC2](#)
- [EC2 部署 - 多可用区](#)
- [ECS 部署 - 多可用区](#)
- [使用 ARC 切换流量](#)

- [使用 Application Load Balancer 和失效转移的 Lambda](#)
- [ACM 复制和失效转移](#)
- [Parameter Store 复制和失效转移](#)
- [ECR 跨区域复制和失效转移](#)
- [Secrets Manager 跨区域复制配置](#)
- [为 EFS 和失效转移启用跨区域复制](#)
- [EFS 跨区域复制和失效转移](#)
- [网络失效转移](#)
- [使用 MRAP 的 S3 端点失效转移](#)
- [为 S3 创建跨区域复制](#)
- [使用 ARC 对区域 API Gateway 进行失效转移](#)
- [使用多区域全球加速器进行失效转移](#)
- [使用 DRS 进行失效转移](#)
- [使用 Amazon Route 53 创建灾难恢复机制](#)

相关示例：

- [Disaster Recovery on AWS](#)
- [Elastic Disaster Recovery on AWS](#)

## REL11-BP03 自动修复所有层

在检测到故障时，使用自动化功能执行修复操作。降级可能会通过内部服务机制自动修复，也可能需要通过补救措施重启或移除资源。

对于自我管理型应用程序和跨区域修复，可以从 [现有最佳实践](#) 中获取恢复设计和自动修复流程。

重启或移除资源是修复故障的重要方法。最佳实践是尽可能使服务为无状态。这可以防止重启资源时数据丢失或可用性受损。在云中，作为重启的一部分，您可以（而且在一般情况下也应该）替换完整的资源（例如计算实例或无服务器函数）。重启本身是从故障恢复的简单而可靠的方法。工作负载中会发生很多不同类型的故障。故障可能发生在硬件、软件、通信和操作上。

重启或重试也适用于网络请求。向网络超时以及依赖项返回错误的依赖性故障应用相同的恢复方法。这两个事件对系统具有类似的影响，可应用类似的采用指数回退和抖动的有限重试策略，而不是尝试将各个事件当作特例进行处理。重启功能是面向恢复的计算和高可用性集群架构的特色恢复机制。

期望的结果：执行自动操作以修复检测到的故障。

常见反模式：

- 预置资源，而不进行自动扩缩。
- 在实例或容器中单独部署应用程序。
- 在不使用自动恢复的情况下，部署无法部署到多个位置的应用程序。
- 手动修复弹性伸缩和自动恢复无法修复的应用程序。
- 缺乏对数据库进行失效转移的自动化功能。
- 缺乏将流量重新路由到新端点的自动化方法。
- 缺乏存储复制功能。

建立此最佳实践的好处：自动修复可以缩短平均恢复时间，并提高可用性。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

Amazon EKS 或其他 Kubernetes 服务的设计应包括最小和最大副本集或有状态集，以及最小集群和节点组大小。这些机制提供了最低数量的持续可用处理资源，同时使用 Kubernetes 控制平面自动修复任何故障。

使用计算集群通过负载均衡器访问的设计模式应利用 Auto Scaling 组。Elastic Load Balancing (ELB) 自动在一个或多个可用区 (AZ) 中的多个目标和虚拟设备之间分配传入的应用程序流量。

对于不使用负载平衡的基于集群计算的设计，其规模至少应能承受一个节点的中断。这将允许服务在恢复新节点时，使自身以可能降低的容量维持运行状态。示例服务有 Mongo、DynamoDB Accelerator、Amazon Redshift、Amazon EMR、Cassandra、Kafka、MSK-EC2、Couchbase、ELK 和 Amazon OpenSearch Service。其中许多服务都可以设计额外的自动修复功能。某些集群技术必须在节点丢失时生成警报，从而触发自动或手动工作流程来重新创建新节点。可以使用 AWS Systems Manager 来自动执行此工作流程，以快速修复问题。

Amazon EventBridge 可用于监控和筛选事件，例如 CloudWatch 警报或其他 AWS 服务中的状态更改。然后，它可以根据事件信息调用 AWS Lambda、Systems Manager Automation 或其他目标，对您的工作负载运行自定义修复逻辑。可以将 Amazon EC2 Auto Scaling 配置为检查 EC2 实例的运行状况。若实例处于正在运行以外的任何状态，或系统状态受损，Amazon EC2 Auto Scaling 会认为实例的运行不正常，并且启动替换实例。针对大规模替换（例如整个可用区丢失），静态稳定性更适合用来实现高可用性。

## 实施步骤

- 使用 Auto Scaling 组在工作负载中部署层。 [Auto Scaling](#) 可以对无状态应用程序执行自我修复，以及添加或移除容量。
- 对于前面提到的计算实例，可使用 [负载均衡](#)，然后选择合适的负载均衡器类型。
- 考虑 Amazon RDS 修复。对于备用实例，请配置 [自动失效转移](#) 到备用实例。针对 Amazon RDS 只读副本，需要自动化工作流程，使只读副本成为主副本。
- 在部署了应用程序的 [EC2 实例上实施自动恢复](#)，这些应用程序无法部署到多个位置，但在故障后允许重新启动。当无法将应用程序部署到多个位置时，可以使用自动恢复来替换发生故障的硬件并重新启动实例。将保留实例元数据和关联的 IP 地址，以及 [EBS 卷](#) 和 [Amazon Elastic File System](#) 或 [适用于 Lustre 的文件系统](#) 和 [适用于 Windows 的文件系统](#) 的挂载点。使用 [AWS OpsWorks](#) 时，您可以在层级别配置 EC2 实例的自动修复。
- 当您无法使用自动扩展或自动恢复时，或者自动恢复失败时，使用 [AWS Step Functions](#) 和 [AWS Lambda](#) 实施自动恢复。当您无法使用自动扩展，并且无法使用自动恢复或自动恢复失败时，可以使用 AWS Step Functions 和 AWS Lambda 进行自动修复。
- [Amazon EventBridge](#) 可用于监控和筛选事件，例如 [CloudWatch 警报](#) 或其他 AWS 服务状态的变化。根据事件信息，它可以调用 AWS Lambda（或其他目标），在您的工作负载上运行自定义修复逻辑。

## 资源

### 相关最佳实践：

- [可用性定义](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)

### 相关文档：

- [AWS Auto Scaling 的工作原理](#)
- [Amazon EC2 Automatic Recovery](#)
- [Amazon Elastic Block Store \( Amazon EBS \)](#)
- [Amazon Elastic File System \( Amazon EFS \)](#)
- [什么是 Amazon FSx for Lustre ?](#)
- [什么是 Amazon FSx for Windows File Server ?](#)
- [AWS OpsWorks：使用自动修复来更换失败的实例](#)



- [什么是 AWS Step Functions ?](#)
- [什么是 AWS Lambda ?](#)
- [什么是 Amazon EventBridge ?](#)
- [使用 Amazon CloudWatch 警报](#)
- [Amazon RDS 失效转移](#)
- [SSM - Systems Manager 自动化](#)
- [弹性架构最佳实践](#)

相关视频：

- [Automatically Provision and Scale OpenSearch Service](#)
- [Amazon RDS Failover Automatically](#)

相关示例：

- [Auto Scaling 研讨会](#)
- [Amazon RDS 失效转移研讨会](#)

相关工具：

- [CloudWatch](#)
- [CloudWatch X-Ray](#)

## REL11-BP04 恢复期间依赖于数据平面而不是控制平面

控制平面提供用于创建、读取和描述、更新、删除和列出 ( CRUDL ) 资源的管理 API，而数据平面则处理日常服务流量。在对可能影响弹性的事件实施恢复或缓解响应时，着眼于使用最少数量的控制平面操作，来实现对服务的恢复、重新扩展、恢复、修复或失效转移。在这些降级事件期间，数据平面操作应凌驾于任何活动之上。

例如，以下是所有控制平面操作：启动新的计算实例、创建数据块存储和描述队列服务。启动计算实例时，控制平面必须执行多项任务，例如查找具有容量的物理主机、分配网络接口、准备本地数据块存储卷、生成凭证以及添加安全规则。控制平面往往是复杂的编排。

期望的结果：当资源进入受损状态时，系统能够通过将流量从受损资源转移到正常运行资源，来自动或手动恢复。

## 常见反模式：

- 依赖于通过更改 DNS 记录来重新路由流量。
- 由于预置资源不足，依赖控制平面扩展操作来替换受损组件。
- 依靠广泛、多服务、多 API 控制平面操作来修复任何类别的受损情况。

建立此最佳实践的好处：提高自动修复的成功率可以缩短平均恢复时间，并提高工作负载的可用性。

未建立这种最佳实践的情况下暴露的风险等级：中。对于某些类型的服务降级，控制平面会受到影响。依赖于大量使用控制平面进行修复可能会增加恢复时间（RTO）和平均恢复时间（MTTR）。

## 实施指导

要限制数据平面操作，请评估每项服务，以了解恢复服务所需的操作。

利用 Amazon Application Recovery Controller (ARC) 来转移 DNS 流量。这些功能持续监控您的应用程序从故障中恢复的能力，并使您能够跨多个 AWS 区域、可用区和本地部署控制您的应用程序恢复。

Route 53 路由策略使用控制平面，因此不要依赖控制平面进行恢复。Route 53 数据平面会回复 DNS 查询，并执行和评估运行状况检查。它们分布在全球各地，专为 [100% 可用性的服务等级协议 \(SLA\)](#) 而设计。

您用于创建、更新和删除 Route 53 资源的 Route 53 管理 API 和控制台是在控制平面上运行的，而这些控制平面设计用于优先考虑您在管理 DNS 时所需的强一致性和持久性。为了实现这一点，控制平面位于单个区域“美国东部（弗吉尼亚州北部）”中。虽然这两个系统都非常可靠，但控制平面不包含在 SLA 中。在极少数情况下，数据平面的弹性设计允许它保持可用性，而控制平面做不到。对于灾难恢复和失效转移机制，使用数据平面功能可提供尽可能高的可靠性。

对于 Amazon EC2，使用静态稳定性设计来限制控制平面操作。控制平面操作包括单独扩展资源，或使用 Auto Scaling 群组（ASG）来扩展资源。要获得最高级别的弹性，请在集群中配置足够的容量用于失效转移。如果必须限制此容量阈值，请在整个端到端系统上设置限制，以安全地限制到达有限资源集的总流量。

对于诸如 Amazon DynamoDB、Amazon API Gateway、负载均衡器和 AWS Lambda 无服务器之类的服务，使用这些服务时可以利用数据平面。但是，创建新函数、负载均衡器、API 网关或 DynamoDB 表是一项控制平面操作，应在降级之前完成，以便为事件和演练失效转移操作做准备。对于 Amazon RDS，数据平面操作允许访问数据。

有关数据平面、控制平面以及 AWS 如何构建服务以满足高可用性目标的更多信息，请参阅 [使用可用区的静态稳定性](#)。

了解哪些操作位于数据平面，哪些位于控制平面。

## 实施步骤

对于降级事件后需要恢复的每个工作负载，请评估失效转移运行手册、高可用性设计、自动修复设计或 HA 资源恢复计划。确定可能被视为控制平面操作的每个操作。

考虑将控制平面操作更改为数据平面操作：

- Auto Scaling（控制平面）与预扩展 Amazon EC2 资源（数据平面）的比较
- 迁移到 Lambda 及其扩展方法（数据平面）或 Amazon EC2 和 ASG（控制平面）
- 评估任何使用 Kubernetes 的设计以及控制平面操作的性质。在 Kubernetes 中，添加 Pod 是一项数据平面操作。操作应仅限于添加 Pod 而不是添加节点。使用 [过度预置的节点](#) 是限制控制平面操作的首选方法

考虑允许数据平面操作影响相同修复措施的其他方法。

- Route 53 记录更改（控制平面）或 ARC（数据平面）
- [Route 53 运行状况检查以获取更多自动更新](#)

如果服务是任务关键型，可考虑使用辅助区域中的某些服务，以便在不受影响的区域内进行更多控制平面和数据平面操作。

- 主区域中的 Amazon EC2 Auto Scaling 或 Amazon EKS 与辅助区域中的 Amazon EC2 Auto Scaling 或 Amazon EKS 的比较，以及将流量路由到辅助区域（控制平面操作）
- 在辅助区域中创建只读副本或在主区域中尝试相同的操作（控制平面操作）

## 资源

相关最佳实践：

- [可用性定义](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)

相关文档：

- [APN 合作伙伴：可以帮助您实现容错自动化的合作伙伴](#)

- [AWS Marketplace](#) : 可以支持容错的产品
- [Amazon Builders' Library](#) : 通过控制较小的服务来避免分布式系统中出现过载
- [Amazon DynamoDB API \( 控制平面和数据平面 \)](#)
- [AWS Lambda 执行](#) ( 拆分为控制平面和数据平面 )
- [AWS Elemental MediaStore 数据平面](#)
- [使用 Amazon Route 53 应用程序恢复控制器构建高弹性应用程序，第 1 部分：单区域堆栈](#)
- [使用 Amazon Route 53 应用程序恢复控制器构建高弹性应用程序，第 2 部分：多区域堆栈](#)
- [使用 Amazon Route 53 创建灾难恢复机制](#)
- [什么是 Route 53 应用程序恢复控制器？](#)
- [Kubernetes 控制平面和数据平面](#)

相关视频：

- [Back to Basics - Using Static Stability](#)
- [Building resilient multi-site workloads using AWS global services](#)

相关示例：

- [Amazon Route 53 应用程序恢复控制器简介](#)
- [Amazon Builders' Library](#) : 通过控制较小的服务来避免分布式系统中出现过载
- [使用 Amazon Route 53 应用程序恢复控制器构建高弹性应用程序，第 1 部分：单区域堆栈](#)
- [使用 Amazon Route 53 应用程序恢复控制器构建高弹性应用程序，第 2 部分：多区域堆栈](#)
- [使用可用区的静态稳定性](#)

相关工具：

- [Amazon CloudWatch](#)
- [AWS X-Ray](#)

REL11-BP05 使用静态稳定性来防止双模态行为

工作负载应具有静态稳定性，并且仅在单一正常模式下运行。双模态行为是指工作负载在正常模式和故障模式下表现出不同的行为。

例如，您可能会尝试在不同的可用区中启动新实例，以便从可用区故障中恢复。在故障模式下，这可能会导致双模态响应。您应该构建静态稳定的工作负载，并且仅在一个模式下运行。在此示例中，这些实例应在出现故障之前，已经在第二个可用区中预置。这种静态稳定性设计可确保工作负载仅在单一模式下运行。

期望的结果：在正常模式和故障模式下，工作负载不会表现出双模态行为。

常见反模式：

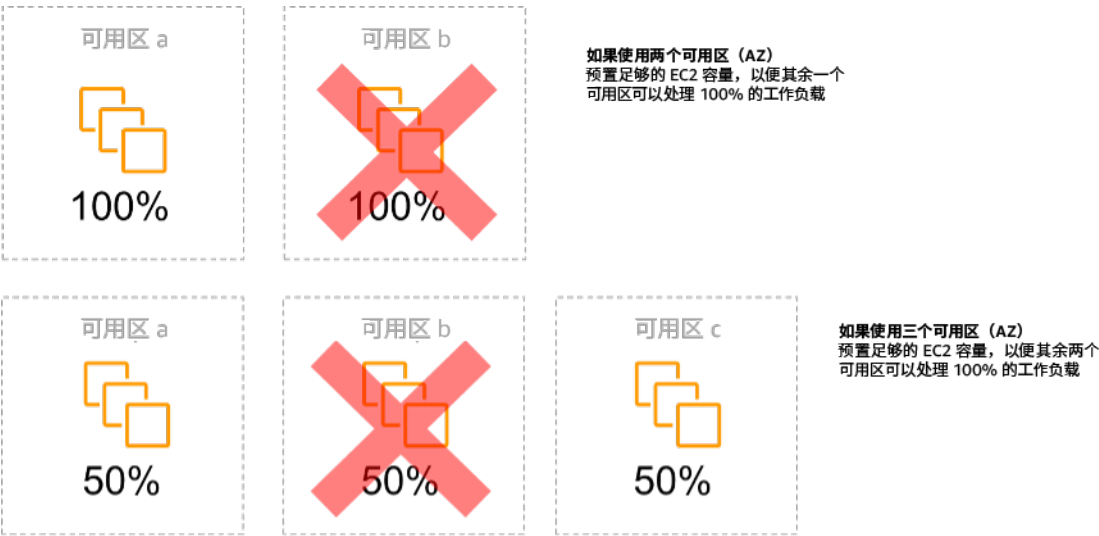
- 假设无论故障范围如何，始终可以预置资源。
- 尝试在故障期间动态获取资源。
- 在出现故障之前，没有跨可用区或跨区域预置足够的资源。
- 仅为计算资源考虑了静态稳定设计。

建立此最佳实践的好处：采用静态稳定设计运行的工作负载，在正常情况和故障事件期间能够得到可预测的结果。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

当工作负载在正常模式和故障模式下展现出不同的行为时，这就是双模态行为（例如，在可用区发生故障时依赖于启动新的实例）。双模态行为的一个例子是，稳定 Amazon EC2 设计在每个可用区中预置了足够的实例，用于处理在移除了一个可用区时的工作负载。Elastic Load Balancing 或 Amazon Route 53 运行状况将执行检查，并将负载从受损实例上移出。在流量转移以后，使用 AWS Auto Scaling 异步替换故障可用区的实例，并在运行正常的可用区内启动。适用于计算部署（如 EC2 实例或容器）的静态稳定性将提供最高水平的可靠性。



## 跨可用区的 EC2 实例的静态稳定性

您必须就这一模型的成本以及在所有情况下保持工作负载弹性的业务价值进行权衡。预置较少的计算容量并依赖于在出现故障时启动新实例，这种方法的成本较低，但是对于大规模故障（例如可用区或区域受损），这种方法效果较差，因为它既依赖于运营层面，也依赖未受影响的可用区或区域中是否有足够的可用资源。

您的解决方案还需要在工作负载的可靠性和成本需求之间做出取舍。静态稳定性架构适用于各种架构，包括跨多个可用区的计算实例、数据库只读副本设计、Kubernetes（Amazon EKS）集群设计和多区域失效转移架构。

您还可以在每个可用区中使用更多资源，从而实施具有更好的静态稳定性的设计。通过添加更多可用区，您可以减少实现静态稳定性所需的额外计算容量。

双模态行为的一个例子是，网络超时可能会导致系统尝试刷新整个系统的配置状态。这会向另一个组件添加意外负载，进而可能导致该组件出现故障，引发其他意外后果。此负面的反馈环路会影响您的工作负载的可用性。相反，您可以构建静态稳定的系统，并且仅在一个模式下运行。静态稳定的设计会持续工作，并且始终定期刷新配置状态。当调用失败时，工作负载将使用先前缓存的值并启动警报。

双模态行为的另一个示例是允许客户端在故障发生时绕过您的工作负载缓存。这看起来似乎是可以满足客户端需求的解决方案，但它会明显改变您的工作负载的需求，而且很有可能导致故障。

评估关键工作负载，以确定哪些工作负载需要这种弹性设计。对于被视为关键的工作负载，必须审核每个应用程序组件。需要静态稳定性评估的服务类型示例包括：

- 计算：Amazon EC2、EKS-EC2、ECS-EC2、EMR-EC2
- 数据库：Amazon Redshift、Amazon RDS、Amazon Aurora
- 存储：Amazon S3（单区）、Amazon EFS（挂载）、Amazon FSx（挂载）
- 负载均衡器：在某些设计下

### 实施步骤

- 构建静态稳定的系统，并且仅在一个模式下运行。在这种情况下，应在每个可用区或区域中预置足够的实例，以便在移除了一个可用区或区域时处理工作负载容量。您可以使用多种服务来路由到正常运行的资源，例如：
  - [跨区域 DNS 路由](#)
  - [MRAP Amazon S3 多区域路由](#)
  - [AWS Global Accelerator](#)

- [Amazon Application Recovery Controller \(ARC\)](#)
- 配置 [数据库只读副本](#)，来应对单个主实例或只读副本丢失的情况。如果流量由只读副本提供服务，则每个可用区和每个区域中的资源数量，应等于可用区或区域出现故障时的总体需求量。
- 在 Amazon S3 存储中配置关键数据，设计为在可用区出现故障时可确保所存储数据的静态稳定性。如果使用 [Amazon S3 One Zone-IA](#) 存储类，则不应将其视为静态稳定，因为丢失该可用区会将对所存储数据的可访问性降到最低。
- [负载均衡器](#) 有时服务于特定可用区，这可能是由于配置不正确，也可能是有意设计成这样。在这种情况下，静态稳定的设计可能需要采用更复杂的设计，将工作负载分布到多个可用区。出于安全、延迟或成本方面的考虑，可能会使用最初的设计来减少区域间流量。

## 资源

相关的 Well-Architected 最佳实践：

- [可用性定义](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)
- [REL11-BP04 恢复期间依赖于数据平面而不是控制平面](#)

相关文档：

- [尽可能减小灾难恢复计划中的依赖关系](#)
- [Amazon Builders' Library：使用可用区的静态稳定性](#)
- [故障隔离界限](#)
- [使用可用区的静态稳定性](#)
- [多可用区 RDS](#)
- [尽可能减小灾难恢复计划中的依赖关系](#)
- [跨区域 DNS 路由](#)
- [MRAP Amazon S3 多区域路由](#)
- [AWS Global Accelerator](#)
- [ARC](#)
- [单区 Amazon S3](#)
- [跨可用区负载均衡](#)



相关视频：

- [Static stability in AWS: AWS re:Invent 2019: Introducing The Amazon Builders' Library \(DOP328\)](#)

相关示例：

- [Amazon Builders' Library：使用可用区的静态稳定性](#)

REL11-BP06 当事件影响可用性时发出通知

在检测到突破阈值时发送通知，即使导致问题的事件已自动解决。

自动修复使您的工作负载变得可靠。不过，它也可能会掩盖需要处理的潜在问题。实施适当的监控和措施，以便检测问题的模式，包括那些被自动修复的问题，从而从根本上解决问题。

弹性系统经过精心设计，可以立即将性能下降事件传达给相应的团队。这些通知应通过一个或多个通信渠道发送。

期望的结果：在突破阈值 [例如错误率、延迟或其他重要的关键绩效指标 ( KPI ) ] 时，系统会立即向运营团队发送警报，以便尽快解决这些问题，从而避免或最大限度地减少对用户的影响。

常见反模式：

- 发送的警报过多。
- 发送不可操作的警报。
- 将警报阈值设置得太高 ( 过于敏感 ) 或太低 ( 不够敏感 ) 。
- 不针对外部依赖项发送警报。
- 在设计监控和警报时，没有考虑 [灰色故障](#) 。
- 执行自动修复，但未通知相应的团队需要进行该修复。

建立此最佳实践的好处：恢复通知可以让运营和业务团队了解到服务性能下降的情况，这样他们就可以立即做出反应，尽可能地缩短平均检测时间 ( MTTD , Mean Time To Detect ) 和平均修复时间 ( MTTR , Mean Time To Repair ) 。恢复事件通知还可确保您不会忽略不经常发生的问题。

未建立这种最佳实践的情况下暴露的风险等级：中。未能实施适当的监控和事件通知机制，会导致未能检测到出现的问题模式，包括那些被自动修复的问题。只有当用户联系客服或者在偶然的情况下，团队才会了解到系统性能下降的情况。

## 实施指导

在定义监控策略时，触发的警报是常见事件。此事件可能包含警报的标识符、警报状态（例如 IN ALARM 或 OK），以及触发该警报的对象的信息。在许多情况下，系统应该检测到警报事件并发送电子邮件通知。这是对警报执行操作的示例。警报通知对于可观测性至关重要，因为它会告知相关人员所存在的问题。不过，当您的可观测性解决方案能够针对事件采取合理的操作时，它可以自动修复问题，而无需人工干预。

建立 KPI 监控警报后，在超过阈值时，应向相应的团队发送警报。这些警报还可用于触发自动流程，尝试修复性能下降问题。

对于更复杂的阈值监控，应考虑使用复合警报。复合警报使用多个 KPI 监控警报，根据运营业务逻辑创建警报。CloudWatch 警报可被配置为发送电子邮件，或使用 Amazon SNS 集成或 Amazon EventBridge 将事件记录到第三方事件跟踪系统。

## 实施步骤

根据监控工作负载的方式，创建各种类型的警报，例如：

- 使用应用程序警报，检测工作负载的任何部分未正常工作的情况。
- [基础设施警报](#) 指明何时扩展资源。警报可以直观地显示在控制面板上，通过 Amazon SNS 或电子邮件发送，并与 Auto Scaling 结合使用来扩展或缩减工作负载资源。
- 您可以创建简单的 [静态警报](#)，用于监控在指定数量的评估周期内，指标突破静态阈值的情况。
- [复合警报](#) 可以处理来自多个来源的复杂警报。
- 创建警报后，请创建相应的通知事件。您可以直接调用 [Amazon SNS API](#) 来发送通知，并关联任何自动化的修复或通信机制。
- 集成 [Amazon Health Aware](#) 监控功能，以便监控可能出现性能下降的 AWS 资源。对于关键业务工作负载，此解决方案可提供对 AWS 服务的主动实时警报的访问。

## 资源

相关的 Well-Architected 最佳实践：

- [可用性定义](#)

相关文档：

- [基于静态阈值创建 CloudWatch 警报](#)

- [什么是 Amazon EventBridge ?](#)
- [什么是 Amazon Simple Notification Service ?](#)
- [发布自定义指标](#)
- [使用 Amazon CloudWatch 警报](#)
- [Amazon Health Aware \( AHA \)](#)
- [设置 CloudWatch 复合警报](#)
- [What's new in AWS Observability at re:Invent 2022](#)

相关工具：

- [CloudWatch](#)
- [CloudWatch X-Ray](#)

REL11-BP07 构造您的产品以满足可用性目标和正常运行时间服务等级协议 ( SLA )

构造您的产品以满足可用性目标和正常运行时间服务等级协议 ( SLA )。如果您公开或私下同意可用性目标或正常运行时间 SLA，请确认您设计了架构和运营流程来支持它们。

期望结果：每个应用程序的性能指标都有明确的可用性和 SLA 目标，可以监控和维护目标，以便符合业务成果。

常见反模式：

- 在不设置任何 SLA 的情况下设计和部署工作负载。
- 在没有理由或业务需求的情况下将 SLA 指标设置得很高。
- 在设置 SLA 时不考虑依赖项及其基础 SLA。
- 创建应用程序设计时不考虑弹性的责任共担模式。

建立此最佳实践的好处：根据关键弹性目标设计应用程序有助于实现业务目标和满足客户期望。这些目标有助于推动评估不同技术并考虑各种权衡的应用程序设计过程。

未建立此最佳实践暴露的风险等级：中

## 实施指导

应用程序设计必须考虑因业务、运营和财务目标而产生的各种要求。根据运营要求，工作负载需要具有特定的弹性指标目标，以便可以适当地监控和支持它们。不得在部署工作负载之后才设置或引出弹性指标。而应该在设计阶段定义这些指标，并帮助指导作出各种决策和权衡。

- 每个工作负载都应有自己的一组弹性指标。这些指标可能与其他业务应用程序的指标不同。
- 减少依赖项会对可用性产生积极影响。每个工作负载都应考虑其依赖项及其 SLA。一般来说，选择可用性目标等于或大于工作负载目标的依赖项。
- 在可能的情况下，考虑采用松散耦合设计，以便您的工作负载可以在依赖项受损的情况下正常运行。
- 减少控制面板依赖项，特别是在恢复或降级期间。评估对于任务关键型工作负载保持静态稳定的设计。使用资源节约来提高工作负载中这些依赖项的可用性。
- 可观测性和仪表化对于通过减少平均检测时间（MTTD）和平均修复时间（MTTR）来实现 SLA 至关重要。
- 更低的故障频率（更长的 MTBF）、更短的故障检测时间（更短的 MTTD）和更短的修复时间（更短的 MTTR）是用于提高分布式系统中的可用性的三个因素。
- 建立和满足工作负载的弹性指标，这是所有有效设计的基础。这些设计必须考虑设计复杂性、服务依赖项、性能、扩展和成本之间的权衡。

## 实施步骤

- 检查并记录工作负载设计，考虑以下问题：
  - 在工作负载中的什么地方使用控制面板？
  - 工作负载如何实施容错？
  - 扩缩、弹性伸缩、冗余和高可用性组件的设计模式是什么？
  - 数据一致性和可用性的要求是什么？
  - 是否考虑了资源节约或资源静态稳定性？
  - 服务依赖项是什么？
- 与利益攸关方合作时，根据工作负载架构定义 SLA 指标。考虑工作负载使用的所有依赖项的 SLA。
- 设置 SLA 目标后，优化架构以满足 SLA。
- 设置满足 SLA 的设计后，实施运营更改、流程自动化和运行手册，这些操作也将侧重于减少 MTTD 和 MTTR。
- 部署之后，监控和报告 SLA。

## 资源

### 相关最佳实践：

- [REL03-BP01 选择如何划分工作负载](#)
- [REL10-BP01 将工作负载部署到多个位置](#)
- [REL11-BP01 监控工作负载的所有组件以检测故障](#)
- [REL11-BP03 自动修复所有层](#)
- [REL12-BP05 使用混沌工程测试弹性](#)
- [REL13-BP01 定义停机和数据丢失的恢复目标](#)
- [了解工作负载运行状况](#)

### 相关文档：

- [通过冗余实现可用性](#)
- [可靠性支柱 - 可用性](#)
- [衡量可用性](#)
- [AWS 故障隔离界限](#)
- [弹性的责任共担模式](#)
- [使用可用区的静态稳定性](#)
- [AWS 服务等级协议 \( SLA \)](#)
- [AWS 上基于 cell 的架构指南](#)
- [AWS 基础设施](#)
- [高级多可用区弹性模式白皮书](#)

### 相关服务：

- [Amazon CloudWatch](#)
- [AWS Config](#)
- [AWS Trusted Advisor](#)

## REL 12.如何测试可靠性？

在为您的工作负载采用弹性设计以应对生产压力以后，测试是确保其按设计预期运行，并且提供您所预期弹性的唯一方式。

### 最佳实践

- [REL12-BP01 使用行动手册调查故障](#)
- [REL12-BP02 执行事后分析](#)
- [REL12-BP03 测试功能要求](#)
- [REL12-BP04 测试扩展和性能要求](#)
- [REL12-BP05 使用混沌工程测试弹性](#)
- [REL12-BP06 定期进行实际试用](#)

### REL12-BP01 使用行动手册调查故障

通过在行动手册中记录调查流程，实现对并不十分了解的故障场景做出一致且及时的响应。行动手册是在确定哪些因素导致故障场景时要执行的预定义步骤。所有流程步骤的结果都将用于确定要采取的后续步骤，直到问题得到确定或上报。

行动手册是您必须要执行的主动计划，以便有效采取响应措施。当在生产中遇到行动手册未涉及的故障场景时，首先要解决问题（灭火）。然后回过头来思考您在解决问题时采取的措施，并将这些措施作为新条目添加到行动手册中。

请注意，行动手册可用于对特定事件做出响应，运行手册则用来达成特定的结果。通常，运行手册适用于例行活动，而行动手册则被用于对非例行事件做出响应。

常见反模式：

- 计划在以下情况下部署工作负载：不清楚诊断问题或响应意外事件的流程。
- 关于在对事件进行调查时从哪些系统收集日志和指标的计划外的决定。
- 指标和事件保留的时间不够长，无法检索到数据。

建立此最佳实践的好处：使用行动手册可确保始终如一地遵循程序。编写行动手册可以减少手动操作导致的错误。通过实现行动手册自动化，可以消除团队成员干预的需要，或者在他们开始干预时便向他们提供更多信息，从而缩短事件响应时间。

未建立此最佳实践暴露的风险等级：高

## 实施指导

- 使用行动手册来发现问题。管理手册是用于调查问题的书面程序。在行动手册中记录流程，实现对故障场景的一致而及时的响应。行动手册必须包含所需的信息和指导，让足够熟练的员工能够收集适用信息、确定故障的潜在来源、隔离故障，并确定成因（在意外事件发生后执行分析）。
- 以代码形式实施行动手册。为行动手册编写脚本，以代码形式执行运营，以确保一致性并减少由手动流程引起的错误。行动手册可以由代表不同步骤的多个脚本组成，这些步骤可能是确定问题成因所必需的。系统可能会在运行手册活动过程中触发或执行行动手册活动，也可能针对响应发现的事件而提示执行行动手册活动。
  - [使用 AWS Systems Manager 自动执行您的运营手册](#)
  - [AWS Systems Manager Run Command](#)
  - [AWS Systems Manager Automation](#)
  - [什么是 AWS Lambda ?](#)
  - [什么是 Amazon EventBridge ?](#)
  - [使用 Amazon CloudWatch 告警](#)

## 资源

### 相关文档：

- [AWS Systems Manager Automation](#)
- [AWS Systems Manager Run Command](#)
- [使用 AWS Systems Manager 自动执行您的运营手册](#)
- [使用 Amazon CloudWatch 告警](#)
- [使用金丝雀 \( Amazon CloudWatch Synthetics \)](#)
- [什么是 Amazon EventBridge ?](#)
- [什么是 AWS Lambda ?](#)

### 相关示例：

- [使用行动手册和运行手册自动完成操作](#)



## REL12-BP02 执行事后分析

审核影响客户的事件，确定这些事件的成因和预防措施。利用这些信息来制定缓解措施，以限制或防止再次发生同类事件。制定程序以迅速有效地做出响应。根据目标受众，适当传达事件成因和纠正措施。如果需要，可将这些原因告知他人。

评估为什么现有测试找不到问题。如果还没有，增设测试。

期望结果：您的团队采用商定的一致方法来进行事后分析。一种机制是[错误更正 \(COE\) 流程](#)。COE 流程有助于您的团队识别、理解和解决事件的根本原因，同时还可以建立防护机制，以限制同一事件再次发生的可能性。

常见反面模式：

- 查找事件成因，但不继续深入探究其他潜在问题和缓解问题的方法。
- 只找出人为错误原因，但不提供任何培训或可防止人为错误的自动化功能。
- 只注重追究责任，而不去了解根本原因，营造恐惧文化，阻碍开诚布公的交流
- 见解分享不畅，事件分析结果仅限于一小群人知道，其他人无法从中吸取经验教训
- 没有收集制度性知识的机制，因此无法以最新最佳实践的形式保存经验教训，从而失去宝贵见解，导致根源相同或相似的事件反复发生

建立此最佳实践的好处：执行事后分析并且分享分析结果，可以减轻其他实施了相同故障因素的工作负载发生故障的风险，并且让它们能够在事件发生前就实施缓解措施或自动恢复。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

有效的事后分析让您有机会针对系统中其他地方使用的架构模式存在的问题提出常见的解决方案。

COE 流程的基础是记录和解决问题。建议定义一种标准化的方法来记录关键的根本原因，并确保其得到分析和处理。为事后分析流程分配明确的责任人。指派负责的团队或个人来监督事件调查和后续行动。

鼓励注重学习和改进而不是互相推脱责任的文化。强调目标是防止将来发生事件，而不是惩罚某些人。

为执行事后分析制定明确定义的程序。这些程序应概述要采取的步骤、要收集的信息以及在分析期间要解决的关键问题。彻底调查事件，除直接原因外，还要找出根本原因和影响因素。使用诸如[“五个为什么”](#)之类的技巧来深入研究潜在问题。

维护从事件分析中吸取的经验教训的存储库。这些制度性知识可以作为未来的事件和预防工作的参考。分享在事后分析中发现的结果和洞察，并考虑召开公开的事后总结会议，讨论经验教训。

## 实施步骤

- 在执行事后分析时，确保整个过程不是以追究责任为目的。这使事件中涉及到的人员能够冷静地看待建议的纠正措施，并促进诚实的自我评测和团队间的协作。
- 定义记录关键问题的标准化方法。此类文档的示例结构如下所示：
  - 发生了什么？
  - 客户和您的业务受到了什么影响？
  - 根本原因是什么？
  - 您有哪些数据来支持这一点？
    - 例如，指标和图表
  - 对关键支柱有什么影响，特别是在安全方面？
    - 在构造工作负载时，您需要基于您的业务环境在各个支柱之间做出权衡。这些业务决策可以确定设计优先事项。在开发环境中，您可能会通过降低可靠性来降低成本；而对于任务关键型解决方案，您可能会通过增加成本来提高可靠性。安全始终是头等大事，因为您必须保护您的客户。
  - 您获得了哪些经验教训？
  - 您采取了哪些纠正措施？
    - 行动项
    - 相关事项
- 为执行事后分析制定明确定义的标准操作程序。
- 设置标准化事件报告流程。全面记录所有事件，包括最初的事件报告、日志、通信和事件期间采取的行动。
- 请记住，并不是发生了停机才叫做事件。这可能是未遂事件，也可能是系统能够履行其业务功能，但以意外的方式运行。
- 根据反馈和经验教训，持续改进事后分析流程。
- 在知识管理系统中记录关键调查发现，并考虑应添加到开发人员指南或部署前清单中的任何模式。

## 资源

### 相关文档：

- [为什么您应该制定错误更正 \( COE \) 措施](#)

## 相关视频：

- [亚马逊成功应对失败的方法](#)
- [AWS re:Invent 2021 - Amazon Builders' Library：亚马逊的卓越运营](#)

## REL12-BP03 测试功能要求

使用的技术包括用于验证所需功能的单元测试和集成测试。

如果这些测试作为构建和部署措施的一部分自动运行，则您可以获得最佳的结果。例如，使用 AWS CodePipeline，开发人员会在 CodePipeline 自动检测到变更时提交对源存储库的更改。执行更改，然后加以测试。在测试完成以后，构建的代码会被部署到用于测试的暂存服务器。CodePipeline 会从暂存服务器运行更多测试，如集成或负载测试等。在成功完成此类测试以后，CodePipeline 会将经过测试并获得批准的代码部署到生产实例。

此外，过去的经验告诉我们，可运行合成事务测试（又被称作金丝雀测试，但不要和金丝雀部署相混淆）模拟用户行为，这是最重要的测试流程之一。从不同的远程位置针对您的工作负载端点持续地运行此类测试。Amazon CloudWatch Synthetics 使您能够 [创建 Canary](#) 以便监控您的终端节点和 API。

未建立此最佳实践暴露的风险等级：高

## 实施指导

- 测试功能要求。这包括用于验证所需功能的单元测试和集成测试。
  - [将 CodePipeline 与 AWS CodeBuild 结合使用以测试代码和运行构建](#)
  - [AWS CodePipeline 增加了对通过 AWS CodeBuild 进行单位和自定义集成测试的支持](#)
  - [持续交付和持续集成](#)
  - [使用金丝雀 \( Amazon CloudWatch Synthetics \)](#)
  - [软件测试自动化](#)

## 资源

## 相关文档：

- [AWS 合作伙伴：可帮助实施持续集成管道的合作伙伴](#)
- [AWS CodePipeline 增加了对通过 AWS CodeBuild 进行单位和自定义集成测试的支持](#)
- [AWS Marketplace：可用于实现持续集成的产品](#)

- [持续交付和持续集成](#)
- [软件测试自动化](#)
- [将 CodePipeline 与 AWS CodeBuild 结合使用以测试代码和运行构建](#)
- [使用金丝雀 \( Amazon CloudWatch Synthetics \)](#)

## REL12-BP04 测试扩展和性能要求

使用的技术包括负载测试以验证工作负载是否满足扩展和性能要求。

在云中，您可以按需为您的工作负载创建生产规模环境。如果在缩减的基础设施上运行这些测试，您必须根据您认为在生产中将会发生的情况扩展您观察到的结果。如果不想影响实际用户，您可以在生产开展负载和性能测试，并且对您的测试数据进行标记，以避免它与真实的用户数据、损坏的使用情况统计或生产报告混在一起。

通过测试确保您的基础资源、扩展设置、服务限额和弹性设计能够在负载之下如预期运行。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

- 测试扩展和性能要求。执行负载测试以验证工作负载是否满足扩展和性能要求。
  - [AWS 上的分布式负载测试：模拟数千个连接的用户](#)
  - [Apache JMeter](#)
    - 将您的应用程序部署在与生产环境相同的环境中，然后执行负载测试。
    - 使用基础设施即代码概念，以创建尽可能类似于生产环境的环境。

### 资源

相关文档：

- [AWS 上的分布式负载测试：模拟数千个连接的用户](#)
- [Apache JMeter](#)

## REL12-BP05 使用混沌工程测试弹性

在处于或尽可能接近生产的环境中定期运行混沌试验，以了解系统如何应对不利条件。

期望结果：

除了在事件期间验证已知预期工作负载行为的弹性测试之外，还可以通过以故障注入实验或注入意外负载的形式应用混沌工程，定期验证工作负载的弹性。将混沌工程和弹性测试结合起来，您可以提升信心，相信工作负载能够经受组件故障，并可从意外中断中恢复，而影响极小甚至没有影响。

常见反模式：

- 进行弹性设计，但不验证故障发生时工作负载如何作为一个整体运行。
- 从不在真实环境和预期负载下进行试验。
- 不将实验视为代码，也不在整个开发周期中维护它们。
- 不将混沌实验作为 CI/CD 管道的一部分，也不在部署之外运行。
- 在确定要对哪些故障进行试验时，没有想到使用过去的意外事件后分析。

建立此最佳实践的好处：注入故障来验证工作负载的弹性，这可以让您提升信心，相信您的弹性设计的恢复程序将在真正发生故障的情况下能够发挥作用。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

利用混沌工程，您的团队能够在服务提供商、基础设施、工作负载和组件级别，以可控的方式不断注入真实世界的干扰（模拟），而对客户的影响极小甚至没有影响。它使您的团队能够从故障中学习，观察、测量和提高工作负载的弹性，并验证在发生事件时，系统会发出警报并通知团队。

当持续执行时，混沌工程会突出工作负载中的缺陷，这些缺陷若不加以解决，可能会对可用性和运营产生负面影响。

#### Note

混沌工程是对系统进行试验以让人们确信系统能够在生产中经受住混乱情形的规范。 – [混沌工程的原则](#)

如果系统能够经受住这些干扰，那么应将混沌实验作为自动回归测试来加以维护。这样一来，应将混沌实验作为系统开发生命周期（SDLC）的一部分，以及作为 CI/CD 管道的一部分来执行。

为了确保您的工作负载能够承受组件故障，请在实验中注入实际事件。例如，对 Amazon EC2 实例的丢失或主 Amazon RDS 数据库实例的失效转移进行试验，并验证您的工作负载没有受到影响（或影响极小）。使用组件故障的组合来模拟可能因可用区中断而引起的事件。

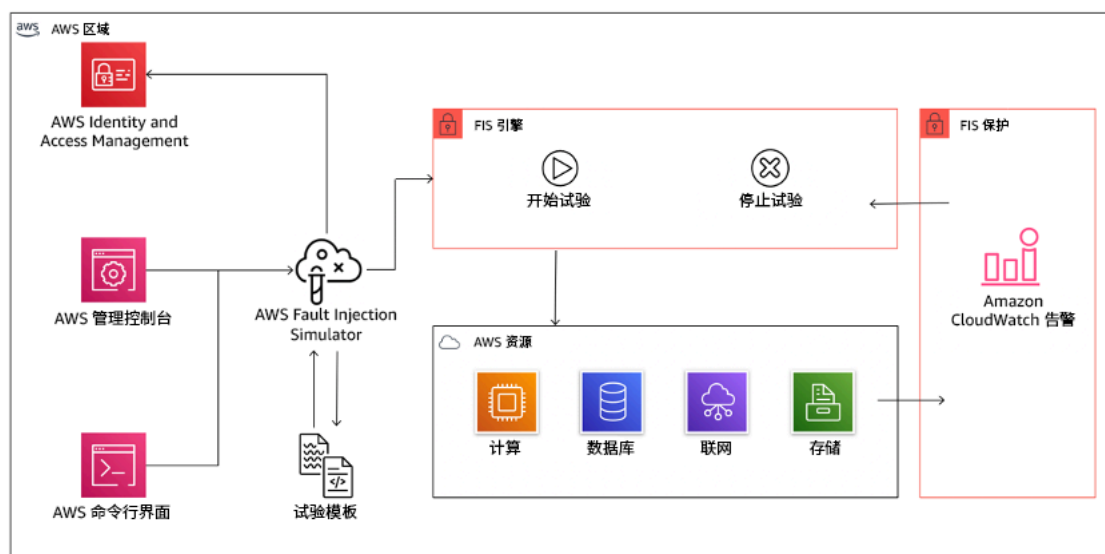
对于应用程序级故障（如崩溃），您可以从内存和 CPU 耗尽等压力源开始。

为了验证因间歇性网络中断而引发的外部依赖项的 [回退或失效转移机制](#)，您的组件应通过在指定时间段（从几秒到几小时不等）内阻止对第三方提供商的访问来模拟此类事件。

其他降级模式可能会影响功能的使用并降低响应速度，这通常会导致服务中断。性能下降的常见原因是，关键服务的延迟增加以及网络通信不可靠（丢包）。对于这些故障（包括延迟、丢弃的消息和 DNS 故障等网络效应）的实验可能包括无法解析名称、无法访问 DNS 服务或无法建立与依赖服务的连接。

混沌工程工具：

AWS Fault Injection Service (AWS FIS) 是一项完全托管式服务，用于运行故障注入实验，而这些实验可用作 CD 管道的一部分，或在管道之外使用。AWS FIS 是在混沌工程实际试用期间使用的一个不错的选择。它支持在不同类型的资源中同时引入故障，包括 Amazon EC2、Amazon Elastic Container Service (Amazon ECS)、Amazon Elastic Kubernetes Service (Amazon EKS) 和 Amazon RDS 等资源。这些故障包括终止资源、强制失效转移、对 CPU 或内存施加压力、节流、延迟和数据包丢失。由于它与 Amazon CloudWatch 警报集成，因此您可以设置停止条件作为防护机制，以在实验导致意外影响时回滚。



AWS Fault Injection Service 与 AWS 资源集成，使您能够为您的工作负载运行故障注入实验。

故障注入实验也有多种第三方选项。其中包括开源工具，例如 [Chaos ToolKit](#)、[Chaos Mesh](#) 和 [Litmus Chaos](#) 以及商业选项，如 Gremlin。为了扩大可在 AWS 上注入的故障范围，AWS FIS [与 Chaos Mesh 和 Litmus Chaos 集成](#)，使您能够在多个工具之间协调故障注入 workflow。例如，您可以使用 Chaos Mesh 或 Litmus 故障对容器组（pod）的 CPU 运行压力测试，同时使用 AWS FIS 故障操作终止随机选择的集群节点百分比。

## 实施步骤

- 确定哪些故障要用于实验。

评估工作负载的设计是否具有弹性。这种设计（使用 [Well-Architected Framework](#) 的最佳实践创建）考虑到了基于关键依赖关系、过去的事件、已知问题和合规性要求的风险。列出每个旨在保持弹性的设计元素及其旨在缓解的故障。有关创建此类列表的更多信息，请参阅 [《运营准备就绪情况审核》白皮书](#)，该白皮书指导您如何创建流程来防止以前的事件再次发生。故障模式与影响分析（FMEA）流程为您提供了一个框架，用于对故障及其对工作负载的影响执行组件级分析。Adrian Cockcroft 在 [《Failure Modes and Continuous Resilience》](#) 中更详细地概述了 FMEA。

- 为每个故障指定一个优先级。

先进行粗略的分类，如高、中或低。要评估优先级，请考虑故障的频率和故障对整体工作负载的影响。

考虑给定故障的频率时，请分析此工作负载的以往数据（如有）。如果没有以往数据，则使用在类似环境中运行的其他工作负载的数据。

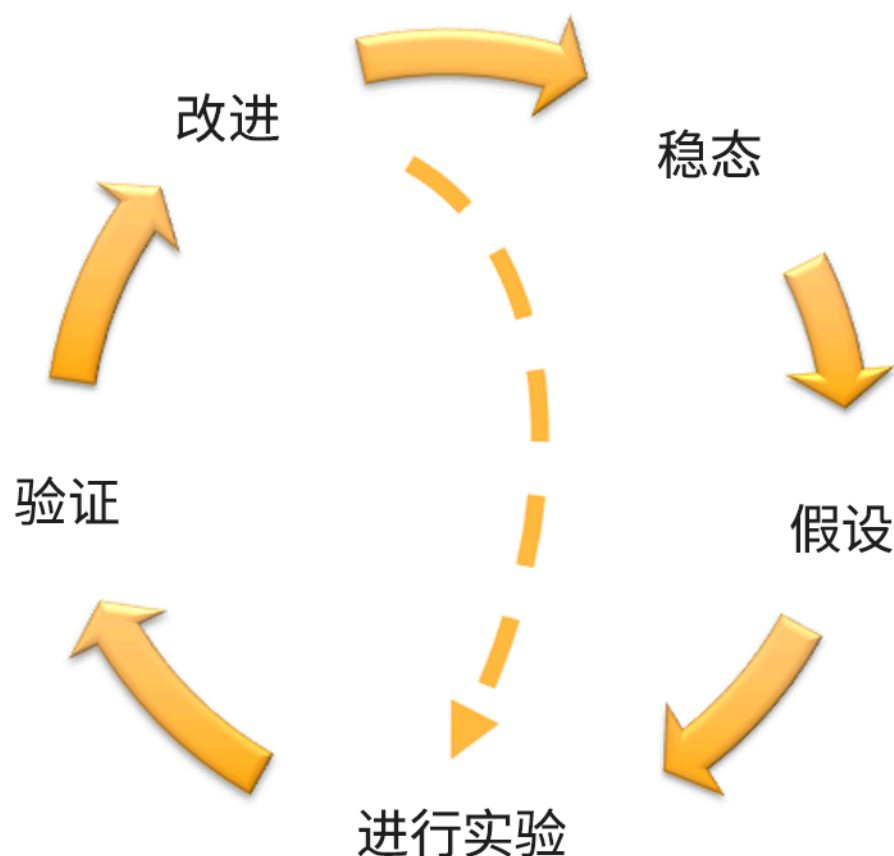
考虑给定故障的影响时，故障的范围越大，一般来说影响也越大。还要考虑工作负载设计和目的。例如，访问源数据存储的能力对于进行数据转换和分析的工作负载至关重要。在这种情况下，您将确定关于访问故障以及节流访问和延迟插入的实验的优先级。

意外事件后分析是了解故障模式的频率和影响的良好数据来源。

使用指定的优先级来确定首先用哪些故障进行实验，以及开发新的故障注入实验的顺序。

- 对于您执行的每项实验，请遵循混沌工程和连续弹性飞轮。





Adrian Hornsby 采用科学方法制作的混沌工程和连续弹性飞轮。

- 将稳态定义为指示正常行为的工作负载的一些可测量输出。

如果工作负载运行可靠且符合预期，则显示为稳态。因此，定义稳态之前，请验证您的工作负载正常运行。稳态并不一定意味着故障发生时对工作负载没有影响，因为一定百分比的故障可能在可接受的范围内。稳态是您在实验期间将观察到的基线，如果您在下一步中定义的假设结果不符合预期，它将突出显示异常。

例如，可以将某个支付系统的稳态定义为处理 300TPS，成功率为 99%，且往返时间为 500ms。

- 形成一个关于工作负载如何应对故障的假设。

一个好的假设是基于工作负载预计如何缓解故障以保持稳态。该假设指出，如果发生特定类型的故障，系统或工作负载将继续保持稳态，因为该工作负载在设计时就有特定的缓解措施。应在假设中具体说明特定的故障类型和缓解措施。

假设可以使用以下模板（但其他措辞也可以接受）：

 Note

如果发生 ##### ，则 ##### 工作负载将 ##### 维持 #####。

例如：

- 如果 Amazon EKS 节点组中 20% 的节点出现故障，则 Transaction Create API 将在不到 100ms 的时间内继续处理 99% 的请求（稳态）。Amazon EKS 节点将在五分钟内恢复，容器组（pod）将在实验开始后八分钟内得到调度并处理流量。警报将在三分钟内发出。
- 如果发生单个 Amazon EC2 实例故障，订单系统的 Elastic Load Balancing 运行状况检查将导致 Elastic Load Balancing 仅向剩余的运行状况良好的实例发送请求，而 Amazon EC2 Auto Scaling 将替换故障实例，从而保持服务器端（5xx）错误增长率低于 0.01%（稳态）。
- 如果主 Amazon RDS 数据库实例发生故障，则供应链数据收集工作负载将失效转移并连接到备用 Amazon RDS 数据库实例，以保持不到 1 分钟的数据库读写错误（稳态）。
- 通过注入故障来进行实验。

默认情况下，实验应具有故障保护机制，可承受工作负载。如果您知道工作负载将发生故障，则不要进行实验。混沌工程应该用于寻找已知的不确定因素或未知的不确定因素。已知的不确定因素是您知道但不完全理解的东西，而未知的不确定因素是您既不知道也不完全理解的东西。对您知道已经发生故障的工作负载进行试验不会为您提供新的见解。您应该对实验仔细规划，明确一个影响范围，并提供一种可在出现意外动荡时应用的回滚机制。如果尽职调查表明您的工作负载应该能经受住实验，那就继续这项实验。有几种注入故障的选项。对于 AWS 上的工作负载，[AWS FIS](#) 提供了许多称为 [操作](#) 的预定义故障模拟。您还可以定义在 AWS FIS 中运行的自定义操作（使用 [AWS Systems Manager 文档](#)）。

我们不鼓励使用自定义脚本进行混沌实验，除非这些脚本能够了解工作负载的当前状态，能够发出日志，并在可能的情况下提供回滚和停止条件的机制。

支持混沌工程的有效框架或工具集应跟踪实验的当前状态，发出日志，并提供回滚机制以支持实验的受控执行。从 AWS FIS 这样的成熟服务开始，该服务支持您在明确定义的范围内和安全机制下进行实验，如果实验引入了意外的动荡，则可以回滚实验。要了解更多信息使用 AWS FIS 的实验，另请参阅 [“通过混沌工程构建弹性且架构完善的应用程序”实验室](#)。此外，[AWS Resilience Hub](#) 将分析您的工作负载，并创建您可以选择在 AWS FIS 中实施和运行的实验。

**Note**

对于每项实验，要清楚地了解其范围及影响。我们建议首先在非生产环境中模拟故障，然后再在生产环境中运行。

应使用实际负载，通过 [金丝雀部署](#) 在生产环境中进行实验，尽可能同时启动控制和实验系统部署。在非高峰时间进行实验是一种很好的做法，可以减小首次在生产环境中试验时的潜在影响。此外，如果使用实际的客户流量会带来太大的风险，您可以在生产基础设施上针对控制和实验部署使用合成流量进行实验。当不能使用生产环境时，在尽可能接近生产环境的预生产环境中进行实验。

您必须建立和监控防护机制，确保实验对生产流量或其他系统的影响不会超过可接受的限度。建立停止条件，以便在实验达到您定义的防护机制指标的阈值时停止实验。这应该包括工作负载的稳态指标，以及针对您要注入故障的组件的指标。A [合成监控器](#)（也称为用户金丝雀）是一个通常应作为用户代理包含的指标。[AWS FIS 的停止条件](#) 应纳入实验模板中，每个模板最多可以有五个停止条件。

混沌的原则之一是尽量缩小实验范围并减小其影响：

虽然必须考虑到一些短期负面影响，但混沌工程师有责任和义务确保实验产生的影响极小且可控。

验证范围和潜在影响的一种方法是首先在非生产环境中进行实验（验证停止条件的阈值在实验期间按预期激活，并且可观测性到位以捕获异常），而不是直接在生产环境中进行实验。

运行故障注入实验时，确保所有责任方均知情。与适当的团队（如运营团队、服务可靠性团队和客户支持团队）沟通，让他们知道实验将在何时运行以及预期会发生什么。为这些团队提供沟通工具，以便在他们看到任何不利影响时通知进行实验的人员。

必须将工作负载及其底层系统恢复到最初的已知良好状态。通常，工作负载的弹性设计会自我修复。但一些故障设计或失败的实验可能会使您的工作负载处于意外的失败状态。在实验结束时，您必须意识到这一点，并恢复工作负载和系统。使用 AWS FIS，您可以在操作参数中设置回滚配置（也称为后期操作）。后期操作将目标返回到运行该操作之前的状态。无论是自动执行（如使用 AWS FIS）还是手动执行，这些后期操作都应包含在描述如何检测和处理故障的行动手册中。

- 验证假设。

[混沌工程的原则](#) 为如何验证工作负载的稳态提供了以下指导：

关注系统的可测量输出，而不是系统的内部属性。短时间内对该输出的测量构成了系统稳态的代理。整个系统的吞吐量、错误率和延迟百分比都可以是代表稳态行为的相关指标。通过关注实验过程中的系统行为模式，混沌工程验证系统确实在工作，而不是试图验证它如何工作。

在之前的两个示例中，我们包括了服务器端 ( 5xx ) 错误增长率低于 0.01% 和数据库读写错误持续时间不到 1 分钟的稳态指标。

5xx 错误是一个很好的指标，因为它们是工作负载客户端将直接经历的故障模式的结果。数据库错误测量适合作为故障的直接结果，但是还应补充一个客户端影响测量，例如失败的客户端请求或向客户端显示的错误。此外，在工作负载客户端直接访问的任何 API 或 URI 上包括一个合成监控器 ( 也称为用户金丝雀 )。

- 改进工作负载设计，以提高弹性。

如果未保持稳态，则调查如何改进工作负载设计以缓解故障，应用 [AWS Well-Architected 可靠性支柱](#) 的最佳实践。可以在 [AWS Builder's Library](#) 中找到其他指导和资源，其中包含有关如何 [改进运行状况检查](#) 或 [在应用程序代码中结合采用重试与回退](#) 的文章，等等。

实施这些更改后，再次进行实验 ( 如混沌工程飞轮中的虚线所示 )，以确定其有效性。如果验证步骤表明假设成立，那么工作负载将处于稳态，循环将继续。

- 定期进行实验。

混沌实验是一个循环，作为混沌工程的一部分，应定期进行实验。在工作负载满足实验的假设后，实验应实现自动化，作为 CI/CD 管道的回归部分持续运行。要了解如何做到这一点，请参阅关于 [如何使用 AWS CodePipeline 进行 AWS FIS 实验](#) 的博客。这个关于反复 [在 CI/CD 管道中进行 AWS FIS 实验](#) 的实验室使您能够动手实践。

故障注入实验也是实际试用的一部分 ( 请参阅 [REL12-BP06 定期进行实际试用](#) )。实际试用会模拟故障或事件，以便验证系统、流程和团队的响应。其目的是实际执行团队在发生意外事件时会执行的操作。

- 捕获和存储实验结果。

必须捕获并持久保存故障注入实验的结果。包括所有必要的信息 ( 如时间、工作负载和条件 )，以便以后能够分析实验结果和趋势。结果示例可能包括控制面板的屏幕截图、从指标数据库进行的 CSV 转储，或实验中事件和观察结果的手写记录。[使用 AWS FIS 进行实验记录](#) 可作为这种数据捕获的一部分。

## 资源

### 相关最佳实践：

- [REL08-BP03 将弹性测试作为部署的一部分进行集成](#)
- [REL13-BP03 测试灾难恢复实施以验证实施效果](#)

### 相关文档：

- [什么是 AWS Fault Injection Service ?](#)
- [什么是 AWS Resilience Hub ?](#)
- [混沌工程的原则](#)
- [混沌工程：规划您的首次实验](#)
- [弹性工程：学会接受故障](#)
- [混沌工程案例](#)
- [避免在分布式系统中回退](#)
- [用于混沌实验的金丝雀部署](#)

### 相关视频：

- [AWS re:Invent 2020：使用混沌工程测试弹性 \( ARC316 \)](#)
- [AWS re:Invent 2019：通过混沌工程提高弹性 \( DOP309-R1 \)](#)
- [AWS re:Invent 2019：在无服务器世界中执行混沌工程 \( CMY301 \)](#)

### 相关示例：

- [Well-Architected 实验室：第 300 级：测试 Amazon EC2、Amazon RDS 和 Amazon S3 的弹性](#)
- [“混沌工程在 AWS 上的应用”实验室](#)
- [“通过混沌工程构建弹性且架构完善的应用程序”实验室](#)
- [“无服务器混沌”实验室](#)
- [“使用 AWS Resilience Hub 测量和提高应用程序弹性”实验室](#)

### 相关工具：

- [AWS Fault Injection Service](#)
- AWS Marketplace : [Gremlin 混沌工程平台](#)
- [Chaos ToolKit](#)
- [Chaos Mesh](#)
- [Litmus](#)

## REL12-BP06 定期进行实际试用

利用实际试用活动，在尽可能接近生产环境的环境中（包括在生产环境中），与将参与实际故障情景的人员一起为应对事件和故障而练习如何使用您的程序。实际试用会强制执行相关措施，以确保生产事件不会影响用户。

实际试用会模拟故障或事件，以便测试系统、流程和团队的响应。其目的是实际执行团队在发生意外事件时会执行的操作。这将帮助您了解可以从哪些方面作出改进，并有助于培养组织处理各种事件的经验。这些操作应该定期进行，让团队建立起关于响应方式的肌肉记忆。

在非生产环境中对您的弹性设计进行测试以后，可通过 Game Day 确保生产中的一切按照计划运行。Game Day，尤其如果是首次开展，是所有人员都应该参加的活动，工程师和运营团队都会得到关于开展时间以及活动内容的信息。运行手册准备就绪。以规定的方式在生产系统中执行模拟事件（包括可能出现的故障事件），并评估影响。如果所有系统如设计运行，检测和自我修复不会产生或只会产生非常轻微的影响。但如果观察到负面影响，测试将会回滚，并且（使用运行手册）修复问题，在必要时手动修复。由于实际试用经常在生产中进行，所以应采取全部预防措施，以确保不会对客户造成可用性影响。

常见反模式：

- 记录您的程序，但不要执行。
- 不要让业务决策者参与测试练习。

建立此最佳实践的好处：定期执行实际试用可确保在发生实际事件时，所有员工都遵守策略和程序，并且能够验证这些策略和程序是否合适。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

- 安排实际试用，以定期运用运行手册和行动手册。生产事件中涉及的所有人员均需参与实际试用：业务负责人、开发人员、运营人员和事件响应团队。

- 运行负载或性能测试，然后运行故障注入。
- 寻找运行手册中的异常，并利用这些异常机会练习使用行动手册。
- 如果您违反运行手册，请完善运行手册或纠正相应行为。如果练习使用行动手册，请确定应使用的运行手册，或者创建一个新运行手册。

## 资源

### 相关文档：

- [什么是 AWS 实际试用？](#)

### 相关视频：

- [AWS re:Invent 2019：通过混沌工程提高弹性 \( DOP309-R1 \)](#)

### 相关示例：

- [AWS Well-Architected 实验室 – 测试弹性](#)

## REL 13.如何规划灾难恢复 ( DR ) ？

拥有适当的备份和冗余工作负载组件是您的 DR 策略的开始。[RTO 和 RPO](#) 是您恢复工作负载的目标。根据业务需求设置这些目标。通过实施策略来实现这些目标，同时考虑工作负载资源和数据的位置和功能。中断概率和恢复成本也是关键因素，有助于了解为工作负载提供灾难恢复的商业价值。

## 最佳实践

- [REL13-BP01 定义停机和数据丢失的恢复目标](#)
- [REL13-BP02 使用定义的恢复策略来实现恢复目标](#)
- [REL13-BP03 测试灾难恢复实施以验证实施效果](#)
- [REL13-BP04 管理 DR 站点或区域的配置偏差](#)
- [REL13-BP05 自动执行恢复](#)

### REL13-BP01 定义停机和数据丢失的恢复目标

工作负载具有恢复时间目标 ( RTO ) 和恢复点目标 ( RPO ) 。



恢复时间目标 (RTO) 是指服务中断和服务恢复之间的最大可接受延迟。这可以确定在服务不可用时被视为可接受的时间窗口。

恢复点目标 (RPO) 是指自上一个数据恢复点以来的最大可接受时间。这可以确定在上一个恢复点和服务中断之间可接受的数据丢失程度。

在为您的工作负载选择合适的灾难恢复 (DR, Disaster Recovery) 策略时, RTO 和 RPO 值是重要的考虑因素。这些目标由业务部门确定, 然后由技术团队用来选择和实施 DR 策略。

期望结果：

每个工作负载都有一个根据业务影响定义的指定 RTO 和 RPO。工作负载被分配到一个预定义的层, 该层定义服务可用性和可接受的数据丢失, 以及关联的 RTO 和 RPO。如果无法进行这样的分层, 那么可以为每个工作负载分配定制的分层, 用于以后创建层。在为工作负载选择灾难恢复策略实施时, 使用 RTO 和 RPO 作为主要考虑因素之一。在选择 DR 策略时还要考虑成本约束、工作负载依赖关系和运维需求。

对于 RTO, 了解基于中断持续时间的影响。是线性的还是非线性的影响? (例如, 四小时后, 您关闭一条生产线, 直到下一班开始)。

如下所示的灾难恢复矩阵可以帮助您了解工作负载的重要性与恢复目标之间的关系。(请注意, X 轴和 Y 轴的实际值应根据您组织的需求进行定制)。

| 灾难恢复矩阵 |          |         |         |         |        |        |
|--------|----------|---------|---------|---------|--------|--------|
|        |          | 恢复点目标   |         |         |        |        |
|        |          | 少于 1 分钟 | 少于 1 小时 | 少于 6 小时 | 少于 1 天 | 多于 1 天 |
| 恢复时间目标 | 少于 10 分钟 | 严重      | 严重      | 高       | 中      | 中      |
|        | 少于 2 小时  | 严重      | 高       | 中       | 中      | 低      |
|        | 少于 8 小时  | 高       | 中       | 中       | 低      | 低      |
|        | 少于 24 小时 | 中       | 中       | 低       | 低      | 低      |
|        | 多于 24 小时 | 中       | 低       | 低       | 低      | 低      |

图 16：灾难恢复矩阵

常见反模式：

- 未定义恢复目标。

- 选择任意恢复目标。
- 选择过于宽松并且不符合业务目标的恢复目标。
- 不了解停机和数据丢失的影响。
- 选择不切实际的恢复目标，如零恢复时间和零数据丢失，这对于您的工作负载配置可能无法实现。
- 选择比实际业务目标更严格的恢复目标。这将强制实施比工作负载所需的成本更高并且更复杂的 DR。
- 选择与所依赖工作负载的恢复目标不兼容的恢复目标。
- 您的恢复目标没有考虑法规合规性要求。
- 为工作负载定义了 RTO 和 RPO，但从未测试过。

建立此最佳实践的好处：在指导您的 DR 实施时，需要您的恢复时间目标和数据丢失恢复目标。

未建立此最佳实践暴露的风险等级：高

### 实施指导

对于给定的工作负载，您必须了解停机和数据丢失对业务的影响。随着停机时间或数据丢失的增加，影响通常会越来越大，但这种增长的形式可能会因工作负载类型而异。例如，您也许可以容忍长达一小时的停机时间而没有多大影响，但在一小时之后影响会迅速上升。对业务的影响表现为多种形式，包括货币成本（如收入损失）、客户信任（以及对声誉的影响）、运维问题（如错过工资发放或生产力下降）和监管风险。使用以下步骤了解这些影响，并为您的工作负载设置 RTO 和 RPO。

### 实施步骤

1. 确定此工作负载的业务利益相关者，并与他们一起实施这些步骤。工作负载的恢复目标是一项业务决策。然后，技术团队与业务利益相关者合作，使用这些目标来选择 DR 策略。

#### Note

对于步骤 2 和 3，您可以使用 [the section called “实施工作表”](#)。

2. 通过回答以下问题，收集必要的信息来做出决策。
3. 在组织中，您是否对工作负载影响的重要性进行了分类或分级？
  - a. 如果有，请将此工作负载分配到一个类别
  - b. 如果没有，则建立这些类别。创建不超过五个类别，并细化每个类别的恢复时间目标范围。类别示例包括：关键、高、中、低。要了解工作负载如何映射到类别，请考虑工作负载是任务关键型、业务重要型还是非业务驱动型。

- c. 根据类别设置工作负载 RTO 和 RPO。始终选择比进入此步骤时计算的原始值更严格的类别（更低的 RTO 和 RPO）。如果这导致值发生了不适当的较大改变，那么考虑创建一个新类别。
4. 根据这些答案，为工作负载分配 RTO 和 RPO 值。这可以直接完成，也可以通过将工作负载分配给预定义的服务层来完成。
5. 在工作负载团队和利益相关者可访问的位置，记录此工作负载的灾难恢复计划（DRP，disaster recovery plan），此计划是组织的 [业务连续性计划（BCP，Business Continuity Plan）](#) 的一部分
  - a. 记录 RTO 和 RPO，以及用于确定这些值的信息。包括用于评估工作负载对业务影响的策略
  - b. 除 RTO 和 RPO 之外，记录您根据灾难恢复目标正在跟踪或计划跟踪的其他指标
  - c. 在进行创建时，您将 DR 策略和运行手册的详细信息添加到此计划中。
6. 通过在如图 15 所示的矩阵中查找工作负载的重要性，您可以开始建立为组织定义的预定义服务层。
7. 根据实施 DR 策略（或 DR 策略的概念验证）之后，[the section called “REL13-BP02 使用定义的恢复策略来实现恢复目标”](#)测试此策略以确定工作负载的实际 RTC（Recovery Time Capability，恢复时间能力）和 RPC（Recovery Point Capability，恢复点能力）。如果这些能力没有达到所预期的恢复目标，那么，要么与您的业务利益相关者一起调整这些目标，要么对 DR 策略进行更改以便实现预期的目标。

## 主要问题

1. 在对业务产生严重影响之前，工作负载可以停止的最长时间是多少
  - a. 确定在工作负载中断时，每分钟业务的货币成本（直接财务影响）。
  - b. 请注意，影响并不总是线性的。影响可能在一开始是有限的，然后在超过一个关键时间点后迅速增加。
2. 在对业务造成严重影响之前，可以丢失的最大数据量是多少
  - a. 对于最关键的数据存储，请考虑此值。确定其他数据存储的各自关键性。
  - b. 如果工作负载数据丢失，是否可以重新创建？如果这在操作上比备份和还原更容易，那么根据用于重新创建工作负载数据的源数据的重要性来选择 RPO。
3. 此工作负载所依赖的工作负载（下游）或依赖于此工作负载的工作负载（上游）的恢复目标和可用性期望是什么？
  - a. 选择使此工作负载能够满足上游依赖项要求的恢复目标
  - b. 根据下游依赖项的恢复能力，选择可实现的恢复目标。非关键的下游依赖项（您可以“绕过”它们）可以排除。或者，处理关键的下游依赖项，在必要时提高其恢复能力。

## 其他问题

考虑以下问题，以及它们如何应用于此工作负载：

4. 根据中断类型（区域与可用区等），您是否有不同的 RTO 和 RPO？
5. 您的 RTO/RPO 是否会在特定时间（季节性、销售活动、产品发布）发生变化？如果是这样，不同的测量和时间边界是什么？
6. 如果工作负载中断，会有多少客户受到影响？
7. 如果工作负载中断，对声誉有何影响？
8. 如果工作负载中断，可能会产生哪些其他运营影响？例如，如果电子邮件系统不可用或工资单系统无法提交事务，则会影响员工的工作效率。
9. 工作负载 RTO 和 RPO 如何与业务线和组织 DR 策略保持一致？
10. 是否存在提供服务的内部合同义务？不履行这些义务会受到处罚吗？
11. 数据的监管或合规性约束是什么？

## 实施工作表

您可以将此工作表用于实施步骤 2 和 3。您可以调整此工作表以满足您的特定需求，例如添加其他问题。

| 步骤 2：主要问题         | 适用于工作负载？      | 工作负载 RTO | 工作负载 RPO | RTO 调整。 | RPO 调整。 | 说明                                                   |
|-------------------|---------------|----------|----------|---------|---------|------------------------------------------------------|
| [1] 工作负载可以停止的最长时间 |               |          |          |         |         | 以从中断开始到恢复的时间进行衡量                                     |
| [2] 可以丢失的最大数据量    |               |          |          |         |         | 以从最后一个已知的可恢复数据集算起的时间进行衡量                             |
| [3a] 上游依赖关系       |               |          |          |         |         | 输入最严格的上游恢复目标                                         |
| [3b] 下游依赖关系       |               |          |          |         |         | 输入最宽松的下游恢复目标                                         |
| [3a] 协调后的上游依赖关系   |               |          |          |         |         | 如果上游值小于当前值，而下游值大于当前值，则使用依赖关系进行协调，并在此处输入协调后的值         |
| [3b] 协调后的下游依赖关系   |               |          |          |         |         |                                                      |
| [3] 依赖关系          |               |          |          |         |         | 降低值以满足上游依赖关系，或者根据下游依赖关系的能力提高值                        |
| 步骤 2：其他问题         |               |          |          |         |         | 指出问题是否适用。如果问题不适用，则跳过                                 |
| 基本 RTO/RPO        |               |          |          |         |         | 将 RTO 和 RPO 值从上方向下移到此处                               |
| [4] 中断类型          | [ ] Y / [ ] N |          |          |         |         | 输入要求最严格的事件类型的恢复目标                                    |
| [5] 基于时间的具体目标     | [ ] Y / [ ] N |          |          |         |         | 输入要求最严格的恢复时间目标                                       |
| [6] 客户中断          | [ ] Y / [ ] N |          |          |         |         | 根据停机时间或数据丢失情况绘制受影响客户的图表。<br>根据客户影响输入允许的最大的 RTO 和 RPO |
| [7] 声誉影响          | [ ] Y / [ ] N |          |          |         |         | 与业务部门合作，根据对声誉的影响确定最大的 RTO 和 RPO                      |
| [8] 运营影响          | [ ] Y / [ ] N |          |          |         |         | 根据运营影响输入最大的 RTO 和 RPO                                |
| [9] 组织一致性         | [ ] Y / [ ] N |          |          |         |         | 根据 LOB 和组织要求，输入此类工作负载的最大 RTO 和 RPO                   |
| [10] 合同义务         | [ ] Y / [ ] N |          |          |         |         | 根据合同义务输入最大的 RTO 和 RPO                                |
| [11] 监管合规性        | [ ] Y / [ ] N |          |          |         |         | 根据适用的监管合规性输入最大的 RTO 和 RPO                            |
| 基于其他问题的目标         |               |          |          |         |         | 从问题 4-11 中取最小值（更严格的值）并在此处输入                          |
| 调整后的目标            |               |          |          |         |         | 如果无法满足上述目标，请与利益相关者一起放松约束，并在此处输入新的最小值                 |
| 调整后的 RTO/RPO      |               |          |          |         |         | 输入基本 RPO/RTO 值或调整后的目标值，以较低者为准                        |
| 步骤 3              |               |          |          |         |         |                                                      |
| 映射到预定义类别或层        |               |          |          |         |         | 将这两个值向下调整（更严格），以符合最接近的定义层                            |

## 工作表

实施计划的工作量级别：低

## 资源

相关最佳实践：

- [the section called “REL09-BP04 定期执行数据恢复以验证备份完整性和流程”](#)
- [the section called “REL13-BP02 使用定义的恢复策略来实现恢复目标”](#)
- [the section called “REL13-BP03 测试灾难恢复实施以验证实施效果”](#)

相关文档：

- [AWS 架构博客：灾难恢复系列](#)
- [AWS 上工作负载的灾难恢复：云中的恢复 \( AWS 白皮书 \)](#)
- [使用 AWS Resilience Hub 管理弹性策略](#)
- [AWS 合作伙伴：可以帮助进行灾难恢复的合作伙伴](#)
- [AWS Marketplace：可以用于灾难恢复的产品](#)

相关视频：

- [AWS re:Invent 2018：适用于多区域主动-主动应用程序的架构模式 \( ARC209-R2 \)](#)
- [AWS 上工作负载的灾难恢复](#)

## REL13-BP02 使用定义的恢复策略来实现恢复目标

定义满足工作负载恢复目标的灾难恢复 ( DR ) 策略。选择一种策略，例如备份和还原、备用 ( 主动/被动 ) 或主动/主动。

期望结果：对于每个工作负载，都有一个已定义和实施的 DR 策略，使该工作负载能够实现 DR 目标。工作负载之间的 DR 策略利用可重用模式 ( 例如，前面描述的策略 )。

常见反模式：

- 为具有类似 DR 目标的工作负载实施不一致的恢复过程。
- 在发生灾难时临时实施 DR 策略。
- 没有针对灾难恢复的计划。

- 恢复期间依赖于控制面板操作。

建立此最佳实践的好处：

- 通过定义恢复策略，您可以使用常用工具和测试步骤。
- 使用定义的恢复策略，改进团队之间的知识共享，并在他们自己的工作负载上实施 DR。

在未建立这种最佳实践的情况下暴露的风险等级：高。若没有经过计划、实施和测试的 DR 策略，在发生灾难时不太可能实现恢复目标。

实施指导

DR 策略依赖于在主位置无法运行工作负载的情况下，在恢复站点中支持工作负载的能力。最常见的恢复目标是 RTO 和 RPO，相关讨论内容位于 [REL13-BP01 定义停机和数据丢失的恢复目标](#)。

跨单个 AWS 区域 内的多个可用区 ( AZ ) 的 DR 策略可以缓解火灾、洪水和重大停电等灾难事件。如果需要实施保护措施，为工作负载无法在给定 AWS 区域 中运行这种不太可能发生的事件提供保护，您可以使用跨多个区域的 DR 策略。

在跨多个区域构建 DR 策略时，您应该选择以下策略之一。这些策略按成本和复杂性升序排列，按 RTO 和 RPO 降序排列。 恢复区域指的是 AWS 区域，而不是用于工作负载的主要区域。

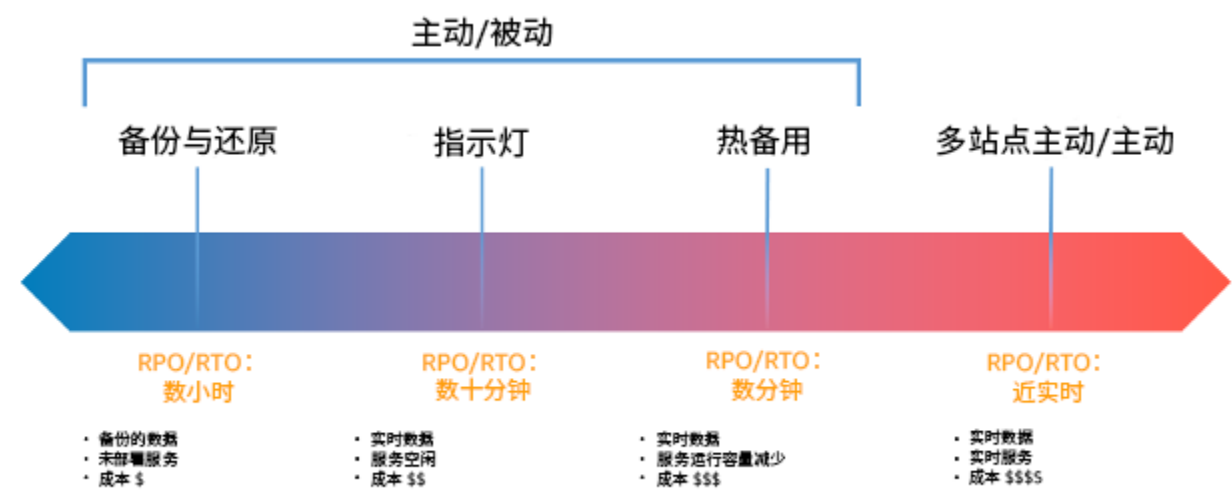


图 17：灾难恢复 ( DR ) 策略

- 备份和还原 ( RPO 以小时为单位，RTO 为 24 小时或更短 )：将您的数据和应用程序备份到恢复区域。使用自动或连续备份可以实现时间点故障恢复，在某些情况下，可以将 RPO 降低到 5 分钟。在



发生灾难的情况下，您将部署基础设施（使用基础设施即代码来减少 RTO）、部署代码并还原备份的数据，以便在恢复区域从灾难中恢复。

- 指示灯（RPO 以分钟为单位，RTO 为数十分钟）：在恢复区域中预置核心工作负载基础设施的副本。将您的数据复制到恢复区域并在那里创建数据备份。支持数据复制和备份所需的资源（如数据库和对象存储）始终处于启用状态。其他元素（如应用程序服务器或无服务器计算）未部署，但可以在需要时使用必要的配置和应用程序代码创建。
- 热备用（RPO 以秒为单位，RTO 以分钟为单位）：保证在恢复区域中始终运行缩减但功能齐全版本的工作负载。业务关键型系统是完全重复，而且始终可用的系统，只是其队列的规模经过缩减。数据在恢复区域中复制并留存。在需要恢复时，系统会快速扩展以处理生产负载。热备用的规模越大，RTO 和控制面板依赖度就越低。当完全扩展时，这称为热备用服务器。
- 多区域（多站点）主动-主动（RPO 接近于零，RTO 可能为零）：您的工作负载部署到多个 AWS 区域，并且主动处理来自这些区域的流量。此策略要求您跨区域同步数据。必须避免或处理在两个不同区域副本中写入同一记录可能引起的冲突，这会很复杂。数据复制对于数据同步非常有用，并且可以防止某些类型的灾难，但是它不能防止数据损坏或破坏，除非您的解决方案还包含时间点故障恢复选项。

#### Note

指示灯和热备用之间的差异有时难以区分。两者都在恢复区域中包含一个环境，其中具有主区域资产的副本。区别在于，如果不先采取额外措施，指示灯无法处理请求，而热备用可以立即处理流量（容量级别降低）。指示灯将要求您启用服务器，可能需要部署额外的（非核心）基础设施并纵向扩展，而热备用只需要您纵向扩展（所有内容都已部署并运行）。根据您的 RTO 和 RPO 需求在两者之间进行选择。

当成本是一个问题，并且您希望实现与热备用策略中定义的类似 RPO 和 RTO 目标时，您可以考虑云原生解决方案（例如，AWS Elastic Disaster Recovery），该解决方案采用指示灯方法并提供改进的 RPO 和 RTO 目标。

## 实施步骤

### 1. 确定将满足此工作负载恢复要求的 DR 策略。

选择 DR 策略是在减少停机时间和数据丢失（RTO 和 RPO）与策略实施的成本和复杂性之间进行权衡。您应该避免实施比所需策略更严格的策略，因为这会产生不必要的成本。



例如，在下图中，企业已经确定了他们允许的最大 RTO 以及他们可以在服务恢复策略上花费的费用限额。鉴于企业目标，指示灯或热备用这样的 DR 策略将同时满足 RTO 和成本标准。

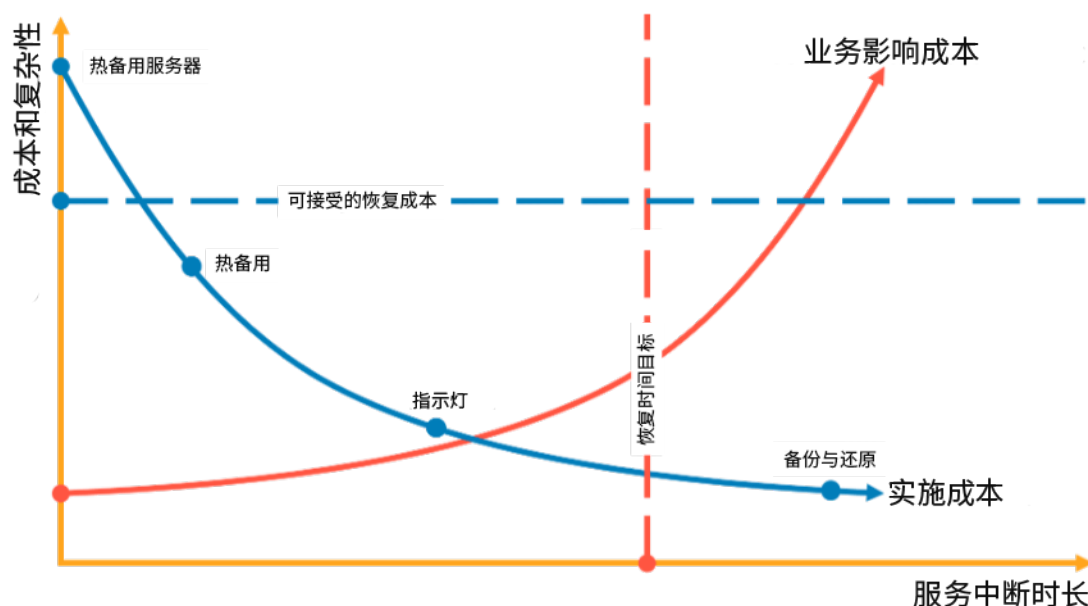


图 18：根据 RTO 和成本选择 DR 策略

如需了解更多信息，请参阅[业务连续性计划 \(BCP\)](#)

## 2. 查看如何实施所选 DR 策略的模式。

这一步是了解如何实施所选策略。这些策略可以解释为使用多个 AWS 区域 作为主要站点和恢复站点。不过，您也可以选择使用单个区域内的多个可用区作为 DR 策略，这将利用多个策略的元素。

在后续步骤中，您可以对特定的工作负载应用策略。

### 备份和还原

备份和还原是实施起来最简单的策略，但需要更多时间和工作来恢复工作负载，从而导致更高的 RTO 和 RPO。最好的做法是，始终备份数据并将数据备份复制到另一个站点（如另一个 AWS 区域）。

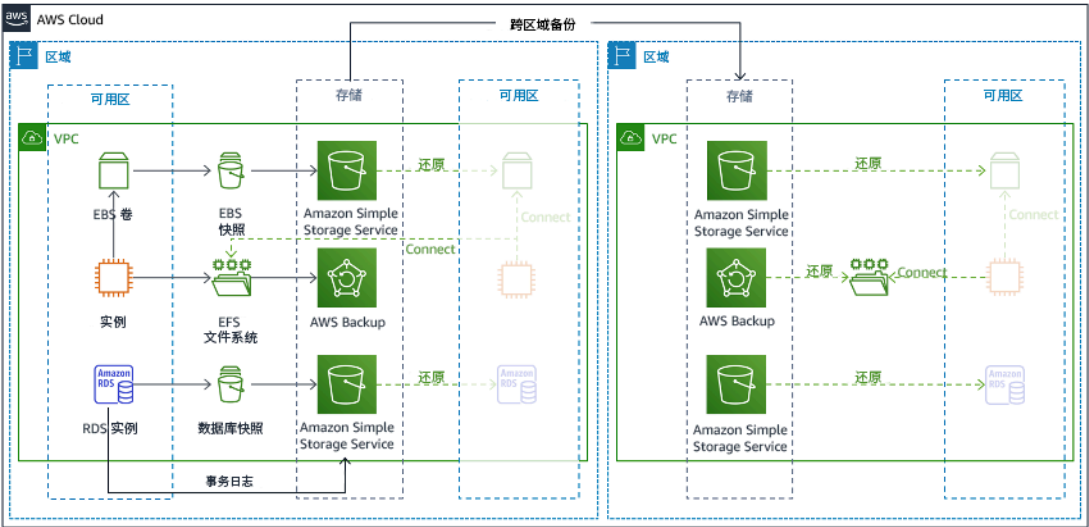


图 19：备份和还原架构

有关此策略的更多详细信息，请参阅 [AWS 上的灾难恢复 \(DR\) 架构，第 II 部分：使用快速恢复功能的备份与还原](#)。

指示灯

利用指示灯方法，您可以将数据从主要区域复制到恢复区域。用于工作负载基础设施的核心资源部署在恢复区域中，但仍需要额外的资源和所有依赖项才能使此恢复区域成为功能堆栈。例如，在图 20 中，没有部署计算实例。

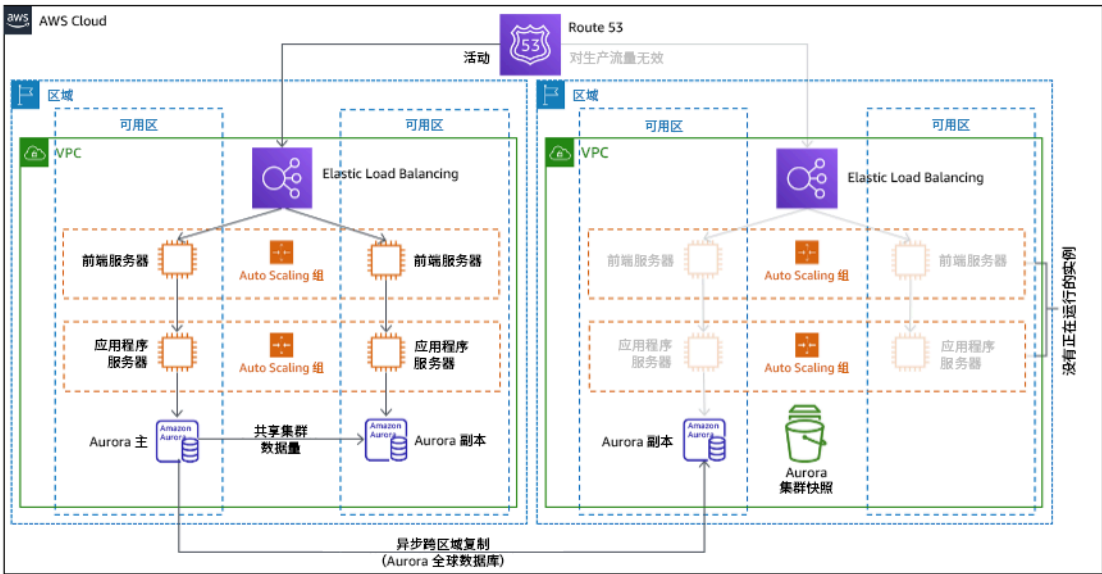


图 20：指示灯架构

有关此策略的更多详细信息，请参阅 [AWS 上的灾难恢复 \(DR\) 架构，第 III 部分：指示灯和热备用。](#)

热备用

热备用方法涉及到确保在另一个区域中存在生产环境的规模缩减但功能齐全的副本。这种方法扩展了指示灯概念并减少了恢复时间，因为您的工作负载始终在另一个区域中运行。如果以全部容量部署恢复区域，那么这种方式称为热备用。

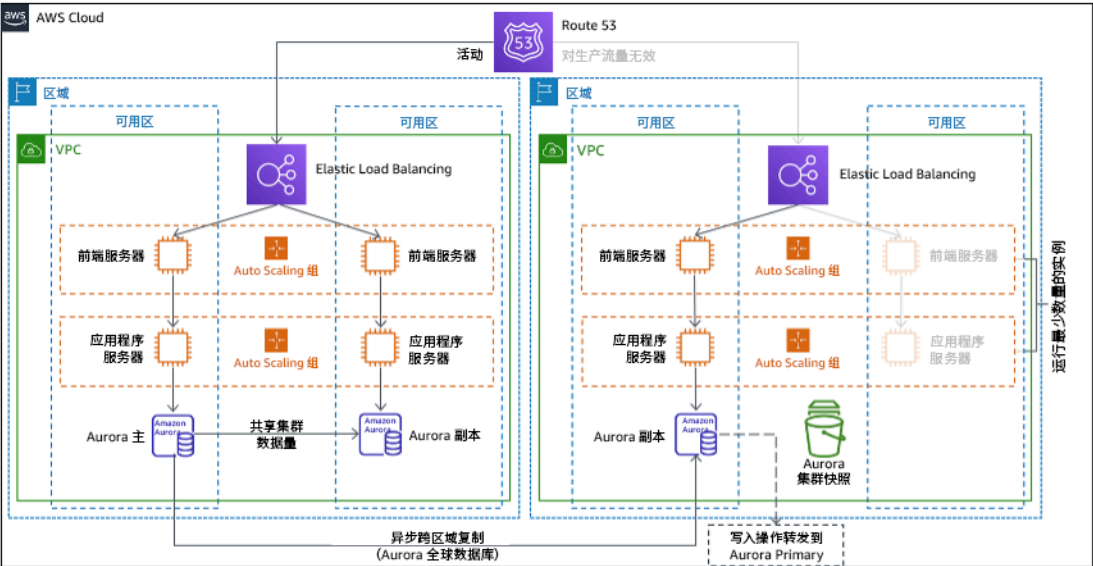


图 21：热备用架构

使用热备用或指示灯需要扩展恢复区域中的资源。为确保在需要时有可用的容量，请考虑使用 EC2 实例的[容量预留](#)。如果使用 AWS Lambda，那么[预置并发](#)可以提供执行环境，以便它们准备好立即响应函数的调用。

有关此策略的更多详细信息，请参阅 [AWS 上的灾难恢复 \(DR\) 架构，第 III 部分：指示灯和热备用。](#)

多站点主动/主动

作为多站点主动/主动策略的一部分，您可以在多个区域中同时运行工作负载。多站点主动/主动策略处理来自其部署到的所有区域的流量。客户可能会出于 DR 以外的原因选择此策略。此策略可以用于提高可用性，或者在向全球受众部署工作负载时（使端点更靠近用户和/或部署针对该区域受众的本地化堆栈）使用此策略。作为一种 DR 策略，如果工作负载在部署此策略的某个 AWS 区域中不能得到支持，那么该区域将被撤出，使用其余区域维持可用性。多站点主动/主动策略是 DR 策略中操作最复杂的策略，只有在业务需求时才应选择它。

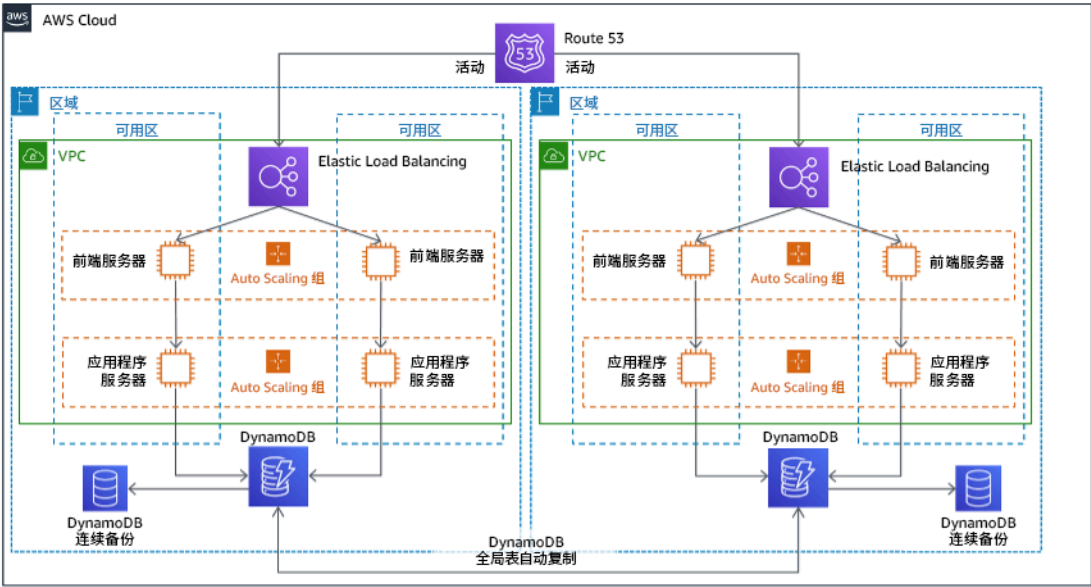


图 22：多站点主动/主动架构

有关此策略的更多详细信息，请参阅 [AWS 上的灾难恢复 \(DR\) 架构，第 IV 部分：多站点主动/主动](#)。

AWS Elastic Disaster Recovery

如果您正考虑为灾难恢复使用指示灯或热备用策略，AWS Elastic Disaster Recovery 可以提供一种带来更多好处的替代方法。Elastic Disaster Recovery 可以提供类似于热备用方法的 RPO 和 RTO 目标，同时保持指示灯方法的低成本。Elastic Disaster Recovery 将数据从主区域复制到恢复区域，使用持续数据保护来实现以秒为单位的 RPO 和以分钟为单位的 RTO。在恢复区域中仅部署复制数据所需的资源，从而降低成本，类似于指示灯策略。使用 Elastic Disaster Recovery 时，如果在失效转移或演练过程中启动，则服务会协调和编排计算资源的恢复。

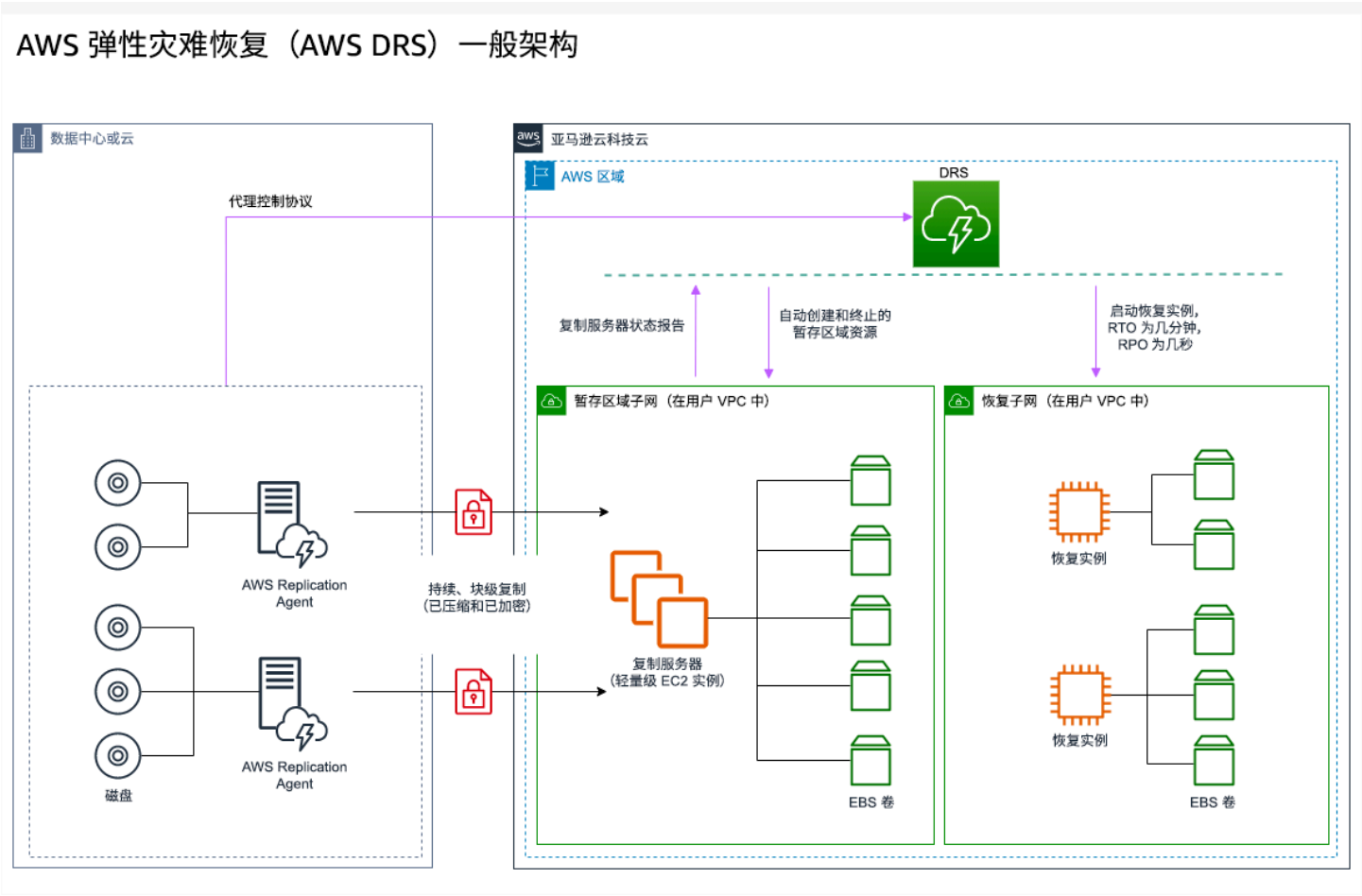


图 23 : AWS Elastic Disaster Recovery 架构

其他保护数据的实践

对于所有这些策略，您还必须减轻数据灾难的影响。持续的数据复制可以防止某些类型的灾难，但它可能无法防止数据损坏或破坏，除非您的策略还包括存储数据的版本控制或用于时间点故障恢复的选项。除了副本之外，您还必须备份恢复站点中的复制数据以创建时间点备份。

使用单个 AWS 区域 内的多个可用区 ( AZ )

使用单个区域内的多个可用区时，您的 DR 实施会使用上述策略的多个元素。首先，您必须使用多个可用区创建一个高可用性 ( HA ) 架构，如图 23 所示。此架构使用多站点主动/主动方法，因为 [Amazon EC2 实例](#)和 [Elastic Load Balancer](#) 在多个可用区中部署了资源，主动处理请求。此架构还演示了热备用服务器方法，如果主 [Amazon RDS](#) 实例出现故障 ( 或可用区本身出现故障 )，则备用实例将提升为主实例。

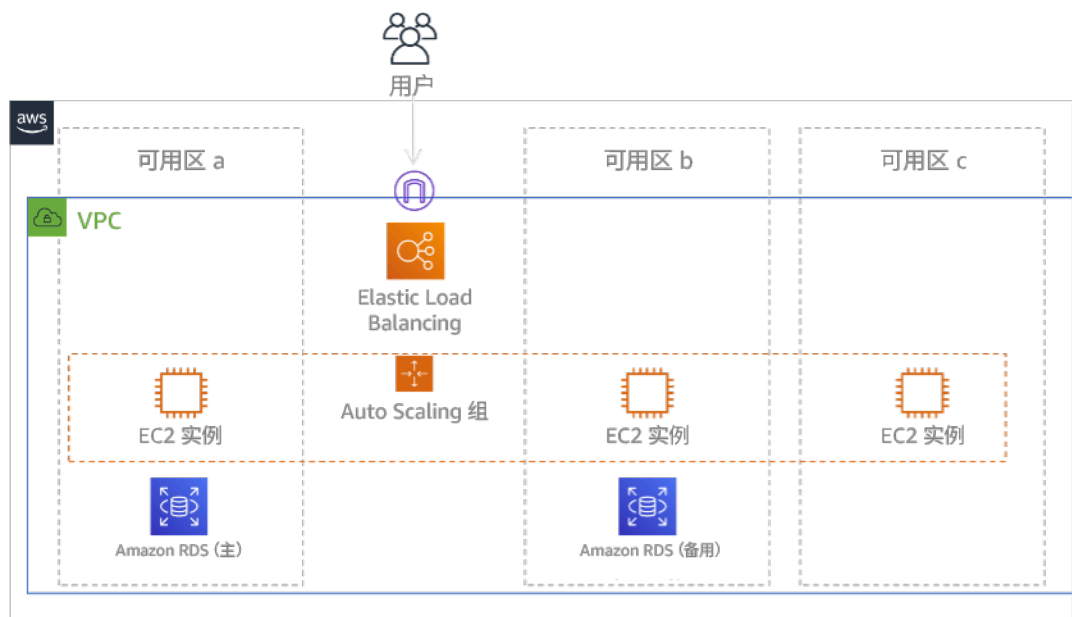


图 24：多可用区架构

除了这种 HA 架构之外，您还需要添加运行工作负载所需的所有数据的备份。这对于限制在单个区的数据尤其重要，例如 [Amazon EBS 卷](#) 或 [Amazon Redshift 集群](#)。如果一个可用区发生故障，您需要将这些数据恢复到另一个可用区。如果可能，您还应该将数据备份复制到另一个 AWS 区域，提供另一层保护。

下面的博客文章中介绍了一种不太常见的单区域多可用区 DR 的替代方法：[使用 Amazon Route 53 Application Recovery Controller 构建高弹性应用程序，第 1 部分：单区域堆栈](#)。这里的策略是尽可能保持可用区之间的隔离，就像区域的运作方式一样。使用这种替代策略，您可以选择主动/主动或主动/被动方法。

#### Note

某些工作负载具有数据驻留法规要求。如果这适用于当前只有一个 AWS 区域的位置的工作负载，那么多区域将不适合您的业务需求。多可用区策略可以很好地抵御大多数灾难。

### 3. 评估工作负载的资源，以及失效转移之前（正常操作期间）恢复区域中的资源配置。

对于基础设施和 AWS 资源，使用基础设施即代码功能（如 [AWS CloudFormation](#)）或第三方工具（如 Hashicorp Terraform）。要使用单个操作跨多个账户和区域部署，您可以使用 [AWS CloudFormation StackSets](#)。对于多站点主动/主动和热备用服务器策略，恢复区域中部署的基础设施具有与主区域相同的资源。对于指示灯和热备用策略，部署的基础设施将需要额外的操作才可用于生产。使用



CloudFormation [参数](#)和[条件逻辑](#)，您可以通过[单个模板](#)控制部署的堆栈是活动还是备用。使用 Elastic Disaster Recovery 时，服务会复制和编排应用程序配置和计算资源的还原。

所有 DR 策略都要求在 AWS 区域内备份数据源，然后将这些备份复制到恢复区域。[AWS Backup](#) 提供了一个集中视图，您可以在其中配置、调度和监控这些资源的备份。对于指示灯、热备用和多站点主动/主动方法，您还应该将数据从主区域复制到恢复区域中的数据资源，例如 [Amazon Relational Database Service \( Amazon RDS \)](#) 数据库实例或 [Amazon DynamoDB](#) 表。因此，这些数据资源处于活动状态，可以随时处理恢复区域中的请求。

要了解更多关于 AWS 服务如何跨区域运行的信息，请参阅以下博客系列：[使用 AWS 服务创建多区域应用程序](#)。

4. 确定并实施措施，让恢复区域在需要时（在灾难事件期间）可以进行失效转移。

对于多站点主动/主动策略，失效转移意味着撤离一个区域，并依赖剩余的活动区域。通常，这些区域已准备好接受流量。对于指示灯和热备用策略，恢复操作将需要部署缺失的资源（如图 20 中的 EC2 实例），以及任何其他缺失的资源。

对于上述所有策略，您可能需要将数据库的只读实例提升为主读/写实例。

对于备份和还原，从备份中还原数据时会为该数据创建资源，例如 EBS 卷、RDS 数据库实例和 DynamoDB 表。您还需要还原基础设施并部署代码。您可以使用 AWS Backup 来还原恢复区域中的数据。请参阅 [REL09-BP01 识别和备份需要备份的所有数据，或从源复制数据](#) 了解更多详细信息。重建基础设施包括创建资源，例如，EC2 实例以及所需的 [Amazon Virtual Private Cloud \( Amazon VPC \)](#)、子网和安全组。您可以自动执行大部分还原过程。要了解具体方法，请参阅[这篇博客文章](#)。

5. 确定并实施措施，以在需要时（在灾难事件期间）可以重新路由流量进行失效转移。

此失效转移操作可以自动或手动启动。应谨慎使用基于运行状况检查或警报自动启动的失效转移，因为不必要的失效转移（误报）会产生不可用和数据丢失等成本。因此，通常会使用手动启动的失效转移。在这种情况下，您仍然应该自动执行失效转移步骤，这样手动启动就像按一下按钮一样简单。

在使用 AWS 服务时，需要考虑几个流量管理选项。一个选项是使用 [Amazon Route 53](#)。使用 Amazon Route 53，您可以将一个或多个 AWS 区域中的多个 IP 端点与一个 Route 53 域名相关联。要实施手动启动的失效转移，您可以使用 [Amazon Route 53 Application Recovery Controller](#)，它提供高度可用的数据面板 API 以将流量重新路由到恢复区域。实施失效转移时，使用数据面板操作并避免控制面板操作，如以下部分所述：[REL11-BP04 恢复期间依赖于数据平面而不是控制平面](#)。

要了解有关此选项及其他选项的更多信息，请参阅[灾难恢复白皮书的这一部分](#)。



## 6. 设计工作负载的失效自动恢复计划。

失效自动恢复是指在灾难事件消除后将工作负载运营恢复到主区域。向主区域预置基础设施和代码通常遵循最初使用的相同步骤，依赖于基础设施即代码和代码部署管道。失效自动恢复面临的挑战是还原数据存储，并确保它们与运行中的恢复区域保持一致。

在失效转移状态下，恢复区域中的数据库处于活动状态，并且具有最新数据。然后，目标是从恢复区域重新同步到主区域，确保主区域是最新的。

某些 AWS 服务会自动执行此操作。如果使用 [Amazon DynamoDB 全局表](#)，即使主区域中的表不可用，当它重新联机时，DynamoDB 也会继续传播任何挂起的写操作。如果使用 [Amazon Aurora 全局数据库](#) 并使用 [托管的计划失效转移](#)，则维护 Aurora 全局数据库的现有复制拓扑。因此，主区域中以前的读/写实例将成为副本，并从恢复区域接收更新。

如果这不是自动执行的，您将需要在主区域中重新建立数据库，作为恢复区域中数据库的副本。在许多情况下，这将涉及删除旧的主数据库，然后创建新的副本。例如，有关如何使用 Amazon Aurora 全局数据库对计划外失效转移执行此操作的说明，请参阅下面的实验：[全局数据库的失效自动恢复](#)。

失效转移后，如果您可以继续在此恢复区域中运行，请考虑将此区域设为新的主区域。您仍然需要执行上述所有步骤，将以前的主区域变成恢复区域。有些组织会进行定期轮换，定期交换其主区域和恢复区域（例如每三个月一次）。

失效转移和失效自动恢复所需的所有步骤都应保存在行动手册且可供所有团队成员使用，并定期进行审查。

使用 Elastic Disaster Recovery 时，服务会协助编排和自动执行失效自动恢复流程。有关更多详细信息，请参阅[执行失效自动恢复](#)。

实施计划的工作量级别：高

资源

相关最佳实践：

- [the section called “REL09-BP01 识别和备份需要备份的所有数据，或从源复制数据”](#)
- [the section called “REL11-BP04 恢复期间依赖于数据平面而不是控制平面”](#)
- [the section called “REL13-BP01 定义停机和数据丢失的恢复目标”](#)

相关文档：

- [AWS 架构博客：灾难恢复系列](#)
- [AWS 上工作负载的灾难恢复：云中的恢复 \( AWS 白皮书 \)](#)
- [云中的灾难恢复选项](#)
- [在一小时内构建无服务器多区域、双活后端解决方案](#)
- [多区域无服务器后端 – 重新加载](#)
- [RDS：跨区域复制只读副本](#)
- [Route 53：配置 DNS 故障转移](#)
- [S3：跨区域复制](#)
- [什么是 AWS Backup？](#)
- [什么是 Route 53 Application Recovery Controller？](#)
- [AWS 弹性灾难恢复](#)
- [HashiCorp Terraform：入门 – AWS](#)
- [AWS 合作伙伴：可以帮助进行灾难恢复的合作伙伴](#)
- [AWS Marketplace：可以用于灾难恢复的产品](#)

相关视频：

- [AWS 上工作负载的灾难恢复](#)
- [AWS re:Invent 2018：适用于多区域主动-主动应用程序的架构模式 \( ARC209-R2 \)](#)
- [开始使用 AWS 弹性灾难恢复 | Amazon Web Services](#)

相关示例：

- [Well-Architected 实验室 - 灾难恢复](#) - 说明 DR 策略的系列研讨会

REL13-BP03 测试灾难恢复实施以验证实施效果

定期测试到恢复站点的失效转移，以验证是否在正常运行，并满足 RTO 和 RPO。

常见反模式：

- 切勿在生产环境中练习失效转移。

建立此最佳实践的好处：定期测试您的灾难恢复计划，确认该计划在需要时能够正常发挥作用，并且您的团队知道如何执行该策略。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

要避免的模式是制定了恢复路径但很少测试。例如，您可能有一个用于只读查询的辅助数据存储。当您写入某个数据存储，却发现主存储故障时，您可能希望将故障转移到辅助数据存储。如果您不经常测试此故障转移，可能会发现您关于辅助数据存储容量的假设是错误的。辅助数据存储容量在您上次测试时可能是足够的，但可能无法再容纳这次情况下的负载。我们的经验表明，唯一有效的错误恢复是您经常测试的路径。因此，最好只开发几条恢复路径。您可以建立恢复模式并定期对其进行测试。如果恢复路径比较复杂或至关重要，您仍需定期在生产环境中测试该故障，确保恢复路径有效。在我们刚才讨论的示例中，您应该定期将故障转移到备用存储，无论是否需要。

## 实施步骤

1. 为灾难恢复设计工作负载。定期测试恢复路径。面向恢复的计算可识别系统中能够增强恢复功能的特性：隔离和冗余，系统范围回滚更改的能力，监控并确定运行状况的能力，提供诊断、自动恢复、模块化设计的能力，以及重启的能力。练习恢复路径，以确认您可以在指定时间内恢复到指定状态。在此恢复过程中使用运行手册来记录问题，并在下一次测试之前找到问题的解决方案。
2. 对于基于 Amazon EC2 的工作负载，使用 [AWS Elastic Disaster Recovery](#) 为 DR 策略实施和启动演练实例。AWS Elastic Disaster Recovery 可以高效地运行演练，从而帮助您为失效转移事件做好准备。您还可以使用 Elastic Disaster Recovery 频繁地启动实例以用于测试和演练目的，无需重定向流量。

## 资源

### 相关文档：

- [AWS 合作伙伴：可以帮助进行灾难恢复的合作伙伴](#)
- [AWS 架构博客：灾难恢复系列](#)
- [AWS Marketplace：可以用于灾难恢复的产品](#)
- [AWS Elastic Disaster Recovery](#)
- [AWS 上工作负载的灾难恢复：云中的恢复 \(AWS 白皮书\)](#)
- [AWS Elastic Disaster Recovery 为失效转移做准备](#)
- [加州大学伯克利分校/斯坦福大学的面向恢复的计算项目](#)

- [什么是 AWS Fault Injection Simulator ?](#)

相关视频：

- [AWS re:Invent 2018：适用于多区域主动-主动应用程序的架构模式](#)
- [AWS re:Invent 2019：AWS 的备份与还原，以及灾难恢复解决方案](#)

相关示例：

- [Well-Architected 实验室 – 测试弹性](#)

## REL13-BP04 管理 DR 站点或区域的配置偏差

确保 DR 站点或区域的基础设施、数据和配置满足需求。例如，检查 AMI 和服务限额是否为最新。

AWS Config 会持续监控和记录 AWS 资源配置。它可以检测到偏差并触发 [AWS Systems Manager Automation](#) 进行修复和发出警报。AWS CloudFormation 还可以在您已部署的堆栈中检测到偏差。

常见反模式：

- 在主位置进行配置或基础设施更改时，未能在恢复位置进行更新。
- 不考虑主位置和恢复位置的潜在限制（如服务区别）。

建立此最佳实践的好处：确保您的 DR 环境与现有环境一致，可保证完整恢复。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

- 确保您的交付管道可交付到主站点和备份站点。用于将应用程序部署到生产中的交付管道必须分布到所有指定的灾难恢复策略位置，包括开发和测试环境。
- 启用 AWS Config 来跟踪潜在偏差位置。使用 AWS Config 规则来创建可强制实施灾难恢复策略并在检测到偏差时生成提醒的系统。
  - [按照 AWS Config 规则 修正不合规 AWS 资源](#)
  - [AWS Systems Manager Automation](#)
- 使用 AWS CloudFormation 部署基础设施。AWS CloudFormation 可以检测 CloudFormation 模板指定的内容和实际部署内容之间的偏差。

- [AWS CloudFormation：在整个 CloudFormation 堆栈上检测偏差](#)

## 资源

### 相关文档：

- [AWS 合作伙伴：可以帮助进行灾难恢复的合作伙伴](#)
- [AWS 架构博客：灾难恢复系列](#)
- [AWS CloudFormation：在整个 CloudFormation 堆栈上检测偏差](#)
- [AWS Marketplace：可以用于灾难恢复的产品](#)
- [AWS Systems Manager Automation](#)
- [AWS 上工作负载的灾难恢复：云中的恢复 \( AWS 白皮书 \)](#)
- [如何在 AWS 上实施基础设施配置管理解决方案？](#)
- [按照 AWS Config 规则 修正不合规 AWS 资源](#)

### 相关视频：

- [AWS re:Invent 2018：适用于多区域主动-主动应用程序的架构模式 \( ARC209-R2 \)](#)

## REL13-BP05 自动执行恢复

利用 AWS 或第三方工具自动进行系统恢复，并将流量路由至 DR 站点或区域。

根据已配置的运行状况检查，Elastic Load Balancing 和 AWS Auto Scaling 等 AWS 服务可将负载分配到运行正常的可用区，而 Amazon Route 53 和 AWS Global Accelerator 等服务则可将负载路由到运行正常的 AWS 区域。Amazon Route 53 Application Recovery Controller 可帮助您使用就绪检查和路由控制功能来管理和协调失效转移操作。这些功能持续监控您的应用程序从故障中恢复的能力，因此您可以跨多个 AWS 区域、可用区和本地部署控制您的应用程序恢复。

对于现有的物理或虚拟数据中心或私有云上的工作负载，[AWS 弹性灾难恢复](#)（通过 AWS Marketplace 提供）使组织能够设置自动向 AWS 进行灾难恢复的策略。CloudEndure 还支持 AWS 中的跨区域/跨可用区灾难恢复。

### 常见反模式：

- 实施相同的自动故障转移和故障恢复可能会导致在故障时发生摆动。

建立此最佳实践的好处：自动恢复通过消除发生手动错误的可能性来缩短恢复时间。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

- 恢复路径自动化。如果恢复时间很短，人工判断和操作无法用于可用性非常高的场景。在这种情况下，系统每次必须自动进行恢复。
- 使用 CloudEndure Disaster Recovery 自动执行失效转移和故障恢复操作。CloudEndure Disaster Recovery 可持续将您的计算机（包括操作系统、系统状态配置、数据库、应用程序和文件）复制到目标 AWS 账户和首选区域中的低成本暂存区域。在发生灾难时，您可以指示 CloudEndure Disaster Recovery 在几分钟内自动启动数千台处于完全预置状态的计算机。
  - [执行灾难恢复故障转移和故障恢复](#)
  - [CloudEndure Disaster Recovery](#)

### 资源

相关文档：

- [AWS 合作伙伴：可以帮助进行灾难恢复的合作伙伴](#)
- [AWS 架构博客：灾难恢复系列](#)
- [AWS Marketplace：可以用于灾难恢复的产品](#)
- [AWS Systems Manager Automation](#)
- [AWS 的 CloudEndure Disaster Recovery](#)
- [AWS 上工作负载的灾难恢复：云中的恢复（AWS 白皮书）](#)

相关视频：

- [AWS re:Invent 2018：适用于多区域主动-主动应用程序的架构模式（ARC209-R2）](#)

## 性能效率

性能效率要素包括有效地使用计算资源以满足系统要求的能力以及在需求变化和技术改进时保持此效率的能力。如需有关具体实施的说明性指导，请参阅 [《性能效率支柱》白皮书](#)。

### 最佳实践领域

- [架构选择](#)
- [计算和硬件](#)
- [数据管理](#)
- [联网和内容分发](#)
- [流程和文化](#)

## 架构选择

### 问题

- [PERF 1.如何为工作负载选择合适的云资源和架构？](#)

### PERF 1.如何为工作负载选择合适的云资源和架构？

针对特定工作负载的最佳解决方案各不相同，而且解决方案通常会结合多种方法。架构完善的工作负载会使用多种解决方案，并且允许使用各种不同的功能来提高性能。

### 最佳实践

- [PERF01-BP01 了解并掌握可用的云服务和功能](#)
- [PERF01-BP02 使用云提供商或合适的合作伙伴提供的指导来了解架构模式和最佳实践](#)
- [PERF01-BP03 制定架构决策时考虑成本因素](#)
- [PERF01-BP04 评估权衡机制对客户和架构效率的影响](#)
- [PERF01-BP05 使用策略和参考架构](#)
- [PERF01-BP06 使用基准测试来推动制定架构决策](#)
- [PERF01-BP07 使用数据驱动的方法进行架构选择](#)

### PERF01-BP01 了解并掌握可用的云服务和功能

不断了解和发现可用的服务和配置，这些服务和配置有助于您做出更好的架构决策，并提高工作负载架构的性能效率。

### 常见反模式：

- 您将云端用作联合数据中心。
- 迁移到云端后，您不对应用程序进行现代化改造。
- 您仅使用一种存储类型来存储所有需要继续保留的内容。



- 您使用的实例类型最接近您当前的标准，但有时候需要使用更大的实例。
- 您部署和管理作为托管服务提供的技术。

建立此最佳实践的好处：通过考虑采用新的服务和配置，可以大大提高性能、降低成本并减少维护工作负载所需的工作量。还有助于您缩短支持云的产品价值实现时间。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

AWS 不断发布新的服务和功能，可提高性能并降低云工作负载的成本。及时了解这些新服务和功能对于保持云端的性能效率至关重要。对工作负载架构进行现代化改造还有助于提高工作效率、推动创新并解锁更多增长机会。

## 实施步骤

- 盘点相关服务的工作负载软件和架构。决定要进一步了解哪一类产品。
- 探索 AWS 产品/服务，查看并了解有助于您提高性能、降低成本和运维复杂性的相关服务和配置选项。
  - [AWS 新增功能](#)
  - [AWS 博客](#)
  - [AWS Skill Builder](#)
  - [AWS 活动和网络研讨会](#)
  - [AWS 培训与认证](#)
  - [AWS YouTube 频道](#)
  - [AWS 研讨会](#)
  - [AWS 社区](#)
- 使用沙盒（非生产）环境来学习和试验新服务，且不会产生额外费用。

## 资源

相关文档：

- [AWS Architecture Center](#)
- [AWS Partner Network](#)
- [AWS 解决方案库](#)

- [AWS Knowledge Center](#)
- [在 AWS 上构建现代化应用程序](#)

相关视频：

- [This is my Architecture](#)

相关示例：

- [AWS 示例](#)
- [AWS 开发工具包示例](#)

PERF01-BP02 使用云提供商或合适的合作伙伴提供的指导来了解架构模式和最佳实践

利用云服务公司提供的资源（如文档、解决方案架构师、专业服务或合适的合作伙伴）来指导您制定架构决策。这些资源有助于您审查和改进架构，以实现最佳性能。

常见反模式：

- 您将 AWS 视为普通的云提供商。
- 您没有按 AWS 服务的既定用途使用这些服务。
- 您在遵循所有指导时没有考虑到业务环境。

建立此最佳实践的好处：使用云提供商或合适的合作伙伴提供的指导，有助于您为工作负载选择合适的架构，并让您对自己的决策充满信心。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

AWS 提供广泛的指导、文档和资源，有利于您构建和管理高效的云工作负载。AWS 文档提供了代码示例、教程和详细的服务说明。除文档外，AWS 还提供培训和认证计划、解决方案架构师和专业服务，可协助客户探索云服务的不同方面，并在 AWS 上实施高效的云架构。

利用这些资源深入了解宝贵的知识和最佳实践，节省时间，并在 AWS Cloud 中取得更好的成果。

## 实施步骤

- 查看 AWS 文档和指南并遵循最佳实践。这些资源有助于您有效地选择和配置服务，并实现更好的性能。
  - [AWS 文档](#)（例如用户指南和白皮书）
  - [AWS 博客](#)
  - [AWS 培训与认证](#)
  - [AWS YouTube 频道](#)
- 参加 AWS 合作伙伴活动（如 AWS 全球峰会、AWS re:Invent、用户群组 and 研讨会），向 AWS 专家学习关于使用 AWS 服务的最佳实践。
  - [AWS 活动和网络研讨会](#)
  - [AWS 研讨会](#)
  - [AWS 社区](#)
- 如需其他指导或产品信息，请联系 AWS 获取帮助。AWS 解决方案架构师和 [AWS 专业服务](#) 提供解决方案实施指导。[AWS 合作伙伴](#) 提供 AWS 专业知识，可帮助您实现业务敏捷性和创新能力。
- 如果您需要技术支持才能高效地使用服务，请使用 [支持](#)。[我们的支持计划](#) 旨在为您提供理想的工具组合以及获取专业知识的渠道，让您可以在优化性能、管理风险和控制成本的同时，使用 AWS 取得成功。

## 资源

### 相关文档：

- [AWS Architecture Center](#)
- [AWS 解决方案库](#)
- [AWS Knowledge Center](#)
- [AWS 企业支持](#)

### 相关视频：

- [This is my Architecture](#)

### 相关示例：

- [AWS 示例](#)

- [AWS 开发工具包示例](#)

## PERF01-BP03 制定架构决策时考虑成本因素

制定架构决策时考虑成本因素，以便提高云工作负载的资源利用率和性能效率。当您意识到云工作负载的成本影响时，您就更有可能充分利用有效资源，减少浪费。

常见反模式：

- 您只使用一个系列的实例。
- 您没有对照开源解决方案对许可的解决方案进行评估。
- 您没有定义存储生命周期策略。
- 您没有查看 AWS Cloud 的新服务和功能。
- 您只使用数据块存储。

建立此最佳实践的好处：通过在制定决策时考虑成本因素，可以让您使用更有效的资源，并探索其他投资方式。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

针对成本优化工作负载能够提高资源利用率，避免在云工作负载中出现浪费。要在制定架构决策时考虑成本因素，通常包括合理调整工作负载组件的大小和实现弹性，从而提高云工作负载的性能效率。

### 实施步骤

- 制定成本目标，如云工作负载的预算限额。
- 确定会增加工作负载成本的关键组件（如实例和存储）。您可以使用 [AWS 定价计算器](#) 和 [AWS Cost Explorer](#) 来确定工作负载中的关键成本驱动因素。
- 使用 [Well-Architected 成本优化最佳实践](#) 来优化这些关键组件的成本。
- 持续监控和分析成本，挖掘工作负载中的成本优化机会。
  - 使用 [AWS Budgets](#)，针对无法接受的成本获取相关提醒。
  - 使用 [AWS Compute Optimizer](#) 或 [AWS Trusted Advisor](#) 获取成本优化建议。

- 使用 [AWS Cost Anomaly Detection](#) 自动进行成本异常检测和根本原因分析。

## 资源

### 相关文档：

- [Cost Intelligence Dashboard 的详细概述](#)
- [AWS Architecture Center](#)
- [AWS Partner Network](#)
- [AWS 解决方案库](#)
- [AWS Knowledge Center](#)

### 相关视频：

- [This is my Architecture](#)
- [Optimize performance and cost for your AWS compute](#)

### 相关示例：

- [AWS 示例](#)
- [AWS 开发工具包示例](#)
- [在启用 Compute Optimizer 和内存利用率的情况下合理调整大小](#)
- [AWS Compute Optimizer 演示代码](#)

## PERF01-BP04 评估权衡机制对客户和架构效率的影响

在评估与性能相关的改进时，确定哪些选择会对客户和工作负载效率产生影响。例如，如果使用键值数据存储可以提高系统性能，那么评估这种更改的最终一致性对客户的影响就非常重要。

### 常见反模式：

- 您认为即便需要实施一些权衡机制，也要实现所有性能收益。
- 在性能问题已经非常严重时，您只评估对工作负载的更改。

建立此最佳实践的好处：当您评估潜在的与性能相关的改进时，必须决定更改时所采用的权衡机制是否符合工作负载要求。在某些情况下，您可能需要实施额外的控制来补偿权衡机制。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

根据性能和客户影响确定架构中的关键领域。确定如何提高性能、性能提高带来的利弊，并确定性能提高对系统和用户体验的影响。例如，缓存数据有助于大幅提高性能，但需要就如何以及何时更新缓存的数据或使其变得无效而制定明确的策略，以防止产生不正确的系统行为。

## 实施步骤

- 了解工作负载要求和 SLA。
- 明确定义评估因素。这些因素可能与工作负载的成本、可靠性、安全性和性能有关。
- 选择可以满足您要求的架构和服务。
- 开展试验工作并执行概念验证（POC），以评估权衡因素以及对客户和架构效率的影响。高度可用、高性能和安全的工作负载往往会消耗更多的云资源，但同时也会提供更好的客户体验。

## 资源

相关文档：

- [Amazon Builders' Library](#)
- [QuickSight KPI](#)
- [Amazon CloudWatch RUM](#)
- [X-Ray 文档](#)
- [Understand resiliency patterns and trade-offs to architect efficiently in the cloud](#)

相关视频：

- [制定监控计划](#)
- [Optimize applications through Amazon CloudWatch RUM](#)
- [Amazon CloudWatch Synthetics 演示](#)

相关示例：

- [使用 Amazon CloudWatch Synthetics 测量页面加载时间](#)
- [Amazon CloudWatch RUM Web 客户端](#)

## PERF01-BP05 使用策略和参考架构

在选择服务和配置时使用内部策略和现有参考架构，以提高设计和实施工作负载的效率。

常见反模式：

- 您允许使用各种各样的技术，而这些技术可能会影响公司的管理开销。

建立此最佳实践的好处：制定架构、技术和供应商选择策略将有助于快速做出决策。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

在选择资源和架构时需要制定内部策略，这样可以提供在进行架构方面的选择时应遵循的标准和指导方针。这些指导方针可以简化选择合适的云服务时的决策过程，并提高性能效率。使用策略或参考架构部署工作负载。将服务集成到云部署中，然后使用性能测试来验证是否能继续满足性能要求。

### 实施步骤

- 清楚了解云工作负载的要求。
- 审查内部和外部策略，找出最有效的策略。
- 使用 AWS 提供的合适参考架构或行业最佳实践。
- 创建一个由策略、标准、参考架构和针对常见情况的规范性指南组成的连续体。这样做可以让团队更快地开展工作。请酌情为您的垂直行业量身定制资产。
- 在沙盒环境中，为您的工作负载验证这些策略和参考架构。
- 随时了解行业标准和 AWS 更新，确保您的策略和参考架构有助于优化云工作负载。

### 资源

相关文档：

- [AWS Architecture Center](#)
- [AWS Partner Network](#)
- [AWS 解决方案库](#)
- [AWS Knowledge Center](#)

相关视频：



- [This is my Architecture](#)

相关示例：

- [AWS 示例](#)
- [AWS 开发工具包示例](#)

## PERF01-BP06 使用基准测试来推动制定架构决策

对现有工作负载的性能进行基准测试，从而了解工作负载在云中的表现情况，并根据这些数据推动制定架构决策。

常见反模式：

- 您启用普通的基准测试，而这些基准测试并不能反映出工作负载的特征。
- 您将客户反馈和看法作为唯一的基准。

建立此最佳实践的好处：对您的当前实施进行基准测试，以便衡量性能改进情况。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

结合使用基准测试与综合测试，评估工作负载组件的性能。相比负载测试，基准测试通常可以更快速地设置，适用于评估特定组件的技术。基准测试通常在新项目开始时进行，因为此时您还没有用于进行负载测试的完整解决方案。

您可以构建自己的自定义基准测试，也可以使用行业标准测试（如 [TPC-DS](#)），来对您的工作负载进行基准测试。行业基准测试适用于比较不同的环境。对于架构中的特定操作类型，自定义基准测试十分有用。

进行基准测试时，为了确保获得有效的结果，预热您的测试环境尤为重要。多次运行同一基准测试，确保捕获一段时间内的所有差异。

由于基准测试运行速度通常比负载测试快，它们可以在部署管道的早期使用，并能更快地提供有关性能偏差的反馈。当您评估一个组件或服务的重要更改时，您可以使用基准测试快速了解您是否有合理的理由来执行更改。结合使用基准测试与负载测试这一点很重要，因为负载测试会告诉您工作负载在生产环境中的表现如何。

## 实施步骤

- 定义指标（如 CPU 利用率、延迟或吞吐量）来评估工作负载性能。
- 确定并设置适合您工作负载的基准测试工具。您可以使用 AWS 服务（如 [Amazon CloudWatch](#)）或与您的工作负载兼容的第三方工具。
- 执行基准测试并在测试期间监控指标。
- 分析并记录基准测试结果，找出任何瓶颈和问题。
- 利用测试结果制定架构决策并调整工作负载。这可能包括更改服务或采用新功能。
- 调整后重新测试工作负载。

## 资源

### 相关文档：

- [AWS Architecture Center](#)
- [AWS Partner Network](#)
- [AWS 解决方案库](#)
- [AWS Knowledge Center](#)
- [Amazon CloudWatch RUM](#)
- [Amazon CloudWatch Synthetics](#)

### 相关视频：

- [This is my Architecture](#)
- [Optimize applications through Amazon CloudWatch RUM](#)
- [Amazon CloudWatch Synthetics 演示](#)

### 相关示例：

- [AWS 示例](#)
- [AWS 开发工具包示例](#)
- [分布式负载测试](#)
- [使用 Amazon CloudWatch Synthetics 测量页面加载时间](#)

- [Amazon CloudWatch RUM Web 客户端](#)

## PERF01-BP07 使用数据驱动的方法进行架构选择

为架构选择确定清晰的数据驱动方法，确保使用合适的云服务和配置来满足您的特定业务需求。

常见反模式：

- 您认为当前的架构是静态的，将来不会更新。
- 您选择架构时是基于猜测和假设。
- 您不断对架构进行更改，而不提供理由。

建立此最佳实践的好处：通过使用明确定义的方法来选择架构，您可以通过分析数据来优化工作负载设计，从而在未来做出明智的决策。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

利用内部经验和云知识或外部资源（如已发布的使用场景或白皮书）来选择架构中的资源和服务。应该有一个明确定义的流程，该流程支持对可能会用于工作负载的不同服务进行试验和基准测试。

关键工作负载的待办事项不仅应包括用户案例（提供与业务和用户相关的功能），还应包括技术案例（创建工作负载的架构跑道）。该跑道依托于技术和新服务领域新的改进，并根据数据和适当的理由应用这些改进。这可以确保架构经得起未来考验，不会停滞不前。

### 实施步骤

- 与主要利益相关者一起确定工作负载要求，包括性能、可用性和成本方面的考量。考虑诸如用户数量和工作负载使用模式之类的因素。
- 创建架构跑道或技术待办事项，统筹确定它们与功能型待办事项的优先级。
- 评估和评测不同的云服务（有关详细信息，请参阅 [PERF01-BP01 了解并掌握可用的云服务和功能](#)）。
- 探索满足性能要求的不同架构模式，如微服务或无服务器（有关详细信息，请参阅 [PERF01-BP02 使用云提供商或合适的合作伙伴提供的指导来了解架构模式和最佳实践](#)）。
- 咨询其他团队，查阅架构图和资源，例如 AWS Solution Architect、[AWS Architecture Center](#)和 [AWS Partner Network](#)，从而为工作负载选择合适的架构。

- 定义吞吐量和响应时间等性能指标，以便于您评估工作负载的性能。
- 进行试验并使用定义的指标来验证所选架构的性能。
- 持续监控并根据需要进行调整，从而使架构保持最佳性能。
- 记录所选架构和决策，作为将来更新和学习的参考。
- 根据经验、新技术以及可表明当前方法需要进行更改或存在问题的指标，不断审查和更新架构选择方法。

## 资源

### 相关文档：

- [AWS 解决方案库](#)
- [AWS Knowledge Center](#)

### 相关视频：

- [This is my Architecture](#)

### 相关示例：

- [AWS 示例](#)
- [AWS 开发工具包示例](#)

## 计算和硬件

### PERF 2.如何在工作负载中选择和使用计算资源？

适合特定工作负载的最佳计算方案会因应用程序设计、使用模式和配置设置而有所不同。架构可能会使用不同的计算方案来支持各种组件，并允许使用不同的功能来提高性能。为架构选择错误的计算方案可能会降低性能效率。

### 最佳实践

- [PERF02-BP01 为工作负载选择最佳计算方案](#)
- [PERF02-BP02 了解可用的计算配置和功能](#)
- [PERF02-BP03 收集与计算相关的指标](#)

- [PERF02-BP04 配置计算资源并合理调整资源规模](#)
- [PERF02-BP05 动态扩展计算资源](#)
- [PERF02-BP06 使用基于硬件的优化型计算加速器](#)

## PERF02-BP01 为工作负载选择最佳计算方案

通过为工作负载选择最合适的计算方案，可以提高性能，减少不必要的基础设施成本，减少维护工作负载所需的运维工作。

常见反模式：

- 使用本地所用的计算方案。
- 对云计算方案、功能和解决方案以及这些解决方案如何提高计算性能缺乏认识。
- 为了满足扩展或性能需求，过度预置现有计算方案，而使用替代计算方案可以更准确地满足您的工作负载特征需求。

建立此最佳实践的好处：通过确定计算需求并对可用方案进行评估，可以使工作负载更有效地利用资源。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

在为了提高性能效率而优化云工作负载时，请务必根据使用场景和性能需求选择最合适的计算方案。AWS 提供了多种计算方案，可满足云中不同工作负载的需求。例如，您可以使用 [Amazon EC2](#) 启动和管理虚拟服务器，[AWS Lambda](#) 运行代码而无需预置或管理服务器，[Amazon ECS](#) 或 [Amazon EKS](#) 运行和管理容器，或 [AWS Batch](#) 并行处理大量数据。您应根据自己的规模和计算需求，选择和配置最适合自己的情况的计算解决方案。您也可以考虑在单个工作负载中使用多种类型的计算解决方案，因为每种解决方案都有自己的优缺点。

以下步骤将指导您根据自身工作负载的特征和性能要求，选择合适的计算方案。

### 实施步骤

1. 了解工作负载计算要求。需要考虑的关键要求包括：处理需求、流量模式、数据访问模式、扩展需求和延迟要求。
2. 了解适用于 AWS 上工作负载的不同计算方案（如 [PERF01-BP01 了解并掌握可用的云服务和功能](#) 中所述）。以下介绍了一些关键的 AWS 计算方案、这些方案的特征和常见使用场景：

| AWS 服务                                                                                                                         | 主要特征                                                                                                                                                                                                    | 常见使用场景                   |
|--------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| <a href="#">Amazon Elastic Compute Cloud (Amazon EC2)</a>                                                                      | 具有硬件专用选项、许可证要求，多种不同实例系列、处理器类型和计算加速器可供选择                                                                                                                                                                 | 直接迁移、整体式应用程序、混合环境、企业应用程序 |
| <a href="#">Amazon Elastic Container Service (Amazon ECS)</a> 、 <a href="#">Amazon Elastic Kubernetes Service (Amazon EKS)</a> | 轻松部署、环境一致、可扩展                                                                                                                                                                                           | 微服务、混合环境                 |
| <a href="#">AWS Lambda</a>                                                                                                     | <a href="#">无服务器计算</a> 服务运行代码以响应事件并自动管理底层计算资源。                                                                                                                                                          | 微服务、事件驱动的应用程序            |
| <a href="#">AWS Batch</a>                                                                                                      | 高效、动态地进行预置和扩展 <a href="#">Amazon Elastic Container Service (Amazon ECS)</a> 、 <a href="#">Amazon Elastic Kubernetes Service (Amazon EKS)</a> 和 <a href="#">AWS Fargate</a> 计算资源，可根据您的任务要求选择使用按需实例或竞价型实例 | HPC，训练机器学习模型             |
| <a href="#">Amazon Lightsail</a>                                                                                               | 预先配置的 Linux 和 Windows 应用程序，用于运行小型工作负载                                                                                                                                                                   | 简单的 Web 应用程序、自定义网站       |

- 评估与每种计算方案相关的成本（如每小时费用或数据传输）和管理开销（如修补和扩展）。
- 在非生产环境中进行试验和基准测试，确定哪种计算方案最能满足您的工作负载需求。
- 试验并确定新的计算解决方案后，规划迁移并验证性能指标。
- 使用 AWS 监控工具（如 [Amazon CloudWatch](#)）和优化服务（如 [AWS Compute Optimizer](#)）根据实际使用模式持续优化您的计算资源。

## 资源

### 相关文档：

- [使用 AWS 进行云计算](#)
- [Amazon EC2 实例类型](#)
- [Amazon EKS 容器：Amazon EKS Worker 节点](#)
- [Amazon ECS 容器：Amazon ECS 容器实例](#)
- [函数：Lambda 函数配置](#)
- [容器规范性指南](#)
- [无服务器规范性指南](#)

### 相关视频：

- [如何为初创企业选择计算方案](#)
- [Optimize performance and cost for your AWS compute](#)
- [Amazon EC2 foundations](#)
- [Powering next-gen Amazon EC2: Deep dive into the Nitro system](#)
- [Deploy ML models for inference at high performance and low cost](#)
- [Better, faster, cheaper compute: Cost-optimizing Amazon EC2](#)

### 相关示例：

- [将 Web 应用程序迁移到容器](#)
- [运行无服务器 Hello World](#)

## PERF02-BP02 了解可用的计算配置和功能

了解计算服务的可用配置选项和功能，从而预置适量的资源并提高性能效率。

### 常见反模式：

- 您没有依据工作负载特征评估计算方案或可用的实例系列。
- 您过度预置计算资源来满足高峰需求。



建立此最佳实践的好处：熟悉 AWS 计算功能和配置，以便您能够使用经过优化的计算解决方案，来满足工作负载特征需求。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

每种计算解决方案都有独特的配置和功能，可支持不同的工作负载特征和需求。了解这些方案如何完善您的工作负载，并确定哪些配置选项最适合您的应用程序。这些选项的示例包括实例系列、规模、功能（GPU、I/O）、突增、超时、函数大小、容器实例和并发度。如果您的工作负载已经使用同一计算方案超过四周，并且您预计这些特征在未来将保持不变，那么您可以使用 [AWS Compute Optimizer](#) 从 CPU 和内存的角度确定当前的计算方案是否适合工作负载。

实施步骤

1. 了解工作负载要求（如 CPU 需求、内存和延迟）。
2. 查看 AWS 文档和最佳实践，了解有助于提高计算性能的推荐配置选项。以下是一些需要考虑的关键配置选项：

| 配置选项 | 示例                                                                                                                                                                                                                                                                      |
|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 实例类型 | <ul style="list-style-type: none"><li>• <a href="#">计算优化型</a> 实例非常适合需要较高 vCPU 与内存比的工作负载。</li><li>• <a href="#">内存优化型</a> 实例提供大量内存以支持内存密集型工作负载。</li><li>• <a href="#">存储优化型</a> 实例专为需要对本地存储进行大量顺序读写访问（IOPS）的工作负载而设计。</li></ul>                                           |
| 定价模式 | <ul style="list-style-type: none"><li>• <a href="#">按需型实例</a> 让您按小时或按秒使用计算容量，而无需做出长期承诺。这些实例非常适合超出性能基线需求的突增情况。</li><li>• <a href="#">Savings Plans</a> 相比按需型实例，能够节省大量成本，以换取在一年或三年内使用特定数量计算能力的承诺。</li><li>• <a href="#">竞价型实例</a> 可让您以折扣价将未使用的实例容量用于无状态的容错工作负载。</li></ul> |

| 配置选项         | 示例                                                                                                                                                                                                                                                   |
|--------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Auto Scaling | 使用 <a href="#">Auto Scaling</a> 配置以使计算资源与流量模式相匹配。                                                                                                                                                                                                    |
| 确定大小         | <ul style="list-style-type: none"> <li>使用 <a href="#">Compute Optimizer</a> 可以就哪种计算配置最符合您的计算特征，获得基于机器学习的建议。</li> <li>使用 <a href="#">AWS Lambda Power Tuning</a> 为 Lambda 函数选择最佳配置。</li> </ul>                                                        |
| 基于硬件的计算加速器   | <ul style="list-style-type: none"> <li><a href="#">加速型计算实例</a> 与基于 CPU 的替代方案相比，执行图形处理或数据模式匹配等功能的效率更高。</li> <li>对于机器学习工作负载，请利用特定于工作负载的专用硬件，例如 <a href="#">AWS Trainium</a>、<a href="#">AWS Inferentia</a> 和 <a href="#">Amazon EC2 DL1</a></li> </ul> |

## 资源

### 相关文档：

- [使用 AWS 进行云计算](#)
- [Amazon EC2 实例类型](#)
- [Amazon EC2 实例的处理器状态控制](#)
- [Amazon EKS 容器：Amazon EKS Worker 节点](#)
- [Amazon ECS 容器：Amazon ECS 容器实例](#)
- [函数：Lambda 函数配置](#)

### 相关视频：

- [Amazon EC2 foundations](#)
- [Powering next-gen Amazon EC2: Deep dive into the Nitro system](#)
- [Optimize performance and cost for your AWS compute](#)

相关示例：

- [在启用 Compute Optimizer 和内存利用率的情况下合理调整大小](#)
- [AWS Compute Optimizer 演示代码](#)

## PERF02-BP03 收集与计算相关的指标

记录和跟踪与计算相关的指标，以便更好地了解计算资源的表现情况，并提高计算资源的性能和利用率。

常见反模式：

- 您只手动搜索日志文件来查找指标。
- 您只使用由监控软件记录的默认指标。
- 您只在出现问题时审查指标。

建立此最佳实践的好处：收集与性能相关的指标有助于您根据业务需求调整应用程序性能，从而确保满足工作负载需求。收集指标还有利于您持续提高工作负载中的资源性能和利用率。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

云工作负载会生成大量数据，例如指标、日志和事件。在 AWS Cloud 中，收集指标是提高安全性、成本效率、性能和可持续性的关键步骤。AWS 使用监控服务（如 [Amazon CloudWatch](#)）提供各种与性能相关的指标，从而为您提供宝贵的洞察。CPU 利用率、内存利用率、磁盘 I/O 以及网络入站和出站等指标有助于您深入了解利用率水平或性能瓶颈。将这些指标用作数据驱动方法的一部分，以便主动调整和优化工作负载的资源。理想情况下，您应该在单一平台上收集与计算资源相关的所有指标，并实施留存策略以支持成本目标和运营目标。

## 实施步骤

1. 确定哪些与性能相关的指标与您的工作负载相关。您应该收集有关资源利用率和云工作负载运行方式的指标（例如响应时间和吞吐量）。
  - a. [Amazon EC2 默认指标](#)
  - b. [Amazon ECS 默认指标](#)
  - c. [Amazon EKS 默认指标](#)
  - d. [Lambda 默认指标](#)

- e. [Amazon EC2 内存和磁盘指标](#)
2. 为工作负载选择并设置合适的日志记录和监控解决方案。
  - a. [AWS 原生可观测性](#)
  - b. [AWS Distro for OpenTelemetry](#)
  - c. [Amazon Managed Service for Prometheus](#)
3. 根据工作负载要求为指标确定所需的筛选和聚合。
  - a. [使用 Amazon CloudWatch Logs 和指标筛选器量化自定义应用程序指标](#)
  - b. [使用 Amazon CloudWatch 策略性标记收集自定义指标](#)
4. 为指标配置数据留存策略，从而符合安全目标和运营目标。
  - a. [CloudWatch 指标的默认数据留存](#)
  - b. [CloudWatch Logs 的默认数据留存](#)
5. 如有需要，可为指标创建提醒和通知，协助您主动应对与性能相关的问题。
  - a. [使用 Amazon CloudWatch 异常检测为自定义指标创建提醒](#)
  - b. [使用 Amazon CloudWatch RUM 为特定网页创建指标和提醒](#)
6. 使用自动化技术来部署指标和日志聚合代理。
  - a. [AWS Systems Manager 自动化](#)
  - b. [OpenTelemetry Collector](#)

## 资源

### 相关文档：

- [Amazon CloudWatch 文档](#)
- [使用 CloudWatch 代理从 Amazon EC2 实例和本地服务器收集指标和日志](#)
- [访问 AWS Lambda 的 Amazon CloudWatch Logs](#)
- [结合使用 CloudWatch Logs 与容器实例](#)
- [发布自定义指标](#)
- [AWS Answers：集中式日志记录](#)
- [发布 CloudWatch 指标的 AWS 服务](#)
- [在 AWS Fargate 上监控 Amazon EKS](#)

### 相关视频：

- [Application Performance Management on AWS](#)

相关示例：

- [第 100 级：使用 CloudWatch 控制面板进行监控](#)
- [第 100 级：使用 CloudWatch 控制面板监控 Windows EC2 实例](#)
- [第 100 级：使用 CloudWatch 控制面板监控 Amazon Linux EC2 实例](#)

## PERF02-BP04 配置计算资源并合理调整资源规模

配置计算资源并合理调整资源规模，使其满足您工作负载的性能要求，避免资源利用不足或过度利用。

常见反模式：

- 您忽略工作负载性能要求，导致计算资源过度预置或预置不足。
- 您只选择适用于所有工作负载的最大或最小实例。
- 为了便于管理，您只使用一个实例系列。
- 您忽略来自 AWS Cost Explorer 或 Compute Optimizer 的关于合理调整规模的建议。
- 您没有重新评估新实例类型是否适合工作负载。
- 您只为组织认证少量实例配置。

建立此最佳实践的好处：合理调整计算资源的规模后，可避免资源过度预置和预置不足，从而确保资源在云端以最佳方式运行。适当调整计算资源的规模通常可以提高性能和改进客户体验，同时还可以降低成本。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

合理调整规模使组织能够以经济高效的方式运营云基础设施，同时满足业务需求。过度预置云资源会导致产生额外成本，而预置不足则会导致性能不佳和负面的客户体验。AWS 提供诸如 [AWS Compute Optimizer](#) 和 [AWS Trusted Advisor](#) 之类的工具，这些工具使用历史数据来给出合理调整计算资源规模的建议。

### 实施步骤

- 选择最能满足您需求的实例类型：

- [如何为我的工作负载选择合适的 Amazon EC2 实例类型？](#)
- [基于属性为 Amazon EC2 实例集选择实例类型。](#)
- [使用基于属性的实例类型选择来创建 Auto Scaling 组](#)
- [通过 Karpenter 整合优化 Kubernetes 计算成本](#)
- 分析您的工作负载的各种性能特性，以及这些特性与内存、网络 and CPU 使用率之间的关系。根据这些数据选择最符合您的工作负载情况和性能目标的资源。
- 使用 AWS 监控工具（如 Amazon CloudWatch）监控资源使用情况。
- 为计算资源选择合适的配置。
  - 对于临时工作负载，请评估 [实例 Amazon CloudWatch 指标](#)（例如 CPUUtilization），以确定实例是利用不足还是利用过度。
  - 对于稳定工作负载，请定期检查 AWS 合理调整规模工具（如 AWS Compute Optimizer 和 AWS Trusted Advisor），从而挖掘优化计算资源和合理调整计算资源规模的机会。
  - [Well-Architected 实验室 - 合理调整大小建议](#)
  - [Well-Architected 实验室 - 使用 Compute Optimizer 合理调整大小](#)
- 在实际环境中实施之前，先在非生产环境中测试配置更改。
- 持续重新评估新的计算产品/服务，并与工作负载的需求进行比较。

## 资源

### 相关文档：

- [使用 AWS 进行云计算](#)
- [Amazon EC2 实例类型](#)
- [Amazon ECS 容器：Amazon ECS 容器实例](#)
- [Amazon EKS 容器：Amazon EKS Worker 节点](#)
- [函数：Lambda 函数配置](#)
- [Amazon EC2 实例的处理器状态控制](#)

### 相关视频：

- [Amazon EC2 foundations](#)
- [Better, faster, cheaper compute: Cost-optimizing Amazon EC2](#)
- [Deploy ML models for inference at high performance and low cost](#)

- [Optimize performance and cost for your AWS compute](#)
- [Powering next-gen Amazon EC2: Deep dive into the Nitro system](#)
- [Simplifying Data Processing to Enhance Innovation with Serverless Tools](#)

相关示例：

- [在启用 Compute Optimizer 和内存利用率的情况下合理调整大小](#)
- [AWS Compute Optimizer 演示代码](#)

## PERF02-BP05 动态扩展计算资源

利用云的弹性根据需求动态增减计算资源，避免为工作负载预置的容量过多或者不足。

常见反模式：

- 通过手动增加容量来对警报做出反应。
- 使用本地所用的规模调整指南（通常是静态基础设施）。
- 在扩展事件之后，您保留增加的容量，而不是缩减容量。

建立此最佳实践的好处：配置和测试计算资源的弹性将有助于您节省资金、维护性能基准，以及在流量变化时提高可靠性。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

AWS 让您能够通过各种扩展机制灵活地动态扩展或缩减资源，以便满足不断变化的需求。动态扩展结合计算相关的指标，可使工作负载自动响应变化，并利用一系列最优的计算资源来实现目标。

您可以使用大量不同方法来实现资源的供需匹配。

- 目标跟踪方法：监控您的扩展指标，并根据需要自动增加或减少容量。
- 预测性扩缩：根据每日和每周的趋势进行扩展。
- 基于计划的方法：根据可预测的负载变化设置自己的扩展计划。
- 服务扩展：选择可根据设计自动扩展的服务（如无服务器）。

您必须确保工作负载部署可以处理扩展和缩减事件。



实施步骤

- 计算实例、容器和函数都能够与自动扩展服务相结合或作为此服务的一项功能来提供可实现弹性的机制。以下是自动扩展机制的一些示例：

| 自动扩展机制                                                  | 使用情形                                                                                                                                  |
|---------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">Amazon EC2 Auto Scaling</a>                 | 确保您拥有适量的 <a href="#">Amazon EC2</a> 实例，可处理您应用程序的用户负载。                                                                                 |
| <a href="#">Application Auto Scaling</a>                | 自动扩缩 Amazon EC2 以外的各项 AWS 服务的资源，比如 <a href="#">AWS Lambda</a> 函数或 <a href="#">Amazon Elastic Container Service ( Amazon ECS )</a> 服务。 |
| <a href="#">Kubernetes Cluster Autoscaler/Karpenter</a> | 自动扩缩 Kubernetes 集群。                                                                                                                   |

- 扩展通常与计算服务（如 Amazon EC2 实例或 AWS Lambda 函数）相关。此外，也请务必考虑非计算服务（如 [AWS Glue](#)）的配置来满足需求。
- 验证扩展指标是否与正在部署的工作负载的特征相匹配。如果您正在部署一个视频转码应用程序，CPU 利用率预计为 100%，并且不应将此作为您的主要指标。改用转码作业队列的深度。如果需要，您可以为您的扩缩策略使用一个 [自定义指标](#)。要选择正确的指标，请考虑以下关于 Amazon EC2 的指导：
  - 该指标应该是有效的利用率指标，并描述实例的繁忙程度。
  - 该指标值必须随 Auto Scaling 组中的实例数量成比例地增加或减少。
- 确保使用 [动态扩展](#) 而不是 [手动扩展](#)（对于 Auto Scaling 组）。我们还建议您在动态扩展中使用 [目标跟踪扩缩策略](#)。
- 验证工作负载部署是否能够同时处理扩展事件和缩减事件。例如，您可以使用 [活动历史记录](#) 来验证 Auto Scaling 组的扩缩活动。
- 评估您的工作负载以获得可预测的模式，并在您预期需求会发生预测和计划的变化时主动扩缩。借助预测性扩缩，您无需过度调配容量。有关更多详细信息，请参阅 [Amazon EC2 Auto Scaling 预测性扩缩](#)。

资源

相关文档：

- [使用 AWS 进行云计算](#)
- [Amazon EC2 实例类型](#)
- [Amazon ECS 容器：Amazon ECS 容器实例](#)
- [Amazon EKS 容器：Amazon EKS Worker 节点](#)
- [函数：Lambda 函数配置](#)
- [Amazon EC2 实例的处理器状态控制](#)
- [深入探讨 Amazon ECS 集群 Auto Scaling](#)
- [正式推出 Karpenter：高性能开源 Kubernetes Cluster Autoscaler](#)

相关视频：

- [Amazon EC2 foundations](#)
- [Better, faster, cheaper compute: Cost-optimizing Amazon EC2](#)
- [Optimize performance and cost for your AWS compute](#)
- [Powering next-gen Amazon EC2: Deep dive into the Nitro system](#)
- [构建成本、能源和资源高效的计算环境](#)

相关示例：

- [Amazon EC2 Auto Scaling 组示例](#)
- [使用 Karpenter 实施自动扩缩](#)

## PERF02-BP06 使用基于硬件的优化型计算加速器

与基于 CPU 的替代方案相比，使用硬件加速器可以更高效地执行某些功能。

常见反模式：

- 在您的工作负载中，您没有对照能够提供更高性能和更低成本的专用实例，对通用实例进行基准测试。
- 使用基于硬件的计算加速器执行任务，而使用基于 CPU 的替代方案能更高效地完成这些任务。
- 不监控 GPU 使用情况。

建立此最佳实践的好处：通过使用基于硬件的加速器 [如图形处理单元 ( GPU ) 和现场可编程门阵列 ( FPGA ) ]，可以更高效地执行某些处理功能。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

加速型计算实例提供对基于硬件的计算加速器 ( 如 GPU 和 FPGA ) 的访问。这些硬件加速器能够比基于 CPU 的替代方案更有效地执行某些功能，例如图形处理或数据模式匹配。许多加速工作负载 ( 如渲染、转码和机器学习 ) 在资源使用方面变化很大。仅在需要时运行此硬件，并在不需要时自动将其停用，从而提高整体性能效率。

## 实施步骤

- 确定哪些 [加速型计算实例](#) 可以满足您的要求。
- 对于机器学习工作负载，请利用特定于工作负载的专用硬件，例如 [AWS Trainium](#)、[AWS Inferentia](#) 和 [Amazon EC2 DL1](#)。Inf2 等 AWS Inferentia 实例 [与同类 Amazon EC2 实例相比，性能功耗比提升了多达 50%](#)。
- 收集加速型计算实例的使用情况指标。例如，您可以使用 CloudWatch 代理，为 GPU 收集各种指标，例如 utilization\_gpu 和 utilization\_memory，如 [使用 Amazon CloudWatch 收集 NVIDIA GPU 指标](#) 中所示。
- 优化硬件加速器的代码、网络运营和设置，确保底层硬件得到充分利用。
  - [优化 GPU 设置](#)
  - [深度学习 AMI 中的 GPU 监控和优化](#)
  - [优化 I/O 以实现 Amazon SageMaker AI 中深度学习训练的 GPU 性能优化](#)
- 使用最新的高性能库和 GPU 驱动程序。
- 使用自动化功能在不使用 GPU 实例时将其释放。

## 资源

### 相关文档：

- [GPU 实例](#)
- [使用 AWS Trainium 的实例](#)
- [使用 AWS Inferentia 的实例](#)
- [让我们来构建！使用自定义芯片和加速器来构建](#)

- [加速计算型](#)
- [Amazon EC2 VT1 实例](#)
- [如何为我的工作负载选择合适的 Amazon EC2 实例类型？](#)
- [选择最佳 AI 加速器和模型编译，以使用 Amazon SageMaker AI 进行计算机视觉推理](#)

相关视频：

- [如何选择 Amazon EC2 GPU 实例进行深度学习](#)
- [部署经济高效的深度学习推理](#)

## 数据管理

### PERF 3.如何存储、管理和访问工作负载中的数据？

针对特定系统的最佳数据管理解决方案往往取决于数据类型（数据块、文件或对象）、访问模式（随机或连续）、所需吞吐量、访问频率（在线、离线、归档）、更新频率（WORM、动态）以及可用性与持久性限制等因素。架构完善的工作负载使用专门构建的数据存储，这些存储允许使用不同的功能来提高性能。

#### 最佳实践

- [PERF03-BP01 使用最能满足数据访问和存储要求的专用数据存储](#)
- [PERF03-BP02 评估数据存储的可用配置选项](#)
- [PERF03-BP03 收集和记录数据存储性能指标](#)
- [PERF03-BP04 实施可提高数据存储查询性能的策略](#)
- [PERF03-BP05 实现利用缓存的数据访问模式](#)

#### PERF03-BP01 使用最能满足数据访问和存储要求的专用数据存储

了解数据特征（如数据的可共享性、大小、缓存大小、访问模式、延迟、吞吐量和持久性），为工作负载选择合适的专用数据存储（存储或数据库）。

#### 常见反模式：

- 由于内部对某种特定类型的数据库解决方案具备相关经验且比较了解，因此坚持使用一种数据存储。
- 您认为所有工作负载都有类似的数据存储和访问要求。
- 您没有实施数据目录来清点您的数据资产。

建立此最佳实践的好处： 通过了解数据特征和要求，您能够确定最高效、性能最高的存储技术，来满足工作负载需求。

未建立这种最佳实践的情况下暴露的风险等级： 高

实施指导

选择和实施数据存储时，要确保查询、扩展和存储特征支持工作负载数据要求。AWS 提供多种数据存储和数据库技术，包括数据块存储、对象存储、流式存储、文件系统、关系数据库、键值数据库、文档数据库、内存数据库、图形数据库、时间序列数据库和分类账数据库等。每种数据管理解决方案都有可供您使用的选项和配置，可支持您的使用场景和数据模型。通过了解数据特征和要求，您可以摆脱单一存储技术以及有很多局限性的一刀切方法，专注于合理管理数据。

实施步骤

- 清点工作负载中存在的各种数据类型。
- 了解并记录数据特征和要求，包括：
  - 数据类型（非结构化、半结构化、关系型）
  - 数据量和增长
  - 数据持久性：持久、短暂、瞬时
  - ACID（原子性、一致性、隔离性、持久性）要求
  - 数据访问模式（读取密集型或写入密集型）
  - 延迟
  - 吞吐量
  - IOPS（每秒输入/输出操作数）
  - 数据留存期
- 了解可用于 AWS 工作负载的不同数据存储，这些存储可以满足您的数据特征要求（如 [PERF01-BP01 了解并掌握可用的云服务和功能](#)中所述）。AWS 存储技术及其关键特征的一些示例包括：

| 类型   | AWS 服务                    | 主要特征                                                                                                              |
|------|---------------------------|-------------------------------------------------------------------------------------------------------------------|
| 对象存储 | <a href="#">Amazon S3</a> | 无限的可扩展性、高可用性以及多种可访问性选项。在 Amazon S3 内外传输和访问对象可以使用诸如 <a href="#">Transfer Acceleration</a> 或 <a href="#">Access</a> |

| 类型     | AWS 服务                                                                                                     | 主要特征                                                                                                                                                      |
|--------|------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|
|        |                                                                                                            | <a href="#">Points</a> 之类的服务，来满足您的位置、安全需求和访问模式。                                                                                                           |
| 存档存储   | <a href="#">Amazon S3 Glacier</a>                                                                          | 专为数据存档而打造。                                                                                                                                                |
| 流媒体存储  | <a href="#">Amazon Kinesis</a><br><a href="#">Amazon Managed Streaming for Apache Kafka ( Amazon MSK )</a> | 高效摄取和存储流媒体数据。                                                                                                                                             |
| 共享文件系统 | <a href="#">Amazon Elastic File System ( Amazon EFS )</a>                                                  | 可装载文件系统，可由多种类型的计算解决方案访问。                                                                                                                                  |
| 共享文件系统 | <a href="#">Amazon FSx</a>                                                                                 | 基于最新 AWS 计算解决方案而构建，支持四种常用文件系统：NetApp ONTAP、OpenZFS、Windows 文件服务器和 Lustre。Amazon FSx <a href="#">延迟、吞吐量和 IOPS</a> 因文件系统而不同，因此，在为您的工作负载需求选择合适的文件系统时应考虑这些因素。 |
| 数据块存储  | <a href="#">Amazon Elastic Block Store ( Amazon EBS )</a>                                                  | 设计用于 Amazon Elastic Compute Cloud ( Amazon EC2 ) 的可扩展、高性能的块存储服务。Amazon EBS 包括用于事务性、IOPS 密集型工作负载的基于 SSD 的存储，以及用于吞吐量密集型工作负载的基于 HDD 的存储。                     |

| 类型    | AWS 服务                                                                                         | 主要特征                                                                                                                                                                            |
|-------|------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 关系数据库 | <a href="#">Amazon Aurora</a> 、 <a href="#">Amazon RDS</a> 、 <a href="#">Amazon Redshift</a> 。 | 旨在支持 ACID ( 原子性、一致性、隔离性、持久性 ) 事务，并保持参照完整性和数据强一致性。许多传统应用程序、企业资源规划 ( ERP , Enterprise Resource Planning ) 、客户关系管理 ( CRM , Customer Relationship Management ) 和电子商务都使用关系数据库来存储其数据。 |
| 键值数据库 | <a href="#">Amazon DynamoDB</a>                                                                | 已针对常见的访问模式进行优化，通常用于存储和检索大量数据。键值数据库的典型使用案例包括高流量 Web 应用程序，电子商务系统和游戏应用程序。                                                                                                          |
| 文档数据库 | <a href="#">Amazon DocumentDB</a>                                                              | 旨在将半结构化数据存储为类似 JSON 的文档。这些数据库可帮助开发人员快速构建和更新应用程序，例如内容管理、目录和用户配置文件。                                                                                                               |
| 内存数据库 | <a href="#">Amazon ElastiCache</a> 、<br><a href="#">适用于 Redis 的 Amazon MemoryDB</a>            | 用于需要实时访问数据、最低延迟和最高吞吐量的应用程序。您可以将内存数据库用于应用程序缓存、会话管理、游戏排行榜、低延迟机器学习功能存储、微服务消息传送系统和高吞吐量流式传输机制。                                                                                       |



| 类型      | AWS 服务                                                       | 主要特征                                                                                 |
|---------|--------------------------------------------------------------|--------------------------------------------------------------------------------------|
| 图形数据库   | <a href="#">Amazon Neptune</a>                               | 用于需要大规模以毫秒延迟在高度连接的图形数据集之间浏览和查询数百万关系的应用程序。许多公司将图形数据库用于欺诈检测、社交网络和推荐引擎。                 |
| 时间序列数据库 | <a href="#">Amazon Timestream</a>                            | 用于高效收集、合成数据，并从不断变化的数据中获得见解。IoT 应用程序、开发运营和工业遥测可以利用时间序列数据库。                            |
| 宽列      | <a href="#">Amazon Keyspaces ( Apache Cassandra 兼容 )</a>     | 使用表、行和列，但是与关系数据库不同的是，同一个表中各行的列名称和格式可能会有所不同。您通常会看到一个宽列存储在大规模工业应用程序中，用于设备维护、队列管理和路线优化。 |
| 分类账     | <a href="#">Amazon Quantum Ledger Database (Amazon QLDB)</a> | 提供可信中央机构，以维护每个应用程序的可扩展、不可变和允许以加密方式进行验证的交易记录。我们看到分类账数据库用于记录系统、供应链、注册甚至银行交易。           |

- 如果您在构建数据平台，请利用 [现代数据架构](#)（基于 AWS），来集成数据湖、数据仓库和专用数据存储。
- 为工作负载选择数据存储时需要考虑的关键问题如下：

| 问题                            | 需要考虑的事项                                                                                                                                                                                                                                                                             |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 数据结构如何？                       | <ul style="list-style-type: none"> <li>如果数据是非结构化的，请考虑使用对象存储，例如 <a href="#">Amazon S3</a> 或 NoSQL 数据库，如 <a href="#">Amazon DocumentDB</a></li> <li>对于键值数据，请考虑使用 <a href="#">DynamoDB</a>、<a href="#">Amazon ElastiCache (Redis OSS)</a> 或 <a href="#">Amazon MemoryDB</a></li> </ul> |
| 需要什么级别的参照完整性？                 | <ul style="list-style-type: none"> <li>对于外键约束，关系数据库（如 <a href="#">Amazon RDS</a> 和 <a href="#">Aurora</a>）可以提供这种级别的完整性。</li> <li>通常，在 NoSQL 数据模型中，您可以将数据去规范化到单个文档或文档集合，以便在单个请求中进行检索，而不是跨各文档或各表联接。</li> </ul>                                                                        |
| 是否要求符合 ACID（原子性、一致性、隔离性、持久性）？ | <ul style="list-style-type: none"> <li>如果需要与关系数据库关联的 ACID 属性，请考虑使用关系数据库，例如 <a href="#">Amazon RDS</a> 和 <a href="#">Aurora</a>。</li> <li>如果 <a href="#">NoSQL 数据库</a> 需要强一致性，则可以使用 <a href="#">DynamoDB</a> 的强一致性读取。</li> </ul>                                                     |
| 存储需求将如何随时间变化？这对可扩展性有何影响？      | <ul style="list-style-type: none"> <li>无服务器数据库（如 <a href="#">DynamoDB</a> 和 <a href="#">Amazon Quantum Ledger Database (Amazon QLDB)</a>）将动态扩展。</li> <li>关系数据库的预置存储空间设有上限，一旦达到这些限制，往往必须使用分片等机制进行水平分区。</li> </ul>                                                                    |
| 读查询与写查询的比例是多少？缓存有可能提高性能吗？     | <ul style="list-style-type: none"> <li>包含大量读操作的工作负载可以受益于缓存层，如 <a href="#">ElastiCache</a> 或 <a href="#">DAX</a>（如果数据库是 <a href="#">DynamoDB</a>）。</li> <li>读取操作也可以通过关系数据库（如 <a href="#">Amazon RDS</a>）分流到只读副本。</li> </ul>                                                          |

| 问题                                                                                                             | 需要考虑的事项                                                                                                                                                                                                                                                                                               |
|----------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 存储和修改（OLTP – Online Transaction Processing，联机事务处理）还是检索和报告（OLAP – Online Analytical Processing，联机分析处理）具有更高的优先级？ | <ul style="list-style-type: none"><li>对于高吞吐量的按原样读取事务处理，可以考虑使用 NoSQL 数据库，例如 DynamoDB。</li><li>对于具有一致性的高吞吐量和复杂的读取模式（如联接），请使用 Amazon RDS。</li><li>对于分析查询，请考虑使用列存数据库（如 <a href="#">Amazon Redshift</a>），或者将数据导出到 Amazon S3 并使用 <a href="#">Athena</a> 或 <a href="#">Amazon QuickSight</a> 执行分析。</li></ul> |
| 数据需要什么级别的持久性？                                                                                                  | <ul style="list-style-type: none"><li>Aurora 自动在一个区域内的三个可用区复制您的数据，这意味着您的数据具有高度的持久性，数据丢失的可能性较小。</li><li>DynamoDB 自动跨多个可用区复制，提供高可用性和数据持久性。</li><li>Amazon S3 提供 11 个 9 的持久性。许多数据库服务（如 Amazon RDS 和 DynamoDB）支持将数据导出到 Amazon S3，以便进行长期留存和存档。</li></ul>                                                   |
| 是否希望摆脱商用数据库引擎或许可成本？                                                                                            | <ul style="list-style-type: none"><li>考虑使用 Amazon RDS 或 Aurora 上的开源引擎，如 PostgreSQL 和 MySQL。</li><li>利用 <a href="#">AWS Database Migration Service</a> 和 <a href="#">AWS Schema Conversion Tool</a> 执行从商用数据库引擎到开源引擎的迁移</li></ul>                                                                       |
| 对数据库的运维有什么期望？迁移到托管服务是主要的关注点吗？                                                                                  | <ul style="list-style-type: none"><li>利用 Amazon RDS 而不是 Amazon EC2，以及利用 DynamoDB 或 Amazon DocumentDB 而不是自行托管 NoSQL 数据库，可以减少运维开销。</li></ul>                                                                                                                                                            |

| 问题                                                                   | 需要考虑的事项                                                                                                                                                               |
|----------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 当前如何访问数据库？是只有应用程序访问，还是有商业智能（BI，Business Intelligence）用户和其他互联的现成应用程序？ | <ul style="list-style-type: none"> <li>如果您依赖于外部工具，那么您可能必须保持与这些工具所支持的数据库的兼容性。Amazon RDS 完全兼容其支持的不同引擎版本，包括 Microsoft SQL Server、Oracle、MySQL 和 PostgreSQL 等。</li> </ul> |

- 在非生产环境中进行试验和基准测试，确定哪种数据存储可以满足工作负载需求。

## 资源

### 相关文档：

- [Amazon EBS 卷类型](#)
- [Amazon EC2 存储](#)
- [Amazon EFS : Amazon EFS 性能](#)
- [Amazon FSx for Lustre 性能](#)
- [Amazon FSx for Windows File Server 性能](#)
- [Amazon S3 Glacier : S3 Glacier 文档](#)
- [Amazon S3 : 请求速率和性能注意事项](#)
- [使用 AWS 进行云存储](#)
- [Amazon EBS I/O 特性](#)
- [AWS 云数据库](#)
- [AWS 数据库缓存](#)
- [DynamoDB Accelerator](#)
- [Amazon Aurora 最佳实践](#)
- [Amazon Redshift 性能](#)
- [Amazon Athena 10 大性能提示](#)
- [Amazon Redshift Spectrum 最佳实践](#)
- [Amazon DynamoDB 最佳实践](#)
- [在 Amazon EC2 和 Amazon RDS 之间选择](#)

- [实施 Amazon ElastiCache 的最佳实践](#)

相关视频：

- [Deep dive on Amazon EBS](#)
- [Optimize your storage performance with Amazon S3](#)
- [Modernize apps with purpose-built databases](#)
- [Amazon Aurora storage demystified: How it all works](#)
- [Amazon DynamoDB deep dive: Advanced design patterns](#)

相关示例：

- [Amazon EFS CSI 驱动程序](#)
- [Amazon EBS CSI 驱动程序](#)
- [Amazon EFS 实用程序](#)
- [Amazon EBS 自动扩展](#)
- [Amazon S3 示例](#)
- [Optimize Data Pattern using Amazon Redshift Data Sharing](#)
- [数据库迁移](#)
- [MS SQL Server - AWS Database Migration Service \(AWS DMS\) Replication Demo](#)
- [数据库现代化动手实践研讨会](#)
- [Amazon Neptune 示例](#)

## PERF03-BP02 评估数据存储的可用配置选项

了解并评估数据存储的各种可用功能和配置选项，从而优化工作负载的存储空间和性能。

常见反模式：

- 您对所有工作负载都只使用一种存储类型，例如 Amazon EBS。
- 您对所有工作负载都使用预调配 IOPS，而没有对所有存储层进行真实测试。
- 您不了解所选数据管理解决方案的配置选项。
- 您只依赖于增加实例大小，而没有考虑其他可用的配置选项。
- 您没有测试数据存储的扩展特征。

建立此最佳实践的好处：通过探索和试用数据存储选项，您也许能够降低基础设施成本，提高性能，并减少维护工作负载所需的工作量。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

根据数据存储和访问要求，一个工作负载能够使用一个或多个数据存储。要优化性能效率和成本，必须评估数据访问模式来确定适当的数据存储配置。在研究数据存储选项时，要考虑存储选项、内存、计算、只读副本、一致性要求、连接池和缓存选项等各个方面。尝试使用这些不同的配置选项来提高性能效率指标。

实施步骤

- 了解数据存储的当前配置（如实例类型、存储大小或数据库引擎版本）。
- 查看 AWS 文档和最佳实践，了解有助于提高数据存储性能的推荐配置选项。需要考虑的关键数据存储选项如下：

| 配置选项               | 示例                                                                                                                                                                                                                                                                                                                                                         |
|--------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 分流读取操作（例如只读副本和缓存）  | <ul style="list-style-type: none"><li>• 对于 DynamoDB 表，您可以使用 DAX 缓存功能来分流读取操作。</li><li>• 您可以创建一个 Amazon ElastiCache (Redis OSS) 集群，并将应用程序配置为首先从缓存中读取，并在请求的项目不存在时使用数据库。</li><li>• 关系数据库（如 Amazon RDS 和 Aurora）以及预置的 NoSQL 数据库（如 Neptune 和 Amazon DocumentDB）全部支持添加只读副本，以便分流工作负载的读取部分。</li><li>• DynamoDB 等无服务器数据库将自动扩展。确保您预置了足够的读取容量单位（RCU）来处理工作负载。</li></ul> |
| 扩展写入（例如分区键分片或引入队列） | <ul style="list-style-type: none"><li>• 对于关系数据库，您可以增加实例的大小来适应增加的工作负载，或增加预调配 IOPS 来增加底层存储的吞吐量。</li></ul>                                                                                                                                                                                                                                                    |

| 配置选项                                                                                                                | 示例                                                                                                                                                                                                                                                                                                                                          |
|---------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                                                                                     | <ul style="list-style-type: none"><li>您还可以在数据库前面引入队列，而不是直接写入数据库。借助此模式，您可以将摄取操作与数据库解耦，并控制流量，这样数据库就不会过载。</li><li>对写入请求进行批处理，而不是创建许多短期事务，这样有助于提高有大量写入的关系数据库的吞吐量。</li><li>像 DynamoDB 这样的无服务器数据库可以自动扩展写入吞吐量，也可以根据容量模式调整预置的写入容量单位（WCU）。</li><li>当达到给定分区键的吞吐量限制时，仍然会遇到热分区问题。这可以通过选择更均匀分布的分区键或对分区键进行写分片来缓解。</li></ul>                               |
| 使用策略来管理数据集的生命周期                                                                                                     | <ul style="list-style-type: none"><li>您可以使用 <a href="#">Amazon S3 生命周期</a> 在对象的整个生命周期中对其进行管理。如果访问模式未知、不断变化或不可预测，可以使用 <a href="#">Amazon S3 Intelligent-Tiering</a>，此功能可监控访问模式，并自动将尚未访问的对象移动到成本较低的访问层。您可以利用 <a href="#">Amazon S3 Storage Lens</a> 指标来识别生命周期管理中的优化机会和差距。</li><li><a href="#">Amazon EFS 生命周期管理</a> 可自动管理文件系统的文件存储。</li></ul> |
| 连接管理和共享                                                                                                             | <ul style="list-style-type: none"><li>Amazon RDS 代理可与 Amazon RDS 和 Aurora 结合使用来管理与数据库的连接。</li><li>DynamoDB 等无服务器数据库没有与之关联的连接，但会考虑根据预置容量和自动扩展策略来处理负载峰值。</li></ul>                                                                                                                                                                            |
| <ul style="list-style-type: none"><li>在非生产环境中进行试验和基准测试，从而确定哪种配置选项可以满足您的工作负载要求。</li><li>试验完成后，规划迁移并验证性能指标。</li></ul> |                                                                                                                                                                                                                                                                                                                                             |



- 使用 AWS 监控工具（如 [Amazon CloudWatch](#)）和优化工具（如 [Amazon S3 Storage Lens](#)），在实际使用模式下持续优化数据存储。

## 资源

### 相关文档：

- [使用 AWS 进行云存储](#)
- [Amazon EBS 卷类型](#)
- [Amazon EC2 存储](#)
- [Amazon EFS : Amazon EFS 性能](#)
- [Amazon FSx for Lustre 性能](#)
- [Amazon FSx for Windows File Server 性能](#)
- [Amazon S3 Glacier : S3 Glacier 文档](#)
- [Amazon S3 : 请求速率和性能注意事项](#)
- [使用 AWS 进行云存储](#)
- [使用 AWS 进行云存储](#)
- [Amazon EBS I/O 特性](#)
- [AWS 云数据库](#)
- [AWS 数据库缓存](#)
- [DynamoDB Accelerator](#)
- [Amazon Aurora 最佳实践](#)
- [Amazon Redshift 性能](#)
- [Amazon Athena 10 大性能提示](#)
- [Amazon Redshift Spectrum 最佳实践](#)
- [Amazon DynamoDB 最佳实践](#)

### 相关视频：

- [Deep dive on Amazon EBS](#)
- [Optimize your storage performance with Amazon S3](#)

- [Modernize apps with purpose-built databases](#)
- [Amazon Aurora storage demystified: How it all works](#)
- [Amazon DynamoDB deep dive: Advanced design patterns](#)

相关示例：

- [Amazon EFS CSI 驱动程序](#)
- [Amazon EBS CSI 驱动程序](#)
- [Amazon EFS 实用程序](#)
- [Amazon EBS 自动扩展](#)
- [Amazon S3 示例](#)
- [Amazon DynamoDB 示例](#)
- [AWS 数据库迁移示例](#)
- [数据库现代化研讨会](#)
- [使用 Amazon RDS for Postgress DB 上的参数](#)

### PERF03-BP03 收集和记录数据存储性能指标

跟踪并记录数据存储的相关性能指标，了解数据管理解决方案的执行情况。这些指标有助于您优化数据存储，验证是否满足工作负载要求，并清晰地概述工作负载的表现情况。

常见反模式：

- 您只手动搜索日志文件来查找指标。
- 您只将指标发布到团队使用的内部工具，而没有全面了解您的工作负载。
- 您只使用由自己选定的监控软件记录的默认指标。
- 您只在出现问题时审查指标。
- 您只监控系统级指标，而不捕获数据访问或使用情况指标。

建立此最佳实践的好处：建立性能基准有助于了解工作负载的正常行为和需求。可以更快地识别和调试异常模式，从而提高数据存储的性能和可靠性。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

要监控数据存储的性能，必须记录一段时间的多项性能指标。这样您便可以检测异常并根据业务指标衡量性能，确保满足您的工作负载需求。

指标既应包括支持数据存储的底层系统指标，也应包括数据库指标。底层系统指标可能包括 CPU 利用率、内存、可用磁盘存储、磁盘 I/O、缓存命中率以及网络入站和出站指标，而数据存储指标可能包括每秒事务数、最多的查询、平均查询速率、响应时间、索引使用情况、表锁定、查询超时和打开的连接数。这些数据对于了解工作负载的表现情况以及数据管理解决方案的使用方式至关重要。在数据驱动方法中使用这些指标，以便调整和优化工作负载的资源。

使用各种工具、库和系统来记录与数据库性能相关的性能测量值。

## 实施步骤

1. 确定要跟踪的数据存储关键性能指标。
  - a. [Amazon S3 指标和维度](#)
  - b. [监控 Amazon RDS 实例中的指标](#)
  - c. [在 Amazon RDS 上使用 Performance Insights 监控数据库负载](#)
  - d. [增强监控概述](#)
  - e. [DynamoDB 指标和维度](#)
  - f. [监控 DynamoDB Accelerator](#)
  - g. [使用 Amazon CloudWatch 监控 Amazon MemoryDB](#)
  - h. [我应该监控哪些指标？](#)
  - i. [监控 Amazon Redshift 集群性能](#)
  - j. [Timestream 指标和维度](#)
  - k. [Amazon Aurora 的 Amazon CloudWatch 指标](#)
  - l. [Amazon Keyspaces \(for Apache Cassandra\) 中的日志记录和监控](#)
  - m. [监控 Amazon Neptune 资源](#)
2. 使用经批准的日志记录和监控解决方案来收集这些指标。[Amazon CloudWatch](#) 可以收集架构中各种资源的指标。您也可以收集和发布自定义指标，用于显示业务指标或派生指标。使用 CloudWatch 或第三方解决方案来设置指示超出阈值的警报。
3. 检查数据存储监控，确定其能否受益于可检测性能异常的机器学习解决方案。
  - a. [Amazon DevOps Guru for Amazon RDS](#) 会显示性能问题，并提出纠正措施的建议。
4. 在监控和日志记录解决方案中配置数据留存，从而满足您的安全和运营目标。

- a. [CloudWatch 指标的默认数据留存](#)
- b. [CloudWatch Logs 的默认数据留存](#)

## 资源

### 相关文档：

- [AWS 数据库缓存](#)
- [Amazon Athena 10 大性能提示](#)
- [Amazon Aurora 最佳实践](#)
- [DynamoDB Accelerator](#)
- [Amazon DynamoDB 最佳实践](#)
- [Amazon Redshift Spectrum 最佳实践](#)
- [Amazon Redshift 性能](#)
- [AWS 云数据库](#)
- [Amazon RDS Performance Insights](#)

### 相关视频：

- [AWS purpose-built databases](#)
- [Amazon Aurora storage demystified: How it all works](#)
- [Amazon DynamoDB deep dive: Advanced design patterns](#)
- [Best Practices for Monitoring Redis Workloads on Amazon ElastiCache](#)

### 相关示例：

- [第 100 级：使用 CloudWatch 控制面板进行监控](#)
- [AWS 数据集摄取指标收集框架](#)
- [Amazon RDS 监控研讨会](#)

## PERF03-BP04 实施可提高数据存储查询性能的策略

实施可优化数据和改进数据查询的策略，从而提高工作负载的可扩展性和性能效率。

## 常见反模式：

- 您没有对数据存储中的数据进行分区。
- 您在数据存储中只以一种文件格式存储数据。
- 您没有在数据存储中使用索引。

建立此最佳实践的好处： 优化数据和查询性能可以提高效率、降低成本并改善用户体验。

未建立这种最佳实践的情况下暴露的风险等级： 中

## 实施指导

数据优化和查询调整是提高数据存储性能效率的关键环节，因为这会影响整个云工作负载的性能和响应能力。如果查询未经优化，则会耗用更多的资源并产生更多的瓶颈，从而降低数据存储的整体效率。

数据优化会涵盖多种技术，旨在确保高效的数据存储和访问，同时还有助于改进在数据存储中的查询性能。关键策略包括数据分区、数据压缩和数据去规范化，这有助于针对存储和访问优化数据。

## 实施步骤

- 了解并分析在数据存储中执行的关键数据查询。
- 识别数据存储中运行速度较慢的查询，并使用查询计划了解当前状态。
  - [在 Amazon Redshift 中分析查询计划](#)
  - [在 Athena 中使用 EXPLAIN 和 EXPLAIN ANALYZE](#)
- 实施可提高查询性能的策略。一些关键策略包括：
  - 使用 [列式文件格式](#)（比如 Parquet 或 ORC）。
  - 压缩数据存储中的数据，以减少存储空间和 I/O 操作。
  - 进行数据分区，将数据分割成更小的部分，减少数据扫描时间。
    - [在 Athena 中对数据进行分区](#)
    - [分区和数据分发](#)
  - 对查询中的常用列编制数据索引。
  - 为查询选择合适的联接操作。联接两个表时，请在联接的左侧指定较大的表，在联接的右侧指定较小的表。
  - 实施分布式缓存解决方案，从而缩短延迟和减少数据库 I/O 操作次数。
  - 定期维护，例如执行统计。

- 在非生产环境中试验和测试策略。

## 资源

### 相关文档：

- [Amazon Aurora 最佳实践](#)
- [Amazon Redshift 性能](#)
- [Amazon Athena 10 大性能提示](#)
- [AWS 数据库缓存](#)
- [实施 Amazon ElastiCache 的最佳实践](#)
- [在 Athena 中对数据进行分区](#)

### 相关视频：

- [Optimize Data Pattern using Amazon Redshift Data Sharing](#)
- [Optimize Amazon Athena Queries with New Query Analysis Tools](#)

### 相关示例：

- [Amazon EFS CSI 驱动程序](#)

## PERF03-BP05 实现利用缓存的数据访问模式

实施可从缓存数据受益的访问模式，以便快速检索经常访问的数据。

### 常见反模式：

- 您缓存经常变化的数据。
- 您依赖缓存的数据，就好像这些数据是持久存储的，并且始终可用。
- 您不考虑缓存数据的一致性。
- 您不监控缓存实现方案的效率。

建立此最佳实践的好处：将数据存储于缓存中可以改善读取延迟、读取吞吐量、用户体验和整体效率，还可以降低成本。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

缓存是一种软件或硬件组件，旨在存储数据，以便将来可以更快或更高效地处理对相同数据的请求。如果存储在缓存中的数据丢失，则可以通过重复先前的计算或其他数据存储中获取数据进行重建。

数据缓存可能是提高应用程序整体性能和减轻底层主数据源负担的最有效策略之一。数据可以在应用程序中的多个层面上缓存，例如在进行远程调用的应用程序内，这称为 **客户端缓存**，或者使用快速的辅助服务来存储数据，这称为 **远程缓存**。

### 客户端缓存

借助客户端缓存，每个客户端（查询后端数据存储的应用程序或服务）都可以在本地将特定查询的结果存储指定的时间。可以通过先检查本地客户端缓存，减少通过网络向数据存储发出的请求数量。如果结果不存在，则应用程序可以查询数据存储并将这些结果存储在本地。这种模式允许每个客户端将数据存储在尽可能近的位置（客户端本身），从而尽可能降低延迟。当后端数据存储不可用时，客户端还可以继续支持某些查询，从而提高整个系统的可用性。

这种方法的一个缺点是，当涉及多个客户端时，它们可能会在本地存储相同的缓存数据。这会导致这些客户端之间存在重复的存储使用情况和数据不一致性。一个客户端可能刚缓存查询结果，而一分钟后，另一个客户端可能运行相同的查询并得到不同的结果。

### 远程缓存

要解决客户端之间数据重复的问题，可使用快速外部服务或 **远程缓存** 来存储查询的数据。在查询后端数据存储之前，每个客户端都将检查远程缓存，而不是检查本地数据存储。这种策略可在客户端之间实现更加一致的响应、更高的存储数据效率以及更高的缓存数据量，因为存储空间可独立于客户端进行扩展。

远程缓存的缺点是整个系统的延迟可能会更高，因为需要额外的网络跃点数来检查远程缓存。客户端缓存可以与远程缓存一起使用，以实现多级缓存，从而缩短延迟。

## 实施步骤

1. 确定可以从缓存中受益的数据库、API 和网络服务。读取工作负载繁重、读写比率高或扩展成本高昂的服务适合使用缓存。
  - [数据库缓存](#)
  - [启用 API 缓存以增强响应能力](#)
2. 确定最适合您的访问模式的适当缓存策略类型。



- [缓存策略](#)
  - [AWS 缓存解决方案](#)
3. 遵循适合您的数据存储的 [缓存最佳实践](#)。
4. 为所有数据配置缓存失效策略，例如生存时间（TTL），以平衡数据的时效性并减轻后端数据存储的压力。
5. 启用诸如自动连接重试、指数回退、客户端超时和客户端连接池等功能（如果有），因为它们可以提高性能和可靠性。
- [最佳实践：Redis 客户端和 Amazon ElastiCache \(Redis OSS\)](#)
6. 监控缓存命中率，目标为 80% 或更高。较低的值可能表示缓存大小不足，或访问模式无法从缓存中受益。
- [我应该监控哪些指标？](#)
  - [在 Amazon ElastiCache 上监控 Redis 工作负载的最佳实践](#)
  - [Monitoring best practices with Amazon ElastiCache \(Redis OSS\) using Amazon CloudWatch](#)
7. 实施 [数据复制](#) 将读取操作分流到多个实例，以提高数据读取性能和可用性。

## 资源

### 相关文档：

- [使用 Amazon ElastiCache Well-Architected Lens](#)
- [Monitoring best practices with Amazon ElastiCache \(Redis OSS\) using Amazon CloudWatch](#)
- [我应该监控哪些指标？](#)
- [《Performance at Scale with Amazon ElastiCache》白皮书](#)
- [缓存挑战和策略](#)

### 相关视频：

- [Amazon ElastiCache Learning Path](#)
- [Design for success with Amazon ElastiCache best practices](#)

### 相关示例：

- [Boosting MySQL database performance with Amazon ElastiCache \(Redis OSS\)](#)

# 联网和内容分发

## PERF 4.如何在工作负载中选择和配置网络资源？

针对特定系统的最有效数据库解决方案取决于您的具体需求，包括可用性、一致性、分区容错性、延迟、持久性、可扩展性以及查询能力等等。许多系统会使用多种不同的数据库解决方案来满足其各个子系统的实际需要，并启用不同的功能来提高性能。为系统选择错误的数据库解决方案和功能可能会导致性能效率降低。

### 最佳实践

- [PERF04-BP01 了解联网对性能的影响](#)
- [PERF04-BP02 评估可用的联网功能](#)
- [PERF04-BP03 为工作负载选择合适的专用连接或 VPN](#)
- [PERF04-BP04 使用负载均衡在多个资源之间分配流量](#)
- [PERF04-BP05 选择网络协议以提高性能](#)
- [PERF04-BP06 根据网络要求选择工作负载的位置](#)
- [PERF04-BP07 根据指标优化网络配置](#)

### PERF04-BP01 了解联网对性能的影响

分析并了解与网络相关的决策如何影响您的工作负载，从而提供更高的性能和更好的用户体验。

#### 常见反模式：

- 所有流量都会流经您现有的数据中心。
- 您通过中央防火墙路由所有流量，而不是使用云原生网络安全工具。
- 您在不了解实际使用要求的情况下预置 AWS Direct Connect 连接。
- 在确立联网解决方案时，您未考虑工作负载特性和加密开销。
- 您将本地概念和策略用于云中的联网解决方案。

建立此最佳实践的好处：通过了解联网如何影响工作负载性能，有助于您识别潜在的瓶颈、改善用户体验、提高可靠性并在工作负载发生变化时减少运营维护。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

网络负责应用程序组件、云服务、边缘网络和本地数据之间的连接，因此，它会严重影响工作负载性能。除了工作负载性能之外，用户体验也会受到网络延迟、带宽、协议、位置、网络拥塞、抖动、吞吐量和路由规则的影响。

清楚记录工作负载的联网要求列表，包括延迟、数据包大小、路由规则、协议和支持的流量模式。查看可用的联网解决方案，并确定哪种服务与您的工作负载联网特性相符。基于云的网络可以快速重建，因此有必要随着时间的推移改进网络架构，以提高性能效率。

### 实施步骤：

1. 定义和记录网络性能要求，包括网络延迟、带宽、协议、位置、流量模式（峰值和频率）、吞吐量、加密、检查和路由规则等指标。
2. 了解关键的 AWS 网络服务，例如 [VPC](#)、[AWS Direct Connect](#)、[Elastic Load Balancing \( ELB \)](#) 和 [Amazon Route 53](#)。
3. 捕获以下关键网络特性：

| 特性       | 工具和指标                                                                                                                                                                                                                     |
|----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 基础网络特性   | <ul style="list-style-type: none"><li>• <a href="#">VPC 流日志</a></li><li>• <a href="#">AWS Transit Gateway 流日志</a></li><li>• <a href="#">AWS Transit Gateway 指标</a></li><li>• <a href="#">AWS PrivateLink 指标</a></li></ul> |
| 应用程序网络特性 | <ul style="list-style-type: none"><li>• <a href="#">Elastic Fabric Adapter</a></li><li>• <a href="#">AWS App Mesh 指标</a></li><li>• <a href="#">Amazon API Gateway 指标</a></li></ul>                                        |
| 边缘网络特性   | <ul style="list-style-type: none"><li>• <a href="#">Amazon CloudFront 指标</a></li><li>• <a href="#">Amazon Route 53 指标</a></li><li>• <a href="#">AWS Global Accelerator 指标</a></li></ul>                                   |
| 混合网络特性   | <ul style="list-style-type: none"><li>• <a href="#">AWS Direct Connect 指标</a></li><li>• <a href="#">AWS Site-to-Site VPN 指标</a></li><li>• <a href="#">AWS Client VPN 指标</a></li></ul>                                     |

| 特性     | 工具和指标                                                                                                                                                                                                                                                                              |
|--------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | <ul style="list-style-type: none"> <li>• <a href="#">AWS Cloud WAN 指标</a></li> </ul>                                                                                                                                                                                               |
| 安全网络特性 | <ul style="list-style-type: none"> <li>• <a href="#">AWS Shield、AWS WAF 和 AWS Network Firewall 指标</a></li> </ul>                                                                                                                                                                   |
| 跟踪特性   | <ul style="list-style-type: none"> <li>• <a href="#">AWS X-Ray</a></li> <li>• <a href="#">VPC Reachability Analyzer</a></li> <li>• <a href="#">Network Access Analyzer</a></li> <li>• <a href="#">Amazon Inspector</a></li> <li>• <a href="#">Amazon CloudWatch RUM</a></li> </ul> |

#### 4. 对网络性能进行基准测试和其他测试：

- a. [对网络吞吐量进行基准测试](#)，因为当实例位于同一 VPC 中时，一些因素可能会影响 Amazon EC2 网络性能。测量同一 VPC 中的 Amazon EC2 Linux 实例之间的网络带宽。
- b. 执行 [负载测试](#) 以试用各种联网解决方案和选项。

## 资源

### 相关文档：

- [Application Load Balancer](#)
- [Linux 上的 EC2 增强联网](#)
- [Windows 上的 EC2 增强联网](#)
- [EC2 置放群组](#)
- [在 Linux 实例上启用弹性网络适配器 \( ENA \) 增强联网](#)
- [Network Load Balancer](#)
- [AWS 联网产品](#)
- [Transit Gateway](#)
- [过渡到 Amazon Route 53 中基于延迟的路由](#)
- [VPC 终端节点](#)
- [VPC 流日志](#)

### 相关视频：

- [连接 AWS 和混合 AWS 网络架构](#)
- [优化 Amazon EC2 实例的网络性能](#)
- [提高应用程序的全球网络性能](#)
- [EC2 实例和性能优化最佳实践](#)
- [优化 Amazon EC2 实例的网络性能](#)
- [使用 Well-Architected Framework 进行联网的最佳实践和技巧](#)
- [大规模迁移中的 AWS 联网最佳实践](#)

相关示例：

- [AWS Transit Gateway 和可扩展的安全解决方案](#)
- [AWS 联网研讨会](#)

#### PERF04-BP02 评估可用的联网功能

评估云中可能提高性能的联网功能。借助测试、指标和分析来衡量这些功能的影响。例如，利用可用的网络级功能来减少延迟、网络距离或抖动。

常见反模式：

- 您一直待在一个区域，因为这是您的总部实际所在的区域。
- 您使用防火墙而不是安全组来过滤流量。
- 您中断 TLS 来进行流量检查，而不是依赖安全组、端点策略和其他云原生功能。
- 您只能使用基于子网的分段，而不是安全组。

建立此最佳实践的好处：评估所有服务功能和选项可以提高您的工作负载性能，降低基础设施的成本，减少维护工作负载所需的工作量，并提升您的整体安全状况。您可以利用 AWS 的全球骨干网，为客户提供出色的联网体验。

未建立这种最佳实践的情况下暴露的风险等级：高

#### 实施指导

AWS 提供 [AWS Global Accelerator](#) 和 [Amazon CloudFront](#) 等有助于提高网络性能的服务，而大多数 AWS 服务都具有相应的产品功能（例如 [Amazon S3 Transfer Acceleration](#) 功能）来优化网络流量。

查看您可以使用哪些与网络相关的配置选项，以及这些配置选项对您的工作负载有何影响。要想优化性能，需要了解这些选项如何与您的架构进行交互，以及它们将对测得的性能和用户体验产生的影响。

实施步骤

- 创建工作负载组件列表。
  - 在搭建统一的全球网络时，考虑使用 [AWS Cloud WAN](#) 来构建、管理和监控您组织的网络。
  - 使用 [Amazon CloudWatch Logs 指标](#) 监控您的全球和核心网络。利用 [Amazon CloudWatch RUM](#)，它提供了有助于识别、理解和增强用户的数字体验的见解。
  - 查看 AWS 区域 和可用区之间以及每个可用区内的聚合网络延迟，使用 [AWS Network Manager](#) 深入了解您的应用程序性能与底层 AWS 网络性能的关系。
  - 使用现有的配置管理数据库 ( CMDB ) 工具或 [AWS Config](#) 等服务创建工作负载清单及其配置方式。
- 如果这是一个现有的工作负载，请确定并记录性能指标的基准，重点关注瓶颈和需要改进之处。基于业务要求和工作负载特征，与性能相关的网络指标将因工作负载而异。首先，对于您的工作负载，检查带宽、延迟、数据包丢失、抖动和重传等指标可能很重要。
- 如果这是一个新的工作负载，请执行 [负载测试](#) 以识别性能瓶颈。
- 对于识别的性能瓶颈，请查看解决方案的配置选项，以确定性能改进机会。查看以下主要联网选项和功能：

| 改进机会    | 解决方案                                                                                                                                                                                                                                          |
|---------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 网络路径或路由 | 使用 <a href="#">Network Access Analyzer</a> 以识别路径或路由。                                                                                                                                                                                          |
| 网络协议    | 请参阅 <a href="#">PERF04-BP05 选择网络协议以提高性能</a>                                                                                                                                                                                                   |
| 网络拓扑    | 当连接多个账户时，请评估 <a href="#">VPC 对等</a> 和 <a href="#">AWS Transit Gateway</a> 之间的运营和性能权衡。AWS Transit Gateway 可简化所有 VPC 之间的互连，这些 VPC 可以跨越数千个 AWS 账户 并接入您的本地网络。在多个账户之间共享您的 AWS Transit Gateway - 通过使用 <a href="#">AWS Resource Access Manager</a> 。 |

| 改进机会   | 解决方案                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | 请参阅 <a href="#">PERF04-BP03 为工作负载选择合适的专用连接或 VPN</a>                                                                                                                                                                                                                                                                                                                                                                                                               |
| 网络服务   | <p><a href="#">AWS Global Accelerator</a> 是一项网络服务，使用 AWS 全球网络基础设施，可将用户流量的性能提高多达 60%。</p> <p><a href="#">Amazon CloudFront</a> 可在全球范围内提高工作负载内容交付性能并减少延迟。</p> <p>使用 <a href="#">Lambda@edge</a> 运行一些函数，这些函数可自定义 CloudFront 在离用户更近的位置提供的内容、减少延迟并提高性能。</p> <p>Amazon Route 53 提供 <a href="#">基于延迟的路由</a>、<a href="#">地理位置路由</a>、<a href="#">地理位置临近度路由</a>和 <a href="#">基于 IP 的路由</a> 选项，以帮助您提高面向全球受众的工作负载性能。当您的工作负载分布在全球时，通过查看您的工作负载流量和用户位置，确定哪种路由选项可以优化您的工作负载性能。</p> |
| 存储资源功能 | <p><a href="#">Amazon S3 Transfer Acceleration</a> 是一项功能，借助该功能，外部用户在向 Amazon S3 上传数据时可以通过 CloudFront 的网络优化获益。这就提高了将大量数据从没有专用连接的远程位置传输到 AWS Cloud 的能力。</p> <p><a href="#">Amazon S3 多区域接入点</a> 将内容复制到多个区域，并通过提供一个接入点简化了工作负载。使用多区域接入点时，您可以使用标识最低延迟桶的服务向 Amazon S3 请求或写入数据。</p>                                                                                                                                                                                      |



| 改进机会   | 解决方案                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|--------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 计算资源功能 | <p><a href="#">弹性网络接口 ( ENI )</a> ( 由 Amazon EC2 实例、容器和 Lambda 函数使用 ) 按流进行限制。查看您的置放群组以优化 <a href="#">EC2 网络吞吐量</a>。为避免每个流上出现瓶颈，请将应用程序设计为使用多个流。要监控和查看与计算相关的网络指标，请使用 CloudWatch 指标和 <a href="#">ethtool</a>。此 ethtool 命令包含在 ENA 驱动程序中，并公开了其他与网络相关的指标，这些指标可作为 <a href="#">自定义指标</a> 发布到 CloudWatch。</p> <p><a href="#">Amazon 弹性网络适配器 ( ENA )</a> 通过为 <a href="#">集群放置组</a>。</p> <p><a href="#">Elastic Fabric Adapter ( EFA )</a> 是 Amazon EC2 实例的网络接口，使您能够在 AWS 上大规模运行需要高级别节点间通信的工作负载。</p> <p><a href="#">Amazon EBS 优化实例</a> 使用经过优化的配置堆栈，可以提供额外的专用容量来提高 Amazon EBS I/O。</p> |

## 资源

### 相关文档：

- [Amazon EBS - 优化实例](#)
- [Application Load Balancer](#)
- [Linux 上的 EC2 增强联网](#)
- [Windows 上的 EC2 增强联网](#)
- [EC2 置放群组](#)
- [在 Linux 实例上启用弹性网络适配器 \( ENA \) 增强联网](#)
- [Network Load Balancer](#)
- [AWS 联网产品](#)
- [AWS Transit Gateway](#)

- [过渡到 Amazon Route 53 中基于延迟的路由](#)
- [VPC 终端节点](#)
- [VPC 流日志](#)

相关视频：

- [连接 AWS 和混合 AWS 网络架构](#)
- [优化 Amazon EC2 实例的网络性能](#)
- [AWS Global Accelerator](#)

相关示例：

- [AWS Transit Gateway 和可扩展的安全解决方案](#)
- [AWS 联网研讨会](#)

PERF04-BP03 为工作负载选择合适的专用连接或 VPN

当需要混合连接来连接本地资源和云资源时，请预置足够的带宽以满足您的性能要求。估算混合工作负载的带宽和延迟要求。这些数字将推动您的规模需求。

常见反模式：

- 您仅根据网络加密要求评估 VPN 解决方案。
- 您不评估备用或冗余连接选项。
- 您不确定全部工作负载要求（加密、协议、带宽和流量需求）。

建立此最佳实践的好处：选择和配置适当的连接解决方案将会提高工作负载的可靠性，并最大限度地提高性能。通过确定工作负载要求、提前规划和评估混合解决方案，您可以最大限度地减少成本高昂的物理网络变更和运营开销，同时加快实现价值的速度。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

根据您的带宽要求建立混合联网架构。[AWS Direct Connect](#) 允许您使用 AWS 私密地连接本地网络。它适用于需要高带宽、低延迟，同时实现一致性能的情况。VPN 连接通过互联网建立安全连接。在以

下情况下可使用 VPN 连接：只需要临时连接、需要考虑成本因素，或者在使用 AWS Direct Connect 的情况下等待建立弹性物理网络连接时作为应急措施。

如果您的带宽要求很高，则可以考虑使用多种 AWS Direct Connect 或 VPN 服务。可以在服务之间对流量进行负载平衡，但由于延迟和带宽差异，我们不建议在 AWS Direct Connect 和 VPN 之间进行负载平衡。

## 实施步骤

1. 估计现有应用程序的带宽和延迟要求。
  - a. 对于迁移到 AWS 的现有工作负载，利用来自内部网络监控系统的数据。
  - b. 对于新工作负载或您没有监控数据的现有工作负载，请咨询产品所有者，以确定足够的性能指标并提供良好的用户体验。
2. 选择专用连接或 VPN 作为连接选项。根据所有工作负载要求（加密、带宽和流量需求），您可以选择 AWS Direct Connect 或 [AWS VPN](#)（或两者）。下图可协助您选择适当的连接类型。
  - a. [AWS Direct Connect](#) 使用专用连接或托管连接，提供到 AWS 环境的专用连接，速度从 50Mbps 到 100Gbps 不等。这样一来，延迟得到管理和控制，并且拥有预置带宽，让您的工作负载能够高效地连接到其他环境。使用 AWS Direct Connect 合作伙伴，您可以从多个环境获得端到端连接，从而提供具有一致性能的扩展网络。AWS 使用原生 100Gbps、链接聚合组（LAG）或 BGP 同等成本多路径（ECMP）提供扩展 Direct Connect 连接带宽。
  - b. AWS [Site-to-Site VPN](#) 提供支持互联网协议安全性（IPsec）的托管 VPN 服务。创建 VPN 连接时，每个 VPN 连接包括两条隧道以实现高可用性。
3. 按照 AWS 文档选择合适的连接选项：
  - a. 如果您决定使用 AWS Direct Connect，请为您的连接选择合适的带宽。
  - b. 如果您在多个位置使用 AWS Site-to-Site VPN 连接到 AWS 区域，请使用 [加速的 Site-to-Site VPN 连接](#)，以便有机会提高网络性能。
  - c. 如果您的网络设计包含基于 [AWS Direct Connect](#) 的 IPsec VPN 连接，请考虑使用专用 IP VPN，来提高安全性并实现分段。[AWS 点对点专用 IP VPN](#) 部署在传输虚拟接口（VIF）之上。
  - d. [AWS Direct Connect SiteLink](#) 允许通过 [AWS Direct Connect 位置](#) 之间最快的路径发送数据，并绕过 AWS 区域，从而在全球各地的数据中心之间创建低延迟的冗余连接。
4. 在部署到生产环境之前，验证您的连接设置。执行安全和性能测试，确保其满足您的带宽、可靠性、延迟和合规性要求。
5. 定期监控您的连接性能和使用情况，并在需要进行优化。

## 确定性性能流程图

### 资源

#### 相关文档：

- [Network Load Balancer](#)
- [AWS 联网产品](#)
- [AWS Transit Gateway](#)
- [过渡到 Amazon Route 53 中基于延迟的路由](#)
- [VPC 终端节点](#)
- [Site-to-Site VPN](#)
- [构建可扩展且安全的多 VPC AWS 网络基础设施](#)
- [AWS Direct Connect](#)
- [Client VPN](#)

#### 相关视频：

- [连接 AWS 和混合 AWS 网络架构](#)
- [优化 Amazon EC2 实例的网络性能](#)
- [AWS Global Accelerator](#)
- [AWS Direct Connect](#)
- [AWS Transit Gateway Connect](#)
- [VPN 解决方案](#)
- [VPN 解决方案的安全性](#)

#### 相关示例：

- [AWS Transit Gateway 和可扩展的安全解决方案](#)
- [AWS 联网研讨会](#)

## PERF04-BP04 使用负载均衡在多个资源之间分配流量

跨多个资源或服务分配流量，以便让工作负载能够利用云提供的弹性。您也可以使用负载均衡机制来卸载加密终端，以便提高性能和可靠性，并有效管理和路由流量。

常见反模式：

- 在选择负载均衡器类型时不考虑工作负载要求。
- 不利用负载均衡器功能进行性能优化。
- 工作负载直接暴露给互联网，而不使用负载均衡器。
- 您通过现有负载均衡器来路由所有互联网流量。
- 您使用通用 TCP 负载均衡，并让每个计算节点处理 SSL 加密。

建立此最佳实践的好处：负载均衡器可在单个可用区内或多个可用区之间处理您的应用程序不断变化的流量负载，并实现高可用性、自动扩展和更高的工作负载利用率。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

负载均衡器充当工作负载的接入点，它们从这里将流量分发到您的后端目标，例如计算实例或容器，以提高利用率。

优化架构的第一步是选择适合的负载均衡器类型。首先列出工作负载特征，例如协议（如 TCP、HTTP、TLS 或 WebSocket）、目标类型（如实例、容器或无服务器）、应用程序要求（如长时间运行的连接、用户身份验证或粘性）和置放（如区域、Local Zone、Outpost 或区域隔离）。

AWS 为您的应用程序提供多种模型，以使用负载均衡。[Application Load Balancer](#) 最适合对 HTTP 和 HTTPS 流量进行负载均衡，面向交付包括微服务和容器在内的现代应用程序架构，提供高级请求路由功能。

[Network Load Balancer](#) 最适合对需要极高性能的 TCP 流量进行负载均衡。网络负载均衡器每秒能够处理数百万请求，同时能保持超低延迟，还针对处理突发和不稳定的流量模式进行了优化。

[Elastic Load Balancing](#) 提供集成的证书管理和 SSL/TLS 解密，使您可以灵活地集中管理负载均衡器的 SSL 设置，并从工作负载中卸载占用大量 CPU 的工作。

选择适合的负载均衡器之后，您可以开始利用其功能来减少后端为提供流量所付出的工作量。

例如，您可以使用 Application Load Balancer ( ALB ) 和 Network Load Balancer ( NLB ) 执行 SSL/TLS 加密卸载，这是一个机会，可以避免目标完成 CPU 密集型 TLS 握手，同时还可以改进证书管理。

当您在负载均衡器中配置 SSL/TLS 卸载时，它负责加密进出客户端的流量，同时将未加密的流量传输到您的后端，释放后端资源和缩短客户端的响应时间。

Application Load Balancer 还可以提供 HTTP/2 流量，无需在您的目标上支持它。因为 HTTP/2 可以更高效地使用 TCP 连接，所以这个简单的决定可以缩短应用程序的响应时间。

在定义架构时应考虑工作负载延迟要求。例如，如果您有延迟敏感型应用程序，则可以决定使用 Network Load Balancer，它提供极低的延迟。或者，您可以决定利用 Application Load Balancer ( 在 [AWS Local Zones](#) 甚至 [AWS Outposts](#) 中 )，让工作负载更靠近客户。

延迟敏感型工作负载的另一个考虑因素是跨区域负载平衡。借助跨区域负载平衡，每个负载均衡器节点在所有允许的可用区中的已注册目标之间分配流量。

使用与负载均衡器集成的 Auto Scaling。高效性能系统的其中一个关键方面与合理调整后端资源的大小有关。为此，您可以为后端目标资源使用负载均衡器集成。使用负载均衡器与 Auto Scaling 组的集成，根据需要在负载均衡器中添加或删除目标，以应对传入流量。负载均衡器还可以与 [Amazon ECS](#) 和 [Amazon EKS](#) 集成，以处理容器化工作负载。

- [Amazon ECS - 服务负载均衡](#)
- [Amazon EKS 上的应用程序负载均衡](#)
- [Amazon EKS 上的网络负载均衡](#)

## 实施步骤

- 定义您的负载平衡要求，包括流量、可用性和应用程序可扩展性。
- 为您的应用程序选择正确的负载均衡器类型。
  - 为 HTTP/HTTPS 工作负载使用 Application Load Balancer。
  - 为在 TCP 或 UDP 上运行的非 HTTP 工作负载使用 Network Load Balancer。
  - 如果您想利用两个产品的功能，则可以使用两者的组合 ( [ALB 作为 NLB 的目标](#) )。例如，如果想要将 NLB 的静态 IP 与 ALB 基于 HTTP 标头的路由结合使用，或者想要将 HTTP 工作负载向 [AWS PrivateLink](#) 公开，则可以使用此组合。
  - 有关负载均衡器的完整比较，请参阅 [ELB 产品比较](#)。
- 如果可能，请使用 SSL/TLS 卸载。

- 配置 HTTPS/TLS 侦听器，将 [Application Load Balancer](#) 和 [Network Load Balancer](#) 与 [AWS Certificate Manager](#) 集成。
- 请注意，出于合规性原因，有些工作负载可能需要端到端加密。在这种情况下，必须允许在目标上启用加密。
- 有关安全最佳实践，请参阅 [SEC09-BP02 在传输中执行加密](#)。
- 选择适合的路由算法（仅限 ALB）。
  - 路由算法会影响后端目标的使用情况，从而决定它们对性能的影响。例如，ALB 提供 [两个路由算法选项](#)：
  - 最少未完成请求：用于在应用程序的请求复杂程度不同或目标处理能力不同的情况下，实现更好的后端目标负载分布。
  - 循环：当请求和目标类似，或需要在目标之间平均分配请求时使用。
- 考虑跨区域和区域隔离。
  - 使用跨区域关闭（区域隔离）来改善延迟和区域故障域。它在 NLB 中默认关闭，而 [在 ALB 中，您可以按目标组来关闭](#)。
  - 使用跨区域开启以提高可用性和灵活性。默认情况下，为 ALB 开启跨区域，而 [在 NLB 中，您可以按目标组来开启](#)。
- 为 HTTP 工作负载开启 HTTP 保持活动（仅限 ALB）。借助此功能，在保持活动超时到期之前，负载均衡器可以重复使用后端连接，从而改进 HTTP 请求和缩短响应时间，还可以降低后端目标的资源利用率。有关如何为 Apache 和 Nginx 执行此操作的详细信息，请参阅 [使用 Apache 或 NGINX 作为 ELB 的后端服务器的最佳设置是什么？](#)
- 为您的负载均衡器开启监控。
  - 启用 [Application Load Balancer](#) 和 [Network Load Balancer](#) 的访问日志。
  - 对于 ALB，要考虑的主要方面是 request\_processing\_time、request\_processing\_time 和 response\_processing\_time。
  - 对于 NLB，要考虑的主要方面是 connection\_time 和 tls\_handshake\_time。
  - 准备好在需要时查询日志。您可以使用 Amazon Athena 来查询 [ALB 日志](#) 和 [NLB 日志](#)。
  - 为性能相关指标（例如 [TargetResponseTime for ALB](#)）创建警报。

## 资源

### 相关文档：

- [ELB 产品比较](#)



- [AWS 全球基础设施](#)
- [使用可用区关联性提高性能和降低成本](#)
- [使用 Amazon Athena 进行日志分析的分步指南](#)
- [查询 Application Load Balancer 日志](#)
- [监控 Application Load Balancers](#)
- [监控 Network Load Balancer](#)
- [使用 Elastic Load Balancing 在 Auto Scaling 组中的不同实例间分配流量](#)

相关视频：

- [AWS re:Invent 2018:Elastic Load Balancing: 深度挖掘和最佳实践](#)
- [AWS re:Invent 2021 - 如何为 AWS 工作负载选择适合的负载均衡器](#)
- [AWS re:Inforce 2022 - 如何使用 Elastic Load Balancing 全面改善安全状况](#)
- [AWS re:Invent 2019：充分利用适用于不同工作负载的 Elastic Load Balancing](#)

相关示例：

- [使用 Amazon Athena 进行日志分析的 CDK 和 AWS CloudFormation 示例](#)

## PERF04-BP05 选择网络协议以提高性能

根据对工作负载性能的影响，做出有关系统与网络之间的通信协议的决策。

延迟和带宽之间的关系可以实现高吞吐量。如果文件传输使用传输控制协议（TCP），则延迟越高，整体吞吐量很可能越低。有一些方法可以使用 TCP 调整和优化的传输协议来解决此问题，但一种解决方案是使用用户数据报协议（UDP）。

常见反模式：

- 无论有怎样的性能要求，您都可以为所有工作负载使用 TCP。

建立此最佳实践的好处：确认已为用户和工作负载组件之间的通信使用适当的协议，有助于改善应用程序的整体用户体验。例如，无连接 UDP 虽然允许较高速度，但不提供重新传输或高可靠性。TCP 虽然是一个功能全面的协议，但它在处理这些数据包时需要较高的开销。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

如果您能够为应用程序选择不同的协议，并且具有该领域的专业知识，请使用不同的协议来优化您的应用程序和最终用户体验。请注意，这种方法难度很大，只有在先用其他方法优化了应用程序后才能尝试。

提高工作负载性能的主要考虑因素是了解延迟和吞吐量要求，然后选择可优化性能的网络协议。

### 何时考虑使用 TCP

TCP 提供可靠的数据传输，并可用于工作负载组件之间的通信，在这种情况下，可靠性和有保证的数据传输很重要。许多基于 Web 的应用程序依赖基于 TCP 的协议（例如 HTTP 和 HTTPS）来打开 TCP 套接字，以便在应用程序组件之间进行通信。电子邮件和文件数据传输也是使用 TCP 的常见应用，因为它是应用程序组件之间简单可靠的传输机制。在 TCP 之上使用 TLS 会增加一些通信开销，进而会导致延迟增加和吞吐量降低，但它也有安全方面的优势。该开销主要来自握手过程（需要多次往返才能完成）的额外开销。握手完成后，加密和解密数据的开销相对较小。

### 何时考虑使用 UDP

UDP 是一种面向无连接的协议，因此适用于需要快速、高效传输的应用，例如日志、监控和 VoIP 数据。此外，如果您的工作负载组件要响应来自大量客户端的小型查询，以确保工作负载实现最佳性能，则可考虑使用 UDP。数据报传输层安全性（DTLS）是传输层安全性（TLS）的 UDP 等效项。在 UDP 上使用 DTLS 时，因为简化了握手过程，所以开销来自加密和解密数据。因为 DTLS 包括额外的字段，用于指明安全性参数和检测篡改，所以它也会给 UDP 数据包增加少量的开销。

### 何时考虑使用 SRD

可扩展的可靠数据报（SRD）是一种针对高吞吐量工作负载而优化的网络传输协议，因为它能够跨多条路径对流量进行负载均衡，并能在发生丢包或链路故障时快速恢复。因此，SRD 非常适合在计算节点之间需要高吞吐量、低延迟通信的高性能计算（HPC）工作负载。这可能包括并行处理任务（例如，涉及在节点间进行大量数据传输的模拟、建模和数据分析）。

## 实施步骤

1. 使用 [AWS Global Accelerator](#) 和 [AWS Transfer Family](#) 服务提高在线文件传输应用程序的吞吐量。AWS Global Accelerator 服务帮助您在客户端设备与 AWS 上的工作负载之间实现更低延迟。借助 AWS Transfer Family，您可以使用基于 TCP 的协议 [安全外壳文件传输协议（SFTP）和基于 SSL 的文件传输协议（FTPS）] 安全地扩展和管理到 AWS 存储服务的文件传输。
2. 使用网络延迟来确定 TCP 是否适合工作负载组件之间的通信。如果客户端应用程序和服务器之间的网络延迟很高，则 TCP 三次握手需要一些时间，因而会影响应用程序的响应能力。可以使用第一个

- 字节的时间 ( TTFB ) 和往返时间 ( RTT ) 等指标来衡量网络延迟。如果您的工作负载向用户提供动态内容, 请考虑使用 [Amazon CloudFront](#), 它将为动态内容建立到每个源的持久连接, 以便减少连接设置时间, 要不然会减慢每个客户端请求的速度。
3. 由于在 TCP 或 UDP 上使用 TLS 会影响加密和解密, 从而导致工作负载的延迟增加和吞吐量降低。对于此类工作负载, 请考虑使用 [Elastic Load Balancing](#) 上的 SSL/TLS 卸载, 允许负载均衡器处理 SSL/TLS 加密和解密过程, 而不是让后端实例来处理, 从而提高工作负载性能。这可以帮助降低后端实例上的 CPU 利用率, 进而可以提高性能并增加容量。
  4. 使用 [Network Load Balancer \( NLB \)](#) 来部署依赖 UDP 协议的服务 ( 例如, 身份验证和授权、日志记录、DNS、IoT 和串流媒体 ), 以便提高工作负载的性能和可靠性。NLB 在多个目标之间分配传入的 UDP 流量, 使您可以横向扩展工作负载、提高容量和减少单个目标的开销。
  5. 对于高性能计算 ( HPC ) 工作负载, 考虑使用 [弹性网络适配器 \( ENA \) Express](#) 功能, 它使用 SRD 协议为 EC2 实例之间的网络流量提供更高的单流带宽 ( 25Gbps ) 和更低的尾延迟 ( 99.9 百分位数 ), 从而提高网络性能。
  6. 使用 [Application Load Balancer \( ALB \)](#) 对工作负载组件之间或 gRPC 客户端与服务之间的 gRPC ( 远程过程调用 ) 流量进行路由和负载平衡。gRPC 使用基于 TCP 的 HTTP/2 协议进行传输, 并且它可带来性能优势, 例如更少的网络占用、压缩、高效的二进制序列化、支持众多语言以及双向流式传输。

## 资源

### 相关文档：

- [Amazon EBS - 优化实例](#)
- [Application Load Balancer](#)
- [Linux 上的 EC2 增强联网](#)
- [Windows 上的 EC2 增强联网](#)
- [EC2 置放群组](#)
- [在 Linux 实例上启用弹性网络适配器 \( ENA \) 增强联网](#)
- [Network Load Balancer](#)
- [AWS 联网产品](#)
- [AWS Transit Gateway](#)
- [过渡到 Amazon Route 53 中基于延迟的路由](#)
- [VPC 终端节点](#)
- [VPC 流日志](#)

相关视频：

- [连接 AWS 和混合 AWS 网络架构](#)
- [优化 Amazon EC2 实例的网络性能](#)

相关示例：

- [AWS Transit Gateway 和可扩展的安全解决方案](#)
- [AWS 联网研讨会](#)

## PERF04-BP06 根据网络要求选择工作负载的位置

评估资源置放选项，以便减少网络延迟和提高吞吐量，通过缩短页面加载和数据传输时间来提供最佳的用户体验。

常见反模式：

- 您将所有工作负载资源整合到一个地理位置中。
- 您选择的是离您的位置最近的区域，而不是离工作负载最终用户最近的区域。

建立此最佳实践的好处：用户与您的应用程序之间的延迟会极大地影响用户体验。通过使用适当的 AWS 区域和 AWS 专用全球网络，您可以减少延迟，为远程用户提供更好的体验。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

Amazon EC2 实例等资源放入 [AWS 区域](#)、[AWS Local Zones](#)、[AWS Outposts](#) 或 [AWS Wavelength](#) 区域内的可用区。选择此位置会影响给定用户位置的网络延迟和吞吐量。诸如 [Amazon CloudFront](#) 和 [AWS Global Accelerator](#) 等边缘服务也可用于在边缘站点缓存内容或为用户提供通过 AWS 全球网络到达工作负载的最佳路径，从而提高网络性能。

Amazon EC2 为联网提供置放群组。置放群组是实例的逻辑分组，可以减少延迟。使用具有支持的实例类型和弹性网络适配器 (ENA) 的置放群组，可使工作负载参与低延迟、低抖动的 25Gbps 网络。建议将置放群组用于可受益于低网络延迟和/或高网络吞吐量的工作负载。

延迟敏感型服务使用 AWS 全球网络在边缘站点交付，例如 [Amazon CloudFront](#) 访问 AWS 资源。这些边缘站点通常提供内容分发网络 (CDN) 和域名系统 (DNS) 等服务。通过在边缘交付这些服务，工作负载可以低延迟响应内容或 DNS 解析请求。这些服务还提供地理定位服务，例如内容地理定位

( 基于最终用户位置提供不同内容 ) ，或基于延迟的路由 ( 将最终用户引导至最近的区域以实现最小延迟 ) 。

可使用边缘服务来减少延迟并启用内容缓存。为 DNS 和 HTTP/HTTPS 正确配置缓存控制，以便通过这些方式获得最大优势。

实施步骤

- 捕获有关传入和传出网络接口的 IP 流量的信息。
  - [使用 VPC 流日志记录 IP 流量](#)
  - [如何将客户端 IP 地址保留在 AWS Global Accelerator 中](#)
- 分析您的工作负载中的网络访问模式，以便确定用户如何使用您的应用程序。
  - 使用监控工具，例如 [Amazon CloudWatch](#) 和 [AWS CloudTrail](#)，以便收集有关网络活动的数据。
  - 分析数据以确定网络访问模式。
- 请根据以下关键元素，为您的工作负载部署选择区域：
  - 数据所在位置：对于数据密集型应用程序（如大数据和机器学习），应用程序代码的运行应尽量接近数据。
  - 用户所在位置：对于面向用户的应用程序，选择接近您工作负载用户的一个或多个区域。
  - 其他制约：考虑成本和合规性等制约，[如 What to Consider when Selecting a Region for your Workloads 中所述。](#)
- 使用 [AWS Local Zones](#) 运行视频渲染等工作负载。Local Zones 使计算和存储资源更接近终端用户，从而使您受益。
- 为需要保留在本地的的工作负载使用 [AWS Outposts](#)，在此您希望该工作负载与 AWS 中的其他工作负载一起无缝运行。
- 高分辨率实时视频流、高保真度音频和增强现实或虚拟现实（AR/VR）等应用要求 5G 设备具有超低延迟。对于此类应用，请考虑使用 [AWS Wavelength](#)。AWS Wavelength 在 5G 网络内嵌入 AWS 计算和存储服务，为开发、部署和扩展超低延迟应用提供移动边缘计算基础设施。
- 对常用资产使用本地缓存或 [AWS 缓存解决方案](#)，以提高性能，减少数据移动并减小对环境的影响。

| 服务                                | 何时使用                                           |
|-----------------------------------|------------------------------------------------|
| <a href="#">Amazon CloudFront</a> | 用于缓存静态内容（如图像、脚本和视频）以及动态内容（如 API 响应或 Web 应用程序）。 |

| 服务                                   | 何时使用                      |
|--------------------------------------|---------------------------|
| <a href="#">Amazon ElastiCache</a>   | 用于缓存 Web 应用程序的内容。         |
| <a href="#">DynamoDB Accelerator</a> | 用于将内存中加速添加到您的 DynamoDB 表。 |

- 使用有助于您在更接近工作负载用户的位置运行代码的服务，例如：

| 服务                                          | 何时使用                                          |
|---------------------------------------------|-----------------------------------------------|
| <a href="#">Lambda@edge</a>                 | 用于执行计算密集型操作，当对象不在缓存中时启动这些操作。                  |
| <a href="#">Amazon CloudFront Functions</a> | 用于处理简单使用场景，如 HTTP(S) 请求或响应操作，这些操作可由短期运行的函数启动。 |
| <a href="#">AWS IoT Greengrass</a>          | 用于为互联设备运行本地计算、消息收发和数据缓存。                      |

- 有些应用程序需要固定入口点，或通过减少第一个字节延迟和抖动以及提高吞吐量来提高性能。在边缘站点提供静态任播 IP 地址和 TCP 终止的网络服务可以使这些应用程序受益。[AWS Global Accelerator](#) 可以将应用程序的性能提高多达 60%，并为多区域架构提供快速失效转移。AWS Global Accelerator 为您提供静态任播 IP 地址，作为一个或多个 AWS 区域中托管的应用程序的固定入口点。这些 IP 地址使流量在尽可能靠近用户的位置进入 AWS 全球网络。AWS Global Accelerator 在客户端和最靠近客户端的 AWS 边缘站点之间建立 TCP 连接，从而缩短初始连接设置时间。检查 AWS Global Accelerator 的使用情况，以提高 TCP/UDP 工作负载的性能，并为多区域架构提供快速失效转移。

## 资源

相关最佳实践：

- [COST07-BP02 根据成本考虑实施区域](#)
- [COST08-BP03 实施服务以便降低数据传输成本](#)
- [REL10-BP01 将工作负载部署到多个位置](#)
- [REL10-BP02 为您的多位置部署选择合适的位置](#)
- [SUS01-BP01 根据业务需求和可持续发展目标选择区域](#)
- [SUS02-BP04 根据其联网要求优化工作负载的地理位置](#)

- [SUS04-BP07 最大限度地减少跨网络的数据移动](#)

相关文档：

- [AWS 全球基础设施](#)
- [AWS Local Zones 和 AWS Outposts，为边缘工作负载选择适合的技术](#)
- [置放群组](#)
- [AWS Local Zones](#)
- [AWS Outposts](#)
- [AWS Wavelength](#)
- [Amazon CloudFront](#)
- [AWS Global Accelerator](#)
- [AWS Direct Connect](#)
- [AWS Site-to-Site VPN](#)
- [Amazon Route 53](#)

相关视频：

- [AWS Local Zones 解说视频](#)
- [AWS Outposts：概述及工作原理](#)
- [AWS re:Invent 2021 - AWS Outposts：将 AWS 体验带到本地](#)
- [AWS re:Invent 2020 - AWS Wavelength：在 5G 边缘以超低延迟运行应用](#)
- [AWS re:Invent 2022 - AWS Local Zones：为分布式边缘构建应用程序](#)
- [AWS re:Invent 2021 - 使用 Amazon CloudFront 构建低延迟网站](#)
- [AWS re:Invent 2022 - 使用 AWS Global Accelerator 提高性能和可用性](#)
- [AWS re:Invent 2022 - 使用 AWS 构建您的全球广域网](#)
- [AWS re:Invent 2020：使用 Amazon Route 53 进行全球流量管理](#)

相关示例：

- [AWS Global Accelerator 研讨会](#)
- [使用边缘函数处理重写和重定向](#)



## PERF04-BP07 根据指标优化网络配置

使用收集和分析的数据做出有关优化网络配置的明智决策。

常见反模式：

- 您应认为所有性能相关的问题都与应用程序有关。
- 您只需从距离已部署工作负载很近的位置测试您的网络性能。
- 为所有网络服务使用默认配置。
- 过度预置网络资源以提供充足的容量。

建立此最佳实践的好处：收集 AWS 网络的必要指标并实施网络监控工具，使您可以了解网络性能和优化网络配置。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

监控进出 VPC、子网或网络接口的流量，这对于了解如何利用 AWS 网络资源以及如何优化网络配置至关重要。通过使用以下 AWS 网络工具，您可以进一步检查有关流量使用、网络访问和日志的信息。

### 实施步骤

- 确定要收集的关键性能指标，例如延迟或丢包。AWS 提供了多种工具，可以协助您收集这些指标。通过使用以下工具，您可以进一步检查有关流量使用、网络访问和日志的信息：

| AWS 工具                                         | 使用情形                                                                                 |
|------------------------------------------------|--------------------------------------------------------------------------------------|
| <a href="#">Amazon VPC IP Address Manager。</a> | 使用 IPAM 来规划、跟踪和监控 AWS 与本地工作负载的 IP 地址。这是优化 IP 地址使用和分配的最佳实践。                           |
| <a href="#">VPC 流日志</a>                        | 使用 VPC 流日志来捕获有关进出 VPC 中网络接口的流量的详细信息。借助 VPC 流日志，您可以诊断过于严格或过于宽松的安全组规则，并确定进出网络接口的流量的方向。 |
| <a href="#">AWS Transit Gateway 流日志</a>        | 使用 AWS Transit Gateway 流日志捕获有关进出中转网关的 IP 流量的信息。                                      |

| AWS 工具                                  | 使用情形                                                                                                                                                                                                                                                    |
|-----------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">DNS 查询日志记录</a>              | 有关 Route 53 收到的公共或私有 DNS 查询的日志信息。借助 DNS 日志，您可以了解请求的域或子域，或了解响应 DNS 查询的 Route 53 边缘站点，从而优化 DNS 配置。                                                                                                                                                        |
| <a href="#">Reachability Analyzer</a>   | Reachability Analyzer 有助于您分析和调试网络可达性。Reachability Analyzer 是一个配置分析工具，使您可以在 VPC 中的源资源和目标资源之间执行连接测试。此工具帮助您确认您的网络配置与预期的连接匹配。                                                                                                                               |
| <a href="#">Network Access Analyzer</a> | Network Access Analyzer 有助于您了解对资源的网络访问权限。您可以使用 Network Access Analyzer 来指定网络访问需求，并确定不能满足指定要求的潜在网络路径。通过优化相应的网络配置，您可以了解和验证网络的状态，并证明 AWS 中的网络满足您的合规性要求。                                                                                                    |
| <a href="#">Amazon CloudWatch</a>       | 使用 <a href="#">Amazon CloudWatch</a> 并启用适当的网络选项指标。确保为工作负载选择适合的网络指标。例如，您可以为 VPC 网络地址使用、VPC NAT Gateway、AWS Transit Gateway、VPN 隧道、AWS Network Firewall、Elastic Load Balancing 和 AWS Direct Connect 启用指标。为了观察和了解网络状态和使用情况，以及便于您根据观察结果优化网络配置，持续监控指标是一种好方法。 |
| <a href="#">AWS Network Manager</a>     | 使用 AWS Network Manager，您可以出于运营和规划目的，监控 <a href="#">AWS 全球网络</a> 的实时和历史性能。Network Manager 提供 AWS 区域和可用区之间以及每个可用区内的聚合网络延迟，使您能够更好地了解应用程序性能与底层 AWS 网络性能的关系。                                                                                                 |

| AWS 工具                                | 使用情形                                                    |
|---------------------------------------|---------------------------------------------------------|
| <a href="#">Amazon CloudWatch RUM</a> | 使用 Amazon CloudWatch RUM 收集指标，以便为您提供洞察，协助您识别、理解和改善用户体验。 |

- 使用 VPC 和 AWS Transit Gateway 流日志识别主要贡献者和应用程序流量模式。
- 评估和优化您当前的网络架构，包括 VPC、子网和路由。例如，您可以评估不同的 VPC 对等互连或 AWS Transit Gateway 如何让您能够改善架构中的联网。
- 评估网络中的路由路径，以验证目的地之间是否始终使用最短路径。Network Access Analyzer 可以协助您做到这一点。

## 资源

### 相关文档：

- [VPC 流日志](#)
- [公共 DNS 查询日志记录](#)
- [什么是 IPAM？](#)
- [什么是 Reachability Analyzer？](#)
- [什么是 Network Access Analyzer？](#)
- [适合 VPC 的 CloudWatch 指标](#)
- [使用 Apache Parquet 格式的 VPC 流日志优化性能并降低网络分析成本](#)
- [Monitoring your global and core networks with Amazon CloudWatch metrics](#)
- [持续监控网络流量和资源](#)

### 相关视频：

- [使用 AWS Well-Architected Framework 进行联网的最佳实践和技巧](#)
- [监控网络流量并排查问题](#)

### 相关示例：

- [AWS 联网研讨会](#)

- [AWS 网络监控](#)

## 流程和文化

### PERF 5.您的组织实践和文化如何助力提高工作负载的性能效率？

在最初构建工作负载时，您可以采用一些原则和实践，来协助您更好地运行高效、高性能的云工作负载。要采用能提高云工作负载性能效率的文化，请考虑以下关键原则和实践：

#### 最佳实践

- [PERF05-BP01 建立关键性能指标 \( KPI \) 来衡量工作负载运行状况和性能](#)
- [PERF05-BP02 使用监控解决方案了解性能最为关键的方面](#)
- [PERF05-BP03 制定流程来提高工作负载性能](#)
- [PERF05-BP04 对工作负载进行负载测试](#)
- [PERF05-BP05 使用自动化技术主动修复与性能相关的问题](#)
- [PERF05-BP06 让工作负载和服务保持最新状态](#)
- [PERF05-BP07 定期检查指标](#)

#### PERF05-BP01 建立关键性能指标 ( KPI ) 来衡量工作负载运行状况和性能

确定用于定量和定性地衡量工作负载性能的 KPI。KPI 有助于您衡量与业务目标相关的工作负载的运行状况和性能。

#### 常见反模式：

- 您只监控系统级指标以深入了解工作负载，而不了解这些指标对业务的影响。
- 您认为 KPI 已作为标准指标数据发布和共享。
- 您没有定义可量化、可衡量的 KPI。
- KPI 与业务目标或策略不符。

建立此最佳实践的好处：确定可反映工作负载运行状况和性能的具体 KPI，有助于调整团队的工作重点，并确定成功的业务成果。与所有部门共享这些指标可让所有人了解并一致认可阈值、期望值和业务影响。

未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

利用 KPI，业务和工程团队可在衡量目标和策略以及如何将这些因素结合来取得业务成果方面达成共识。例如，网站工作负载可能会将页面加载时间用作总体性能指示。该指标将是用来衡量用户体验的多个数据点之一。除了确定页面加载时间阈值之外，您还应记录未达到理想性能要求时的预期成果或业务风险。较长的页面加载时间会直接影响最终用户的体验，降低他们的用户体验评分，并会导致客户流失。在定义 KPI 阈值时，请结合考虑行业基准和最终用户期望。例如，如果当前行业基准是两秒内加载网页，而您的最终用户希望网页在一秒内加载，那么您在建立 KPI 时应考虑这两个数据点。

您的团队必须使用实时的精细数据和历史数据作为参考来评估工作负载 KPI，并创建控制面板来对 KPI 数据执行指标计算，从而获得运维和利用率方面的洞察。应记录 KPI，包括支持业务目标和策略的阈值，并且应与所监控的指标对应起来。当业务目标、策略或最终用户需求发生变化时，应重新审视 KPI。

## 实施步骤

1. 确定并记录关键业务利益相关者。
2. 与这些利益相关者合作，确定并记录您的工作负载目标。
3. 查看行业最佳实践，确定与您的工作负载目标相一致的相关 KPI。
4. 使用行业最佳实践和工作负载目标为工作负载 KPI 设定目标。使用这些信息设置 KPI 阈值的严重性或警报级别。
5. 确定并记录未满足 KPI 时带来的风险和影响。
6. 确定并记录有助于您建立 KPI 的指标。
7. 使用监控工具，例如 [Amazon CloudWatch](#) 或 [AWS Config](#) 收集指标并衡量 KPI。
8. 使用控制面板直观显示 KPI 并与利益相关者进行沟通。
9. 定期审查和分析指标，确定需要在哪些方面改进工作负载。
10. 当业务目标或工作负载性能发生变化时，重新评估 KPI。

## 资源

相关文档：

- [CloudWatch 文档](#)
- [监控、日志记录和性能 AWS Partner](#)
- [X-Ray 文档](#)

- [使用 Amazon CloudWatch 控制面板](#)
- [QuickSight KPI](#)

相关视频：

- [AWS re:Invent 2019: Scaling up to your first 10 million users](#)
- [Cut through the chaos: Gain operational visibility and insight](#)
- [制定监控计划](#)

相关示例：

- [使用 QuickSight 创建控制面板](#)

PERF05-BP02 使用监控解决方案了解性能最为关键的方面

了解并确定在哪些方面提高工作负载性能，会对效率或客户体验产生积极的影响。例如，拥有大量客户交互的网站会因为使用边缘服务在距离客户更近的位置向客户分发内容而受益。

常见反模式：

- 您认为标准计算指标（例如，CPU 利用率或内存压力）足够捕获性能问题。
- 您只使用由自己选定的监控软件记录的默认指标。
- 您只在出现问题时审查指标。

建立此最佳实践的好处：了解关键性能领域可以帮助工作负载负责人监控 KPI 并确定具有高影响力的优先改进。

未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

设置端到端的跟踪，用于确定流量模式、延迟和关键性能领域。针对速度缓慢的查询或性能欠佳的碎片和分区数据，监控数据访问模式。使用负载测试或监控来确定受约束的工作负载领域。

通过了解架构、流量模式和数据访问模式，提高性能效率，并确定延迟和处理时间。确定随着工作负载增长可能会影响客户体验的潜在瓶颈。在研究了这些方面之后，再看看可以通过部署哪项解决方案来解决这些性能问题。

实施步骤

1. 设置端到端的监控，用于收集所有工作负载组件和指标。以下是 AWS 监控解决方案的示例。

| 服务                                                                      | 使用情形                                                       |
|-------------------------------------------------------------------------|------------------------------------------------------------|
| <a href="#">Amazon CloudWatch 真实用户监控 ( RUM , Real-User Monitoring )</a> | 收集真实用户客户端和前端会话的应用程序性能指标。                                   |
| <a href="#">AWS X-Ray</a>                                               | 通过应用程序层跟踪流量，并确定组件间的延迟以及依赖关系。使用 X-Ray 服务地图查看工作负载组件之间的关系和延迟。 |
| <a href="#">Amazon Relational Database Service Performance Insights</a> | 查看数据库性能指标并确定性能改进机会。                                        |
| <a href="#">Amazon RDS 增强监控</a>                                         | 查看数据库 OS 性能指标。                                             |
| <a href="#">Amazon DevOps Guru</a>                                      | 检测异常运营模式，以便您可以在运营问题影响客户之前发现它们。                             |

2. 执行测试以生成指标，确定流量模式、瓶颈和关键性能领域。以下是一些有关如何执行测试的示例：
- 设置 [CloudWatch Synthetic Canary](#)，使用 Linux cron 作业或 rate 表达式，通过编程方式模拟浏览器端的用户活动，从而生成一段时间内的稳定指标。
  - 使用 [AWS 分布式负载测试](#) 解决方案生成峰值流量，或者在预期增长速率下测试工作负载。
3. 评估指标和遥测数据，确定您的关键性能领域。与团队一起审查这些方面，讨论监控和解决方案以避免瓶颈。
4. 试验性能改进，并利用数据来衡量这些更改。例如，您可以使用 [CloudWatch Evidently](#) 测试新的改进以及对工作负载性能的影响。

资源

相关文档：

- [Amazon Builders' Library](#)
- [X-Ray 文档](#)
- [Amazon CloudWatch RUM](#)



- [Amazon DevOps Guru](#)

相关视频：

- [The Amazon Builders' Library: 25 years of Amazon operational excellence](#)
- [Visual Monitoring of Applications with Amazon CloudWatch Synthetics](#)

相关示例：

- [使用 Amazon CloudWatch Synthetics 测量页面加载时间](#)
- [Amazon CloudWatch RUM Web 客户端](#)
- [适用于 Node.js 的 X-Ray 开发工具包](#)
- [适用于 Python 的 X-Ray 开发工具包](#)
- [适用于 Java 的 X-Ray 开发工具包](#)
- [适用于 .Net 的 X-Ray 开发工具包](#)
- [适用于 Ruby 的 X-Ray 开发工具包](#)
- [X-Ray 进程守护程序](#)
- [AWS 上的分布式负载测试](#)

## PERF05-BP03 制定流程来提高工作负载性能

制定相应流程，以在新的服务、设计模式、资源类型和配置推出后，对它们进行评估。例如，对新实例产品运行现有性能测试，以确定它们是否有潜力改进工作负载。

常见反模式：

- 您认为当前的架构是静态的，将来不会更新。
- 您不断对架构进行更改，而不提供任何指标方面的依据。

建立此最佳实践的好处：通过制定架构更改流程，您可以使用所收集的数据来影响以后的工作负载设计。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

工作负载的性能会面临一些关键约束。记录这些约束，以便您了解哪些创新可以改进工作负载的性能。当您知道有新的服务或技术推出时，借助这些信息来确定消除约束或瓶颈的方法。

确定针对工作负载的关键性能约束。记录您的工作负载的性能约束，以便您了解哪类创新可以提高工作负载的性能。

## 实施步骤

- 确定工作负载性能 KPI ( 如 [PERF05-BP01 建立关键性能指标 \( KPI \) 来衡量工作负载运行状况和性能](#) 中所述 )，为工作负载建立基准。
- 使用 [AWS 可观测性工具](#) 收集性能指标和衡量 KPI。
- 进行深入分析，确定工作负载中性能欠佳的方面 ( 如配置和应用程序代码 )，如 [PERF05-BP02 使用监控解决方案了解性能最为关键的方面](#) 中所述。
- 使用分析和性能工具来确定性能优化策略。
- 使用沙盒环境或预生产环境来验证策略的有效性。
- 在生产环境中实施变更并持续监控工作负载的性能。
- 记录改进内容并将其传达给利益相关者。

## 资源

相关文档：

- [AWS 博客](#)
- [AWS 新增功能](#)

相关视频：

- [AWS 事件 YouTube 频道](#)
- [AWS 在线技术讲座 YouTube 频道](#)
- [Amazon Web Services YouTube 频道](#)

相关示例：

- [AWS Github](#)

- [AWS Skill Builder](#)

## PERF05-BP04 对工作负载进行负载测试

对工作负载进行负载测试，从而验证工作负载能否处理生产负载，并找出任何性能瓶颈。

常见反模式：

- 您对工作负载的各个部分进行单独负载测试，而不是测试整个工作负载。
- 您在与生产环境不同的基础设施上进行负载测试。
- 您只对预期负载，而不对其他负载进行负载测试，来预测未来可能会出现问题的方面。
- 您执行负载测试时未查阅 [Amazon EC2 测试策略](#)，也未提交模拟活动提交表。这会导致您的测试无法运行，因为它看起来像是拒绝服务事件。

建立此最佳实践的好处：通过负载测试来衡量您的性能，可向您说明随着负载的增加，您将在哪些方面受到影响。这样您便可以在更改影响您的工作负载之前，对所需进行的更改进行预测。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

云端负载测试是在预期用户负载的实际条件下衡量云工作负载性能的过程。这一过程包括：预置类似于生产的云环境，使用负载测试工具生成负载，分析各个指标来评测工作负载处理实际负载的能力。必须使用生产数据的合成或净化版本（删除敏感信息或身份识别信息）运行负载测试。作为交付管道的一部分，自动执行负载测试，并将结果与预定义的 KPI 和阈值进行比较。这一过程有利于您持续实现所需的性能。

### 实施步骤

- 根据生产环境设置测试环境。您可以使用 AWS 服务来运行生产规模的环境，进而测试架构。
- 选择和配置适合您工作负载的负载测试工具。
- 定义负载测试场景和参数（如测试持续时间和用户数量）。
- 大规模执行测试场景。利用 AWS Cloud 来测试工作负载，发现工作负载的哪些部分无法扩展或者是否以非线性方式扩展。例如，您可以使用竞价型实例以很低的成本生成负载，并在投入生产前发现瓶颈。
- 监控和记录性能指标（如吞吐量和响应时间）。Amazon CloudWatch 可以收集架构中各种资源的指标。您也可以收集和发布自定义指标，用于显示业务指标或派生指标。

- 对结果进行分析，确定性能瓶颈和需要改进的地方。
- 记录和报告负载测试过程和结果。

## 资源

### 相关文档：

- [AWS CloudFormation](#)
- [Amazon CloudWatch RUM](#)
- [Amazon CloudWatch Synthetics](#)
- [AWS 上的分布式负载测试](#)

### 相关视频：

- [Solving with AWS Solutions: Distributed Load Testing](#)
- [Optimize applications through Amazon CloudWatch RUM](#)
- [Amazon CloudWatch Synthetics 演示](#)

### 相关示例：

- [AWS 上的分布式负载测试](#)

## PERF05-BP05 使用自动化技术主动修复与性能相关的问题

使用关键性能指标 ( KPI ) 并结合监控和警报系统，主动解决与性能相关的问题。

### 常见反模式：

- 您只允许运营人员对工作负载进行运营更改。
- 您通过设置筛选条件将所有没有主动修复行为的警报发送给运营团队。

建立此最佳实践的好处：主动修复警报行为使支持人员能够集中精力处理那些无法自动完成的工作。这样一来，操作人员只需集中精力处理关键警报，从而避免因处理所有警报而变得应接不暇。

未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

使用警报触发自动操作，以便在可能的情况下修复问题。如果无法实现自动响应，则将警报上报给能够响应的人员。例如，您的系统在关键性能指标 ( KPI ) 超出特定阈值时，能够预测预期 KPI 值并发出警报；或者您的工具在 KPI 超出预期值时，能够自动停止或回滚部署。

实施相应流程，让您在工作负载运行期间了解其性能。构建监控控制面板并确定预期性能基准，以确定工作负载的性能是否达到最佳。

## 实施步骤

- 识别并了解可以自动修复的性能问题。使用 AWS 监控解决方案 ( 例如 [Amazon CloudWatch](#) 或 AWS X-Ray ) 来更好地了解问题的根本原因。
- 创建可用于自动修复问题的分步修复计划和流程。
- 将触发器配置为自动启动修复过程。例如，您可以定义一个触发器，以便在实例达到特定 CPU 利用率阈值时自动重启实例。
- 使用 AWS 服务和技术实现修复过程自动化。例如，[AWS Systems Manager Automation](#) 提供了一种安全且可扩展的方法来自动执行修复过程。
- 在预生产环境中测试自动修复流程。
- 测试完成后，在生产环境中实施修复过程，并持续进行监控，以便及时发现哪些地方需要改进。

## 资源

相关文档：

- [CloudWatch 文档](#)
- [监控、日志记录和性能 AWS Partner Network 合作伙伴](#)
- [X-Ray 文档](#)
- [在 CloudWatch 中使用警报和警报操作](#)

相关视频：

- [智能自动化云运维](#)
- [Setting up controls at scale in your AWS environment](#)
- [Automating patch management and compliance using AWS](#)
- [Amazon 如何使用更好的指标来提高网站性能](#)

相关示例：

- [CloudWatch Logs 自定义警报](#)

## PERF05-BP06 让工作负载和服务保持最新状态

随时了解新的云服务和功能，积极采用高效的功能，解决出现的问题并提高工作负载的整体性能效率。

常见反模式：

- 您认为当前的架构是静态的，将来不会更新。
- 您没有任何系统（也不定期）评估更新的软件和软件包是否与您的工作负载兼容。

建立此最佳实践的好处：通过建立流程来及时了解新服务和产品的最新情况，您可以采用新的特性和功能，解决问题并提高工作负载性能。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

随着新的服务、设计模式和产品功能的推出，评估可提高性能的方法。通过评估、内部讨论或外部分析来确定哪些方法可以提高工作负载的性能或效率。制定相应流程，评估与工作负载相关的更新、新功能和服务。例如，使用新技术构建概念验证或咨询内部团队。在尝试新想法或新服务时，运行性能测试，以衡量这些新想法或新服务对工作负载性能的影响。

### 实施步骤

- 盘点工作负载软件和架构，并确定需要更新的组件。
- 确定与您工作负载组件相关的资讯和更新来源。例如，您可以订阅 [AWS 新增功能博客](#)，了解与您的工作负载组件相匹配的产品。您可以订阅 RSS 源或管理您的 [电子邮件订阅](#)。
- 制定一个计划来评估工作负载的新服务和功能。
  - 您可以使用 [AWS Systems Manager 清单](#) 从 Amazon EC2 实例中收集操作系统（OS）、应用程序和实例元数据，并快速了解哪些实例正在运行您的软件策略所需的软件和配置，以及哪些实例需要更新。
- 了解如何更新工作负载的组件。利用云中的敏捷性，快速测试新功能如何改善工作负载，从而提高性能效率。
- 采用自动化更新流程，以减少部署新功能的工作量，并减少手动过程引起的错误。
  - 您可以使用 [CI/CD](#) 自动更新 AMI、容器映像以及其他与云应用程序相关的构件。

- 您可以使用 [AWS Systems Manager Patch Manager](#) 等工具自动执行系统更新流程，并使用 [AWS Systems Manager 维护时段](#) 安排活动。
- 记录评估更新和新服务的流程。为负责人提供所需的时间和空间来研究、测试、试验和验证更新及新服务。回顾记录的业务需求和 KPI，帮助优先确定哪些更新可以带来积极的业务影响。

## 资源

### 相关文档：

- [AWS 博客](#)
- [AWS 新增功能](#)

### 相关视频：

- [AWS 事件 YouTube 频道](#)
- [AWS 在线技术讲座 YouTube 频道](#)
- [Amazon Web Services YouTube 频道](#)

### 相关示例：

- [Well-Architected 实验室 - 清单和补丁管理](#)
- [实验室：AWS Systems Manager](#)

## PERF05-BP07 定期检查指标

作为例行维护的一部分或为了应对事件或意外事件，请检查收集到了哪些指标。通过这些检查，找出哪些指标对于解决问题至关重要，以及跟踪哪些其他指标会有助于发现、解决或预防问题。

### 常见反模式：

- 您让指标保持警报状态较长时间。
- 您创建自动化系统无法操作的警报。

建立此最佳实践的好处：不断检查收集的指标，以确认它们能够帮助正确地发现、解决问题或预防问题发生。如果您让指标保持警报状态过长时间，这些指标也会过时。

未建立这种最佳实践的情况下暴露的风险等级：中



## 实施指导

不断改进指标收集和监控。在响应意外事件或事件的过程中，评估哪些指标有助于解决问题、哪些目前没有跟踪的指标会有助于解决问题。通过这种方法，您可以提高收集的指标的质量，从而预防或更快速地解决未来发生的意外事件。

在响应意外事件或事件的过程中，评估哪些指标有助于解决问题、哪些目前没有跟踪的指标会有助于解决问题。这样，您可以提高收集的指标的质量，从而预防或更快速地解决未来发生的意外事件。

## 实施步骤

1. 根据您的工作负载目标，定义要监控的关键性能指标。
2. 为每个指标设置基准和期望值。
3. 建立定期机制（例如每周或每月）来审核关键指标。
4. 在每次审核期间，评测趋势以及与基准值的偏差。找出任何性能瓶颈或异常情况。
5. 对于已发现的问题，开展深入的根本原因分析，了解问题背后的主要原因。
6. 记录您的调查发现，使用策略来处理已发现的问题和瓶颈。
7. 持续评测和改进指标审核流程。

## 资源

### 相关文档：

- [CloudWatch 文档](#)
- [使用 CloudWatch 代理从 Amazon EC2 实例和本地服务器收集指标和日志](#)
- [监控、日志记录和性能 AWS Partner Network 合作伙伴](#)
- [X-Ray 文档](#)

### 相关视频：

- [Setting up controls at scale in your AWS environment](#)
- [Amazon 如何使用更好的指标来提高网站性能](#)

### 相关示例：

- [使用 QuickSight 创建控制面板](#)

- [第 100 级：使用 CloudWatch 控制面板进行监控](#)

## 成本优化

成本优化支柱包括以最低价格运行系统来交付商业价值的能力。如需有关具体实施的说明性指导，请参阅 [部分](#) 访问 AWS 资源。

### 最佳实践领域

- [践行云财务管理](#)
- [支出和使用情况意识](#)
- [具有成本效益的资源](#)
- [管理需求和供应资源](#)
- [持续优化](#)

## 践行云财务管理

### 问题

- [COST 1. 如何实施云财务管理？](#)

### COST 1. 如何实施云财务管理？

实施云财务管理有助于组织在 AWS 上优化成本和使用情况并进行扩展，从而实现商业价值和财务成功。

### 最佳实践

- [COST01-BP01 确立成本优化的责任归属模式](#)
- [COST01-BP02 在财务和技术人员之间建立合作关系](#)
- [COST01-BP03 建立云预算和预测](#)
- [COST01-BP04 在组织流程中落实成本意识](#)
- [COST01-BP05 报告和通知成本优化](#)
- [COST01-BP06 主动监控成本](#)
- [COST01-BP07 及时了解新发布的服务](#)
- [COST01-BP08 建立对成本敏感的文化](#)

- [COST01-BP09 量化通过成本优化实现的业务价值](#)

## COST01-BP01 确立成本优化的责任归属模式

创建一个团队（云业务办公室、云卓越中心或 FinOps 团队），负责在整个组织内建立并维护成本意识。成本优化的负责人可以是了解整个组织和云财务的个人或团队（需要来自财务、技术和业务团队的人员）。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

这是云业务办公室（CBO）或云卓越中心（CCoE）部门或团队的介绍，此部门或团队负责建立并维护一种对云计算成本敏感的文化。此部门可以是一个现有的个人、组织内的一个团队，也可以是一个由整个组织的关键财务、技术和组织利益相关者组成的新团队。

此部门（个人或团队）会排定成本管理和成本优化活动的优先级，并根据需要为这些活动投入一定比率的时间。相对于较大型企业中的全职部门，小型组织的这一部门在此方面花费的时间可能更少。

此部门（个人或团队）会排定成本管理和成本优化活动的优先级，并根据需要为这些活动投入一定比率的时间。相对于较大型企业中的全职部门，小型组织的这一部门在成本管理和优化活动方面花费的时间可能更少。

此部门需要采取多学科方法，并具备项目管理、数据科学、财务分析和软件或基础设施开发的能力。此部门可通过三种不同的责任归属模式来执行成本优化，以提高工作负载的效率：

- 集中式：通过 FinOps 团队、云财务管理（CFM）团队、云业务办公室（CBO）或云卓越中心（CCoE）等指定团队，客户可以设计和实施监管机制，并在全公司范围内推广最佳实践。
- 分散式：由具有影响力的技术团队执行成本优化。
- 混合式：同时利用集中式和分散式团队，两者可以合作执行成本优化。

可以对照成本优化目标（例如工作负载效率指标）来衡量此部门的执行和交付能力。

您必须为此部门获得高管支持，这是取得成功的一个关键因素。支持者负责倡导注重成本效益的云消费理念，为成本优化团队提供升级上报支持，确保按组织确定的优先级开展成本优化活动。否则，相关方面会忽视指导意见，并且不会优先考虑节省成本的机会。支持者与团队共同确保组织有效利用云资源并创造业务价值。

如果您制定了商业、企业入门或企业 [支持计划](#)，并需要协助来建立此团队或部门，请通过您的客户团队联系云财务管理（CFM）专家。

## 实施步骤

- **定义主要成员：** 组织的所有相关部门都应该关注成本管理，并做出自己的贡献。组织中的常见团队通常包括：财务、应用程序或产品负责人、管理和技术团队（DevOps）。有些是全职工作（财务或技术），另一些则是根据需要定期工作。执行 CFM 的个人或团队需要掌握以下技能：
  - **软件开发：** 在需要构建脚本和自动化流程时。
  - **基础设施工程：** 部署脚本，自动化流程，并了解如何预置服务或资源。
  - **运营敏锐性：** CFM 的宗旨是通过测量、监控、修改、规划和扩展对云的有效使用，在云上高效地运行。
- **定义目标和指标：** 该部门需要以不同的方式为组织创造价值。相关目标已确定，并将随着组织的发展而不断完善。常见活动包括：在整个组织内创建并执行关于成本优化的培训计划、制定组织范围标准，例如监控和报告成本优化，以及设置关于优化的工作负载目标。该部门还需要定期向组织报告其成本优化能力。

您可以定义基于价值或成本的关键绩效指标（KPI，Key Performance Indicator）。定义 KPI 时，可以根据效率和预期业务成果计算预期成本。基于价值的 KPI 将成本和使用情况指标与业务价值驱动因素联系起来，有助于合理调整 AWS 支出。获得基于价值的 KPI 的第一步是跨组织协作，选择并商定一组标准 KPI。

- **建立定期沟通机制：** 该小组（财务、技术和业务团队）应该定期开会，以审查目标和指标。典型的定期沟通机制包括审核组织的状态，审核当前执行的任何计划，审核整体财务和优化指标。随后，更详细地报告关键工作负载。

在这些定期审核中，您可以审核工作负载效率（成本）和业务成果。例如，工作负载的成本增加 20% 可能与客户使用量的增加相一致。在这种情况下，这 20% 的成本增加可以理解为一项投资。这些定期沟通通话有助于团队识别价值 KPI，为整个组织带来意义。

## 资源

相关文档：

- [AWS CCOE 博客](#)
- [创建云业务办公室](#)
- [CCOE - 云卓越中心](#)

相关视频：

- [Vanguard CCOE 成功案例](#)

相关示例：

- [使用云卓越中心 \( CCoE \) 实现整个企业转型](#)
- [构建 CCOE 以实现整个企业转型](#)
- [构建 CCOE 时应避免的 7 个陷阱](#)

COST01-BP02 在财务和技术人员之间建立合作关系

在云之旅的所有阶段，都让财务和技术团队参与成本和使用情况的讨论。团队定期开会，讨论组织目标、成本和使用情况的当前状态以及财务和会计实务等主题。

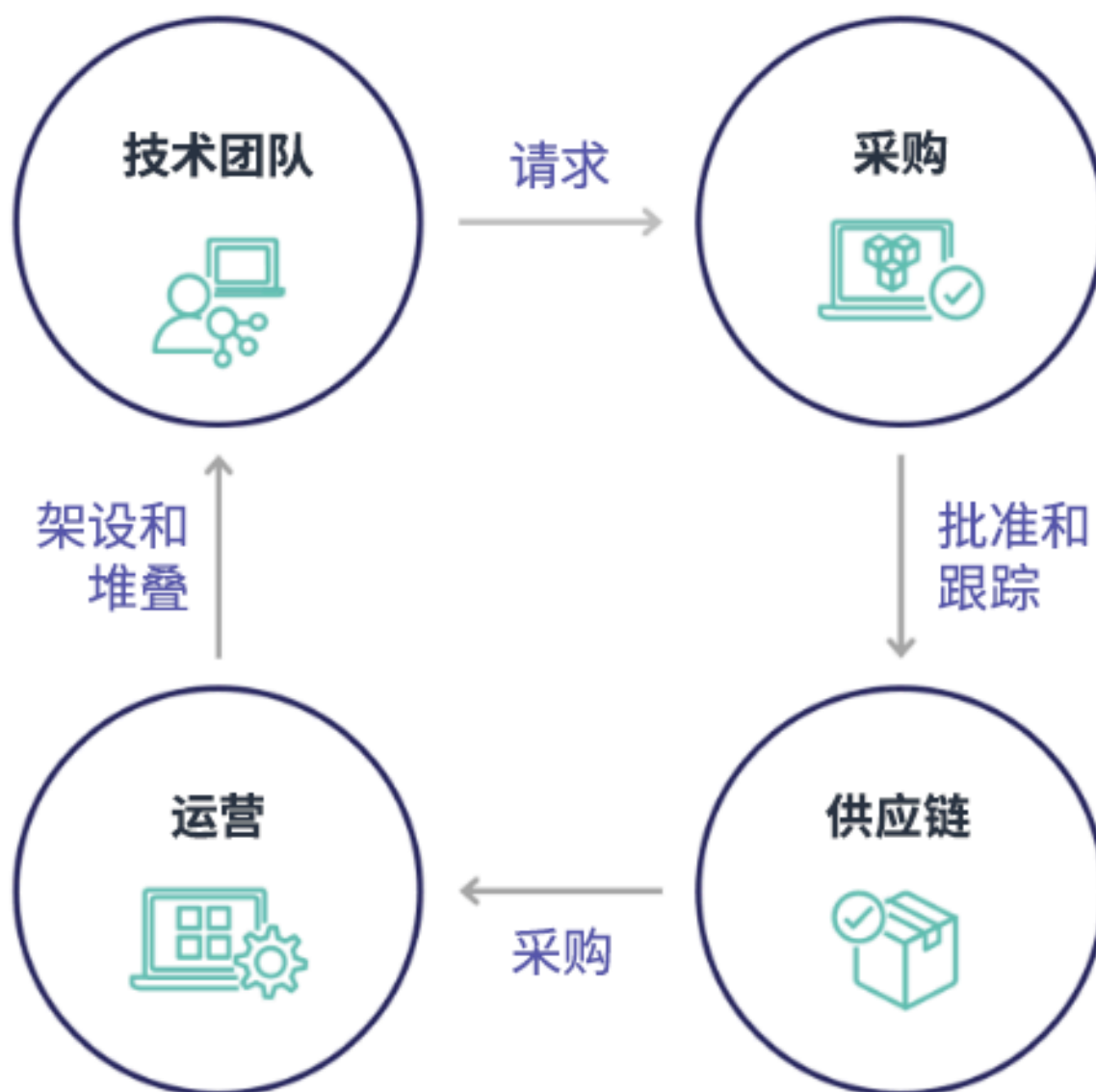
未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

由于缩短了审批、采购和基础设施部署周期，技术团队在云端的创新速度更快。这可能是对财务组织的一种调整，以前，他们习惯于执行耗时的资源密集型流程，以便在数据中心和本地环境中获取和部署资金，并且只在项目批准时进行成本分配。

从财务和采购组织的角度来看，资本预算编制、资本请求、审批、采购和安装物理基础设施这一流程是我们几十年来一直在学习和标准化的流程：

- 工程团队或 IT 团队通常是请求者
- 不同的财务团队充当审批者和采购者
- 运营团队架设、堆叠并移交可供使用的基础设施



随着云的采用，基础设施的采购和消费不再受制于一连串的依赖关系。在云模式下，技术和产品团队不再仅仅是构建者，还是产品的运营者和负责人，他们负责过去与财务和运营团队有关的大部分活动，包括采购和部署。

预置云资源真正需要的只是一个用户账户和一组正确的权限。这也有助于降低 IT 和财务风险，因为团队只需点击几下鼠标或进行几次 API 调用，就可以终止空闲或不必要的云资源。这也使技术团队能够更快创新，因为他们获得了启动实验以及之后拆除实验的敏捷性和能力。虽然从资本预算编制和预测的角度来看，云消费的可变性质可能会影响可预测性，但云为组织提供了降低过度预置成本的能力，以及降低与保守预置不足相关的机会成本的能力。



在关键的财务和技术利益相关者之间建立合作关系，让他们就组织目标达成共识，并开发在云计算的可变支出模型中获得财务成功的机制。组织内的相关团队必须在云之旅的各个阶段参与成本和使用量讨论，包括：

- **财务领导：**首席财务官、财务总监、财务规划师、业务分析师、采购、供应商开发人员和应付账款负责人必须了解消费、采购选项和每月开票流程的云模型。财务部门需要与技术团队合作创建有关 IT 价值的沟通内容并广泛传播，帮助业务团队了解技术支出与业务成果之间的联系。这样，技术支出就不会被视为成本，而会被视为投资。由于云运营（如使用量的变化速率、即付即用的定价、分级定价、定价模型以及详细的计费和使用信息）与本地运营之间存在根本差异，财务组织必须了解云的使用对业务方面的影响，包括采购流程、激励跟踪、成本分配和财务报表。
- **技术领导：**技术领导（包括产品和应用程序负责人）必须了解财务要求（如预算限制）和业务要求（如服务水平协议）。如此才能实施工作负载并实现组织的预期目标。



财务与技术人员的合作可带来以下好处：

- 财务和技术团队几乎可以实时看到成本和使用量。
- 财务和技术团队建立了标准的操作程序来处理云支出差异。
- 在如何使用资本购买承诺折扣（例如预留实例或 AWS Savings Plans）以及如何使用云来发展组织方面，财务利益相关者充当战略顾问。
- 将现有的应付账款和采购流程用于云部署。
- 财务和技术团队协作预测未来的 AWS 成本和使用量，以调整和建立组织预算。
- 通过共通的语言以及对财务概念的一致理解，更好地进行跨组织沟通。

组织中应参与成本和使用量讨论的其他利益相关者包括：

- 业务部门负责人：业务部门负责人必须了解云业务模式，从而为业务部门和整个公司提供指导。在需要预测增长和工作负载使用情况，以及评估不同购买选项（例如预留实例或 Savings Plans）时，这方面的云知识至关重要。
- 工程团队：在财务和技术团队之间建立合作关系对于建立对成本敏感的文化至关重要，这可以鼓励工程师在云财务管理（CFM）上采取行动。CFM 或财务运营从业者和财务团队的一个常见问题是让工程师了解云上的整个业务，遵循最佳实践，并采取建议的行动。
- 第三方：如果您的组织使用第三方（例如顾问或工具），请确保他们与您的财务目标一致，并且可以通过参与模式和投资回报（ROI）证明一致性。第三方通常会帮助报告和分析其管理的任何工作负载，并且提供他们设计的任何工作负载的成本分析。

要想实施 CFM 并取得成功，需要财务、技术和业务团队之间彼此协作，并在整个组织内就如何转变云支出进行沟通和评估。将工程团队包括在内，让他们在所有阶段参与这些成本和使用情况的讨论；还要鼓励他们遵循最佳实践并采取相应的商定行动。

## 实施步骤

- 定义主要成员：确认财务和技术团队的所有相关成员都参与合作。相关财务成员将是与云账单进行交互的人员。通常是首席财务官、财务总监、财务规划师、业务分析师和采购员。技术成员通常是产品和应用程序负责人、技术经理和所有在云上执行构建的团队的代表。其他成员可能包括业务部门负责人（例如影响产品使用的市场营销部门）和第三方（例如顾问），以确保与您的目标和机制保持一致，并协助进行报告。
- 定义讨论主题：定义团队之间的共同主题，或者需要达成共识的主题。从生成成本之时跟踪成本，直到账单已付为止。请注意所涉及的任何成员，以及需要应用的组织流程。了解经过的每个步骤或流程以及相关信息，例如可用的定价模式、分级定价、折扣模型、预算和财务要求。

- 建立定期沟通机制：为了让财务和技术人员展开合作，建立定期沟通机制，以促进并保持一致。该小组需要定期聚在一起，讨论目标和指标。典型的定期沟通机制包括审核组织的状态，审核当前运行的任何计划，审核整体财务和优化指标。然后，更详细地报告关键工作负载。

## 资源

### 相关文档：

- [AWS 新闻博客](#)

## COST01-BP03 建立云预算和预测

调整现有的组织预算和预测流程，使之适应云成本和使用情况的易变特性。流程必须是动态的，可以使用基于趋势或基于业务驱动因素的算法，也可以将两者结合使用。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

客户使用云来提高效率、速度和敏捷性，这导致成本和使用量的变化速度极快。随着工作负载效率的提高或者新工作负载和功能的部署，成本可以降低（有时也会提高）。工作负载可以扩展以服务更多的客户，这会增加云的使用量和成本。现在比以前更容易获得资源。云的弹性也使成本和预测变得具有弹性。必须修改现有的组织预算流程，将这种变化因素考虑在内。

预算通常针对一年进行编制，并且保持固定，要求所有相关人员严格遵守。相比之下，预测更为灵活，可以在全年重新调整，并提供一年、两年或三年的动态预测。在确定各技术和业务利益相关者的财务预期方面，预算制定和预测都起着至关重要的作用。准确的预测和实施还要求直接负责预置成本的利益相关者承担责任，也能提高他们的整体成本意识。

使用基于趋势（将历史成本用作输入）的算法或者基于驱动因素（例如新产品发布、区域扩张或用于工作负载的新环境）的算法（适用于动态和可变支出环境），或者将趋势和业务驱动因素相结合，调整现有的预算和预测流程，使其更为灵活。

您可以使用 [AWS Cost Explorer](#)，根据历史支出，对确定的未来时间范围进行基于趋势的预测。AWS Cost Explorer 的预测引擎会根据付费类型（例如，预留实例）对历史数据进行细分，并结合使用机器学习和基于规则的模型来分别预测所有付费类型的支出。

确定可能影响使用成本的业务驱动因素，并分别针对每个驱动因素进行预测，以确保提前计算出预期使用量。其中一些驱动因素与组织中的 IT 和产品团队有关。其他业务驱动因素（如营销活动、促销、合

并和收购)则是销售、营销和业务领导者所了解的,因此与他们合作,将所有这些需求驱动因素考虑在内也很重要。您需要与他们密切合作,以了解对新的内部驱动因素的影响。

使用 Cost Explorer 或任何其他工具确定了基于趋势的预测后,请使用 [AWS 定价计算器](#),根据预期使用情况(流量、每秒请求数、所需 Amazon EC2 实例)估计 AWS 使用场景和未来成本。您也可以使用此工具来协助自己计划支出方式,找到节省成本的机会,并在使用 AWS 时做出明智的决定。跟踪预测的准确性非常重要,因为预算的制定需要基于这些预测计算和估算值。

使用 [AWS Budgets](#) 设置精细的自定义预算,通过指定时间段、重复发生或金额(固定或可变),以及添加筛选条件(如服务、AWS 区域和标签)来实现。为了随时了解现有预算的执行情况,您可以创建 [AWS Budgets 报告](#) 并安排好时间表,定期以电子邮件的形式发送给您和您的利益相关者。您还可以创建 [AWS Budgets 警报](#),该警报可以根据实际成本(本质上是被动的)创建或根据预测成本(从而留出时间缓解潜在的成本超支情况)创建。您的成本或使用量超出或预计将超出预算金额时,系统会向您发送警报。

使用 [AWS Cost Anomaly Detection](#) 防止或减少意外成本,加强控制,同时不放慢创新速度。AWS Cost Anomaly Detection 利用机器学习来识别异常支出并找出根本原因,使您能够快速采取行动。[只需简单三步](#),您就可以创建自己的情境化监控器,在检测到任何异常支出时接收警报。

正如 Well-Architected 成本优化支柱的[“财务与技术人员合作”部分](#)所述,在 IT 部门、财务部门和其他利益相关者之间建立合作关系和沟通机制非常重要,以确保他们都使用相同的工具或流程来保持一致性。在预算可能需要更改的情况下,增加沟通机制接触点有助于更快地应对这些更改。

## 实施步骤

- 分析基于趋势的预测:使用偏好的基于趋势的预测工具,例如 AWS Cost Explorer 和 Amazon Forecast。从服务、账户、标签和成本类别等不同维度分析使用成本。如果需要进行高级预测,请将 AWS 成本和使用情况报告数据导入 Amazon Forecast(该工具将线性回归作为机器学习的一种形式应用于预测)。
- 分析基于驱动因素的预测:确定业务驱动因素对云使用情况的影响,并分别针对每个驱动因素进行预测,以预先计算预期的使用成本。与业务单位负责人和利益相关者密切合作,了解新驱动因素的影响,计算预期成本变化以确定准确的预算。
- 更新现有的预测和预算流程:根据所采用的预测方法(如基于趋势的预测方法、基于业务驱动因素的预测方法,或者结合使用两种预测方法),确定预测预算编制流程。预算应根据这些预测流程进行计算,并符合实际情况。
- 配置警报和通知:使用 AWS Budgets 警报和 AWS Cost Anomaly Detection 获取警报和通知。
- 与主要利益相关者定期审核:例如,与 IT、财务、平台团队和其他业务领域的利益相关者进行定期审核,以与业务方向和使用方面的变化保持一致。

## 资源

### 相关文档：

- [AWS Cost Explorer](#)
- [AWS 成本和使用情况报告](#)
- [QuickSight 预测](#)
- [Amazon Forecast](#)
- [AWS Budgets](#)
- [AWS 新闻博客](#)

### 相关视频：

- [How can I use AWS Budgets to track my spending and usage](#)
- [AWS Cost Optimization Series: AWS Budgets](#)

### 相关示例：

- [Understand and build driver-based forecasting](#)
- [How to establish and drive a forecasting culture](#)
- [How to improve your cloud cost forecasting](#)
- [Using the right tools for your cloud cost forecasting](#)

## COST01-BP04 在组织流程中落实成本意识

在影响使用量的新流程或现有流程中落实成本意识、创建成本透明度和成本问责制，并利用现有流程提高成本意识。在员工培训中贯彻成本意识。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

必须新的和现有的组织流程中建立成本意识。它是其他最佳实践的基础性先决条件之一。建议在可能的情况下重用和修改现有流程，这可以更大限度地减少对敏捷性和速度的影响。向技术团队以及业务和财务团队的决策者报告云成本，以提高成本意识，并为财务和业务利益相关者建立效率关键性能指标（KPI）。以下建议有助您在工作负载中建立成本意识：

- 验证变更管理包含成本度量，以量化变更对财务的影响。这有助于主动解决与成本相关的问题，并强调成本节省。
- 验证成本优化是您运营能力的核心组成部分。例如，您可以利用现有的意外事件管理流程来调查和确定成本及使用量异常或成本超支的根本原因。
- 通过自动化或工具加快节省成本和实现业务价值。在考虑实施成本时，请在对话中加入投资回报（ROI）信息，以证明投入时间或资金的合理性。
- 通过对云支出（包括在基于承诺的购买选项、共享服务和市场购买上的支出）实施 showback 或 chargeback 来分配云成本，以推动极具成本意识的云消费。
- 扩展现有的培训和发展计划，在整个组织开展成本意识培训。建议在其中加入持续的培训和认证。这样有助建立一个能够自我管理成本和使用量的组织。
- 利用免费的 AWS 原生工具，比如 [AWS Cost Anomaly Detection](#)，[AWS Budgets](#) 和 [AWS Budgets 报告](#)。

当组织持续采用 [云财务管理](#)（CFM）实践时，这些行为就会在工作和决策过程中落地生根。最终带来的是在从架构新的云原生应用程序的开发人员，到分析这些新的云投资的 ROI 的财务经理之间建立起更加注重成本的企业文化。

## 实施步骤

- 识别相关的组织流程：每个组织单位审核自己的流程，并确定影响成本和使用情况的流程。任何导致资源创建或终止的流程都需要进行审核。查找能够在企业中支持成本意识的流程，例如事件管理和培训。
- 建立自我维持的成本意识文化：确保所有相关的利益相关者都认同变更原因和影响是一种成本，这样他们就能理解云成本。这将使贵组织能够在创新方面建立一种自我维持的成本意识文化。
- 更新流程，增加成本意识：每个流程都进行修改，将成本意识纳入其中。流程可能需要执行额外的预检查，例如评估成本的影响，或执行后期检查，以验证成本和使用情况是否发生了预期变化。可以扩展支持流程（如培训和事件管理），以包括成本和使用情况项目。

要获得帮助，请通过您的客户团队与 CFM 专家联系，或浏览以下资源和相关文档。

## 资源

相关文档：

- [AWS 云财务管理](#)

相关示例：

- [高效云成本管理战略](#)
- [成本控制博客系列 3：如何应对成本冲击](#)
- [AWS Cost Management 初学者指南](#)

## COST01-BP05 报告和通知成本优化

设置云预算并配置检测异常使用情况的机制。配置相关工具，根据预定义目标发出成本和使用量警报，并在任何使用量超过这些目标时接收通知。定期召开会议，分析工作负载的成本效益，提高成本意识。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

必须定期报告组织内成本和使用量的优化情况。您可以安排专门的会议来讨论成本绩效，或者将成本优化纳入工作负载的常规运营报告周期。利用服务和工具定期监控成本绩效，并抓住机会实施节约成本措施。

使用 [AWS Cost Explorer](#)，按多种筛选条件和粒度查看成本和使用量，此工具提供控制面板和报告，例如按服务或按账户统计的成本、每日成本或市场成本。通过 [AWS Budgets 报告](#) 根据配置的预算，跟踪成本和使用量情况。

使用 [AWS Budgets](#) 设置自定义预算以跟踪成本和使用情况，并在您超过阈值时快速响应从电子邮件或 Amazon Simple Notification Service ( Amazon SNS ) 通知收到的警报。[将首选预算](#) 期设置为每日、每月、每季度或每年，并创建具体的预算限制，以随时了解实际或预测成本和使用量相对于预算阈值的情况。您还可以将 [警报](#) 和 [操作](#) 设置为自动运行，或在超出预算目标时通过审批流程运行。

启用关于成本和使用情况的通知，以确保在成本和使用情况发生意外变化时能够迅速采取行动。[AWS Cost Anomaly Detection](#) 使您能够减少意外成本，加强控制，同时不放慢创新速度。AWS Cost Anomaly Detection 可识别异常支出并找出根本原因，这有助于降低账单意外的风险。只需简单三步，您就可以创建自己的情境化监控器，在检测到任何异常支出时接收警报。

您也可以结合使用 [QuickSight](#) 与 AWS 成本和使用情况报告 ( CUR ) 数据，以提供包含更精细数据的高度定制的报告。利用 QuickSight，您可以安排报告，并定期收到关于历史成本和使用情况或成本节省机会的成本报告电子邮件。看看我们的 [Cost Intelligence Dashboard](#) ( CID ) 解决方案，该解决方案基于 QuickSight 构建，为您提供高级监控能力。

使用 [AWS Trusted Advisor](#)，它可提供指导，以验证预置的资源是否符合 AWS 的成本优化最佳实践。



通过可视化图表，对照精细的成本和使用量数据检查Savings Plans建议。每小时图表显示按需支出和建议的Savings Plans承诺，让用户深入了解预计节省的费用、Savings Plans覆盖范围和Savings Plans使用情况。这有助于组织了解Savings Plans如何应用于每小时的支出，而不必投入时间和资源来构建分析支出的模型。

定期创建报告，其中包含来自 AWS Cost Explorer 的Savings Plans、预留实例和 Amazon EC2 合理调整大小建议的提要，以开始降低与稳态工作负载、闲置和未充分利用的资源相关的成本。识别并收回与已部署资源的云浪费有关的支出。当创建的资源大小不正确，或者观察到不同于预期的使用模式时，就会发生云浪费。遵循 AWS 最佳实践以减少浪费，或者请您的客户团队和合作伙伴协助您 [优化并节省](#) 云成本。

定期生成报告，为您的资源提供更好的购买选项，以降低工作负载的单位成本。诸如Savings Plans、预留实例或 Amazon EC2 竞价型实例等购买选项可为具备容错能力的工作负载节省大量成本，并使利益相关者（业务负责人、财务和技术团队）能够参与有关这些承诺的讨论。

分享包含可能有助于降低云总拥有成本（TCO）的机会或新发布公告的报告。采用新的服务、区域、功能、解决方案或新的方式来进一步削减成本。

## 实施步骤

- 配置 AWS Budgets：在所有账户中为您的工作负载配置 AWS Budgets。通过使用标签设置账户总支出预算和工作负载预算。
  - [Well-Architected 实验室：成本和治理使用情况](#)
- 报告成本优化：定期讨论和分析工作负载的效率。使用已确立的指标，报告实现的指标和实现成本。找出任何负面趋势，加以解决，并确定可以在整个组织中推广的正面趋势。报告的参与者应该包括应用程序团队和负责人、财务团队以及云支出方面的关键决策者的代表。

## 资源

### 相关文档：

- [AWS Cost Explorer](#)
- [AWS Trusted Advisor](#)
- [AWS Budgets](#)
- [AWS 成本和使用情况报告](#)
- [AWS Budgets 最佳实践](#)
- [Amazon S3 分析](#)



## 相关示例：

- [Well-Architected 实验室：成本和治理使用情况](#)
- [开始优化 AWS 云成本的关键方法](#)

## COST01-BP06 主动监控成本

利用工具和控制面板主动监控工作负载的成本。定期用已配置的工具或开箱即用的工具审核成本，不要只在收到通知时才查看成本和类别。主动监控和分析成本有助于识别积极趋势，使您能够在整个组织中推广这些趋势。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

建议在组织内部主动监控成本和使用量，而不仅仅是在出现异常或意外时。在整个办公室或工作环境中，一目了然的仪表板能确保关键人员可以访问所需信息，并凸显组织对成本优化的重视程度。通过可见的控制面板，您可以积极推动成功的结果，并在整个组织中加以实施。

创建一个每日或经常性的例程，以使用 [AWS Cost Explorer](#) 或任何其他控制面板（如 [Amazon QuickSight](#)）来查看成本并主动分析。通过分组和筛选，在 AWS 账户级、工作负载级或特定 AWS 服务级分析 AWS 服务的使用情况和成本，并验证它们是否符合预期。使用小时级和资源级粒度和标签来筛选和识别主要资源所产生的成本。您还可以使用 [Cost Intelligence Dashboard](#)（一种 [Amazon QuickSight](#) 解决方案，由 AWS 解决方案架构师构建）构建自己的报告，并将预算与实际成本和使用情况进行比较。

## 实施步骤

- 报告成本优化：定期讨论和分析工作负载的效率。使用已确立的指标，报告实现的指标和实现成本。找出任何负面趋势，加以解决，并确定积极趋势，以便在整个组织中推广。报告应该包括应用程序团队和负责人、财务团队和管理团队的代表。
- 为成本和使用情况创建并启用每日粒度的 [AWS Budgets](#)，以便及时采取措施防止任何潜在的成本超支情况：您可以使用 AWS Budgets 配置警报通知，以便在任何预算类型超出预配置阈值时，都会得到通知。利用 AWS Budgets 的最佳方式是将预期成本和使用量设置为您的限值，这样一来，任何超出预算的情况均视为超支。
- 为成本监控器创建 AWS Cost Anomaly Detection：[AWS Cost Anomaly Detection](#) 使用先进的机器学习技术来识别异常支出和根本原因，以便您快速采取行动。您可以利用它来配置成本监控器，以定义您想要评估的支出部分（例如，各项 AWS 服务、成员账户、成本分配标签和成本类别），并设置

何时、何地以及如何接收警报通知。对于每个监控器，为业务负责人和技术团队附加多个警报订阅，包括每个订阅的名称、成本影响阈值和警报频率（单个警报、每日总结、每周总结）。

- 使用 AWS Cost Explorer 或将您的 AWS 成本和使用情况报告（CUR）数据与 Amazon QuickSight 控制面板集成，使贵组织的成本可视化：AWS Cost Explorer 有一个易于使用的界面，使您能够可视化、理解和管理随时间变化的 AWS 成本和使用情况。使用 [Cost Intelligence Dashboard](#)，这是一个可定制且可访问的控制面板，可帮助创建您自己的成本管理和优化工具的基础。

## 资源

### 相关文档：

- [AWS Budgets](#)
- [AWS Cost Explorer](#)
- [每日成本和使用情况预算](#)
- [AWS Cost Anomaly Detection](#)

### 相关示例：

- [Well-Architected 实验室：可视化](#)
- [Well-Architected 实验室：高级可视化](#)
- [Well-Architected 实验室：云智能控制面板](#)
- [Well-Architected 实验室：成本可视化](#)
- [AWS Cost Anomaly Detection 警报与 Slack 集成](#)

## COST01-BP07 及时了解新发布的服务

定期咨询专家或 AWS 合作伙伴，以便确定哪些服务和功能的成本更低。查看 AWS 博客和其他信息来源。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

AWS 不断地添加新的功能，以便您可以利用前沿技术来更快地进行试验和创新。您或许可以实施新的 AWS 服务和功能，以提高工作负载的成本效率。定期查看 [AWS 成本管理](#)、[AWS 新闻博客](#)、[AWS](#)

[成本管理博客](#)和 [AWS 新增功能](#) 以了解有关新服务和功能发布的信息。新增功能博客文章简要概述了 AWS 发布的所有服务、功能和区域扩展公告。

## 实施步骤

- 订阅博客：转到 AWS 博客页面，订阅新增功能博客和其他相关博客。您可以在 [通信偏好](#) 页面上用您的电子邮件地址进行注册。
- 订阅 AWS 新闻：定期查看 [AWS 新闻博客](#) 和 [AWS 新增功能](#) 以了解有关新服务和功能发布的信息。订阅 RSS 源，或通过电子邮件关注公告和发布的信息。
- 关注 AWS 降价：我们会定期对各项服务进行降价，这是 AWS 将我们的规模所带来的经济效益传递给客户的一种标准方式。截至 2022 年 4 月，AWS 自 2006 年推出以来已经降价 115 次。如果您有任何因价格问题而悬而未决的业务决策，可在降价和新服务整合后再次考虑这些决策。您可以了解以前的降价情况，包括 Amazon Elastic Compute Cloud ( Amazon EC2 ) 实例 ([在 AWS 新闻博客的降价类别中了解这些情况](#))。
- AWS 活动和交流会：参加当地的 AWS 峰会，以及与当地其他组织的任何当地交流会。如果您不能亲自参加，请尝试参加虚拟活动，从 AWS 专家和其他客户的业务案例中了解更多信息。
- 与客户团队交流：定期与您的客户团队交流，讨论行业发展趋势和 AWS 服务。与您的客户经理、解决方案架构师和支持团队沟通。

## 资源

相关文档：

- [AWS 成本管理](#)
- [AWS 新增功能](#)
- [AWS 新闻博客](#)

相关示例：

- [Amazon EC2 – 15 年来为您优化和节省 IT 成本](#)
- [AWS 新闻博客 - 降价](#)

## COST01-BP08 建立对成本敏感的文化

在整个组织中实施更改或计划，以建立对成本敏感的文化。建议先从小范围着手，然后随着能力的增强和组织对云的使用的增加，在更广泛的范围实施更大型的计划。

未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

在对成本敏感的文化中，您可以在整个组织中以有机和分散的方式执行最佳实践，从而扩展成本优化和云财务管理（财务运营、云卓越中心、云运营团队等）。与严格的自上而下的集中式方法相比，只需要很少的工作，成本意识就能让您在整个组织中培养起较高的能力水平。

建立云计算成本意识（尤其对于云计算中的主要成本驱动因素）可以让团队了解成本角度的任何更改的预期结果。访问云环境的团队应了解定价模式，以及传统的本地数据中心与云计算之间的区别。

对成本敏感的文化的主要好处是，技术团队会主动且持续地优化成本（例如，在设计新工作负载或更改现有工作负载时，优化成本会被视为一项非功能性需求），而不是根据需要被动地进行成本优化。

文化上的细微改变可对您当前和将来的工作负载的效率产生重大影响。这种情况的示例包括：

- 在工程团队中提供能见度并提高意识，以了解他们所做的事情，以及他们在成本方面的影响。
- 以游戏方式展示整个组织的成本和使用量。这可以通过一个公开可见的仪表板或比较各个团队的规范化成本和使用量的报告（例如，每个工作负载的成本和每笔交易的成本）来完成。
- 认可成本效率。公开或私下奖励自愿或主动实现的成本优化成果，并从错误中吸取教训，以免日后重蹈覆辙。
- 创建自上而下的组织要求，确保工作负载按预定义的预算运行。
- 询问变更的业务要求，以及所请求的架构基础设施或工作负载配置变更的成本影响，以确保您只需为所需的内容付费。
- 确保变更规划者了解对成本有影响的预期变更，并确保这些变更得到利益相关者的确认，以经济高效的方式实现业务成果。

## 实施步骤

- 向技术团队报告云成本：提高成本意识，为财务和业务利益相关者建立效率 KPI。
- 将计划变更告知利益相关者或团队成员：创建一个议程项目，在每周变更会议期间讨论计划变更及其对工作负载的成本效益影响。
- 与客户团队交流：与您的客户团队建立定期沟通机制，讨论行业发展趋势和 AWS 服务。与您的客户经理、架构师和支持团队沟通。
- 分享成功案例：分享关于任何工作负载、AWS 账户或组织的成本降低的成功案例，以便围绕成本优化树立积极的态度并给予鼓励。
- 培训：确保技术团队或团队成员接受了关于 AWS Cloud 资源成本意识方面的培训。

- AWS 活动和交流会：参加当地的 AWS 峰会，以及与当地其他组织的任何当地交流会。
- 订阅博客：转到 AWS 博客页面，订阅 [新增功能博客](#) 和其他相关博客，以关注 AWS 分享的新发布、实施、示例和更改。

## 资源

### 相关文档：

- [AWS 博客](#)
- [AWS 成本管理](#)
- [AWS 新闻博客](#)

### 相关示例：

- [AWS 云财务管理](#)
- [AWS Well-Architected 实验室：云财务管理](#)

## COST01-BP09 量化通过成本优化实现的业务价值

通过量化成本优化带来的业务价值，您可以了解组织取得的全部效益。由于成本优化是一项必要的投资，因此量化业务价值之后，您就可以向利益相关者说明投资回报。如果能够量化业务价值，在未来的成本优化投资中，就可以从利益相关者那里得到更多支持，并获得一个框架来衡量组织成本优化活动的成果。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

量化业务价值是指衡量企业从他们所采取的行动和作出的决策中获得的收益。业务价值可以有形的（例如减少支出或增加利润），也可以是无形的（例如改善品牌声誉或提高客户满意度）。

量化成本优化带来的业务价值，意味着要确定您从提高支出效率的努力中获得了多少价值或收益。例如，如果一家公司花费 10 万 USD 在 AWS 上部署工作负载，然后对其进行优化，则在不牺牲质量或产出的情况下，新成本仅为 8 万 USD。在这种情况下，成本优化带来的量化业务价值是节省 2 万 USD。但是，除了节约成本外，企业还可以从缩短交付时间、提高客户满意度或成本优化工作带来的其他指标等方面量化价值。利益相关者需要就成本优化的潜在价值、优化工作负载的成本和回报价值作出决定。

除了报告通过成本优化节省的费用外，建议量化其创造的附加价值。成本优化的效益通常以每项业务成果降低的成本来量化。例如，当您购买 Savings Plans 时，可以量化 Amazon Elastic Compute Cloud ( Amazon EC2 ) 节省的成本，这可以降低成本并维持工作负载输出水平。当闲置的 Amazon EC2 实例删除，或者未连接的 Amazon Elastic Block Store ( Amazon EBS ) 卷删除时，可以量化削减的 AWS 支出成本。

然而，成本优化带来的好处绝不仅仅在于降低或规避成本。考虑捕获其他数据来衡量效率提升值和业务价值。

## 实施步骤

- 评估业务效益：这是分析和调整 AWS Cloud 成本的过程，使每一美元的花费都能产生最大的效益。与其专注于降低成本而忽略业务价值，不如考虑业务效益和成本优化的投资回报，这样可能会让您所花的钱产生更大价值。这就是要明智地花钱，在能产生最大回报的领域进行投资和支出。
- 分析预测 AWS 成本：预测使得财务利益相关者可以与其他内部和外部组织的利益相关者一同设定期望，并有助于提高组织的财务可预测性。[AWS Cost Explorer](#) 可用于预测成本和使用量。

## 资源

相关文档：

- [AWS Cloud 经济性](#)
- [AWS 博客](#)
- [AWS 成本管理](#)
- [AWS News Blog](#)
- [Well-Architected 可靠性支柱白皮书](#)
- [AWS Cost Explorer](#)

相关视频：

- [在 AWS 上使用 Windows 挖掘业务价值](#)

相关示例：

- [客户 360 的商业价值衡量及最大化](#)
- [采用 Amazon Web Services 托管数据库的业务价值](#)

- [Amazon Web Services 为独立软件供应商带来的业务价值](#)
- [云现代化的业务价值](#)
- [迁移到 Amazon Web Services 的业务价值](#)

## 支出和使用情况意识

### 问题

- [COST 2.您如何管理使用情况？](#)
- [COST 3.您如何监控成本和使用情况？](#)
- [COST 4.您如何停用资源？](#)

### COST 2.您如何管理使用情况？

制定各种策略和机制，确保花费适当的成本来达到目标。采用制约与平衡方法，您可以在不超支的情况下进行创新。

### 最佳实践

- [COST02-BP01 根据组织的要求制定各种策略](#)
- [COST02-BP02 实施方向性目标和执行性目标](#)
- [COST02-BP03 实施账户结构](#)
- [COST02-BP04 实施组和角色](#)
- [COST02-BP05 实施成本控制](#)
- [COST02-BP06 跟踪项目生命周期](#)

### COST02-BP01 根据组织的要求制定各种策略

制定策略，确定贵组织应该如何管理资源，并定期执行检查。策略应该涵盖资源和工作负载的成本，包括在资源生命周期内的创建、修改和停用。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

了解组织的成本和驱动因素，对于有效管理成本和使用量以及找到降低成本的机会至关重要。在组织中，通常会有多个团队运行多个工作负载。这些团队可能在不同的部门，每个部门都有其自己的收入来源。将资源成本分摊到工作负载、各个组织或产品负责人，这样既推动更高效的资源使用模式，又能



减少浪费。准确的成本和使用量监控有助于您了解如何优化工作负载，以及各部门和产品的盈利能力。利用这些知识，您可以就在组织内的何处分配资源做出更明智的决策。组织中各层级的人员都了解使用量是推动变化的关键，因为使用量变化会导致成本变化。考虑采用多元方法来了解您的使用量和支出情况。

执行治理的第一步是按照组织要求来针对云的使用制定策略。这些策略定义组织如何使用云以及如何管理资源。策略应涵盖与成本或使用量有关的资源和工作负载的所有方面，包括资源生命周期内的创建、修改和停用。确认云环境中的任何更改都是遵循相应策略和程序实施的。在 IT 变更管理会议期间，提出问题以了解计划变更的成本影响（无论是增加还是减少）、业务理由以及预期结果。

策略应该简单易懂，能够在整个组织中有效实施。策略还需要易于遵守和解释（以便落实），并且要具体（不会在各团队之间造成误解）。此外，需要定期审查这些策略（例如我们的机制），并在客户业务状况或业务优先级发生变化时及时更新策略，以免导致策略过时。

从广泛的、高层级的策略开始，例如使用哪个地理区域，或者一天中应该运行资源的时间。逐步为各组织部门和工作负载细化策略。常见策略包括可以使用哪些服务和功能（例如，测试和开发环境中性能较低的存储），哪些类型的资源可供不同团队使用（例如，开发账户中最大规模的资源为中等大小），以及这些资源将使用多长时间（可能是临时使用、短期使用或在特定时间段内使用）。

### 策略示例

以下是侧重于成本优化的策略示例，您可以参考该策略来创建自己的云治理策略。请确保根据组织和利益相关者的要求调整策略。

- **策略名称：** 定义清晰的策略名称，例如“资源优化和成本削减策略”。
- **目的：** 解释为什么应使用此策略以及预期效果如何。此策略的目标是确认，在为了满足业务要求而部署和运行所需工作负载方面，有最低的成本要求。
- **范围：** 明确定义谁应使用此策略以及在什么情况下使用此策略，例如 DevOps X 团队，为美国东部的客户在 X 环境（生产或非生产）中使用此策略。

### 策略声明

1. 根据工作负载的环境和业务需求（开发、用户验收测试、预生产或生产），选择美国东部区域 1 或多个美国东部区域。
2. 安排 Amazon EC2 和 Amazon RDS 实例在早上 6 点到晚上 8 点之间运行 [美国东部标准时间 (EST)]。
3. 在处于不活动状态 8 小时之后，停止所有未使用的 Amazon EC2 实例，在处于不活动状态 24 小时之后，停止未使用的 Amazon RDS 实例。

4. 在非生产环境中处于不活动状态 24 小时之后，终止所有未使用的 Amazon EC2 实例。提醒 Amazon EC2 实例拥有者（根据标签）查看其生产环境中已停止的 Amazon EC2 实例，并告知他们，如果其 Amazon EC2 实例在 72 小时内未使用，将会被终止。
5. 使用通用实例系列和大小（如 m5.large），然后使用 AWS Compute Optimizer，根据 CPU 和内存利用率调整实例大小。
6. 优先使用自动扩缩，根据流量动态调整运行的实例数量。
7. 为非关键工作负载使用竞价型实例。
8. 查看容量需求，为可预测的工作负载使用实惠配套或预留实例，并通知云财务管理团队。
9. 使用 Amazon S3 生命周期策略，将不经常访问的数据移动到更便宜的存储层。如果未定义保留策略，请使用 Amazon S3 Intelligent Tiering，自动将对象移动到存档层。
10. 使用 Amazon CloudWatch 监控资源利用率并设置警报来触发扩展事件。
11. 对于每个 AWS 账户，使用 AWS Budgets，根据成本中心和业务单位为您的账户设置成本和使用量预算。
12. 使用 AWS Budgets 为账户设置成本和使用量预算，有助于您控制支出和避免意外账单，从而更好地控制成本。

过程：提供实施本策略的详细过程，或参考介绍如何实施各个策略声明的其他文档。此部分应提供了关于如何执行策略要求的分步说明。

要实施此策略，您可以使用各种第三方工具或 AWS Config 规则来检查是否符合策略声明，并使用 AWS Lambda 函数触发自动修复操作。您也可以使用 AWS Organizations 来强制执行策略。此外，您应定期查看资源使用情况，并在必要时调整策略，以确保策略继续满足您的业务需求。

## 实施步骤

- 与利益相关者会面：要制定策略，让组织内的利益相关者（云业务办公室、工程师或实施策略的职能部门决策者）详细说明他们的要求并记录下来。采用迭代方法，首先大致进行，然后在每一步中不断细化到最小单元。团队成员包括与工作负载切身相关的人员（例如组织单位或应用程序负责人）以及支持小组（例如安全和财务团队）。
- 完成确认：确保团队就哪些人可以访问和部署到 AWS Cloud 的策略达成一致。确保这些人遵守组织的策略，并确认其资源创建符合商定的策略和程序。
- 创建入职培训课程：要求新组织成员完成入职培训课程，以建立成本意识并了解组织要求。他们可能会采取不同于以往经验的策略，或者根本不考虑这些策略。
- 定义工作负载的位置：定义工作负载的运行位置，包括国家/地区以及国家/地区中的区域。此信息用于映射到 AWS 区域和可用区。

- **定义和分组服务和资源：** 定义工作负载所需的服务。对于每项服务，指定类型、大小和所需资源数量。按职能定义资源组，如应用程序服务器或数据库存储。资源可属于多个组。
- **按职能定义和分组用户：** 定义与工作负载交互的用户，侧重于用户的工作范畴及其使用工作负载的方式，而不是侧重于他们的身份或其在组织中的职位。将类似用户或职能分组在一起。您可以使用 AWS 托管策略作为指南。
- **定义操作：** 使用前面确定的位置、资源和用户，定义每项在其生命周期（开发、运行和停用）内实现工作负载成果所需的操作。根据每个位置的组（而不是组中的个别元素）确定操作。首先广泛读写，然后细化到每项服务的具体操作。
- **定义审核期：** 工作负载和组织要求可能会随时间而变化。定义工作负载审核计划，以确保其与组织重点保持一致。
- **将策略编制成档：** 验证已定义的策略是否可按组织的要求访问。这些策略用于实施、维护和审计对环境的访问。

## 资源

### 相关文档：

- [云中的变更管理](#)
- [针对工作职能的 AWS 托管策略](#)
- [AWS 多账户计费策略](#)
- [AWS 服务的操作、资源和条件键](#)
- [AWS 管理和治理](#)
- [使用 IAM 策略控制对 AWS 区域的访问](#)
- [全球基础设施区域和可用区](#)

### 相关视频：

- [大规模的 AWS 管理和治理](#)

### 相关示例：

- [VMware – 什么是云策略？](#)

## COST02-BP02 实施方向性目标和执行性目标

实施工作负载的成本和使用量方向性目标和执行性目标。方向性目标为组织在预期结果方面指明了方向，而执行性目标则提供了要为工作负载实现的具体可衡量结果。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

为组织制定成本和使用量的方向性目标及执行性目标。作为一家在 AWS 上不断发展壮大的组织，针对成本优化设定方向性目标并进行跟踪，这一点非常重要。这些方向性目标，或者说是 [关键绩效指标 \(KPI\)](#)，可以是按需支出百分比等内容，也可以是采用某些优化服务（比如 AWS Graviton 实例或 gp3 EBS 卷类型）。设定可衡量且可实现的方向性目标有助于您持续衡量效率的提高情况，这对于日常业务运营至关重要。方向性目标为组织提供有关预期结果的指引和方向。执行性目标则提供要实现的具体可衡量的结果。简言之，方向性目标是指您前进的方向，而执行性目标就是在这个方向上的进展情况，以及何时实现这个目标（使用具体、可衡量、可分配、切合实际、适时的指导原则，即 SMART 指导原则）。方向性目标的一个示例是：在略微（非线性）增加成本的情况下，显著提升平台使用量。执行性目标的一个示例是：在成本增长不到 5% 的情况下，将平台使用量提升 20%。另一个常见的方向性目标是每 6 个月提高一次工作负载的效率。相应的执行性目标则是，需要每 6 个月将各项业务指标的成本缩减 5%。

就成本优化而言，其方向性目标是提高工作负载效率，即逐步降低每项业务成果的工作负载成本。建议为所有工作负载实施此方向性目标，同时设定执行性目标，例如每 6 至 12 个月将效率提高 5%。通过在云端构建成本优化能力，以及发布新服务和功能，可以实现这一目标。

您需要能够近乎实时地监控 KPI 及相关的成本节省机会，并不断跟踪进度，这非常重要。要着手定义和跟踪 KPI 方向性目标，我们建议使用 [Cloud Intelligence 控制面板 \(CID\) 框架中提供的 KPI 控制面板](#)。根据来自 AWS 成本和使用情况报告的数据，KPI 控制面板提供了一系列建议的成本优化 KPI，能够设定自定义的方向性目标并不断跟踪进度。

如果您通过其他解决方案来设定和跟踪 KPI 目标，请务必让组织中的所有云财务管理利益相关者都采用该解决方案。

### 实施步骤

- 定义预期使用量水平：首先，要重点关注使用量水平。与应用程序负责人、市场营销团队和更大的业务团队交流，了解工作负载的预期使用量水平。客户需求如何随时间而变化？是否会因季节性增长或营销活动而发生变化？

- **定义工作负载资源和成本：** 定义使用量水平后，量化满足这些使用量水平所需的工作负载资源变化。您可能需要增加工作负载组件的资源大小或数量，增加数据传输，或者在特定级别将工作负载组件更改为不同的服务。详细说明每项要点的成本，以及当使用量发生变化时成本的变化。
- **定义业务目标：** 从预期的使用量和成本变化中获取输出，将其与预期的技术变化或正在运行的任何计划相结合，制定工作负载目标。目标必须阐明使用量、成本以及两者之间的关系。目标必须简单、具有较高的水平级别，并能让人们了解企业在结果方面的期望（例如确保未使用的资源保持在一定的成本水平以下）。您不需要为每种未使用的资源类型定义方向性目标，也不需要定义会导致方向性目标和执行性目标损失的成本。确认制定有组织的计划（例如培训和教育等能力培养项目），以防成本呈预期变化，而使用量无变化。
- **定义执行性目标：** 对于定义的每个方向性目标，指定一个可衡量的执行性目标。如果方向性目标是提高工作负载的效率，则执行性目标将量化改进量（通常为每一美元支出的业务产出）及获益时间。例如，如果您设定了一个方向性目标，即最大限度地减少由于过度资源调配而造成的浪费，那么您的执行性目标可能是：第一层生产工作负载中由于计算过度资源调配造成的浪费不应超过层计算成本的 10%；第二层生产工作负载中由于计算过度资源调配导致的浪费不应该超过层计算成本的 5%。

## 资源

### 相关文档：

- [针对工作职能的 AWS 托管策略](#)
- [针对 AWS Control Tower 登录区的 AWS 多账户策略](#)
- [使用 IAM 策略控制对 AWS 区域的访问](#)
- [SMART 方向性目标](#)

### 相关视频：

- [Well-Architected 实验室：方向性目标和执行性目标（第 100 级）](#)

### 相关示例：

- [Well-Architected 实验室：停用资源（方向性目标和执行性目标）](#)
- [Well-Architected 实验室：资源类型、大小和数量（方向性目标和执行性目标）](#)

## COST02-BP03 实施账户结构

实施与您的组织对应的账户结构。这有助于在整个组织内分摊和管理成本。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

AWS Organizations 允许您创建多个 AWS 账户，这有助于您在扩展 AWS 上的工作负载时集中管理环境。您可以通过在组织单位 ( OU ) 结构中分组 AWS 账户，并在每个 OU 下创建多个 AWS 账户来对组织层次结构进行建模。要创建账户结构，您首先需要确定哪个 AWS 账户将成为管理账户。之后，您可以按照[管理账户最佳实践](#)和[成员账户最佳实践](#)，根据您的账户结构创建新 AWS 账户或选择现有账户作为成员账户。

建议无论组织规模或使用情况如何，始终至少有一个管理账户和一个与之链接的成员账户。所有工作负载资源应仅驻留在成员账户中，不应在管理账户中创建任何资源。对于您应该拥有多少 AWS 账户这一问题，没有标准答案。评估当前和未来的运营和成本模型，以确保您的 AWS 账户结构反映了组织的目标。有些公司出于业务原因会创建多个 AWS 账户，例如：

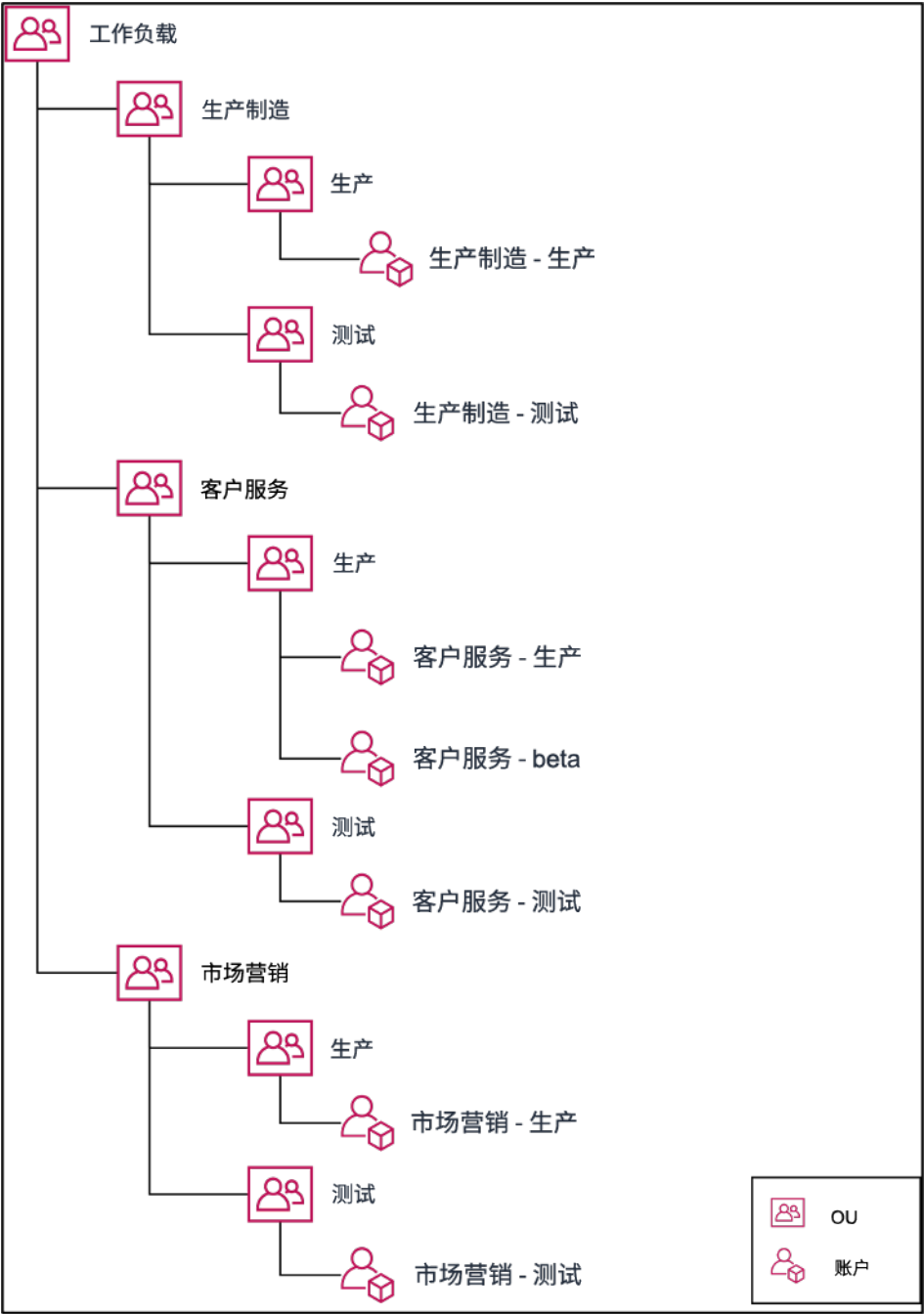
- 需要在组织部门、成本中心或特定工作负载之间实施管理或财务和计费隔离。
- AWS 服务限制设置为特定于特殊工作负载。
- 必须对工作负载和资源进行隔离和分离。

在 [AWS Organizations](#) 内，[整合账单](#) 在一个或多个成员账户和管理账户之间创建构造。通过成员账户，您可以按团队隔离和区分成本和使用量。常见做法是每个组织部门 ( 如财务、营销和销售 )、每个环境生命周期 ( 如开发、测试和生产 ) 或每个工作负载 ( 工作负载 a、b 和 c ) 具有单独的成员账户，然后使用整合账单将这些关联账户汇总在一起。

通过整合账单，您可以将多个成员 AWS 账户的付款整合至一个管理账户下，同时仍可查看每个关联账户的活动。由于成本和使用量在管理账户中汇总，因此，您可以最大限度地提高服务量折扣，并最大限度地利用承诺折扣 ( Savings Plans 和预留实例 ) 来获得最高折扣。

下图显示了如何对组织单位 ( OU ) 使用 AWS Organizations，以对多个账户进行分组，并在每个 OU 下放置多个 AWS 账户。建议将 OU 用于各种应用场景和工作负载，这提供了组织账户的模式。





对组织部门下的多个 AWS 账户进行分组的示例。

[AWS Control Tower](#) 可以快速设置和配置多个 AWS 账户，确保治理符合您组织的要求。

实施步骤

- 定义分离要求：分离要求涉及多项因素，包括安全性、可靠性和财务结构。按顺序阐明每项因素，并详细说明工作负载或工作负载环境是否应与其他工作负载分开。安全性有助于满足访问和数据要求。可靠性管理限制，以便环境和工作负载不会影响其他环境和工作负载。定期查看架构完善的框架的安



全性和可靠性支柱，并遵循提供的最佳实践。财务结构创建严格的财务分离（不同的成本中心、工作负载所有权和问责制）。常见的分离示例包括在单独的账户中运行生产和测试工作负载，或使用单独的账户，以便可以将发票和计费数据提供给组织中的单个业务单位或部门，或拥有该账户的利益相关者。

- 定义分组要求：分组要求并不覆盖分离要求，而是用于协助管理。将无需分离的类似环境或工作负载分组在一起。例如，将一个或多个工作负载的多个测试或开发环境分组在一起。
- 定义账户结构：使用这些分离和分组，为每个组指定一个账户，并确保持续满足分离要求。这些账户有成员账户或关联账户。通过将这些成员账户分组到一个管理账户或付款人账户下，可以合并使用量，从而可以跨所有账户享有更大的批量折扣，这为所有账户提供一个账单。可以分离账单数据，并为每个成员账户提供其账单数据的单独视图。如果成员账户不能让任何其他账户看到自己的使用情况或账单数据，或者，如果需要 AWS 提供单独的账单，请定义多个管理账户或付款人账户。在这种情况下，每个成员账户都有自己的管理账户或付款人账户。资源应始终放置在成员账户或关联账户中。管理账户或付款人账户应只用于管理。

## 资源

### 相关文档：

- [使用成本分配标签](#)
- [针对工作职能的 AWS 托管策略](#)
- [AWS 多账户计费策略](#)
- [使用 IAM 策略控制对 AWS 区域的访问](#)
- [AWS Control Tower](#)
- [AWS Organizations](#)
- [管理账户和成员账户的最佳实践](#)
- [使用多个账户组织 AWS 环境](#)
- [启用共享预留实例和 Savings Plans 折扣](#)
- [整合账单](#)
- [整合账单](#)

### 相关示例：

- [拆分 CUR 和共享访问](#)

相关视频：

- [AWS Organizations 简介](#)
- [设置使用 AWS Organizations 最佳实践的多账户 AWS 环境](#)

相关示例：

- [架构完善的实验室：创建 AWS 组织（级别 100）](#)
- [拆分 AWS 成本和使用情况报告 和共享访问](#)
- [为电信公司定义 AWS 多账户策略](#)
- [优化 AWS 账户的最佳实践](#)
- [使用 AWS Organizations 管理组织部门的最佳实践](#)

## COST02-BP04 实施组和角色

实施与策略一致的组和角色，控制每个组中谁可以创建、修改或停用实例和资源。例如，实施开发组、测试组和生产组。这适用于 AWS 服务和第三方解决方案。

在未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

用户角色和组是设计和实施安全、高效的系统的基本构件。角色和组有助于组织在控制力需求与灵活性和生产率要求之间取得平衡，最终满足组织目标和用户需求。根据 AWS Well-Architected Framework 安全性支柱的[身份和访问权限管理](#)部分的建议，您需要制定强健的身份管理和权限策略，以便在适当的条件下为合适的人员提供对正确资源的访问权限。用户仅获得完成任务所需的访问权限。这可最大限度地降低与未经授权访问或滥用相关的风险。

制定策略后，可以在组织内创建逻辑组和用户角色。这使您可以分配权限、控制使用情况，并有助于实施强健的访问控制机制，防止未经授权访问敏感信息。从大致的人员分组开始，这通常与组织部门和岗位角色（例如，IT 部门的系统管理员、财务主管或业务分析师）对应。这些组将执行相似任务并需要相似访问权限的人员划分在一起。角色定义组必须做什么。管理群和角色的权限比管理单个用户的权限更容易。角色和组可一致且系统性地为所有用户分配权限，防止出现错误和不一致。

当用户的角色发生变化时，管理员可以在角色或组级别上调整访问权限，而不是重新配置单个用户账户。例如，IT 部门的系统管理员需要创建所有资源的权限，而分析团队成员仅需要创建分析资源。

## 实施步骤

- 实施组：如有必要，请使用组织策略中定义的用户组实施相应的组。有关用户、组和身份验证的最佳实践，请参阅 AWS Well-Architected Framework 的[安全性支柱](#)。
- 实施角色和策略：使用组织策略中定义的操作，创建所需的角色和访问策略。有关角色和策略的最佳实践，请参阅 AWS Well-Architected Framework 的[安全性支柱](#)。

## 资源

### 相关文档：

- [针对工作职能的 AWS 托管策略](#)
- [AWS 多账户计费策略](#)
- [AWS Well-Architected Framework 的安全性支柱](#)
- [AWS Identity and Access Management \(IAM\)](#)
- [AWS Identity and Access Management 策略](#)

### 相关视频：

- [为什么要使用身份和访问权限管理](#)

### 相关示例：

- [Well-Architected 实验室：基本身份和访问权限](#)
- [使用 IAM 策略控制对 AWS 区域的访问](#)
- [开始您的云财务管理之旅：云成本运营](#)

## COST02-BP05 实施成本控制

根据组织策略以及定义的组和角色来实施控制。这样可以确保成本只根据组织要求的规定产生，例如，控制用户对区域或资源类型的访问。

在未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

实施成本控制的第一步通常是进行相关设置，以便在发生成本或使用量超出策略的事件时触发通知。您可以迅速采取行动，并验证是否需要采取纠正措施，而不会限制工作负载或新活动，抑或是对它们产生负面影响。了解工作负载和环境限制后，可以强制实施治理。[AWS Budgets](#) 允许您为 AWS 成本、使用量和承诺折扣（Savings Plans 和预留实例）设置通知并定义每月预算。可以在总成本级别（如所有成本）创建预算，也可以在更细粒度的级别创建预算，其中只包含特定的维度，如关联的账户、服务、标记或可用区。

使用 AWS Budgets 设置预算限制后，可以使用 [AWS Cost Anomaly Detection](#) 来减少意外成本。AWS Cost Anomaly Detection 是一种成本管理服务，它使用机器学习持续监控成本和使用情况，以检测异常支出。它可以帮助您识别异常支出和根本原因，以便您可以快速采取行动。首先，在 AWS Cost Anomaly Detection 中创建成本监视器，然后通过设置美元阈值来选择提醒首选项（例如，对影响大于 1,000 美元的异常情况发出提醒）。收到提醒后，可以分析造成异常情况的根本原因，以及对成本的影响。您还可以在 AWS Cost Explorer 中监控并执行自己的异常分析。

在 AWS 中通过 [AWS Identity and Access Management](#) 和 [AWS Organizations 服务控制策略（SCP）](#) 强制实施治理策略。IAM 允许您安全地管理对 AWS 服务和资源的访问。您可以使用 IAM 控制谁能创建和管理 AWS 资源、可创建的资源类型，以及可在何处创建。这最大限度地降低了在定义的策略之外创建资源的可能性。使用先前创建的角色和组，并分配 [IAM 策略](#)，以强制实施正确的使用量。SCP 用于集中管控组织中所有账户的最大可用权限，从而确保您的账户始终在访问控制准则允许的范围内。SCP 仅在启用了所有功能的组织中可用，并且您可以将 SCP 配置为默认情况下拒绝或允许对成员账户执行操作。有关实施访问管理的更多详细信息，请参阅 [《架构完善的安全性支柱》白皮书](#)。

还可以通过管理 [AWS 服务限额](#) 来实施治理。通过确保为服务限额设置最低开销并进行准确维护，您可以最大限度地减少组织要求以外的资源创建。要实现这一点，您必须了解要求的改变速度，了解正在进行的项目（资源的创建和停用），以及影响可以实施的配额更改速度的因素。必要时，可以使用 [服务限额](#) 来增加配额。

## 实施步骤

- 实施支出通知：使用定义的组织策略创建 [AWS Budgets](#) 以在支出超出策略时通知您。配置多个成本预算，每个账户一个，以通知您账户的总支出情况。在每个账户中为该账户内的较小单元配置额外的成本预算。这些单元因您的账户结构而异。一些常见示例包括 AWS 区域、工作负载（使用标记）或 AWS 服务。将电子邮件通讯组列表配置为通知的收件人，而不是个人的电子邮件账户。可以为超出金额的情况配置实际预算，或者使用预测预算通知预测使用量。还可以预配置 AWS 预算操作，以强制实施特定 IAM 或 SCP 策略，或停止目标 Amazon EC2 或 Amazon RDS 实例。可以自动执行预算操作，也可以要求工作流程审批。

- 实施异常支出通知：可以使用 [AWS Cost Anomaly Detection](#) 降低组织中的意外成本，并分析潜在异常支出的根本原因。创建成本监视器以指定粒度识别异常支出，并在 AWS Cost Anomaly Detection 中配置通知后，它会在检测到异常支出时向您发送提醒。这将使您能够分析导致异常的根本原因，并了解对成本的影响。配置 AWS Cost Anomaly Detection 时使用 AWS 成本类别来确定哪个项目团队或业务部门团队可以分析导致意外成本的根本原因，并及时采取必要的措施。
- 实施使用量控制：使用定义的组织策略，实施 IAM 策略和角色，以指定用户可执行的操作和不能执行的操作。一个 AWS 策略中可能包含多个组织策略。采用定义策略时所用的方式，首先大致进行，然后在每一步施加更细粒度的控制。服务限制也是一种有效的使用量控制措施。对所有账户实施正确的服务限制。

## 资源

### 相关文档：

- [针对工作职能的 AWS 托管策略](#)
- [AWS 多账户计费策略](#)
- [使用 IAM 策略控制对 AWS 区域的访问](#)
- [AWS Budgets](#)
- [AWS Cost Anomaly Detection](#)
- [控制 AWS 成本](#)

### 相关视频：

- [如何使用 AWS Budgets 来跟踪我的支出和使用情况](#)

### 相关示例：

- [IAM 访问管理策略示例](#)
- [服务控制策略示例](#)
- [AWS 预算操作](#)
- [创建 IAM 策略以使用标记控制对 Amazon EC2 资源的访问](#)
- [限制 IAM 身份对特定 Amazon EC2 资源的访问](#)
- [创建 IAM 策略以按系列限制 Amazon EC2 使用](#)
- [架构完善的实验室：成本和使用情况治理（级别 100）](#)

- [架构完善的实验室：成本和使用情况治理（级别 200）](#)
- [使用 Amazon Q Developer in chat applications 的 Cost Anomaly Detection Slack 集成](#)

## COST02-BP06 跟踪项目生命周期

跟踪、衡量并审计项目、团队和环境的生命周期，以避免使用不必要的资源并为此付费。

在未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

通过有效跟踪项目生命周期，组织可以通过改进规划和管理并优化资源、时间和质量，来更好地控制成本。通过跟踪获得的见解对于作出明智的决策非常宝贵，有助于提高项目的成本效益和整体成功率。

跟踪工作负载的整个生命周期有助于您了解何时不再需要工作负载或工作负载组件。现有的工作负载和组件可能看起来仍在使用中，但是当 AWS 发布新的服务或功能时，它们可能会被停用或采用。检查工作负载的先前阶段。在工作负载进入生产之后，可以停用以前的环境或大幅降低其容量，直到再次需要它们为止。

AWS 提供了许多可用于实体生命周期跟踪的管理和治理服务。可以使用 [AWS Config](#) 或 [AWS Systems Manager](#) 提供一份详尽的 AWS 资源和配置清单。建议集成现有项目或资产管理系统来跟踪组织内的活动项目和产品。将当前系统与 AWS 提供的丰富事件集和指标结合起来，您就可以构建大量生命周期事件的视图并主动管理资源，以减少不必要的成本。

与[应用程序生命周期管理（ALM）](#)类似，跟踪项目生命周期应涉及多个流程、工具和团队的协同工作，例如设计和开发、测试、生产、支持和工作负载冗余。

通过仔细监控项目生命周期的每个阶段，组织可以获得重要的见解并增强掌控力，从而促进成功的项目规划、实施和完成。这种仔细的监督可以确保项目不仅符合质量标准，而且在预算范围内按时交付，从而提高总体成本效率。

有关实施实体生命周期跟踪的更多详细信息，请参阅 [AWS Well-Architected 卓越运营支柱白皮书](#)。

### 实施步骤

- 建立项目生命周期监控流程：[云卓越中心团队](#)必须建立项目生命周期监控流程。建立结构化的系统化方法来监控工作负载，以改善项目的控制、可见性和性能。使监控过程透明化、协作化并注重持续改进，以最大限度地提高其效率和价值。
- 执行工作负载审查：根据组织政策的规定，定期审计现有项目并执行工作负载审查。在审计方面投入的工作量应与组织的大致风险、价值或成本成比例。主要审计领域包括组织面临的事件或中断风险，

或对组织所做的贡献（以收入或品牌声誉进行衡量）、工作负载的成本（以资源的总成本和运营成本进行衡量）和工作负载的使用量（以单位时间的组织产出量进行衡量）。如果这些领域在生命周期内发生变化，则需要对工作负载进行调整，例如全部停用或部分停用。

## 资源

### 相关文档：

- [AWS 上的标记指南](#)
- [什么是 ALM（应用程序生命周期管理）？](#)
- [针对工作职能的 AWS 托管策略](#)

### 相关示例：

- [使用 IAM 策略控制对 AWS 区域的访问](#)

### 相关工具：

- [AWS Config](#)
- [AWS Systems Manager](#)
- [AWS Budgets](#)
- [AWS Organizations](#)
- [AWS CloudFormation](#)

## COST 3.您如何监控成本和使用情况？

建立策略和程序以便监控并适当分配您的成本。这让您能够衡量和改进工作负载的成本效益。

## 最佳实践

- [COST03-BP01 配置详细信息源](#)
- [COST03-BP02 在成本和使用情况中添加组织信息](#)
- [COST03-BP03 确定成本归属类别](#)
- [COST03-BP04 建立组织指标](#)
- [COST03-BP05 配置账单和成本管理工具](#)



- [COST03-BP06 根据工作负载指标分配成本](#)

## COST03-BP01 配置详细信息源

按小时粒度配置成本管理和报告工具，以提供详细的成本和使用量信息，从而实现更深入的分析 and 透明度。配置工作负载，为交付的每个业务成果生成或提供日志条目。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

利用成本管理工具中的详细账单信息（如每小时粒度），组织能够更详细地跟踪使用情况，并有助于组织找出一些成本增加的原因。这些数据源最确切地反映了整个组织中的成本和使用量。

AWS 成本和使用情况报告 提供所有收费 AWS 服务的每日或每小时使用粒度、费率、成本和使用属性。CUR 中的所有可能维度包括：标记、位置、资源属性和账户 ID。

使用以下自定义项配置 CUR：

- 包括资源 ID
- 自动刷新 CUR
- 每小时粒度
- 版本控制：覆盖现有报告
- 数据集成：Athena（Parquet 格式和压缩）

使用 [AWS Glue](#) 准备分析数据、使用 [Amazon Athena](#) 执行数据分析、使用 SQL 查询数据。您也可以使用 [QuickSight](#) 构建复杂的自定义视图，并在整个组织内分发。

### 实施步骤

- 配置成本和使用情况报告：使用账单控制台，至少配置一个成本和使用情况报告。配置以每小时为粒度的报告，以便包括所有标识符和资源 ID。还可以创建采用不同粒度的其他报告，以提供概括性摘要信息。
- 在 Cost Explorer 中配置每小时粒度：启用 每小时 和 资源级别数据，以按照每小时粒度访问过去 14 天的成本和使用情况数据，并按资源级别粒度访问这些数据。
- 配置应用程序日志记录：确认应用程序记录所交付的每项业务成果，以便进行跟踪和衡量。确保该数据的粒度至少为每小时一次，以便与成本和使用情况数据匹配。有关日志记录和监控的更多详细信息，请参阅 [Well-Architected 卓越运营支柱](#)。

## 资源

相关文档：

- [AWS 成本和使用情况报告](#)
- [AWS Glue](#)
- [QuickSight](#)
- [AWS 成本管理定价](#)
- [标记 AWS 资源](#)
- [使用 AWS Budgets 分析成本](#)
- [使用 Cost Explorer 分析成本](#)
- [管理 AWS 成本和使用情况报告](#)
- [Well-Architected 卓越运营支柱](#)

相关示例：

- [AWS Account Setup](#)
- [AWS Cost Explorer's New Look and Common Use Cases](#)

COST03-BP02 在成本和使用情况中添加组织信息

根据您的组织、工作负载属性和成本分配类别定义标记方案，以便您可以在成本管理工具中筛选和搜索资源或监控成本和使用情况。在可能的情况下，根据目的、团队、环境或与您的业务相关的其他标准，在所有资源中实施一致的标记。

在未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

[在 AWS 中实施标记](#)，以将组织信息添加到您的资源中，然后将其添加到成本和使用量信息中。标记是键值对 — 键是定义的，必须在整个组织中唯一，值则对于一组资源唯一。键值对的一个示例是键为 Environment，值为 Production。生产环境中的所有资源都有这个键值对。借助标记，您可以使用有意义、相关的组织信息对成本进行分类和跟踪。您可以应用代表组织类别（例如成本中心、应用名称、项目或拥有者）的标记，标识工作负载和工作负载的特征（例如测试或生产），以在整个组织中分摊成本和使用量。

当您将标记应用于 AWS 资源（如 Amazon Elastic Compute Cloud 实例或 Amazon Simple Storage Service 存储桶）并激活标记后，AWS 会将此信息添加到成本和使用情况报告中。您可以在带标记和无标记的资源上运行报告并执行分析，以更好地遵守内部成本管理策略，并确保准确归属。

跨组织账户创建和实施 AWS 标记标准之后，您将能够一致且统一地管理和治理 AWS 环境。使用 AWS Organizations 中的[标记策略](#)定义有关如何对 AWS Organizations 中的账户中的 AWS 资源使用标记的规则。借助标记策略，您可以采用标准化方法轻松标记 AWS 资源

[AWS 标记编辑器](#)可用于为多个资源添加、删除和管理标记。通过标记编辑器，您可以搜索要标记的资源，然后在搜索结果中管理资源的标记。

[AWS Cost Categories](#) 可用于向成本分配组织含义，而无需在资源上添加标记。您可以将成本和使用量信息映射到唯一的内部组织结构。您可以定义类别规则，以使用账单维度（例如账户和标签）对成本进行映射和分类。除了标记之外，这还提供了另外一个级别的管理能力。您还可以将特定账户和标记映射到多个项目。

## 实施步骤

- 定义标记方案：召集整个业务的所有利益相关者来定义方案。这通常包括担任技术、财务和管理角色的人员。定义所有资源必须具有的标签列表，以及资源应该具有的标签列表。验证标记名称和值在整个组织中是否一致。
- 标记资源：使用定义的成本归属类别，根据类别在工作负载中的所有资源上[放置标记](#)。使用 CLI、标记编辑器或 AWS Systems Manager 等工具提高效率。
- 实施 AWS Cost Categories：您可以创建 [Cost Categories](#)，而无需实施标记。Cost Categories 使用现有的成本和使用量维度。根据方案创建类别规则，并在 Cost Categories 中加以实施。
- 自动标记：要验证您是否在所有资源中保持高水平的标记，请自动标记，以便在创建资源时自动标记资源。使用服务（如 [AWS CloudFormation](#)）验证资源在创建时是否已标记。还可以创建一个自定义解决方案，使用 Lambda 函数[自动标记](#)，或使用微服务定期扫描工作负载并删除任何未标记的资源，这是测试和开发环境的理想选择。
- 监控和报告标记：要验证您是否在整个组织中保持高水平的标记，请监控并报告工作负载中的标记。可以使用 [AWS Cost Explorer](#) 查看标记资源和未标记资源的成本，也可以使用[标记编辑器](#)等服务。定期审核未标记资源的数量，并执行操作添加标记，直到达到所需的标记级别。

## 资源

### 相关文档：

- [标记最佳实践](#)

- [AWS CloudFormation 资源标记](#)
- [AWS Cost Categories](#)
- [标记 AWS 资源](#)
- [使用 AWS Budgets 分析成本](#)
- [使用 Cost Explorer 分析成本](#)
- [管理 AWS 成本和使用情况报告](#)

相关视频：

- [如何对我的 AWS 资源进行标记，以便按成本中心或项目划分账单](#)
- [标记 AWS 资源](#)

相关示例：

- [根据标识或角色自动对新的 AWS 资源进行标记](#)

### COST03-BP03 确定成本归属类别

确定组织类别，例如业务单位、部门或项目，这些类别可用于在组织中按照内部使用实体分配成本。利用这些类别来执行支出问责制，树立成本意识，并推动高效的使用行为。

未建立这种最佳实践的情况下暴露的风险等级：高

#### 实施指导

成本分类过程在预算编制、会计、财务报告、决策、基准制定和项目管理中至关重要。通过对费用进行分级和分类，团队可以更好地了解自己整个云之旅中会产生成本类型，从而有助于团队做出明智的决策并高效地管理预算。

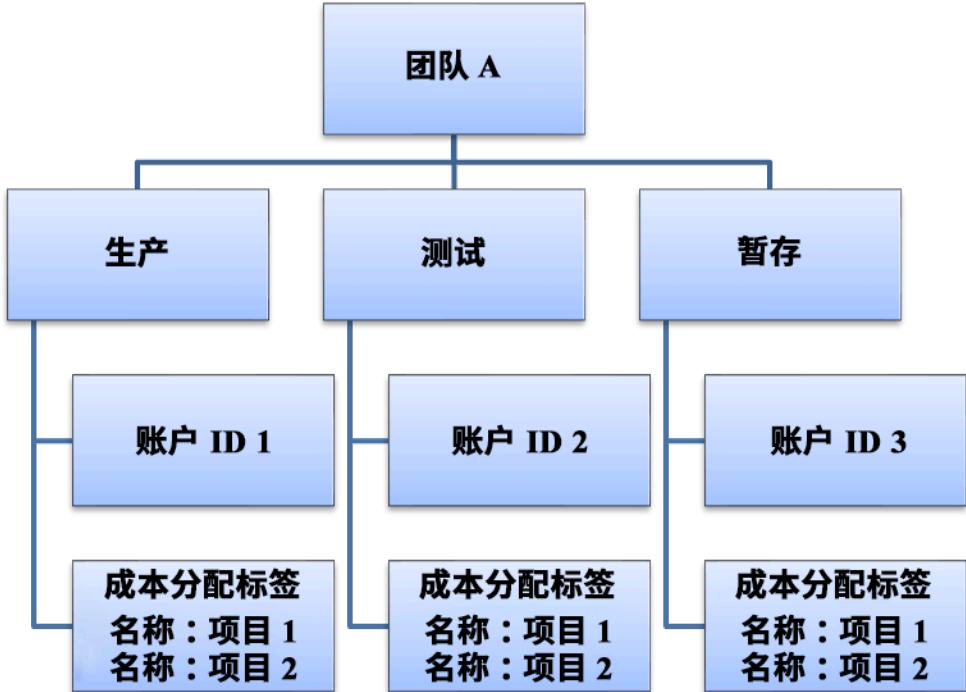
云支出问责制为严明的需求和成本管理提供了强有力的激励措施。最终，组织在将大部分云支出分配到使用这些资源的业务单位或团队后，可以显著地节省云成本。此外，通过分配云支出，有助于组织采用更多集中式云监管的最佳实践。

与您的财务团队和其他利益相关者合作，在定期沟通通话期间，了解必须如何在组织内部分配成本的要求。必须将工作负载成本分配至整个生命周期，包括开发、测试、生产和停用。了解组织如何对学习、员工培养和创意构思进行成本归类。这有助于将用于此目的的账户正确分配给培训和开发预算，而不是一般的 IT 成本预算。

在与组织中的利益相关者定义成本归属类别后，使用 [AWS Cost Categories](#) 将您在 AWS Cloud 中的成本和使用情况信息分成有意义的类别，例如特定项目的成本，或者部门或业务单位的 AWS 账户的成本。您可以创建自定义类别，并根据使用各种维度（如账户、标记、服务或费用类型）定义的规则，将成本和使用量信息映射到这些类别中。设置成本类别后，您可以按这些类别查看成本和使用情况信息，从而使贵组织能够做出更好的战略和采购决策。这些类别也可在 AWS Cost Explorer、AWS Budgets 和 AWS 成本和使用情况报告 中查看。

例如，为业务单位（DevOps 团队）创建成本类别，并在每个类别下创建多个规则（每个子类别的规则），这些规则具有基于所定义分组的多个维度（AWS 账户、成本分配标记、服务或费用类型）。通过成本类别，您可以使用基于规则的引擎整理成本。您配置的规则可将成本分类。在这些规则中，您可以使用每个类别的多个维度进行筛选，例如特定 AWS 账户、AWS 服务或费用类型。然后，您可以在 [AWS 账单与成本管理](#) 和 [成本管理控制台](#) 中，跨多种产品使用这些类别。这些产品包括 AWS Cost Explorer、AWS Budgets、AWS 成本和使用情况报告 和 AWS Cost Anomaly Detection。

下图举例说明了如何通过多个团队（成本类别）、多个环境（规则）以及每个环境具有多个资源或资产（维度），来对组织中的成本和使用信息进行分组。



成本和使用量组织结构图

您也可以使用成本类别创建成本分组。创建成本类别后（为使用量记录创建成本类别之后，允许在 24 小时内更新相应的值），这些类别将显示在 [AWS Cost Explorer](#)、[AWS Budgets](#)、[AWS 成本和使用](#)

[情况报告](#)和 [AWS Cost Anomaly Detection](#) 中。在 AWS Cost Explorer 和 AWS Budgets 中，成本类别显示为额外的计费维度。您可以使用此选项筛选特定的成本类别值，或按成本类别进行分组。

## 实施步骤

- 定义组织类别：与内部利益相关者和业务单位会面，确定反映组织结构和要求的类别。这些类别应直接对应于现有财务类别的结构，例如业务单位、预算、成本中心或部门。了解云为您带来的业务成果（例如培训或教育），因为这些也是组织类别。
- 定义功能类别：与内部利益相关者和业务单位会面，确定反映业务所含功能的类别。这可以是工作负载名称或应用程序名称以及环境类型（例如生产、测试或开发）。
- 定义 AWS Cost Categories：使用 [AWS Cost Categories](#) 创建成本类别以整理成本和使用情况信息，并将 AWS 成本和使用情况映射到 [有意义的类别](#)。可以将多个类别分配给一个资源，并且一个资源可以位于多个不同的类别中，因此可以根据需要定义任意数量的类别，以便您在分类的结构中 [管理成本](#)（使用 AWS Cost Categories）。

## 资源

相关文档：

- [标记 AWS 资源](#)
- [使用成本分配标签](#)
- [使用 AWS Budgets 分析成本](#)
- [使用 Cost Explorer 分析成本](#)
- [管理 AWS 成本和使用情况报告](#)
- [AWS Cost Categories](#)
- [使用 AWS Cost Categories 管理成本](#)
- [创建成本类别](#)
- [标记成本类别](#)
- [在成本类别中拆分费用](#)
- [AWS Cost Categories 的功能](#)

相关示例：

- [使用 AWS Cost Categories 整理成本和使用量数据](#)
- [使用 AWS Cost Categories 管理成本](#)

- [Well-Architected 实验室：成本和使用量可视化](#)
- [Well-Architected 实验室：成本类别](#)

## COST03-BP04 建立组织指标

建立此工作负载需要的组织指标。生成的客户报告或提供给客户的 Web 页面都属于工作负载指标。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

了解如何根据业务成功来衡量工作负载的输出。每个工作负载通常有一组表示性能的主要输出。如果您的工作负载复杂且包含许多组件，则可以对列表进行优先级排序，或者为每个组件定义和跟踪指标。与团队合作，了解要使用哪些指标。此部分将用于了解工作负载的效率，或每项业务输出的成本。

### 实施步骤

- 定义工作负载结果：与业务利益相关者召开会议，定义工作负载成果。这些主要用于衡量客户使用情况，因此必须是业务指标，而不是技术指标。每个工作负载应该有少量的概要指标（少于 5 个）。如果工作负载针对不同的使用案例产生多个结果，请将其分组为一个指标。
- 定义工作负载组件结果：如果工作负载大而复杂，或者可以轻松地将工作负载分为输入和输出定义明确的多个组件（例如微服务），则可以选择为每个组件定义指标。这项工作应反映组件的价值和成本。按照从大到小的顺序，从最大的组件开始，逐步处理较小的组件。

### 资源

相关文档：

- [标记 AWS 资源](#)
- [使用 AWS Budgets 分析成本](#)
- [使用 Cost Explorer 分析成本](#)
- [管理 AWS 成本和使用情况报告](#)

## COST03-BP05 配置账单和成本管理工具

根据您的组织策略配置成本管理工具，以管理和优化云支出。这包括用于整理和跟踪成本和使用量数据的服务、工具和资源，通过整合账单和访问权限增强控制，通过预算和预测改进规划，接收通知或提醒，并通过资源和定价优化进一步降低成本。



未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

为了建立强有力的问责制，首先应将您的账户策略视为成本分配策略的一部分。做好这一点，可能会为您省下许多工作。否则就会出现意识不足的情况，造成更多痛点。

为了鼓励针对云支出建立问责制，用户应有权使用可查看其成本和使用量的工具。建议为所有工作负载和团队都配置相关工具，以获取以下详细信息和了解用途：

- **整理：** 使用您自己的标记策略和分类建立您的成本分配和治理基准。
- **整理：** 使用您自己的标记策略和分类方法，建立您的成本分配和治理基准。标记支持的 AWS 资源，并根据您的组织结构（业务单位、部门或项目）对其进行有目的性的分类。为特定成本中心标记账户名称，并将其与 AWS Cost Categories 对应起来，以为特定业务单位就其成本中心进行账户分组，这样业务单位负责人就可以在一个位置查看多个账户的使用量。
- **访问：** 在 [整合账单](#) 中跟踪组织范围内的计费信息，并确认相应的利益相关者和业务负责人是否拥有访问权限。
- **控制：** 使用服务控制策略（SCP，Service Control Policy）、标签策略和预算警报时，使用适当的防护机制建立有效的治理机制，以防止出现意外情况。例如，您可以通过使用有效的控制机制，只允许团队在首选区域创建资源。
- **当前状态：** 配置显示当前成本和使用量水平的控制面板。控制面板应位于工作环境中的显眼位置，类似于操作控制面板。您可以使用 [Cloud Intelligence 控制面板（CID）](#) 或任何其他支持的产品来实现这种方便查看的效果。
- **通知：** 当成本或使用量超出定义的限值时，以及在 AWS Budgets 和 AWS Cost Anomaly Detection 中出现异常时，触发通知。
- **报告：** 汇总所有成本和使用量信息，并通过详细的、可归因的成本数据提高对云支出的认识和加强问责制。报告应与使用报告的团队相关，理想情况下应包含建议。
- **跟踪：** 对照配置的方向性目标或执行性目标显示当前的成本和使用量。
- **分析：** 让团队成员可按照所有可能的维度，执行详尽至每小时粒度的自定义和深入分析。
- **检查：** 及时了解资源部署和成本优化机会。获取有关组织级别资源部署的通知（使用 Amazon CloudWatch、Amazon SNS 或 Amazon SES），以及查看成本优化建议（例如，AWS Compute Optimizer 或 AWS Trusted Advisor）。
- **趋势分析：** 以所需的粒度显示所需时间段内成本和使用量的变化。
- **预测：** 使用您创建的预测控制面板，显示估计的未来成本、估计资源使用量和支出。

您可以使用 AWS 工具（例如 [AWS Cost Explorer](#)、[AWS 账单与成本管理](#) 或 [AWS Budgets](#)）来获取基本功能，您也可以将 CUR 数据与 [Amazon Athena](#) 和 [QuickSight](#) 集成来提供此功能，以获取更详细的视图。如果贵组织不具备必备技能或带宽，则可以与 [AWS ProServ](#)、[AWS Managed Services \(AMS\)](#) 或 [AWS Partner](#) 合作，并使用他们的工具。您也可以使用第三方工具，但首先要验证成本是否为贵组织提供了价值。

## 实施步骤

- 允许对工具进行团队性质的访问：配置您的账户并创建组，这些组可以访问所需的成本和使用量报告来了解其使用情况，并使用 [AWS Identity and Access Management](#) 来 [控制](#) 对 AWS Cost Explorer 等工具的访问。这些组必须包括负责或管理应用程序的所有团队的代表。这证明每个团队都可以访问他们的成本和使用量信息，以跟踪他们的使用情况。
- 配置 AWS Budgets：[在所有账户中为您的工作负载配置 AWS Budgets](#)。通过使用标记设置账户总支出预算和工作负载预算。在 AWS Budgets 中配置通知，以便在您超出预算金额或估计成本超出预算时收到警报。
- 配置 AWS Cost Explorer：为您的工作负载和账户配置 [AWS Cost Explorer](#)，从而实现您的成本数据可视化，以便进一步分析。为工作负载创建一个控制面板，用于跟踪总体支出、工作负载的关键使用指标，以及基于历史成本数据的未来成本预测。
- 配置 AWS Cost Anomaly Detection：将 [AWS Cost Anomaly Detection](#) 用于您创建的账户、核心服务或成本类别，以监控您的成本和使用量，并检测异常支出。您可以在汇总的报告中单独收到警报，以及在电子邮件或 Amazon SNS 主题中收到警报，以便您分析和确定异常的根本原因，并确定导致成本增加的原因。
- 配置高级工具：您可以选择为组织创建自定义工具，以便提供额外详细信息和所需的粒度。您可以使用 [Amazon Athena](#) 实施高级分析功能，使用 [QuickSight](#) 实施控制面板。考虑使用 [CID 解决方案](#)，它具有预配置的高级控制面板。您还可以与 [AWS Partner](#) 合作并采用其云管理解决方案，以便在一个方便的位置激活云账单监控和优化功能。

## 资源

### 相关文档：

- [AWS 成本管理](#)
- [标记 AWS 资源](#)
- [使用 AWS Budgets 分析成本](#)
- [使用 Cost Explorer 分析成本](#)
- [管理 AWS 成本和使用情况报告](#)

- [AWS Cost Categories](#)
- [使用 AWS 管理云财务](#)
- [服务控制策略示例](#)
- [AWS APN 合作伙伴 – 成本管理](#)

相关视频：

- [部署 Cloud Intelligence 控制面板](#)
- [获取任何 FinOps 或成本优化指标或 KPI 的警报](#)

相关示例：

- [Well-Architected 实验室 – AWS 账户设置](#)
- [Well-Architected 实验室：账单可视化](#)
- [Well-Architected 实验室：成本和治理使用情况](#)
- [Well-Architected 实验室：成本和使用量分析](#)
- [Well-Architected 实验室：成本和使用量可视化](#)
- [Well-Architected 实验室：Cloud Intelligence 控制面板](#)
- [如何使用 SCP 跨账户设置权限防护机制](#)

## COST03-BP06 根据工作负载指标分配成本

根据使用量指标或业务成果分配工作负载的成本，以便衡量工作负载的成本效益。实施一个流程，使用分析服务来分析成本和使用量数据，以便深入了解成本因素和退款功能。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

成本优化旨在以最低的价格实现业务成果，这只能通过按工作负载指标分配工作负载成本（按工作负载效率衡量）来实现。通过日志文件或其他应用程序监控来监控定义的工作负载指标。将此数据与工作负载成本（可通过查看具有特定标签值或账户 ID 的成本获得）相结合。建议每小时进行一次分析。如果有一些静态成本要素（例如，持续运行的后端数据库）且请求率不同（例如，使用量高峰在上午 9 点至下午 5 点，晚间的请求数量很少），则效率通常会变化。了解静态成本和可变成本之间的关系有助于您将精力集中在优化活动上。

与 Amazon Elastic Container Service ( Amazon ECS ) 和 Amazon API Gateway 上的容器化应用程序等资源相比，为共享资源创建工作负载指标可能并非易事。但是，您可以通过某些方法来分类使用量并跟踪成本。如果您需要跟踪 Amazon ECS 和 AWS Batch 共享资源，则可以在 AWS Cost Explorer 中启用拆分成本分配数据。通过拆分成本分配数据，您可以了解并优化容器化应用程序的成本和使用量，并根据共享计算和内存资源的使用情况，将应用程序成本分配给各个业务实体。如果您有共享的 API Gateway 和 AWS Lambda 函数使用量，那么您可以使用 [AWS Application Cost Profiler](#)，根据函数的 ## ID 或 ## ID 对使用情况分类。

## 实施步骤

- 将成本分配到工作负载指标：使用定义的指标和配置的标记，创建结合工作负载输出和工作负载成本的指标。使用 Amazon Athena 和 Amazon QuickSight 等分析服务，为整个工作负载和任何组件创建效率控制面板。

## 资源

相关文档：

- [标记 AWS 资源](#)
- [使用 AWS Budgets 分析成本](#)
- [使用 Cost Explorer 分析成本](#)
- [管理 AWS 成本和使用量报告](#)

相关示例：

- [利用 AWS 拆分成本分配数据提高 Amazon ECS 和 AWS Batch 的成本可见性。](#)

## COST 4. 您如何停用资源？

在从项目开始到结束的过程中实施变更控制和资源管理。这可以确保您关闭或终止未使用的资源，以便减少浪费。

## 最佳实践

- [COST04-BP01 在资源生命周期内跟踪资源](#)
- [COST04-BP02 实施停用流程](#)
- [COST04-BP03 停用资源](#)
- [COST04-BP04 自动停用资源](#)

- [COST04-BP05 执行数据留存策略](#)

## COST04-BP01 在资源生命周期内跟踪资源

制定和实施一种方法，在资源生命周期内跟踪资源及其与系统的关联。您可以使用标记来标识资源的工作负载或功能。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

停用不再需要的工作负载资源。一个常见的示例是用于测试的资源：在测试完成后，可以将资源删除。使用标记跟踪资源（并对这些标记运行报告），可以帮助您识别要停用的资产，因为将不会使用这些资产，或者这些资产的许可证将过期。使用标记是跟踪资源的一种有效方法，它通过标记资源的功能或资源的已知可停用日期来跟踪资源。然后，可以对这些标记运行报告。功能标记的示例值是 `feature-X testing`，用于根据工作负载生命周期标识资源的用途。另一个示例是对资源使用 `LifeSpan` 或 `TTL`，例如要删除的标记键名称和值，以定义停用的时间段或具体时间。

### 实施步骤

- 实施标记方案：实施标记方案，标识资源所属的工作负载，从而验证相应地标记了工作负载中的所有资源。标记可以帮助您按用途、团队、环境或与您的业务相关的其他条件对资源进行分类。有关标记应用场景、策略和技术的更多详细信息，请参阅 [AWS 标记最佳实践](#)。
- 实施工作负载吞吐量或输出监控：实施工作负载吞吐量监控或警报，在输入请求或输出完成时触发。将其配置为在工作负载请求或输出下降到零时发出通知，指示不再使用工作负载资源。如果在正常情况下，工作负载周期性地下降到零，则加入时间因素。有关未使用或未充分利用的资源的更多详细信息，请参阅 [AWS Trusted Advisor 成本优化检查](#)。
- 对 AWS 资源进行分组：为 AWS 资源创建分组。您可以使用 [AWS Resource Groups](#) 来组织和管理位于同一 AWS 区域中的 AWS 资源。您可以将标记添加到您的大多数资源中，以帮助识别资源，并为您组织内的资源进行排序。可以使用 [标记编辑器](#) 将标记批量添加到支持的资源。考虑使用 [AWS Service Catalog](#) 来创建、管理已批准的产品组合并将其分发给最终用户，以及管理产品生命周期。

### 资源

#### 相关文档：

- [AWS Auto Scaling](#)
- [AWS Trusted Advisor](#)

- [AWS Trusted Advisor 成本优化检查](#)
- [标记 AWS 资源](#)
- [发布自定义指标](#)

相关视频：

- [如何使用 AWS Trusted Advisor 优化成本](#)

相关示例：

- [组织 AWS 资源](#)
- [使用 AWS Trusted Advisor 优化成本](#)

## COST04-BP02 实施停用流程

实施一个流程来确定和停用未使用的资源。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

在整个组织中实施标准化流程，以识别和删除未使用的资源。该流程应该定义执行搜索的频率以及删除资源的流程，以验证满足所有组织要求。

### 实施步骤

- 创建并实施停用流程：与工作负载开发人员和负责人合作，为工作负载及其资源构建停用流程。该流程应涵盖一种方法，验证工作负载是否正在使用以及每个工作负载资源是否正在使用。详细说明停用资源所需的步骤，将资源从服务中删除同时确保符合任何法规要求。应包含任何关联的资源，例如许可证或附加的存储。向工作负载负责人发送已执行停用流程的通知。

使用以下停用步骤来指导您确定在流程中应检查的内容：

- 确定要停用的资源：确定符合 AWS Cloud 中停用条件的资源。记录所有必要的信息并安排停用。在您的时间表中，请务必考虑在此过程中是否（以及何时）会出现意外问题。
- 协调和沟通：与工作负载负责人合作，确认要停用的资源
- 记录元数据并创建备份：如果生产环境中的资源需要元数据或元数据是关键资源，则记录元数据（例如公有 IP、区域、可用区、VPC、子网和安全组）并创建备份（例如 Amazon Elastic Block Store 快照或执行 AMI、密钥导出和证书导出）。

- 验证基础设施即代码：确定资源是使用 AWS CloudFormation、Terraform、AWS Cloud Development Kit (AWS CDK)，还是任何其他基础设施即代码部署工具部署的，以便在必要时可以重新部署它们。
- 阻止访问：在一段时间内应用限制性控制，以防止在确定是否需要资源时使用资源。验证资源环境是否可以在需要时恢复到其原始状态。
- 遵循内部停用流程：遵循组织的管理任务和停用流程，例如从组织域中删除资源，删除 DNS 记录，以及从配置管理工具、监控工具、自动化工具和安全工具中删除资源。

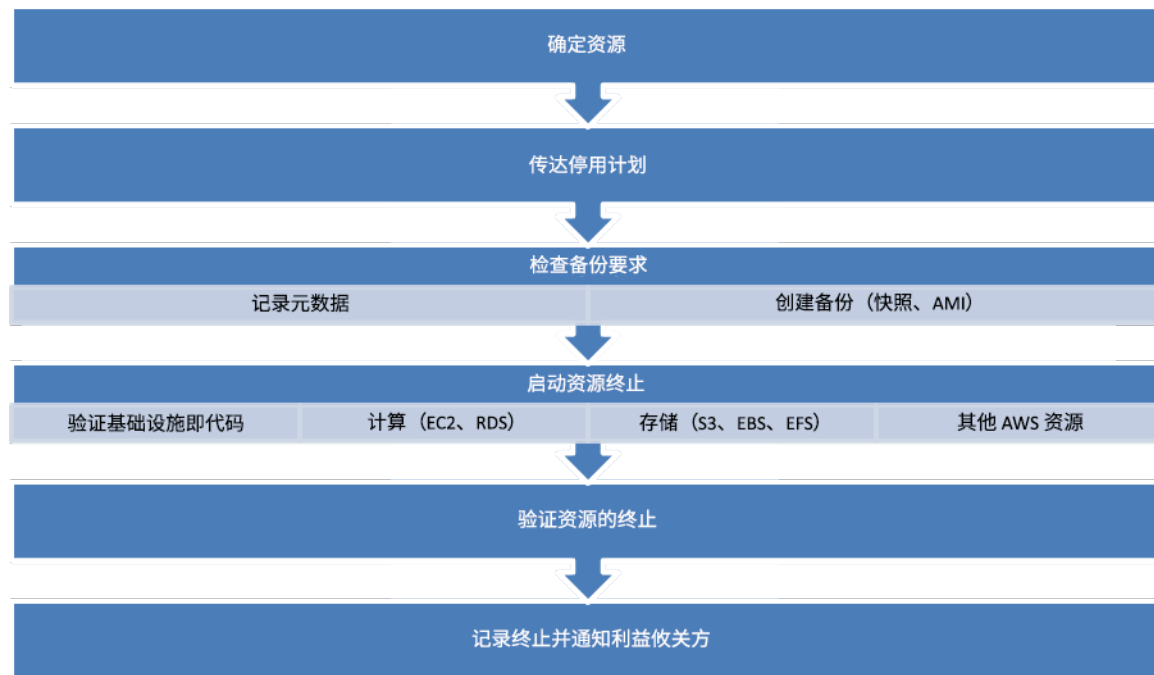
如果资源是 Amazon EC2 实例，请参阅以下列表。[有关更多信息，请参阅“如何删除或终止我的 Amazon EC2 资源？”](#)

- 停止或终止所有 Amazon EC2 实例和负载均衡器。Amazon EC2 实例在终止后会在控制台中短暂显示。您不需要为任何未处于运行状态的实例付费
- 删除您的 Auto Scaling 基础设施。
- 释放所有专属主机。
- 删除所有 Amazon EBS 卷和 Amazon EBS 快照。
- 释放所有弹性 IP 地址。
- 取消注册所有亚马逊云机器镜像 (AMI)。
- 终止所有 AWS Elastic Beanstalk 环境。

如果资源是 Amazon S3 Glacier 存储中的对象，并且您在满足最短存储持续时间之前删除了存档，将按比例收取提前删除费用。Amazon S3 Glacier 最短存储持续时间取决于所使用的存储类。有关每个存储类的最短存储持续时间的摘要，请参阅[跨 Amazon S3 存储类的性能](#)。有关如何计算提前删除费用的详细信息，请参阅[Amazon S3 定价](#)。

下面的简单停用流程图概述了停用步骤。在停用资源之前，请验证组织是否未使用已确定要停用的资源。





资源停用流程。

资源

相关文档：

- [AWS Auto Scaling](#)
- [AWS Trusted Advisor](#)
- [AWS CloudTrail](#)

相关视频：

- [删除 CloudFormation 堆栈，但保留某些资源](#)
- [了解哪个用户启动了 Amazon EC2 实例](#)

相关示例：

- [删除或终止 Amazon EC2 资源](#)
- [了解哪个用户启动了 Amazon EC2 实例](#)

## COST04-BP03 停用资源

停用由定期审计或使用情况发生变化等事件触发的资源。停用通常定期执行，可以手动停用，也可以自动停用。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

搜索未使用资源的频率和工作量应反映潜在的节省额，因此，与成本较高的账户相比，对成本较低的账户进行分析的频率应该更低。搜索和停用事件可由工作负载中的状态更改触发，比如产品生命周期结束或被更换。搜索和停用事件也可由外部事件触发，如市场条件发生变化或产品终止。

### 实施步骤

- 停用资源：这是 AWS 资源的折旧阶段，不再需要这些资源或许可协议终止。在进入处置阶段并停用资源之前完成所有最终检查，以防止任何意外中断，例如拍摄快照或备份。使用停用流程，停用每项被确定为未使用的资源。

### 资源

相关文档：

- [AWS Auto Scaling](#)
- [AWS Trusted Advisor](#)

相关示例：

- [架构完善的实验室：停用资源 \( 级别 100 \)](#)

## COST04-BP04 自动停用资源

设计您的工作负载，使其在您发现并停用非关键资源、不需要的资源或使用率低的资源时妥善处理资源的终止。

在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

使用自动化技术可以减少或消除停用流程中的相关成本。将工作负载设计为执行自动化停用将减少工作负载在其整个生命周期内的总成本。可以使用 [AWS Auto Scaling](#) 执行停用流程。还可以使用 [API 或开发工具包](#) 来实施自定义代码，以自动停用工作负载资源。

[现代应用程序](#) 首先构建为无服务器，这是一种优先采用无服务器服务的策略。AWS 为堆栈的所有三个层开发了 [无服务器服务](#)：计算层、集成层和数据存储层。使用无服务器架构将允许您在低流量期间通过自动纵向扩展和缩减来节省成本。

## 实施步骤

- 实施 AWS Auto Scaling：对于受支持的资源，请使用 [AWS Auto Scaling](#) 进行配置。AWS Auto Scaling 可以帮助您在使用 AWS 服务时优化利用率和成本效率。当需求下降时，AWS Auto Scaling 将自动删除任何多余的资源容量，以避免超支。
- 配置 CloudWatch 以终止实例：可以将实例配置为使用 [CloudWatch 警报终止](#)。使用停用流程的指标，实施包含 Amazon Elastic Compute Cloud 操作的警报。在推出之前，在非生产环境中验证操作。
- 在工作负载中实施代码：可以使用 AWS 软件开发工具包或 AWS CLI 停用工作负载资源。在与 AWS 集成的应用程序中实施代码，并终止或删除不再使用的资源。
- 使用无服务器服务：在 AWS 上优先构建 [无服务器架构](#) 和 [事件驱动的架构](#)，以生成并运行应用程序。AWS 提供多种无服务器技术服务，这些服务本身提供自动优化的资源利用率和自动停用功能（横向缩减和横向扩展）。通过无服务器应用程序，可自动优化资源利用率，您无需为过度预置付费。

## 资源

相关文档：

- [AWS Auto Scaling](#)
- [AWS Trusted Advisor](#)
- [AWS 上的无服务器](#)
- [创建停止、终止、重启或恢复实例的警报](#)
- [开始使用 Amazon EC2 Auto Scaling](#)
- [向 Amazon CloudWatch 警报添加终止操作](#)

相关示例：

- [计划自动删除 AWS CloudFormation 堆栈](#)
- [架构完善的实验室 – 自动停用资源 \( 级别 100 \)](#)
- [Servian AWS 自动清理](#)

## COST04-BP05 执行数据留存策略

在支持的资源上定义数据留存策略，以根据您组织的要求处理对象的删除事宜。识别并删除非必要的或不再需要的孤立资源和对象。

在未建立这种最佳实践的情况下暴露的风险等级：中等

使用数据留存策略和生命周期策略，来降低停用过程的相关成本以及已确定资源的存储成本。定义您的数据留存策略和生命周期策略来执行自动存储类迁移和删除，将会降低其生命周期内的整体存储成本。您可以使用 Amazon Data Lifecycle Manager 来自动创建和删除 Amazon Elastic Block Store 快照及基于 Amazon EBS 的亚马逊云机器镜像（AMI），并使用 Amazon S3 Intelligent-Tiering 或 Amazon S3 生命周期配置来管理您的 Amazon S3 对象的生命周期。您还可以使用 [API 或 SDK](#) 实施自定义代码，以创建生命周期策略和策略规则，来自动删除对象。

### 实施步骤

- 使用 Amazon Data Lifecycle Manager：使用 Amazon Data Lifecycle Manager 中的生命周期策略来自动删除 Amazon EBS 快照和基于 Amazon EBS 的 AMI。
- 在桶上设置生命周期配置：对桶使用 Amazon S3 生命周期配置，以根据您的业务要求，定义 Amazon S3 要在对象生命周期内采取的操作，以及在对象生命周期结束时的删除操作。

### 资源

相关文档：

- [AWS Trusted Advisor](#)
- [Amazon Data Lifecycle Manager](#)
- [如何在 Amazon S3 桶上设置生命周期配置](#)

相关视频：

- [使用 Amazon Data Lifecycle Manager 实现 Amazon EBS 快照自动化](#)
- [使用生命周期配置规则清空 Amazon S3 桶](#)

相关示例：

- [使用生命周期配置规则清空 Amazon S3 桶](#)
- [Well-Architected 实验室：自动停用资源（级别 100）](#)

## 具有成本效益的资源

问题

- [COST 5.您在选择服务时如何评估成本？](#)
- [COST 6.在选择资源类型、规模和数量时，如何实现成本目标？](#)
- [COST 7.您如何使用定价模式来降低成本？](#)
- [COST 8.您如何规划数据传输费用？](#)

### COST 5.您在选择服务时如何评估成本？

Amazon EC2、Amazon EBS 和 Amazon S3 属于构建块 AWS 服务。托管服务（如 Amazon RDS 和 Amazon DynamoDB）属于更高级别或应用程序级别的 AWS 服务。通过选择适当的基础服务和托管服务，您可以优化工作负载，从而降低成本。例如，使用托管服务，您可以节省或消除大部分管理和运营开销，从而使您有精力从事应用程序和业务相关活动。

最佳实践

- [COST05-BP01 确定组织对成本的要求](#)
- [COST05-BP02 分析此工作负载的所有组件](#)
- [COST05-BP03 对每个组件进行彻底分析](#)
- [COST05-BP04 选择具有成本效益许可的软件](#)
- [COST05-BP05 选择此工作负载的组件，以便根据组织的优先事项优化成本](#)
- [COST05-BP06 对不同时间的不同使用情况执行成本分析](#)

#### COST05-BP01 确定组织对成本的要求

与团队成员合作，为此工作负载确定成本优化与其他支柱（例如性能和可靠性）之间的平衡。

在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

在大多数组织中，信息技术（IT）部门由多个小型团队组成，每个小团队都有自己的议程和重点领域，这反映了其团队成员的专长和技能。您需要了解组织的总体目标、优先事项、具体目标以及每个部门或项目如何为这些总体目标做出贡献。对所有基本资源（包括人员、设备、技术、材料和外部服务）进行分类，是实现组织目标和全面预算规划的重要一环。采用这种系统化的方法来确定和了解成本，对于组织制定切合实际且稳健的成本计划至关重要。

在为工作负载选择服务时，了解组织的优先要务至关重要。在成本优化和 AWS Well-Architected Framework 的其他支柱（例如性能和可靠性）之间取得平衡。这一过程应系统性地定期进行，以反映组织目标、市场条件和运营动态的变化。完全成本优化的工作负载是最符合组织需求的解决方案，但不一定是成本最低的。与组织内的所有团队（例如产品、业务、技术和财务）会面以收集信息。在有冲突的利益或替代方法之间做出权衡并评估其影响，以便在确定工作重心或选择行动方案时作出明智的决策。

例如，加快新功能上市的速度可能会比成本优化更重要，或者您可以为非关系数据选择关系数据库，以简化迁移系统的工作，而不是迁移到针对您的数据类型优化的数据库和更新您的应用程序。

## 实施步骤

- 确定组织对成本的要求：与组织中的团队成员会面，这些成员包括产品管理、应用程序负责人、开发和运营团队、管理和财务角色。对此工作负载及其组件的 Well-Architected 支柱进行优先级排序，输出应该是一个按顺序排列的支柱列表。您还可以为每个支柱添加一个权重，以指示该支柱体现的额外关注程度，或者两个支柱之间的关注点的相似程度。
- 解决技术债务并将其记录在案：在工作负载审查期间，解决技术债务。记录待办事项以便将来重新审视工作负载，目标是重构或重新构造以进一步优化工作负载。必须向其他利益相关者明确说明所做的权衡取舍。

## 资源

相关最佳实践：

- [REL11-BP07 构造您的产品以满足可用性目标和正常运行时间服务等级协议（SLA）](#)
- [OPS01-BP06 评估权衡](#)

相关文档：

- [AWS 总拥有成本（TCO）计算器](#)

- [Amazon S3 存储类](#)
- [云产品](#)

## COST05-BP02 分析此工作负载的所有组件

确认已分析工作负载的每个组件，无论当前大小或当前成本如何。审核工作应该体现出可能带来的好处，例如当前成本和预期成本。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

旨在为组织提供业务价值的工作负载组件可能包含各种服务。对于每个组件，组织可以选择特定的 AWS Cloud 服务来满足业务需求。对这些服务的熟悉程度或以往的使用经验等因素可能会影响这种选择。

确定组织的需求（如 [COST05-BP01 确定组织对成本的要求](#) 中提到的）后，对工作负载中的所有组件进行全面分析。根据当前和预计的成本和规模，分析每个组件。结合工作负载在其生命周期内可能实现的节省来考虑分析成本。分析该工作负载的所有组件所花费的精力应与优化该特定组件预期能产生的潜在节省或改进相匹配。例如，如果拟议资源的成本为每月 10 USD，在预测的负载下不会超过每月 15 USD，则花一天的时间将成本降低 50%（每月 5 USD）可能会超过系统使用寿命内的潜在收益。使用更快、更有效的基于数据的预估可为该组件带来最佳的总体结果。

工作负载可能会随时间变化，如果工作负载架构或使用量发生变化，原本合适的服务集可能不再是最优之选。为甄选服务进行分析时，必须考虑工作负载当前和未来的状态以及使用量水平。为将来的工作负载状态或使用量实施服务可以减少或消除未来进行更改所需的工作量，从而降低总体成本。例如，最初使用 Amazon EMR Serverless 可能是合适的选择。但是，随着该服务使用量的增加，过渡到 Amazon EC2 上的 Amazon EMR 可以降低该工作负载组件的成本。

对所有工作负载组件进行战略审查，无论其目前的属性如何，都有可能随着时间的推移带来显著的改进和财务节约。应仔细考量审查过程中投入的精力，同时仔细考虑可能实现的潜在优势。

[AWS Cost Explorer](#) 和 [AWS 成本和使用情况报告](#)（CUR）可以分析概念验证（PoC）或运行环境的成本。还可以使用 [AWS 定价计算器](#) 来估算工作负载成本。

### 实施步骤

- 列出工作负载组件：生成工作负载组件的列表。这用于验证是否分析了每个组件。投入的工作量应体现出组织优先事项所规定的工作负载的关键性。如果有多个数据库，按功能（例如生产数据库存储）将资源分组可以提高效率。



- 对组件列表进行优先级排序：获取组件列表，按工作顺序进行优先级排序。通常按照组件的成本从最昂贵到最便宜的顺序排列，或者按照组织优先事项规定的的关键性排列。
- 执行分析：对于列表中的每个组件，检查可用的选项和服务，然后选择最符合组织优先事项的选项。

## 资源

### 相关文档：

- [AWS 定价计算器](#)
- [AWS Cost Explorer](#)
- [Amazon S3 存储类](#)
- [云产品](#)

## COST05-BP03 对每个组件进行彻底分析

分析组织为每个组件付出的总体成本。通过考虑运营和管理成本（尤其是在使用云提供商提供的托管服务时）来计算总拥有成本。审核工作量应该体现可能带来的好处（例如分析时间与组件成本成正比）。

在未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

利用节省下来的时间，您的团队将能够专注于解决技术债务、创新和增值功能，并打造企业的竞争优势。例如，您可能需要尽快将数据库从本地环境直接迁移（也称为更换主机）到云，然后再进行优化。值得探索的是，通过使用 AWS 上可消除或减少许可证成本的托管服务，您可以节省多少成本。AWS 上的托管服务消除了维护服务的运营和管理负担（例如修补或更新 OS），让您可以专注于创新和业务。

由于托管服务在云级运行，它们可以提供更低的单位事务或服务成本。您可以在不改变应用程序核心架构的情况下进行潜在优化，以获得一些切实的优势。例如，您可能希望通过迁移到数据库即服务平台 [如 [Amazon Relational Database Service \(Amazon RDS\)](#)]，或将您的应用程序迁移到完全托管式平台（如 [AWS Elastic Beanstalk](#)），来减少管理数据库实例所花的时间。

通常，可以设置托管服务的部分属性，以确保容量足够。您必须设置和监控这些属性，以便最大限度地减少多余容量，并最大限度地提高性能。您可以使用 AWS Management Console 或 AWS API 和 SDK 来修改 AWS Managed Services 的属性，以使资源需求匹配不断变化的要求。例如，您可以增加或减少 Amazon EMR 集群（或 Amazon Redshift 集群）上的节点数量，以扩展或缩减集群。

您还可以在 AWS 资源上打包多个实例，以实现更高密度的使用量。例如，您可以在单个 Amazon Relational Database Service ( Amazon RDS ) 数据库实例上预置多个小型数据库。随着使用量的增长，您可以使用快照和还原过程将其中一个数据库迁移到专用 Amazon RDS 数据库实例。

在托管服务上预置工作负载时，您必须了解调整服务容量的要求。这些要求通常是时间、工作量和对正常工作负载运营的任何影响。预置的资源必须留出时间来进行任何更改，并预置必要的开销以允许这样做。通过使用与系统和监控工具（如 Amazon CloudWatch）集成的 API 和开发工具包，可以将修改服务所需的持续工作量减少至接近零。

[Amazon RDS](#)、[Amazon Redshift](#) 和 [Amazon ElastiCache](#) 提供托管数据库服务。[Amazon Athena](#)、[Amazon EMR](#) 和 [Amazon OpenSearch Service](#) 提供托管分析服务。

[AMS](#) 是代表企业客户和合作伙伴运营 AWS 基础设施的服务。该服务提供了一个安全且合规的环境，您可以将工作负载部署到其中。AMS 使用具有自动化功能的企业云运营模型，可以满足组织要求，更快地迁移到云中并降低持续管理的成本。

## 实施步骤

- 执行彻底分析：使用组件列表，从最高优先级到最低优先级遍历每个组件。对于优先级较高且成本较高的组件，执行额外分析并评估所有可用选项及其长期影响。对于优先级较低的组件，评估使用情况的变化是否会更改组件的优先级，然后以适当的工作量进行分析。
- 比较托管和未托管资源：考虑您所管理的资源的运营成本，并将其与 AWS 托管的资源进行比较。例如，审查您在 Amazon EC2 实例上运行的数据库，并与 Amazon RDS 选项（AWS 托管的服务）进行比较，或将 Amazon EMR 与在 Amazon EC2 上运行 Apache Spark 进行比较。当从自我管理的工作负载迁移到 AWS 完全托管的工作负载时，请仔细研究您的选择。需要考虑的三个最重要的因素是您想使用的[托管服务类型](#)、您将用于[迁移数据](#)的过程，以及了解[AWS 责任共担模式](#)。

## 资源

### 相关文档：

- [AWS 总拥有成本 \( TCO \) 计算器](#)
- [Amazon S3 存储类](#)
- [AWS Cloud 产品](#)
- [AWS 责任共担模式](#)

### 相关视频：

- [为何要迁移到托管数据库？](#)
- [什么是 Amazon EMR，如何才能使用它来处理数据？](#)

相关示例：

- [为何要迁移到托管数据库？](#)
- [使用 AWS DMS 将相同 SQL Server 数据库中的数据整合到单个 Amazon RDS for SQL Server 数据库中](#)
- [将数据大规模交付到 Amazon Managed Streaming for Apache Kafka \( Amazon MSK \)](#)
- [将 ASP.NET Web 应用程序迁移到 AWS Elastic Beanstalk](#)

## COST05-BP04 选择具有成本效益许可的软件

开源软件无需软件许可成本，从而大大节省了工作负载的成本。如果需要许可软件，应避免使用绑定到任意属性（如 CPU）的许可证，而应使用绑定到输出或结果的许可证。这些许可证的成本与所提供的效益更为相当。

在未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

开源起源于软件开发，表示软件符合某些自由发布的标准。任何人都可以检查、修改和增强开源软件的源代码。根据业务要求、工程师技能、预测的使用情况或其他技术依赖关系，组织可以考虑在 AWS 上使用开源软件来最大限度地降低许可成本。换句话说，使用[开源软件](#)可以降低软件许可成本。随着工作负载规模的扩展，这可能对工作负载成本产生重大影响。

将许可软件能够带来的好处与总成本进行比较，以优化工作负载。对许可中的任何更改及其对工作负载成本的影响建模。如果供应商更改了数据库许可证的成本，请调查这会如何影响工作负载的整体效率。考虑供应商的历史定价公告，了解其产品中的许可更改趋势。许可成本也可以独立于吞吐量或使用量进行扩缩，例如按硬件扩缩的许可证（CPU 绑定许可证）。应避免使用这些许可证，因为成本会迅速增加，而且无法取得相应的结果。

例如，与在 Windows 上运行 Amazon EC2 实例相比，使用 Linux 操作系统在 us-east-1 中运行另一个 Amazon EC2 实例可以将成本降低大约 45%。

[AWS 定价计算器](#) 提供了一种全面的方法，可以比较具有不同许可选项的各种资源（例如 Amazon RDS 实例和不同的数据库引擎）的成本。此外，AWS Cost Explorer 为现有工作负载（尤其是附带不同许可证的工作负载）的成本提供了宝贵的视角。对于许可证管理，[AWS License Manager](#) 提供了一种监督和处理软件许可证的简化方法。客户可以在 AWS Cloud 中部署和运行他们首选的开源软件。

## 实施步骤

- 分析许可选项：查看可用软件的许可条款。查看具有所需功能的开源版本，以及许可软件提供的效益是否大于成本。优惠条款可确保软件成本与所提供的效益相符。
- 分析软件提供商：查看供应商的任何历史定价或许可变更。了解与成果不符的任何变化，例如在特定供应商硬件或平台上运行的惩罚性条款。此外，还要了解他们如何执行审计和处罚。

## 资源

### 相关文档：

- [AWS 开源](#)
- [AWS 总拥有成本 \( TCO \) 计算器](#)
- [Amazon S3 存储类](#)
- [云产品](#)

### 相关示例：

- [开源博客](#)
- [AWS 开源博客](#)
- [优化与许可评测](#)

COST05-BP05 选择此工作负载的组件，以便根据组织的优先事项优化成本

为工作负载选择所有组件时考虑成本因素。这包括使用应用程序级别的托管服务或无服务器服务、容器或事件驱动型架构，以降低整体成本。使用开源软件、没有许可证费用的软件或替代方案来尽可能减少许可证成本，从而减少开支。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

在选择所有组件时考虑服务和选项的成本。这包括使用应用程序级别的托管服务，例如 [Amazon Relational Database Service](#) ( Amazon RDS )、[Amazon DynamoDB](#)、[Amazon Simple Notification Service](#) ( Amazon SNS ) 和 [Amazon Simple Email Service](#) ( Amazon SES )，来降低组织的总体成本。

使用无服务器服务和容器进行计算，例如用于静态网站的 [AWS Lambda](#) 和 [Amazon Simple Storage Service](#)（Amazon S3）。尽可能将您的应用程序容器化，并使用 AWS 托管的容器服务，例如 [Amazon Elastic Container Service](#)（Amazon ECS）或 [Amazon Elastic Kubernetes Service](#)（Amazon EKS）。

使用开源软件或没有许可证费用的软件，尽可能减少许可证成本，例如，对计算工作负载使用 Amazon Linux 或将数据库迁移到 Amazon Aurora。

您可以使用无服务器或应用程序级服务，如 [Lambda](#)、[Amazon Simple Queue Service \(Amazon SQS\)](#)、[Amazon SNS](#) 和 [Amazon SES](#)。这些服务剔除了管理资源的需要，并提供代码执行、排队服务和消息传递功能。另一个好处是，这些服务可以根据使用量扩展性能和成本，从而实现有效的成本分配和归属。

还可以通过无服务器服务来使用 [事件驱动型架构](#)。事件驱动型架构是推送式的，所以当事件在路由器中出现时，一切都按需发生。这样，您就不会为检查事件的连续轮询而付费。这意味着更少的网络带宽消耗、更低的 CPU 使用率、更少的闲置实例集容量，以及更少的 SSL/TLS 握手次数。

有关无服务器的更多信息，请参阅 [架构完善的无服务器应用程序剖析白皮书](#)。

## 实施步骤

- 选择每个服务以优化成本：使用经过优先级排序的列表和分析，选择最符合组织优先事项的每个选项。与其增加容量来满足需求，不如考虑其他可以提供更高性价比的选项。例如，您需要审查您的数据库在 AWS 上的预期流量，并考虑增加实例大小或使用 Amazon ElastiCache 服务（Redis 或 Memcached），来为您的数据库提供缓存机制。
- 评估事件驱动型架构：通过使用无服务器架构，您还可以为基于微服务的分布式应用程序构建事件驱动型架构，这有助于您构建可扩展、有弹性、敏捷且具有成本效益的解决方案。

## 资源

相关文档：

- [AWS 总拥有成本 \(TCO\) 计算器](#)
- [AWS 无服务器](#)
- [什么是事件驱动型架构？](#)
- [Amazon S3 存储类](#)
- [云产品](#)
- [Amazon ElastiCache \(Redis OSS\)](#)

## 相关示例：

- [Getting started with event-driven architecture](#)
- [事件驱动型架构](#)
- [How Statsig runs 100x more cost-effectively using Amazon ElastiCache \(Redis OSS\)](#)
- [Best practices for working with AWS Lambda functions](#)

## COST05-BP06 对不同时间的不同使用情况执行成本分析

工作负载可能会随时间而变化。某些服务或功能在不同的使用水平下更具成本效益。通过根据每个组件在一段时间内的预期使用情况执行分析，工作负载可在其生命周期内保持成本效益。

在未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

随着 AWS 发布新的服务和功能，适用于您的工作负载的最佳服务可能会发生变化。所需的工作量应反映出可能带来的好处。工作负载审核频率取决于您的组织要求。如果工作负载的成本很高，则尽早实施新服务可最大限度地节省成本，因此提高审核频率可能是有利的。审核的另一个触发因素是使用模式发生变化。使用量发生重大变化可能表明备用服务更加理想。

如果您需要将数据移入 AWS Cloud，可以选择 AWS 提供的任何种类的服务和合作伙伴工具，来帮助您迁移数据集，无论它们是文件、数据库、机器镜像、数据块卷，还是磁带备份。例如，要将大量数据移入和移出 AWS，或者要在边缘处理数据，您可以使用 AWS 专用设备之一，经济高效地离线移动 PB 级数据。再例如，对于更高的数据传输率，直接连接服务可能比 VPN 更便宜，而 VPN 可以为您的业务提供所需的稳定一致的连接。

根据一段时间内针对不同使用情况的成本分析，审查您的扩缩活动。分析结果，看看是否可以调整扩缩策略，以增加具有多种实例类型和购买选项的实例。审查您的设置，看看是否可以减小最小值，用较小的实例集大小满足用户请求，并增加更多资源来满足预期的高需求。

通过与组织中的利益相关者讨论，针对一段时间内的不同使用情况进行成本分析，并使用 [AWS Cost Explorer](#) 的预测功能来预测服务变化的潜在影响。使用 AWS Budgets、CloudWatch 计费警报和 AWS Cost Anomaly Detection 监视使用水平触发因素，以尽早确定和实施最具成本效益的服务。

### 实施步骤

- 定义预计使用模式：与组织中的相关人员（例如市场营销部门和产品负责人）合作，记录哪些预期和预计使用模式适用于工作负载。与业务利益相关者讨论历史和预测的成本及使用量的增加，并确保增



加幅度与业务需求一致。确定预计会有更多用户使用您的 AWS 资源的日历日、周或月，这表明您应该增加现有资源的容量，或采用额外的服务来降低成本和提高性能。

- 对预测的使用情况执行成本分析：使用定义的使用模式，对每个点执行分析。分析工作量应该体现潜在的成果，例如，如果使用情况变化很大，应执行彻底分析，以验证任何成本和变化。换句话说，当成本增加时，企业的使用量也应该增加。

## 资源

### 相关文档：

- [AWS 总拥有成本 \( TCO \) 计算器](#)
- [Amazon S3 存储类](#)
- [云产品](#)
- [Amazon EC2 Auto Scaling](#)
- [云数据迁移](#)
- [AWS Snow Family](#)

### 相关视频：

- [AWS OpsHub for Snow Family](#)

## COST 6.在选择资源类型、规模和数量时，如何实现成本目标？

确保选择适合当前任务的资源规模和资源数量。选择最经济实惠的资源类型、规模和数量可以尽可能减少浪费。

### 最佳实践

- [COST06-BP01 执行成本建模](#)
- [COST06-BP02 根据数据选择资源类型、规模和数量](#)
- [COST06-BP03 根据指标自动选择资源类型、规模和数量](#)

### COST06-BP01 执行成本建模

确定组织要求（如业务需求和现有承诺），并对工作负载及其每个组件进行成本建模（总体成本）。对不同预计负载下的工作负载执行基准测试活动，并比较成本。建模工作量应该体现可能带来的好处，例如花费的时间与组件成本成正比。



在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

对工作负载及其每个组件执行成本建模，以了解资源之间的平衡，并在给定的具体性能水平下，确定工作负载中各项资源的适当规模。在评估计划的工作负载部署的价值实现结果时，了解成本考虑因素可为您组织的业务案例和决策过程提供依据。

对不同预计负载下的工作负载执行基准测试活动，并比较成本。建模工作量应该体现可能带来的好处，例如花费的时间与组件成本或预计可节省的成本成正比。有关最佳实践，请参阅 [《AWS Well-Architected Framework – 性能效率支柱》白皮书的“审核”部分](#)。

例如，要为由计算资源组成的工作负载创建成本模型，[AWS Compute Optimizer](#) 可以协助对所运行的工作负载进行成本建模。它根据历史使用量为计算资源提供合理调整大小的建议。确保将 CloudWatch 代理部署到 Amazon EC2 实例，以收集内存指标，这会通过 AWS Compute Optimizer 内更准确的建议为您提供帮助。这是计算资源的理想数据源，因为它是一项免费的服务，并且使用机器学习根据风险等级提出多个建议。

有[多个服务](#)可以与自定义日志一起用作数据源，针对其他服务和工作负载组件合理调整规模，例如 [AWS Trusted Advisor](#)、[Amazon CloudWatch](#) 和 [Amazon CloudWatch Logs](#)。AWS Trusted Advisor 会检查资源并标记出利用率低的资源，这可以帮助您合理调整资源规模，并建立成本模型。

以下是成本建模数据和指标的建议：

- 监控必须准确反映用户体验。为时间段选择正确的粒度，并仔细选择最大值或第 99 个百分位值而不是平均值。
- 为覆盖任何工作负载周期所需的分析时间段选择正确的粒度。例如，如果执行为期两周的分析，您可能会忽略高利用率的月度周期，这可能导致预置不足。
- 通过考虑您现有的承诺、为其他工作负载选择的定价模式，以及更快地创新和专注于核心商业价值的能力，为您计划的工作负载选择合适的 AWS 服务。

## 实施步骤

- 对资源执行成本建模：将工作负载或概念验证部署到具有特定资源类型和规模的单独账户，然后执行测试。使用测试数据运行工作负载，并记录输出结果以及运行测试时段的成本数据。然后，重新部署工作负载或更改资源类型和规模，并再次运行测试。在创建成本模型时，考虑您可能与这些资源一起使用的任何产品的许可证费用，以及用于部署和管理这些资源的估计运营（劳动力或工程师）成本。考虑对一个时间段（每小时、每天、每月、每年或三年）进行成本建模。

## 资源

### 相关文档：

- [AWS Auto Scaling](#)
- [确定合理调整规模的机会](#)
- [Amazon CloudWatch 功能](#)
- [成本优化：合理调整 Amazon EC2 的规模](#)
- [AWS Compute Optimizer](#)
- [AWS 定价计算器](#)

### 相关示例：

- [执行数据驱动的成本建模](#)
- [估算计划的 AWS 资源配置的成本](#)
- [选择合适的 AWS 工具](#)

## COST06-BP02 根据数据选择资源类型、规模和数量

根据工作负载和资源特征的相关数据选择资源规模或类型，例如计算、内存、吞吐量或写入密集型资源。选择的依据通常是使用工作负载的上一个版本（本地版本）、文档或关于工作负载的其他信息源。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

Amazon EC2 提供许多实例类型，它们具有不同的 CPU、内存、存储和联网容量级别，适合不同的使用案例。这些实例类型具有不同的 CPU、内存、存储和联网容量组合，让您在为项目选择正确的资源组合时有更多的备选方案。每种实例类型都有多种大小，因此您可以根据工作负载的需求调整资源。要确定您需要哪种实例类型，请根据您计划在实例上运行的应用程序或软件，收集相关的系统要求详细信息。这些详细信息应包括以下各项：

- 操作系统
- CPU 内核数量
- GPU 内核数
- 系统内存 (RAM) 容量

- 存储类型和空间
- 网络带宽要求

确定计算要求的目的以及需要哪个实例，然后探索各种 Amazon EC2 实例系列。Amazon 提供以下实例类型系列：

- 通用型
- 计算优化型
- 内存优化型
- 存储优化型
- 加速计算型
- HPC 优化型

要更深入地了解特定 Amazon EC2 实例系列适用的特定目的和使用案例，请参阅 [AWS 实例类型](#)。

想要选择最能满足您需求的特定实例系列和实例类型，系统要求的收集至关重要。实例类型名称由系列名称和实例大小组成。例如，t2.micro 实例属于 T2 系列，而且是微型实例。

根据工作负载和资源特征（例如计算、内存、吞吐量或写入密集型）选择资源规模或类型。选择的依据通常是使用成本建模、工作负载的上一个版本（例如本地版本）、文档或关于工作负载的其他信息源（白皮书或发布的解决方案）。使用 AWS 定价计算器或成本管理工具有助于就实例类型、规模和配置作出明智的决策。

## 实施步骤

- 根据数据选择资源：使用您的成本建模数据来选择预期的工作负载使用水平，然后选择指定的资源类型和规模。根据成本建模数据，确定虚拟 CPU 的数量、总内存 (GiB)、本地实例存储容量 (GB)、Amazon EBS 卷和网络性能级别，同时考虑实例所需的数据传输速率。务必根据详细的分析和准确的数据做出选择，这样不仅能够优化性能，还能有效管理成本。

## 资源

相关文档：

- [AWS 实例类型](#)
- [AWS Auto Scaling](#)
- [Amazon CloudWatch 功能](#)

- [成本优化：合理调整 EC2 的大小](#)

相关视频：

- [为您的工作负载选择正确的 Amazon EC2 实例](#)
- [合理调整服务规模](#)

相关示例：

- [让发现和比较 Amazon EC2 实例类型变得更容易](#)

### COST06-BP03 根据指标自动选择资源类型、规模和数量

使用当前运行的工作负载的指标选择正确的规模和类型，从而优化成本。为计算、存储、数据和网络服务适当地预置吞吐量、规模和存储。这可以通过自动扩展等反馈环路进行，也可以在工作负载中使用自定义代码来实现。

在未建立这种最佳实践的情况下暴露的风险等级：低

#### 实施指导

在工作负载中创建一个反馈循环，此循环使用正在运行的工作负载中的活动指标来对该工作负载进行更改。您可以使用托管服务（例如 [AWS Auto Scaling](#)），您可以配置该服务来执行正确的规模调整操作。AWS 还提供了 [API、软件开发工具包](#) 和功能，能够以最小的工作量修改资源。您可以对工作负载进行编程以停止和启动 Amazon EC2 实例，从而允许更改实例大小或实例类型。这带来双重好处：既合理调整了大小，又几乎消除了进行更改所需的全部运营成本。

某些 AWS 服务内置了自动类型或大小选择功能，例如 [Amazon Simple Storage Service 智能分层](#)。Amazon S3 智能分层会根据您的使用情况模式，自动在两个访问层之间移动数据：频繁访问和不频繁访问。

#### 实施步骤

- 通过配置工作负载指标来提高可观测性：捕获工作负载的关键指标。这些指标指明了客户体验（例如工作负载输出），并适应资源类型和规模之间的差异（例如 CPU 和内存使用情况）。对于计算资源，请分析性能数据以适当调整 Amazon EC2 实例大小。识别空闲实例和未充分利用的实例。要查找的关键指标是 CPU 使用率和内存利用率（例如，按照[在启用 AWS Compute Optimizer 和内存利用率的情况下调整大小](#)中的说明，在 90% 的时间内 40% 的 CPU 利用率）。确定四周内最大 CPU 使用率和内存利用率低于 40% 的实例。这些是大小适当能够降低成本的实例。对于存储资源（例如

Amazon S3 )，您可以使用 [Amazon S3 Storage Lens](#)，它允许您在存储桶级别查看各个类别的 28 个指标，默认情况下可在控制面板中查看 14 天的历史数据。您可以按摘要和成本优化或事件筛选 Amazon S3 Storage Lens 控制面板，以分析特定指标。

- 查看合理调整大小建议：使用成本管理控制台中的 AWS Compute Optimizer 和 Amazon EC2 合理调整大小工具中的合理调整大小建议，或查看 AWS Trusted Advisor 如何合理调整资源大小，以对工作负载进行调整。在合理调整不同资源的大小时，请务必使用[正确的工具](#)，并遵循[合理调整大小准则](#)，无论是 Amazon EC2 实例、AWS 存储类，还是 Amazon RDS 实例类型。对于存储资源，您可以使用 Amazon S3 Storage Lens，它可以让您了解对象存储使用情况、活动趋势，并提出可行的建议，以优化成本并应用数据保护最佳实践。使用 [Amazon S3 Storage Lens](#) 从整个组织的指标分析中得出的上下文建议，您可以立即采取措施来优化存储。
- 根据指标自动选择资源类型和规模：使用工作负载指标，手动或自动选择工作负载资源。对于计算资源，配置 AWS Auto Scaling 或在应用程序中实施代码可以减少频繁更改所需的工作量，而且实现更改的速度可能比手动操作更快。您可以在单个 Auto Scaling 组中启动并自动扩展按需型实例和竞价型实例的实例集。除了获得使用竞价型实例的折扣外，您还可以使用预留实例或 Savings Plans 来享受常规按需型实例定价的折扣率。将所有这些因素相结合，可帮助您优化 Amazon EC2 实例的成本节约，并确定应用程序所需的规模和性能。您还可以在 [Auto Scaling 组 \( ASG \)](#) 中使用[基于属性的实例类型选择 \( ABS \)](#) 策略，该策略允许您将实例要求表示为一组属性，例如 vCPU、内存和存储。您可以在发布新一代实例类型时自动使用它们，并通过 Amazon EC2 竞价型实例访问更广泛的容量。Amazon EC2 实例集和 Amazon EC2 Auto Scaling 会选择并启动符合指定属性的实例，无需手动选择实例类型。对于存储资源，您可以使用 [Amazon S3 Intelligent Tiering](#) 和 [Amazon EFS 不频繁访问](#) 功能，这些功能允许您自动选择存储类，以便在数据访问模式更改时自动节省存储成本，而不会影响性能或增加运营开销。

## 资源

### 相关文档：

- [AWS Auto Scaling](#)
- [AWS 合理调整大小](#)
- [AWS Compute Optimizer](#)
- [Amazon CloudWatch 功能](#)
- [设置 CloudWatch](#)
- [CloudWatch 发布自定义指标](#)
- [开始使用 Amazon EC2 Auto Scaling](#)
- [Amazon S3 Storage Lens](#)

- [Amazon S3 智能分层](#)
- [Amazon EFS 不频繁访问](#)
- [使用 SDK 启动 Amazon EC2 实例](#)

相关视频：

- [适当调整服务大小](#)

相关示例：

- [基于属性为 Amazon EC2 实例集的 Auto Scaling 选择实例类型](#)
- [使用计划的扩缩优化 Amazon Elastic Container Service 成本](#)
- [使用 Amazon EC2 Auto Scaling 的预测性扩缩](#)
- [使用 Amazon S3 Storage Lens 优化成本并提高可见性](#)
- [架构完善的实验室：合理调整大小建议（级别 100）](#)
- [架构完善的实验室：在启用 AWS Compute Optimizer 和内存利用率的情况下合理调整大小（级别 200）](#)

## COST 7.您如何使用定价模式来降低成本？

使用最适合的资源定价模式可以尽可能减少支出。

### 最佳实践

- [COST07-BP01 执行定价模式分析](#)
- [COST07-BP02 根据成本选择区域](#)
- [COST07-BP03 选择具有经济实惠条款的第三方协议](#)
- [COST07-BP04 针对此工作负载的所有组件实施定价模型](#)
- [COST07-BP05 在管理账户级别执行定价模式分析](#)

### COST07-BP01 执行定价模式分析

分析工作负载的每个组件。确定组件和资源是长时间运行（享受承诺折扣），还是短时间动态运行（采用 Spot 或按需定价）。使用成本管理工具中的建议对工作负载执行分析，并对这些建议应用业务规则以实现高回报。



在未建立这种最佳实践的情况下暴露的风险等级：高

## 实施指导

AWS 有多种[定价模式](#)，让您可以通过符合组织需求、最具成本效益的方式支付资源费用，具体取决于产品。与您的团队一起确定最适合的定价模式。您的定价模式通常包含多种选项的组合，根据供应情况来决定。

使用按需型实例时，您按小时或按秒（最少 60 秒）支付计算或数据库容量的费用，具体取决于您运行哪些实例，而无需长期承诺或预付款。

Savings Plans 是一种灵活的定价模式，以低廉的价格使用 Amazon EC2、Lambda 和 AWS Fargate，而无需承诺在一年或三年内保持一致的使用量（以每小时美元金额计算）。

竞价型实例是一种 Amazon EC2 定价机制，允许您以折扣的小时费率（与按需型实例价格相比，最多可减少 90%）申请备用计算容量，无需前期投入。

使用预留实例时，通过预付容量费用，您可以享受高达 75% 的折扣。有关更多详细信息，请参阅[使用预留优化成本](#)。

您可以选择为与生产、质量和开发环境相关的资源包括 Savings Plans。或者，因为沙盒资源仅在需要时才会启用，所以您可以为该环境中的资源选择按需模式。使用亚马逊[竞价型实例](#)降低 Amazon EC2 成本，或使用[计算 Savings Plans](#)降低 Amazon EC2、Fargate 和 Lambda 成本。[AWS Cost Explorer](#) 建议工具通过 Saving Plans 提供承诺折扣的机会。

如果您过去一直为 Amazon EC2 购买[预留实例](#)，或者在组织内确立了成本分配做法，则目前可以继续使用 Amazon EC2 预留实例。但是，我们建议制定一项战略，以便在未来使用 Savings Plans，它是一种更灵活的成本节省机制。您可以在 AWS Cost Management 中刷新 Savings Plans（SP）建议，以便随时生成新的 Savings Plans 建议。使用预留实例（RI）来降低 Amazon RDS、Amazon Redshift、Amazon ElastiCache 和 Amazon OpenSearch Service 成本。Saving Plans 和预留实例提供三个选项：全额预付、部分预付和无预付款。使用 AWS Cost Explorer RI 和 SP 购买建议中提供的建议。

要为 Spot 工作负载寻找机会，请查看总体使用量的小时视图，并确定使用量或弹性的定期变化周期。您可以为各种容错和灵活应用程序使用竞价型实例。示例包括无状态 Web 服务器、API 端点、大数据和分析应用程序、容器化工作负载、CI/CD 及其他灵活工作负载。

分析您的 Amazon EC2 和 Amazon RDS 实例，确定在您不使用时（下班后和周末）是否可以将其关闭。与 24/7 全天候使用这些实例相比，您可以使用这种方法将成本降低 70% 或更多。如果您的 Amazon Redshift 集群只需要在特定时间内可用，则可以暂停集群并在稍后恢复它。停止 Amazon Redshift 集群或 Amazon EC2 和 Amazon RDS 实例时，计算计费会停止，仅收取存储费用。



请注意，[按需容量预留](#) ( ODCR ) 不是一项价格折扣。不管您是否在预留容量中运行实例，容量预留均按照与按需费率同等的费率来收费。当您需要为计划运行的资源提供足够的容量时，应该考虑使用这种模式。ODCR 不需要做出长期承诺，因为当您不再需要时可以取消它们，但它们也可以享受 Savings Plans 或预留实例提供的折扣。

## 实施步骤

- 分析工作负载弹性：在 Cost Explorer 或自定义控制面板中使用每小时粒度，分析您的工作负载的弹性。确定正在运行的实例数量的规律性变化。短期实例比较适合采用竞价型实例或竞价型实例集。
- [Well-Architected 实验室：Cost Explorer](#)
- [Well-Architected 实验室：成本可视化](#)
- 查看现有定价合同：查看当前合同或对长期需求的承诺。分析您目前拥有的承诺以及这些承诺中有多少正在使用。利用已有的合同折扣或企业协议。[企业协议](#)让客户可以选择定制最适合其需求的协议。对于长期承诺，请考虑预留定价折扣、特定实例类型的预留实例或 Savings Plans、实例系列、AWS 区域 和可用区。
- 执行承诺折扣分析：在账户中使用 Cost Explorer 查看 Savings Plans 和预留实例建议。要验证您是否实施了具有所需折扣和风险的正确建议，请按照 [Well-Architected 实验室](#) 操作。

## 资源

### 相关文档：

- [获取预留实例建议](#)
- [实例购买选项](#)
- [AWS Enterprise](#)

### 相关视频：

- [节省高达 90% 并在竞价型实例上运行生产工作负载](#)

### 相关示例：

- [Well-Architected 实验室：Cost Explorer](#)
- [Well-Architected 实验室：成本可视化](#)
- [Well-Architected 实验室：定价模式](#)

## COST07-BP02 根据成本选择区域

资源定价在每个区域中可能各不相同。确定区域成本差异，仅当需要满足延迟、数据驻留和数据主权要求时，才在成本较高的区域部署。考虑区域成本有助于您为此工作负载支付最低的总体费用。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

此 [AWS Cloud 基础设施](#) 是全球性的，托管在 [世界范围内的多个位置](#)，并围绕 AWS 区域、可用区、Local Zone、AWS Outposts 和 Wavelength Zone 构建。区域是世界上的一个物理位置，每个区域都是一个独立的地理位置，而 AWS 有多个可用区。可用区是每个区域内多个相互隔离的位置，由一个或多个独立的数据中心组成，每个数据中心都拥有冗余的电力、联网和连接。

每个 AWS 区域都在当地市场条件下运作，由于土地、光纤、电力和税收等成本的不同，每个区域的资源定价也不同。选择特定区域来运行解决方案组件或整个解决方案，以便您可以在全球范围内以尽可能低的价格运行。使用 [AWS 计算器](#)，按位置类型（区域、Wavelength Zone 和 Local Zone）和区域搜索服务，估算您的工作负载在不同区域的成本。

在架构解决方案时，最佳实践是设法将计算资源放在更接近用户的位置，以提供更低的延迟和强大的数据主权。根据您的业务、数据隐私、性能和安全要求来选择地理位置。对于具有全球终端用户的应用程序，请使用多个位置。

如果您不存在数据隐私、安全和业务要求方面的义务，请使用 AWS 服务价格较低的区域来部署您的工作负载。例如，如果您的默认区域是 ap-southeast-2（悉尼），并且不存在使用其他区域的限制（例如数据隐私、安全），则在 north-east-1（弗吉尼亚州北部）区域部署非关键（开发和测试）Amazon EC2 实例成本更少。

|      | 合规性 | 延迟    | 成本     | 服务/功能 |
|------|-----|-------|--------|-------|
| 区域 1 | ✓   | 15 ms | \$\$   | ✓     |
| 区域 2 | ✓   | 20 ms | \$\$\$ | X     |
| 区域 3 | ✓   | 80 ms | \$     | ✓     |
| 区域 4 | ✓   | 15 ms | \$\$   | ✓     |
| 区域 5 | ✓   | 20 ms | \$\$\$ | X     |
| 区域 6 | ✓   | 15 ms | \$     | ✓     |
| 区域 7 | ✓   | 80 ms | \$     | ✓     |
| 区域 8 | ✓   | 15 ms | \$     | X     |

区域特征矩阵表

前面的矩阵表显示，区域 4 是这种给定场景的最佳选择，因为与其他区域相比，其延迟较低，可以使用相应服务，而且是最便宜的区域。

实施步骤

- 查看 AWS 区域定价：分析当前区域的工作负载成本。首先使用按服务和使用类型划分的最高成本，计算其他可用区域的成本。如果预测的节省超过迁移组件或工作负载的成本，则迁移到新区域。
- 审查多区域部署的要求：分析您的业务要求和义务（数据隐私、安全或性能），以确认是否存在任何会阻止您使用多个区域的限制。如果不存在要求您只能使用单个区域的限制，那么就使用多个区域。
- 分析所需的数据传输：在选择区域时要考虑数据传输成本。让您的数据靠近您的客户和资源。在数据流动和数据传输非常少时，选择成本较低的 AWS 区域。根据您的数据传输的业务要求，可以使用 [Amazon CloudFront](#)、[AWS PrivateLink](#)、[AWS Direct Connect](#)和 [AWS Virtual Private Network](#) 来降低您的联网成本，提高性能，并增强安全性。

资源

相关文档：

- [获取预留实例建议](#)
- [Amazon EC2 定价](#)
- [实例购买选项](#)
- [区域表](#)

相关视频：

- [节省高达 90% 并在竞价型实例上运行生产工作负载](#)

相关示例：

- [常见架构的数据传输成本概述](#)
- [全球部署的成本考虑因素](#)
- [为工作负载选择区域时应考虑的事项](#)
- [Well-Architected 实验室：按区域限制服务使用（级别 200）](#)

## COST07-BP03 选择具有经济实惠条款的第三方协议

经济实惠的协议和条款可确保这些服务的成本与所提供的效益相称。选择与可为组织带来额外效益相称的协议和定价。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

市场上有多种产品可用于管理云环境中的成本。根据客户需求，它们在功能方面可能有一些差异，例如有些侧重于成本治理或成本可见性，而有些则侧重于成本优化。想要得到有效的成本优化和治理，一个关键因素是使用具有必要功能的正确工具和正确的定价模型。这些产品有不同的定价模型。有些产品按每月账单的一定比例收费，有些则按所实现节省额的一定比例收费。理想情况下，您应该只为您需要的资源付费。

当您在云中使用时，确保定价结构契合期望结果非常重要。定价应与其带来的结果和价值成比例。例如，可带来一定百分比节省额的软件中，节省额（结果）越高，其价格也就越高。许可协议规定，随着开支的增加，您需要支付更多的费用，这可能并不总是有利于您优化成本。但是，如果供应商为您的账单的所有部分都提供明显的优惠，那么这种按比例收费可能是合理的。

例如，如果您使用的其他服务没有带来任何好处，提供 Amazon EC2 相关建议并按整体账单一定比例收取费用的解决方案将会变得更加昂贵。另一个示例是根据托管资源的成本，按一定百分比收费的托管

服务。实例越大并不一定意味着需要更多的管理工作，但会收取更多费用。验证这些服务定价安排是否在其服务中包含了成本优化计划或功能，以提高效率。

客户可能会发现市场上的这些产品更先进或更易于使用。您需要考虑这些产品的成本，并考虑长期潜在的成本优化结果。

## 实施步骤

- 分析第三方协议和条款：查看第三方协议中的定价。基于不同的使用情况水平执行建模，并将新成本（例如使用新服务，或当前服务由于工作负载增长导致使用量增加）纳入考量。确定额外成本能否为业务提供所需效益。

## 资源

相关文档：

- [获取预留实例建议](#)
- [实例购买选项](#)

相关视频：

- [节省高达 90% 并在竞价型实例上运行生产工作负载](#)

## COST07-BP04 针对此工作负载的所有组件实施定价模型

永久运行的资源应利用预留容量，如Savings Plans或预留实例。短期容量配置为使用竞价型实例或竞价型实例集。按需型实例仅用于无法中断并且运行时长不足以使用预留容量的短期工作负载（根据具体的资源类型，时长介于 25% 到 75% 之间）。

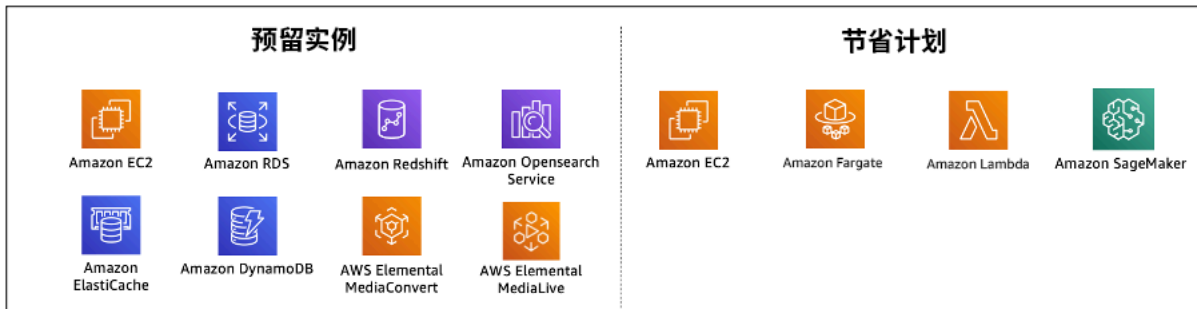
在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

为了提高成本效率，AWS 根据您过去的使用情况提供多项承诺建议。您可以使用这些建议，了解可以节省的金额以及如何使用承诺。您使用这些服务的方式可以是按需型实例、竞价型实例，或在一定时间内承诺使用量，通过预留实例（RI）和Savings Plans（SP）降低按需成本。您不仅需要了解每个工作负载组件和多项 AWS 服务，还需要了解这些服务的承诺折扣、购买选项和竞价型实例，以优化您的工作负载。

考虑您的工作负载组件的要求，并了解这些服务的不同定价模型。定义这些组件的可用性要求。确定工作负载中是否存在执行功能的多个独立资源，以及工作负载随着时间推移的需求情况。使用默认的按需定价模型和其他适用模型比较资源成本。考虑资源或工作负载组件的任何潜在更改。

例如，让我们看一下 AWS 上的这个 Web 应用程序架构。此示例工作负载由多个 AWS 服务组成，例如 Amazon Route 53、AWS WAF、Amazon CloudFront、Amazon EC2 实例、Amazon RDS 实例、负载均衡器、Amazon S3 存储和 Amazon Elastic File System ( Amazon EFS )。您需要审查每项服务，并确定使用不同定价模型带来的潜在成本节省机会。其中一些服务可能有资格使用 RI 或 SP，而另一些可能只能使用按需方案。如下图所示，一些 AWS 服务可以使用 RI 或 SP 承诺。



## 使用预留实例和Savings Plans承诺的 AWS 服务

### 实施步骤

- **实施定价模型：**根据分析结果，购买Savings Plans、预留实例或实施竞价型实例。如果是首次承诺购买，请选择列表中的前 5 个或 10 个建议，然后在接下来的一到两个月内监控并分析结果。AWS Cost Management Console会指导您完成整个过程。查看控制台中的 RI 或 SP 建议，自定义建议（类型、付款和期限），查看每小时承诺（例如每小时 20 USD），然后添加到购物车。折扣将自动应用于符合条件的使用量。定期购买少量承诺折扣（例如每 2 周或每月一次）。对可以中断或者无状态的工作负载实施竞价型实例。最后，选择按需型 Amazon EC2 实例并为其需求分配资源。
- **工作负载审核周期：**实施工作负载审核周期，用于专门分析定价模型覆盖范围。工作负载达到所需覆盖范围后，可部分（每隔几个月）购买额外的承诺折扣，或者随着组织的使用情况变化进行购买。

### 资源

#### 相关文档：

- [了解您的 Savings Plans 建议](#)
- [访问预留实例建议](#)
- [如何购买预留实例](#)
- [实例购买选项](#)

- [竞价型实例](#)
- [其他 AWS 服务的预留模式](#)
- [Savings Plans支持的服务](#)

相关视频：

- [节省高达 90% 并在竞价型实例上运行生产工作负载](#)

相关示例：

- [在购买Savings Plans之前我应该考虑什么？](#)
- [如何使用 Cost Explorer 来分析我的支出和使用情况？](#)

## COST07-BP05 在管理账户级别执行定价模式分析

检查账单和成本管理工具，并查看承诺和预留的建议折扣，以便在管理账户级别执行定期分析。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

定期执行成本建模帮助您跨多个工作负载进行优化。例如，如果总体上多个工作负载使用按需型实例，则变更的风险较低，并且实施基于承诺的折扣可降低总体成本。建议每两周到一个月定期执行一次分析。这样您就可以进行少量调整性采购，因此定价模型的覆盖范围会随着工作负载及其组件的变化而不断变化。

使用 [AWS Cost Explorer](#) 建议工具，寻找在您的管理账户中享受承诺折扣的机会。在计算管理账户级别的建议时，考虑 AWS 组织中启用了预留实例（RI）或 Savings Plans（SP）的所有账户的使用情况。它们也是在启用折扣共享的情况下计算的，以便推荐可在不同账户中最大限度实现节省的承诺。

虽然在许多情况下，在管理账户级别购买可以优化以最大限度地节省开支，但在某些情况下，您可能会考虑在关联账户级别购买 SP，例如当您希望首先对该特定关联账户中的资源使用量应用折扣时。在个人账户级别计算成员账户建议，以便为每个独立账户实现最大限度的节省。如果您的账户同时拥有 RI 和 SP 承诺，则它们将按以下顺序应用：

1. 区域 RI
2. 标准 RI
3. 可转换 RI



## 4. 实例 Savings Plan

## 5. 计算 Savings Plan

如果您在管理账户级别购买 SP，则将根据从最高到最低的折扣百分比来应用实惠。管理账户级别的 SP 会查看所有关联账户，并在任何可以实现最高折扣的地方应用实惠。如果您希望限制应用实惠的位置，则可以在关联账户级别购买 Savings Plan，而只要该账户在运行符合条件的计算服务，就会首先在其上应用折扣。当账户未运行符合条件的计算服务时，将在同一管理账户下的其他关联账户之间共享折扣。默认情况下，折扣共享处于启用状态，不过您可以根据需要关闭。

在整合账单系列中，Savings Plans 首先应用于拥有者账户的使用量，然后应用于其他账户的使用量。只有当您启用了共享时才会出现这种情况。您的 Savings Plans 将首先应用到能够实现最高节省百分比的地方。如果多个使用量具有相同的节省百分比，则 Savings Plans 将应用于 Savings Plans 比率最低的第一个使用量。Savings Plans 将继续应用，直到没有剩余的使用量或您的承诺用量用完为止。剩余的所有使用量均按照按需费率收取费用。您可以在 AWS Cost Management 中刷新 Savings Plans 建议，以便随时生成新的 Savings Plans 建议。

分析实例的灵活性之后，您可以按照建议进行承诺。创建成本模型，使用可能的不同资源选项分析工作负载的短期成本、分析 AWS 定价模式并使其与您的业务需求保持一致，发现总拥有成本和 [成本优化](#) 机会。

### 实施步骤

执行承诺折扣分析：在账户中使用 Cost Explorer 查看 Savings Plans 和预留实例建议。确保了解实惠配套建议，估算每月支出以及每月节省。查看管理账户级别的建议，这些建议根据 AWS 组织中启用了 RI 或 Savings Plans 折扣共享的所有成员账户的使用情况计算得出，以便各账户实现最大限度的节省。您可以按照 Well-Architected 实验室操作，确认您实施了具有所需折扣和风险的正确建议。

### 资源

相关文档：

- [AWS 定价如何运作？](#)
- [实例购买选项](#)
- [实惠配套概览](#)
- [实惠配套建议](#)
- [获取预留实例建议](#)
- [了解您的 Saving Plans 建议](#)
- [Savings Plans 如何应用于您的 AWS 使用](#)

- [实惠配套与整合账单](#)
- [启用共享预留实例和Savings Plans折扣](#)

相关视频：

- [节省高达 90% 并在竞价型实例上运行生产工作负载](#)

相关示例：

- [AWS Well-Architected 实验室：定价模式（级别 200）](#)
- [AWS Well-Architected 实验室：定价模式分析（级别 200）](#)
- [在购买 Savings Plan 之前我应该考虑什么？](#)
- [如何使用滚动 Savings Plans 来降低承诺风险？](#)
- [何时使用竞价型实例](#)

## COST 8. 您如何规划数据传输费用？

确保规划并监护您的数据传输费用，以便制定架构决策，尽可能降低成本。持续以小步迭代的方式进行架构优化可以实现运营成本的大幅降低。

最佳实践

- [COST08-BP01 执行数据传输建模](#)
- [COST08-BP02 选择组件以便优化数据传输成本](#)
- [COST08-BP03 实施服务以便降低数据传输成本](#)

### COST08-BP01 执行数据传输建模

收集组织要求，并对工作负载及其每个组件执行数据传输建模。这样可以确定满足当前数据传输要求的最低成本点。

在未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

在云端设计解决方案时，由于习惯于设计使用本地数据中心的架构或者缺乏相应的知识，数据传输费用常常会被忽视。AWS 中的数据传输费用由流量的来源、目的地和数量决定。在设计阶段将这些费用考

虑在内可以节省成本。了解工作负载中的哪些环节需要进行数据传输、传输成本及其相关好处，对于准确估算总拥有成本 ( TCO ) 非常重要。因此，您可以作出明智的决定来修改或接受架构决策。例如，您可能有一个多可用区配置，可以在可用区之间复制数据。

您可以对在工作负载中传输数据的服务组件进行建模，并确定这是可接受的成本（类似于在两个可用区中支付计算和存储费用），以实现所需的可靠性和韧性。对不同使用级别的成本进行建模。工作负载的使用量可能随时间而变化，不同的服务可能在不同的级别上更具有成本效益。

在对数据传输进行建模时，须考虑接收了多少数据以及这些数据来自哪里。此外，还须考虑处理了多少数据以及需要多少存储或计算容量。在建模期间，应遵循工作负载架构的联网最佳实践，以优化潜在的数据传输成本。

可利用 AWS 定价计算器 查看特定 AWS 服务和预期数据传输的估计成本。如果您的工作负载已经在运行（用于测试目的或在预生产环境中），请使用 [AWS Cost Explorer](#) 或 [AWS 成本和使用情况报告](#) ( CUR ) 来了解您的数据传输成本并对其进行建模。配置概念证明 (PoC) 或测试您的工作负载，并在实际的模拟负载下运行测试。您可以根据不同的工作负载需求对成本进行建模。

## 实施步骤

- 确定需求：计划在来源和目的地之间传输数据的主要目标和业务需求是什么？最后预期的业务成果是什么？收集业务需求并定义预期结果。
- 确定来源和目的地：数据传输的数据来源和目的地是什么，例如是在 AWS 区域内传输、传入 AWS 服务还是传出到互联网？
  - [AWS 区域内的数据传输](#)
  - [AWS 区域之间的数据传输](#)
  - [将数据传输到互联网](#)
- 识别数据分类：此数据传输的数据分类是什么？这是什么样的数据？数据量有多大？必须多久传输一次数据？数据是否敏感？
- 确定要使用的 AWS 服务或工具：此数据传输使用哪些 AWS 服务？是否可以将已经预置的服务用于其他工作负载？
- 计算数据传输成本：结合使用 [AWS 定价](#) 和您之前创建的数据传输建模，计算工作负载的数据传输成本。计算不同使用情况水平的数据传输成本，包括工作负载使用量的增加和减少这两种情况。如果工作负载架构有多个传输选项，则计算每个选项的成本以便进行比较。
- 将成本与成果相关联：对于产生的每项数据传输成本，明确要实现的工作负载成果。如果在组件之间传输，可能是为了实现解耦；如果在可用区之间传输，则可能是为了实现冗余。
- 创建数据传输建模：收集所有信息后，为多个使用案例和不同工作负载创建概念性基础数据传输建模。

## 资源

### 相关文档：

- [AWS 缓存解决方案](#)
- [AWS 定价](#)
- [Amazon EC2 定价](#)
- [Amazon VPC 定价](#)
- [了解数据传输费用](#)

### 相关视频：

- [监控和优化数据传输成本](#)
- [S3 Transfer Acceleration](#)

### 相关示例：

- [常见架构的数据传输成本概述](#)
- [AWS 规范性指南 - 联网](#)

## COST08-BP02 选择组件以便优化数据传输成本

选择所有组件然后设计架构，以便降低数据传输成本。其中包括使用广域网（WAN）优化和多可用区（AZ）配置等组件

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

针对数据传输进行构造，可最大限度地降低数据传输成本。这可能涉及使用内容分发网络来定位更靠近用户的数据，或者使用从您的本地设施到 AWS 的专用网络链接。您还可以使用 WAN 优化和应用程序优化来减少组件之间传输的数据量。

向 AWS Cloud 传输数据或在其中传输数据时，必须根据不同的使用案例、数据的性质和可用的网络资源来了解目的地，以便选择正确的 AWS 服务来优化数据传输。AWS 提供一系列针对不同数据迁移要求量身定制的数据传输服务。根据组织内部的业务需求，选择正确的[数据存储](#)和[数据传输](#)选项。

在规划或审查您的工作负载架构时，请考虑以下几点：

- 在 AWS 内使用 VPC 端点：VPC 端点允许在您的 VPC 和支持的 AWS 服务之间建立私有连接。这使您可以避免使用公共互联网，而使用公共互联网会产生数据传输成本。
- 使用 NAT 网关：使用 [NAT 网关](#)，以便私有子网中的实例可以连接到互联网或 VPC 外部的服务。检查发送流量最多的 NAT 网关后面的资源是否与 NAT 网关位于同一可用区。如果未在同一可用区，则在与资源相同的可用区中创建新的 NAT 网关，从而降低跨可用区传输数据的费用。
- 使用 AWS Direct Connect AWS Direct Connect 会绕过公共互联网，在您的本地网络和 AWS 之间直接建立私有连接。与通过互联网传输大量数据相比，这可能更具成本效益且更加一致。
- 避免跨区域边界传输数据：在 AWS 区域 之间（从一个区域到另一个区域）传输数据通常会产生费用。采用多区域方法应是一项深思熟虑的决定。有关更多详细信息，请参阅[多区域场景](#)。
- 监控数据传输：使用 Amazon CloudWatch 和 [VPC 流日志](#) 来捕获有关您的数据传输和网络使用情况的详细信息。分析 VPC 中捕获的网络流量信息，例如进出网络接口的 IP 地址或范围。
- 分析您的网络使用情况：使用计量和报告工具（例如 AWS Cost Explorer、CUDOS 控制面板，或 CloudWatch）了解工作负载的数据传输成本。

## 实施步骤

- 选择数据传输组件：使用[COST08-BP01 执行数据传输建模](#)中所述的数据传输建模，关注产生最多数据传输成本之处，或者工作负载使用情况发生变化时产生最多数据传输成本之处。寻找替代架构或其他组件，以消除或减少数据传输的需要（或降低其成本）。

## 资源

### 相关最佳实践：

- [COST08-BP01 执行数据传输建模](#)
- [COST08-BP03 实施服务以便降低数据传输成本](#)

### 相关文档：

- [云数据迁移](#)
- [AWS 缓存解决方案](#)
- [使用 Amazon CloudFront 更快地交付内容](#)

### 相关示例：

- [常见架构的数据传输成本概述](#)
- [AWS 网络优化技巧](#)
- [使用 Apache Parquet 格式的 VPC 流日志优化性能并降低网络分析成本](#)

## COST08-BP03 实施服务以便降低数据传输成本

实施服务以减少数据传输。例如，使用边缘站点或内容分发网络 (CDN) 向最终用户交付内容，在应用程序服务器或数据库前构建缓存层，并使用专用网络连接而不是 VPN 连接到云端。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

您可以利用多种 AWS 服务来优化网络数据传输使用情况。根据工作负载组件、类型和云架构，这些服务有助于您在云端压缩、缓存、共享和分发流量。

- [Amazon CloudFront](#) 是一个全球内容分发网络，可提供低延迟、高传输速度的数据。它在世界各地的边缘站点缓存数据，从而减少资源负担。通过使用 CloudFront，您可以减少向全球大量用户分发内容的管理工作，同时将延迟降到最低。此 [安全实惠服务包](#) 有助于您节省高达 30% 的 CloudFront 使用量 - 如果您计划随着时间的推移增加使用量。
- [AWS Direct Connect](#) 允许您建立到 AWS 的专用网络连接。与基于互联网的连接相比，这可以降低网络成本、增加带宽并提供更一致的网络体验。
- [AWS VPN](#) 可让您在专用网络和 AWS 全球网络之间建立安全的专用连接。此网络是小型办公室或业务合作伙伴的理想之选，因其提供简化的连接，并且是完全托管的弹性服务。
- [VPC 端点](#) 允许通过专用网络在 AWS 服务之间建立连接，可用于减少公共数据传输和 [NAT 网关](#) 成本。[网关 VPC 端点](#) 不按小时收费，支持 Amazon S3 和 Amazon DynamoDB。[接口 VPC 端点](#) 由 [AWS PrivateLink](#) 提供，有小时费和每 GB 使用成本。
- [NAT 网关](#) 与独立的 NAT 实例相比，提供内置的扩展和管理功能，可降低成本。将 NAT 网关放置在与高流量实例所在的同一可用区中，并考虑对需要访问 Amazon DynamoDB 或 Amazon S3 的实例使用 VPC 端点，以降低数据传输和处理成本。
- 使用具有计算资源的 [AWS Snow Family](#) 设备在边缘收集和处理数据。AWS Snow Family 设备 ( [Snowball Edge](#)、[Snowball Edge](#) 和 [Snowmobile](#) ) 让您以经济高效地将 PB 级数据离线迁移到 AWS Cloud。

## 实施步骤

- 实施服务：使用数据传输建模并查看 VPC 流日志，根据服务和工作负载类型选择适用的 AWS 网络服务。了解产生最多成本和最多数据流之处。了解 AWS 服务并评测是否存在一种服务，可以减少或消除传输，特别是联网和内容交付。如有需要重复访问数据或存在大量数据的情况，则应了解缓存服务。

## 资源

### 相关文档：

- [AWS Direct Connect](#)
- [AWS 探索我们的产品](#)
- [AWS 缓存解决方案](#)
- [Amazon CloudFront](#)
- [AWS Snow Family](#)
- [Amazon CloudFront 安全实惠服务包](#)

### 相关视频：

- [监控和优化数据传输成本](#)
- [AWS Cost Optimization Series: CloudFront](#)
- [如何降低我的 NAT 网关的数据传输费用？](#)

### 相关示例：

- [How-to chargeback shared services: An AWS Transit Gateway example](#)
- [Understand AWS data transfer details in depth from cost and usage report using Athena query and QuickSight](#)
- [Overview of Data Transfer Costs for Common Architectures](#)
- [Using AWS Cost Explorer to analyze data transfer costs](#)
- [Cost-Optimizing your AWS architectures by utilizing Amazon CloudFront features](#)
- [如何降低我的 NAT 网关的数据传输费用？](#)



## 管理需求和供应资源

### 问题

- [COST 9. 如何管理需求和供应资源？](#)

### COST 9. 如何管理需求和供应资源？

为了工作负载的性能与支出实现平衡，请确保您支付过费用的所有资源都得到利用，并避免出现资源利用率过低的情况。无论是从运营成本（由于过度使用导致性能下降）还是从浪费 AWS 支出（由于超额配置）的角度衡量，利用率指标过高或过低都会对您的组织产生负面影响。

### 最佳实践

- [COST09-BP01 对工作负载需求执行分析](#)
- [COST09-BP02 实施缓冲区或节流来管理需求](#)
- [COST09-BP03 动态供应资源](#)

### COST09-BP01 对工作负载需求执行分析

分析工作负载需求随时间的变化。确认分析涵盖季节性趋势，并准确反映整个工作负载生命周期内的运行条件。分析工作应该体现出可能带来的好处，例如花费的时间与工作负载成本成正比。

未建立这种最佳实践的情况下暴露的风险等级：高

### 实施指导

分析云计算的工作负载需求涉及了解在云环境中启动的计算任务的模式和特征。这种分析有助于用户优化资源分配、管理成本并验证性能是否达到要求的水平。

了解工作负载的要求。组织需求应指出工作负载对于请求的响应时间。响应时间可用于确定需求是否得到了管理，或者是否应改变资源供应以满足需求。

分析应包括需求的可预测性和可重复性、需求的变化速率以及需求的变化量。对足够长的时间执行分析，以纳入任何季节性变化，例如月末处理或假期高峰。

分析工作应反映实施扩展的潜在好处。查看组件的预期总成本，以及在工作负载生命周期内使用量和成本的任何增加和减少。

以下是执行云计算的工作负载需求分析时，需要考虑的一些关键方面：

1. 资源利用率和性能指标：分析一段时间内 AWS 资源的使用情况。确定高峰和非高峰期使用模式，以优化资源分配和扩展策略。监控性能指标，如响应时间、延迟、吞吐量和错误率。这些指标有助于评测云基础设施的整体运行状况和效率。
2. 用户和应用程序扩展行为：了解用户行为及其对工作负载需求的影响。检查用户流量模式有助于增强内容的交付和应用程序的响应能力。分析工作负载如何随着需求的增长而扩展。确定弹性伸缩参数的配置是否正确有效，以处理负载波动。
3. 工作负载类型识别云中运行的不同类型的工作负载，如批处理、实时数据处理、Web 应用程序、数据库或机器学习。每种类型的工作负载都可能有不同的资源需求和性能特征。
4. 服务水平协议 (SLA)：将实际绩效与 SLA 进行比较，以确保合规性并确定需要改进的领域。

您可以使用 [Amazon CloudWatch](#) 来收集和跟踪指标，监控日志文件，设置警报，并自动应对 AWS 资源变化。您还可以使用 Amazon CloudWatch 了解系统范围内的资源使用率、应用程序性能和运行状况。

利用 [AWS Trusted Advisor](#)，您可以按照最佳实践预置资源，以提高系统性能和可靠性，增强安全性，并寻找节省资金的机会。您也可以关闭非生产实例，并使用 Amazon CloudWatch 和 Auto Scaling 来匹配需求的增加或减少。

最后，您可以将 [AWS Cost Explorer](#) 或 [QuickSight](#) 与 AWS 成本和使用情况报告 (CUR) 文件或应用程序日志一起使用，对工作负载需求执行高级分析。

总之，通过全面的工作负载需求分析，组织可以就资源预置、扩展和优化做出明智的决策，从而提高性能、成本效益和用户满意度。

## 实施步骤

- 分析现有工作负载数据：分析现有工作负载中的数据、以前工作负载版本中的数据或预测使用模式中的数据。使用 Amazon CloudWatch、日志文件和监控数据来深入了解工作负载的使用情况。分析整个工作负载周期，并收集任何季节性变化数据，如月末或年末活动。分析中反映的工作应该体现出工作负载特征。应将最多的精力放在需求变化最大的高价值工作负载上。应将最少的精力放在需求变化最小的低价值工作负载上。
- 预测外部影响：与组织中会影响或更改工作负载需求需求的团队成员会面。通常涉及的团队包括销售、营销或业务拓展团队。与他们合作，了解其运作周期，以及是否有改变工作负载需求的任何活动。使用这些数据预测工作负载需求。

## 资源

相关文档：

- [Amazon CloudWatch](#)
- [AWS Trusted Advisor](#)
- [AWS X-Ray](#)
- [AWS Auto Scaling](#)
- [AWS Instance Scheduler](#)
- [开始使用 Amazon SQS](#)
- [AWS Cost Explorer](#)
- [QuickSight](#)

相关视频：

相关示例：

- [监控、跟踪和分析以实现成本优化](#)
- [Searching and analyzing logs in CloudWatch](#)

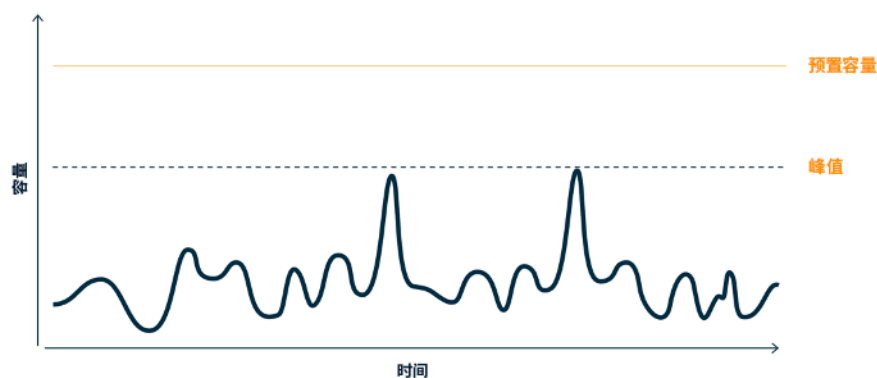
## COST09-BP02 实施缓冲区或节流来管理需求

缓冲区和节流可修改工作负载需求，从而避免出现任何峰值情形。在客户端执行重试时实施节流。实施缓冲以存储请求并将处理任务往后推迟一段时间。确认设计节流和缓冲区时客户端能够在所需的时间内收到响应。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

在云计算领域，实施缓冲区或节流对于管理需求和减少工作负载所需的预置容量至关重要。为了获得最佳性能，必须衡量总需求，包括峰值、请求变化的速度和必需的响应时间。当客户端能够重新发送请求时，应用节流比较切实可行。相反，对于缺乏重试功能的客户端，理想的方法是实施缓冲区解决方案。此类缓冲区可舒缓请求的涌入，并优化具有不同操作速度的应用程序之间的交互。



需求曲线具有两个不同的峰值，需要高预置容量

假设工作负载的需求曲线如上图所示。此工作负载有两个峰值，为了处理这些峰值，如橙色线所示预置资源容量。因为需要预置容量来处理这两个峰值，所以此工作负载所使用的资源和能源不是由需求曲线下的区域表示，而是由预置容量线下面的区域表示。展平工作负载需求曲线有助于降低工作负载的预置容量和减少对环境的影响。为了平滑峰值，可以考虑实施节流或缓冲解决方案。

为了更好地理解它们，让我们探讨一下节流和缓冲。

**节流：**如果需求源具有重试功能，可以实施节流。节流会告诉需求源，如果当前无法处理请求，则应稍后再试。源将等待一段时间，然后重试请求。实施节流的优点是可限制最大资源量和工作负载成本。在 AWS 中，您可以使用 [Amazon API Gateway](#) 来实施节流。

**基于缓冲区：**基于缓冲区的方法使用产生器（向队列发送消息的组件）、使用器（从队列接收消息的组件）和队列（保存消息）来存储消息。然后消息将由使用器读取并处理，这样消息就能够以满足使用器业务需求的速率运行。通过使用以缓冲区为中心的方法，产生器发出的消息被存储在队列或流中，随时可供使用器以符合其运营需求的速度进行访问。

在 AWS 中，您可以从多个服务中进行选择来实施缓冲方法。[Amazon Simple Queue Service \(Amazon SQS\)](#) 是一项托管服务，它提供队列，允许单个使用器读取单个消息。[Amazon Kinesis](#) 提供了一个允许许多使用器读取相同消息的流。

缓冲和节流可以通过修改工作负载需求来平滑任何峰值。在客户端重试操作时使用节流，并使用缓冲来保存请求以供以后处理。使用基于缓冲区的方法时，请构造工作负载以在所需时间内处理请求，并验证您是否能够处理重复的工作请求。分析总体需求、变化率和所需的响应时间，以使所需节流或缓冲的大小适宜。

## 实施步骤

- 分析客户端需求：分析客户端请求以确定它们是否能够重试。对于无法执行重试的客户端，需要实施缓冲区。分析总体需求、变化率和所需的响应时间，以确定所需的节流或缓冲区大小。
- 实施缓冲区或节流：在工作负载中实施缓冲区或节流。Amazon Simple Queue Service ( Amazon SQS ) 之类的队列可以为工作负载组件提供缓冲区。Amazon API Gateway 可以为工作负载组件提供节流。

## 资源

### 相关最佳实践：

- [SUS02-BP06 实施缓冲区和节流以展平需求曲线](#)
- [REL05-BP02 限制请求](#)

### 相关文档：

- [AWS Auto Scaling](#)
- [AWS 实例调度器](#)
- [Amazon API Gateway](#)
- [Amazon Simple Queue Service](#)
- [开始使用 Amazon SQS](#)
- [Amazon Kinesis](#)

### 相关视频：

- [为分布式应用程序选择适合的消息传递服务](#)

### 相关示例：

- [管理和监控工作负载中的 API 节流](#)
- [使用 API Gateway 大规模节流分层的多租户 REST API](#)
- [使用 Amazon API Gateway 在多租户 Amazon EKS SaaS 解决方案中启用分层和节流](#)
- [使用队列和消息的应用程序集成](#)

## COST09-BP03 动态供应资源

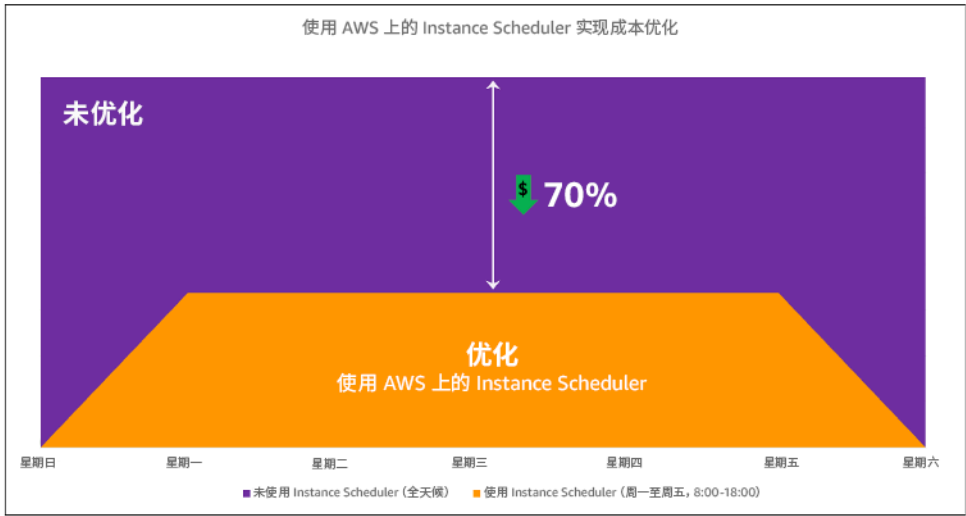
资源按计划预置。这种预置可以基于需求（例如通过自动扩缩来实现），也可以基于时间（需求可以预测，基于时间提供资源）。这些方法可以尽可能减少过度预置或预置不足的情况。

未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

AWS 客户可以通过多种方式增加可供应用程序使用的资源，并提供资源以满足需求。其中一个选项是使用 AWS Instance Scheduler，其可以自动启动和停止 Amazon Elastic Compute Cloud（Amazon EC2）以及 Amazon Relational Database Service（Amazon RDS）实例。另一个选项是使用 AWS Auto Scaling，该服务让您可以根据应用程序或服务的需求，自动扩展计算资源。根据需求提供资源，这样您就只需要为使用的资源付费，并且仅在有需要时启动资源，在不需要时终止资源，从而降低成本。

[AWS Instance Scheduler](#) 让您可以将 Amazon EC2 和 Amazon RDS 实例配置为在指定的时间停止和启动，这样您就可以通过一致的时间模式满足对相同资源的需求，例如用户在每天早上八点访问 Amazon EC2 实例，晚上六点后就不再需要访问。该解决方案可停止不使用的资源，并在需要时启动它们，来实现运营成本的降低。



使用 AWS Instance Scheduler 优化成本。

您还可以通过简单的用户界面（UI）使用 AWS Systems Manager 快速设置，轻松地跨账户和区域来为您的 Amazon EC2 实例配置计划。您可以使用 AWS Instance Scheduler 来调度 Amazon EC2 或 Amazon RDS 实例，也可以停止和启动现有实例。然而，您不能停止和启动属于 Auto Scaling 组（ASG）或管理 Amazon Redshift 或 Amazon OpenSearch Service 等服务的实例。Auto Scaling 组对组内创建的实例有自己的计划。

[AWS Auto Scaling](#) 有助于您调整容量以维持稳定、可预测的性能，并确保成本最低，来满足不断变化的需求。这是一项用来扩展应用程序容量的完全托管式免费服务，与 Amazon EC2 实例和竞价型实例集、Amazon ECS、Amazon DynamoDB 和 Amazon Aurora 集成。Auto Scaling 提供自动资源发现功能，以便您在工作负载中找到可以配置的资源，其具有内置的扩展策略来优化性能、成本或者在两者之间取得平衡，并提供预测性扩展来协助应对定期出现的峰值。

您可以通过多种扩展选项来扩展 Auto Scaling 组：

- 始终保持当前的实例水平
- 手动扩展
- 根据计划扩展
- 根据需求扩展
- 使用预测性扩缩

Auto Scaling 策略各不相同，可以分为动态扩缩策略和计划扩缩策略。动态策略为手动扩缩或动态扩缩，这可以是计划扩缩或者预测性扩缩。您可以针对动态、计划和预测性扩缩使用扩缩策略。您还可以使用来自 [Amazon CloudWatch](#) 的指标和警报触发工作负载的扩缩事件。我们建议您使用 [启动模板](#)，这可让您访问最新的功能和改进。使用启动配置时，并非所有 Auto Scaling 功能均可用。例如，您不能创建一个 Auto Scaling 组，然后在其中同时启动竞价型实例和按需型实例，也不能指定多种实例类型。这些功能必须使用启动模板来配置。使用启动模板时，建议您对每个模板进行版本控制。通过启动模板的版本控制，您可以创建完整参数集的子集。然后，您可以重复使用参数子集来创建相同启动模板的其他版本。

您可以使用 AWS Auto Scaling，也可以通过 [AWS API 或 SDK](#) 在代码中纳入扩缩功能。这样省去了手动更改环境的操作成本，因而工作负载的总体成本得以降低，而且可以更快地执行更改。这还使您的工作负载资源配置随时与您的需求相匹配。为了遵循这一最佳实践，并为您的组织动态供应资源，您应该了解 AWS Cloud 中的横向和纵向扩展，以及 Amazon EC2 实例上运行的应用程序的性质。您的云财务管理团队最好与技术团队合作，以遵循这一最佳实践。

[Elastic Load Balancing \( Elastic Load Balancing \)](#) 通过在多种资源之间分配需求来助力您扩展规模。通过使用 ASG 和 Elastic Load Balancing，您可以按照最优方式路由流量来管理传入的请求，这样就可以避免 Auto Scaling 组中某个实例负载过高的情况。请求将以轮询方式，在目标组的所有目标上分配，而不考虑容量或利用率。

典型的指标可以是标准 Amazon EC2 指标，例如 CPU 利用率、网络吞吐量以及 Elastic Load Balancing 观察到的请求和响应延迟。如果可能，应该使用指示客户体验的指标，通常是来自工作负载中的应用程序代码的自定义指标。在本文档中，为了详细说明如何动态满足需求，我们将 Auto Scaling 划分为两类：基于需求的供应模型和基于时间的供应模型，并分别深入探讨这两种模型。



**基于需求的供应：** 利用云的弹性来提供资源，根据近实时的需求状态来满足不断变化的需求。对于基于需求的供应，请使用 API 或服务功能，以编程方式改变架构中云资源的数量。这使您能够在架构中扩展组件，并在需求高峰期间增加资源数量以保持性能，也可以在需求量降低时减少容量以降低成本。

## 基于需求的供应（动态扩缩策略）



**简单/分步扩展**



**目标跟踪**

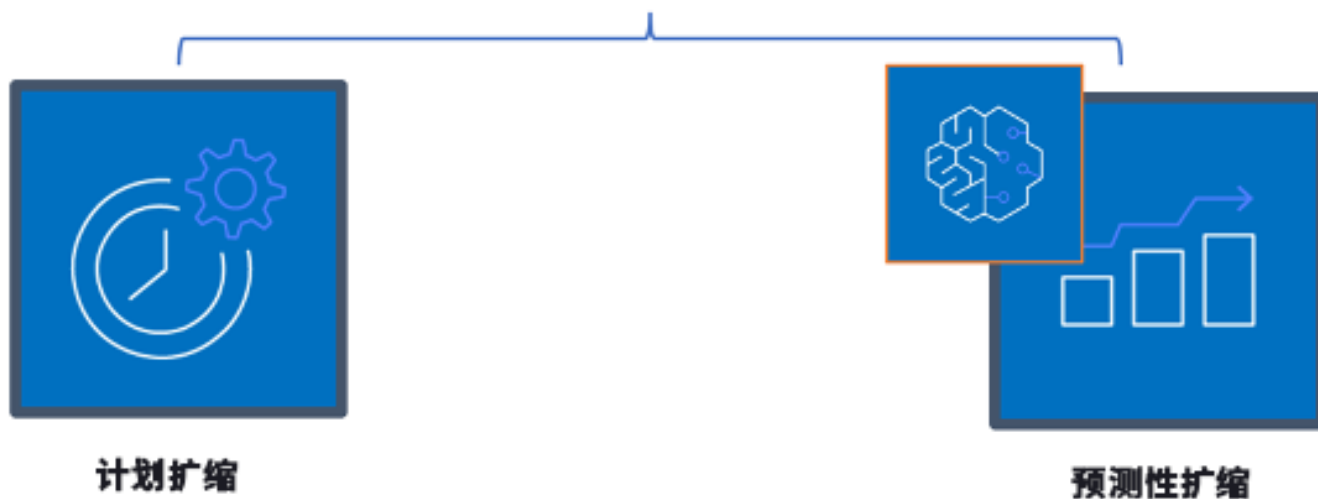
### 基于需求的动态扩缩策略

- **简单/分步扩展：** 监控指标，按照客户定义的步骤手动添加/删除实例。
- **目标跟踪：** 类似恒温器的控制机制，可自动添加或删除实例，以维护客户定义的目标指标。

当构建基于需求的方法时，请注意两个重要事项。首先，了解您必须以多快的速度预置新资源。其次，了解供应和需求之间的差额将发生变化。您必须准备好应对需求变化的速度，并准备好应对资源故障。

**基于时间的供应：** 基于时间的方法可以协调资源容量以满足可预测或时间明确定义的需求。此方法通常不依赖资源的利用水平。基于时间的方法可以确保资源在需要的特定时间可用，并且提供时不会因启动流程和系统或一致性检查而发生延迟。使用基于时间的方法，您可以在繁忙时段提供额外的资源或增加容量。

## 基于时间的供应 (计划扩缩策略和预测性扩缩策略)



### 基于时间的扩缩策略

您可以使用计划性或预测性自动扩缩，实施基于时间的方法。工作负载可以在定义的时间按计划扩展或缩减（例如办公时间开始时），从而确保用户就位或需求增加时资源可用。预测性扩缩使用模式进行横向扩展，而计划扩展在预先定义的时间进行横向扩展。您也可以将 [基于属性的实例类型选择 \(ABS\) 策略](#) 用于 Auto Scaling 组，该策略允许您将实例要求表示为一组属性，例如 vCPU、内存和存储。这还允许您在新一代实例类型发布时自动使用它们，并通过 Amazon EC2 竞价型实例访问更广泛的容量。Amazon EC2 实例集和 Amazon EC2 Auto Scaling 会选择并启动符合指定属性的实例，无需手动选择实例类型。

您还可以利用 [AWS API 与 SDK](#) 和 [AWS CloudFormation](#) 在需要时自动预置和停用整个环境。此方法非常适合仅在定义的办公时间或时间段运行的开发或测试环境。您可以使用 API 来扩展环境中的资源大小（纵向扩展）。例如，可以通过更改实例大小或分类纵向扩展生产工作负载。这可以通过停止和启动实例，以及选择不同的实例大小或分类来实现。这种技巧也可以应用于其他资源，如 Amazon EBS 弹性卷，您可以在使用时对其进行修改以增加大小、调整性能（IOPS）或更改卷类型。

当采用基于时间的方法进行架构设计时，请注意两个重要事项。首先，使用模式的一致性如何？其次，如果模式发生更改会产生什么影响？您可以通过两种方式提高预测的准确性：监控工作负载和使用商业智能。如果您发现使用模式发生重大更改，可以调整时间，以确保提供覆盖范围。

## 实施步骤

- **配置计划扩缩：** 对于可预测的需求变化，基于时间的扩缩可以及时提供正确的资源量。如果资源创建和配置的速度不够快，无法响应需求变化，也可使用这种方法。根据工作负载分析，使用 AWS Auto Scaling 配置计划扩缩。要配置基于时间的计划，您可以使用预测性扩缩而不是计划扩缩，根据预期或可预测的负载变化，提前增加Auto Scaling组中的 Amazon EC2 实例数量。
- **配置预测性扩缩：** 通过预测性扩缩，您可以根据流量的每日和每周模式，提前增加Auto Scaling组中的 Amazon EC2 实例数量。如果您有定期的流量高峰和需要很长时间才能启动的应用程序，则应该考虑使用预测性扩缩。与本质上属于被动应对的单纯动态扩缩相比，预测性扩缩可帮助您在预测的负载到来之前初始化容量，从而更快地扩展。例如，如果用户在开始上班时开始使用工作负载，而在下班后不再使用，那么预测性扩缩可以在上班之前增加容量，这就消除了动态扩缩对流量变化作出反应的延迟。
- **配置动态自动扩缩：** 要根据活动的工作负载指标配置扩缩，请使用Auto Scaling。使用分析并配置 Auto Scaling以在正确的资源级别上启动，并确认工作负载在所需的时间内扩展。您可以在单个Auto Scaling组中启动并自动扩展按需型实例和竞价型实例的实例集。除了获得使用竞价型实例的折扣外，您还可以使用预留实例或实惠配套来享受常规按需型实例定价的折扣率。将所有这些因素相结合，有助于您优化 Amazon EC2 实例的成本节约，并帮助您获得应用程序所需的规模和性能。

## 资源

### 相关文档：

- [AWS Auto Scaling](#)
- [AWS Instance Scheduler](#)
- 扩缩Auto Scaling组的大小
- [开始使用 Amazon EC2 Auto Scaling](#)
- [开始使用 Amazon SQS](#)
- [Amazon EC2 Auto Scaling 的计划扩缩](#)
- [Amazon EC2 Auto Scaling 的预测性扩缩](#)

### 相关视频：

- [Auto Scaling的目标跟踪扩缩策略](#)
- [AWS Instance Scheduler](#)

相关示例：

- [基于属性为 Amazon EC2 实例集的 Auto Scaling 选择实例类型](#)
- [使用计划扩缩优化 Amazon Elastic Container Service 成本](#)
- [Amazon EC2 Auto Scaling 预测性扩缩](#)
- [如何通过将 Instance Scheduler 和 AWS CloudFormation 结合使用来调度 Amazon EC2 实例？](#)

## 持续优化

问题

- [COST 10.如何评估新服务？](#)
- [COST 11.如何评估工作量成本？](#)

### COST 10.如何评估新服务？

AWS 不断发布新服务和功能，因此您最好不断审视现有架构决策，以便确保其始终最具成本效益。

最佳实践

- [COST10-BP01 制定工作负载审核流程](#)
- [COST10-BP02 定期审核和分析此工作负载](#)

#### COST10-BP01 制定工作负载审核流程

制定流程，定义工作负载的审核标准和流程。审核工作量应该体现可能带来的好处，例如，核心工作负载或费用占比超过 10% 的工作负载每季度或每六个月审核一次，而费用占比低于 10% 的工作负载每年审核一次。

在未建立这种最佳实践的情况下暴露的风险等级：高

实施指导

为了让工作负载始终最具成本效益，您必须定期对其进行审核，以了解是否有机会实施新的服务、功能和组件。为了降低整体成本，审核工作量必须与潜在的节省额成比例。例如，与占总支出 5% 的工作负载相比，应更经常、更彻底地审核占总支出 50% 的工作负载。考虑任何外部因素或波动。如果工作负载服务于特定的地理位置或市场领域，并且您已预测出该领域会出现的变化，则提高审核频率可能会节省成本。审核时要考虑的另一个因素是实施更改的工作量。如果测试和验证变更的成本很高，则审核的频率应该降低。

考虑维护过时和旧式组件及资源的长期成本，以及无法在其中实施新功能的事实。当前的测试和验证成本可能会超过预计的收益。但是，随着时间的推移，工作负载和当前技术之间的差距会增大，进行更改的成本可能会大幅增加，导致成本升高。例如，迁移到新的编程语言的成本当前可能不具成本效益。然而，五年之后，熟练使用该语言的人员的成本可能会增加，并且由于工作负载的扩展，迁移到新语言的系统规模更大，这其中涉及的工作量甚至高于以前。

将工作负载分解成多个组件，分配组件的成本（估算即可），然后在每个组件旁边列出因素（例如工作量和外部市场）。使用这些指示信息来确定每个工作负载的审核频率。例如，您可能觉得 Web 服务器的成本高、变更的工作量小、外部因素多，因而审核频率很高。而中央数据库的成本可能中等、变更的工作量很大、外部因素较少，因而审核频率也为中等。

制定相应流程，在新的服务、设计模式、资源类型和配置推出后，对它们进行评估，以优化您的工作负载成本。与[性能支柱审核](#)和[可靠性支柱审核](#)过程类似，识别、验证和优先考虑优化和改进活动以及问题补救，并将其纳入您的待办事项列表。

## 实施步骤

- **定义审核频率：**定义工作负载及其组件的审核频率。分配时间和资源，以持续改进并保持审核频率，来优化工作负载并提高工作负载的效率。应考虑多种因素，且这些因素可能会因组织中的工作负载以及工作负载中的组件而异。常见的因素包括：从收入或品牌角度来讲对组织的重要性、运行工作负载的总成本（包括运营和资源成本）、工作负载的复杂性、实施变革的难易程度、任何软件许可协议以及变革是否会因惩罚性的许可而显著增加许可成本。可以在功能上或技术上定义组件，例如 Web 服务器和数据库，或者计算和存储资源。相应地权衡这些因素，并为工作负载及其组件制定一个周期。您可能决定每 18 个月审核一次完整的工作负载，每 6 个月审核一次 Web 服务器，每 12 个月审核一次数据库，每 6 个月审核一次计算资源和短期存储，每 12 个月审核一次长期存储。
- **定义审核的彻底性：**定义在工作负载或工作负载组件的审核上投入的工作量。与审核频率类似，这也需要权衡多种因素。评估各种改进机会并确定其优先顺序，以便将精力集中在可以实现最大收益的工作上，同时估计这些活动所需的工作量。如果预期成果不符合目标，并且所需工作会花费更多成本，则寻求其他行动方案。审核流程中应该分配专用的时间和资源，以便实现持续增量改进。例如，您可能决定投入一周的时间对数据库组件进行分析，投入一周的时间对计算资源进行分析，投入四小时的时间进行存储审核。

## 资源

相关文档：

- [AWS 新闻博客](#)
- [云计算类型](#)

- [AWS 新增功能](#)

相关示例：

- [AWS 支持主动式服务](#)
- [定期审核 SAP 工作负载](#)

## COST10-BP02 定期审核和分析此工作负载

根据每个明确的流程定期审核现有工作负载，以便确定是采用新服务、替换现有服务还是重新构建工作负载。

在未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

AWS 在不断地增加新功能，以便您可以使用最新技术更快地进行实验和创新。[AWS 新增功能](#)详细介绍了 AWS 如何实现这一点，并在 AWS 服务、功能和区域扩展公告发布时进行简要概述。您可以深入了解已经宣布的产品发布，并使用它们来审核和分析您现有的工作负载。为了享受新 AWS 服务和功能带来的好处，请对工作负载进行审核，并根据需要实施新服务和功能。这意味着您可能需要替换用于工作负载的现有服务，或对工作负载进行现代化改造，以便采用这些新的 AWS 服务。例如，您可以审核工作负载，并使用 Amazon Simple Email Service 替换消息收发组件。这省去了运行和维护实例集的成本，同时能以更低的成本提供所有功能。

为了分析您的工作负载并突显潜在机会，您不仅应考虑新服务，而且还要考虑构建解决方案的新方法。观看 AWS 上的[这是我的架构](#)视频，了解其他客户的架构设计、面临的挑战及其解决方案。查看[全包系列](#)，了解 AWS 服务的现实世界应用和客户案例。您还可以观看[回归基础](#)视频系列。该系列说明、检查和分解基本的云架构模式最佳实践。另一个资料来源是[如何构建](#)视频。这些视频旨在帮助有好想法的人使用 AWS 服务将他们的最小可行产品（MVP）变为现实。来自世界各地有着坚定想法的构建者可以利用这种方式从经验丰富的 AWS 解决方案架构师那里获得架构指导。最后，您可以查看[使用入门](#)资源材料，其中包含循序渐进的教程。

在执行审核流程之前，请遵照您的企业对工作负载的要求、安全性和数据隐私要求，以便在遵循商定的审核流程时，运用特定的服务或区域和性能要求。

### 实施步骤

- 定期审核工作负载：使用明确的流程，按照指定的频率执行审核。确认在每个组件上投入适当的工作量。此流程类似于您选择服务进行成本优化的初始设计流程。分析服务及其带来的优势，这一次需考虑实施更改所产生的成本，而不仅仅是长期优势。

- 实施新服务：如果分析结果表明可以实施更改，请先执行工作负载基线，以了解每项产出的当前成本。实施更改，然后执行分析以确认每项产出的新成本。

## 资源

### 相关文档：

- [AWS 新闻博客](#)
- [AWS 新增功能](#)
- [AWS 文档](#)
- [AWS 入门](#)
- [AWS 一般资源](#)

### 相关视频：

- [AWS - 这是我的架构](#)
- [AWS - 回归基础](#)
- [AWS - 全包系列](#)
- [如何构建](#)

## COST 11.如何评估工作量成本？

### 最佳实践

- [COST11-BP01 执行运营自动化](#)

### COST11-BP01 执行运营自动化

评估云上运营工作量的成本。量化使用自动化之后管理任务、部署和其他运营工作减少的时间和工作量。评估运营工作所需的时间和成本，并自动执行管理任务，以尽可能减少人工工作量。

在未建立这种最佳实践的情况下暴露的风险等级：低

自动执行运营可解放人力资源和改善指标，从而提高一致性和可扩展性、实现更高的可见性、可靠性和灵活性、降低成本以及加快创新。还可以在部署、管理或运营工作负载时提供一致且可靠的体验，从而降低手动任务的频率、提高效率以及使企业受益。您可以将基础设施资源从手动操作任务中解放出来，



并将它们用于更高价值的任务和创新，从而改善业务成果。企业需要行之有效、经过测试的方法来管理他们在云中的工作负载。该解决方案必须安全、快速和经济高效，风险最小且可靠性最高。

首先着眼于云中的总体运营成本，根据所需的工作量确定运营优先级。例如，在云中部署新资源、对现有资源进行优化更改或实施必要的配置需要多长时间？考虑运营和管理成本，看看人工行为的总成本。优先考虑管理任务的自动化，以减少人工工作量。审核工作应该体现可能带来的好处。例如，与自动执行任务相比，手动执行任务所花费的时间。优先考虑自动执行重复、高价值的活动。从那些人为错误风险更高的活动开始实现自动化通常会更好，因为风险通常会带来不必要的额外运营成本（例如，运营团队加班产生的成本）。

使用 AWS 服务、工具或第三方产品，您可以选择要实施哪些 AWS 自动化，并根据您的特定需求进行定制。下表显示了您为了自动执行管理和运营而可以通过 AWS 服务实现的一些核心运营职能和能力：

- [AWS Audit Manager](#)：持续审计您的 AWS 使用情况，以便简化风险和合规性评测
- [AWS Backup](#)：集中管理和自动执行数据保护。
- [AWS Config](#)：配置计算资源、评测、审计和评估配置与资源清单。
- [AWS CloudFormation](#)：使用基础设施即代码启动高可用性资源。
- [AWS CloudTrail](#)：IT 变更管理、合规性和控制。
- [Amazon EventBridge](#)：安排事件并触发 AWS Lambda 以采取行动。
- [AWS Lambda](#)：通过事件触发流程或使用 Amazon EventBridge 按固定时间表运行流程，使重复流程实现自动化。
- [AWS Systems Manager](#)：启动和停止工作负载、修补操作系统、自动配置和持续管理。
- [AWS Step Functions](#)：安排作业和自动执行工作流。
- [AWS Service Catalog](#)：具有合规性和控制的模板使用和基础设施即代码。

利用节省下来的时间，您的团队将能够专注于解决技术债务、创新和增值功能。例如，您可能需要尽快将本地环境直接迁移到云，然后再进行优化。值得探索的是，通过使用可消除或减少许可证成本的完全托管式 AWS 服务（例如，[Amazon Relational Database Service](#)、[Amazon EMR](#)、[Amazon WorkSpaces](#) 和 [Amazon SageMaker AI](#)），您可以节省多少成本。托管服务消除了维护服务的运营和管理负担，让您可以专注于创新。此外，由于托管服务在云级别运行，因此可以提供更低的单位事务或服务成本。

如果您希望使用 AWS 产品和服务立即采用自动化，而您的组织中没有相应的技能，请联系 [AWS Managed Services \(AMS\)](#)、[AWS Professional Services](#) 或 [AWS 合作伙伴](#)，以便提高自动化的采用率，以及改进云中的卓越运营。

[AWS Managed Services \( AMS \)](#) 是代表企业客户和合作伙伴运营 AWS 基础设施的服务。该服务提供了一个安全且合规的环境，您可以将工作负载部署到其中。AMS 使用具有自动化功能的企业云运营模型，可以满足组织要求，更快地迁移到云中并降低持续管理的成本。

[AWS Professional Services](#) 还可以帮助您实现期望的业务成果并通过 AWS 实现运营自动化。AWS Professional Services 提供全球专业实践服务，以支持您在企业云计算重点领域的工作。这些专业实践服务通过解决方案、技术和行业主题领域的最佳实践、框架、工具和服务提供有针对性的指导。它们帮助客户部署自动、稳健、敏捷的 IT 运营，还提供针对云中心进行优化的治理能力。

## 实施步骤

- 一次构建，多次部署：使用基础设施即代码 [例如 AWS CloudFormation、AWS SDK 或 AWS Command Line Interface ( AWS CLI ) ]，对于同一环境或灾难恢复场景，只部署一次，多次使用。在部署时进行标记，以便按照其他最佳实践中的规定跟踪您的使用情况。使用 [AWS Launch Wizard](#) 以缩短部署许多常用企业工作负载的时间。AWS Launch Wizard 会指导您按照 AWS 最佳实践完成企业工作负载的大小调整、配置和部署。您还可以使用 [AWS Service Catalog](#) 帮助创建和管理经批准的基础设施即代码模板，在 AWS 上使用，以便任何人都可以发现经批准的自助式云资源。
- 自动操作：自动运行日常操作，无需人工干预。使用 AWS 服务和工具，您可以选择要实施哪些 AWS 自动化，并根据您的特定需求进行定制。例如，使用 [EC2 Image Builder](#) 来构建、测试和部署虚拟机和容器映像，以在 AWS 或在本地使用。如果无法使用 AWS 服务完成您的期望操作或您需要使用筛选资源的更复杂操作，则使用 [AWS CLI](#) 或 AWS SDK 工具自动执行操作。AWS CLI 可以通过脚本自动完成控制和管理 AWS 服务的整个过程，而无需使用 AWS 控制台。选择您首选的 AWS SDK 与 AWS 服务交互。有关其他代码示例，请参阅 [AWS SDK 代码示例库](#)。

## 资源

### 相关文档：

- [在 AWS Cloud 中实现运营现代化](#)
- [AWS 服务自动化](#)
- [AWS Systems Manager 自动化](#)
- [SAP 管理和运营的 AWS 自动化](#)
- [AWS Managed Services](#)
- [AWS Professional Services](#)
- [基础设施和自动化](#)

相关示例：

- [重塑自动化运营（第 I 部分）](#)
- [重塑自动化运营（第 II 部分）](#)
- [SAP 管理和运营的 AWS 自动化](#)
- [使用 AWS Lambda 实现 IT 自动化](#)
- [AWS 代码示例库](#)
- [AWS 示例](#)

## 可持续性

可持续性支柱包括在构建云工作负载时，了解所使用服务的影响，量化整个工作负载生命周期的影响，并应用设计原则和最佳实践来减少这些影响。如需有关具体实施的说明性指导，请参阅 [《可持续性支柱》白皮书](#) 访问 AWS 资源。

最佳实践领域

- [区域选择](#)
- [符合需求](#)
- [软件和架构](#)
- [数据](#)
- [硬件和服务](#)
- [流程和文化](#)

## 区域选择

问题

- [SUS 1 您如何为工作负载选择区域？](#)

### SUS 1 您如何为工作负载选择区域？

为工作负载选择区域会显著影响其 KPI，包括性能、成本和碳足迹。为了有效提高这些 KPI，您应该根据业务需求和可持续发展目标为工作负载选择区域。

最佳实践

- [SUS01-BP01 根据业务需求和可持续发展目标选择区域](#)

## SUS01-BP01 根据业务需求和可持续发展目标选择区域

根据您的业务需求和可持续发展目标为您的工作负载选择一个区域，以优化其 KPI，包括性能、成本和碳足迹。

常见反模式：

- 您可以根据自己所在的位置选择工作负载的区域。
- 您可以将所有工作负载资源整合到一个地理位置中。

建立此最佳实践的好处：将工作负载放置在 Amazon 可再生能源项目或已发布碳强度较低的区域附近，有助于降低云工作负载的碳足迹。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

AWS Cloud 是一个不断扩展的区域和入网点（PoP）网络，其全球网络基础设施将它们连接在一起。为工作负载选择区域会显著影响其 KPI，包括性能、成本和碳足迹。为了有效提高这些 KPI，您应该根据业务需求和可持续发展目标为工作负载选择区域。

### 实施步骤

- 按照以下步骤进行操作，根据您的业务需求（包括法规遵从性、可用功能、成本和延迟）评估工作负载的潜在区域并列入候选名单：
  - 根据您所需的当地法规，确认这些区域合规。
  - 使用 [AWS 区域性服务列表](#)，检查区域是否具有运行工作负载所需的服务和功能。
  - 使用 [AWS 定价计算器](#) 计算每个区域的工作负载成本。
  - 测试最终用户位置与每个 AWS 区域之间的网络延迟。
- 选择亚马逊可再生能源项目附近的区域和其电网公布的碳强度低于其他位置（或区域）的区域。
  - 确定您的相关可持续发展准则，以根据[温室气体核算协议](#)（基于市场和基于位置的方法）跟踪和比较逐年碳排放量。
  - 根据用于跟踪碳排放的方法选择区域。有关根据可持续性准则选择区域的更多详细信息，请参阅[如何根据可持续性目标为工作负载选择区域](#)。

## 资源

相关文档：

- [了解碳排放估算](#)
- [Amazon 遍布全球](#)
- [可再生能源方法](#)
- [为工作负载选择区域时应考虑的事项](#)

相关视频：

- [可持续地构建并减少 AWS 碳足迹](#)

## 符合要求

问题

- [SUS 2 如何将云资源与您的需求相匹配？](#)

### SUS 2 如何将云资源与您的需求相匹配？

用户和应用程序使用您的工作负载及其他资源的方式可以帮助您找出改进措施，以实现可持续性目标。扩展基础设施以持续匹配需求，并验证您是否仅使用了支持用户所需的最少资源。使服务水平与客户需求保持一致。定位资源以限制用户和应用程序使用这些资源所需的网络。删除未使用的资产。为您的团队成员提供满足其需求的设备，并尽可能降低他们的可持续发展影响。

最佳实践

- [SUS02-BP01 动态扩缩工作负载基础设施](#)
- [SUS02-BP02 使 SLA 与可持续性目标保持一致](#)
- [SUS02-BP03 停止创建和维护未使用的资产](#)
- [SUS02-BP04 根据其联网要求优化工作负载的地理位置](#)
- [SUS02-BP05 针对执行的活动优化团队成员资源](#)
- [SUS02-BP06 实施缓冲和节流以展平需求曲线](#)

## SUS02-BP01 动态扩缩工作负载基础设施

利用云的弹性并动态扩缩基础设施，以使云资源的供应与需求相匹配，避免在工作负载中过度调配容量。

常见反模式：

- 您没有扩缩基础设施以匹配用户负载。
- 您一直在手动扩缩基础设施。
- 在扩展事件之后，您将保留增加的容量，而不是缩减容量。

建立此最佳实践的好处：配置和测试工作负载弹性有助于有效地将云资源的供应与需求相匹配，并避免过度调配容量。您可以利用云中的弹性，在需求高峰期间和之后自动扩缩容量，以确保您只使用满足业务需求所需的适当数量的资源。

在未建立这种最佳实践的情况下暴露的风险等级：中等

### 实施指导

云让您能够通过各种机制灵活地动态扩展或缩减资源，以便满足不断变化的需求。供应与需求的最佳匹配提供了最低的工作负载环境影响。

需求可以是固定的，也可以是变化的，需要指标和自动化来确保管理不会变成沉重负担。应用程序可以通过修改实例大小来纵向扩展或缩减，通过修改实例数量来横向扩展或缩减，或者组合使用这两种方式。

您可以使用大量不同方法来实现资源的供需匹配。

- 目标跟踪方法：监控您的扩缩指标，并根据需要自动增加或减少容量。
- 预测性扩缩：根据每日和每周的趋势进行扩缩。
- 基于计划的方法：根据可预测的负载变化设置自己的扩缩计划。
- 服务扩缩：选择按设计可以原生扩缩或者将自动扩缩作为一项功能提供的服务（如无服务器）。

确定利用率低或无利用率的时段，缩减资源以消除过剩容量并提高效率。

### 实施步骤

- 弹性可根据对您拥有的资源的需求来提供这些资源。实例、容器和函数提供了弹性机制，可以与自动扩缩结合使用，也可以作为服务的一项功能。AWS 提供了一系列自动扩缩机制，以确保工作负载可以在低用户负载期间快速轻松地缩减。以下是自动扩缩机制的一些示例：

| Auto scaling mechanism                        | Where to use                                                                                      |
|-----------------------------------------------|---------------------------------------------------------------------------------------------------|
| <a href="#">Amazon EC2 Auto Scaling</a>       | 用于验证您拥有适量的 Amazon EC2 实例，可处理您应用程序的用户负载。                                                           |
| <a href="#">Application Auto Scaling</a>      | 用于自动扩缩 Amazon EC2 以外的各项 AWS 服务的资源，比如 Lambda 函数或 Amazon Elastic Container Service (Amazon ECS) 服务。 |
| <a href="#">Kubernetes Cluster Autoscaler</a> | 用于自动扩缩 AWS 上的 Kubernetes 集群。                                                                      |

- 扩缩通常与计算服务（如 Amazon EC2 实例或 AWS Lambda 函数）相关。考虑使用非计算服务配置（如 [Amazon DynamoDB](#) 读写容量单元或 [Amazon Kinesis Data Streams](#) 分片）来满足需求。
- 验证衡量扩展或缩减的指标已根据所部署的工作负载类型进行了验证。如果您正在部署一个视频转码应用程序，CPU 利用率预计为 100%，并且不应将此作为您的主要指标。如果需要，可以对扩缩策略使用[自定义指标](#)（如内存利用率）。要选择正确的指标，请考虑以下关于 Amazon EC2 的指导：
  - 该指标应该是有效的利用率指标，并描述实例的繁忙程度。
  - 该指标值必须随 Auto Scaling 组中的实例数量成比例地增加或减少。
- 对 Auto Scaling 组使用[动态扩缩](#)而不是[手动扩缩](#)。我们还建议您在动态扩缩中使用[目标跟踪扩缩策略](#)。
- 确认工作负载部署可以处理扩展事件和缩减事件。为缩减事件创建测试场景，以确认工作负载的行为符合预期，并且不会影响用户体验（如丢失粘滞会话）。您可以使用[活动历史记录](#)验证 Auto Scaling 组的扩缩活动。
- 评估您的工作负载以获得可预测的模式，并在您预期需求会发生预测和计划的变化时主动扩缩。借助预测性扩缩，您无需过度调配容量。有关详细信息，请参阅[使用 Amazon EC2 Auto Scaling 进行预测性扩缩](#)。

## 资源

### 相关文档：

- [开始使用 Amazon EC2 Auto Scaling](#)
- [由机器学习提供支持的 EC2 预测式扩缩](#)
- [使用 Amazon OpenSearch Service、Amazon Data Firehose 和 Kibana 分析用户行为](#)



- [什么是 Amazon CloudWatch ?](#)
- [在 Amazon RDS 上使用性能详情监控数据库负载](#)
- [介绍对 Amazon EC2 Auto Scaling 预测式扩缩的原生支持](#)
- [介绍 Karpenter - 高性能开源 Kubernetes Cluster Autoscaler](#)
- [深入探讨 Amazon ECS 集群 Auto Scaling](#)

相关视频：

- [构建成本、能源和资源高效的计算环境](#)
- [更好、更快、更便宜的计算：成本优化Amazon EC2 \( CMP202-R1 \)](#)

相关示例：

- [实验室：Amazon EC2 Auto Scaling 组示例](#)
- [实验室：使用 Karpenter 实施自动扩缩](#)

SUS02-BP02 使 SLA 与可持续性目标保持一致

根据您的可持续发展目标审查和优化工作负载服务水平协议（SLA），以便在继续满足业务需求的同时，尽量减少支持您的工作负载所需的资源。

常见反模式：

- 工作负载 SLA 未知或模棱两可。
- 只针对可用性和性能定义您的 SLA。
- 对所有工作负载使用相同设计模式（如多可用区架构）。

建立此最佳实践的好处：使 SLA 与可持续发展目标一致，在满足业务需求的同时实现最佳资源使用率。

在未建立这种最佳实践的情况下暴露的风险等级：低

实施指导

SLA 定义云工作负载的预期服务水平，如响应时间、可用性和数据留存。它们影响云工作负载的架构、资源使用率和环境影响。定期审查 SLA，并做出权衡，显著减少资源使用，以换取可接受的服务水平降低幅度。

## 实施步骤

- 定义或重新设计 SLA，在支持可持续性目标的同时满足而不是超出您的业务需求。
- 做出权衡，显著降低可持续性影响，以换取可接受的服务水平降低幅度。
  - 可持续性和可靠性：高可用性工作负载往往会消耗更多资源。
  - 可持续发展和性能：使用更多资源来提升性能可能会对环境产生更大影响。
  - 可持续发展和安全：过度安全的工作负载可能会对环境产生更大影响。
- 使用优先考虑业务关键功能的设计模式（例如 [AWS](#) 上的微服务），并允许非关键功能具有较低的服务水平（例如响应时间或恢复时间目标）。

## 资源

### 相关文档：

- [AWS 服务水平协议 \( SLA \)](#)
- [Importance of Service Level Agreement for SaaS Providers](#)

### 相关视频：

- [提供可持续、高性能的架构](#)
- [构建成本、能源和资源高效的计算环境](#)

## SUS02-BP03 停止创建和维护未使用的资产

停用您的工作负载中未使用的资产，以便减少支持您的需求所需的云资源数量，并最大限度地减少浪费。

### 常见反模式：

- 您没有分析应用程序以查找冗余或不再需要的资产。
- 您没有移除冗余或不再需要的资产。

建立此最佳实践的好处：移除未使用的资产可释放资源并提高工作负载的整体效率。

在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

未使用的资产会消耗存储空间和计算能力等云资源。通过识别和消除这些资产，您可以释放这些资源，从而形成更高效的云架构。定期分析应用程序资产（例如预编制的报告、数据集和静态图像）和资产访问模式，以识别冗余、利用率低下的情况和潜在的淘汰目标。移除这些冗余资产以减少工作负载中的资源浪费。

## 实施步骤

- 使用监控工具来识别不再需要的静态资产。
- 在移除任何资产之前，评估移除它会对架构产生什么影响。
- 制定计划并移除不再需要的资产。
- 整合生成的重叠资产以消除冗余处理。
- 更新应用程序，以便不再产生和存储不需要的资产。
- 指示第三方停止生成和存储代您管理但不再需要的资产。
- 指示第三方整合代表您生成的多余资产。
- 定期审核工作负载以识别和移除未使用的资产。

## 资源

### 相关文档：

- [优化您的 AWS 基础设施以实现可持续性，第 II 部分：存储](#)
- [如何终止我的 AWS 账户中不再需要的活动资源？](#)

### 相关视频：

- [如何检查我的 AWS 账户中是否有不再需要的活动资源，然后移除它们？](#)

## SUS02-BP04 根据其联网要求优化工作负载的地理位置

为工作负载选择可缩短网络流量必须传输的距离的云位置和服务，并减少支持您的工作负载所需的总网络资源。

### 常见反模式：

- 您根据自己所在的位置选择工作负载的区域。

- 您可以将所有工作负载资源整合到一个地理位置中。
- 所有流量都会流经您现有的数据中心。

建立此最佳实践的好处：将工作负载放在接近用户的地方可以提供极低的延迟，同时减少网络中的数据移动并减小对环境的影响。

未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

AWS Cloud 基础设施围绕区域、可用区、置放群组 and 边缘站点（例如，[AWS Outposts](#) 和 [AWS Local Zones](#)）等位置选项而构建。这些位置选项负责维护应用程序组件、云服务、边缘网络和本地数据中心之间的连接。

分析您的工作负载中的网络访问模式，以便确定如何使用这些云位置选项和缩短网络流量必须传输的距离。

实施步骤

- 分析您的工作负载中的网络访问模式，以便确定用户如何使用您的应用程序。
  - 使用监控工具，例如 [Amazon CloudWatch](#) 和 [AWS CloudTrail](#)，以便收集有关网络活动的数据。
  - 分析数据以确定网络访问模式。
- 请根据以下关键元素，为您的工作负载部署选择区域：
  - 您的可持续发展目标：如 [区域选择](#) 中所述。
  - 数据所在位置：对于数据密集型应用程序（如大数据和机器学习），应用程序代码的运行应尽量接近数据。
  - 用户所在位置：对于面向用户的应用程序，选择接近您工作负载用户的一个或多个区域。
  - 其他制约：考虑成本和合规性等制约，如 [为工作负载选择区域时应考虑的事项](#) 中所述。
- 对常用资产使用本地缓存或 [AWS 缓存解决方案](#)，以提高性能，减少数据移动并减小对环境的影响。

| 服务                                | 何时使用                                           |
|-----------------------------------|------------------------------------------------|
| <a href="#">Amazon CloudFront</a> | 用于缓存静态内容（如图像、脚本和视频）以及动态内容（如 API 响应或 Web 应用程序）。 |

| 服务                                   | 何时使用                      |
|--------------------------------------|---------------------------|
| <a href="#">Amazon ElastiCache</a>   | 用于缓存 Web 应用程序的内容。         |
| <a href="#">DynamoDB Accelerator</a> | 用于将内存中加速添加到您的 DynamoDB 表。 |

- 使用有助于您在更接近工作负载用户的位置运行代码的服务：

| 服务                                          | 何时使用                                          |
|---------------------------------------------|-----------------------------------------------|
| <a href="#">Lambda@Edge</a>                 | 用于执行计算密集型操作，当对象不在缓存中时启动这些操作。                  |
| <a href="#">Amazon CloudFront Functions</a> | 用于处理简单使用场景，如 HTTP(S) 请求或响应操作，这些操作可由短期运行的函数启动。 |
| <a href="#">AWS IoT Greengrass</a>          | 用于为互联设备运行本地计算、消息收发和数据缓存。                      |

- 使用连接池来允许连接重用并减少所需资源。
- 使用不依赖于持久连接和同步更新的分布式数据存储来保持一致性，从而为区域人口提供服务。
- 用共享的动态容量代替预先配置的静态网络容量，并与其他订阅用户共享网络容量的可持续性影响。

## 资源

### 相关文档：

- [优化您的 AWS 基础设施以实现可持续性，第 III 部分：联网](#)
- [Amazon ElastiCache 文档](#)
- [什么是 Amazon CloudFront？](#)
- [Amazon CloudFront 主要功能](#)

### 相关视频：

- [揭秘 AWS 上的数据传输](#)
- [在新一代 Amazon EC2 实例上扩展网络性能](#)

相关示例：

- [AWS 联网研讨会](#)
- [针对可持续性设计 – 最大限度地减少跨网络的数据移动](#)

## SUS02-BP05 针对执行的活动优化团队成员资源

优化提供给团队成员的资源，在支持其需求的同时最大程度地降低对环境可持续性的影响。

常见反模式：

- 忽略了团队成员使用的设备对云应用程序整体效率的影响。
- 手动管理和更新团队成员使用的资源。

建立此最佳实践的好处：优化团队成员资源可以提高支持云的应用程序的整体效率。

在未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

了解您的团队成员用来使用服务的资源、它们的预期生命周期，以及财务和可持续性影响。实施战略以优化这些资源。例如，在利用率高的可扩展基础设施上，而不是在利用率不高的强力单用户系统上，执行渲染和编译等复杂的操作。

### 实施步骤

- 按照工作站和其他设备的使用方式对它们进行预置。
- 使用虚拟桌面和应用程序串流来限制升级和设备要求。
- 将处理器或内存密集型任务移至云端以利于其弹性。
- 评估流程和系统对您的设备生命周期的影响，并选择在满足业务需求的同时最大限度减少设备更换需求的解决方案。
- 对设备实施远程管理以减少所需的商务旅行。
  - [AWS Systems Manager Fleet Manager](#) 是一种统一的用户界面 ( UI ) 体验，帮助您远程管理在 AWS 上或在本地运行的节点。

### 资源

相关文档：

- [什么是 Amazon WorkSpaces ?](#)
- [Amazon WorkSpaces 的成本优化器](#)
- [Amazon AppStream 2.0 文档](#)
- [NICE DCV](#)

相关视频：

- [在 AWS 上管理 Amazon WorkSpaces 的成本](#)

## SUS02-BP06 实施缓冲和节流以展平需求曲线

缓冲和节流可展平需求曲线，并降低工作负载所需的预置容量。

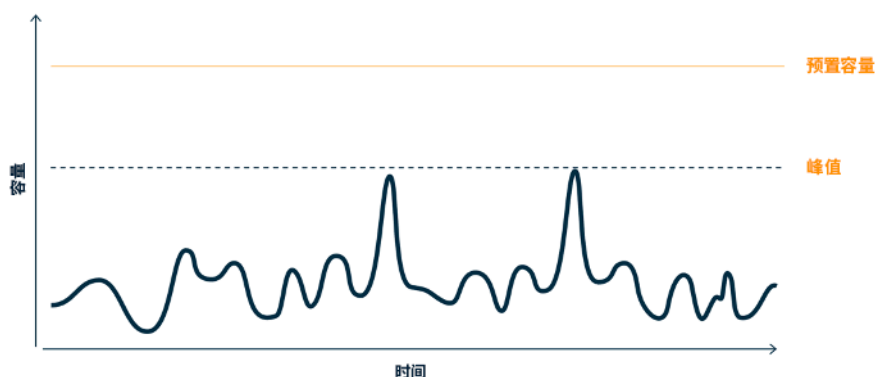
常见反模式：

- 在不需要的时候立即处理客户端请求。
- 没有分析客户端请求的要求。

建立此最佳实践的好处：展平需求曲线可降低工作负载所需的预置容量。降低预置容量即可减少能源消耗和减少对环境的影响。

在未建立这种最佳实践的情况下暴露的风险等级：低

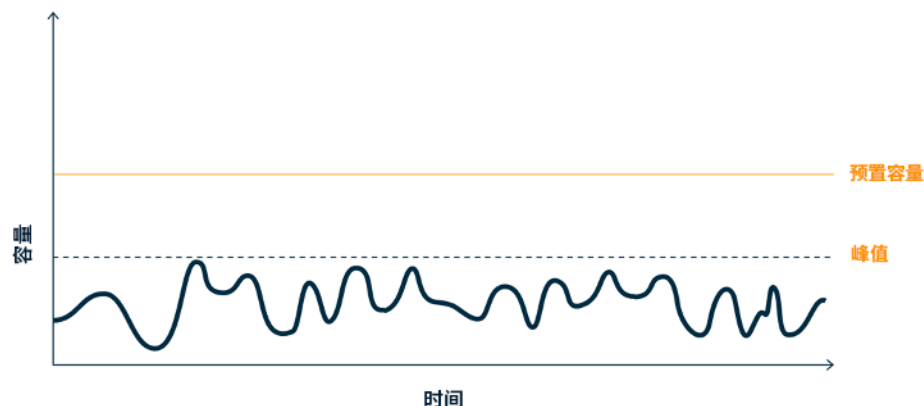
展平工作负载需求曲线有助于降低工作负载的预置容量和减少对环境的影响。假设工作负载的需求曲线如下图所示。此工作负载有两个峰值，为了处理这些峰值，如橙色线所示预置资源容量。因为需要预置容量来处理这两个峰值，所以此工作负载所使用的资源和能量不是由需求曲线下的区域表示，而是由预置容量线下面的区域表示。





需求曲线具有两个不同的峰值，需要高预置容量。

您可以使用缓冲和节流来修改需求曲线和弄平峰值，这意味着可以减少预置容量和消耗的能量。在客户端可以执行重试时实施节流。实施缓冲以存储请求并将处理任务往后推迟一段时间。



节流对需求曲线和预置容量的影响。

## 实施步骤

- 分析客户端请求以确定如何对它们作出响应。要考虑的问题包括：
  - 是否可以异步处理此请求？
  - 客户端是否具有重试能力？
- 如果客户端有重试能力，则您可以实施节流，它会告诉需求源，如果当前无法处理请求，则应稍后再试。
  - 您可以使用 [Amazon API Gateway](#) 来实施节流。
- 对于无法执行重试的客户端，则需要实施缓冲以展平需求曲线。缓冲会延迟请求处理，从而让以不同速率运行的应用程序可以有效通信。基于缓冲的方法使用队列或流来接受来自生产方的消息。然后消息将由使用方读取并处理，这样消息就能够以满足使用方业务需求的速率运行。
  - [Amazon Simple Queue Service \( Amazon SQS \)](#) 是一项托管服务，提供允许单个使用方读取单个消息的队列。
  - [Amazon Kinesis](#) 提供允许众多使用方读取相同消息的流。
- 分析总体需求、变化率和所需的响应时间，以使所需节流或缓冲的大小适宜。

## 资源

相关文档：

- [开始使用 Amazon SQS](#)
- [使用队列和消息的应用程序集成](#)

相关视频：

- [为分布式应用程序选择适合的消息传递服务](#)

## 软件和架构

问题

- [SUS 3 您如何利用软件和架构模式来支持您的可持续发展目标？](#)

### SUS 3 您如何利用软件和架构模式来支持您的可持续发展目标？

实施用于执行负载平滑和保持已部署资源始终如一的高利用率的模式，以最大限度地减少资源消耗。由于用户行为会随着时间的推移而发生变化，因此组件可能会因缺乏使用而变得空闲。修改模式和架构以整合未充分利用的组件，从而提高整体利用率。停用不再需要的组件。了解工作负载组件的性能，并优化消耗资源最多的组件。注意客户用来访问您服务的设备，并实施相应的模式以最大限度地减少设备升级需要。

最佳实践

- [SUS03-BP01 针对异步和计划作业优化软件和架构](#)
- [SUS03-BP02 删除或重构很少或没有使用的工作负载组件](#)
- [SUS03-BP03 优化消耗最多时间或资源的代码区域](#)
- [SUS03-BP04 优化对设备的影响](#)
- [SUS03-BP05 使用最能支持数据访问和存储模式的软件模式和架构](#)

#### SUS03-BP01 针对异步和计划作业优化软件和架构

使用高效的软件和架构模式（如队列驱动）来保持所部署资源的始终如一的高利用率。

常见反模式：

- 为了应对不可预见的需求高峰，您过度预置云工作负载中的资源。
- 架构不会通过消息传递组件分离异步消息的发送方和接收方。

建立此最佳实践的好处：

- 高效的软件和架构模式可以最大程度地减少工作负载中未使用的资源，并提高整体效率。
- 可以独立于异步消息的接收来扩缩处理。
- 通过消息传递组件，可以放宽可用性要求，从而能够用更少的资源来满足这些要求。

在未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

使用高效的架构模式（如[事件驱动型架构](#)），均匀地利用组件并最大程度地减少工作负载中的过度预留。使用高效的架构模式可以最大程度地减少由于需求随时间变化而导致的闲置资源。

了解工作负载组件的要求，并采用可提高资源总体利用率的架构模式。停用不再需要的组件。

实施步骤

- 分析工作负载的需求，以确定如何响应这些需求。
- 对于不需要同步响应的请求或作业，请使用队列驱动型架构和自动扩缩工作线程来最大限度地提高利用率。以下是一些可以考虑采用队列驱动型架构的示例：

| Queuing mechanism                                              | Description                                    |
|----------------------------------------------------------------|------------------------------------------------|
| <a href="#">AWS Batch 作业队列</a>                                 | AWS Batch 作业将提交到作业队列，并一直驻留在队列中，直到可以计划在计算环境中运行。 |
| <a href="#">Amazon Simple Queue Service 和 Amazon EC2 竞价型实例</a> | 将 Amazon SQS 实例和竞价型实例配对，构建容错又高效的架构。            |

- 对于可以随时处理的请求或作业，请使用调度机制批量处理作业以提高效率。以下是 AWS 上的调度机制的一些示例：

| Scheduling mechanism                    | Description                                                    |
|-----------------------------------------|----------------------------------------------------------------|
| <a href="#">Amazon EventBridge 计划程序</a> | <a href="#">Amazon EventBridge</a> 提供的一项功能，允许您大规模创建、运行和管理计划任务。 |

| Scheduling mechanism                                               | Description                                                                 |
|--------------------------------------------------------------------|-----------------------------------------------------------------------------|
| <a href="#">AWS Glue 基于时间的计划</a>                                   | 在 AWS Glue 中为爬网程序和作业定义基于时间的计划。                                              |
| <a href="#">Amazon Elastic Container Service (Amazon ECS) 计划任务</a> | Amazon ECS 支持创建计划任务。计划任务使用 Amazon EventBridge 规则按计划或响应 EventBridge 事件来运行任务。 |
| <a href="#">实例计划程序</a>                                             | 为您的 Amazon EC2 和 Amazon Relational Database Service 实例配置启动和停止计划。            |

- 如果在架构中使用轮询和 Webhook 机制，请将它们替换为事件。使用[事件驱动型架构](#)构建高效的工作负载。
- 利用 [AWS 上的无服务器](#)来消除过量配置的基础设施。
- 适当调整架构中各个组件的大小，以防止等待输入的闲置资源。

## 资源

### 相关文档：

- [什么是 Amazon Simple Queue Service ?](#)
- [什么是 Amazon MQ ?](#)
- [基于 Amazon SQS 进行扩缩](#)
- [什么是 AWS Step Functions ?](#)
- [什么是 AWS Lambda ?](#)
- [将 AWS Lambda 与 Amazon SQS 配合使用](#)
- [什么是 Amazon EventBridge ?](#)

### 相关视频：

- [迁移到事件驱动型架构](#)

## SUS03-BP02 删除或重构很少或没有使用的工作负载组件

移除未使用且不再需要的组件，并重构利用率低的组件，以最大限度减少工作负载中的浪费。

## 常见反模式：

- 没有定期检查工作负载的各个组件的利用率水平。
- 没有检查和分析 AWS 合理调整大小工具（如 [AWS Compute Optimizer](#)）的建议。

建立此最佳实践的好处：移除未使用的组件可最大限度减少浪费并提高云工作负载的整体效率。

在未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

检查您的工作负载以识别空闲或未使用的组件。这是一个迭代改进过程，可以通过需求变化或新云服务的发布来触发。例如，[AWS Lambda](#) 函数执行时间显著缩短，这可能是需要降低内存大小的指标。此外，随着 AWS 发布新的服务和功能，适用于您的工作负载的最佳服务和架构可能会发生变化。

持续监控工作负载活动并寻找机会来提高单个组件的利用水平。通过删除空闲组件并执行合理调整大小活动，您就可以使用最少的云资源来满足您的业务需求。

## 实施步骤

- 监控和捕获工作负载关键组件的利用率指标（例如 [Amazon CloudWatch 指标](#) 中的 CPU 利用率、内存利用率或网络吞吐量）。
- 对于稳定的工作负载，定期检查 AWS 合理调整大小工具（如 [AWS Compute Optimizer](#)），以便识别空闲、未使用或未充分利用的组件。
- 对于临时工作负载，请评估利用率指标以识别空闲、未使用或未充分利用的组件。
- 停用不再需要的组件及关联资产（如 Amazon ECR 映像）。
- 重构未充分利用的组件或将其与其他资源整合以提高利用效率。例如，您可以在单个 [Amazon RDS](#) 数据库实例中预置多个小数据库，而不是在单个未充分利用的实例中运行数据库。
- 了解[您的工作负载为完成一个工作单元而预置的资源](#)。

## 资源

### 相关文档：

- [AWS Trusted Advisor](#)
- [什么是 Amazon CloudWatch？](#)
- [自动清理 Amazon ECR 中未使用的映像](#)

相关示例：

- [Well-Architected 实验室 - 使用 AWS Compute Optimizer 合理调整大小](#)
- [Well-Architected 实验室 - 优化硬件模式并遵守可持续性 KPI](#)

### SUS03-BP03 优化消耗最多时间或资源的代码区域

优化在架构的不同组件中运行的代码，以最大限度地减少资源使用和提高性能。

常见反模式：

- 忽略为资源使用优化代码。
- 通常通过增加资源来应对性能问题。
- 代码审核和开发过程不会跟踪性能变化。

建立此最佳实践的好处：使用高效的代码可以最大限度地减少资源使用量并提高性能。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

至关重要的是检查每个功能区域（包括云架构应用程序的代码）以优化其资源使用 and 性能。持续监控工作负载在构建环境和生产中的性能，并确定改进资源使用率特别高的代码片段的机会。采用定期审核流程来识别代码中资源使用效率低下的错误或反模式。利用可为您的使用场景产生相同结果的简单和高效算法。

### 实施步骤

- 在开发工作负载时，采用自动化代码审核流程来提高质量并识别错误和反模式。
  - [使用 Amazon CodeGuru Reviewer 自动审查代码](#)
  - [使用 Amazon CodeGuru 检查并发错误](#)
  - [使用 Amazon CodeGuru 提高 Python 应用程序的代码质量](#)
- 在运行工作负载时，监测资源以将单个工作单元的资源需求高的组件确定为代码审核目标。
- 对于代码审核，使用代码分析器确定使用时间最长或使用资源最多的代码区域作为优化目标。
  - [借助 Amazon CodeGuru Profiler 减少组织的碳排放](#)
  - [使用 Amazon CodeGuru Profiler 了解 Java 应用程序中的内存使用](#)
  - [通过 Amazon CodeGuru Profiler 改进客户体验并降低成本](#)

- 对工作负载使用最高效的操作系统和编程语言。有关节能编程语言（包括 Rust）的详细信息，请参阅 [Rust 可持续性](#)。
- 使用可产生相同结果的更简单、更高效算法取代计算密集型算法。
- 删除排序和格式等不必要的代码。

## 资源

### 相关文档：

- [什么是 Amazon CodeGuru Profiler？](#)
- [FPGA 实例](#)
- [在 AWS 上进行构建所需工具的 AWS SDK](#)

### 相关视频：

- [使用 Amazon CodeGuru Profiler 提高代码效率](#)
- [使用 Amazon CodeGuru 自动提供代码审核和应用程序性能建议](#)

## SUS03-BP04 优化对设备的影响

了解您的架构中使用的设备，并使用策略来减少其使用。这可以最大限度地减少云工作负载对环境的影响。

### 常见反模式：

- 忽略客户所用设备对环境的影响。
- 手动管理和更新客户使用的资源。

建立此最佳实践的好处：实施针对客户设备优化的软件模式和功能可以减少云工作负载对环境的影响。

在未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

实施针对客户设备优化的软件模式和功能可以从几个方面减少对环境的影响。

- 实施向后兼容的新功能可以减少硬件更换次数。



- 优化应用程序以在设备上高效运行，这有助于降低能耗和延长电池寿命（如果它们由电池供电）。
- 针对设备优化应用程序还可以减少网络上的数据传输。

了解架构中使用的设备、其预期生命周期以及更换这些组件产生的影响。实施软件模式和功能，有助于最大程度地降低设备能耗、减少客户更换设备和手动升级设备的需求。

## 实施步骤

- 列出您架构中使用的设备。设备可以是移动设备、平板电脑、物联网设备、智能灯，甚至是工厂中的智能设备。
- 优化设备上运行的应用程序：
  - 使用策略（例如在后台运行任务）来降低能耗。
  - 在构建有效负载时考虑网络带宽和延迟，并实施有助于您的应用程序在低带宽、高延迟链路上良好运行的功能。
  - 将有效负载和文件转换为设备所需的优化格式。例如，您可以使用 [Amazon Elastic Transcoder](#) 或 [AWS Elemental MediaConvert](#) 将大型、高质量的数字媒体文件转换为用户可以在移动设备、平板电脑、Web 浏览器和联网电视上播放的格式。
  - 在服务器端执行计算密集型活动（例如图像渲染），或使用应用程序串流来改善旧设备上的用户体验。
  - 对输出进行分段和分页，尤其是对于交互式会话，以管理有效负载并限制本地存储要求。
- 使用自动化空中下载（OTA）机制将更新部署到一个或多个设备。
  - 您可以使用 [CI/CD 管道](#) 更新移动应用程序。
  - 您可以使用 [AWS IoT Device Management](#) 大规模地远程管理互联设备。
- 要测试新功能和更新，请使用具有代表性硬件集的托管式设备场，并迭代开发以最大限度增加支持的设备数。有关更多详细信息，请参阅 [SUS06-BP04 使用托管式设备场进行测试](#)。

## 资源

### 相关文档：

- [什么是 AWS Device Farm？](#)
- [Amazon AppStream 2.0 文档](#)
- [NICE DCV](#)
- [在运行 FreeRTOS 的设备上更新固件的 OTA 教程](#)

相关视频：

- [AWS Device Farm 简介](#)

## SUS03-BP05 使用最能支持数据访问和存储模式的软件模式和架构

了解数据在工作负载中的使用方式、用户使用数据的方式，以及数据的传输和存储方式。使用最能支持数据访问和存储的软件模式和架构，最大限度地减少支持工作负载所需的计算、网络 and 存储资源。

常见反模式：

- 假设所有工作负载都具有相似的数据存储和访问模式。
- 假设所有工作负载都位于一个存储层，且只使用该存储层。
- 假设数据访问模式会随着时间的推移保持一致。
- 您的架构支持潜在的高数据访问突发，这会导致资源大部分时间都处于空闲状态。

建立此最佳实践的好处：根据数据访问和存储模式选择和优化架构将有助于降低开发复杂性并提高总体利用率。了解何时使用全局表、数据分区和缓存将帮助您减少运营开销，并根据您的工作负载需求进行扩展。

在未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

使用最符合您的数据特性和访问模式的软件和架构模式。例如，使用 [AWS 上的现代数据架构](#) 使您可以使用为您的独特分析应用场景优化的专用服务。这些架构模式可提高数据处理效率和减少资源使用。

### 实施步骤

- 分析您的数据特性和访问模式，以便确定云资源的适合配置。要考虑的关键特征包括：
  - 数据类型：结构化、半结构化、非结构化
  - 数据增长：有界、无界
  - 数据耐久性：持久、短暂、瞬时
  - 访问模式：读或写、更新频率、峰值或一致
- 使用最能支持数据访问和存储模式的架构模式。
  - [让我们来构建！现代数据架构](#)
  - [AWS 上的数据库：根据作业选择合适工具](#)

- 使用可以原生处理压缩数据的技术。
- 为您架构中的数据处理使用专用[分析服务](#)。
- 使用最能支持您的主导查询模式的数据库引擎。管理您的数据库索引以确保高效的查询执行。有关更多详情，请参阅 [AWS 数据库](#)。
- 选择可减少架构中所用网络容量的网络协议。

## 资源

### 相关文档：

- [Athena 压缩支持文件格式](#)
- [使用 Amazon Redshift 从列数据格式复制](#)
- [在 Firehose 中转换您的输入记录格式](#)
- [AWS Glue 中 ETL 输入和输出的格式选项](#)
- [通过转换为列格式提高 Amazon Athena 上的查询性能](#)
- [使用 Amazon Redshift 从 Amazon S3 加载压缩数据文件](#)
- [使用 Amazon Aurora 上的 Performance Insights 监视数据库负载](#)
- [使用 Amazon RDS 上的 Performance Insights 监视数据库负载](#)
- [Amazon S3 Intelligent-Tiering 存储类](#)

### 相关视频：

- [在 AWS 上构建现代数据架构](#)

## 数据

### 问题

- [SUS 4 您如何利用数据管理策略和模式来支持可持续发展目标？](#)

### SUS 4 您如何利用数据管理策略和模式来支持可持续发展目标？

实施数据管理实践以减少支持工作负载所需的预置存储，以及使用存储所需的资源。了解您的数据，并使用能够更有效地支持数据的商业价值及其使用方式的存储技术和配置。当需求减少时，将数据移到更高效、性能更低的存储中，并删除不再需要的数据。

## 最佳实践

- [SUS04-BP01 实施数据分类策略](#)
- [SUS04-BP02 使用支持数据访问和存储模式的技术](#)
- [SUS04-BP03 使用策略管理数据集的生命周期](#)
- [SUS04-BP04 使用弹性和自动化来扩展数据块存储或文件系统](#)
- [SUS04-BP05 删除不需要或多余的数据](#)
- [SUS04-BP06 使用共享文件系统或存储来访问通用数据](#)
- [SUS04-BP07 最大限度地减少跨网络的数据移动](#)
- [SUS04-BP08 仅在难以重新创建时备份数据](#)

### SUS04-BP01 实施数据分类策略

对数据进行分类，以了解其对业务成果的重要性，并选择合适的节能存储层来存储数据。

常见反模式：

- 您没有识别正在处理或存储的具有类似特征（如敏感性、业务关键性或监管要求）的数据资产。
- 您没有实施数据目录来清点您的数据资产。

建立此最佳实践的好处：实施数据分类策略使您能够为数据确定最节能的存储层。

在未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

数据分类涉及识别在由组织拥有或运营的信息系统中正在处理和存储的数据类型。它还涉及到对数据的重要性以及数据泄露、丢失或滥用的可能影响进行判断。

实施数据分类策略时，要从数据的使用情境进行反推，并创建一个分类方案，该方案考虑到特定数据集对组织运营的重要程度。

### 实施步骤

- 对您的工作负载存在的各种数据类型进行清点。
  - 有关数据分类类别的更多详情，请参阅 [《数据分类》白皮书](#)。
- 根据给组织带来的风险，确定数据的重要性、机密性、完整性和可用性。使用这些要求将数据分组到您采用的数据分类层之一。

- 例如，请参阅[将数据分类并保护初创公司的四个简单步骤](#)。
- 针对未标记和未分类的数据定期审核您的环境，并对数据进行适当的分类和标记。
  - 例如，请参阅 [AWS Glue 中的数据目录和爬网程序](#)。
- 建立一个提供审计和治理功能的数据目录。
- 确定并记录每个数据类的处理过程。
- 使用自动化来持续审计您的环境，以识别未标记和未分类的数据，并适当地对这些数据进行分类和标记。

## 资源

### 相关文档：

- [利用 AWS Cloud 支持数据分类](#)
- [AWS Organizations 中的标记策略](#)

### 相关视频：

- [在 AWS 上实现敏捷的数据治理](#)

## SUS04-BP02 使用支持数据访问和存储模式的技术

使用最能支持您的数据访问和存储方式的存储技术，以在支持您的工作负载的同时最大限度地减少预置资源。

### 常见反模式：

- 假设所有工作负载都具有相似的数据存储和访问模式。
- 假设所有工作负载都位于一个存储层，且只使用该存储层。
- 假设访问模式始终保持不变。

建立此最佳实践的好处：根据数据访问和存储模式选择和优化您的存储技术，有助于您减少满足业务需求所需的云资源，并提高云工作负载的整体效率。

未建立这种最佳实践的情况下暴露的风险等级：低

实施指导

选择最适合您的访问模式的存储解决方案，或者考虑根据存储解决方案更改访问模式，以便尽可能提高性能和效率。

- 评估您的数据特征和访问模式，以收集您的存储需求的关键特征。要考虑的关键特征包括：
  - 数据类型：结构化、半结构化、非结构化
  - 数据增长：限界、不限界
  - 数据持久性：持久、短暂、瞬时
  - 访问模式：读或写、频率、峰值或一致
- 将数据迁移到支持您的数据特征和访问模式的适当存储技术。下面是 AWS 存储技术的一些示例以及它们的关键特征：

| 类型     | 技术                                                        | 主要特征                                                                                                                                                        |
|--------|-----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 对象存储   | <a href="#">Amazon S3</a>                                 | 一项对象存储服务，具有无限的可扩展性、高可用性和多种可访问性选项。在 Amazon S3 内外传输和访问对象可以使用诸如 <a href="#">Transfer Acceleration</a> 或 <a href="#">Access Points</a> 之类的服务，来满足您的位置、安全需求和访问模式。 |
| 存档存储   | <a href="#">Amazon S3 Glacier</a>                         | Amazon S3 的存储类，用于数据归档。                                                                                                                                      |
| 共享文件系统 | <a href="#">Amazon Elastic File System ( Amazon EFS )</a> | 可由多种类型的计算解决方案访问的可挂载文件系统。Amazon EFS 会自动增大和缩小存储，并进行性能优化以提供一致的低延迟。                                                                                             |
| 共享文件系统 | <a href="#">Amazon FSx</a>                                | 基于最新 AWS 计算解决方案而构建，支持四种常用文件系统：NetApp ONTAP、OpenZFS、Windows 文件服务器                                                                                            |

| 类型    | 技术                                                                                           | 主要特征                                                                                                                                  |
|-------|----------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|
|       |                                                                                              | 和 Lustre。Amazon FSx <a href="#">延迟、吞吐量和 IOPS</a> 因文件系统而不同，因此，在为您的工作负载需求选择合适的文件系统时应考虑这些因素。                                             |
| 数据块存储 | <a href="#">Amazon Elastic Block Store ( Amazon EBS )</a>                                    | 设计用于 Amazon Elastic Compute Cloud ( Amazon EC2 ) 的可扩展、高性能的块存储服务。Amazon EBS 包括用于事务性、IOPS 密集型工作负载的基于 SSD 的存储，以及用于吞吐量密集型工作负载的基于 HDD 的存储。 |
| 关系数据库 | <a href="#">Amazon Aurora</a> 、 <a href="#">Amazon RDS</a> 、 <a href="#">Amazon Redshift</a> | 旨在支持 ACID ( 原子性、一致性、隔离性、持久性 ) 事务，并保持参照完整性和数据强一致性。许多传统应用程序、企业资源规划 ( ERP ) 、客户关系管理 ( CRM ) 和电子商务系统都使用关系数据库来存储数据。                        |
| 键值数据库 | <a href="#">Amazon DynamoDB</a>                                                              | 已针对常见的访问模式进行优化，通常用于存储和检索大量数据。键值数据库的典型使用案例包括高流量 Web 应用程序，电子商务系统和游戏应用程序。                                                                |

- 对于固定大小的存储系统 ( 例如 Amazon EBS 或 Amazon FSx ) ，请监控可用的存储空间，并在达到阈值时自动分配存储空间。您可以利用 Amazon CloudWatch 来收集和分析 [Amazon EBS](#) 和 [Amazon FSx](#)的不同指标。
- Amazon S3 存储类可以在对象级别进行配置，一个桶可以包含存储在所有存储类中的对象。



- 您也可以使用 Amazon S3 生命周期策略，在不对应用程序进行任何更改的情况下，于存储类之间自动转换对象或删除数据。通常来说，在考虑这些存储机制时，您必须在资源效率、访问延迟和可靠性之间做出取舍。

## 资源

### 相关文档：

- [Amazon EBS 卷类型](#)
- [Amazon EC2 实例存储](#)
- [Amazon S3 Intelligent-Tiering](#)
- [Amazon EBS I/O 特性](#)
- [使用 Amazon S3 存储类](#)
- [什么是 Amazon S3 Glacier？](#)

### 相关视频：

- [AWS 上数据湖的架构模式](#)
- [深入讨论 Amazon EBS \( STG303-R1 \)](#)
- [利用 Amazon S3 优化存储性能 \( STG343 \)](#)
- [在 AWS 上构建现代数据架构](#)

### 相关示例：

- [Amazon EFS CSI 驱动程序](#)
- [Amazon EBS CSI 驱动程序](#)
- [Amazon EFS 实用程序](#)
- [Amazon EBS 自动扩展](#)
- [Amazon S3 示例](#)

## SUS04-BP03 使用策略管理数据集的生命周期

管理所有数据的生命周期并自动执行删除，以最大限度地减少工作负载所需的总存储。

### 常见反模式：

- 手动删除数据。
- 不删除任何工作负载数据。
- 不根据数据的保留和访问要求将数据移动到更节能的存储层。

建立此最佳实践的好处：使用数据生命周期策略可确保在工作负载中高效访问和保留数据。

在未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

数据集在其生命周期中通常具有不同的保留和访问要求。例如，应用程序可能需要在有限的时间段内频繁访问某些数据集。之后，这些数据集很少被访问。

要在数据集的整个生命周期内高效管理数据集，请配置生命周期策略，这些策略是定义如何处理数据集的规则。

使用生命周期配置规则，您可以指示特定存储服务将数据集转换到更节能的存储层、将其存档或删除。

实施步骤

- [对工作负载中的数据集进行分类。](#)
- 定义每个数据类的处理过程。
- 设置自动化生命周期策略以强制实施生命周期规则。以下是如何为不同 AWS 存储服务设置自动化生命周期策略的一些示例：

| Storage service                            | How to set automated lifecycle policies                                                                                                                                                                             |
|--------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <a href="#">Amazon S3</a>                  | 您可以使用 <a href="#">Amazon S3 生命周期</a> 在对象的整个生命周期中对其进行管理。如果访问模式未知、不断变化或不可预测，可以使用 <a href="#">Amazon S3 智能分层</a> ，此功能可监控访问模式，并自动将尚未访问的对象移动到成本较低的访问层。您可以利用 <a href="#">Amazon S3 Storage Lens</a> 指标来识别生命周期管理中的优化机会和差距。 |
| <a href="#">Amazon Elastic Block Store</a> | 您可以使用 <a href="#">Amazon Data Lifecycle Manager</a> 自动创建、保留和删除 Amazon EBS 快照和 Amazon EBS 支持的 AMI。                                                                                                                   |

| Storage service                                   | How to set automated lifecycle policies                     |
|---------------------------------------------------|-------------------------------------------------------------|
| <a href="#">Amazon Elastic File System</a>        | <a href="#">Amazon EFS 生命周期管理</a> 可自动管理文件系统的文件存储。           |
| <a href="#">Amazon Elastic Container Registry</a> | <a href="#">Amazon ECR 生命周期策略</a> 通过根据期限或计数使映像过期来自动清理容器映像。  |
| <a href="#">AWS Elemental MediaStore</a>          | 您可以使用 <a href="#">对象生命周期策略</a> 来控制对象应在MediaStore 容器中存储多长时间。 |

- 删除未使用的卷、快照和超出保留期的数据。利用本机服务功能（如 Amazon DynamoDB 生存时间或 Amazon CloudWatch 日志保留）进行删除。
- 在适当情况下根据生命周期规则汇总和压缩数据。

资源

相关文档：

- [通过 Amazon S3 存储类分析优化 Amazon S3 生命周期规则](#)
- [使用 AWS Config 规则 评估资源](#)

相关视频：

- [利用 Amazon S3 生命周期简化数据生命周期并优化存储成本](#)
- [使用 Amazon S3 Storage Lens 降低存储成本](#)

SUS04-BP04 使用弹性和自动化来扩展数据块存储或文件系统

随着数据的增长，使用弹性和自动化来扩展数据块存储或文件系统，以便最大限度减少总预置存储。

常见反模式：

- 购买大型数据块存储或文件系统以备将来需要。
- 过度预置文件系统的每秒输入和输出操作数（IOPS）。
- 不监控数据卷的利用率。

建立此最佳实践的好处：最大限度地减少存储系统的过度预置可减少空闲资源并提高工作负载的整体效率。

在未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

根据适合工作负载的大小分配、吞吐量和延迟，创建数据块存储和文件系统。随着数据的增长，使用弹性和自动化来扩展数据块存储或文件系统，而无需过度预置这些存储服务。

## 实施步骤

- 对于固定大小存储（例如 [Amazon EBS](#)），请确保您正在监控已使用存储量与总体存储量大小之间的关系，如果可以，请创建自动化流程，以便在达到阈值时增加存储大小。
- 使用弹性卷和托管式数据块存储服务，随着持久性数据的增长自动分配额外的存储。例如，您可以使用 [Amazon EBS 弹性卷](#) 来更改卷大小、卷类型或调整 Amazon EBS 卷的性能。
- 为您的文件系统选择适合的存储类、性能模式和吞吐量模式，以满足您的业务需求，不要超过这个需求。
  - [Amazon EFS 性能](#)
  - [Linux 实例上的 Amazon EBS 卷性能](#)
- 为您的数据卷设置目标利用率水平，并调整超出预期范围的卷大小。
- 合理调整只读卷的大小以适应数据。
- 将数据迁移到对象存储，以避免使用数据块存储上的固定卷大小预置多余容量。
- 定期检查弹性卷和文件系统，终止空闲卷并缩减过度预置的资源，以适应当前数据大小。

## 资源

相关文档：

- [Amazon FSx 文档](#)
- [什么是 Amazon Elastic File System？](#)

相关视频：

- [深入了解 Amazon EBS 弹性卷](#)
- [提高性能和节省成本的 Amazon EBS 和快照优化策略](#)

• [使用最佳实践优化 Amazon EFS 以节省成本和提高性能](#)

SUS04-BP05 删除不需要或多余的数据

删除不需要或多余的数据，以最大程度地减少存储数据集所需的存储资源。

常见反模式：

- 复制可以轻松获取或重新创建的数据。
- 备份所有数据时不考虑其重要性。
- 只不定期地删除数据、操作事件时删除数据，或者根本不删除数据。
- 无论存储服务的持久性如何，都冗余地存储数据。
- 在没有任何业务理由的情况下启用 Amazon S3 版本控制。

建立此最佳实践的好处：删除不需要的数据可减少工作负载所需的存储大小和工作负载对环境的影响。

在未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

请勿存储不需要的数据。自动删除不需要的数据。使用技术在文件和数据块级别进行重复数据删除。利用服务的本机数据复制和冗余功能。

实施步骤

- 评估是否可以通过使用 [AWS Data Exchange](#) 中的现有公开可用数据集，以及 [AWS 上的开放数据](#) 来避免存储数据。
- 使用可以在数据块和对象级别删除重复数据的机制。以下是有关如何删除 AWS 上的重复数据的一些示例：

| Storage service            | Deduplication mechanism                                                                            |
|----------------------------|----------------------------------------------------------------------------------------------------|
| <a href="#">Amazon S3</a>  | 使用 <a href="#">AWS Lake Formation FindMatches</a> 通过新的 FindMatches ML 转换在数据集（包括没有标识符的数据集）中查找匹配的记录。 |
| <a href="#">Amazon FSx</a> | 在 Amazon FSx for Windows 上启用 <a href="#">重复数据删除</a> 。                                              |

| Storage service                               | Deduplication mechanism              |
|-----------------------------------------------|--------------------------------------|
| <a href="#">Amazon Elastic Block Store 快照</a> | 快照是增量备份，这意味着只保存拍摄最新快照之后出现更改的设备上的数据块。 |

- 分析数据访问以识别不需要的数据。自动执行生命周期策略。利用本机服务功能 ( 如 [Amazon DynamoDB 生存时间](#)、[Amazon S3 生命周期](#) 或 [Amazon CloudWatch 日志保留](#) ) 进行删除。
- 使用 AWS 上的数据虚拟化功能在源头维护数据并避免数据重复。
  - [AWS 上的云原生数据虚拟化](#)
  - [实验：使用 Amazon Redshift 数据共享优化数据模式](#)
- 使用可进行增量备份的备份技术。
- 利用 [Amazon S3](#) 的持久性和 [Amazon EBS 的复制性](#) 来实现持久性目标，而不是使用自我管理技术 [如独立磁盘冗余阵列 ( RAID ) ]。
- 集中日志和跟踪数据，对相同的日志条目进行重复数据删除，并在需要时建立调整详细程度的机制。
- 仅在合理的情况下预填充缓存。
- 建立缓存监控和自动化以相应地调整缓存大小。
- 推送新版本的工作负载时，从对象存储和边缘缓存中删除过时的部署和资产。

## 资源

### 相关文档：

- [更改 CloudWatch Logs 中的日志数据留存](#)
- [Amazon FSx for Windows File Server 上的重复数据删除](#)
- [Amazon FSx for ONTAP 的功能，包括重复数据删除](#)
- [使 Amazon CloudFront 上的文件失效](#)
- [使用 Amazon EFS 备份和还原 AWS Backup 文件系统](#)
- [什么是 Amazon CloudWatch Logs ?](#)
- [在 Amazon RDS 上使用备份](#)

### 相关视频：

- [使用适用于 AWS Lake Formation 的 ML Transforms 进行模糊匹配和重复数据删除](#)

相关示例：

- [如何使用 Amazon Athena 分析我的 Amazon S3 服务器访问日志？](#)

SUS04-BP06 使用共享文件系统或存储来访问通用数据

采用共享文件系统或存储以避免数据重复，并为您的工作负载提供更高效的基础设施。

常见反模式：

- 为每个客户端预置存储。
- 未卸下不活动的客户端的数据卷。
- 不提供跨平台和系统的存储访问。

建立此最佳实践的好处：使用共享文件系统或存储实现将数据共享到一个或多个使用者，而无需复制数据。这有助于减少工作负载所需的存储资源。

在未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

如果您有多个用户或应用程序访问同一个数据集，则使用共享存储技术很重要，这可以为工作负载提供高效的基础设施。共享存储技术提供一个位置来集中存储和管理数据集并避免数据重复。它还加强了不同系统之间数据的一致性。此外，因为多个计算资源会同时并行访问和处理数据，所以利用共享存储技术可以更高效地使用计算能力。

仅在需要时才从这些共享存储服务中提取数据，并卸下未使用的卷以释放资源。

实施步骤

- 当数据具有多个使用者时，将数据迁移到共享存储。下面是 AWS 上的共享存储技术的一些示例：

| Storage option                  | When to use                                                       |
|---------------------------------|-------------------------------------------------------------------|
| <a href="#">Amazon EBS 多重挂载</a> | Amazon EBS 多重挂载使您可以将单个预调配 IOPS SSD ( io1 或 io2 ) 卷挂载到同一可用区中的多个实例。 |
| <a href="#">Amazon EFS</a>      | 请参阅 <a href="#">何时选择 Amazon EFS</a> 。                             |



| Storage option             | When to use                                                        |
|----------------------------|--------------------------------------------------------------------|
| <a href="#">Amazon FSx</a> | 请参阅 <a href="#">选择 Amazon FSx 文件系统</a> 。                           |
| <a href="#">Amazon S3</a>  | 不需要文件系统结构而旨在与对象存储一起使用的应用程序可以使用 Amazon S3 作为可大规模扩展、持久、低成本的对象存储解决方案。 |

- 仅在需要将数据复制到共享文件系统或从共享文件系统提取数据。例如，您可以创建[采用 Amazon S3 的 Amazon FSx for Lustre 文件系统](#)，并仅将处理作业所需的数据子集加载到 Amazon FSx。
- 根据您的使用模式适当删除数据，如[SUS04-BP03 使用策略管理数据集的生命周期](#)所述。
- 将卷与未积极使用它们的客户端分离。

## 资源

### 相关文档：

- [将文件系统链接到 Amazon S3 存储桶](#)
- [在无服务器应用程序中使用适用于 AWS Lambda 的 Amazon EFS](#)
- [Amazon EFS Intelligent-Tiering 通过改变访问模式优化工作负载成本](#)
- [将 Amazon FSx 与本地数据存储库配合使用](#)

### 相关视频：

- [使用 Amazon EFS 进行存储成本优化](#)

## SUS04-BP07 最大限度地减少跨网络的数据移动

使用共享文件系统或对象存储来访问通用数据，并最大限度地减少支持工作负载数据移动所需的总网络资源。

### 常见反模式：

- 不管数据用户位于何处，将所有数据存储在同一 AWS 区域。
- 在网络中移动数据之前不优化数据大小和格式。

建立此最佳实践的好处： 优化跨网络的数据移动可以减少工作负载所需的总网络资源，并降低对环境的影响。

未建立这种最佳实践的情况下暴露的风险等级： 中

实施指导

在组织中移动数据需要计算、网络 and 存储资源。使用相应的技术最大程度地减少数据移动并提高工作负载的整体效率。

实施步骤

- 在 [为工作负载选择区域](#) 时，考虑将与数据或用户的接近程度作为决策因素。
- 按区域对使用的服务进行分区，以便将其特定于区域的数据存储在使用它的区域内。
- 使用高效的文件格式（如 Parquet 或 ORC），并在通过网络移动数据之前先对其进行压缩。
- 不移动未使用的数据。一些让您能够避免移动未使用数据的示例：
  - 将 API 响应缩减到仅针对相关数据。
  - 聚合详细数据（不需要记录级别信息）。
  - 请参阅 [Well-Architected 实验室 – 使用 Amazon Redshift 数据共享优化数据模式](#)。
  - 考虑 [AWS Lake Formation 中的跨账户数据共享](#)。
- 使用有助于您在更接近工作负载用户的位置运行代码的服务。

| 服务                                   | 何时使用                                         |
|--------------------------------------|----------------------------------------------|
| <a href="#">Lambda@Edge</a>          | 用于计算密集型操作，当对象不在缓存中时会运行这些操作。                  |
| <a href="#">CloudFront Functions</a> | 用于简单使用场景（如 HTTP(S) 请求/响应操作），这些操作可由短期运行的函数启动。 |
| <a href="#">AWS IoT Greengrass</a>   | 为互联设备运行本地计算、消息收发和数据缓存。                       |

资源

相关文档：

- [优化您的 AWS 基础设施以实现可持续性，第 III 部分：联网](#)

- [AWS 全球基础设施](#)
- [Amazon CloudFront 主要功能，包括 CloudFront 全球边缘网络](#)
- [在 Amazon OpenSearch Service 中压缩 HTTP 请求](#)
- [使用 Amazon EMR 进行中间数据压缩](#)
- [将压缩数据文件从 Amazon S3 加载到 Amazon Redshift](#)
- [通过 Amazon CloudFront 提供压缩文件](#)

相关视频：

- [揭秘 AWS 上的数据传输](#)

相关示例：

- [针对可持续性设计 – 最大限度地减少跨网络的数据移动](#)

SUS04-BP08 仅在难以重新创建时备份数据

避免备份没有商业价值的数据，尽量减少工作负载的存储资源需求。

常见反模式：

- 没有为数据制定备份策略。
- 备份可以轻松重新创建的数据。

建立此最佳实践的好处：避免备份非关键数据可减少工作负载所需的存储资源并降低其对环境的影响。

在未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

避免备份不必要的数据有助于降低成本和减少工作负载使用的存储资源。仅备份具有商业价值或满足合规性要求所必需的数据。检查备份策略并在恢复方案中排除没有价值的临时存储。

实施步骤

- 实施数据分类策略，如[SUS04-BP01 实施数据分类策略](#)中所述。
- 根据重要性对数据进行分类，并根据[恢复时间目标 \( RTO \) 和恢复点目标 \( RPO \)](#)设计备份策略。避免备份非关键数据。

- 排除可以轻松重新创建的数据。
- 从备份中排除临时数据。
- 排除数据的本地副本，除非从公共位置恢复该数据所需的时间会超过您的服务等级协议（SLA）。
- 使用自动化解决方案或托管服务来备份关键业务数据。
  - [AWS Backup](#) 是一项完全托管式服务，可让您轻松集中自动化云端和本地不同 AWS 服务的数据保护。有关如何使用 AWS Backup 创建自动备份的动手实践指导，请参阅 [Well-Architected 实验室 - 测试数据的备份和恢复](#)。
  - [使用 AWS Backup 自动备份和优化 Amazon EFS 的备份成本](#)。

## 资源

### 相关最佳实践：

- [REL09-BP01 识别和备份需要备份的所有数据，或从源复制数据](#)
- [REL09-BP03 自动执行数据备份](#)
- [REL13-BP02 使用定义的恢复策略来实现恢复目标](#)

### 相关文档：

- [使用 AWS Backup 备份和还原 Amazon EFS 文件系统](#)
- [Amazon EBS 快照](#)
- [在 Amazon Relational Database Service 上使用备份](#)
- [APN 合作伙伴：可以帮助进行备份的合作伙伴](#)
- [AWS Marketplace：可以用于备份的产品](#)
- [备份 Amazon EFS](#)
- [适用于 Windows File Server 的备份 Amazon FSx](#)
- [Amazon ElastiCache \(Redis OSS\) 的备份和还原](#)

### 相关视频：

- [AWS re:Invent 2021 - 使用 AWS 进行备份、灾难恢复和勒索软件防护](#)
- [AWS Backup 演示：跨账户和跨区域备份](#)
- [AWS re:Invent 2019：深入了解 AWS Backup，主讲：Rackspace \(STG341\)](#)

相关示例：

- [Well-Architected 实验室 - 测试数据的备份与还原](#)
- [Well-Architected 实验室 - 面向分析工作负载的备份和还原 \( 具备失效自动恢复功能 \)](#)
- [Well-Architected 实验室 - 灾难恢复 - 备份与还原](#)

## 硬件和服务

问题

- [SUS 5 您如何选择并使用架构中的云硬件和服务来支持自己的可持续发展目标？](#)

### SUS 5 您如何选择并使用架构中的云硬件和服务来支持自己的可持续发展目标？

寻找机会，通过更改硬件管理实践来降低工作负载可持续性影响。最大限度地减少预置和部署所需的硬件数量，并为您的各项工作负载选择最高效的硬件和服务。

最佳实践

- [SUS05-BP01 使用最少的硬件来满足您的需求](#)
- [SUS05-BP02 使用影响最小的实例类型](#)
- [SUS05-BP03 使用托管服务](#)
- [SUS05-BP04 优化基于硬件的计算加速器的使用](#)

#### SUS05-BP01 使用最少的硬件来满足您的需求

为您的工作负载使用最少的硬件，高效地满足您的业务需求。

常见反模式：

- 不监控资源使用率。
- 架构中有利用率较低的资源。
- 没有检查静态硬件的利用率以确定是否应调整大小。
- 没有根据业务 KPI 为计算基础设施设置硬件利用率目标。

建立此最佳实践的好处：合理调整云资源的大小有助于减少工作负载对环境的影响、节省资金并保持性能基准。

在未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

以最佳方式选择工作负载所需的硬件总数，以提高其整体效率。AWS Cloud 让您能够通过各种机制（如 [AWS Auto Scaling](#)）灵活地动态扩展或缩减资源数，并满足不断变化的需求。它还提供 [API 和 SDK](#)，让您可以轻松修改资源。使用这些功能经常更改工作负载实施。此外，按照 AWS 工具中的合理调整大小准则高效地运营您的云资源和满足您的业务需求。

## 实施步骤

- 选择尽可能满足您的需求的实例类型。
  - [如何为我的工作负载选择适当的 Amazon EC2 实例类型？](#)
  - [基于属性为 Amazon EC2 实例集选择实例类型。](#)
  - [使用基于属性的实例类型选择来创建 Auto Scaling 组。](#)
- 通过小增量扩缩来适应可变的工作负载。
- 使用多个计算购买选项，在实例灵活性、可扩展性和成本节省方面实现平衡。
  - [按需型实例](#)最适合新的、有状态和突增工作负载，这些工作负载不能灵活地调整实例类型、位置或时间。
  - [竞价型实例](#)为容错和灵活的应用程序提供了很好的补充选择。
  - 将 [Compute Savings Plans](#) 用于稳定状态的工作负载，当您的需求（如可用区、区域、实例系列或实例类型）发生变化时，这种工作负载可以允许灵活性。
- 使用实例和可用区多样性最大限度地提高应用程序可用性和尽可能利用过剩的容量。
- 使用来自 AWS 工具的合理调整大小建议来调整工作负载。
  - [AWS Compute Optimizer](#)
  - [AWS Trusted Advisor](#)
- 协商服务等级协议（SLA），允许暂时减少容量，同时利用自动化功能部署替换资源。

## 资源

相关文档：

- [优化您的 AWS 基础设施以实现可持续性，第 I 部分：计算](#)
- [基于属性为 Amazon EC2 实例集的 Auto Scaling 选择实例类型](#)
- [AWS Compute Optimizer 文档](#)

- [运行 Lambda：性能优化](#)
- [弹性伸缩文档](#)

相关视频：

- [构建成本、能源和资源高效的计算环境](#)

相关示例：

- [Well-Architected 实验室 - 在启用 AWS Compute Optimizer 和内存利用率的情况下合理调整大小 \( 级别 200\)](#)

### SUS05-BP02 使用影响最小的实例类型

持续监控和使用新实例类型以充分利用能源效率改进。

常见反模式：

- 您只使用一个系列的实例。
- 您只使用 x86 实例。
- 您在 Amazon EC2 Auto Scaling 配置中指定一种实例类型。
- 您使用 AWS 实例的方式与其预期用途不匹配（例如，您将计算优化的实例用于内存密集型工作负载）。
- 您没有定期评估新的实例类型。
- 您不查看 AWS 合理调整大小工具（如 [AWS Compute Optimizer](#)）的建议。

建立此最佳实践的好处：通过使用节能且大小合适的实例，您可以大大减小工作负载对环境的影响并降低其成本。

未建立这种最佳实践的情况下暴露的风险等级：中

### 实施指导

在云工作负载中使用高效的实例对于降低资源使用率和成本效益至关重要。持续监控新实例类型的发布并利用能效改进，包括那些旨在支持特定工作负载（例如机器学习训练和推理以及视频转码）的实例类型。



## 实施步骤

- 学习和探索可以减小工作负载对环境影响的实例类型。
  - 订阅 [AWS 新增功能](#) 及时了解新增的 AWS 技术和实例。
  - 了解不同的 AWS 实例类型。
  - 通过观看如下视频，了解基于 AWS Graviton 的实例（这些实例在 Amazon EC2 中每瓦能耗方面提供出色性能）：[re:Invent 2020 - 深入了解 AWS Graviton2 处理器提供支持的 Amazon EC2 实例](#) 和 [深入了解 AWS Graviton3 和 Amazon EC2 C7g 实例](#)。
- 规划工作负载并将其转换为影响极小的实例类型。
  - 定义一个流程来评估工作负载的新功能或实例。利用云中的敏捷性，快速测试新的实例类型如何改善工作负载的环境可持续性。使用代理指标来衡量完成一个单元的工作需要多少资源。
  - 如有可能，修改工作负载以使用不同数量的 vCPU 和不同数量的内存，以最大限度地增加您的实例类型选项。
  - 考虑将工作负载转换为基于 Graviton 的实例，以提高工作负载的性能效率。
    - [AWS Graviton Fast Start](#)
    - [将工作负载转换为基于 AWS Graviton 的 Amazon Elastic Compute Cloud 实例时的注意事项](#)
    - [适用于 ISV 的 AWS Graviton2](#)
  - 考虑选择 AWS Graviton 选项（在使用 [AWS 托管服务时](#)）。
  - 将工作负载迁移到提供对可持续性影响极小的实例且仍满足您的业务要求的区域。
  - 对于机器学习工作负载，请利用特定于工作负载的专用硬件，例如 [AWS Trainium](#)、[AWS Inferentia](#) 和 [Amazon EC2 DL1](#)。Inf2 等 AWS Inferentia 实例，相比同类同类 Amazon EC2 实例，性能功耗比提升了 50%。
  - 使用 [Amazon SageMaker AI Inference Recommender](#) 来合理调整机器学习推理端点的大小。
  - 对于突增工作负载（不经常需要额外容量的工作负载），请使用 [可突增性能实例](#)。
  - 对于无状态和容错工作负载，请使用 [Amazon EC2 竞价型实例](#) 提高云的整体利用率并减少未使用资源对可持续性的影响。
- 运营和优化您的工作负载实例。
  - 对于临时工作负载，请评估 [实例 Amazon CloudWatch 指标](#)（例如 CPUUtilization），以确定实例是空闲还是未充分利用。
  - 对于稳定的工作负载，请定期检查 AWS 合理调整大小工具（如 [AWS Compute Optimizer](#)），以确定优化和合理调整实例大小的机会。
    - [Well-Architected 实验室 - 合理调整大小建议](#)
    - [Well-Architected 实验室 - 使用 Compute Optimizer 合理调整大小](#)

- [Well-Architected 实验室 - 优化硬件模式并观察可持续性 KPI](#)

## 资源

### 相关文档：

- [优化您的 AWS 基础设施以实现可持续性，第 I 部分：计算](#)
- [AWS Graviton](#)
- [Amazon EC2 DL1](#)
- [Amazon EC2 容量预留实例集](#)
- [Amazon EC2 竞价型实例集](#)
- [函数：Lambda 函数配置](#)
- [基于属性为 Amazon EC2 实例集选择实例类型。](#)
- [在 AWS 上构建可持续、高效且优化成本的应用程序](#)
- [Contino 可持续发展控制面板如何助力客户减少碳排放](#)

### 相关视频：

- [深入了解 AWS Graviton2 处理器提供支持的 Amazon EC2 实例](#)
- [深入了解 AWS Graviton3 和 Amazon EC2 C7g 实例](#)
- [构建成本、能源和资源高效的计算环境](#)

### 相关示例：

- [解决方案：关于在 AWS 上优化深度学习工作负载以实现可持续性的指导](#)
- [Well-Architected 实验室 - 合理调整大小建议](#)
- [Well-Architected 实验室 - 使用 Compute Optimizer 合理调整大小](#)
- [Well-Architected 实验室 - 优化硬件模式并观察可持续性 KPI](#)
- [Well-Architected 实验室 - 将服务迁移到 Graviton](#)

## SUS05-BP03 使用托管服务

使用托管服务在云中更高效地运营。

常见反模式：

- 使用利用率低的 Amazon EC2 实例来运行应用程序。
- 内部团队仅管理工作负载，而没有时间专注于创新或简化。
- 为可在托管服务上更高效运行的任务部署和维护技术。

建立此最佳实践的好处：

- 使用托管服务将责任转移给 AWS，其拥有对数百万客户的洞察，可以帮助推动新的创新和提高效率。
- 由于使用了多租户控制面板，托管服务将服务的环境影响分散到许多用户。

在未建立这种最佳实践的情况下暴露的风险等级：中

实施指导

托管服务将维持已部署硬件的高利用率和可持续性优化的责任转移给 AWS。托管服务还消除了维护服务的运营和管理负担，让您的团队有更多时间专注于创新。

审核您的工作负载，以便确定可由 AWS 托管服务替换的组件。例如，[Amazon RDS](#)、[Amazon Redshift](#) 和 [Amazon ElastiCache](#) 提供托管数据库服务。[Amazon Athena](#)、[Amazon EMR](#) 和 [Amazon OpenSearch Service](#) 提供托管分析服务。

实施步骤

1. 清点工作负载的服务和组件。
2. 评测和确定可由托管服务替换的组件。以下是一些可以考虑采用托管服务的示例：

| Task     | What to use on AWS                                                                                                                                              |
|----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 托管数据库    | 使用托管 <a href="#">Amazon Relational Database Service ( Amazon RDS )</a> 实例而不是在 <a href="#">Amazon Elastic Compute Cloud ( Amazon EC2 )</a> 上维护自己的 Amazon RDS 实例。 |
| 托管容器工作负载 | 使用 <a href="#">AWS Fargate</a> ，而不是实施您自己的容器基础设施。                                                                                                                |

| Task      | What to use on AWS                                                       |
|-----------|--------------------------------------------------------------------------|
| 托管 Web 应用 | 使用 <a href="#">AWS Amplify 托管</a> 作为完全托管式 CI/CD，为静态网站和服务端呈现的 Web 应用托管服务。 |

3. 确定依赖项和创建迁移计划。相应地更新运行手册和行动手册。

- [AWS Application Discovery Service](#) 会自动收集并提供有关应用程序依赖项和利用率的详细信息，帮助您在制定迁移计划时做出更明智的决策

4. 迁移到托管服务之前测试服务。

5. 使用迁移计划，用托管服务来代替自托管服务。

6. 迁移完成后持续监控服务，以便根据需要进行调整并优化服务。

## 资源

相关文档：

- [AWS Cloud 产品](#)
- [AWS 总拥有成本 \( TCO \) 计算器](#)
- [Amazon DocumentDB](#)
- [Amazon Elastic Kubernetes Service \( EKS \)](#)
- [Amazon Managed Streaming for Apache Kafka \( Amazon MSK \)](#)

相关视频：

- [使用 AWS Managed Services 实现大规模云运营](#)

## SUS05-BP04 优化基于硬件的计算加速器的使用

优化加速型计算实例的使用，以减少工作负载的物理基础架构需求。

常见反模式：

- 不监控 GPU 使用情况。
- 将通用实例用于工作负载，而专用实例可以提供更高的性能、更低的成本和更高的性能功耗比。
- 使用基于硬件的计算加速器来完成任务，而使用基于 CPU 的替代方案能更高效地完成任务。

建立此最佳实践的好处：通过优化基于硬件的加速器的使用，您能够减少工作负载对物理基础设施的需求。

未建立这种最佳实践的情况下暴露的风险等级：中

## 实施指导

如果需要高处理能力，可以受益于使用加速型计算实例，这些实例提供对基于硬件的计算加速器的访问，例如图形处理单元（GPU）和现场可编程门阵列（FPGA）。这些硬件加速器能够比基于 CPU 的替代方案更有效地执行某些功能，例如图形处理或数据模式匹配。许多加速工作负载（如渲染、转码和机器学习）在资源使用方面变化很大。仅在需要时运行此硬件，并在不需要时自动停用它们，以最大限度地减少资源消耗。

## 实施步骤

- 确定哪些 [加速型计算实例](#) 可以满足您的要求。
- 对于机器学习工作负载，请利用特定于工作负载的专用硬件，例如 [AWS Trainium](#)、[AWS Inferentia](#) 和 [Amazon EC2 DL1](#)。Inf2 等 AWS Inferentia 实例相比 [同类 Amazon EC2 实例](#)，性能功耗比提升了 50%。
- 收集加速型计算实例的使用情况指标。例如，您可以使用 CloudWatch 代理，为 GPU 收集各种指标，例如 utilization\_gpu 和 utilization\_memory，如 [使用 Amazon CloudWatch 收集 NVIDIA GPU 指标](#) 中所示。
- 优化硬件加速器的代码、网络运营和设置，确保底层硬件得到充分利用。
  - [优化 GPU 设置](#)
  - [深度学习 AMI 中的 GPU 监控和优化](#)
  - [优化 I/O 以实现 Amazon SageMaker AI 中深度学习训练的 GPU 性能优化](#)
- 使用最新的高性能库和 GPU 驱动程序。
- 使用自动化功能在不使用 GPU 实例时将其释放。

## 资源

### 相关文档：

- [加速计算](#)
- [让我们来构建！使用自定义芯片和加速器来构建](#)
- [如何为我的工作负载选择合适的 Amazon EC2 实例类型？](#)
- [Amazon EC2 VT1 实例](#)

- [选择最佳 AI 加速器和模型编译，以使用 Amazon SageMaker AI 进行计算机视觉推理](#)

相关视频：

- [如何选择 Amazon EC2 GPU 实例进行深度学习](#)
- [部署经济高效的深度学习推理](#)

## 流程和文化

问题

- [SUS 6 您的组织流程如何支持您的可持续发展目标？](#)

### SUS 6 您的组织流程如何支持您的可持续发展目标？

寻找机会，通过对开发、测试和部署实践进行更改来降低可持续性影响。

最佳实践

- [SUS06-BP01 采用可以快速引入可持续性改进的方法](#)
- [SUS06-BP02 让您的工作负载保持最新状态](#)
- [SUS06-BP03 提高构建环境的利用率](#)
- [SUS06-BP04 使用托管式设备场进行测试](#)

#### SUS06-BP01 采用可以快速引入可持续性改进的方法

采用方法和流程来验证潜在的改进、最大限度降低测试成本和带来一些小改进。

常见反模式：

- 仅在项目开始时才完成一次审核应用程序可持续性。
- 工作负载变得过时，因为发布过程过于繁琐，无法为提高资源效率而引入微小的更改。
- 未制定相应的机制来提高工作负载的可持续性。

建立此最佳实践的好处：通过建立流程以引入和跟踪可持续性改进，您将能够不断采用新特性和功能、消除问题和提高工作负载效率。

在未建立这种最佳实践的情况下暴露的风险等级：中等

## 实施指导

在将潜在可持续性改进部署到生产环境之前，对其进行测试和验证。在计算改进的潜在未来收益时，考虑测试成本。开发低成本的测试方法，以实现细微的改进。

## 实施步骤

- 在您的开发待办事项中添加可持续性改进要求。
- 使用迭代[改进流程](#)来识别、评测、测试和部署这些改进并确定其优先顺序。
- 持续改进和简化您的开发流程。例如，[使用持续集成和持续交付（CI/CD）管道测试和部署潜在的改进，自动完成软件交付过程](#)，从而减少工作量和减少手动操作引起的错误。
- 使用最小可行代表性组件开发和测试潜在改进，从而降低测试成本。
- 持续评测改进的影响并根据需要作出调整。

## 资源

### 相关文档：

- [AWS 支持可持续性解决方案](#)
- [基于 AWS CodeCommit 的可扩展敏捷开发实践](#)

### 相关视频：

- [提供可持续、高性能的架构](#)

### 相关示例：

- [Well-Architected 实验室 - 将成本和使用情况报告转化为效率报告](#)

## SUS06-BP02 让您的工作负载保持最新状态

让您的工作负载保持最新状态，采用高效功能、消除问题和提高工作负载的整体效率。

### 常见反模式：

- 假设当前的架构是静态的，不会随着时间的推移而更新。
- 您没有任何系统（也不会定期）评估更新的软件和软件包是否与您的工作负载兼容。



建立此最佳实践的好处：通过建立一个及时更新工作负载的流程，您可以采用新的特性和功能，解决问题，并提高工作负载效率。

在未建立这种最佳实践的情况下暴露的风险等级：低

实施指导

最新的操作系统、运行时、中间件、库和应用程序可以提高工作负载效率，并简化更高效技术的采用。最新的软件可能还包括更准确地衡量工作负载对可持续性的影响的功能，因为供应商提供的功能是为了满足其自身的可持续性目标。定期更新，以便使用最新的功能和版本让您的工作负载保持最新。

实施步骤

- 定义一个流程和计划来评估工作负载的新功能或实例。利用云中的敏捷性，快速测试新功能如何改善工作负载以：
  - 减小对可持续性的影响。
  - 提升性能效率。
  - 为计划改进消除障碍。
  - 提高衡量和管理可持续性影响的能力。
- 盘点工作负载软件和架构，并确定需要更新的组件。
  - 您可以使用 [AWS Systems Manager 清单](#) 从 Amazon EC2 实例收集操作系统 ( OS )、应用程序和实例元数据，并快速了解哪些实例正在运行您的软件策略所需的软件和配置，以及哪些实例需要更新。
- 了解如何更新工作负载的组件。

| Workload component | How to update                                                                                                                                     |
|--------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| 机器映像               | 使用 <a href="#">EC2 Image Builder</a> 来管理适用于 Linux 或 Windows Server 映像的 <a href="#">亚马逊云机器镜像 ( AMI )</a> 更新。                                       |
| 容器映像               | 将 <a href="#">Amazon Elastic Container Registry (Amazon ECR)</a> 与现有管道配合使用以 <a href="#">管理 Amazon Elastic Container Service (Amazon ECS) 映像</a> 。 |
| AWS Lambda         | AWS Lambda 包括 <a href="#">版本管理功能</a> 。                                                                                                            |

- 采用自动化更新流程，以减少部署新功能的工作量，并减少手动过程引起的错误。
- 您可以使用 [CI/CD](#) 自动更新 AMI、容器映像以及与您的云应用程序相关的其他构件。
- 您可以使用 [AWS Systems Manager Patch Manager](#) 等工具自动执行系统更新流程，并使用 [AWS Systems Manager 维护窗口](#) 安排活动。

## 资源

### 相关文档：

- [AWS 架构中心](#)
- [AWS 新增功能](#)
- [AWS 开发人员工具](#)

### 相关示例：

- [Well-Architected 实验室 - 清单和补丁管理](#)
- [实验室：AWS Systems Manager](#)

## SUS06-BP03 提高构建环境的利用率

提高资源利用率，以开发、测试和构建工作负载。

### 常见反模式：

- 手动预置或终止构建环境。
- 使构建环境保持独立于测试、构建或发布活动运行（例如，在开发团队成员的工作时间之外运行环境）。
- 为构建环境过度预置资源。

建立此最佳实践的好处：通过提高构建环境的利用率，您可以提高云工作负载的整体效率，同时将资源分配给构建者，以便高效地进行开发、测试和构建。

在未建立这种最佳实践的情况下暴露的风险等级：低

## 实施指导

使用自动化和基础设施即代码功能，在需要时启动构建环境，并在不使用时将其关闭。一种常见模式是安排与开发团队成员的工作时间相吻合的可用时段。您的测试环境与生产配置非常相似。但是，寻找机

会使用具有突增容量的实例类型、Amazon EC2 竞价型实例、自动扩展数据库服务、容器和无服务器技术，以使开发和测试容量与使用容量保持一致。限制数据量，使之刚好满足测试要求。如果在测试中使用生产数据，请探索共享生产数据，而无需四处移动数据的可能性。

## 实施步骤

- 使用基础设施即代码来预置构建环境。
- 使用自动化功能来管理开发和测试环境的生命周期，并最大限度地提高构建资源的效率。
- 使用策略来最大程度地利用开发和测试环境。
  - 使用最小可行代表性环境来开发和测试潜在的改进。
  - 如果可能，请使用无服务器技术。
  - 使用按需型实例来补充您的开发人员设备。
  - 使用具有容量暴增的实例类型、竞价型实例和其他技术，使构建容量与使用容量保持一致。
  - 采用原生云服务来实现安全的实例 Shell 访问，而不是部署堡垒主机群。
  - 根据构建作业自动扩展构建资源。

## 资源

### 相关文档：

- [AWS Systems Manager Session Manager](#)
- [Amazon EC2 可突增性能实例](#)
- [什么是 AWS CloudFormation？](#)
- [什么是 AWS CodeBuild？](#)
- [AWS 上的实例计划程序](#)

### 相关视频：

- [持续集成最佳实践](#)

## SUS06-BP04 使用托管式设备场进行测试

使用托管式设备场在一组具有代表性的硬件上高效地测试新功能。

### 常见反模式：

- 在各个物理设备上手动测试和部署应用程序。
- 未在真实的物理设备上使用应用测试服务进行测试以及与应用（例如，Android、iOS 和 Web 应用）互动。

建立此最佳实践的好处：使用托管式设备场来测试具有云功能的应用程序，这提供了许多好处：

- 包括可在各种设备上测试应用程序的更高效功能。
- 无需使用内部基础设施进行测试。
- 提供多种设备类型（包括不太常用的较旧硬件），从而不需要进行不必要的设备升级。

在未建立这种最佳实践的情况下暴露的风险等级：低

### 实施指导

使用托管式设备场有助于简化在一组有代表性的硬件上测试新功能的过程。托管式设备场提供多种设备类型，包括不太常用的较旧硬件，并避免不必要的设备升级对客户可持续性的影响。

### 实施步骤

- 定义您的测试要求和计划（例如，测试类型、操作系统和测试时间表）。
  - 您可以使用 [Amazon CloudWatch RUM](#) 收集和分析客户端数据并制定测试计划。
- 选择可支持您的测试要求的托管式设备场。例如，您可以使用 [AWS Device Farm](#) 来测试和了解您的更改对一组具有代表性的硬件的影响。
- 使用持续集成/持续部署（CI/CD）来安排和运行测试。
  - [将 AWS Device Farm 与 CI/CD 管道集成，以便运行跨浏览器的 Selenium 测试](#)
  - [使用 AWS DevOps 和移动服务构建和测试 iOS 和 iPadOS 应用](#)
- 持续审核测试结果，必要时进行改进。

### 资源

相关文档：

- [AWS Device Farm 设备列表](#)
- [查看 CloudWatch RUM 控制面板](#)

相关示例：

- [适用于 Android 的 AWS Device Farm 示例应用](#)
- [适用于 iOS 的 AWS Device Farm 示例应用](#)
- [适用于 AWS Device Farm 的 Appium Web 测试](#)

相关视频：

- [使用 Amazon CloudWatch RUM 通过最终用户洞察优化应用程序](#)

# 声明

客户负责对本文档中的信息进行独立评估判断。本文档：( a ) 仅供参考，( b ) 代表 AWS 当前的产品和服务和实践，如有变更，恕不另行通知，以及 ( c ) 不构成 AWS 及其附属公司、供应商或授权商的任何承诺或保证。AWS 产品或服务均“按原样”提供，没有任何明示或暗示的担保、声明或条件。AWS 对其客户的责任和义务由 AWS 协议规定，本文档与 AWS 和客户之间签订的任何协议无关，亦不影响任何此类协议。

版权所有 © 2021 , Amazon Web Services, Inc. 或其附属公司。