



生成式 AI 工作量评估

AWS 规范性指导



AWS 规范性指导: 生成式 AI 工作量评估

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

Amazon 的商标和商业外观不得用于任何非 Amazon 的商品或服务，也不得以任何可能引起客户混淆、贬低或诋毁 Amazon 的方式使用。所有非 Amazon 拥有的其他商标均为各自所有者的财产，这些所有者可能附属于 Amazon、与 Amazon 有关联或由 Amazon 赞助，也可能不是如此。

Table of Contents

简介	1
本指南的目的	1
目标受众和福利	2
范围	2
目标业务成果	4
评估注意事项和先决条件	6
从清晰的用例开始	6
确保业务协调一致	7
实施治理和监督	7
满足数据和技术先决条件	7
考虑计算资源需求	7
解决隐私和安全隐患	7
尽早让利益相关者参与	7
迭代和学习	8
生成式 AI 工作量评估问卷	9
准备就绪	9
使用案例	11
架构	13
存储	14
法规与合规	14
集成	15
测试	17
部署和自动化	18
数据策略	19
将评估见解转化为切实可行的结果	22
后续步骤	24
常见问题解答	25
主要目标是什么？	25
谁应该使用这个评估？	25
关键组件有哪些？	25
这如何帮助定义架构？	25
有什么好处？	25
我们怎样才能成功实现呢？	25
面临哪些挑战？	26

监管和合规要求是什么？	26
利益相关者的作用是什么？	26
我们如何衡量成功？	26
根据组织规模，这种方法有何不同？	26
资源	28
文档历史记录	29
术语表	30
#	30
A	30
B	33
C	34
D	37
E	40
F	42
G	43
H	44
我	45
L	47
M	48
O	52
P	54
Q	56
R	57
S	59
T	62
U	63
V	64
W	64
Z	65
	lxvi

生成式 AI 工作量评估

Tabby Ward 和 Deepak Dixit , Amazon Web Services (AWS)

2024 年 11 月 ([文件历史记录](#))

生成式 AI 工作负载评估是一种战略方法，旨在评估和改善组织创建或更新其生成式 AI 工作负载的准备情况。这种评估很重要，因为将生成式人工智能纳入业务运营可以极大地改变事物的运作方式，并可以提供新的效率和能力。但是，要成功采用生成式人工智能，就必须彻底了解当前的系统并为未来制定明确的计划。

生成式 AI 工作负载是指涉及使用人工智能模型的计算任务，这些模型可以创建新内容，例如文本、图像、代码或其他数据类型。这些工作负载通常需要大量的计算能力、专门的硬件（例如）以及 GPUs 用于训练和推理的大型数据集。将生成式 AI 工作负载集成到运营中会带来一些挑战：

- 基础设施要求：配置生成式 AI 模型所需的大量计算资源和专用硬件。
- 数据管理：在处理大型数据集的同时确保数据质量、隐私和合规性。
- 技能差距：缺乏人工智能技术和模型部署方面的专业知识。
- 道德考量：解决人工智能生成内容中的偏见、公平和透明问题。
- 集成复杂性：将生成式 AI 无缝整合到现有工作流程和传统系统中。
- 成本管理：在潜在收益与高昂的实施和运营成本之间取得平衡。

克服这些挑战需要精心规划、基础设施和人才投资以及战略实施方法。

本指南的目的

生成式人工智能正迅速成为许多行业的关键组成部分。它提供了变革性机会，但也带来了集成、合规性和可扩展性方面的挑战。由于技术基础薄弱、变革阻力和数据质量问题，许多组织难以充分利用人工智能。生成式人工智能工作负载评估通过确定现代化的要求、定义实施范围以及挑战传统系统和思维来应对这些挑战。它还有助于确定最低可行产品 (MVPs)，并帮助您开发目标解决方案架构，从而确保采用结构化和战略性的方法来采用 AI。

本指南是一种结构化方法，可帮助组织应对采用生成式人工智能技术的复杂性。该指南没有从一开始就明确界定要求，而是提供以下帮助：

- 确定组织内生成式 AI 的潜在用例。

- 评估您的组织是否准备好采用生成式 AI。
- 定义和完善用例目标和延伸目标。
- 确定生成式 AI 实现的范围和要求。
- 开发目标解决方案架构。

目标受众和福利

该评估专为想要评估生成式 AI 工作负载现代化技术方面的解决方案架构师、企业架构师和应用程序架构师而设计。对于想要衡量团队总体准备情况、资源分配和能力需求的项目和人事经理来说，这也很有价值。行业最佳实践强调了全面评估的重要性，以确保为采用人工智能做好准备。这包括评估架构、存储、合规性、集成、测试、部署和自动化。

范围

以下主题属于生成式 AI 工作负载评估方法的范围：

- 当前的生成式 AI 技术和模型（例如，大型语言模型、图像生成模型）
- 使用生成技术的狭窄人工智能应用程序
- 将生成式 AI 与现有系统和工作流程集成
- 用于训练和微调生成式 AI 模型的数据策略
- 当前生成式人工智能应用的道德考虑因素和负责任的人工智能实践
- 生产环境中生成式 AI 的测试和部署策略
- 生成式 AI 实现的安全和隐私注意事项
- 生成式 AI 工作负载的性能优化和可扩展性
- 生成式人工智能在各行各业的用例和应用
- 评估生成式 AI 输出和质量保证流程

以下主题不在范围内：

- 人工通用智能 (AGI) 和人工超级智能 (ASI) 场景
- 除了当前的生成模型之外，人工智能的未来可预测性进步
- 量子计算在人工智能中的应用
- 神经形态计算和脑机接口

- 人工智能系统中的意识和自我意识
- 超越当前生成式人工智能应用的高级人工智能对社会的长期影响
- 假设的未来 AI 技术的监管框架
- 关于机器中智力和意识本质的哲学辩论
- 人工智能的极端边缘案例或高度投机性的用例
- 专有 AI 模型或架构的详细技术规格

目标业务成果

生成式 AI 工作负载评估旨在提供多个有针对性的结果，这些结果对于成功实现生成式 AI 工作负载的现代化至关重要。这些结果确保了组织为有效和高效地整合人工智能技术做好了充分的准备。

对于每项目标结果，生成式 AI 工作量评估侧重于：

- **相互依存关系**：确定并阐明现代化过程的结果与其他方面之间的任何相互依存关系。这包括了解一种结果可能如何影响或受其他结果的影响，以确保采用整体方法实现现代化。
- **利益相关者协调**：概述战略，使不同的利益相关者与每项结果保持一致。这包括向不同的组织级别和部门传达每项结果的价值和影响，以促进认同和支持。
- **优先级**：在确定了多个用例或结果的情况下，请提供一个框架，以便根据业务影响、资源需求和战略调整等因素对它们进行优先排序。
- **持续改进**：针对每项结果，建立持续评估和完善的机制。这确保了现代化工作能够适应和响应不断变化的技术格局和业务需求。

以下是对每项目标结果的详细讨论：

目标架构

- **定义**：该评估有助于为生成式 AI 工作负载定义清晰且可扩展的目标架构。
- **组件**：这包括选择合适的云服务、设计数据管道和确保系统互操作性。
- **优点**：定义明确的架构支持可扩展性、可靠性和性能优化，并为现代化提供了坚实的基础。

客户准备就绪

- **评估**：评估组织基础架构、流程和文化的现状，以确定采用生成式人工智能现代化的准备情况。
- **标准**：这包括评估技术能力、数据质量和组织接受变革的意愿。
- **成果**：找出差距和需要改进的领域，确保组织为顺利过渡到现代解决方案和技术做好准备。

用例目标和延伸目标

- **用例目标**为目标解决方案的实施设定了明确的目标，重点是特定的业务问题或机会。

生成式人工智能现代化背景下的用例目标是指组织旨在通过实施生成式人工智能解决方案来实现的具体、可衡量的目标。这些目标通常与更广泛的业务目标保持一致，并侧重于应对组织内部的特定挑战或机遇。用例目标的示例可能包括：

- 通过使用生成式 AI 驱动的聊天机器人，将客户服务响应时间缩短 50%。
- 通过生成式 AI 辅助代码分析，将代码审查效率提高 30%。
- 通过使用生成式 AI 模式识别，将欺诈检测准确性提高 25%。
- 延伸目标定义了雄心勃勃的目标，这些目标突破了生成式人工智能现代化在组织内可以实现的目标。
- 影响：设定可实现的目标和理想的目标有助于使生成式人工智能现代化计划与战略业务目标保持一致，并鼓励创新。

工作量估算

- 目的：准确的工作量估算有助于资源规划，并确保项目在预算范围内按时交付。
- 范围：估算实施生成式人工智能现代化计划所需的资源、时间和预算。
- 因素：考虑技术复杂性、集成挑战和潜在风险。

支持需求

- 培训和发展：确定成功采用生成式人工智能现代化所需的技能和知识。
- 资源：确定是否需要培训计划、研讨会和其他支持活动。
- 成果：确保员工具备必要的技能，可以提高生成式人工智能现代化计划的有效性，并支持长期成功。

实施计划

- 路线图：制定详细计划，概述实现生成式 AI 现代化所需的步骤。
- 里程碑：定义关键里程碑和可交付成果以跟踪进度。
- 好处：明确的实施计划提供了方向和问责制，并促进了生成式人工智能现代化的结构化方法。

评估注意事项和先决条件

从清晰的用例开始

确定生成式 AI 可以解决的具体业务问题或机会。重点关注与战略业务目标一致并提供可衡量收益的用例。优先考虑针对组织内常见挑战的用例，以确保解决方案架构可以作为多种场景的模式。

在对潜在的生成式人工智能应用有大致了解的情况下启动评估过程是有益的，但不是强制性的。本指南中包含的[问卷](#)可以适应不同的准备水平，从具有明确用例定义的组织到只有广泛想法的组织。评估过程有助于：

- 完善并阐明这些最初的用户想法。
- 确定新的潜在用例。
- 为每个用例制定具体、可衡量的目标。
- 评估每个用例的可行性和潜在影响。

让我们考虑一个假设的例子：一家金融服务公司决定探索生成式人工智能现代化。他们从改善客户服务和欺诈检测流程的宏伟想法开始。

- 初步评估：问卷可帮助他们评估其当前的系统、数据质量以及采用生成式人工智能的组织准备情况。
- 用例完善：通过评估流程，他们将最初的想法完善为两个具体的用例：
 - 为客户查询实现人工智能驱动的生成式聊天机器人
 - 使用生成式 AI 进行实时交易欺诈检测
- 目标设定：对于每个用例，他们都定义了具体的目标：
 - 在 6 个月内将客户服务响应时间缩短 40%
 - 欺诈检测准确率提高 20%，误报率减少 15%
- 延伸目标：他们还设定了以下雄心勃勃的目标：
 - 通过 AI 辅助响应实现 80% 的客户满意度
 - 开发可识别新欺诈模式的预测性欺诈检测模型
- MVP 定义：问卷可帮助他们确定每个用例的 MVP，重点关注能够带来即时价值的基本功能。
- 目标架构：最后，他们开发了一种目标架构，该架构支持一个或两个用例，并确保可扩展性以及有系统的集成。

确保业务协调一致

使生成式 AI 计划与整体业务战略和目标保持一致。对于每个用例，都要制定明确的价值主张，展示生成式人工智能如何为业务增长、效率或创新做出贡献。建立衡量生成式 AI 实施对关键绩效指标影响的指标 (KPIs)。

实施治理和监督

成立一个跨职能指导委员会来监督生成式人工智能计划。制定负责任地使用人工智能的政策和指导方针，解决道德考虑因素和潜在偏见。为生成式 AI 项目建立审查流程，以确保符合组织标准和监管要求。

满足数据和技术先决条件

评估和改善数据质量，并实施数据治理实践，以确保生成式 AI 模型的可靠输入。制定数据策略，以满足特定于生成人工智能需求的数据收集、存储和管理。评估和增强数据基础架构，以支持生成式 AI 工作负载所需的数据量和速度。

考虑计算资源需求

评估当前的 IT 基础架构，找出生成式 AI 工作负载的计算能力差距。考虑云服务或本地高性能计算集群等选项，规划可扩展的计算资源。优化资源分配，平衡训练和推理工作负载的性能和成本效益。

解决隐私和安全隐患

实施强有力的安全措施，保护生成式 AI 训练和操作中使用的敏感数据。在处理个人信息时，确保遵守数据保护法规，例如《通用数据保护条例》(GDPR) 或《加州消费者隐私法》(CCPA)。开发用于安全模型部署和监控的协议，以防止未经授权的访问或滥用生成式 AI 功能。

尽早让利益相关者参与

从一开始就让关键利益相关者参与进来，以获得领导层的支持和支持。明确传达现代化计划的好处和潜在影响，特别是针对生成式 AI 工作负载。提供培训和资源，帮助利益相关者了解生成式人工智能技术及其影响。

迭代和学习

采用渐进式方法，让您可以完善目标解决方案。使用反馈回路持续改进工作负载架构和流程。定期评估生成式 AI 实施的性能和影响，并根据实际结果和不断变化的业务需求根据需要调整策略。

生成式 AI 工作量评估问卷

以下各节提供了一些问题，您可以使用这些问题来评估组织生成式 AI 工作负载现代化的不同方面。这份全面的问卷调查评估贵组织采用和实施生成式 AI 工作负载的准备情况，并针对关键领域提出问题，包括用例、架构、存储、合规性、集成、测试、部署和数据策略。通过探讨生成式人工智能实施的关键方面，从技术基础设施到监管方面的考虑，本问卷可帮助您确定人工智能现代化之旅中的优势、差距和机遇。

部分：

- [准备就绪](#)
- [使用案例](#)
- [架构](#)
- [存储](#)
- [法规与合规](#)
- [集成](#)
- [测试](#)
- [部署和自动化](#)
- [数据策略](#)

你也可以下载 Microsoft Excel 格式的问卷并用它来记录你的信息。



[下载问卷](#)

准备就绪

问题	响应示例
您是否有可以用来处理这些工作负载的 AWS 账户？	是还是不是。
您与之签订了现有的企业协议 AWS 吗？	是还是不是。

问题	响应示例
您当前的云基础架构在处理生成式 AI 工作负载方面的可扩展性如何？	我们的云基础设施具有高度的可扩展性，具有计算资源和分布式存储系统的自动扩展功能，旨在高效处理大规模的生成式 AI 工作负载。
您是否具有用于大规模预处理和特征工程的数据管道功能？	我们的数据管道使用分布式处理框架（例如 Apache Spark）进行大规模数据预处理和功能工程，同时支持批处理和流式数据处理。
你有账户配置和管理能力吗？	是还是不是。
你会如何描述你的组织的人工智能素养和采用生成式人工智能技术的意愿？	我们的组织在人工智能教育项目上投入了大量资金，大多数技术人员已经完成了基本的人工智能/机器学习培训。该组织的创新文化包含新技术，包括生成式人工智能。
您的组织中存在哪些 AI/ML 专业知识，这些专业知识是如何分布的？	我们有一个专门的 AI 卓越中心，由经验丰富的数据科学家和机器学习工程师组成。我们提高不同业务部门的领域专家的技能，以提高人工智能素养并识别生成式人工智能用例。
您是否有阐明云计划目标、优势和成本的高级业务案例？	是还是不是。
你将解决方案投入生产的时间表是什么？	几周、几个月等等。
您的主要利益相关者（例如首席财务官、CIT/CTO、首席运营官）是否已做出资金承诺？	是还是不是。
在生成式人工智能计划中，如何确保遵守数据保护法规？	我们有一个专门的合规团队，与我们的人工智能团队密切合作。我们定期进行隐私影响评估，按照设计原则实施数据保护，并维护所有生成式人工智能项目的详细数据处理记录。
与新的生成式人工智能技术集成的现有系统有多成熟？	我们的 IT 架构基于微服务 APIs，允许灵活集成新的生成式 AI 技术。这些系统采用通用数据格式和协议进行标准化，以确保互操作性。

问题	响应示例
你在操作机器学习模型方面有什么经验，这会如何应用于生成式人工智能系统？	我们已经建立了 MLOps 实践，包括自动模型部署管道、监控系统和 A/B 测试框架。这些实践正在进行调整，以满足大规模生成式人工智能模型的独特要求。

使用案例

问题	响应示例
该用例的主要目标或成功标准是什么？	缩短客户支持响应时间，提高销售转化率，增强产品推荐。另外：提高用户满意度、任务完成率、响应质量等。
此用例如何与贵组织的战略目标保持一致？	这符合我们的战略目标，即通过缩短客户服务响应时间来提高客户满意度。
该用例的预期数据量或请求量是多少？	每秒 500 笔交易 (TPS)。
需要哪些类型的数据源来支持您的生成式 AI 工作负载？	内部结构化数据库（客户记录、销售数据等）；来自文档、电子邮件和社交媒体的非结构化文本数据；用于语音和图像识别任务的音频和视频文件；来自物联网设备和传感器的实时流数据；公共数据集和 APIs 用于丰富的数据集。
您需要多久更新或刷新一次来自这些来源的数据？	交易数据库：近乎实时更新；文档存储库：每日批量更新；社交媒体提要：每小时更新；物联网传感器数据：持续实时流；公共数据集：每月或每季度更新。
您的生成式 AI 模型需要哪些数据格式作为输入？	结构化数据：CSV、JSON 和 SQL 数据库表；文本数据：纯文本、PDF 和 HTML；图像数据：JPEG、PNG 和 TIFF；音频数据：WAV 和 MP3；视频数据：MP4 以及 AVI。

问题	响应示例
对于生成式 AI 工作负载，您最关心的数据质量问题是什么？	完整性：确保不遗漏任何关键字段；准确性：验证数据的正确性并消除错误；一致性：在不同来源之间保持统一的格式和值；及时性：确保数据是最新的，以便进行实时推断；相关性：确认数据与特定的生成人工智能任务一致。
关键性能要求是什么（例如，响应时间、吞吐量、准确性）？	准确率为 95%；响应时间小于 500 毫秒；能够每秒处理 1000 个请求。精度高（95% 以上）、中等精度（80-90%）、尽力而为，等等。
你还有其他 KPIs 方法可以衡量这个用例的成功吗？	关键 KPIs 包括错误率降低、每笔交易节省的时间以及客户满意度得分。
需要多少模型精度，它如何与成本保持平衡？	精度高 (> 90%)，成本适中，精度适中 (70-80%)，成本低，等等。
生成式 AI 解决方案的主要用例或场景是什么？	客户服务聊天机器人、内容生成、产品推荐等。
生成式人工智能系统的目标用户或角色是什么？	客户服务代理、营销团队、员工、最终用户等。
预计的请求量或用户量是多少？	每天 1,000 个请求；每月活跃用户 10,000 个。
是否有任何特定的用例限制或要求？	实时响应、多语言支持、数据隐私等。
您是否有用于开发和维护生成式人工智能解决方案的预算？	初始开发成本估计为 20 万美元，每年的维护成本为 50,000 美元。
此用例的预计投资回报率 (ROI) 和投资回报期是多少？	预计三年内投资回报率为 150%，投资回收期为 18 个月。
是否应该考虑任何隐性成本或潜在的节约？	潜在的节省包括减少加班成本。隐性成本可能涉及对工作人员的额外培训。
这种生成式人工智能解决方案的可扩展性和未来扩展可能性有哪些？	该解决方案旨在随着我们的运营而扩展，将来有可能扩展到其他部门。
如何确保生成人工智能模型中的公平性并减少偏见？	我们计划通过多样化的数据收集、定期的偏见审计和偏见缓解技术的实施来缓解偏见。

问题	响应示例
你们制定了哪些流程来解决道德问题或意外后果？	我们将通过既定的人工智能事件响应计划、定期的道德风险评估、员工的匿名报告系统、与外部伦理专家的合作以及根据反馈持续监控和调整已部署的模型来管理道德问题。
您如何对组织中不同项目和部门的生成式 AI 工作量评估进行优先级排序和排序？	通过对所有部门进行高级别调查，以确定潜在的生成式人工智能用例，并根据三个关键标准对其进行评估：业务影响、技术可行性和道德考虑。优先考虑潜在影响力大、技术壁垒较低且道德问题最少的项目。

架构

问题	响应示例
正在考虑哪种类型的生成式 AI 模型或架构？	变压器、卷积神经网络 (CNN)、循环神经网络 (RNN)、决策树等。
预期的数据和计算规模或数量是多少？	数百万用户、PB 级数据等。
训练和推理的硬件要求（例如 CPUs 或 GPUs）是什么？	高端 GPUs、CPU 集群、云实例等。
随着时间的推移，生成式 AI 模型将如何更新或重新训练？	通过持续学习、定期再培训、手动更新等。
数据预处理和特征工程要求是什么？	文本清理、图像增强、功能选择等。
生成式 AI 系统将如何处理边缘情况、异常值或低可信度输入？	通过回退到人工监督，请求澄清等。
生成式 AI 应用程序的延迟要求是什么？	实时、近实时、批处理等。

存储

问题	响应示例
训练数据将存储在哪里？	在云存储（例如 Amazon S3、文件存储、块存储或对象存储）、本地存储等中。
训练数据和模型工件（例如容量、耐久性、可用性）的存储要求是什么？	PB 级存储、高耐久性（99.999999999% 的耐久性）、高可用性等。
训练数据和模型工件的数据保留和备份要求是什么？	数据保留 x 年、每日备份、异地备份等。
哪些文件格式主要用于存储 AI 训练数据集（例如 CSV、JSON、Parquet 等 HDF5）？	Parquet 文件 HDF5 用于结构化数据、大型多维数组和非结构化数据，例如图像和文本。我们使用专门的格式，TFRecord 例如在训练期间优化数据加载。
您的训练数据集是如何组织的：作为单个文件、数据库或使用专门的 AI 数据格式？	为了灵活起见，中小型数据集作为单独的 Parquet 文件存储在对象存储中。大型数据集存储在分布式数据库 (Cassandra) 中以应对规模。
您是否使用任何专门用于生成式 AI 训练数据的数据压缩或编码技术？	对于表格数据，我们使用 Parquet 中提供的字典编码和位打包技术。对于图像，我们使用有损 JPEG 压缩，其质量设置针对我们的模型进行了优化。
您如何处理训练数据集不同迭代的版本控制和存储？这对您的整体存储需求有什么影响？	我们使用与机器学习平台集成的数据版本控制系统 (DVC)。

法规与合规

问题	响应示例
生成式人工智能解决方案（例如 GDPR、HIPA A、PCI-DSS）的相关法规或合规要求是什么？	GDPR 用于处理个人数据，HIPAA 用于医疗保健数据，PCI-DSS 用于支付数据，等等。

问题	响应示例
您的组织采用了哪些合乎道德的生成式人工智能指导方针或框架？	我们实施了自己的负责任的人工智能指导方针。所有生成式人工智能项目在批准和部署之前都要经过伦理审查。
生成式 AI 系统的安全要求是什么？	数据加密、安全网络通信、定期安全审计。
数据隐私和保护的要求是什么？	数据匿名化、加密、访问控制等。
处理敏感或机密数据的解决方案有哪些要求？	严格的访问控制、数据屏蔽、数据驻留要求等。
如何处理用户身份验证和授权？	通过使用 OAuth API 密钥、单点登录 (SSO) 和基于角色的访问控制 (RBAC)。
在生产环境中将如何监控和管理解决方案？	通过使用 Prometheus 和 Datadog 等监控工具、ELK Stack 等日志工具、警报系统等。

集成

问题	响应示例
将生成式 AI 解决方案与现有系统或数据源集成的要求是什么？	REST APIs、消息队列、数据库连接器等。
生成式人工智能解决方案将如何摄取和预处理数据？	通过使用批处理、流数据、数据转换和特征工程。
生成式人工智能解决方案的输出将如何被消耗或与下游系统集成？	通过 API 端点、消息队列、数据库更新等。
哪些事件驱动的集成模式可用于生成式 AI 解决方案？	消息队列（例如 Amazon SQS、Apache Kafka、RabbitMQ）、发布/订阅系统、网络挂钩、事件流平台。
哪些基于 API 的集成方法可用于将生成式 AI 解决方案与其他系统连接起来？	RESTful APIs、GraphQL APIs、SOAP APIs（适用于遗留系统）。

问题	响应示例
哪些微服务架构组件可用于生成式 AI 解决方案集成？	用于服务间通信的服务网格、API 网关、容器编排（例如 Kubernetes）。
如何为生成式 AI 解决方案实现混合集成？	通过结合事件驱动的模式进行实时更新、历史数据的批处理和外部系统 APIs 集成。
如何将生成式 AI 解决方案输出与下游系统集成？	通过 API 端点、消息队列、数据库更新、Webhook 和文件导出。
要集成生成式人工智能解决方案，应考虑采取哪些安全措施？	身份验证机制（例如 OAuth 或 JWT）、加密（传输中和静态加密）、API 速率限制和访问控制列表 (ACLs)。
您计划如何将开源框架（例如 LlamaIndex 或整合）集成 LangChain 到现有的数据管道和生成式人工智能工作流程中？	我们计划使用它 LangChain 来构建复杂的生成式 AI 应用程序，特别是其代理和内存管理功能。我们的目标是在未来 6 个月 LangChain 内让 60% 的生成式人工智能项目投入使用。
您将如何确保所选的开源框架与现有数据基础设施之间的兼容性？	我们正在组建一个专门的集成团队，以确保流畅的兼容性。到第三季度，我们的目标是建立一个完全集成的管道，LlamaIndex 用于在我们当前的数据湖结构中进行高效的数据索引和检索。
您计划如何利用框架的模块化组件，例如 LangChain 用于快速原型设计和实验？	我们正在建立一个沙盒环境，开发人员可以在其中使用 LangChain 组件快速制作原型。
在这些快速发展的开源框架中，你有什么策略可以跟上更新和新功能的步伐？	我们已经指派了一个团队来监控和的 GitHub 仓库 LangChain 和社区论坛 LlamaIndex。我们计划每季度评估和整合主要更新，重点是性能改进和新功能。

测试

问题	响应示例
测试要求是什么（例如，单元测试、集成测试、end-to-end测试）？	单个组件的单元测试、与外部系统的集成 end-to-end测试、关键场景的测试等。
如何确保生成式 AI 训练的不同来源的数据质量和一致性？	我们通过自动数据分析工具、定期数据审计和集中式数据目录来维护数据质量。我们实施了数据治理政策，以确保不同来源的一致性并维护数据沿袭。
如何评估和验证生成式人工智能模型？	通过使用抵制数据集、人工评估、A/B 测试等。
评估生成式人工智能模型的性能和准确性的标准是什么？	精度、召回率、F1 分数、困惑、人为评估等。
如何识别和处理边缘案例和极端情况？	通过使用全面的测试套件、人工评估、对抗测试等。
您将如何测试生成式人工智能模型中的潜在偏差？	通过使用人口均等分析、机会均等测试、对抗性消除偏见技术、反事实测试等。
哪些指标将用于衡量模型输出的公平性？	不同的影响比率、均衡的赔率、人口平等、个人公平性指标等。
您将如何确保偏见检测的测试数据集具有多样化的表现形式？	通过使用跨人口群体的分层抽样、与多元化专家的合作、使用合成数据来填补空白等。
部署后将实施哪个流程来持续监测模型的公平性？	定期进行公平性审计、自动偏见检测系统、用户反馈分析、使用更新的数据集进行定期再培训等。
你将如何解决生成式人工智能模型中的交叉偏差？	通过使用交叉公平性分析、子组测试、与领域专家合作研究交叉性等。
你将如何测试模型在不同语言和文化背景下的表现？	通过使用多语言测试集、与文化专家合作、本地化的公平性指标、跨文化比较研究等。

部署和自动化

问题	响应示例
扩展和负载平衡的要求是什么？	智能请求路由；自动扩展系统；通过采用模型缓存、延迟加载和分布式存储系统等技术来优化快速冷启动；设计系统以处理突发的、不可预测的流量模式。
更新和推出新版本有什么要求？	蓝/绿部署、金丝雀版本、滚动更新等。
灾难恢复和业务连续性的要求是什么？	Backup 和恢复程序、故障转移机制、高可用性配置等。
自动化生成式 AI 模型的训练、部署和管理有哪些要求？	自动训练管道、持续部署、自动扩展等。
随着新数据的出现，生成式人工智能模型将如何更新和重新训练？	通过定期再训练、增量学习、迁移学习等。
自动化监控和管理的要求是什么？	自动警报、自动扩展、自我修复等。
对于生成式 AI 工作负载，您首选的部署环境是什么？	一种混合方法，它使用 AWS 进行模型训练，使用我们的本地基础设施进行推理，以满足数据驻留要求。
对于生成式 AI 部署，你更喜欢哪些特定的云平台吗？	AWS 服务，尤其是用于模型开发和部署的 SageMaker Amazon AI，以及用于基础模型的 Amazon Bedrock。
对于生成式 AI 工作负载，您正在考虑使用哪些容器化技术？	我们希望在使用 Kubernetes 编排的 Docker 容器上实现标准化，以确保混合环境中的可移植性和可扩展性。
在你的生成人工智能管道中，你有首选的 CI/CD 工具吗？	GitLab 用于版本控制和 CI/CD 管道，与 Jenkins 集成，用于自动测试和部署。
您正在考虑使用哪些编排工具来管理生成式 AI 工作流程？	Apache Airflow 用于工作流程编排，特别是用于数据预处理和模型训练管道。

问题	响应示例
您对支持生成式 AI 工作负载的本地基础设施有什么具体要求吗？	我们正在投资 GPU 加速的服务器和高速网络，以支持本地推理工作负载。
您计划如何管理不同环境中的模型版本控制和部署？	我们计划使用 MLflow 模型跟踪和版本控制，并将其与我们的 Kubernetes 基础架构集成，以便跨环境进行无缝部署。
您正在考虑使用哪些监控和可观察性工具进行生成式 AI 部署？	Prometheus 用于指标收集，Grafana 用于可视化，还有用于特定模型监控的其他自定义日志解决方案。
您如何解决混合部署模式中的数据移动和同步问题？	我们将使用 AWS DataSync 本地存储和根据训练周期安排的自动同步作业 AWS，在本地存储之间进行高效的数据传输。
您正在为跨不同环境的生成式 AI 部署实施哪些安全措施？	我们将将 IAM 用于云资源，并将其与本地 Active Directory 集成，以实现 end-to-end 加密和网络分段，以保护数据流。

数据策略

问题	响应示例
哪些特定的数据类型对你的生成式 AI 工作负载至关重要，其中有多少百分比目前可以访问？	客户通话记录和产品评论数据至关重要。目前，这些数据类型中有85%可供我们的生成式人工智能项目访问。
您如何确保和衡量数据的质量？	我们已经实施了数据质量指标，包括完整性、准确性、一致性和及时性。我们使用自动化工具定期评估这些指标，并有一个专门的团队来清理和充实数据。
您的数据中有多少百分比符合生成式人工智能使用的质量标准？	目前，我们 78% 的数据符合我们的质量标准。我们的目标是通过改进数据清理流程，在未来 12 个月内实现 95% 的目标。

问题	响应示例
您计划如何建立利益相关者对生成式人工智能中数据使用情况的信任？	我们正在组建人工智能伦理委员会，为人工智能决策提供明确的解释，并每季度进行人工智能审计，以确保透明度和公平性。
您的数据源和世系文档有多全面？	我们维护一个详细的数据目录，其中包括所有数据源的元数据，包括来源、更新频率和使用情况。我们使用数据沿袭工具来跟踪数据在系统中的流动和转换情况。
如何确保数据集的多样性以防止 AI 模型出现偏差？	我们积极从不同的人群中获取数据，并定期审计我们的数据集是否存在代表性偏差。我们还使用合成数据生成技术来平衡代表性不足的类别。
你的关键生成式人工智能模型的数据刷新率是多少？如何确定这个频率？	关键模型每周更新一次。此频率由 A/B 测试性能指标决定，我们的目标是两次刷新之间的降级不超过 2%。
您维护了多少个版本的关键数据集，持续了多长时间？	我们维护每个关键数据集的最后五个版本，每个版本的保留期为 18 个月。
有多少跨职能团队参与了你的生成式人工智能计划并可以访问你的数据？	我们三个跨职能团队。每个团队都包括数据科学家、领域专家、伦理学家和业务分析师。
您制定了哪些数据治理政策和实践？	我们有一个跨职能的数据治理委员会，负责监督我们的数据政策。我们实施了基于角色的访问控制、数据分类方案和定期审计，以确保遵守我们的治理框架。
您采取了哪些措施来确保数据隐私、获得适当同意和保持机密性？	我们已经实施了符合 GDPR 和 CCPA 的全面数据隐私框架。这包括获得对数据使用的明确同意、实施数据匿名化技术以及定期进行隐私影响评估。
在上个季度，您的 AI 训练数据集中有多少百分比经过了偏见审计？	上个季度，我们对 70% 的人工智能训练数据集进行了偏见审计。我们正在实施自动偏见检测工具，以实现 100% 的季度审计。

问题	响应示例
你目前的数据处理能力是多少？你预计未来的生成式 AI 工作负载需要多少能力？	我们目前的容量在一年TB/day. We project needing 30 TB/day内为10个，并且正在扩展我们的基础设施以满足这一需求。
在平衡数据隐私和生成式人工智能模型的数据需求方面，你的策略是什么？	我们正在实施先进的匿名化技术和合成数据生成。我们的目标是在明年将人工智能的可用数据增加40%，同时将隐私风险降低60%。
您的机器学习 (ML) 数据集中有多少百分比被准确标注，您的目标准确率是多少？	目前，我们 85% 的机器学习数据集都已精确标注。我们的目标是通过采用人工和自动标签技术，在下个季度内实现95%的准确率。

将评估见解转化为切实可行的结果

本节提供了一个框架，用于分析问卷答复，并利用这些见解来塑造生成式人工智能现代化计划的目标架构和其他关键可交付成果。该框架弥合了数据收集和实施之间的差距，并确保评估可以直接为您的现代化战略提供依据和推动。

目标架构定义：

- 使用问卷答复为云服务的选择和数据管道的设计提供信息。
- 确保架构设计支持指南中强调的可扩展性和互操作性。

客户准备情况评估：

- 分析与当前基础架构、流程和组织文化相关的问卷答复。
- 找出差距并制定解决这些差距的计划。优先考虑对于 MVP 成功至关重要的差距。

用例和延伸目标：

- 从问卷答复中提取具体的业务问题，以定义明确的用例目标。
- 设定与贵组织生成式 AI 现代化的长期愿景相一致的延伸目标。

工作量估计：

- 使用问卷数据估算 MVP 和全面实施所需的资源、时间和预算。
- 创建从 MVP 开始的分阶段方法，并概述后续阶段。

支持需求：

- 根据调查问卷的答复，确定技能差距和培训需求。
- 制定培训计划，支持即时的 MVP 需求和长期采用生成式 AI。

实施计划：

- 创建一份全面的路线图，从 MVP 开始，概述实现全面生成式 AI 现代化的步骤。
- 为实施的每个阶段定义明确的里程碑和可交付成果。

实际步骤：

- 优先级矩阵：创建一个矩阵，将问卷答复映射到[六个结果](#)，以帮助确定功能和工作的优先顺序。
- 迭代方法：将 MVP 设计为一系列计划版本中的第一个迭代，其中每个版本都朝着完整的目标架构构建。
- 利益相关者协调：使用问卷调查结果使利益相关者了解最有价值球员的范围和实现所有结果的分阶段方法。
- 持续反馈循环：实施机制，在 MVP 部署后收集反馈，并利用见解来完善后续阶段的计划。
- 敏捷实施：采用敏捷方法，允许灵活地处理随时间推移的所有结果，从 MVP 中最关键的结果开始。

后续步骤

完成生成式 AI 工作负载评估后，请按照以下步骤操作：

1. 提供详细的目标架构

- 目标：解决方案架构师创建与组织目标和评估结果相一致的全面目标架构。
- 组件：该架构包括数据摄取、集成点和系统互操作性的设计，以确保可扩展性、可靠性和性能优化。

2. 解释具体 AWS 服务 适合用例的程度

- 服务映射：识别和映射最适合已确定的用例的特定 AWS 服务 服务。
- 优势：重点介绍这些服务如何满足特定的业务需求、提高效率和提供可扩展性。

3. 提供具有优缺点的可选替代解决方案

- 备选方案：提出也可以满足组织要求的替代解决方案。
- 分析：通过考虑成本、复杂性以及与业务目标的一致性等因素，对每种替代方案的优缺点进行详细分析。

4. 提供详细的价格估算 AWS 服务

- 成本分析：提供拟议方案的详细成本估算 AWS 服务，包括潜在的使用场景和定价模型。
- 预算调整：确保成本与组织的预算限制保持一致，并清楚地了解所涉财务问题。

5. 获取有关拟议架构的反馈

- 利益相关者参与：与利益相关者接触，介绍拟议架构并收集反馈。
- 迭代改进：使用反馈来完善和改进解决方案，并确认它符合所有利益相关者的需求和期望。

常见问题解答

生成式 AI 工作量评估的主要目标是什么？

评估的主要目的是评估组织是否准备好对其生成式 AI 工作负载进行现代化改造，确定用例，并开发目标解决方案架构。它旨在定义现代化需求，确定实施范围，并为成功的生成式人工智能现代化做好准备。

谁应该使用这个评估？

此评估适用于想要评估生成式人工智能现代化技术方面的解决方案架构师、企业架构师和应用程序架构师。对于项目经理和人事经理来说，衡量总体准备情况、资源分配和能力需求也很有用。

评估中评估的关键组成部分有哪些？

评估涵盖总体就绪性、用例、架构、存储、法规和合规性、集成、测试、部署自动化和数据策略。这些组件对于确定采用生成式人工智能现代化的技术和组织准备程度至关重要。

评估如何帮助定义目标架构？

该评估提供了一种结构化的方法来评估当前系统并确定改进之处。它可以帮助您选择合适的技术并设计符合业务目标和用例要求的可扩展架构。

进行生成式 AI 工作量评估有什么好处？

好处包括提高效率、改善决策、合规保证、创新促进和可扩展性准备。该评估建立了生成式人工智能现代化的战略方法，并在降低风险的同时最大限度地提高了潜在收益。

各组织如何确保评估后成功实施？

Organizations应制定明确的实施计划，包括明确的里程碑，尽早与利益相关者接触，并采用迭代方法。建立卓越中心 (CoE) 并专注于人才培养也是推荐的最佳做法。

组织在评估期间可能面临哪些挑战？

挑战可能包括变更阻力、数据质量问题和合规复杂性。应对这些挑战需要培养创新文化，确保数据准备就绪，并实施强有力的安全措施。

评估如何满足监管和合规要求？

该评估评估了当前的合规措施并找出了差距。它确保目标解决方案遵守相关法规和数据隐私法，并采用安全最佳实践来保护敏感信息。

利益相关者的参与在评估过程中扮演什么角色？

利益相关者的参与对于获得支持、使现代化计划与业务目标保持一致以及确保成功实施至关重要。尽早参与和明确的利益沟通是克服阻力和促进支持的关键。

评估后，组织如何衡量其生成式人工智能现代化计划的成功？

可以通过使用与业务目标一致的关键绩效指标 (KPIs) 来衡量成功。定期监测和评估这些指标有助于指导决策，并向利益相关者展示生成式人工智能现代化的价值。

对于不同规模（小型、中型或企业）或行业的组织，评估方法有何不同？

小型组织：

- 进行全面评估的资源和专业知识可能有限
- 可能将重点放在特定的高影响力用例上，而不是企业范围内的采用
- 可能更多地依赖第三方工具和服务进行评估
- 评估过程可能不那么正式，更敏捷

中型组织：

- 通常有专门的 IT 或数据团队，但可能缺乏专业的 AI 专业知识
- 可能采取分阶段的方法，首先在关键部门开展试点项目
- 需要在创新与现有系统和流程之间取得平衡

- 评估可能涉及跨职能团队

企业组织：

- 通常有专门的 AI/ML 团队和更多资源进行全面评估
- 需要考虑与现有企业系统的复杂集成
- 可能有行业特定的监管要求需要考虑在内
- 评估通常涉及正式的治理流程

资源

- [生成式 AI 已开启 AWS](#)
- [AWS 提供新的人工智能、机器学习和生成式 AI 指南，用于规划你的 AI 策略](#) (AWS 博客文章)
- [构建生成式 AI 应用程序的最佳实践 AWS](#) (AWS 博客文章)
- [生成式 AI 应用程序生成器 AWS](#) (AWS 解决方案库)
- [生成式 AI 功能](#) (AWS 安全参考架构)
- [AWS 生成式 AI 最佳实践框架](#) (AWS Audit Manager 用户指南)
- [选择生成式 AI 服务](#) (AWS 决策指南)
- [什么是 Amazon Bedrock ?](#) (亚马逊 Bedrock 用户指南)
- [什么是亚马逊 A SageMaker I ?](#) (亚马逊 SageMaker AI 开发者指南)

文档历史记录

下表介绍了本指南的一些重要更改。如果您希望收到有关未来更新的通知，可以订阅 [RSS 源](#)。

变更	说明	日期
初次发布	—	2024 年 11 月 6 日

AWS 规范性指导词汇表

以下是 AWS 规范性指导提供的策略、指南和模式中的常用术语。若要推荐词条，请使用术语表末尾的提供反馈链接。

数字

7 R

将应用程序迁移到云中的 7 种常见迁移策略。这些策略以 Gartner 于 2011 年确定的 5 R 为基础，包括以下内容：

- 重构/重新架构 - 充分利用云原生功能来提高敏捷性、性能和可扩展性，以迁移应用程序并修改其架构。这通常涉及到移植操作系统和数据库。示例：将您的本地 Oracle 数据库迁移到兼容 Amazon Aurora PostgreSQL 的版本。
- 更换平台 - 将应用程序迁移到云中，并进行一定程度的优化，以利用云功能。示例：在中将您的本地 Oracle 数据库迁移到适用于 Oracle 的亚马逊关系数据库服务 (Amazon RDS) AWS Cloud。
- 重新购买 - 转换到其他产品，通常是从传统许可转向 SaaS 模式。示例：将您的客户关系管理 (CRM) 系统迁移到 Salesforce.com。
- 更换主机 (直接迁移) - 将应用程序迁移到云中，无需进行任何更改即可利用云功能。示例：在中的 EC2 实例上将您的本地 Oracle 数据库迁移到 Oracle AWS Cloud。
- 重新定位 (虚拟机监控器级直接迁移)：将基础设施迁移到云中，无需购买新硬件、重写应用程序或修改现有操作。您可以将服务器从本地平台迁移到同一平台的云服务。示例：将 Microsoft Hyper-V 应用程序迁移到 AWS。
- 保留 (重访) - 将应用程序保留在源环境中。其中可能包括需要进行重大重构的应用程序，并且您希望将工作推迟到以后，以及您希望保留的遗留应用程序，因为迁移它们没有商业上的理由。
- 停用 - 停用或删除源环境中不再需要的应用程序。

A

ABAC

请参阅[基于属性的访问控制](#)。

抽象服务

参见[托管服务](#)。

ACID

参见[原子性、一致性、隔离性、持久性](#)。

主动-主动迁移

一种数据库迁移方法，在这种方法中，源数据库和目标数据库保持同步（通过使用双向复制工具或双写操作），两个数据库都在迁移期间处理来自连接应用程序的事务。这种方法支持小批量、可控的迁移，而不需要一次性割接。与[主动-被动迁移](#)相比，它更灵活，但需要更多的工作。

主动-被动迁移

一种数据库迁移方法，在这种方法中，源数据库和目标数据库保持同步，但在将数据复制到目标数据库时，只有源数据库处理来自连接应用程序的事务。目标数据库在迁移期间不接受任何事务。

聚合函数

一个 SQL 函数，它对一组行进行操作并计算该组的单个返回值。聚合函数的示例包括SUM和MAX。

AI

参见[人工智能](#)。

AI Ops

参见[人工智能操作](#)。

匿名化

永久删除数据集中个人信息的过程。匿名化可以帮助保护个人隐私。匿名化数据不再被视为个人数据。

反模式

一种用于解决反复出现的问题的常用解决方案，而在这类问题中，此解决方案适得其反、无效或不如替代方案有效。

应用程序控制

一种安全方法，仅允许使用经批准的应用程序，以帮助保护系统免受恶意软件的侵害。

应用程序组合

有关组织使用的每个应用程序的详细信息的集合，包括构建和维护该应用程序的成本及其业务价值。这些信息是[产品组合发现和分析过程](#)的关键，有助于识别需要进行迁移、现代化和优化的应用程序并确定其优先级。

人工智能 (AI)

计算机科学领域致力于使用计算技术执行通常与人类相关的认知功能，例如学习、解决问题和识别模式。有关更多信息，请参阅[什么是人工智能？](#)

人工智能操作 (AIOps)

使用机器学习技术解决运营问题、减少运营事故和人为干预以及提高服务质量的过程。有关如何在 AIOps AWS 迁移策略中使用的更多信息，请参阅[操作集成指南](#)。

非对称加密

一种加密算法，使用一对密钥，一个公钥用于加密，一个私钥用于解密。您可以共享公钥，因为它不用于解密，但对私钥的访问应受到严格限制。

原子性、一致性、隔离性、持久性 (ACID)

一组软件属性，即使在出现错误、电源故障或其他问题的情况下，也能保证数据库的数据有效性和操作可靠性。

基于属性的访问权限控制 (ABAC)

根据用户属性 (如部门、工作角色和团队名称) 创建精细访问权限的做法。有关更多信息，请参阅 AWS Identity and Access Management (I [AM](#)) 文档 [AWS 中的 AB AC](#)。

权威数据源

存储主要数据版本的位置，被认为是最可靠的信息源。您可以将数据从权威数据源复制到其他位置，以便处理或修改数据，例如对数据进行匿名化、编辑或假名化。

可用区

中的一个不同位置 AWS 区域，不受其他可用区域故障的影响，并向同一区域中的其他可用区提供低成本、低延迟的网络连接。

AWS 云采用框架 (AWS CAF)

该框架包含指导方针和最佳实践 AWS，可帮助组织制定高效且有效的计划，以成功迁移到云端。AWS CAF 将指导分为六个重点领域，称为视角：业务、人员、治理、平台、安全和运营。业务、人员和治理角度侧重于业务技能和流程；平台、安全和运营角度侧重于技术技能和流程。例如，人员角度针对的是负责人力资源 (HR)、人员配置职能和人员管理的利益相关者。从这个角度来看，AWS CAF 为人员发展、培训和沟通提供了指导，以帮助组织为成功采用云做好准备。有关更多信息，请参阅 [AWS CAF 网站](#) 和 [AWS CAF 白皮书](#)。

AWS 工作负载资格框架 (AWS WQF)

一种评估数据库迁移工作负载、推荐迁移策略和提供工作估算的工具。AWS WQF 包含在 AWS Schema Conversion Tool (AWS SCT) 中。它用来分析数据库架构和代码对象、应用程序代码、依赖关系和性能特征，并提供评测报告。

B

坏机器人

旨在破坏个人或组织或对其造成伤害的[机器人](#)。

BCP

参见[业务连续性计划](#)。

行为图

一段时间内资源行为和交互的统一交互式视图。您可以使用 Amazon Detective 的行为图来检查失败的登录尝试、可疑的 API 调用和类似的操作。有关更多信息，请参阅 Detective 文档中的[行为图中的数据](#)。

大端序系统

一个先存储最高有效字节的系统。另请参见[字节顺序](#)。

二进制分类

一种预测二进制结果（两个可能的类别之一）的过程。例如，您的 ML 模型可能需要预测诸如“该电子邮件是否为垃圾邮件？”或“这个产品是书还是汽车？”之类的问题

bloom 筛选条件

一种概率性、内存高效的数据结构，用于测试元素是否为集合的成员。

蓝/绿部署

一种部署策略，您可以创建两个独立但完全相同的环境。在一个环境中运行当前的应用程序版本（蓝色），在另一个环境中运行新的应用程序版本（绿色）。此策略可帮助您在影响最小的情况下快速回滚。

自动程序

一种通过互联网运行自动任务并模拟人类活动或互动的软件应用程序。有些机器人是有用或有益的，例如在互联网上索引信息的网络爬虫。其他一些被称为恶意机器人的机器人旨在破坏个人或组织或对其造成伤害。

僵尸网络

被[恶意软件](#)感染并受单方（称为[机器人](#)牧民或机器人操作员）控制的机器人网络。僵尸网络是最著名的扩展机器人及其影响力的机制。

分支

代码存储库的一个包含区域。在存储库中创建的第一个分支是主分支。您可以从现有分支创建新分支，然后在新分支中开发功能或修复错误。为构建功能而创建的分支通常称为功能分支。当功能可以发布时，将功能分支合并回主分支。有关更多信息，请参阅[关于分支](#)（GitHub 文档）。

破碎的玻璃通道

在特殊情况下，通过批准的流程，用户 AWS 账户 可以快速访问他们通常没有访问权限的内容。有关更多信息，请参阅 Well [-Architected](#) 指南中的“[实施破碎玻璃程序](#)”指示 AWS 器。

棕地策略

您环境中的现有基础设施。在为系统架构采用棕地策略时，您需要围绕当前系统和基础设施的限制来设计架构。如果您正在扩展现有基础设施，则可以将棕地策略和[全新](#)策略混合。

缓冲区缓存

存储最常访问的数据的内存区域。

业务能力

企业如何创造价值（例如，销售、客户服务或营销）。微服务架构和开发决策可以由业务能力驱动。有关更多信息，请参阅[在 AWS 上运行容器化微服务](#)白皮书中的[围绕业务能力进行组织](#)部分。

业务连续性计划（BCP）

一项计划，旨在应对大规模迁移等破坏性事件对运营的潜在影响，并使企业能够快速恢复运营。

C

CAF

参见[AWS 云采用框架](#)。

金丝雀部署

向最终用户缓慢而渐进地发布版本。当你有信心时，你可以部署新版本并全部替换当前版本。

CCoE

参见[云卓越中心](#)。

CDC

请参阅[变更数据捕获](#)。

更改数据捕获 (CDC)

跟踪数据来源 (如数据库表) 的更改并记录有关更改的元数据的过程。您可以将 CDC 用于各种目的, 例如审计或复制目标系统中的更改以保持同步。

混沌工程

故意引入故障或破坏性事件来测试系统的弹性。您可以使用 [AWS Fault Injection Service \(AWS FIS\)](#) 来执行实验, 对您的 AWS 工作负载施加压力并评估其响应。

CI/CD

查看[持续集成和持续交付](#)。

分类

一种有助于生成预测的分类流程。分类问题的 ML 模型预测离散值。离散值始终彼此不同。例如, 一个模型可能需要评估图像中是否有汽车。

客户端加密

在目标 AWS 服务 收到数据之前, 对数据进行本地加密。

云卓越中心 (CCoE)

一个多学科团队, 负责推动整个组织的云采用工作, 包括开发云最佳实践、调动资源、制定迁移时间表、领导组织完成大规模转型。有关更多信息, 请参阅 AWS Cloud 企业战略博客上的 [CCoE 帖子](#)。

云计算

通常用于远程数据存储和 IoT 设备管理的云技术。云计算通常与[边缘计算](#)技术相关。

云运营模型

在 IT 组织中, 一种用于构建、完善和优化一个或多个云环境的运营模型。有关更多信息, 请参阅[构建您的云运营模型](#)。

云采用阶段

组织迁移到以下阶段时通常会经历四个阶段 AWS Cloud :

- 项目 - 出于概念验证和学习目的，开展一些与云相关的项目
- 基础 — 进行基础投资以扩大云采用率（例如，创建 landing zone、定义 CCo E、建立运营模式）
- 迁移 - 迁移单个应用程序
- 重塑 - 优化产品和服务，在云中创新

Stephen Orban 在 AWS Cloud 企业战略博客的博客文章 [《云优先之旅和采用阶段》](#) 中定义了这些阶段。有关它们与 AWS 迁移策略的关系的信息，请参阅 [迁移准备指南](#)。

CMDB

参见 [配置管理数据库](#)。

代码存储库

通过版本控制过程存储和更新源代码和其他资产（如文档、示例和脚本）的位置。常见的云存储库包括 GitHub 或 Bitbucket Cloud。每个版本的代码都称为一个分支。在微服务结构中，每个存储库都专门用于一个功能。单个 CI/CD 管道可以使用多个存储库。

冷缓存

一种空的、填充不足或包含过时或不相关数据的缓冲区缓存。这会影响性能，因为数据库实例必须从主内存或磁盘读取，这比从缓冲区缓存读取要慢。

冷数据

很少访问的数据，且通常是历史数据。查询此类数据时，通常可以接受慢速查询。将这些数据转移到性能较低且成本更低的存储层或类别可以降低成本。

计算机视觉 (CV)

一种 [AI](#) 领域，它使用机器学习来分析和提取数字图像和视频等视觉格式中的信息。例如，Amazon SageMaker AI 为 CV 提供了图像处理算法。

配置偏差

对于工作负载，配置会从预期状态发生变化。这可能会导致工作负载变得不合规，而且通常是渐进的，不是故意的。

配置管理数据库 (CMDB)

一种存储库，用于存储和管理有关数据库及其 IT 环境的信息，包括硬件和软件组件及其配置。您通常在迁移的产品组合发现和分析阶段使用来自 CMDB 的数据。

合规性包

一系列 AWS Config 规则和补救措施，您可以汇编这些规则和补救措施，以自定义您的合规性和安全性检查。您可以使用 YAML 模板将一致性包作为单个实体部署在 AWS 账户 和区域或整个组织中。有关更多信息，请参阅 AWS Config 文档中的 [一致性包](#)。

持续集成和持续交付 (CI/CD)

自动执行软件发布过程的源代码、构建、测试、暂存和生产阶段的过程。CI/CD 通常被描述为管道。CI/CD 可以帮助您实现流程自动化、提高生产力、提高代码质量和更快地交付。有关更多信息，请参阅[持续交付的优势](#)。CD 也可以表示持续部署。有关更多信息，请参阅[持续交付与持续部署](#)。

CV

参见[计算机视觉](#)。

D

静态数据

网络中静止的数据，例如存储中的数据。

数据分类

根据网络中数据的关键性和敏感性对其进行识别和分类的过程。它是任何网络安全风险管理策略的关键组成部分，因为它可以帮助您确定对数据的适当保护和保留控制。数据分类是 Well-Architected AWS d Framework 中安全支柱的一个组成部分。有关详细信息，请参阅[数据分类](#)。

数据漂移

生产数据与用来训练机器学习模型的数据之间的有意义差异，或者输入数据随时间推移的有意义变化。数据漂移可能降低机器学习模型预测的整体质量、准确性和公平性。

传输中数据

在网络中主动移动的数据，例如在网络资源之间移动的数据。

数据网格

一种架构框架，可提供分布式、去中心化的数据所有权以及集中式管理和治理。

数据最少化

仅收集并处理绝对必要数据的原则。在中进行数据最小化 AWS Cloud 可以降低隐私风险、成本和分析碳足迹。

数据边界

AWS 环境中的一组预防性防护措施，可帮助确保只有可信身份才能访问来自预期网络的可信资源。有关更多信息，请参阅在[上构建数据边界](#)。AWS

数据预处理

将原始数据转换为 ML 模型易于解析的格式。预处理数据可能意味着删除某些列或行，并处理缺失、不一致或重复的值。

数据溯源

在数据的整个生命周期跟踪其来源和历史的过程，例如数据如何生成、传输和存储。

数据主体

正在收集和处理其数据的人。

数据仓库

一种支持商业智能（例如分析）的数据管理系统。数据仓库通常包含大量历史数据，通常用于查询和分析。

数据库定义语言（DDL）

在数据库中创建或修改表和对对象结构的语句或命令。

数据库操作语言（DML）

在数据库中修改（插入、更新和删除）信息的语句或命令。

DDL

参见[数据库定义语言](#)。

深度融合

组合多个深度学习模型进行预测。您可以使用深度融合来获得更准确的预测或估算预测中的不确定性。

深度学习

一个 ML 子字段使用多层人工神经网络来识别输入数据和感兴趣的目标变量之间的映射。

defense-in-depth

一种信息安全方法，经过深思熟虑，在整个计算机网络中分层实施一系列安全机制和控制措施，以保护网络及其中数据的机密性、完整性和可用性。当你采用这种策略时 AWS，你会在 AWS

Organizations 结构的不同层面添加多个控件来帮助保护资源。例如，一种 defense-in-depth 方法可以结合多因素身份验证、网络分段和加密。

委托管理员

在中 AWS Organizations，兼容的服务可以注册 AWS 成员帐户来管理组织的帐户并管理该服务的权限。此账户被称为该服务的委托管理员。有关更多信息和兼容服务列表，请参阅 AWS Organizations 文档中[使用 AWS Organizations 的服务](#)。

后

使应用程序、新功能或代码修复在目标环境中可用的过程。部署涉及在代码库中实现更改，然后在应用程序的环境中构建和运行该代码库。

开发环境

参见[环境](#)。

侦测性控制

一种安全控制，在事件发生后进行检测、记录日志和发出警报。这些控制是第二道防线，提醒您注意绕过现有预防性控制的安全事件。有关更多信息，请参阅在 AWS 上实施安全控制中的[侦测性控制](#)。

开发价值流映射 (DVSM)

用于识别对软件开发生命周期中的速度和质量产生不利影响的限制因素并确定其优先级的流程。DVSM 扩展了最初为精益生产实践设计的价值流映射流程。其重点关注在软件开发过程中创造和转移价值所需的步骤和团队。

数字孪生

真实世界系统的虚拟再现，如建筑物、工厂、工业设备或生产线。数字孪生支持预测性维护、远程监控和生产优化。

维度表

在[星型架构](#)中，一种较小的表，其中包含事实表中定量数据的数据属性。维度表属性通常是文本字段或行为类似于文本的离散数字。这些属性通常用于查询约束、筛选和结果集标注。

灾难

阻止工作负载或系统在其主要部署位置实现其业务目标的事件。这些事件可能是自然灾害、技术故障或人为操作的结果，例如无意的配置错误或恶意软件攻击。

灾难恢复 (DR)

您用来最大限度地减少[灾难](#)造成的停机时间和数据丢失的策略和流程。有关更多信息，请参阅 Well-Architected Framework AWS work 中的“[工作负载灾难恢复：云端 AWS 恢复](#)”。

DML

参见[数据库操作语言](#)。

领域驱动设计

一种开发复杂软件系统的方法，通过将其组件连接到每个组件所服务的不断发展的领域或核心业务目标。Eric Evans 在其著作领域驱动设计：软件核心复杂性应对之道 (Boston: Addison-Wesley Professional, 2003) 中介绍了这一概念。有关如何将领域驱动设计与 strangler fig 模式结合使用的信息，请参阅[使用容器和 Amazon API Gateway 逐步将原有的 Microsoft ASP.NET \(ASMX \) Web 服务现代化](#)。

DR

参见[灾难恢复](#)。

漂移检测

跟踪与基准配置的偏差。例如，您可以使用 AWS CloudFormation 来[检测系统资源中的偏差](#)，也可以使用 AWS Control Tower 来[检测着陆区中可能影响监管要求合规性的变化](#)。

DVSM

参见[开发价值流映射](#)。

E

EDA

参见[探索性数据分析](#)。

EDI

参见[电子数据交换](#)。

边缘计算

该技术可提高位于 IoT 网络边缘的智能设备的计算能力。与[云计算](#)相比，边缘计算可以减少通信延迟并缩短响应时间。

电子数据交换 (EDI)

组织之间自动交换业务文档。有关更多信息，请参阅[什么是电子数据交换](#)。

加密

一种将人类可读的纯文本数据转换为密文的计算过程。

加密密钥

由加密算法生成的随机位的加密字符串。密钥的长度可能有所不同，而且每个密钥都设计为不可预测且唯一。

字节顺序

字节在计算机内存中的存储顺序。大端序系统先存储最高有效字节。小端序系统先存储最低有效字节。

端点

参见[服务端点](#)。

端点服务

一种可以在虚拟私有云 (VPC) 中托管，与其他用户共享的服务。您可以使用其他 AWS 账户 或 AWS Identity and Access Management (IAM) 委托人创建终端节点服务，AWS PrivateLink 并向其授予权限。这些账户或主体可通过创建接口 VPC 端点来私密地连接到您的端点服务。有关更多信息，请参阅 Amazon Virtual Private Cloud (Amazon VPC) 文档中的[创建端点服务](#)。

企业资源规划 (ERP)

一种自动化和管理企业关键业务流程（例如会计、[MES](#) 和项目管理）的系统。

信封加密

用另一个加密密钥对加密密钥进行加密的过程。有关更多信息，请参阅 AWS Key Management Service (AWS KMS) 文档中的[信封加密](#)。

环境

正在运行的应用程序的实例。以下是云计算中常见的环境类型：

- 开发环境 — 正在运行的应用程序的实例，只有负责维护应用程序的核心团队才能使用。开发环境用于测试更改，然后再将其提升到上层环境。这类环境有时称为测试环境。
- 下层环境 — 应用程序的所有开发环境，比如用于初始构建和测试的环境。

- 生产环境 — 最终用户可以访问的正在运行的应用程序的实例。在 CI/CD 管道中，生产环境是最后一个部署环境。
- 上层环境 — 除核心开发团队以外的用户可以访问的所有环境。这可能包括生产环境、预生产环境和用户验收测试环境。

epic

在敏捷方法学中，有助于组织工作和确定优先级的功能类别。epics 提供了对需求和实施任务的总体描述。例如，AWS CAF 安全史诗包括身份和访问管理、侦探控制、基础设施安全、数据保护和事件响应。有关 AWS 迁移策略中 epics 的更多信息，请参阅[计划实施指南](#)。

ERP

参见[企业资源规划](#)。

探索性数据分析 (EDA)

分析数据集以了解其主要特征的过程。您收集或汇总数据，并进行初步调查，以发现模式、检测异常并检查假定情况。EDA 通过计算汇总统计数据和创建数据可视化得以执行。

F

事实表

[星形架构](#)中的中心表。它存储有关业务运营的定量数据。通常，事实表包含两种类型的列：包含度量的列和包含维度表外键的列。

失败得很快

一种使用频繁和增量测试来缩短开发生命周期的理念。这是敏捷方法的关键部分。

故障隔离边界

在中 AWS Cloud，诸如可用区 AWS 区域、控制平面或数据平面之类的边界，它限制了故障的影响并有助于提高工作负载的弹性。有关更多信息，请参阅[AWS 故障隔离边界](#)。

功能分支

参见[分支](#)。

特征

您用来进行预测的输入数据。例如，在制造环境中，特征可能是定期从生产线捕获的图像。

特征重要性

特征对于模型预测的重要性。这通常表示为数值分数，可以通过各种技术进行计算，例如 Shapley 加法解释 (SHAP) 和积分梯度。有关更多信息，请参阅使用[机器学习模型的可解释性 AWS](#)。

功能转换

为 ML 流程优化数据，包括使用其他来源丰富数据、扩展值或从单个数据字段中提取多组信息。这使得 ML 模型能从数据中获益。例如，如果您将“2021-05-27 00:15:37”日期分解为“2021”、“五月”、“星期四”和“15”，则可以帮助学习与不同数据成分相关的算法学习精细模式。

少量提示

在要求[法学硕士](#)执行类似任务之前，向其提供少量示例，以演示该任务和所需的输出。这种技术是情境学习的应用，模型可以从提示中嵌入的示例 (镜头) 中学习。对于需要特定格式、推理或领域知识的任务，Few-shot 提示可能非常有效。另请参见[零镜头提示](#)。

FGAC

请参阅[精细的访问控制](#)。

精细访问控制 (FGAC)

使用多个条件允许或拒绝访问请求。

快闪迁移

一种数据库迁移方法，它使用连续的数据复制，通过[更改数据捕获](#)在尽可能短的时间内迁移数据，而不是使用分阶段的方法。目标是将停机时间降至最低。

FM

参见[基础模型](#)。

基础模型 (FM)

一个大型深度学习神经网络，一直在广义和未标记数据的大量数据集上进行训练。FMs 能够执行各种各样的一般任务，例如理解语言、生成文本和图像以及用自然语言进行对话。有关更多信息，请参阅[什么是基础模型](#)。

G

生成式人工智能

[人工智能](#)模型的子集，这些模型已经过大量数据训练，可以使用简单的文本提示来创建新的内容和工件，例如图像、视频、文本和音频。有关更多信息，请参阅[什么是生成式 AI](#)。

地理封锁

请参阅[地理限制](#)。

地理限制 (地理阻止)

在 Amazon 中 CloudFront，一种阻止特定国家/地区的用户访问内容分发的选项。您可以使用允许列表或阻止列表来指定已批准和已禁止的国家/地区。有关更多信息，请参阅 CloudFront 文档[中的限制内容的地理分布](#)。

GitFlow 工作流程

一种方法，在这种方法中，下层和上层环境在源代码存储库中使用不同的分支。Gitflow 工作流程被认为是传统的，而[基于主干的工作流程](#)是现代的首选方法。

金色影像

系统或软件的快照，用作部署该系统或软件的新实例的模板。例如，在制造业中，黄金映像可用于在多个设备上配置软件，并有助于提高设备制造运营的速度、可扩展性和生产力。

全新策略

在新环境中缺少现有基础设施。在对系统架构采用全新策略时，您可以选择所有新技术，而不受对现有基础设施 (也称为[棕地](#)) 兼容性的限制。如果您正在扩展现有基础设施，则可以将棕地策略和全新策略混合。

防护机制

一项高级规则，可帮助管理各组织单位的资源、策略和合规性 (OUs)。预防性防护机制会执行策略以确保符合合规性标准。它们是使用服务控制策略和 IAM 权限边界实现的。侦测性防护机制会检测策略违规和合规性问题，并生成警报以进行修复。它们通过使用 AWS Config、Amazon、AWS Security Hub GuardDuty AWS Trusted Advisor、Amazon Inspector 和自定义 AWS Lambda 支票来实现。

H

HA

参见[高可用性](#)。

异构数据库迁移

将源数据库迁移到使用不同数据库引擎的目标数据库 (例如，从 Oracle 迁移到 Amazon Aurora)。异构迁移通常是重新架构工作的一部分，而转换架构可能是一项复杂的任务。[AWS 提供了 AWS SCT](#) 来帮助实现架构转换。

高可用性 (HA)

在遇到挑战或灾难时，工作负载无需干预即可连续运行的能力。HA 系统旨在自动进行故障转移、持续提供良好性能，并以最小的性能影响处理不同负载和故障。

历史数据库现代化

一种用于实现运营技术 (OT) 系统现代化和升级以更好满足制造业需求的方法。历史数据库是一种用于收集和存储工厂中各种来源数据的数据库。

抵制数据

从用于训练[机器学习](#)模型的数据集中扣留的一部分带有标签的历史数据。通过将模型预测与抵制数据进行比较，您可以使用抵制数据来评估模型性能。

同构数据库迁移

将源数据库迁移到共享同一数据库引擎的目标数据库（例如，从 Microsoft SQL Server 迁移到 Amazon RDS for SQL Server）。同构迁移通常是更换主机或更换平台工作的一部分。您可以使用本机数据库实用程序来迁移架构。

热数据

经常访问的数据，例如实时数据或近期的转化数据。这些数据通常需要高性能存储层或存储类别才能提供快速的查询响应。

修补程序

针对生产环境中关键问题的紧急修复。由于其紧迫性，修补程序通常是在典型的 DevOps 发布工作流程之外进行的。

hypercure 周期

割接之后，迁移团队立即管理和监控云中迁移的应用程序以解决任何问题的时间段。通常，这个周期持续 1-4 天。在 hypercure 周期结束时，迁移团队通常会将应用程序的责任移交给云运营团队。

我

IaC

参见[基础设施即代码](#)。

基于身份的策略

附加到一个或多个 IAM 委托人的策略，用于定义他们在 AWS Cloud 环境中的权限。

空闲应用程序

90 天内平均 CPU 和内存使用率在 5% 到 20% 之间的应用程序。在迁移项目中，通常会停用这些应用程序或将其保留在本地。

IIoT

参见 [工业物联网](#)。

不可变的基础架构

一种为生产工作负载部署新基础架构，而不是更新、修补或修改现有基础架构的模型。[不可变基础架构本质上比可变基础架构更一致、更可靠、更可预测](#)。有关更多信息，请参阅 Well-Architected Framework 中的[使用不可变基础架构 AWS 部署](#)最佳实践。

入站 (入口) VPC

在 AWS 多账户架构中，一种接受、检查和路由来自应用程序外部的网络连接的 VPC。[AWS 安全参考架构](#)建议设置您的网络帐户，包括入站、出站和检查，VPCs 以保护您的应用程序与更广泛的互联网之间的双向接口。

增量迁移

一种割接策略，在这种策略中，您可以将应用程序分成小部分进行迁移，而不是一次性完整割接。例如，您最初可能只将几个微服务或用户迁移到新系统。在确认一切正常后，您可以逐步迁移其他微服务或用户，直到停用遗留系统。这种策略降低了大规模迁移带来的风险。

工业 4.0

该术语由[克劳斯·施瓦布 \(Klaus Schwab \)](#)于2016年推出，指的是通过连接、实时数据、自动化、分析和人工智能/机器学习的进步实现制造流程的现代化。

基础设施

应用程序环境中包含的所有资源和资产。

基础设施即代码 (IaC)

通过一组配置文件预置和管理应用程序基础设施的过程。IaC 旨在帮助您集中管理基础设施、实现资源标准化和快速扩展，使新环境具有可重复性、可靠性和一致性。

工业物联网 (IIoT)

在工业领域使用联网的传感器和设备，例如制造业、能源、汽车、医疗保健、生命科学和农业。有关更多信息，请参阅[制定工业物联网 \(IIoT\) 数字化转型战略](#)。

检查 VPC

在 AWS 多账户架构中，一种集中式 VPC，用于管理对 VPCs（相同或不同 AWS 区域）、互联网和本地网络之间的网络流量的检查。[AWS 安全参考架构](#)建议设置您的网络帐户，包括入站、出站和检查，VPCs 以保护您的应用程序与更广泛的互联网之间的双向接口。

物联网 (IoT)

由带有嵌入式传感器或处理器的连接物理对象组成的网络，这些传感器或处理器通过互联网或本地通信网络与其他设备和系统进行通信。有关更多信息，请参阅[什么是 IoT ?](#)

可解释性

它是机器学习模型的一种特征，描述了人类可以理解模型的预测如何取决于其输入的程度。有关更多信息，请参阅使用[机器学习模型的可解释性 AWS](#)。

IoT

参见[物联网](#)。

IT 信息库 (ITIL)

提供 IT 服务并使这些服务符合业务要求的一套最佳实践。ITIL 是 ITSM 的基础。

IT 服务管理 (ITSM)

为组织设计、实施、管理和支持 IT 服务的相关活动。有关将云运营与 ITSM 工具集成的信息，请参阅[运营集成指南](#)。

ITIL

请参阅[IT 信息库](#)。

ITSM

请参阅[IT 服务管理](#)。

L

基于标签的访问控制 (LBAC)

强制访问控制 (MAC) 的一种实施方式，其中明确为用户和数据本身分配了安全标签值。用户安全标签和数据安全标签之间的交集决定了用户可以看到哪些行和列。

登录区

landing zone 是一个架构精良的多账户 AWS 环境，具有可扩展性和安全性。这是一个起点，您的组织可以从这里放心地在安全和基础设施环境中快速启动和部署工作负载和应用程序。有关登录区的更多信息，请参阅[设置安全且可扩展的多账户 AWS 环境](#)。

大型语言模型 (LLM)

一种基于大量数据进行预训练的深度学习 [AI](#) 模型。法学硕士可以执行多项任务，例如回答问题、总结文档、将文本翻译成其他语言以及完成句子。有关更多信息，请参阅[什么是 LLMs](#)。

大规模迁移

迁移 300 台或更多服务器。

LBAC

请参阅[基于标签的访问控制](#)。

最低权限

授予执行任务所需的最低权限的最佳安全实践。有关更多信息，请参阅 IAM 文档中的[应用最低权限许可](#)。

直接迁移

见 [7 R](#)。

小端序系统

一个先存储最低有效字节的系统。另请参见[字节顺序](#)。

LLM

参见[大型语言模型](#)。

下层环境

参见[环境](#)。

M

机器学习 (ML)

一种使用算法和技术进行模式识别和学习的人工智能。ML 对记录的数据 (例如物联网 (IoT) 数据) 进行分析和学习，以生成基于模式的统计模型。有关更多信息，请参阅[机器学习](#)。

主分支

参见[分支](#)。

恶意软件

旨在危害计算机安全或隐私的软件。恶意软件可能会破坏计算机系统、泄露敏感信息或获得未经授权的访问。恶意软件的示例包括病毒、蠕虫、勒索软件、特洛伊木马、间谍软件和键盘记录器。

托管服务

AWS 服务 它 AWS 运行基础设施层、操作系统和平台，您可以访问端点来存储和检索数据。亚马逊简单存储服务 (Amazon S3) Service 和 Amazon DynamoDB 就是托管服务的示例。这些服务也称为抽象服务。

制造执行系统 (MES)

一种软件系统，用于跟踪、监控、记录和控制将原材料转化为成品的生产过程。

MAP

参见[迁移加速计划](#)。

机制

一个完整的过程，在此过程中，您可以创建工具，推动工具的采用，然后检查结果以进行调整。机制是一种在运行过程中自我增强和改进的循环。有关更多信息，请参阅在 Well-Architect AWS ed 框架中[构建机制](#)。

成员账户

AWS 账户 除属于组织中的管理账户之外的所有账户 AWS Organizations。一个账户一次只能是一个组织的成员。

MES

参见[制造执行系统](#)。

消息队列遥测传输 (MQTT)

[一种基于发布/订阅模式的轻量级 machine-to-machine \(M2M\) 通信协议，适用于资源受限的物联网设备。](#)

微服务

一种小型的独立服务，通过明确的定义进行通信 APIs，通常由小型的独立团队拥有。例如，保险系统可能包括映射到业务能力（如销售或营销）或子域（如购买、理赔或分析）的微服务。微服务

的好处包括敏捷、灵活扩展、易于部署、可重复使用的代码和恢复能力。有关更多信息，请参阅[使用 AWS 无服务器服务集成微服务](#)。

微服务架构

一种使用独立组件构建应用程序的方法，这些组件将每个应用程序进程作为微服务运行。这些微服务使用轻量级通过定义明确的接口进行通信。APIs 该架构中的每个微服务都可以更新、部署和扩展，以满足对应用程序特定功能的需求。有关更多信息，请参阅[在上实现微服务](#)。AWS

迁移加速计划 (MAP)

AWS 该计划提供咨询支持、培训和服务，以帮助组织为迁移到云奠定坚实的运营基础，并帮助抵消迁移的初始成本。MAP 提供了一种以系统的方式执行遗留迁移的迁移方法，以及一套用于自动执行和加速常见迁移场景的工具。

大规模迁移

将大部分应用程序组合分波迁移到云中的过程，在每一波中以更快的速度迁移更多应用程序。本阶段使用从早期阶段获得的最佳实践和经验教训，实施由团队、工具和流程组成的迁移工厂，通过自动化和敏捷交付简化工作负载的迁移。这是 [AWS 迁移策略](#) 的第三阶段。

迁移工厂

跨职能团队，通过自动化、敏捷的方法简化工作负载迁移。迁移工厂团队通常包括运营、业务分析师和所有者、迁移工程师、开发人员和冲刺 DevOps 领域的专业人员。20% 到 50% 的企业应用程序组合由可通过工厂方法优化的重复模式组成。有关更多信息，请参阅本内容集中[有关迁移工厂的讨论](#)和[云迁移工厂](#)指南。

迁移元数据

有关完成迁移所需的应用程序和服务器信息。每种迁移模式都需要一套不同的迁移元数据。迁移元数据的示例包括目标子网、安全组和 AWS 账户。

迁移模式

一种可重复的迁移任务，详细列出了迁移策略、迁移目标以及所使用的迁移应用程序或服务。示例：EC2 使用 AWS 应用程序迁移服务重新托管向 Amazon 的迁移。

迁移组合评测 (MPA)

一种在线工具，可提供信息，用于验证迁移到的业务案例。AWS Cloud MPA 提供了详细的组合评测（服务器规模调整、定价、TCO 比较、迁移成本分析）以及迁移计划（应用程序数据分析和数据收集、应用程序分组、迁移优先级排序和波次规划）。所有 AWS 顾问和 APN 合作伙伴顾问均可免费使用 [MPA 工具](#)（需要登录）。

迁移准备情况评测 (MRA)

使用 AWS CAF 深入了解组织的云就绪状态、确定优势和劣势以及制定行动计划以缩小已发现差距的过程。有关更多信息，请参阅[迁移准备指南](#)。MRA 是 [AWS 迁移策略](#)的第一阶段。

迁移策略

用于将工作负载迁移到的方法 AWS Cloud。有关更多信息，请参阅此词汇表中的“[7 R](#)”条目，并参阅[调动组织以加快大规模迁移](#)。

ML

参见[机器学习](#)。

现代化

将过时的（原有的或单体）应用程序及其基础设施转变为云中敏捷、弹性和高度可用的系统，以降低成本、提高效率和利用创新。有关更多信息，请参阅[中的应用程序现代化策略](#)。AWS Cloud

现代化准备情况评估

一种评估方式，有助于确定组织应用程序的现代化准备情况；确定收益、风险和依赖关系；确定组织能够在多大程度上支持这些应用程序的未来状态。评估结果是目标架构的蓝图、详细说明现代化进程发展阶段和里程碑的路线图以及解决已发现差距的行动计划。有关更多信息，请参阅[中的评估应用程序的现代化准备情况](#) AWS Cloud。

单体应用程序 (单体式)

作为具有紧密耦合进程的单个服务运行的应用程序。单体应用程序有几个缺点。如果某个应用程序功能的需求激增，则必须扩展整个架构。随着代码库的增长，添加或改进单体应用程序的功能也会变得更加复杂。若要解决这些问题，可以使用微服务架构。有关更多信息，请参阅[将单体分解为微服务](#)。

MPA

参见[迁移组合评估](#)。

MQTT

请参阅[消息队列遥测传输](#)。

多分类器

一种帮助为多个类别生成预测（预测两个以上结果之一）的过程。例如，ML 模型可能会询问“这个产品是书、汽车还是手机？”或“此客户最感兴趣什么类别的产品？”

可变基础架构

一种用于更新和修改现有生产工作负载基础架构的模型。为了提高一致性、可靠性和可预测性，Well-Architect AWS ed Framework 建议使用[不可变基础设施](#)作为最佳实践。

O

OAC

请参阅[源站访问控制](#)。

OAI

参见[源访问身份](#)。

OCM

参见[组织变更管理](#)。

离线迁移

一种迁移方法，在这种方法中，源工作负载会在迁移过程中停止运行。这种方法会延长停机时间，通常用于小型非关键工作负载。

OI

参见[运营集成](#)。

OLA

参见[运营层协议](#)。

在线迁移

一种迁移方法，在这种方法中，源工作负载无需离线即可复制到目标系统。在迁移过程中，连接工作负载的应用程序可以继续运行。这种方法的停机时间为零或最短，通常用于关键生产工作负载。

OPC-UA

参见[开放流程通信-统一架构](#)。

开放流程通信-统一架构 (OPC-UA)

一种用于工业自动化的 machine-to-machine (M2M) 通信协议。OPC-UA 提供了数据加密、身份验证和授权方案的互操作性标准。

运营级别协议 (OLA)

一项协议，阐明了 IT 职能部门承诺相互交付的内容，以支持服务水平协议 (SLA)。

运营准备情况审查 (ORR)

一份问题清单和相关的最佳实践，可帮助您理解、评估、预防或缩小事件和可能的故障的范围。有关更多信息，请参阅 Well-Architecte AWS d Frame [work 中的运营准备情况评估 \(ORR\)](#)。

操作技术 (OT)

与物理环境配合使用以控制工业运营、设备和基础设施的硬件和软件系统。在制造业中，OT 和信息技术 (IT) 系统的集成是[工业 4.0](#) 转型的重点。

运营整合 (OI)

在云中实现运营现代化的过程，包括就绪计划、自动化和集成。有关更多信息，请参阅[运营整合指南](#)。

组织跟踪

由此创建的跟踪 AWS CloudTrail，用于记录组织 AWS 账户 中所有人的所有事件 AWS Organizations。该跟踪是在每个 AWS 账户 中创建的，属于组织的一部分，并跟踪每个账户的活动。有关更多信息，请参阅 CloudTrail文档中的[为组织创建跟踪](#)。

组织变革管理 (OCM)

一个从人员、文化和领导力角度管理重大、颠覆性业务转型的框架。OCM 通过加快变革采用、解决过渡问题以及推动文化和组织变革，帮助组织为新系统和战略做好准备和过渡。在 AWS 迁移策略中，该框架被称为人员加速，因为云采用项目需要变更的速度。有关更多信息，请参阅[OCM 指南](#)。

来源访问控制 (OAC)

在中 CloudFront，一个增强的选项，用于限制访问以保护您的亚马逊简单存储服务 (Amazon S3) 内容。OAC 全部支持所有 S3 存储桶 AWS 区域、使用 AWS KMS (SSE-KMS) 进行服务器端加密，以及对 S3 存储桶的动态PUT和DELETE请求。

来源访问身份 (OAI)

在中 CloudFront，一个用于限制访问权限以保护您的 Amazon S3 内容的选项。当您使用 OAI 时，CloudFront 会创建一个 Amazon S3 可以对其进行身份验证的委托人。经过身份验证的委托人只能通过特定 CloudFront 分配访问 S3 存储桶中的内容。另请参阅[OAC](#)，其中提供了更精细和增强的访问控制。

ORR

参见[运营准备情况审查](#)。

OT

参见[运营技术](#)。

出站 (出口) VPC

在 AWS 多账户架构中，一种处理从应用程序内部启动的网络连接的 VPC。[AWS 安全参考架构](#)建议设置您的网络帐户，包括入站、出站和检查，VPCs 以保护您的应用程序与更广泛的互联网之间的双向接口。

P

权限边界

附加到 IAM 主体的 IAM 管理策略，用于设置用户或角色可以拥有的最大权限。有关更多信息，请参阅 IAM 文档中的[权限边界](#)。

个人身份信息 (PII)

直接查看其他相关数据或与之配对时可用于合理推断个人身份的信息。PII 的示例包括姓名、地址和联系信息。

PII

查看[个人身份信息](#)。

playbook

一套预定义的步骤，用于捕获与迁移相关的工作，例如在云中交付核心运营功能。playbook 可以采用脚本、自动化运行手册的形式，也可以是操作现代化环境所需的流程或步骤的摘要。

PLC

参见[可编程逻辑控制器](#)。

PLM

参见[产品生命周期管理](#)。

policy

一个对象，可以在中定义权限（参见[基于身份的策略](#)）、指定访问条件（参见[基于资源的策略](#)）或定义组织中所有账户的最大权限 AWS Organizations（参见[服务控制策略](#)）。

多语言持久性

根据数据访问模式和其他要求，独立选择微服务的数据存储技术。如果您的微服务采用相同的数据存储技术，它们可能会遇到实现难题或性能不佳。如果微服务使用最适合其需求的数据存储，则可以更轻松地实现微服务，并获得更好的性能和可扩展性。有关更多信息，请参阅[在微服务中实现数据持久性](#)。

组合评测

一个发现、分析和确定应用程序组合优先级以规划迁移的过程。有关更多信息，请参阅[评估迁移准备情况](#)。

谓词

返回true或的查询条件false，通常位于子WHERE句中。

谓词下推

一种数据库查询优化技术，可在传输前筛选查询中的数据。这减少了必须从关系数据库检索和处理的数据量，并提高了查询性能。

预防性控制

一种安全控制，旨在防止事件发生。这些控制是第一道防线，帮助防止未经授权的访问或对网络的意外更改。有关更多信息，请参阅在 AWS 上实施安全控制中的[预防性控制](#)。

主体

中 AWS 可以执行操作和访问资源的实体。此实体通常是 IAM 角色的根用户或用户。AWS 账户有关更多信息，请参阅 IAM 文档中[角色术语和概念](#)中的主体。

通过设计保护隐私

一种在整个开发过程中考虑隐私的系统工程方法。

私有托管区

一个容器，其中包含有关您希望 Amazon Route 53 如何响应针对一个或多个 VPCs 域名及其子域名的 DNS 查询的信息。有关更多信息，请参阅 Route 53 文档中的[私有托管区的使用](#)。

主动控制

一种[安全控制](#)措施，旨在防止部署不合规的资源。这些控件会在资源配置之前对其进行扫描。如果资源与控件不兼容，则不会对其进行配置。有关更多信息，请参阅 AWS Control Tower 文档中的[控制参考指南](#)，并参见在上实施安全[控制中的主动](#)控制 AWS。

产品生命周期管理 (PLM)

在产品的整个生命周期中，从设计、开发和上市，到成长和成熟，再到衰落和移除，对产品进行数据和流程的管理。

生产环境

参见[环境](#)。

可编程逻辑控制器 (PLC)

在制造业中，一种高度可靠、适应性强的计算机，用于监控机器并实现制造过程自动化。

提示链接

使用一个 [LLM](#) 提示的输出作为下一个提示的输入，以生成更好的响应。该技术用于将复杂的任务分解为子任务，或者迭代地完善或扩展初步响应。它有助于提高模型响应的准确性和相关性，并允许获得更精细的个性化结果。

假名化

用占位符值替换数据集中个人标识符的过程。假名化可以帮助保护个人隐私。假名化数据仍被视为个人数据。

publish/subscribe (pub/sub)

一种支持微服务间异步通信的模式，以提高可扩展性和响应能力。例如，在基于微服务的 [MES](#) 中，微服务可以将事件消息发布到其他微服务可以订阅的频道。系统可以在不更改发布服务的情况下添加新的微服务。

Q

查询计划

一系列步骤，例如指令，用于访问 SQL 关系数据库系统中的数据。

查询计划回归

当数据库服务优化程序选择的最佳计划不如数据库环境发生特定变化之前时。这可能是由统计数据、约束、环境设置、查询参数绑定更改和数据库引擎更新造成的。

R

RACI 矩阵

参见[负责任、负责、咨询、知情 \(RACI \)](#)。

RAG

请参见[检索增强生成](#)。

勒索软件

一种恶意软件，旨在阻止对计算机系统或数据的访问，直到付款为止。

RASCI 矩阵

参见[负责任、负责、咨询、知情 \(RACI \)](#)。

RCAC

请参阅[行和列访问控制](#)。

只读副本

用于只读目的的数据库副本。您可以将查询路由到只读副本，以减轻主数据库的负载。

重新架构师

见 [7 R](#)。

恢复点目标 (RPO)

自上一个数据恢复点以来可接受的最长时间。这决定了从上一个恢复点到服务中断之间可接受的数据丢失情况。

恢复时间目标 (RTO)

服务中断和服务恢复之间可接受的最大延迟。

重构

见 [7 R](#)。

区域

地理区域内的 AWS 资源集合。每一个 AWS 区域 都相互隔离，彼此独立，以提供容错、稳定性和弹性。有关更多信息，请参阅[指定 AWS 区域 您的账户可以使用的账户](#)。

回归

一种预测数值的 ML 技术。例如，要解决“这套房子的售价是多少？”的问题 ML 模型可以使用线性回归模型，根据房屋的已知事实（如建筑面积）来预测房屋的销售价格。

重新托管

见 [7 R](#)。

版本

在部署过程中，推动生产环境变更的行为。

搬迁

见 [7 R](#)。

更换平台

见 [7 R](#)。

回购

见 [7 R](#)。

故障恢复能力

应用程序抵御中断或从中断中恢复的能力。在中规划弹性时，[高可用性](#)和[灾难恢复](#)是常见的考虑因素。AWS Cloud有关更多信息，请参阅[AWS Cloud 弹性](#)。

基于资源的策略

一种附加到资源的策略，例如 AmazonS3 存储桶、端点或加密密钥。此类策略指定了允许哪些主体访问、支持的操作以及必须满足的任何其他条件。

责任、问责、咨询和知情 (RACI) 矩阵

定义参与迁移活动和云运营的所有各方的角色和责任的矩阵。矩阵名称源自矩阵中定义的责任类型：负责 (R)、问责 (A)、咨询 (C) 和知情 (I)。支持 (S) 类型是可选的。如果包括支持，则该矩阵称为 RASCI 矩阵，如果将其排除在外，则称为 RACI 矩阵。

响应性控制

一种安全控制，旨在推动对不良事件或偏离安全基线的情况进行修复。有关更多信息，请参阅在 AWS 上实施安全控制中的[响应性控制](#)。

保留

见 [7 R](#)。

退休

见 [7 R](#)。

检索增强生成 (RAG)

一种[生成式人工智能](#)技术，其中[法学硕士](#)在生成响应之前引用其训练数据源之外的权威数据源。

例如，RAG 模型可以对组织的知识库或自定义数据执行语义搜索。有关更多信息，请参阅[什么是 RAG](#)。

轮换

定期更新[密钥](#)以使攻击者更难访问凭据的过程。

行列访问控制 (RCAC)

使用已定义访问规则的基本、灵活的 SQL 表达式。RCAC 由行权限和列掩码组成。

RPO

参见[恢复点目标](#)。

RTO

参见[恢复时间目标](#)。

运行手册

执行特定任务所需的一套手动或自动程序。它们通常是为了简化重复性操作或高错误率的程序而设计的。

S

SAML 2.0

许多身份提供商 (IdPs) 使用的开放标准。此功能支持联合单点登录 (SSO)，因此用户无需在 IAM 中为组织中的所有人创建用户即可登录 AWS Management Console 或调用 AWS API 操作。有关基于 SAML 2.0 的联合身份验证的更多信息，请参阅 IAM 文档中的[关于基于 SAML 2.0 的联合身份验证](#)。

SCADA

参见[监督控制和数据采集](#)。

SCP

参见[服务控制政策](#)。

secret

在中 AWS Secrets Manager，您以加密形式存储的机密或受限信息，例如密码或用户凭证。它由密钥值及其元数据组成。密钥值可以是二进制、单个字符串或多个字符串。有关更多信息，请参阅 [Secrets Manager 密钥中有什么？](#) 在 Secrets Manager 文档中。

安全性源于设计

一种在整个开发过程中考虑安全性的系统工程方法。

安全控制

一种技术或管理防护机制，可防止、检测或降低威胁行为体利用安全漏洞的能力。安全控制主要有四种类型：[预防性](#)、[探测](#)、[响应式](#)和[主动式](#)。

安全加固

缩小攻击面，使其更能抵御攻击的过程。这可能包括删除不再需要的资源、实施授予最低权限的最佳安全实践或停用配置文件中不必要的功能等操作。

安全信息和事件管理 (SIEM) 系统

结合了安全信息管理 (SIM) 和安全事件管理 (SEM) 系统的工具和服务。SIEM 系统会收集、监控和分析来自服务器、网络、设备和其他来源的数据，以检测威胁和安全漏洞，并生成警报。

安全响应自动化

一种预定义和编程的操作，旨在自动响应或修复安全事件。这些自动化可作为[探测](#)或[响应式](#)安全控制措施，帮助您实施 AWS 安全最佳实践。自动响应操作的示例包括修改 VPC 安全组、修补 Amazon EC2 实例或轮换证书。

服务器端加密

由接收数据的人在目的地对数据 AWS 服务 进行加密。

服务控制策略 (SCP)

一种策略，用于集中控制组织中所有账户的权限 AWS Organizations。SCPs 定义防护措施或限制管理员可以委托给用户或角色的操作。您可以使用 SCPs 允许列表或拒绝列表来指定允许或禁止哪些服务或操作。有关更多信息，请参阅 AWS Organizations 文档中的[服务控制策略](#)。

服务端点

的入口点的 URL AWS 服务。您可以使用端点，通过编程方式连接到目标服务。有关更多信息，请参阅 AWS 一般参考 中的 [AWS 服务 端点](#)。

服务水平协议 (SLA)

一份协议，阐明了 IT 团队承诺向客户交付的内容，比如服务正常运行时间和性能。

服务级别指示器 (SLI)

对服务性能方面的衡量，例如其错误率、可用性或吞吐量。

服务级别目标 (SLO)

代表服务运行状况的目标指标，由服务[级别指标](#)衡量。

责任共担模式

描述您在云安全与合规方面共同承担 AWS 的责任的模型。AWS 负责云的安全，而您则负责云中的安全。有关更多信息，请参阅[责任共担模式](#)。

SIEM

参见[安全信息和事件管理系统](#)。

单点故障 (SPOF)

应用程序的单个关键组件出现故障，可能会中断系统。

SLA

参见[服务级别协议](#)。

SLI

参见[服务级别指标](#)。

SLO

参见[服务级别目标](#)。

split-and-seed 模型

一种扩展和加速现代化项目的模式。随着新功能和产品发布的定义，核心团队会拆分以创建新的产品团队。这有助于扩展组织的能力和服务，提高开发人员的工作效率，支持快速创新。有关更多信息，请参阅[中的分阶段实现应用程序现代化的方法。AWS Cloud](#)

恶作剧

参见[单点故障](#)。

星型架构

一种数据库组织结构，它使用一个大型事实表来存储交易数据或测量数据，并使用一个或多个较小的维度表来存储数据属性。此结构专为在[数据仓库](#)中使用或用于商业智能目的而设计。

strangler fig 模式

一种通过逐步重写和替换系统功能直至可以停用原有的系统来实现单体系统现代化的方法。这种模式用无花果藤作为类比，这种藤蔓成长为一棵树，最终战胜并取代了宿主。该模式是由 [Martin Fowler](#) 提出的，作为重写单体系统时管理风险的一种方法。有关如何应用此模式的示例，请参阅[使用容器和 Amazon API Gateway 逐步将原有的 Microsoft ASP.NET \(ASMX \) Web 服务现代化](#)。

子网

您的 VPC 内的一个 IP 地址范围。子网必须位于单个可用区中。

监控和数据采集 (SCADA)

在制造业中，一种使用硬件和软件来监控有形资产和生产操作的系统。

对称加密

一种加密算法，它使用相同的密钥来加密和解密数据。

综合测试

以模拟用户交互的方式测试系统，以检测潜在问题或监控性能。您可以使用 [Amazon S CloudWatch ynthetic](#) 来创建这些测试。

系统提示符

一种向[法学硕士提供上下文、说明或指导方针](#)以指导其行为的技术。系统提示有助于设置上下文并制定与用户交互的规则。

T

tags

键值对，用作组织资源的元数据。AWS 标签可帮助您管理、识别、组织、搜索和筛选资源。有关更多信息，请参阅[标记您的 AWS 资源](#)。

目标变量

您在监督式 ML 中尝试预测的值。这也被称为结果变量。例如，在制造环境中，目标变量可能是产品缺陷。

任务列表

一种通过运行手册用于跟踪进度的工具。任务列表包含运行手册的概述和要完成的常规任务列表。对于每项常规任务，它包括预计所需时间、所有者和进度。

测试环境

参见[环境](#)。

训练

为您的 ML 模型提供学习数据。训练数据必须包含正确答案。学习算法在训练数据中查找将输入数据属性映射到目标（您希望预测的答案）的模式。然后输出捕获这些模式的 ML 模型。然后，您可以使用 ML 模型对不知道目标的新数据进行预测。

中转网关

一个网络传输中心，可用于将您的网络 VPCs 和本地网络互连。有关更多信息，请参阅 AWS Transit Gateway 文档中的[什么是公交网关](#)。

基于中继的工作流程

一种方法，开发人员在功能分支中本地构建和测试功能，然后将这些更改合并到主分支中。然后，按顺序将主分支构建到开发、预生产和生产环境。

可信访问权限

向您指定的服务授予权限，该服务可以代表您在其账户中执行任务。AWS Organizations 当需要服务相关的角色时，受信任的服务会在每个账户中创建一个角色，为您执行管理任务。有关更多信息，请参阅 AWS Organizations 文档中的[AWS Organizations 与其他 AWS 服务一起使用](#)。

优化

更改训练过程的各个方面，以提高 ML 模型的准确性。例如，您可以通过生成标签集、添加标签，并在不同的设置下多次重复这些步骤来优化模型，从而训练 ML 模型。

双披萨团队

一个小 DevOps 团队，你可以用两个披萨来喂食。双披萨团队的规模可确保在软件开发过程中充分协作。

U

不确定性

这一概念指的是不精确、不完整或未知的信息，这些信息可能会破坏预测式 ML 模型的可靠性。不确定性有两种类型：认知不确定性是由有限的、不完整的数据造成的，而偶然不确定性是由数据中固有的噪声和随机性导致的。有关更多信息，请参阅[量化深度学习系统中的不确定性](#)指南。

无差别任务

也称为繁重工作，即创建和运行应用程序所必需的工作，但不能为最终用户提供直接价值或竞争优势。无差别任务的示例包括采购、维护和容量规划。

上层环境

参见[环境](#)。

V

vacuum 操作

一种数据库维护操作，包括在增量更新后进行清理，以回收存储空间并提高性能。

版本控制

跟踪更改的过程和工具，例如存储库中源代码的更改。

VPC 对等连接

两者之间的连接 VPCs，允许您使用私有 IP 地址路由流量。有关更多信息，请参阅 Amazon VPC 文档中的[什么是 VPC 对等连接](#)。

漏洞

损害系统安全的软件缺陷或硬件缺陷。

W

热缓存

一种包含经常访问的当前相关数据的缓冲区缓存。数据库实例可以从缓冲区缓存读取，这比从主内存或磁盘读取要快。

暖数据

不常访问的数据。查询此类数据时，通常可以接受中速查询。

窗口函数

一个 SQL 函数，用于对一组以某种方式与当前记录相关的行进行计算。窗口函数对于处理任务很有用，例如计算移动平均线或根据当前行的相对位置访问行的值。

工作负载

一系列资源和代码，它们可以提供商业价值，如面向客户的应用程序或后端过程。

工作流

迁移项目中负责一组特定任务的职能小组。每个工作流都是独立的，但支持项目中的其他工作流。例如，组合工作流负责确定应用程序的优先级、波次规划和收集迁移元数据。组合工作流将这些资产交付给迁移工作流，然后迁移服务器和应用程序。

蠕虫

参见[一次写入，多读](#)。

WQF

参见[AWS 工作负载资格框架](#)。

一次写入，多次读取 (WORM)

一种存储模型，它可以一次写入数据并防止数据被删除或修改。授权用户可以根据需要多次读取数据，但他们无法对其进行更改。这种数据存储基础架构被认为是[不可变的](#)。

Z

零日漏洞利用

一种利用未修补[漏洞](#)的攻击，通常是恶意软件。

零日漏洞

生产系统中不可避免的缺陷或漏洞。威胁主体可能利用这种类型的漏洞攻击系统。开发人员经常因攻击而意识到该漏洞。

零镜头提示

向[法学硕士](#)提供执行任务的说明，但没有示例（镜头）可以帮助指导任务。法学硕士必须使用其预先训练的知识来处理任务。零镜头提示的有效性取决于任务的复杂性和提示的质量。另请参阅[few-shot 提示](#)。

僵尸应用程序

平均 CPU 和内存使用率低于 5% 的应用程序。在迁移项目中，通常会停用这些应用程序。

本文属于机器翻译版本。若本译文内容与英语原文存在差异，则一律以英文原文为准。