代理人工智能的经济学 AWS

AWS 规范性指导

AWS 规范性指导: 代理人工智能的经济学 AWS

Table of Contents

代理人工智能的经济学 AWS

Hans Schabert 和 Prasanta Roy，Amazon Web Services

2026 年 1 月（文件历史记录）

采用人工智能驱动的自动化和代理人工智能系统的组织需要在人工和智能代理之间做出明智的经济决策。这对于可持续的云运营至关重要。本指南可帮助您评估、实施和优化人类劳动力和代理人工智能系统之间的经济权衡。 AWS您可以最大限度地提高投资回报率 (ROI)，同时保持卓越运营。

没有哪个系统是 100% 正确的。这一基本原理推动了对人类和代理人工智能系统的经济分析。组织必须超越简单的成本比较，来评估总体经济影响、风险状况、决策质量要求和长期战略价值创造。

客户行为正在从传统的前期技术投资急剧转变为使成本与业务结果保持一致的 pay-per-outcome模式。这种转变需要采用新的方法来评估、实施和优化人机协作。

成功之路遵循一个明确的模式：从合适的工作开始，衡量所有事情，然后扩展行之有效的方法。采用这种方法的组织通过智能资源分配和注重结果的自动化来实现可持续的竞争优势。

目标受众

本指南适用于以下内容：

- 正在做出战略投资决策的高管 (CEOs CTOs,, CFOs)
- 正在设计组织自动化策略的企业架构师
- 正在优化云财务管理的财务运营从业人员
- 正在评估 AI 实施方法的技术领导者
- 想要了解自动化投资回报率的业务部门负责人
- 正在探索全新 AI 定价模式的采购专业人员

要理解本指南中的概念，我们建议您查看 agentic AI 的基础。 AWS

目标

本指南可帮助您了解以下内容：

- 如何评估职位的代理自动化潜力

- 将人力成本与代理人工智能系统投资进行比较的经济模型
- Pay-per-outcome 定价模型及其对人工智能项目经济的影响
- 用于展示投资回报率和管理风险的衡量技术
- 将固定成本转化为可变结果的扩展策略

关于此内容系列

本指南是关于代理人工智能的系列文章的一部分。 AWS要了解更多信息并查看本系列中的其他指南，请参阅 AWS 规范性指导网站上的 Agentic AI。

了解代理人工智能经济学 AWS

关键原则之一是确定何时使用 AI 代理以及何时使用传统的确定性方法。组织必须系统地评估哪些工作需要机构自动化，哪些工作应该使用传统的自动化或持续的人工操作。做出这一决定需要了解任务特征、风险承受能力和操作方法之间的关系。

在决定实施代理人工智能之前，您应该使用决策框架来了解经济影响。决策框架包括以下三个关键问题：

1. 任务评估 — 此任务是否适合 AI 代理？
2. 风险影响评估 — 涉及哪些风险？
3. 投资回报 — 是否具有成本效益？

任务评估

具有高度复杂性、标准化决策规则的任务可以从代理人工智能方法中受益。传统的自动化或机器人流程自动化可以更好地完成高度标准化、简单的任务。Agentic AI 系统擅长推理、理解上下文或适应性地做出决策，它们在基于规则的处理之外还能增加价值。成功的代理人工智能实施需要能够学习和适应的系统。

评估任务时，请考虑以下因素：

- 复杂性 — 所需的推理程度和上下文理解。与传统的自动化相比，需要情境理解、细致入微的解释或对不断变化的条件做出适应性反应的任务更倾向于代理方法，而纯粹的机械或计算任务可能不需要代理智能。
- 标准化 — 存在明确的模式和规则。如果任务需要情境理解，则建议使用 Agentic AI。如果不需要适应或学习，可以考虑传统的自动化。
- 音量-任务执行频率。建议将 Agentic AI 用于自主活动。对于高容量、一致的任务，建议使用传统的自动化。但是，光靠音量并不能决定方法。少量、高价值的决策可能证明机构为提高决策质量而不是降低成本提供援助是合理的。
- 价值-每次任务完成的业务影响。考虑使用代理人工智能来实现需要类似人类的自主能力的高价值成果。对于重复的、一致的任务，可以考虑使用传统的自动化，这些任务可以通过确定性方式完成。

风险影响评估

目前有四种代理人工智能部署方法：完全自主、人类在环、副驾驶或在代理支持下由人主导。每个人都有自己的风险状况和容错能力，它们都以某种身份涉及人类。下表描述了这些方法的风险细节。

自治级别	风险概况	容错能力	用例示例	人类参与
完全自主	低风险	1-2% 可以接受	• 基本数据分类 • 文档路由 • 标准报告生成	• 尽量减少监督 • 定期审计
Human 在圈子里	中等风险	低于 0.5%	• 答复草稿 • 内容审核 • 初步索赔处理	• 定期审查 • 异常处理 • 质量保证
副驾驶	高风险	接近零	• 战略规划投入 • 风险评测 • 投资决策	• 人类做出最终决定 • 代理提供建议
人为主导，支持代理	严重风险	零容忍	• 法律决定 • 医学诊断 • 监管合规	• 人为驱动流程 • Agent 仅提供研究或分析以及支持信息

下表描述了在这些方法之间进行选择时的关键注意事项。

考虑	完全自主	Human 在圈子里	副驾驶	人为主导

成本效益	最高	高	中	低
可扩展性	无限制	高	中	有限
处理速度	最快	快速	中	慢
风险管理	Basic	增强	很强	最强
复杂性处理	简单的任务	中等复杂的任务	复杂任务	关键任务

该考虑框架可帮助组织将自主权级别与风险状况相匹配，适当扩展运营，在效率与控制之间取得平衡，实施适当的治理并优化资源分配。

投资回报率

计算代理人工智能系统的投资回报率首先要进行全面的成本分析。组织必须首先计算其当前的人力成本，包括工资、福利和工作场所费用，以及特定流程的费用和隐性成本，例如培训、保险和停机时间。

为了进行收支平衡分析，组织应考虑实施成本、持续运营费用以及证明投资合理性所需的数量。同样重要的是要考虑季节性变化以及随着时间的推移系统成熟和改进而出现的学习曲线收益。

在评估人工智能代理时，组织应记住，与人工操作相比，这些系统的前期成本通常更高，但每笔交易的成本更低。此外，与人工团队相比，AI 代理的性能会随着时间的推移而提高，并且可扩展性更好。随着部署规模和运营经验的积累，这使得它们的成本效益越来越高。

衡量代理人工智能系统的成功和投资回报率

衡量代理人工智能系统实施的成功需要采用系统的方法。本节为评估和持续优化提供了一种清晰的方法，该方法使用您现有的分析，而不是从头开始。

第 1 步：使用你现有的粉底

首先根据评估当前流程成本部分中的建议进行全面的成本评估。这为您的 ROI 计算提供了运营基准。如风险影响评估部分所述，在四个自主级别（完全自主、人工在环、副驾驶方法、由代理支持的人为主导）之间进行选择，以便为每个过程确定适当的测量标准和容错阈值。

第 2 步：设定明确的成功目标

制定架构和成功目标，强调具有学习能力的系统，如实施代理人工智能系统的成功模式部分所述。专注于持续改进，而不是静态性能。使用案例研究：比较招聘运营的人工和代理人工智能成本中演示的收支平衡分析方法来设定投资回报率时间表。包括终止不良代理的明确决策要点。

第 3 步：跟踪关键指标

根据既定基准监控财务业绩，并跟踪成本节省和战略价值的改善。衡量运营指标，包括所选自治级别的可接受阈值内的错误率、处理速度的提高和一致性的提高。将重点放在能够证明学习能力和随时间推移适应能力的战略指标上。

第 4 步：使用 AgentOps

应用 “将人类反馈纳入代理人工智能系统” 部分中的持续学习框架，通过系统的人类反馈集成来优化决策。创建包含人类洞察力的实时学习系统，以提高绩效。监控向基于结果的商业模式的转型，如代理人工智能系统的经济转型到基于结果的定价中所述。 AWS

评估您当前的人工流程成本

了解您的真实流程成本是就代理人工智能系统投资做出明智决策的基础。首先，您必须为当前的流程成本建立准确的基准，包括所有隐性费用、失败率和机会成本。这可以帮助您进行精确的投资回报率计算并做出战略决策。这种全面的成本评估是评估代理人工智能系统能否作为高效伴侣提供真正价值的关键基础。

基准成本评估之所以必不可少，主要原因如下：

- 投资回报率的准确性 — 准确的成本基准支持真实的投资回报率预测，这些预测考虑了当前运营支出的各个方面。
- 代理实施策略 — 全面的成本了解可帮助组织确定最有前途的初始代理人工智能系统部署流程。
- 绩效衡量 — 既定的基准提供了衡量框架，用于跟踪代理人工智能实施带来的实际和预计收益。

在比较人工和代理替代方案之前，组织必须系统地识别和评估影响流程经济性的所有成本因素。该评估通过考虑显而易见和隐性的成本驱动因素，确保了准确的基准计算。它特别强调故障成本、历史失败率和错失的商机，这些都是当前流程的真实总成本。

本节介绍如何收集每个成本类别的数据，以便为当前流程建立准确的基准衡量标准。它讨论了信息来源，并提供了以下成本类别的示例：

- 劳动力成本
- 人员绩效和一致性成本
- 技术和基础设施成本
- 失去的商机成本
- 风险和缺陷成本

劳动力成本

提取 24 个月的工资数据，包括基本工资、加班费、福利和培训成本。使用您的人力资源信息系统 (HRIS) 来跟踪招聘费用和人员流失率。时间跟踪系统显示实际生产率与计划工时的比较。绩效管理平台显示技能水平与薪酬成本之间的相关性。计算分配给管理开销的满负荷小时费率。

以下是劳动力成本驱动因素的示例列表。

成本驱动因素	业务影响
基本补偿	满载时速每小时 25—150 美元
福利税和工资税	基本工资的 25-40%
培训和发展	每年劳动力成本的 5-15%
管理开销	直接人工成本的 15-25%

人员绩效和一致性成本

将显示任务完成情况变化的项目管理系统中的数据与考勤系统相结合。这可以揭示缺勤模式和季节性变化。客户服务平台通过解决方案指标展示个人绩效范围，销售客户关系管理 (CRM) 数据可以显示交易完成过程中的效率差异。质量管理系统提供跨团队和地点的缺陷率和流程合规数据。工作流程系统会捕获完成时间、批准延迟和异常处理频率。沟通分析通过会议频率和协作模式揭示了协调开销。

以下是人类绩效和一致性成本驱动因素的示例列表。

成本驱动因素	业务影响
生产率波动	20— 50% 的性能范围
缺勤和保险	需要增加 15— 25% 的容量
疲劳和动机周期	10— 30% 的生产率差异
程序不一致	10— 40% 的效率损失
质量控制差异	占总成本的 10-30%
协调开销	运营成本的 15-25%

技术和基础设施成本

许可证管理平台显示软件成本和利用率。基础设施监控提供正常运行时间数据、性能指标和维护成本。服务台系统跟踪支持开销和反复出现的技术问题。供应商管理系统捕获技术关系总成本，包括集成费用和服务级别绩效。

以下是技术和基础设施成本驱动因素的示例列表。

成本驱动因素	业务影响
技术系统	每位用户每月 50-500 美元
工作空间和设备	每位员工每月 200—1,000 美元

失去的商机成本

CRM 平台包含潜在客户响应时间、转化率和机会损失文档。营销自动化显示后续延迟会对潜在客户转化产生影响。客户支持系统揭示了运营问题如何影响满意度和留存率。竞争分析提供了将运营绩效与收入结果联系起来的市场响应要求和赢亏数据。

以下是失去商机的成本驱动因素的示例列表。

成本驱动因素	业务影响
市场反应延迟	每延迟一天的收入
容量限制	失去的商机
创新资源分配	日常工作的机会成本
客户获取延迟	反应缓慢导致的铅流失 50-90%

风险和缺陷成本

保险单文件显示了一般责任、职业责任、工伤赔偿和网络责任保险的费用。内部风险评估报告确定了操作漏洞和相关的缓解成本。缺陷跟踪系统记录产品或服务故障，包括检测成本、更换费用和保修索赔。资产更换计划显示设备故障率和更换成本。安全事故报告跟踪工作场所事故和相关的工伤赔偿索赔。业务连续性计划详细说明了备份系统成本和灾难恢复投资。

以下是风险和缺陷的成本驱动因素的示例列表。

成本驱动因素	业务影响
保险费用	运营预算的 1-5%
错误的代价	每起错误事件 50-5,000 美元
人为错误影响	占总运营成本的 2-15%
错误率和返工	原校正费用的 1.5—4 倍

在上实施代理人工智能系统的成功模式 AWS

企业人工智能采用现状（ISG 2025 报告）显示，成功实施人工智能的主要障碍不是技术能力，而是学习差距。该术语是指无法适应、记住背景或无法随着时间的推移而改进的系统。实施静态 AI 工具的组织的失败率很高。以下是取得成功的代理人工智能系统的共同特征：

- 情境记忆 — 保留对话历史记录和用户偏好的系统
- 反馈整合 — 能够从更正中吸取教训并提高绩效
- 工作流程调整 — 自动调整以适应不断变化的业务需求
- 持续改进 — 通过运营经验实现可衡量的改进

成功实现人工智能实施的组织通常会优先考虑以下几点：

- 使用全面的合作伙伴生态系统，而不是独立构建和探索人工智能能力
- 与静态工具相比，具有学习能力的系统
- 将业务结果重点放在技术功能比较上
- 工作流程集成而不是独立工具
- 持续调整而不是一次性实施

这些模式与许多 AWS 服务 功能一致，尤其是 Amazon Bedrock 中的基础模型访问、中的事件驱动架构以及通过 Amazon 提供的全面监控。AWS Lambda CloudWatch有关集成人工反馈和支持学习的系统的更多信息，请参阅本指南中的将人类反馈整合到代理人工智能系统中。

将人工反馈整合到代理人工智能系统中

没有哪个系统能百分之百成功，失败必然会发生。每一次失败，都会产生相关的变革成本。Human in the Loop 是一种人工智能方法，在这种方法中，人工智能执行任务，但需要人工干预或批准。当失败的成本高于 human-in-the-loop解决方案的成本时，必须使用这种方法。

代理人工智能系统的成功从根本上取决于代理通过人工反馈进行学习和改进的能力。必须考虑人力成本，这取决于所需努力的程度。与执行预定规则的静态自动化工具不同， human-in-the-loop解决方案具有具有学习能力的代理系统，可以在自主代理和人类之间建立动态的伙伴关系。在代理大规模处理日常处理的同时，人类的专业知识不断提高试剂的性能。这种协作方法将 AI 实施从一次性部署转变为持续的优化过程。该系统适应组织模式，内部化质量标准，并根据现实世界的运营经验完善其决策能力。通过系统地捕获人工更正、批准和见解，组织可以构建能够理解背景、识别模式并随着时间的推移越来越多地与业务目标保持一致的人工智能代理。

对于不需要人工干预或支持的解决方案，无需将人为成本考虑在代理经济学中。

向人类操作员学习行为

人类操作员会提供关键反馈，代理人工智能系统可以利用这些反馈来学习、适应和改善他们的反应。这种反馈回路创建了一个协作环境，在这种环境中，人类的专业知识可以增强代理能力，而代理则可以处理例行处理。

通过人类行为模式识别，代理可以从人际交互模式中学习，以反映成功的沟通方法。这有助于他们适应组织决策模式和风险承受能力。系统通过人工更正和批准将质量预期内在化。他们还可以学习针对不同客户群和业务环境的适当回应。

有效的反馈收集机制可以系统地捕获人工编辑和对代理响应的修改。他们分析了人工审阅者在代理推荐中批准、拒绝或修改的内容。通过了解为什么某些案例需要人工干预，并纳入对不同场景和复杂程度下代理表现的人工评估，这些系统不断完善其能力，以更紧密地与组织标准和期望保持一致。

持续学习操作

实时学习集成使代理人工智能系统能够整合人工反馈，并通过动态模型更新立即改善代理的响应。这些系统使用人类洞察力来识别新的模式和边缘案例。这增强了他们的模式识别能力，同时通过人工指导的学习体验来建立组织记忆。基于人工操作员反馈和业务结果的持续改进推动了持续的绩效优化。

以人为本的培训可以捕捉专业知识，以增强代理决策能力。它将经验丰富的操作员的关键专业知识转移到人工智能系统。通过基于场景的学习，系统使用人为创建的示例来改善对复杂情况的处理。他们还通过质量校准使代理商的性能标准与人类的质量期望保持一致。这种方法融合了人类对组织文化和客户期望的见解。这种文化适应可以帮助代理人在不同的背景下做出适当的反应。

通过人机协作，实现卓越运营

自动风险感知优化可以持续评估操作条件和错误概率，并对高风险场景进行人工监督。这有助于系统从人类风险评估中吸取教训并改进 future 的决策。Amazon Bedrock 提供对具有不同功能和成本状况的多种基础模型的访问权限。这可以实现智能路由，既考虑成本又考虑风险状况，同时纳入人工反馈以优化模型选择。性能调整通过整合人工对质量标准的反馈和可接受的性能权衡来平衡效率和错误率最小化。自动决策考虑了经风险调整后的总拥有成本。运营商提供有关组织风险承受能力和业务优先权衡的指导。这可以帮助您优化成本，同时与组织目标保持一致。

人工增强型学习系统根据错误影响和业务后果来优先考虑人工输入。这创建了通过风险加权反馈来了解技术准确性和业务背景的学习系统。定期的绩效分析包括风险指标和错误成本分析，而人工洞察提供了自动化系统无法捕捉的背景信息。最佳实践开发通过将自动模式识别与人类专业知识和判断相结合，强

调风险管理和错误预防。通过培训计划进行组织能力建设，既可以培养管理代理人工智能系统的人类技能，也可以培养支持人类决策的代理能力。这确保了人机与人工智能合作的全面方法，从而加强了伙伴关系的两个组成部分。

经济转型，为代理人工智能系统提供基于结果的定价 AWS

从传统的固定成本模式向基于结果的定价模式的转变代表了组织如何构建经济运营和管理风险的根本性转变。这种转型是现有流程持续现代化的途径，同时为代理人工智能转型提供资金。它使组织能够从静态的资源密集型运营演变为动态的、以结果为导向的商业模式。

传统、前期模式

部门通常作为成本中心运营，直接的人力成本由成本分配融资。Organizations 通常希望减少这种成本分配。如果流程不现代化，该部门必须以较少的员工队伍实现同样的结果。这通常会降低质量。传统的商业模式带来了重大挑战，包括：

- 成本随着数量的增加而线性扩展 — 这要求组织雇用更多员工来处理增加的业务量。
- 固定成本承诺 — 无论业务绩效和流程效率如何，这些承诺都将持续下去。
- 高级规划 — 在经济衰退和产能限制期间，灵活性有限，需要提前规划。
- 质量下降周期 — 如果在不改进流程的情况下削减成本，则预算减少会导致服务质量降低。

基于结果的模型

基于结果的现代模型将付款与可衡量的业务结果直接挂钩，例如成功完成的员工、实现的质量指标、流程效率的提高或实现的生产率提高。这从根本上将财务风险从业务部门转移到服务提供商，同时形成了自然的激励调整。 以下是基于结果的模型的主要优点：

- 成本直接随着业务价值的产生而扩展
- 运营费用和收入之间的自然一致
- 可根据市场情况灵活调整容量
- Pay-per-success 模型通过将财务风险从前期投资转移到持续运营绩效来降低财务风险
- 专注于能够随着时间的推移而不断改进的具有学习能力的系统，而不是静态的替代方案

这种转型远远超出了内部成本中心，从根本上重塑了组织与外部合作伙伴和服务提供商的互动方式。通过将基于结果的定价应用于合作伙伴协作，组织可以推动长期的质量改进并降低成本，同时间接强调代理人工智能现代化。

Organizations 可以快速进行实验，清晰地衡量绩效，并根据产生的实际业务价值进行扩展，而不是传统的固定资源承诺。这种方法可以实现以下功能：

- 供应商关系的演变 — 合作伙伴投资于客户成功，而不仅仅是服务交付。
- 标准化结果指标 — 简化多个供应商的采购流程。
- 市场响应能力 — 快速适应不断变化的市场条件和客户需求。
- 竞争优势 — 卓越的资源利用率和增强的运营能力。
- 以质量为导向的合作伙伴关系 — 长期合作侧重于持续改进和可衡量的结果。

AWS Marketplace 用作 pay-per-outcome启用码

这种转型的关键推动因素是 AWS Marketplace，它可以作为代理工作和基于结果的定价的交易工具。它允许访问数百个预先构建的 AI 代理和代理解决方案，这些代理和代理解决方案具有透明、基于使用量的定价模型。它可以帮助消除前期许可成本，降低实施复杂性，并使组织能够专注于能够随着时间的推移而调整和改进的具有学习能力的系统，而不是静态的替代方案

使用 AWS Marketplace 可以提供以下好处：

- 快速实验 — 无需大量资本投资即可测试多种解决方案
- 透明定价 — 基于使用量的成本，明确归因于业务成果
- 久经考验的解决方案 — 从经验丰富的提供商那里获得久经考验的代理
- 内置集成 — 与现有设备无缝连接 AWS 服务
- 风险缓解 — 能够根据性能切换提供商
- 学习能力访问权限 — 自适应系统的可用性，无需内部开发成本

这种方法使组织能够根据结果交付和学习能力（而不是功能列表）比较多个选项。它还可以帮助您建立明确的成功标准和衡量方法，并协商与业务结果和系统改进相关的基于结果的定价。通过基于结果的模型为代理人工智能转型提供资金，组织可以持续实现流程现代化，同时只需为可衡量的改进和成功的成果付费。

案例研究：比较招聘运营的人工和代理人工智能成本

招聘运营为评估人类和代理人工智能系统之间的经济权衡提供了一个令人信服的案例研究，但是投资回报率的计算在很大程度上取决于您当前的运营基准。评估代理人工智能投资的组织经常会问一个基本问题：“如果我们只是优化现有的人工流程会怎样？”为了直接解决这个问题，本分析提出了两种不同的情景，它们涵盖了人类运营效率的范围。

方案 A 模拟了 45 分钟的简历 (CV) 或简历筛选时间。场景 B 演示了优化的人工操作，每个应用程序 15 分钟，效率提高了 66%。例如，这种改进可以通过简化的流程、经验丰富的招聘人员或专业工具来实现。

通过将相同的代理系统功能与这些不同的人工绩效基准进行比较，我们揭示了现有流程效率如何影响投资回报率计算、收支平衡时间表和战略实施决策。这种双场景方法有多种用途。它假设仅靠流程优化就足够了，从而防止组织忽略代理人工智能。它还可以帮助流程已经很高效的组织了解其具体的经济状况。此外，这些场景凸显了非财务优势（例如全天候可用性和可扩展性）何时成为主要的决策因素。了解不同效率基准下的这些经济动态，使组织能够就何时何地部署代理人工智能系统做出明智的决定，以最大限度地提高业务影响。

场景 A：放映时间 45 分钟

方案 A 代表招聘业务，其中人工招聘人员花费 45 分钟筛选每份简历。此情景模拟了一名中层招聘人员，其年度满负荷费用为112,250美元。该招聘人员在标准工作时间内处理具有典型人员绩效特征的申请。相比之下，代理人工智能系统需要23,000美元的初始投资用于开发、定制和ATS集成，而云基础架构的最低每月运营成本为500美元。该代理在短短 5 分钟内即可处理应用程序，实现了 2% 的错误率和超过 8,600 个应用程序的月容量。这是一个巨大的效率差距，即代理每个应用程序的运行速度提高了 9 倍，每月的容量提高了 39 倍。本节探讨运营前六个月的成本结构分析、运营指标、基于数量的比较以及累积投资回报率的计算。

基本成本结构

下表显示了方案 A 的初始设置成本。

组件	人工操作	Agentic AI 系统
代理开发和定制	不适用	15,000 美元
申请人跟踪系统 (ATS) 集成	不适用	5,000 美元

训练和优化	不适用	3,000
初始设置总成本	0 美元	23,000 美元

下表显示了方案 A 的年度固定成本。

组件	人工操作	Agentic AI 系统
基本工资	65,000	不适用
福利 (30%)	19,500 美元	不适用
工作空间和设备	12,000 美元	不适用
管理监督 (15%)	9,750 美元	不适用
培训和发展	6,000	不适用
年度固定成本总额	112,250 美元	不适用

下表显示了方案 A 的每月运营成本。

组件	人工操作	Agentic AI 系统
云计算	不适用	$200
仓储服务	不适用	100 USD
数据库操作	不适用	100 USD
监控	不适用	100 USD
每月固定成本总额	9,354 美元	500 美元

运行指标

下表显示了方案 A 的运行指标。

指标	人工操作	Agentic AI 系统
每个申请的处理时间	45 minutes	5 分钟
每小时容量	1.33 个应用程序	12 个应用程序
每日容量（24 小时）	10-11 个应用程序	288 个应用程序
每月容量	220 个应用程序	8,640 个应用程序
每个应用程序的费用	45 美元	2.50 美元
每次成功招聘的费用	2,200 美元	$125
错误率	5%	2%
纠错成本	每个错误 90 美元	每次上报 45 美元

基于数量的成本分析

下表显示了方案 A 的基于数量的成本分析。在此示例中，代理人工智能系统成本包括固定成本和在 12 个月内每月 1,917 美元的摊销设置成本。

每月交易量	人力成本	Agentic AI 系统成本	每月储蓄
100 个应用程序	4,500 美元	750 美元	3,750 美元
500 个应用程序	22,500 美元	2,667 美元	19,833 美元
1,000 个应用程序	45,000 美元	4,917 美元	40,083 美元

投资回报率分析

下表显示了方案 A 的投资回报率分析，该分析基于每月处理 500 个应用程序。

指标	值
每月的人力成本	22,500 美元

每月代理费用	2,667 美元
每月储蓄	19,833 美元
每年节省费用	237,996 美元
收支平衡期	1.16 个月

累积成本比较

下表显示了方案 A 在前六个月的累积成本比较，假设每月有 500 份申请。

月份	人力成本	Agentic AI 系统成本	累积节省
1	22,500 美元	25,667 美元	-3,167 美元
2	45,000 美元	28,334 美元	16,666 美元
3	67,500 美元	31,001 美元	36,499 美元
4	90,000 美元	33,668 美元	56,332 美元
5	112,500 美元	36,335 美元	76,165 美元
6	135,000 美元	39,002 美元	95,998 美元

代理人工智能系统的其他好处

以下是场景 A 中代理人工智能系统提供的其他好处：

- 可扩展性 — 无需额外成本即可应对音量峰值
- 可用性 — 全天候运行，可即时响应
- 一致性 — 采用统一的筛选标准
- 时间效率 — 显著降低 time-to-hire
- 用户体验 — 即时向候选人提供反馈

方案 B：放映时间 15 分钟

情景 B 模型优化了招聘操作，其中人工招聘人员将筛选流程简化为每份申请 15 分钟。这比方案A的效率提高了66%。这种情况保持了中层招聘人员每年112,250美元的满负荷成本。但是，它显著提高了人类生产力，在 8 小时轮班期间，日容量增加到 32 个应用程序，每月吞吐量达到 660 个应用程序。提高的人力效率使每个应用程序的成本从45美元降低到15美元，从而缩小了与代理人工智能系统的经济差距。但是，该代理保持了其结构优势：5分钟的处理时间，24/7的可用性，支持288个日常应用程序，与人为5%相比，错误率降低了2%，每月容量超过8,600个应用程序。尽管这种效率提高将收支平衡期从1.16个月延长至4.76个月，并将每月节省的费用从19,833美元减少到4,833美元，但分析表明，即使与高度优化的人工操作竞争，代理系统在经济上仍然可行，这是评估其当前流程效率水平是否值得进行代理人工智能投资的组织的重要见解。

基本成本结构

下表显示了情景B的年度固定成本。

组件	人工操作	Agentic AI 系统
基本工资	65,000	不适用
福利 (30%)	19,500 美元	不适用
工作空间和设备	12,000 美元	不适用
管理监督 (15%)	9,750 美元	不适用
培训与发展	6,000	不适用
年度固定成本总额	112,250 美元	不适用

下表显示了方案 B 的实施成本。

组件	人工操作	Agentic AI 系统
初始 设置	不适用	23,000 美元
每月固定成本	9,354 美元	500 美元

运行指标

下表显示了方案 B 的运营指标。

指标	人工操作	Agentic AI 系统
每个申请的处理时间	15 分钟	5 分钟
每小时容量	4 个应用程序	12 个应用程序
每日容量（8 小时轮班）	32 个应用程序	288 个应用程序
每月容量	660 个应用程序	8,640 个应用程序
每个应用程序的费用	15 美元	2.50 美元
每次成功招聘的费用	2,200 美元	$125
错误率	5%	2%
纠错成本	每个错误 30 美元	每次上报 45 美元

基于数量的成本分析

下表显示了方案 B 的基于数量的成本分析。在此示例中，代理人工智能系统成本包括固定成本和在 12 个月内每月 1,917 美元的摊销设置成本。

每月交易量	人力成本	Agentic AI 系统成本	每月储蓄
100 个应用程序	1,500 美元	750 美元	750 美元
500 个应用程序	7,500 美元	2,667 美元	4,833 美元
1,000 个应用程序	15,000 美元	4,917 美元	10,083 美元

投资回报率分析

下表显示了方案 B 的投资回报率分析，该分析基于每月处理 500 个应用程序。

指标	值
每月的人力成本	7,500 美元
每月代理人工智能系统成本	2,667 美元
每月储蓄	4,833 美元
每年节省费用	57,996 美元
收支平衡期	4.76 个月

累积成本比较

下表显示了方案B在前六个月的累积成本比较，假设每月有500份申请。

月份	人力成本	Agentic AI 系统成本	累积节省
1	7,500 美元	25,667 美元	-18,167 美元
2	15,000 美元	28,334 美元	-13,334 美元
3	22,500 美元	31,001 美元	-8,501 美元
4	30,000 美元	33,668 美元	-3,668 美元
5	37,500 美元	36,335 美元	1,165 美元
6	45,000 美元	39,002 美元	5,998 美元

比较每种情景的成本和收益

指标	场景 A	场景 B	Impact
放映时间	45 minutes	15 分钟	提高了 66%
每日容量	10—11 个应用程序	32 个应用程序	增长 200%

每个应用程序的费用	45 美元	15 美元	减少 66%
每月节省开支（500 个应用程序）	19,833 美元	4,833 美元	减少了 76%
收支平衡期	1.16 个月	4.76 个月	延长 310%

方案 B 展示了人工操作的效率显著提高，处理时间的缩短可在不增加人员的情况下增加容量，并大大降低每个应用程序的成本。但是，财务影响揭示了一幅更加细致入微的画面：尽管投资回报率仍然为正，但与情景A相比，组织面临着更长的收支平衡期和每月节省的减少。这些结果凸显了实施的关键决策因素——即使在优化的人工运营下，代理系统仍然具有财务可行性，但是组织在评估部署时间表和预期回报时必须采取长期投资视角，仔细考虑数量波动和可扩展性需求。

但是，代理人工智能系统仍然保持着关键的运营优势，这些优势不仅限于纯粹的成本节约。无论时区或工作时间如何，它都提供全天候可用性，便于候选人即时参与。它通过对每种应用应用统一的标准来提供稳定的筛查质量，在不产生额外成本的情况下进行缩放以应对体积峰值。它可以立即为候选人提供响应，从而增强雇主品牌和候选人体验，并且其运行疲劳系数为零，可确保第一份申请与第一千份申请具有相同的高质量表现。

人为错误通常是由疲劳、分心或知识差距引起的，通常涉及沟通不畅或信息不正确。Agentic AI 系统错误通常源于边缘情况、输入不明确或训练数据限制。这些错误本质上往往更加一致。

质量和体验指标揭示了人员和代理能力之间的明显权衡：

- 客户满意度 — 人类在同理心和复杂问题解决方面表现出色，客服人员为例行查询提供一致、准确的信息。
- 响应时间 — 响应时间有利于客服人员随时待命。人工提供营业时间支持，但可能出现排队延迟。
- 一致性-代理对相似的查询提供相同的响应。人类在方法和知识应用上可能有所不同。
- 升级处理 — 需要判断力、创造力或情商的复杂问题仍然是人类的优势。

结论和资源

与代理人工智能系统相比，人类系统的经济学所代表的不仅仅是一个技术决策。它反映了组织如何创造价值、管理风险和实现竞争优势的根本性转变。成功需要系统地评估工作特征，全面衡量结果（包括风险因素），并根据已证实的结果进行战略扩展。

企业人工智能采用现状（ISG 2025 报告）显示，大多数人工智能实现失败的原因是学习差距，即系统无法适应、记住背景或随着时间的推移而改进。取得成功的组织侧重于具有学习能力的系统，这些系统可以深度集成到工作流程中，并通过人工反馈和运营经验展示持续改进。

了解这些原则的组织——从合适的工作开始，将工作分解为任务，衡量包括风险影响在内的所有内容，以及扩展行之有效的方法——将通过最佳的资源利用率和注重结果的自动化来实现可持续的竞争优势，这种自动化会随着业务的成功而发展。

future 属于能够智能地将人类专业知识与代理人工智能能力相结合的组织。这创建了混合模型，既能提供卓越的成果，又能保持动态市场条件所需的灵活性、学习能力和协作优势。

资源

以下资源可以帮助你规划、设计和实施代理人工智能系统： AWS

- 为代理人工智能构建无服务器架构 AWS（AWS 规范性指南）
- 开@@ 启代理人工智能 AWS（AWS 规范性指导）
- A@@ gentic AI 模式和工作流程开启 AWS（AWS 规范性指导）
- Agentic AI（AWS 规范性指导）
- AWS 成本优化中心 (AWS 服务)
- 亚马逊 Bedrock 文档 ()AWS 服务
- 成本优化支柱（Well-Architect AWS ed Framework）
- AI 代理和解决方案 (AWS Marketplace)

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	—	2026 年 1 月 28 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：将本地 Oracle 数据库迁移到 Amazon Aurora PostgreSQL 兼容版。
- 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：将本地 Oracle 数据库迁移到 AWS 云中的 Amazon Relational Database Service（Amazon RDS）for Oracle。
- 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：将客户关系管理（CRM）系统迁移到 Salesforce.com。
- 重新托管（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：将本地 Oracle 数据库迁移到 AWS 云中 EC2 实例上的 Oracle。
- 重新放置（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。您将服务器从本地平台迁移到同一平台的云服务中。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

请参阅托管服务。

ACID

请参阅原子性、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。它比主动-被动迁移更灵活，但工作量更大。

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

一种 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

请参阅人工智能。

AIOps

请参阅人工智能运营。

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。

权威数据来源

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来源复制到其他位置，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

恶意机器人

一种旨在扰乱或伤害个人或组织的机器人。

BCP

请参阅业务连续性计划。

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。另请参阅字节顺序。

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应用程序版本（蓝色），在另一个环境中运行新应用程序版本（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交互的软件应用程序。有些机器人是有用或有益的，例如在互联网上索引信息的 Web 爬网程序。还有一些被称为恶意机器人的机器人，其目的是扰乱或伤害个人或组织。

僵尸网络

被恶意软件感染并受单方（称为僵尸网络控制者或僵尸网络操作者）控制的僵尸网络。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）。

紧急（break-glass）访问

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 AWS Well-Architected Guidance 中的 Implement break-glass procedures 指示器。

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

请参阅 AWS 云采用框架。

金丝雀部署

缓慢而渐进地向最终用户发布版本。当您确信无误后，即可部署新版本，并完全替换当前版本。

CCoE

请参阅云卓越中心。

CDC

请参阅更改数据捕获。

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韧性。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

请参阅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 云 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。云计算通常连接到边缘计算技术。

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 AWS 云中时通常会经历四个阶段：

- 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，创建着陆区、定义 CCo E、建立运营模型）
- 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 云 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

请参阅配置管理数据库。

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 管线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。

配置偏移

对于工作负载而言，一种偏离预期状态的配置更改。这可能会导致工作负载变得不合规，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

请参阅计算机视觉。

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-Architecte 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 云 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个人。

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

请参阅数据库定义语言。

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。

委派管理员

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。

部署

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

请参阅环境。

侦测性控制

一种安全控制，在事件发生后进行检测、记录日志和发出提醒。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

星型架构中的一种较小的表，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）

您用来最大程度地减少由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 Fr ame AWS work 中的 “工作负载灾难恢复：云端 AWS 恢复”。

DML

请参阅数据库操作语言。

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

请参阅灾难恢复。

偏差检测

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

请参阅开发价值流映射。

E

EDA

请参阅探索性数据分析。

EDI

请参阅电子数据交换。

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。与云计算比较时，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）

组织之间业务文件的自动交换。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加密文字的计算流程。

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

请参阅服务端点。

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

请参阅企业资源规划。

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

星型架构中的中心表。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。

快速失效机制

一种使用频繁且增量式的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 云，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

请参阅分支。

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。

少样本提示

在要求 LLM 执行类似任务之前，先向其提供少量示例，以演示任务和预期输出。此技术是上下文内学习的一种应用，其中模型可以从提示中嵌入的示例（样本）中学习。对于需要特定格式、推理或领域知识的任务，少样本提示可能非常有效。另请参阅零样本提示。

FGAC

请参阅精细访问控制。

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，通过更改数据捕获使用连续数据复制，在极短的时间内迁移数据，而非使用分阶段方法。目标是将停机时间降至最低。

FM

请参阅基础模型。

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

AI 模型的一个子集，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构件，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人工智能。

地理阻止

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工作流程，而基于中继的工作流程则是现代的、首选的方法。

黄金映像

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预调配软件，并有助于提高设备制造操作的速度、可扩展性和生产效率。

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

帮助管理各组织单位的资源、策略和合规性的高级规则 (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。侦测性护栏会检测策略违规和合规性问题，并生成提醒以进行修复。它们通过使用 AWS Config、、Amazon、 AWS Security Hub CSPM GuardDuty AWS Trusted Advisor、Amazon Inspector 和自定义 AWS Lambda 支票来实现。

H

HA

请参阅高可用性。

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。

保留数据

从用于训练机器学习模型的数据集中保留的一部分标注的历史数据。通过将模型预测与保留数据进行比较，您可以使用保留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。

hypercare 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 hypercare 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

请参阅基础设施即代码。

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 云 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

请参阅工业物联网。

不可变基础设施

一种模型，可为生产工作负载部署新的基础设施，而不是更新、修补或修改现有基础设施。不可变基础设施本质上比可变基础设施更一致、更可靠、更可预测。有关更多信息，请参阅 AWS Well-Architected Framework 中的使用不可变基础设施进行部署最佳实践。

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 Klaus Schwab 在 2016 年提出，指的是通过连接、实时数据、自动化、分析和 AI/ML 的进步来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调配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。

物联网

请参阅物联网。

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。LLM 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

请参阅 7 R。

小端序系统

一个先存储最低有效字节的系统。另请参阅字节顺序。

LLM

请参阅大型语言模型。

下层环境

请参阅环境。

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

请参阅分支。

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权限。恶意软件的示例包括病毒、蠕虫、勒索软件、木马、间谍软件和键盘记录器。

托管式服务

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。Amazon Simple Storage Service（Amazon S3）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、记录和控制在车间将原材料转化为成品的生产过程。

MAP

请参阅迁移加速计划。

机制

一个完整的过程，您可以在其中创建工具，推动工具的采用，然后检查结果以进行调整。机制是一种在运作过程中自我强化和改善的循环。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

请参阅制造执行系统。

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的信息。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：使用 AWS 应用程序迁移服务重新托管向 Amazon EC2 的迁移。

迁移组合评测（MPA）

一种在线工具，提供了用于验证迁移到 AWS 云的业务案例的信息。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

将工作负载迁移到 AWS 云的方法。有关更多信息，请参见术语表中的 7 R 词条，以及动员您的组织以加快大规模迁移。

ML

请参阅机器学习。

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，请参阅在 AWS 云中实现应用程序现代化的策略。

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，请参阅在 AWS 云中评估应用程序的现代化准备情况。

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

请参阅迁移组合评测。

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？”

可变基础设施

一种用于更新和修改生产工作负载的现有基础设施的模型。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

请参阅来源访问控制。

OAI

请参阅来源访问身份。

OCM

请参阅组织变革管理。

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

请参阅运营集成。

OLA

请参阅运营级别协议。

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

请参阅开放流程通信 – 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 提供了一个包含数据加密、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）

一份问题核对清单和关联的最佳实践，可帮助您了解、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 AWS Well-Architected Framework 中的运营准备情况审查（ORR）。

运营技术（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 转型的关键重点。

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 AWS CloudTrail 此创建的跟踪记录组织 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

请参阅运营准备情况审查。

OT

请参阅运营技术。

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

请参阅个人身份信息。

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

请参阅可编程逻辑控制器。

PLM

请参阅产品生命周期管理。

policy

一个对象，可以定义权限（请参阅基于身份的策略）、指定访问条件（请参阅基于资源的策略）或定义 AWS Organizations 的组织中所有账户的最大权限（请参阅服务控制策略）。

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。

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 `false` 的查询条件，通常位于 `WHERE` 子句中。

谓词下推

一种数据库查询优化技术，可在传输之前筛选查询中的数据。这将减少从关系数据库检索和处理的数据量，并提高查询性能。

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。

隐私设计

一种在整个开发过程中都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，旨在防止部署不合规资源。这些控制会在资源预置之前对其进行扫描。如果资源与控制不兼容，则不会预置它。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）

对产品在其整个生命周期内的数据和流程的管理，从设计、开发和发布，到增长和成熟，再到衰退和淘汰。

生产环境

请参阅环境。

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。

提示串接

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，可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

一系列用于访问 SQL 关系数据库系统中的数据的步骤，类似于指令。

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

请参阅责任、问责、咨询和知情（RACI）。

RAG

请参阅检索增强生成。

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

请参阅责任、问责、咨询和知情（RACI）。

RCAC

请参阅行列访问控制。

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

请参阅 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

请参阅 7 R。

Region

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，请参阅指定您的账户可以使用的 AWS 区域。

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

请参阅 7 R。

版本

在部署过程中，推动生产环境变更的行为。

重新放置

请参阅 7 R。

更换平台

请参阅 7 R。

重新购买

请参阅 7 R。

韧性

应用程序抵御中断或从中断中恢复的能力。在 AWS 云中规划韧性时，高可用性和灾难恢复是常见的考虑因素。有关更多信息，请参阅 AWS 云 韧性。

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

请参阅 7 R。

停用

请参阅 7 R。

检索增强生成（RAG）

一种生成式人工智能技术，其中 LLM 在生成响应之前引用其训练数据来源之外的权威数据来源。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证的过程。

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

请参阅恢复点目标。

RTO

请参阅恢复时间目标。

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 管理控制台 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

请参阅监督控制和数据采集。

SCP

请参阅服务控制策略。

机密密钥

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 文档中的什么是 Amazon Secrets Manager 密钥？。

安全设计

一种在整个开发过程中都考虑安全的系统工程方法。

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。安全控制有以下四种类型：预防性、检测性、响应性和主动性。

安全固化

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

一种预定义的程序化操作，旨在自动响应或修复安全事件。这些自动化可作为侦探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 实例或轮换凭证。

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。

服务水平指示器（SLI）

对服务性能方面的衡量，例如错误率、可用性或吞吐量。

服务水平目标（SLO）

代表服务运行状况的目标指标，由服务水平指示器衡量。

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

请参阅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

请参阅服务水平协议。

SLI

请参阅服务水平指示器。

SLO

请参阅服务水平目标。

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，请参阅在 AWS 云中实现应用程序现代化的分阶段方法。

SPOF

请参阅单点故障。

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事务数据或测量数据，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。

监督控制和数据采集（SCADA）

在制造业中，一种使用硬件和软件来监控实物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。你可以使用 Amazon S CloudWatch ynthetics 来创建这些测试。

系统提示

一种为 LLM 提供上下文、说明或准则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

标签

键值对，用作组织资源的元数据。 AWS 标签有助于您管理、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

请参阅环境。

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。

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

请参阅环境。

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

一种对与当前记录有某种关联的一组行执行计算的 SQL 函数。窗口函数对于处理任务很有用，例如计算移动平均值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。

WORM

请参阅一次写入多次读取。

WQF

请参阅 AWS 工作负载资格鉴定框架。

一次写入多次读取（WORM）

一种存储模型，可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，但无法对其进行更改。此数据存储基础设施被认为不可变。

Z

零日漏洞利用

一种利用零日漏洞的攻击，通常为恶意软件。

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。

零样本提示

为 LLM 提供执行任务的说明，但没有可以帮助指导的示例（样本）。LLM 必须使用预先训练的知识来处理任务。零样本提示的有效性取决于任务的复杂性和提示的质量。另请参阅少样本提示。

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。