亚马逊云科技 | NVIDIA

利用生成式人工智能和机器学习加速创新

借助由生成式人工智能和机器学习驱动的产品、服务和客户体验，更好地满足客户需求



简介

机器学习（ML）不再是全球科技巨擘和数据科学专家的专利，现已成为主流技术。幸赖于云技术，机器学习广泛应用的障碍正在迅速消失。云将数据、低成本存储、安全性和机器学习服务，以及基于 CPU 和 GPU 的高性能且经济高效的计算实例，外加优化型软件，统统融为一体，这对于机器学习的成功至关重要。云还提供了一种随用随付的成本模型，可进一步协助客户控制成本。

目前最前沿的复杂深度学习模型由灵感来自人脑功能的多层深度神经网络组成，因此迫切需要更加强大的计算资源。这些高级模型需要安全、可扩展且经济高效的 CPU、功能强大的 GPU 和 GPU 优化型软件，以及 GB 或 TB 级存储。

有了云，您可以选择完全托管式服务，自动管理您的基础设施，从而为您省去软硬件维护的麻烦，或者您也可以选择自行管理机器学习生命周期，能够更加亲力亲为地定制基础设施的规模并受益于其中。

无论您最终做何选择，只要有了云，您便无需在前期为一切可能的选项一一投资。资源是按需提供的，且始终保持在最新状态，随时可为您提供专用的机器学习工具、计算、存储、网络连接和最新的基础设施创新成果。

借助生成式人工智能和机器学习更好地满足客户需求

有几个常见的生成式人工智能和机器学习使用场景可协助客户实现业务转型



智能文档处理

简介：

各企业通常都有多种类型的文档，比如发票、患者信息表、贷款申请和合同等，这些文档包含申请人姓名、实体（地址或品牌）或患者健康史等数据，而这些数据对企业的业务又至关重要，需要进行处理。借助于使用文本处理算法构建的机器学习模型，智能文档处理（IDP）可从数以百万计的文档中提取文本，可以明确数据的情绪或数据之间的关系，并增加了一道验证、纠正或增强机器学习结果的人工工序，以求提高准确性和合规性。此外，借助生成式人工智能，用户还可以轻松地生成所有这些文档的摘要，并从这些文档中提取关键洞察。

用途：

IDP 从数字文档中提取数据，以执行诸如处理贷款申请、分析客户情绪、确定患者治疗方式，或者从发票中筛选出不合规的购买行为等任务。

成果：

文档处理基于机器学习进行，因此数据准确性更高，数据处理速度更快。它还可以提高客户满意度，提供更准确的信息，并协助各公司更快、更恰当地响应请求。

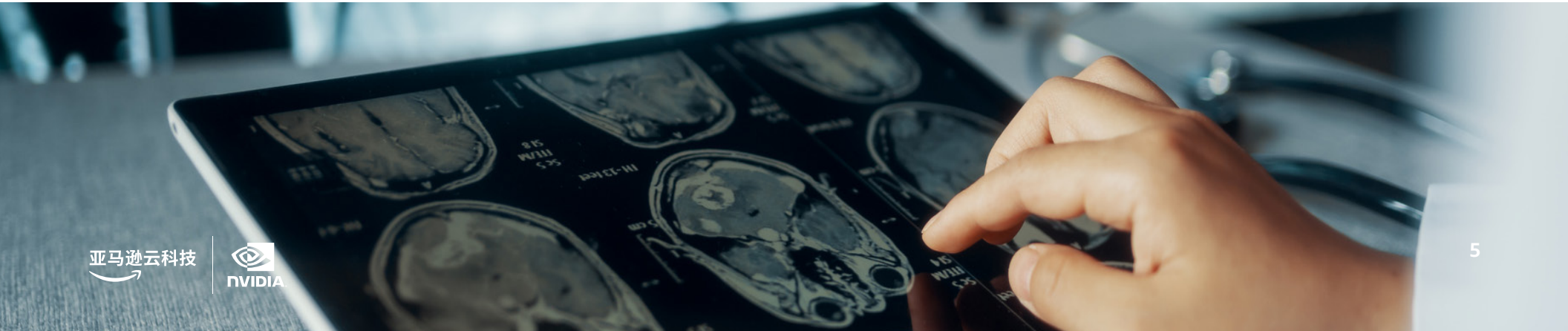
IDP 提高了员工的工作效率，让员工可以将更多的时间花在业务关键型任务上，并减少为获得洞察而需要人工浏览各种文档，以及执行手动数据输入所耗费的时间。自动化文档工作流降低了数据提取和分析的复杂性，使企业能够减少用于执行这些劳动密集型操作的预算和资源。

客户成功案例

Mayo Clinic 的一项研究表明，近 90% 的美国放射科医生都在满负荷或超负荷工作，而 **Rad AI** 有望协助减轻他们的工作负担。该公司训练机器学习模型，替放射科医生阅读详细文档并自动总结结果，医生用这些结果来诊断患者的病症，然后制定出治疗方案。Rad AI 选择将其文档摘要应用程序从基于 GPU 的 **Amazon Elastic Compute Cloud（Amazon EC2）**旧实例迁移到由 NVIDIA A100 Tensor Core GPU 提供技术支持的 **Amazon EC2 P4d** 最新实例。Rad AI 通过在 Amazon EC2 P4d 实例上部署其应用程序，将机器学习推理时间缩短了 60%，从而能够向放射科医生提供更快、更准确的报告，进而改善患者的疗效。

汤森路透社是世界上最值得信赖的资讯提供商之一，其专家团队汇集信息、创新和洞察，为全球各地的企业揭示复杂情况。在法律、税务、新闻等其他领域，汤森路透社拥有 150 多年以来人工加以注释的各类数据。2018 年，该公司选择了 **Amazon SageMaker** 来加快其研发工作。汤森路透社自部署 SageMaker 以来，能够充分利用诸如即时回答生成、长文本摘要以及完全交互式、对话式问答等高级功能。这些功能让汤森路透社得以构建全面的辅助人工智能（AI）系统，引导用户找到可满足其所有信息需求的最佳解决方案。

Rad AI 通过在 Amazon EC2 P4d 实例上部署其应用程序，将机器学习推理时间缩短了 60%，从而能够向放射科医生提供更快、更准确的报告，进而改善患者的疗效。





计算机视觉

简介：

借助计算机视觉（CV），机器能够识别图像中的人物、地点和事物，准确度达到或超过人类水平，并且速度和效率要高得多。CV 自动从单个图像或图像序列中提取、分析、分类并理解有用信息。图像数据可以采用多种形式，例如单张图像、视频序列、来自多个摄像机的视图或三维数据等。

用途：

一些公司使用 CV 来检测媒体和娱乐节目中的不当内容，协助提高品牌声誉和安全性。在医护领域，CV 应用程序用于分析医学图像，并识别需要额外分析的图像，从而改善患者疗效。在制造领域，许多公司正在使用 CV 自动检测装配线上的组件是否存在缺陷。

客户成功案例

牛津大学的**花园、图书馆和博物馆（GLAM, Gardens, Libraries & Museums）**收藏了 2100 万件藏品，这些都是世界上特别重要的文物和标本。利用 SageMaker，牛津大学目前正在使用由基于 NVIDIA GPU 的 **Amazon EC2 P3 实例**提供支持的 CV 来构建增强版图像识别系统，并加快对其收藏的大量硬币进行编目的过程。以前，分析一枚硬币需要志愿者 10 分钟到几个小时的时间，今后，一旦构建完成图像识别系统，预计分析过程只需几分钟。[1]

Aerobotics 是一家农业科技公司，业务遍及全球 18 个国家/地区，总部位于南非开普敦。该公司的使命是提供智能工具来回馈世界。为达成这一使命，该公司在其 Aeroview 平台上为农民提供富有实用价值的数据和洞察，以便农民们可以在生长季节的适当时间进行必要的介入。公司的主要数据来源是无人机航拍图像：捕捉果园中树木和水果的视觉和多光谱图像。

Aerobotics 公司使用 SageMaker 改进其 Tree Insights 产品，该产品提供每棵树的重要参数的测量值（如树冠面积和健康状况等），还能提供枯死和缺失树木的位置。农民利用这些信息进行精确的干预，如修理灌溉线、以不同速率施肥，以及订购新的树木补种等。

以前，分析一枚硬币需要志愿者 10 分钟到几个小时的时间，今后，一旦构建完成图像识别系统，预计分析过程只需几分钟。[1]

[1]《University of Oxford Introduces a Sector-Leading Image Recognition ML Prototype to Augment Digitization in Numismatics》，亚马逊云科技，2021 年



个性化推荐

简介：

如今，消费者在考虑、购买和使用产品和服务时，都希望通过数字化渠道获得实时的、经策管的体验。机器学习算法可用于根据不同渠道的个人喜好和行为，扩展并创造个性化的客户体验。

用途：

企业可以构建能够提供广泛的个性化体验的应用程序，这些体验包括特定的产品推荐、个性化的产品重新排名，以及定制型直接营销，全部由生成式人工智能驱动。通过基于机器学习的个性化，企业可以超越僵化的静态规则，使用推荐系统，向客户提供高度个性化的建议。

成果：

机器学习有助于企业提供高度个性化的体验，从而提高客户参与度、转化率、收入和利润率，并在数字世界中创造差异化。

客户成功案例

跨国电子商务公司 **Zalando** 决定在云端实现其机器学习工作负载标准化，进而改善客户体验、提高工作效率，同时推动公司业务发展。借助 SageMaker，Zalando 得以更好地开展营销活动、生成个性化服装推荐并提供更具吸引力的客户体验，同时将工程师和数据科学家的工作效率提高 20%。[2]

NerdWallet 是一家个人理财初创公司，通过提供适当的工具和建议，协助客户轻松偿还债务、选择最佳金融产品和服务并且实现主要的人生目标，例如购房或储蓄退休金等。该公司非常依赖数据科学及机器学习，为客户提供个性化的金融产品。以前，NerdWallet 会以列表的形式为客户推荐他们可能喜欢的信用卡，但无法预测被客户接受的可能性。借助 SageMaker，该公司得以更有效地为客户匹配合适的金融产品。

Amazon EC2 P3 实例使用 SageMaker 并配备 NVIDIA V100 Tensor Core GPU，还提高了 NerdWallet 的灵活性和性能，从而将数据科学家训练机器学习模型所需的时间从几个月缩短到几天。

借助 SageMaker，Zalando 得以更好地开展营销活动、生成个性化服装推荐并提供更具吸引力的客户体验，同时将工程师和数据科学家的工作效率提高 20%。[2]

[2] 亚马逊云科技客户成功案例：Zalando，亚马逊云科技





生成式人工智能

简介：

生成式人工智能是一种可以生成新内容和想法（包括对话、故事、图像、视频和音乐）的人工智能。与所有人工智能一样，生成式人工智能也是由基于大量数据进行预训练的大型模型提供支持，这些模型通常称为根基模型（FM）。

用途：

通过生成式人工智能，您可以利用机器学习快速推动业务发展，并将其应用于更广泛的使用场景。可以在工程、市场营销、客户服务、金融和销售等所有业务领域应用生成式人工智能。

利用生成式人工智能，可以通过聊天机器人、虚拟助理、智能联络中心和个性化等功能改善客户体验。您还可以利用由生成式人工智能驱动的对话式搜索、内容创建、文本摘要和代码生成等功能，提高员工的工作效率。

此外，还能通过智能文档处理、维护助手、质量控制和视觉检测以及合成训练数据生成，来改善业务运营。最后，您还可以使用生成式人工智能来加速各类创意内容的制作：从艺术和音乐到文本、图像、动画和视频等。

生成式人工智能使用场景

生成式人工智能的使用场景以及带来的可能性惠及各个行业以及每个人。以下是迄今为止最常见的应用领域（按行业区分）：



生命科学

- **创建全新蛋白质序列：**通过创建具有特定属性的序列来设计抗体、酶、疫苗和基因疗法，从而加速药物研发和研究
- **设计合成基因序列：**医疗保健和生命科学公司可将由人工智能生成的基因序列应用于合成生物学和代谢工程，例如创建新的生物合成途径或优化基因表达，以实现生物制造目的



医疗保健

- **创建合成患者和医疗保健数据：**利用模拟数据集，即使无法或难以获取真实数据，企业也可以训练人工智能模型、模拟临床试验或研究罕见病，同时确保遵守行业严格的安全和隐私要求
- **改善患者体验：**生成式人工智能可以提供个性化的患者出院指导和治疗计划。对话助理和聊天机器人可减少临床医生的工作量，提高患者满意度，并有助于为高风险社区提供积极主动的医疗保健服务



金融服务

- **改善体验：**金融服务公司可以部署能够更快解决问题的聊天机器人、提供个性化产品和推荐，并自动完成内部任务，同时仍能提供行业所需的强大数据加密和隐私控制，以此来更好地服务客户和员工
- **提高知识型员工的效率：**金融公司的知识型员工可以更快地处理申请，更深入地洞察客户行为，改善协作，并部署强大的训练计划和模拟
- **分析市场情绪:**通过对社交媒体、新闻报道和金融数据进行更快、更全面的分析，金融服务公司可以了解市场评论，更快地发现商机，并主动降低风险



媒体与娱乐

- **加快内容创建：**从故事情节设计和构思，到后期制作流程，媒体和娱乐公司可以实现低级别任务的自动化，从而提高制作速度，让创意人才能够更快地进行迭代，实现导演的创作愿景
- **提高音乐质量：**艺术家可以借助人工智能生成的音乐来补充和改进他们的专辑，从而创造全新的音乐流派
- **辅助媒体供应链：**生成式人工智能应用程序可以辅助或自动执行本地化、内容审核甚至是内容恢复等任务



教育

- **生成文本摘要：**学生和教师可以创建研究文档、讲课记录和课堂笔记的简明摘要，使其更易于搜索和浏览
- **自动创建内容：**人工智能可将信息转化为试题范例，加快评分速度，通过各种因素衡量学生的表现，并向老师和学生提供个性化反馈和建议
- **提供个性化学习环境：**教育工作者可以为学生群体甚至学生个人，创建个性化学习路径，并利用模拟和虚拟现实，让学习变得更有吸引力



汽车制造

- **改进产品设计：**制造商可以利用人工智能优化机械零件的设计，或创造全新的材料、芯片和零件设计，从而提高质量和耐用性、降低成本并简化生产
- **打造个性化车内体验：**虚拟助手和个性化路线推荐可提升驾驶员和乘客的体验
- **测试和维护：**人工智能可通过生成数据表中缺失的信息来改进产品测试，并实现新的辅助维护使用场景，以更好地维护和维修机器，包括消费者使用的产品
- **提高工厂整体的设备效率：**实现历史机器维护数据、维修数据、设备手册、生产数据，以及可能来自其他制造商的数据的数字化并捕获这些数据，以生成维护、维修或设备参数方面的建议，并提高生产率、可用性和质量

鉴于生成式人工智能的快速发展和商业价值的不断提升，今后，这一变革性技术的使用场景和应用领域的数量只会不断增加。

数据扩充

生成式人工智能可以通过一种叫做标注高效学习（label-efficient learning）的过程来创建合成数据，事实证明，这种能力在所有模式和使用场景中都非常有用。如果真实数据不存在、受限制或根本无法准确地处理极端情况时，便可以利用合成数据来训练人工智能模型。用于创建生成式人工智能应用程序的根基模型，既可以通过自动生成额外的扩充训练数据来降低标注成本，也可以通过学习数据的内部表示形式，以使用较少的标注数据来训练人工智能模型，继而降低标注成本。

使用场景不断增多

除了这些常见使用场景之外，我们还观察到，
有一些新兴使用场景正在迅速涌现



自主操作系统可降低成本、提高生产力

自主操作系统使用许多不同的机器学习模型来感知环境，并在没有人工干预的情况下运行。自主操作系统依靠传感器、执行器、复杂算法、机器学习系统和强大的处理器，来快速执行软件。

机器人就是自主操作系统的一个例子。作为尖端技术的提供者，**Amazon Robotics** 早前就意识到，如果能使用人工智能和机器学习来自动处理货品配送流程中的关键工作，则意味着有可能为公司带来巨额收益，因此在 2017 年，公司派出多个团队来实现这一点。随着公司迭代机器学习项目，其选择了亚马逊云科技和 SageMaker，SageMaker 是一项托管式服务，有助于数据科学家和开发人员快速准备、构建、训练并部署高质量的机器学习模型。这样一来，Amazon Robotics 团队便从维护和管理 NVIDIA GPU 的艰巨任务中解放了出来，同时仍能使用大量 GPU 跨多个区域大规模运行推理。截至 2021 年 1 月，该解决方案为公司节省了近 50% 的机器学习推理成本，并使生产力提高了 20%，同时节省了比例差不多的总体成本。



通过预测型医护拯救生命

目前，由于患者数据不足，或者没有足够的时间来分析并关联大型患者数据集，医疗人员对严重疾病的及时诊断常常被延误。**CloudMedx** 正在通过开发机器学习模型，来协助应对这一挑战，机器学习模型能够理解不同的疾病、症状和药物之间的相互关系，有助于预测疾病恶化，还能预测出患者可能出现并发症的概率。[3]



利用机器学习改善废水管理

对于公共卫生以及保护宝贵的水资源而言，废水管理至关重要。**Opseyes** 利用 AI 开发了第一个用于废水处理厂的快速显微镜测试。这种测试能便于用户立即检查工厂状况，迅速消除污染威胁，从而避免停机。[4]



机器学习让音乐家如虎添翼

Sunhouse 是一家由音乐家和技术专家创立的 AI 初创公司，该公司正在创建一个基于机器学习的系统，使鼓手能够将架子鼓变成用于巡演、作曲和即兴演奏的整个生产套件。Sunhouse 的技术采用定制的鼓传感器和 AI 支持的声学映射，使鼓成为一种富有表现力的工具，能够使用样本、效果和 MIDI 进行创作和表演。Sensory Percussion 使鼓手能够用鼓槌控制电子设备，为打造创意和创作音乐开辟了一条全新的道路。Sunhouse 的解决方案自建立以来，已被纽约爵士乐界的现场表演者和当今许多著名鼓手广泛使用，其中包括 Marcus Gilmore 和 Wilco 的 Glenn Kotche。[5]

[3] Darrow, B.，《This AI Software {CloudMedx} Aims to Conquer Heart Disease》，CloudMedx，2016 年 10 月
[4] Caufield, B.，《In the Drink of an AI: Startup Opseyes Instantly Analyzes Wastewater》，NVIDIA 博客，2021 年 2 月
[5] Finkle, L.，《Drum Roll, Please: AI Startup Sunhouse Founder Tlacael Esparaza Finds His Rhythm》，NVIDIA 播客，2021 年 3 月

来自亚马逊云科技和 NVIDIA 的解决方案

亚马逊云科技和 NVIDIA 合作已超过 12 年，不断为客户提供强大、经济高效且灵活的基于 GPU 的解决方案。**基于 NVIDIA GPU 的 Amazon EC2 实例**提供了快速准确地训练机器学习和深度学习模型所需的高性能、成本优化的基础设施。借助由 NVIDIA A100 Tensor Core GPU 支持的最新 **Amazon EC2 P4d 实例**，开发人员能够将训练模型的时间，从几天缩短到几分钟。您能够以高效率和优化的成本，运行最复杂的多节点训练。这使您能够快速试验、训练并调整模型，从而加速创新。

它们是第一批在云中支持 400 Gbps 实例联网的实例。与上一代 Amazon EC2 P3 和 Amazon EC2 P3dn 实例相比，Amazon EC2 P4d 实例使深度学习模型的性能平均得以提高 2.5 倍。

Amazon EC2 P4d 实例还部署在称为 EC2 UltraClusters 的超大规模集群中，这种集群由云中具有最高性能的计算、联网和存储资源组成。每个 EC2 UltraCluster 都是全球最强大的超级计算机之一，使您能够运行最复杂的多节点机器学习模型训练。在 EC2 UltraClusters 中，您可以根据机器学习项目需求，从几个 NVIDIA A100 Tensor Core GPU 快速扩展到数千个。

Amazon EC2 G5 实例是最新一代基于 NVIDIA GPU 的实例，可用于广泛的机器学习使用场景。与 Amazon EC2 G4dn 实例相比，这些实例的机器学习推理性能提升高达 3 倍，机器学习训练的性能提升高达 3.3 倍。

客户可以使用 G5 实例获得高性能且经济高效的基础设施，来为自然语言处理（NLP）、CV 和推荐器引擎使用场景训练和部署更大且更复杂的模型。

G5 实例配备了多达 8 个 NVIDIA A10G Tensor Core GPU 和第二代 AMD EPYC 处理器。这些实例还支持多达 192 个 vCPU、高达 100Gbps 的网络带宽和高达 7.6TB 的本地 NVMe SSD 存储空间。

亚马逊云科技 Marketplace 上提供的 **NVIDIA AI Enterprise** 是一个全栈软件库，包括 100 多个框架、预训练模型、人工智能工作流和基础设施优化。该方案可以简化生成式人工智能、语音人工智能、视觉人工智能、网络安全等领域的生产就绪型应用的开发和部署。NVIDIA AI Enterprise 会进行持续的漏洞监控，以及定期及时修补关键和常见的漏洞和暴露（CVE），同时提供 API 稳定性、具有 SLA 的企业支持以及与 NVIDIA AI 专家进行沟通的机会，从而解决企业在构建和维护自己的高性能、安全、云原生人工智能软件平台时面临的复杂问题。

立即借助 Amazon SageMaker 这一利器开始利用强大的机器学习基础架构

要从亚马逊云科技云端强大的机器学习基础设施中获益，最简单快捷的方法就是部署 **Amazon SageMaker**。这种完全托管式服务将数据标注、数据准备、特征工程、统计数据偏差检测、AutoML、训练、优化、托管、可解释性、监控和工作流等广泛的基本功能融为一体。

Amazon SageMaker JumpStart 是一个机器学习中心，其中提供了算法、模型和机器学习解决方案。利用 SageMaker JumpStart，客户可以发现、探索和部署 **Amazon Bedrock** 中所没有的开源 FM，如 OpenLLaMA、RedPajama、Mosaic MPT-7B、FLAN-T5/UL2、GPT-J-6B/Neox-20B 和 Bloom/BloomZ。我们还在不断增加更多模型，仅在今年，我们提供的 FM 就增加了一倍。对于希望创建自己的 FM 的客户，Amazon SageMaker 可提供托管基础设施和工具，以加快构建、训练和部署可扩展、可靠且安全的模型。

亚马逊云科技上的 NVIDIA AI 平台包括 NVIDIA GPU 加速型实例、NVIDIA GPU 优化型软件和 NVIDIA AI Enterprise，有助于开发人员大幅提高 SageMaker 中机器学习算法的开发速度。该平台使开发人员能够更快地运行模型训练和推理，更早地部署机器学习应用程序，并实现成本节约。

为了进一步加快人工智能模型的部署并降低推理成本，SageMaker 集成了 **NVIDIA Triton Inference Server**。此举让诸多功能得以实现，例如多框架支持、动态批处理和并发模型执行，从而最大限度地提高基于亚马逊云科技的 CPU 和 GPU 实例的性能。

为实现这些优势，NVIDIA 在**亚马逊云科技 Marketplace** 上提供了 NVIDIA AI Enterprise，这全面集合了 GPU 优化库、用于 CV 的预训练人工智能模型、对话式人工智能和推荐器、应用程序框架和推理服务解决方案（例如 Triton Inference Server），以简化机器学习模型在 CPU 和 GPU 上的部署。NVIDIA AI Enterprise 中的软件不断优化，不仅安全，还由 NVIDIA 提供支持，让开发人员能够轻松利用最新的 NVIDIA 创新。软件会定期发行更新版，让用户能够体验最新功能和性能改进。

将机器学习应用于您的业务

通过在云中运行机器学习工作负载，企业可以按需访问最强大的 GPU 实例和机器学习工具，这些实例和工具可以在几分钟内启动，从一个实例扩展到数以千计的实例，同时还能控制基础设施成本。

NVIDIA 支持的亚马逊云科技服务和基础设施，适用于所有经验水平的企业——从那些擅长构建机器学习工作负载且希望管理其基础设施的企业，到那些更喜欢完全托管式方案的企业。亚马逊云科技在机器学习开发生命周期的每个步骤中，使用计算、联网、存储和机器学习工具为您的企业提供支持，包括收集和准备数据，选择正确的算法，调整模型，旨在获得最大的准确性，以及长期部署并监控模型的性能和质量。

了解亚马逊云科技和 NVIDIA 的潜力 ›

开始使用 Amazon SageMaker ›

*前述特定亚马逊云科技生成式人工智能相关的服务仅在亚马逊云科技海外区域可用，亚马逊云科技中国仅为帮助您发展海外业务和/或了解行业前沿技术选择推荐该服务。

亚马逊云科技 | NVIDIA

©2023，Amazon Web Services, Inc. 或其关联公司。保留所有权利。