亚马逊云科技AI创新重塑生成式AI与机器学习模型的构建与扩展

通过Amazon SageMaker HyperPod的三项新功能，以及直接在Amazon SageMaker中整合亚马逊云科技合作伙伴的热门AI应用产品，亚马逊云科技帮助客户消除AI开发生命周期中无差别的繁重工作，从而更快速、更轻松地构建、训练和部署模型

北京2024年12月16日 /美通社/ -- 亚马逊云科技在2024 re:Invent全球大会上，宣布推出Amazon SageMaker AI四项创新，助力企业更快使用热门的公开模型，最大化训练效率、降低成本，并使用其首选工具加速生成式人工智能（AI）模型的开发。Amazon SageMaker AI是一项端到端的服务，数十万客户使用它来构建、训练和部署各种用例的AI模型，它提供完全托管的基础设施、工具和工作流。

- Amazon SageMaker HyperPod新增三项强大功能，帮助客户更轻松地快速开始训练时下流行的公开可用模型，通过灵活训练计划节省数周的模型训练时间，并最大化计算资源利用率，将成本降低高达40%。
- 现在，客户可以直接在Amazon SageMaker中轻松安全地发现、部署和使用来自亚马逊云科技合作伙伴的完全托管的生成式AI和机器学习（ML）开发应用，例如Comet、Deepchecks、Fiddler AI和Lakera，从而灵活选择最适合的工具。
- Articul8、澳大利亚联邦银行、富达、Hippocratic AI、Luma AI、NatWest、NinjaTech AI、OpenBabylon、Perplexity、Ping Identity、Salesforce和汤森路透等客户正在使用Amazon SageMaker的新功能，加速生成式AI模型开发。

亚马逊云科技人工智能和机器学习服务与基础设施副总裁Baskar Sridharan博士表示："亚马逊云科技在七年前推出Amazon SageMaker，以简化构建、训练和部署AI模型的过程，帮助各种规模的组织访问和扩展其对AI和机器学习的使用。随着生成式AI的兴起，Amazon SageMaker不断快速创新，自2023年以来已经推出了超过140项功能，帮助Intuit、Perplexity和Rocket Mortgage等企业更快地构建基础模型。通过此次发布，我们将为客户提供更高性能、更具成本效益的模型开发基础设施，帮助他们加速将生成式AI工作负载部署到生产环境中。"

Amazon SageMaker HyperPod：训练生成式AI模型的首选基础设施

随着生成式AI的出现，构建、训练和部署机器学习模型的过程变得更加困难，这需要深厚的AI专业知识、访问大量数据以及创建和管理大型计算集群。此外，客户需要开发专门的代码来实现跨集群分布式训练，持续检查和优化模型，并手动处理硬件故障，同时尽量控制时间进度和成本。亚马逊云科技为此推出Amazon SageMaker HyperPod，帮助客户在数千个AI加速器上高效扩展生成式AI模型开发，将训练基础模型的时间缩短高达40%。无论是Writer、Luma AI、Perplexity等领先的初创公司，还是汤森路透、Salesforce等大型企业，都在利用Amazon SageMaker HyperPod加速模型开发。亚马逊还使用Amazon SageMaker HyperPod训练新的Amazon Nova模型，不仅降低了训练成本，提高了训练基础设施的性能，还节省了数月手动设置和管理集群的时间。

现在，越来越多的企业希望微调热门的公开可用模型，或训练自己的专用模型，以利用生成式AI改造业务和应用。Amazon SageMaker HyperPod将持续创新，帮助客户更轻松、更快速、更具成本效益地大规模构建、训练和部署这些模型，具体创新包括：

- 新训练配方帮助客户更快上手：许多客户希望基于Llama和Mistral等热门的公开可用模型，使用内部数据为特定用例进行定制。然而，优化训练性能可能需要数周的反复测试，包括尝试不同的算法、调整参数、观察训练效果、调试问题和设定性能基准。为了帮助客户在几分钟内快速入门，Amazon SageMaker HyperPod现在提供30多个精选的模型训练配方，可适用于时下热门的一些公开可用模型，包括Llama 3.2 90B、Llama 3.1 405B和Mistral 8x22B。这些配方极大地简化了客户的入门过程，自动加载训练数据集、应用分布式训练技术，并配置系统以实现高效的检查点管理和基础设施故障恢复。不同技能水平的客户能够从一开始就在亚马逊云科技基础架构上优化模型训练的性价比，省去了数周的反复评估和测试的时间。客户可以通过Amazon SageMaker GitHub存储库浏览可用的训练配方，根据定制需求调整参数，并在几分钟内完成部署。此外，客户只需一行简单编辑，即可在基于GPU或Trainium的实例之间无缝切换，进一步优化性价比。

  Salesforce的研究人员一直在寻求一种快速启动基础模型训练和微调的解决方案，希望能够在不用过多关注基础设施的情况下，避免为每个新模型耗费数周时间进行训练堆栈优化。通过Amazon SageMaker HyperPod的定制模板，他们现在能够快速开展基础模型的原型设计。目前，Salesforce的AI研究团队可以在短短几分钟内启动各种

预训练和微调流程，并成功实现基础模型的高效运营。

- 灵活训练计划可轻松满足训练时限和预算要求：尽管基础设施创新有助于降低成本并提高训练效率，但客户仍需规划并管理所需计算资源，以确保在预算范围内按时完成训练任务。因此，亚马逊云科技为Amazon SageMaker HyperPod推出了灵活训练计划。客户只需轻松点击几下，就能指定预算、截止日期和所需的最大计算资源量。Amazon SageMaker HyperPod会自动预留容量、设置集群并创建模型训练作业，帮助团队节省数周的训练时间，减少客户在获取大型计算集群以完成模型开发任务时的不确定性。如果提议的训练计划无法满足指定的时间、预算或计算要求，Amazon SageMaker HyperPod会提供替代方案，如延长日期范围、增加计算资源或选择不同的亚马逊云科技区域进行训练。一旦计划获批，Amazon SageMaker会自动配置基础设施并启动训练作业。Amazon SageMaker使用 Amazon Elastic Compute Cloud (EC2)容量块预留所需的加速计算实例，确保训练任务按时完成。根据容量块的可用时间，Amazon SageMaker HyperPod通过有效的暂停和恢复训练作业，确保客户始终拥有按时完成任务所需的计算资源，无需人工干预。

  Hippocratic AI为医疗保健开发以安全为重点的大语言模型（LLM）。为了训练多个模型，Hippocratic AI采用了Amazon SageMaker HyperPod灵活训练计划，获得了按时完成训练任务所需的加速计算资源。这帮助他们将模型训练速度提高了4倍，并更有效地扩展其解决方案，以适应数百个用例。

- 任务治理功能最大化加速器利用率：越来越多的企业为模型训练配置大量加速计算资源。这些计算资源昂贵且有限，因此客户需要一种管理资源使用率的方法，以确保其计算资源优先用于最关键的模型开发任务，避免任何浪费或利用率不足。如果没有对任务优先级和资源分配的有效控制，一些项目最终会因资源不足而停滞，而同时其他项目却资源利用率不足。这给管理员带来了巨大负担，他们必须不断重新规划资源分配，而数据科学家则难以取得进展。这不仅阻碍了企业将AI创新快速推向市场，还可能导致成本超支。通过Amazon SageMaker HyperPod任务治理功能，客户可以在模型训练、微调和推理过程中最大化加速器的利用率，将模型开发成本降低最多 40%。只需点击几下，客户就可以轻松为不同任务定义优先级，并为每个团队或项目可以使用的计算资源设置限制。一旦客户在不同团队和项目之间设置了限制，Amazon SageMaker HyperPod将分配相关资源，自动管理任务队列以确保最关键的工作优先进行。例如，如果客户紧急需要更多计算资源来支持面向客户的推理任务，但所有计算资源都已被占用，Amazon SageMaker HyperPod会自动释放未充分利用的资源或非紧急任务的资源，以确保紧急推理任务获得所需资源。在这种情况下，Amazon SageMaker HyperPod会自动暂停非紧急任务，保存检查点以保证已完成的工作完好无损，并在更多资源可用时从最后保存的检查点恢复任务，确保客户最大化计算资源的利用。

  Articul8 AI是一家快速成长的初创企业，致力于帮助企业构建自己的生成式AI应用产品，因此需要不断优化计算环境，以尽可能高效地分配资源。通过使用Amazon SageMaker HyperPod中的新任务治理功能，该公司的GPU利用率有了显著提高，减少了空闲时间，并加速了端到端模型开发。自动将资源转移到高优先级任务的能力提高了团队的生产力，使他们能够更快地推出生成式AI创新成果。

在Amazon SageMaker中使用亚马逊云科技合作伙伴的热门AI应用产品，加速模型开发和部署

许多客户在使用Amazon SageMaker AI的同时，也在使用业界一流的生成式AI和机器学习模型开发工具来执行专业任务，如跟踪和管理实验、评估模型质量、监控性能和保护AI应用产品。然而，将热门的AI应用产品集成到团队的工作流程中是一个耗时的多步骤过程。这包括寻找合适的解决方案、执行安全和合规性评估、监控跨多个工具的数据访问、配置和管理必要的基础设施、构建数据集成以及验证是否符合治理要求。现在，亚马逊云科技帮助客户更轻松地将专业AI应用产品的强大功能与Amazon SageMaker AI的托管能力和安全性结合起来。这项新功能让客户能够直接在Amazon SageMaker中轻松发现、部署和使用来自领先合作伙伴（如Comet、Deepchecks、Fiddler和Lakera Guard）的最佳生成式AI和机器学习开发应用，从而消除其中的阻碍繁重的工作。

Amazon SageMaker是首个为一系列生成式AI和机器学习开发任务提供精选的、完全托管且安全的合作伙伴应用集的服务。这为客户构建、训练和部署模型提供了更大的灵活性和控制权，同时将AI应用产品的上线时间从数月缩短到数周。每个合作伙伴应用都由Amazon SageMaker AI完全托管，因此客户不必担心设置应用或持续监控以确保有足够的容量。通过Amazon SageMaker可直接访问这些应用，客户无需将数据从安全的亚马逊云科技环境中移出，同时可以减少在不同界面之间切换的时间。客户只需浏览Amazon SageMaker合作伙伴AI应用产品目录，了解他们想要使用的应用的功能、用户体验和定价。然后，他们可以轻松选择和部署应用，并使用Amazon Identity and Access Management（Amazon IAM）管理整个团队的访问权限。

Amazon SageMaker在Ping Identity自研的AI和机器学习基础设施的开发和运营中也发挥着关键作用。借助Amazon SageMaker中的合作伙伴AI应用产品，Ping Identity将能够通过私有、完全托管的服务，为其客户提供更快速、更高效的机器学习驱动的功能，同时满足严格的安全和隐私要求，并减少运营开销。

Amazon SageMaker全部创新技术现已全面可用。

原文地址：http://www.china-nengyuan.com/news/218890.html