亚马逊云科技

亚马逊云科技 AIGC 加速企业创新实践指南

前言

在人工智能发展的漫长历程中，如何让机器学会创作一直被视为难以逾越的天堑，“创造力”也因此被视为人类与机器最本质的区别之一。然而，人类的创造力也终将赋予机器创造力，把世界送入智能创作的新时代。采用机器学习的新范式赋能业务不断探索已经播种了几十年，但随着**足够的可伸缩算力的就位、海量数据的爆炸，**以及**机器学习技术的快速进步，**各行各业的客户开始对业务进行重塑。最近，以 ChatGPT 为代表的智能对话应用引起了广泛的关注，引发了诸多想象。我们正处在一个机器学习被大规模采用的转折点上，我们也相信人工智能将会重塑大量客户体验和应用程序。从机器学习到智能创造，从**专业生产内容（PGC, Professional-generated Content)，用户生产内容（UGC, User-generated Content）**到**人工智能生成内容 (AIGC, AI Generated Content)**，我们见证了一场深刻的生产力变革，而这份变革也开始影响到我们工作与生活的方方面面，AIGC 也慢慢的演变成了一场技术和艺术碰撞的盛宴，不断释放人类创造力，提高艺术设计领域的数字化创新效率。

本白皮书将结合 AIGC 领域最新技术趋势和真实行业客户案例，向所有 AIGC 的决策者，开发者、创业者和使用者展示 AIGC 给各行各业带来的创新与变革，帮助用户更好的理解 AIGC 带给企业的价值，以及如何借助亚马逊云科技的产品和服务快速高效地构建差异化的 AIGC 应用，增强企业在 AIGC 时代的敏捷性与竞争力。

如果您有任何问题，欢迎拨打亚马逊云科技热线电话

亚马逊云科技海外区域: 1010 0866
亚马逊云科技中国（宁夏）区域-由西云数据运营: 1010 0966
亚马逊云科技中国（北京）区域-由光环新网运营: 1010 0766

- 1 键-申请账号及产品咨询
- 2 键-云创计划及联合创新中心
- 3 键-账号账单问题
- 4 键-备案咨询（仅由（宁夏）区域和（北京）区域热线支持）
- 5 键-培训与认证
- 6 键-市场活动查询
- 7 键-亚马逊云科技合作伙伴网络（仅由海外区域热线支持）
- 8 键-Marketplace 产品咨询-仅由（宁夏）区域热线支持

扫码或点击了解更多亚马逊云科技
AIGC 技术能力与解决方案

目　录

篇章一
AIGC 介绍与典型行业应用场景

AIGC，生成式 AI(Generative AI) 与基础模型 (Foundation Models)

Gartner 将生成式 AI 列为最有商业前景的人工智能技术之一。根据其发布的 2022 年人工智能技术成熟度曲线，预计生成式 AI 将在 2-5 年内进入生产成熟期，发展潜力与应用空间巨大。**2025 年，30% 大型组织对外消息将由生成式 AI 生成。2025 年，50% 的药物发现与研发将使用生成式 AI。2027 年，30% 的制造商将使用生成式 AI 提高产品研发效率。**从数字内容生产，到引领产业变革，商业前景，加速产业融合与变革。

AIGC（生成式人工智能）是指可生成全新内容的人工智能技术

从字面意思来看，AIGC 是继 PGC，UGC 之后的新型内容创作方式，可以在创意、表现力、迭代、传播、个性化等方面，充分发挥技术优势，**打造新的数字内容生成与交互形态。**

因此，AIGC 的狭义概念是利用人工智能自动生成内容的生产方式。但是 AIGC 已经代表了人工智能技术发现的新趋势，过去传统的人工智能偏向于分析能力，即通过分析一组数据，发现其中的规律和模式并用于多种用途，比如应用最为广泛的个性化推荐算法。而现在的人工智能正在生成新的东西，而不是仅仅局限于分析已经存在的东西，从而实现了人工智能从感知理解到生成创造的跃迁。广义的 AIGC 可以看作是像人类一样具备生成创造能力的人工智能技术，即生成式人工智能，它可以基于训练数据和生成算法模型，自主生成创造新的文本、图像、音乐、视频、3D 交互内容（如虚拟人、虚拟物品、虚拟环境等）等各种形式的内容和数据，以及包括开启科学新发现、创造新的价值和意义等。因此，AIGC 已经加速成为了人工智能领域的新疆域，推动人工智能迎来下一个时代。

- 人工智能，可为现实世界的任务制作足够接近人类生成内容的原创内容
- 由大量数据预先训练的基础模型驱动
- 只需微调，即可用于特定领域自定义任务
- 适用于文本摘要、问答、数字艺术创作、代码生成等多种用例
- 减少机器学习模型开发和加速创新的时间与成本

与所有人工智能技术一样，AIGC 的能力由机器学习模型提供，这些模型是**基于大量数据进行预先训练的基础模型，**通常被称为**基础模型（Foundation Models）**。机器学习的最新进展（特别是基于 transformer 的神经网络架构的发明）直接带来这一类模型的爆发式增长，这类模型通常包含数十亿个参数或变量。如今的基础模型，例如**大型语言模型 GPT3.5** 或 **BLOOM**，以及由 Stability AI 开发的**文生图模型 Stable Diffusion，**可以执行跨多个领域的多种任务，例如**撰写博客文章、生成图像、解决算术问题、对话聊天，基于文档回答问题等。**

尽管**预训练基础模型**所带来的功能和可能性已足够令人惊叹，而真正让业界兴奋不已的是，**这些通用模型也可以被定制化加工，执行专属于其业务领域的特定功能，帮助业务建立差异化竞争优势，**与从零开始训练模型相比，**仅需使用一小部分数据和计算资源**。定制化的基础模型可以带来独有的顾客体验，体现公司的观点、风格和服务，适用于众多消费者行业，如金融银行、旅行和医疗等。例如，一家金融公司如果需要使用所有相关交易自动生成活动日报以供内部流通，它可以使用包括既往报告在内的专有数据来定制模型，以便基础模型了解如何阅读报告和使用哪些数据来生成日报。

基础模型拥有巨大的潜力，但我们仍处在初级阶段。基础模型的优点在于它们能够通过预训练学习大量的语言表征，从而可以在未经过标记的数据上进行微调，并在各种任务上表现出色。但是，基础模型也有一些缺点，包括计算成本高和数据偏差等问题。

计算成本是基础模型的一个主要缺点。由于这些模型具有数十亿个参数，因此它们需要大量的计算资源才能进行训练和推理。对于中小型企业来说，训练自己的基础模型非常困难，而且在推理时也需要多个 GPU 进行计算，因此运行成本非常高。另一个缺点是数据偏差。由于基础模型是在互联网上的未经筛选数据上进行预训练的，因此这些数据可能包含偏见、仇恨言论等有害信息。即使有人工标注员，也难以检查每个数据点，因此这可能导致基础模型的信任度不高。

尽管存在这些缺点，基础模型的发展仍然是一种重要趋势，它可以提高自然语言处理领域的效率和灵活性。将来，我们可能会看到更多的基础模型应用于各种任务和应用程序，从而推动人工智能技术的进一步发展。

AIGC 典型应用场景与行业分布

按照模态区分，AIGC 又可分为音频生成、文本生成、图像生成、视频生成及图像、视频、文本间的跨模态生成，细分场景众多，其中跨模态生成值得重点关注。

AIGC 塑造数字内容生产与交互的新范式

伴随数字技术与实体经济的深度融合、互联网企业数字化场景拓展至元宇宙，人类对数字内容的总量和丰富程度的需求不断提高，AIGC 作为当前重要的内容生产方式，已率先在**游戏、营销、电商、传媒、影视娱乐**等领域取得进展，伴随 AIGC 在各个行业的渗透，AIGC 作为 AI 数字商业的探路者，有望开启下一场数字商业模式的新篇章。

生成图片

媲美专业画师的精美图片

Stability.ai，Midjourney，OpenAI，RunwayML，Tiamat

生成文字

人机交互、写邮件、写广告、剧本和小说

ChatGPT – 对话形式人机交互；Copy.ai - 广告和营销文案；Jasper.ai - 营销推广文案及博客

生成音频

人工智能作曲 & 编曲、人工智能音乐生成、人工智能演唱、声音克隆人工智能音乐团队 Amper，人工智能播客 Podcast.ai，灵动音科技，行者人工智能

生成视频

文字生成视频、视频内容创作、动态面部编辑、画质增强修复

Make-A-Video(Meta)，Imagen Video(Google)， Phenaki (Google)， Synthesia ， Hour One

AIGC 典型行业应用场景

行业适用性

人工智能生成视频

- 广告 – 品牌宣传视频生成
- 媒体 – 影视预告片生成、特效制作
- 软件 – 视频换脸编辑

人工智能生成文字

- 游戏 – 游戏策略生成、营销文案生成
- 社交 – 营销文案生成
- 电商 – 营销文案生成、智能商品详情
- 媒体 – 影视剧本生成、视频脚本生成

人工智能生成音频

- 游戏 – BGM 生成、语音克隆
- 媒体 – 影视作品配乐

人工智能 生成图片

- 游戏 – IP 角色生成 、场景图片图标生成
- 社交 – 用户头像生成
- 电商 – 商品展示页配图、商品个性化图案设计
- 广告 – 营销素材生成、营销文案配图
- Web3 – NFT 生成
- 软件 – 去除水印、超分辨率、滤镜更替

技术实现成熟度

注：该应用场景是基于日常客户交流与市场洞察总结而成，其中，行业适应性反映的是同一应用场景在不同行业的普适性，技术实现成熟度反映的是产品落地和产品化的难易程度。

篇章二
AIGC 技术生态与典型客户需求

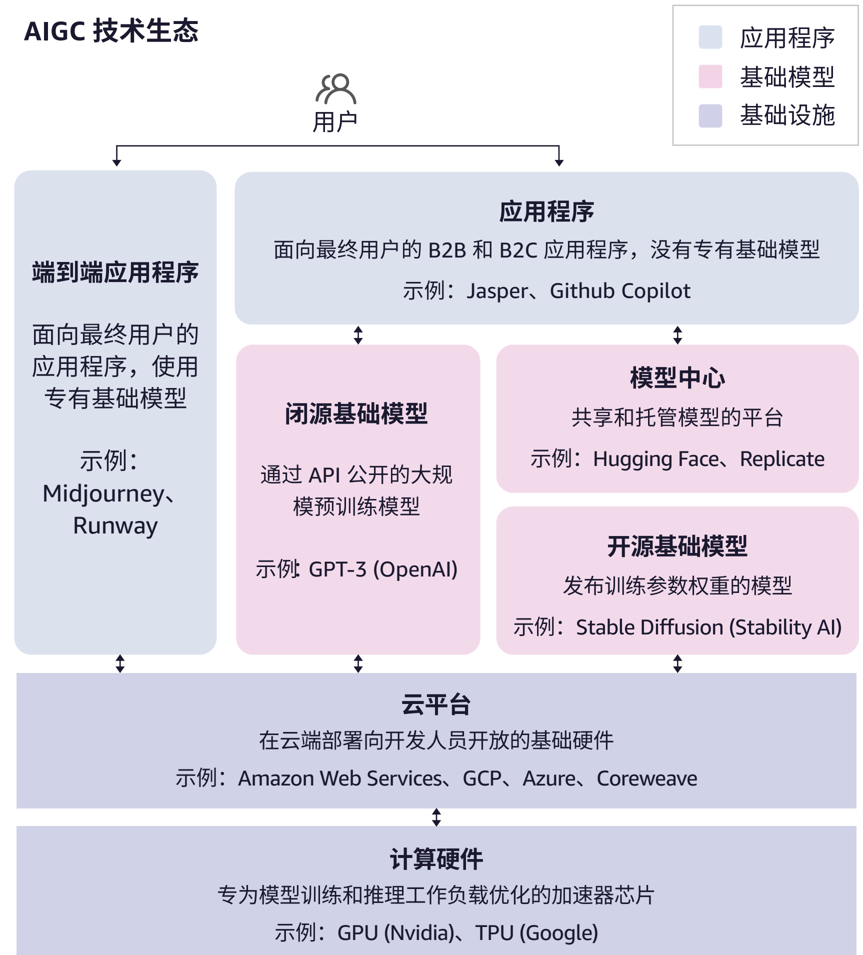
AIGC 技术生态

AIGC 技术生态加速形成与发展，目前整个 AIGC 应用的技术生态大致可以分为三层：分别为**基础设施、基础模型和应用程序。**

AIGC 技术生态：

- **将 AIGC 基础模型集成到面向用户的产品中的应用程序，**运行自己的模型管道（“端到端应用程序”）或依赖第三方 API，比如炙手可热的人工智能文本生成领域独角兽 Jasper.ai, 提供营销文案生成的 SaaS 服务如广告文案、博客、外发邮件等，人工智能绘画软件 Midjourney 等；
- **为 AIGC 应用提供支持的基础模型，**可以通过闭源专有 API （如 GPT-3）或开源模型 (如 Stable Diffusion)，或者提供开源模型托管平台（如 HuggingFace）；
- **为 AIGC 基础模型运行训练和推理工作负载的基础设施**供应商（即云平台和硬件制造商），如亚马逊云科技，英伟达等；

AIGC 技术生态



图片来源：A16Z: who-owns-the-generative-ai-platform/

需要注意的是，我们讲的并不是整个市场的生态图，而是一个分析市场的框架，本文在每个类别中都列出了一些知名厂商的例子，但没有囊括目前所有最先进的 AIGC 应用，也没有深入讨论 MLOps 或 LLMOps 工具，因为现在还没有达到完全成熟的标准化，有机会我们会继续讨论。

AIGC 基础模型生态与典型客户需求

作为云服务提供商，亚马逊云科技专注于为基础模型开发者和使用者提供全面、可靠、安全的工具与服务，助力 AIGC 应用开发加速与成本优化。基于此，我们将基础模型生态的**主要参与者**分为三类：**基础模型提供商** (Model Provider)，**使用基础模型调优提供服务的厂商** (Model Tuner) 和**直接使用基础模型厂商** (Model Consumer)，我们总结的每一类用户的需求和面临的主要挑战如下：

AIGC 应用构建者和使用者的主要需求

基础模型提供商 (Model Providers)	**使用基础模型调优提供服务的厂商 (Model Tuners)**	**直接使用基础模型厂商 (Model Consumers)**
从 0 到 1 构建基础模型	基于开源基础模型如 Stable Diffusion，大语言模型进行 fine-tune 以更好适用业务需求	直接部署开源基础模型进行推理或者调用闭源基础模型 API
技术门槛高，需要大量的金钱和时间成本	需要专有数据进行持续的模型优化工作	缺乏定制支持，若使用闭源 API 则无法控制数据

基础模型产业化所面临的主要技术挑战：

数据准备
- 数据规模大
- 数据质量参差不齐

训练调参
- 模型体积大
- 训练难度高

托管推理
- 推理延迟大
- 调用成本高

算力资源
- 算力规模大
- 性能要求高

安全合规
- 数据隐私安全
- 内容合规

篇章三
亚马逊云科技 AIGC 技术能力与解决方案

亚马逊云科技 AIGC 技术能力概览

20 多年来，人工智能和机器学习一直是亚马逊云科技关注的焦点，可以说，在机器学习领域的发明创新已经深刻在亚马逊云科技的 DNA 里。当前，用户在亚马逊云科技上使用的许多功能都是由其机器学习驱动的，比如电子商务推荐引擎、优化机器人拣选路线、无人机 Prime Air。还有语音助手 Alexa，这也得益于来自 30 多种不同的机器学习系统的支持，每周回应客户数十亿次管理智能家居、购物、获取信息和娱乐的请求。亚马逊有数千名工程师专注于机器学习研究，这既是我们的宝贵资产，也是我们现在最关注的理念和面向未来的实力之所在。

在亚马逊云科技，我们**致力于不断降低机器学习的使用门槛**。截至目前，我们已经帮助超过 10 万家来自各行各业的不同规模的客户使用机器学习进行创新。我们**在人工智能和机器学习堆栈的三个层级都拥有至深至广的产品组合**。长期以来，通过不断投入、持续创新，我们为机器学习提供**高性能、可伸缩的基础设施**和**极具性价比的机器学习训练和推理**；我们研发了 **Amazon SageMaker**，为所有开发人员构建、训练和部署模型提供最大的便利；我们还推出了大量服务，使客户通过简单的 API 调用就可添加 AI 功能到应用程序中，如图像识别、预测和智能搜索；同样，在 AIGC 技术上，亚马逊云科技也迈出了重要的一步，让这项技术也将赋能千行百业。亚马逊云科技能做的就是，**让更多客户能够访问基础模型能力、为机器学习推理和训练提供基础设施、提高所有开发人员的编码效率，帮助我们的客户更简单、更容易地在业务中使用 AIGC。**

亚马逊云科技人工智能与机器学习技术栈 最广泛、最完整的机器学习套件

人工智能服务

专用

业务流程优化
Amazon Personalize
Amazon Forecast
Amazon Fraud Detector
Amazon Lookout for Metrics

搜索
Amazon Kendra

对话
Amazon Lex
Amazon Transcribe Call Analytics
Contact Lens
Voice ID

代码 + DEVOPS
Amazon CodeGuru
Amazon CodeWhisperer
Amazon DevOps Guru

工业
Amazon Monitron
Amazon Lookout for Equipment
Amazon Lookout for Vision

生命健康
Amazon HealthLake
Amazon Comprehend Medical
Amazon Transcribe Medical
Amazon Omics

通用

文本
Amazon Translate
Amazon Comprehend

语音
Amazon Polly
Amazon Transcribe

视觉
Amazon Textract
Amazon Rekognition
Amazon Panorama

Amazon Bedrock

文本 Amazon Titan Text Amazon Titan Embeddings	**AI21** Jurassic-2	**Anthropic** Claude	**Stability AI** Stable Diffusion	**More…**

Amazon SageMaker

Canvas 无代码机器学习

Jumpstart 模型与解决方案库

GroundTruth 数据标记

CI/CD | 数据治理 | 负责任的人工智能

数据准备 Studio IDE 特征存储	地理信息 机器学习	Notebook 模型开发	训练模型 参数调优	部署至生产 管理和监控

EDGE MANAGER 管理边缘设备

机器学习框架与基础设施

PyTorch, Apache MXNet, TensorFlow	Amazon EC2	CPUs	GPUs	Amazon Inferentia	Amazon Trainium	Habana Gaudi	FPGA

面向不同的基础模型生态伙伴，亚马逊云科技提供了不同层次的产品与服务帮助用户提升开发效率，主要的产品和服务如下：

面向基础模型提供商，使用基础模型调优提供服务的厂商，直接使用基础模型厂商，提供全面深入的产品与服务

基础模型提供商

构建自己的基础模型

Amazon SageMaker

EC2 P4 instances
EC2 P3 instances
EC2 G4 instances

基于 Nvidia GPU 的 EC2 实例

Trn1
Inf2

基于亚马逊云科技自研 Trainium 和 Inferentia 芯片的 EC2 实例

使用基础模型调优提供服务的厂商

使用 Amazon SageMaker 基础模型 Hub—SageMaker JumpStart

JumpStart 提供的最先进的第三方预训练基础模型

LightOn
AI21labs
co:here
Hugging Face
alexa
stability.ai

OpenAI GPT-J

来自 HuggingFace 的 Bloom

直接使用基础模型厂商

AIGC 解决方案推理调用基础模型提供的 API

来自亚马逊的自研大语言基础模型

Titan Text
Titan Embeddings

来自 AI21 Labs、Anthropic 和 stability AI 的第三方基础模型

Jurassic-2
Claude
Stable Diffusion

Amazon Bedrock

Amazon CodeWhisperer

产品亮点 I ：

面向模型提供商提供适用于每种工作负载的高性能、经济高效、可扩展基础设施，尤其是**两款基于自研 AI 训练 (Trainium) 与推理 (Inferentia) 芯片专门针对 AIGC 应用优化的高性价比 EC2 实例 Trn1 和 Inf2，帮助企业大幅节省 AIGC 训练和推理的成本**

Amazon Trainium

Amazon Inferentia

无论**运行、构建还是定制基础模型**，客户都需要高性能、低成本且为机器学习专门构建的基础设施。亚马逊云科技提供基于**英伟达最新 GPU 芯片（如 H100, A100, A10, T4 等）**的虚机实例，满足用户对基础模型训练和微调的算力资源需求。除此之外，过去五年，亚马逊云科技持续加大在自研芯片方面的投入，不断突破性能和价格的极限，以支持对此有极高要求的机器学习训练与推理等工作负载。亚马逊云科技 **Trainium 和 Inferentia 芯片可以提供在云上训练模型和运行推理的最低成本。**正是因为我们在成本和性能方面的优势，像 AI21 Labs、Anthropic、Cohere、Grammarly、HuggingFace、Runway、Stability AI 等领先的 AI 初创公司都选择运行在亚马逊云科技平台上。

今天，基础模型花费的时间和金钱主要用于训练，这是因为许多客户才刚刚开始将基础模型部署到生产中。**由 Trainium 支持的 Trn1 计算实例与其他任何 Amazon EC2 实例相比，可以节省高达 50% 的训练成本，经过优化后可以在与高达 800Gbps 的第二代 EFA（弹性结构适配器）网络相连的多个服务器上分发训练任务。**客户可以在超大规模集群（UltraClusters）中部署 Trn1 实例，数量可以扩展到在同一可用区中 3 万个 Trainium 芯片，相当于超过 6 exaflops 的计算能力，并具有 PB 级网络。许多亚马逊云科技客户，包括 Helixon、Money Forward 和亚马逊的搜索团队，都使用 Trn1 实例将训练最大规模的深度学习模型所需的时间从几个月缩短到几周甚至几天，并且降低了成本。800 Gbps 的带宽已经很大，但我们仍不断创新、拓展带宽，推出全新的、网络优化型 Trn1n 实例，它可以提供 1600 Gbps 的网络带宽，专为大型网络密集型模型设计，其性能比 Trn1 高出 20%。

但是，未来，当基础模型进入大规模部署时，我们相信，大部分机器学习成本将来自运行推理。因此，我们推出了**由 Amazon Inferentia2 提供支持的 Inf2 实例，**这些实例专门针对运行数千亿个参数的基础模型驱动的 AIGC 应用程序进行了优化。与上一代相比，**Inf2 实例不仅吞吐量提高了 4 倍，延迟降低了 10 倍，还可实现加速器之间的超高速连接以支持大规模分布式推理。与同类 Amazon EC2 实例相比，这些能力将推理性价比提高了 40%，并把云中的推理成本降到最低。**

亚马逊云科技面向 AIGC 应用的专用加速芯片

Amazon Inferentia	Amazon Trainium	Amazon Inferentia2
在云端运行深度学习（DL）模型时单次推理的最低成本	大语言模型和 diffusion 模型最经济高效的训练方式	大语言模型和 diffusion 模型最经济高效的推理方式
与同类 Amazon EC2 实例相比，将推理性价比提高了 70%	**与同类 Amazon EC2 实例相比，将推理性价比提高了 50%**	**与同类 Amazon EC2 实例相比，将推理性价比提高了 50%**

产品亮点 II ：

面向基础模型提供商以及使用基础模型调优提供服务的厂商，提供全托管的一站式机器学习开发平台 Amazon SageMaker，助力用户高效实现 AIGC 基础模型的训练，推理，自定义、微调，部署和管理

Amazon SageMaker

面向机器学习工程化：Amazon SageMaker

无需自行构建、端到端机器学习平台，最大限度节省 AIGC 模型开发和应用成本，提升整体生产效率。

降低训练成本
内置竞价型实例训练模型
成本优化高达 **90%**

针对大规模的 AIGC
提供异步推理
算力自动伸缩最低至 **0**

内置 MLOps 套件
将 AIGC 实验**自动化**
部署至生产

针对基础模型训练

- Amazon SageMaker 可以轻松访问包括 **Nvidia GPU，Amazon Trainium** 在内的最新的基础设施资源，而且这些实例之间实现了超高速网络通信与高性能存储，方便算法人员聚焦模型调试的工作；
- Amazon SageMaker 提供了包括 **Studio，Notebook** 等一系列调试、分析以及追踪模型效果的工具，可以帮助算法人员尽快完成模型调整，此外 Amazon SageMaker 提供包含**数据标注、模型训练、微调、部署的标准化、自动化的端到端流程和管理工具，轻松实践机器学习运维 MLOps 和大规模集群协调；**
- 对 **TensorFlow、PyTorch 和 HuggingFace** 等框架和库进行了针对亚马逊云科技的优化，提供了显著的性能改善；
- Amazon SageMaker **自带分布式训练库**，支持**数据并行以及模型并行**（管道并行和张量并行）等模式，使得基础模型的分布式训练更加容易上手。除了 Amazon SageMaker 自带的分布式训练库外，还支持 **DeepSpeed 以及 FSDP 等开源分布式训练框架，**适配不同的客户需求；
- 超大规模基础模型训练成本节省：借助 Amazon SageMaker，您可以使用**托管式 Spot 竞价实例轻松训练机器学习模型**，与按需实例相比，**使用托管的 Spot 实例训练模型，可以将成本优化高达 90%。**

针对基础模型推理

- 针对不同的客户需求场景，Amazon SageMaker 支持多种推理方式，例如**实时推理**模式以满足超低时延的业务，批量推理以满足大型数据集的离线业务，**异步推理**以满足长时延的场景等；
- 相较利用 Amazon EC2 或 Amazon EKS，利用 Amazon SageMaker 进行模型部署和推理，可以省去对于计算实例、网络以及存储等基础设置的运营支出，减少运营成本；
- Amazon SageMaker 可以轻松将大语言模型进行模型**并行化处理**，并将模型切片放入单个 GPU 卡内存中，从而实现**单机多卡**模式下的推理，达到低至几百毫秒的推理延迟以及大规模的吞吐量；
- Amazon SageMaker 中的**大型模型推理容器（LMI）与 DeepSpeed、HFAccelerate** 等开源模型并行框架集成，此外还配备了 **BF16 量化能力**，有助于在不显著影响准确性的情况下缩小模型的大小，从而实现低延迟。

低代码构建机器学习模型：SageMaker JumpStart

内置包含 AIGC 场景在内的主流开源模型和算法库

1、“一键式”部署和 Fine-tune Stable Diffusion，Bloom、FLAN-T5、Alexa TM 等主流的 AIGC 基础模型；

2、内置 300+ 种开源模型，10+ 种预设场景解决方案；

3、提供支持 TensorFlow、PyTorch、Hugging Face 和 MXNet 等主流框架的先进 SOTA) 的开源模型；

4、用户可以通过 JumpStart 一键部署或微调众多预训练模型，轻松开发高质量模型并缩短部署时件；

5、支持通过应用场景、行业场景、模型名称以及资源类型进行搜索；

6、持续增加模型和场景。

目前在 SageMaker JumpStart 上可用的预训练基础模型

公开模型

Stability AI

Models
- Text2Image
- Upscaling

Features
- Fine-tuning on SD 2.1 model

支持的任务
- 输入文本生成图片
- 提升生成图片质量

AlexaTM

Models
- AlexaTM20B

支持的任务
- 机器翻译
- 注释生成
- 问答
- 数据生成
- 摘要生成

HuggingFace

Models
- FLAN T5 XL (8 variants)
- DistilGPT2

支持的任务
- 机器翻译
- 注释生成
- 问答
- 数据生成
- 摘要生成

专有模型

Co:here

Models
- Cohere generate Model - medium

支持的任务
- 文本生成
- 问答
- 信息提取
- 摘要生成

LightOn

Models
- Lyra-Fr 10B

支持的任务
- 文本生成
- 问答
- 情感分析
- 关键词提取
- 摘要生成
- 文本分类
- 信息提取

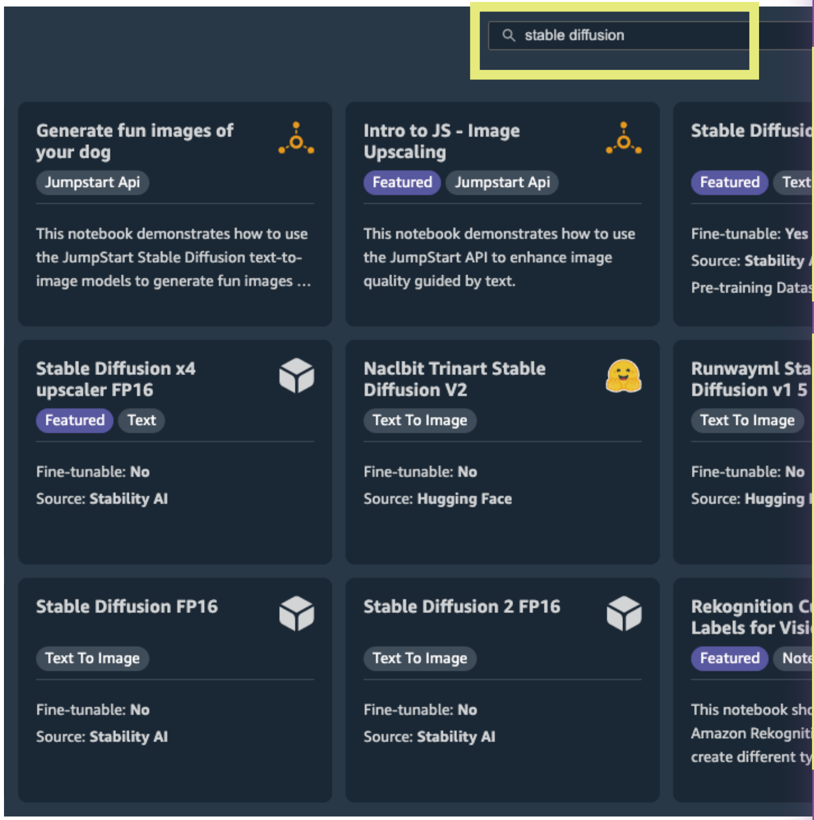
AI21

Models
- Jurassic-1 Grande 17B

支持的任务
- 文本生成
- 聊天
- 问答
- 长文章生成
- 信息提取
- 文本分类
- 摘要生成

SageMaker JumpStart

支持模型部署、训练（微调）



产品亮点 III：

面向使用基础模型调优服务厂商和直接使用基础模型的厂商，提供 Amazon Bedrock 服务和 Amazon Titan 大语言模型，助力构建低门槛，开放，安全的 AIGC 应用

Amazon Bedrock

针对**使用基础模型调优服务厂商和直接使用基础模型厂商**的主要需求：首先，他们需要能直接找到并访问高性能基础模型，这些模型需要能够给出最匹配业务场景的优秀反馈结果。其次，客户希望无缝与应用程序集成，且无需管理大量基础设施集群，也不会增加过高的成本。最后，目前基础模型定位是通用场景公用的能力，缺乏使用客户私有数据构建的差异化应用程序。客户希望能够轻松基于基础模型，利用自己的数据（可多可少）构建差异化的应用程序。由于客户进行定制化的数据是非常有价值的 IP，因此需要在处理过程中确保数据安全和隐私保护。同时，客户还希望能控制数据共享和使用。

基于客户以上的需求，我们推出了 **Amazon BedRock**。Amazon Bedrock 是客户使用基础模型构建和扩展生成式人工智能应用程序的最简单方法，为所有开发者降低使用门槛。凭借 Bedrock 所带来的无服务器体验，客户可以轻松找到适合自身业务的模型，快速上手，在确保数据安全和隐私保护的前提下，使用自有数据基于基础模型进行定制，并使用他们已经熟悉的亚马逊云科技工具和能力，将定制化模型集成并部署到他们的应用程序中，同时无需管理任何基础设施。

目前，Amazon Bedrock 包含两大类能供客户使用的基础模型，第一类为亚马逊自研的 Titan 模型，包括文本生成的 Titan Text 模型和做矢量编码的 Titan Embedding 模型；第二类为第三方合作伙伴的模型，包括 AI21 的 Jurassic-2，Anthropic 的 Claude 以及 Stability.ai 的 Stable Diffusion 模 型，Jurassic-2, Claude 模型为大语言模型，Stable Diffusion 模型为文本生成图片模型。任何规模的企业都可以通过 Amazon Bedrock 访问基础模型，加速机器学习在组织内部的应用，并凭借其轻松上手的特性，构建自己的生成式 AI 应用程序。我们相信，Amazon Bedrock 将是基础模型普惠化进程中的一大进步。

Amazon Bedrock 另外一个优势是**极其容易定制模型**。客户只需向 Amazon Bedrock 展示 Amazon S3 中的几个标注好的数据示例，Amazon Bedrock 就可以针对特定任务微调模型，最少仅需 20 个示例即可，而无需标注大量数据。假设一位时装零售行业的内容营销经理，想为即将推出的手提包新品系列开发新的、针对目标用户的广告创意。他向 Amazon Bedrock 提供了一些标注过的表现最佳的既往营销广告示例，以及新品的相关描述，Amazon Bedrock 将能自动为这些新品生成有效的社交媒体推文内容、展示广告和产品网页。任何客户数据不会被用于底层模型的训练，所有数据都将被加密，且不会离开客户的虚拟私有网络（VPC），可以确保全方位的数据安全和隐私保护。

Amazon Bedrock 的主要优势

- 通过 API 加速开发使用基础模型的生成式人工智能应用，而不需要管理基础设施
- 从 AI21 Labs、Anthropic、Stability AI 以及亚马逊云科技自研大语言模型中进行选择，找到合适客户案例的基础模型
- 使用客户的私有数据定制基础模型
- 利用全面的亚马逊云科技安全功能加强对客户的数据保护
- 使用客户熟悉的亚马逊云科技工具和功能来部署可扩展、可靠且安全的生成式人工智能应用

Amazon Bedrock 支持广泛的基础模型

来自亚马逊的自研大语言基础模型

Titan Text　Titan Embeddings

来自 AI21 Labs、Anthropic 和 stability Al 的第三方基础模型

Jurassic-2　Claude　Stable Diffusion

产品亮点 IV:

面向直接使用基础模型的代码编写者提供 AI 代码助手 Amazon CodeWhisperer，面向所有个人用户免费开放，不设任何资质或使用时长的限制，助力更快捷、更安全地构建应用程序

Amazon CodeWhisperer

我们预见到，**编程将是生成式 AI 技术得到快速应用的领域之一**。今天，软件开发者需要花费大量时间编写相当浅显和无差别的代码。他们还需要花费不少时间学习复杂的新工具和技术，而这些工具和技术总在不断演进。因此，开发者真正用于开发创新的功能与服务的时间少之又少。为应对这一难题，开发者会尝试从网上复制代码片段再进行修改，但可能无意中就复制了无效代码，有安全隐患的代码，或对开源代码的使用没有进行有效的追溯。而且这种搜索和复制的方式也浪费了开发者用于业务构建的时间。

AIGC 可以通过“编写”大部分无差别的代码来大大减少这种繁重的工作，让开发人员能够更快地编写代码，同时让他们有时间专注在更具创造性的编程工作上。因此，我们推出了 **Amazon CodeWhisperer**，一款 AI 编程助手，通过内嵌的基础模型，可以根据开发者用自然语言描述的注释和集成开发环境（IDE）中的既有代码实时生成代码建议，从而提升开发者的生产效率。

使用 CodeWhisperer 生成不同编程语言的代码示例：

```python
import boto3
from botocore.exceptions import ClientError


# Function to upload a file to an S3 bucket
def upload_file(file_name, bucket, object_name=None):
    """Upload a file to an S3 bucket

    :param file_name: File to upload
    :param bucket: Bucket to upload to
    :param object_name: S3 object name. If not specified then file_name is used
    :return: True if file was uploaded, else False
    """

    # If S3 object_name was not specified, use file_name
    if object_name is None:
        object_name = file_name

    # Upload the file
    s3_client = boto3.client('s3')
    try:
        response = s3_client.upload_file(file_name, bucket, object_name)
    except ClientError as e:
        logging.error(e)
        return False
    return True
```

Python 代码示例

```java
import java.io.*;
import org.json.JSONObject;
import org.json.XML;

// Function to convert JSON format to XML format
public class JSONtoXML {
    public String JSONtoXML(string jsonString) {
        // Create a JSONObject from the JSON String
        var jsonObject = new JSONObject(jsonString);

        // Convert the JSONObject to XML
        var xmlString = XML.toString(jsonObject);

        return xmlString;
    }
}
```

Java 代码示例

```javascript
import React from "react";
import ReactDOM from "react-dom";

const products = ["Apple", "Orange", "Peach"];
const prices = [1.5, 2, 3.5];

// Function to render products and prices in HTML divs
function renderProducts(products, prices) {
  const productElements = products.map((product, index) => {
    return (
      <div key={index}>
        {product} - ${prices[index]}
      </div>
    );
  });
  return productElements;
}
```

JavaScript 代码示例

目前，**Amazon CodeWhisperer 对所有个人用户免费，并不设任何资质或使用时长的限制！任何人都可以通过邮箱账户在几分钟内注册 Amazon CodeWhisperer 进行使用，而无需亚马逊云科技账号。对于企业客户，我们则提供了 Amazon CodeWhisperer 专业版，其中包括更多高级管理功能，如集成了身份与访问管理服务 (IAM) 的单点登录 (SSO)，以及使用更高限额的安全扫描。**

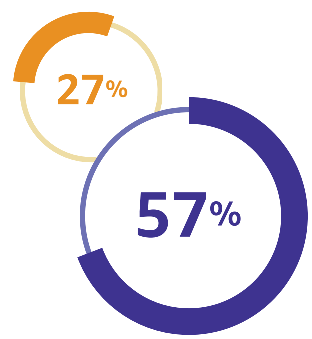
构建像 Amazon CodeWhisperer 这样强大的应用程序对开发人员和我们所有的客户来说都是变革性的。我们还有更多创新性的产品在规划中，也期待更多的客户和开发者在亚马逊云科技上构建更加创新和颠覆性的生成式人工智能应用。我们的使命是，让各种技能水平的开发人员和各种规模的组织都有机会使用生成式 AI 进行创新。我们相信，新一波机器学习技术创新才刚刚开始、方兴未艾，未来还有无限可能。

Amazon CodeWhisperer： 目前正式上线，并免费提供给个人开发者使用！

- 实时生成代码建议
- 扫描代码以发现隐蔽漏洞
- 标记与开源训练数据相似的代或默认为过滤器

在预览期间，亚马逊云科技开展了一项效率挑战，使用 Amazon CodeWhisperer 的参与者**成功完成任务的可能性增加 27%**，并且比不使用 Amazon CodeWhisperer 的

参与者
成功完成任务的可能性增加



参与者
平均加快的效率

亚马逊云科技 AIGC 参考解决方案介绍：

1. 基于 Amazon SageMaker Notebook 快速搭建托管的 Stable Diffusion – AI 作画可视化环境 WebUI

方案优势

1、本方案使用 **Amazon CloudFormation 一键部署（15~20 分钟）**，仅需点击几个步骤，即可快速在您的亚马逊云科技账户中成功部署一套 AIGC 图片方向的 web 应用环境，交互界面简洁友好，帮助用户快速体验文本生成图片、图片生成图片等人工智能服务。

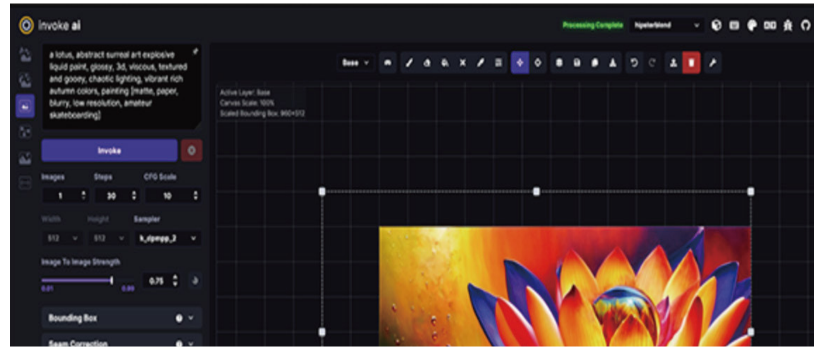
2、底层使用 Amazon SageMaker Notebook Instance 托管的 Jupyter 笔记本实例，**您无需关心基础环境**（如存储、网络等）的构建及底层基础设施运维。

3、方案采用全托管的 Stable Diffusion AI 模型服务（1.5 轻量版本，包含 **Stable Diffusion WebUI 与 InvokeAI 两个界面**），具备良好的开源项目使用体验，并支持根据您的需要安装插件以扩展使用场景（如 ControlNet）。

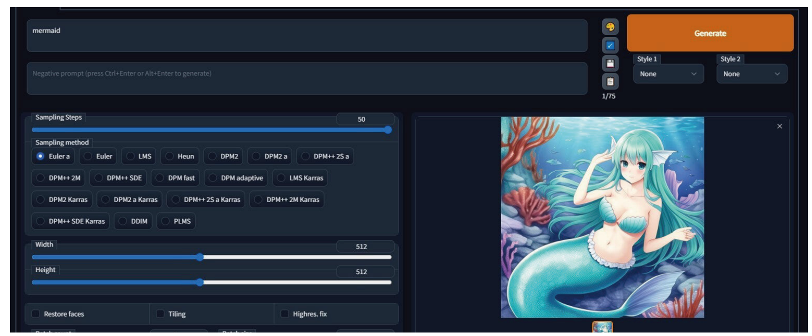
4、用户**可使用自带的图片数据微调模型**，产出的模型可基于自动化流水线工具快速部署上线至推理节点，方便规模化的图片推理调用。

5、本方案完全开源，用户可以在 Amazon SageMaker Notebook 中对模型和扩展（extension）定制开发，满足自身业务需求。

提供多种 UI 界面



InvokeAI



AUTOMATIC1111

2. 基于 Amazon SageMaker 构建的 AIGC 解决方案 -Industry AI

方案优势

1. 界面友好，基于 **Stable Diffusion webUI 优化**，提供了 WebUI，REST APIs 和 Notebook，一站式完成模型训练和部署任务；

2. 支持定制化开发，自主选择模型（Stable Diffusion, StyleGAN2 等），支持 SD1.X/SD2.X 模型和模型切换；

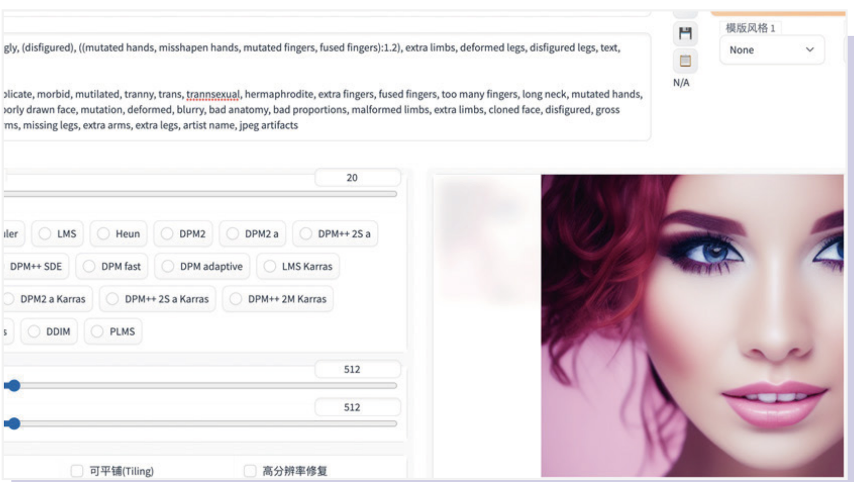
3. 全托管，自动弹性伸缩，前端采用 Amazon ECS 部署，**训练和推理采用 Amazon SageMaker 全托管机器学习平台**；

4. 支持 Embedding，Hypernetwork， Dreambooth，ControlNet 多种方式来训练对象或风格；

5. 支持**多用户管理**，方便配置后端推理和训练资源；

6. 效率高，无需编写代码，即可体验文本生成图像模型，使用 ml.g4dn.xlarge 机型，**在 3 秒内可生成 1 张效果图 (512x512), 图片数量及分辨率均可设置**；

7. 适合对 AIGC 有不同群体需求的企业客户，包括**算法工程师、应用开发者，画师 / 运营人员**等。



3. 亚马逊云科技 Stable Diffusion Extension

方案优势

通过为**社区提供插件和云资源模版**的方式，帮助客户将现有 Stable Diffusion 的模型训练，推理和调优等任务负载从本地服务器迁移至 Amazon SageMaker，利用云上弹性资源加速模型迭代，避免单机部署所带来的性能瓶颈。

1. **安装便捷**。本解决方案使用 Amazon CloudFormation 一键部署亚马逊云科技中间件，搭配**社区原生 Stable Diffusion WebUI 插件**安装形式一键安装，即可赋能用户快速使用 Amazon SageMaker 云上资源，进行推理和训练工作；

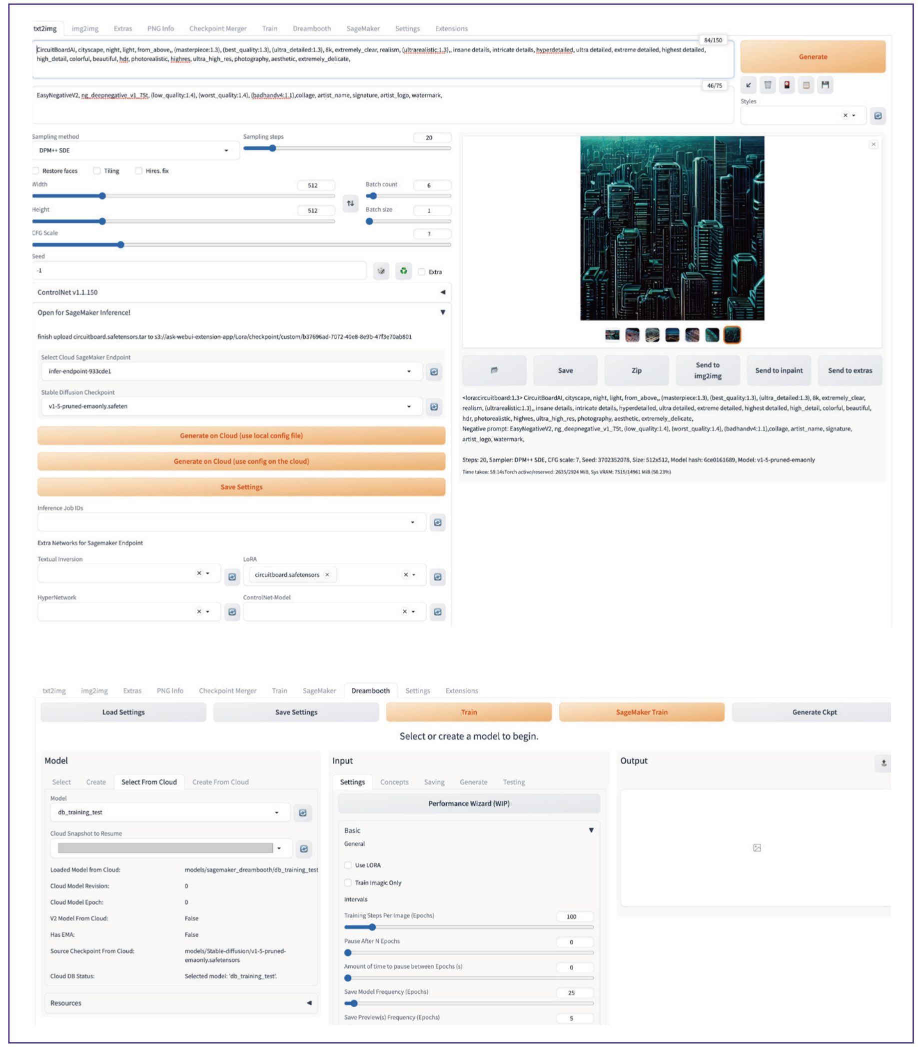
2. **社区原生**。WebUI 界面与后端分离，用户无需改变现有 Stable Diffusion WebUI 的使用习惯，WebUI 可以在任何支持的终端启动而没有 GPU 的限制，原有训练，推理等任务通过插件所提供的功能迁移到 Amazon SageMaker；

3. **可扩展性强**。方案插件以及中间件代码开源，采取非侵入式设计，有助于用户快速跟上社区相关功能的迭代，**从 WebUI 本体到广受欢迎的 Dreambooth、ControlNet、LoRa 等插件**。

4. **优化资源配置**。用户可按需选择云上资源，进行批量推理及模型训练，极大提升效率；

5. **协作性强**。依托于开源社区强大资源，通过插件形式可以与其他开发者合作，有助于更快速迭代产品，为用户提供更有用、易用的产品；

6. 适合对 AI 生图有业务需求的企业客户，包括**算法工程师，制作人，画师，设计师**等。



篇章四
AIGC 客户案例分享

四月科技

游戏·基于亚马逊云科技构建二次元 AIGC 产品 Anime AI

“过去我们需要将素材的设计外包给原画团队，不仅制作周期长，也是企业不可忽视的一项业务成本。在亚马逊云科技上应用 AIGC 进行填色素材自动化创作之后，我们不仅可以根据玩家喜好生成更多的资源，也帮助企业节约了超过 60% 的素材外包设计成本。”

——成都四月科技有限公司 CEO 袁海林

项目背景

四月科技专注在移动端休闲游戏的研发和运营，其中多款二次元游戏产品获得市场认可。在 AIGC 逐渐成熟之后，四月科技 CEO 希望利用 AIGC 打造二次元绘画产品，并将 AIGC 用于二次元原创作品的创作，以降低人工绘画的成本。

项目挑战

1、为了抢占市场，客户希望在 **1 个月之内**完成整个项目的开发和上线；

2、客户只有移动端工程师，没有专业的运维和后端开发团队，也没有 AI 相关的技术经验。

我们的方案

利用 Stable Diffusion2.0 模型，抽取图生图和文字生图的 API 接口，利用 API Gateway 等无服务器架构作为应用后端，将 WebUI 包装成 BYOC 的模式部署到 Amazon SageMaker，并利用 Amazon SageMaker 的异步推理和内部队列实现高可扩展性和高可用性的架构。

最终成功

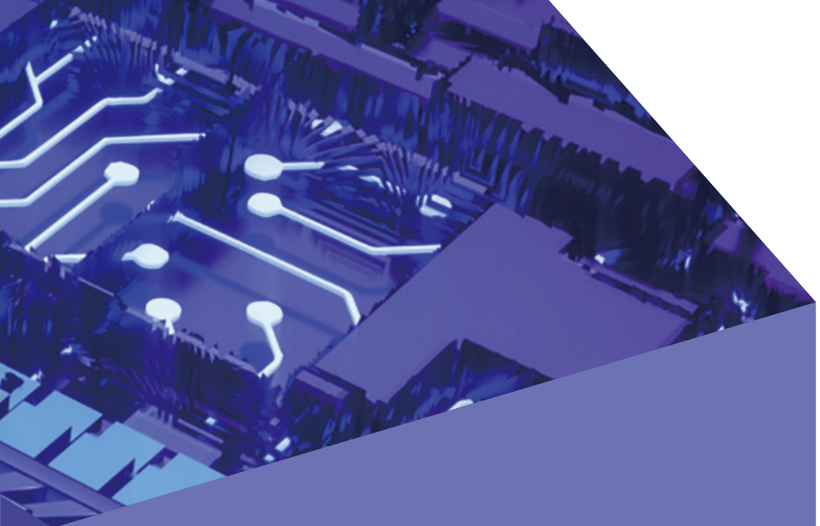
1. 最终产品在**一个月内准时**上线，目前已经获得**百万用户；**

2. 借助 AI 绘画功能，将填色游戏的图片素材**成本降低 60%；**

3.Anime AI 目前成为客户营收最高的业务，占公司总营收的 1/3。

利用 AI 绘画生成个性化二次元填色游戏的素材图片。人工成本降低 60%

利用 AI 绘画结合用户照片生成个性化的二次元风格头像



易点天下

电商 · 广告创意主题营销场景应用

"Amazon SageMaker 助力易点天下提高训练和调参的效率，整合模型训练交付成本下降 60% 以上，并实现在机器学习方面的运营成本节省超过了 75%。"

——易点天下网络科技股份有限公司

项目背景

针对电商客户，生成“穿戴”产品图的高质量 AI 模特，同时提供不同商业应用场景。帮助节省成本，提高商品展示多样性，提升成交总额。

项目挑战

1、广告营销团队没有相关实践经验，涉及内部团队多；

2、工程化解耦耗时长，没有整体解决方案；

3、无法预知未来 C 端用户流量。

我们的方案

上传多张不同角度眼镜产品图，基于 AIGC 技术，生成佩戴假发 / 眼镜且符合客户目标受众特点的 AI 模特图，用于客户站点产品主图展示。

在模型训练阶段，采用 Amazon SageMaker 机器学习平台，动态灵活从 0-1 训练机器学习模型，包含 ML 全生命周期能力。在推理阶段使用 Amazon EC2 结合 Auto Scaling 的能力，快速响应需求变化，实现动态扩缩容，推理成本降低 50%。 整体方案基于亚马逊云科技服务实现，无需管理基础设施，服务稳定可保障。

最终成功

1、上架 AI 模特产品图，某客户订单数据提升至原有 8 倍 +。

2、采用 AI 广告创意素材，某客户 CTR 提升 35.3%，CPC 降至 44.8%。

项目背景

Canva 是一个在线平台，用于创建和编辑从演示文稿到社交媒体帖子、视频、文档，甚至网站的所有内容。该公司的目标是让内容创作民主化，让每个人，从企业到规模最小的博主，都能使用先进的视觉传播工具。

项目挑战

1. 该公司希望推出一项基于 Stable Diffusion 的可以让用户输入文本提示，并获得人工智能生成的图像的新功能，但独自完成这项工作需要至少 6 个月的时间和大量的 GPU；
2. 对于某些 AIGC 的新模型，短期内快速交付是一个沉重的负担，在亚马逊云科技之前，Canva 无法快速交付大型、现代、前沿的模型；
3. Canva 关注的不仅仅是上市速度，更重要的是用户信任和安全，人工智能生成艺术的出现，为用户创造不确定的内容带来了新的方式，在某些情况下，这些 AI 甚至可能自行创建攻击性图像。

我们的方案

Canva 设置其图像创建序列，以便在用户输入文本提示后，使之使用 **Amazon SageMaker Real-Time Inference（实时推理）端点**来生成图像。当图像生成时，系统**通过 Amazon Rekognition 模型对其进行过滤**。在管道的末端，Canva 会将图像展示给最终用户，供其选择。借助这种前沿的文本到图像技术，用户可以在几秒钟内创建独特的高质量图像，而不是几小时或几天。

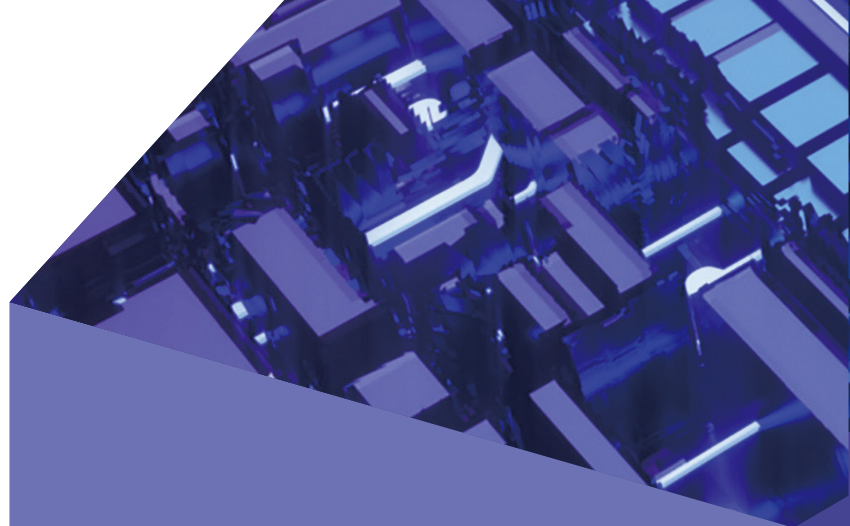
https://aws.amazon.com/cn/solutions/case-studies/canva-sagemaker-case-study/?nc1=h_ls

最终成功

通过使用 Amazon SageMaker，Canva 可以在 3 周内将新的文本到图像功能交付给用户；

使用 Amazon SageMaker 制作超过 60 个 ML 模型；

借助这种前沿的文本到图像技术，用户可以在几秒钟内创建独特的高质量图像。



Canva

可画

使用 Amazon Sagemaker 快速生成高质量图像

“我们在使用亚马逊云科技后 Canva ML 环境在扩展到大量用户方面做得很好，帮助我们实现了 AIGC 新模型的快速交付，交付实践从 6 个月缩短到 3 周以内。”

—— Glen Pink | Canva 的 ML 总监

stability.ai

借助亚马逊云科技实现功能强大的多模态机器学习

“我们于 2021 年开始与亚马逊云科技合作，使用 Amazon EC2 P4d 实例构建的文本到图像扩散模型 StableDiffusion，我们将该模型部署在大规模环境下，将模型训练时间从数月缩短到数周。并且在使用第二代 EC2 UltraCluster 中的 Amazon EC2 P5 实例后我们预计 P5 实例会进一步将我们的模型训练时间缩短 4 倍，从而使我们能够以更低的成本更快地提供突破性的 AI。”

—— Emad Mostaque | Stability AI CEO

项目背景

Stability AI 是一家总部位于加利福尼亚州的初创公司，专注于授权超过 20 万名成员的开发者社区为语言、音频、视频、3D 和生物构建开放的人工智能模型。

项目挑战

Stable Diffusion 这样的模型训练起来非常困难，需要使用数千个 GPU 或 Amazon Trainium 机器学习训练专用芯片。

我们的方案

通过使用 Amazon Sagemaker 托管的基础设施和优化库，Stability AI 能够使其模型训练具有更高韧性和性能

模型

- 多模态模型。NLP+ 图像；
- Stable Diffusion 是一种文本到图像的模型，可使数十亿用户在几秒钟内创作令人惊叹的艺术作品；
- 该模型是在 4000 个 A100 超级集群训练的，这是后续一系列基础模型使用该方式训练的首次探索；

训练

- 计算：Amazon EC2 P4d 实例；
- 规模：4000 个 GPU，用于 1 个训练工作；
- 业务流程：Amazon EKS；
- PyTorch 库：Torch. nn.DataParallel

推理

- 计算：Amazon EC2 G5 实例；
- 业务流程：Kubernetes；
- 存储：Amazon S3；
- PyTorch 库：TorchServe；

最终成功

1、能够在 Amazon EC2 P4d 实例上训练基于 PyTorch 的大规模机器学习模型，利用 Amazon FSx for Lustre 和 Amazon Batch，借助云规模对 GAN 计算机视觉和转换器模型进行分布式训练。

2、使用云服务拓展训练的能力将这些大型模型的训练时间从几个月缩短到几天，并将其发布到开源社区。

3、使用 Amazon SageMaker 及其模型库，Stability AI 可减少 58% 的训练时间和成本。

项目背景

AI21 Labs 成立于 2017 年，总部位于以色列特拉维夫，开发专注于语义和上下文的大规模语言模型，并通过旗舰产品 Wordtune 提供基于人工智能的写作辅助。

项目挑战

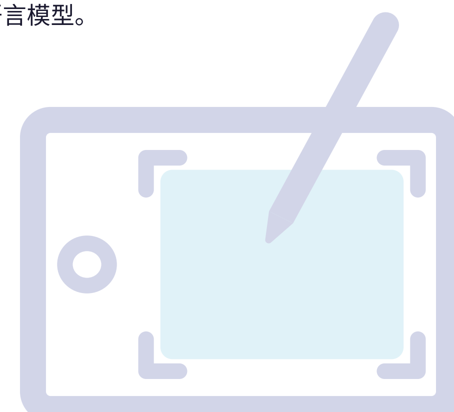
在训练 Jurasci-1 Jumbo（一种具有 1780 亿个参数的自回归语言模型）时，AI21 Labs 希望实现功能强大的计算和网络功能，并最大限度地提高效率。

我们的方案

该公司使用 Amazon EC2 P4d 实例，通过在数百个 GPU 中分发模型训练，获得所需的性能和内存，从而提供自然语言处理即服务。

最终成功

- 开发一个包含 1780 亿个参数和 25.6 万个词汇的语言模型。
- 高效且经济地扩展到数百个 GPU。
- 为大规模开发模型积累知识。

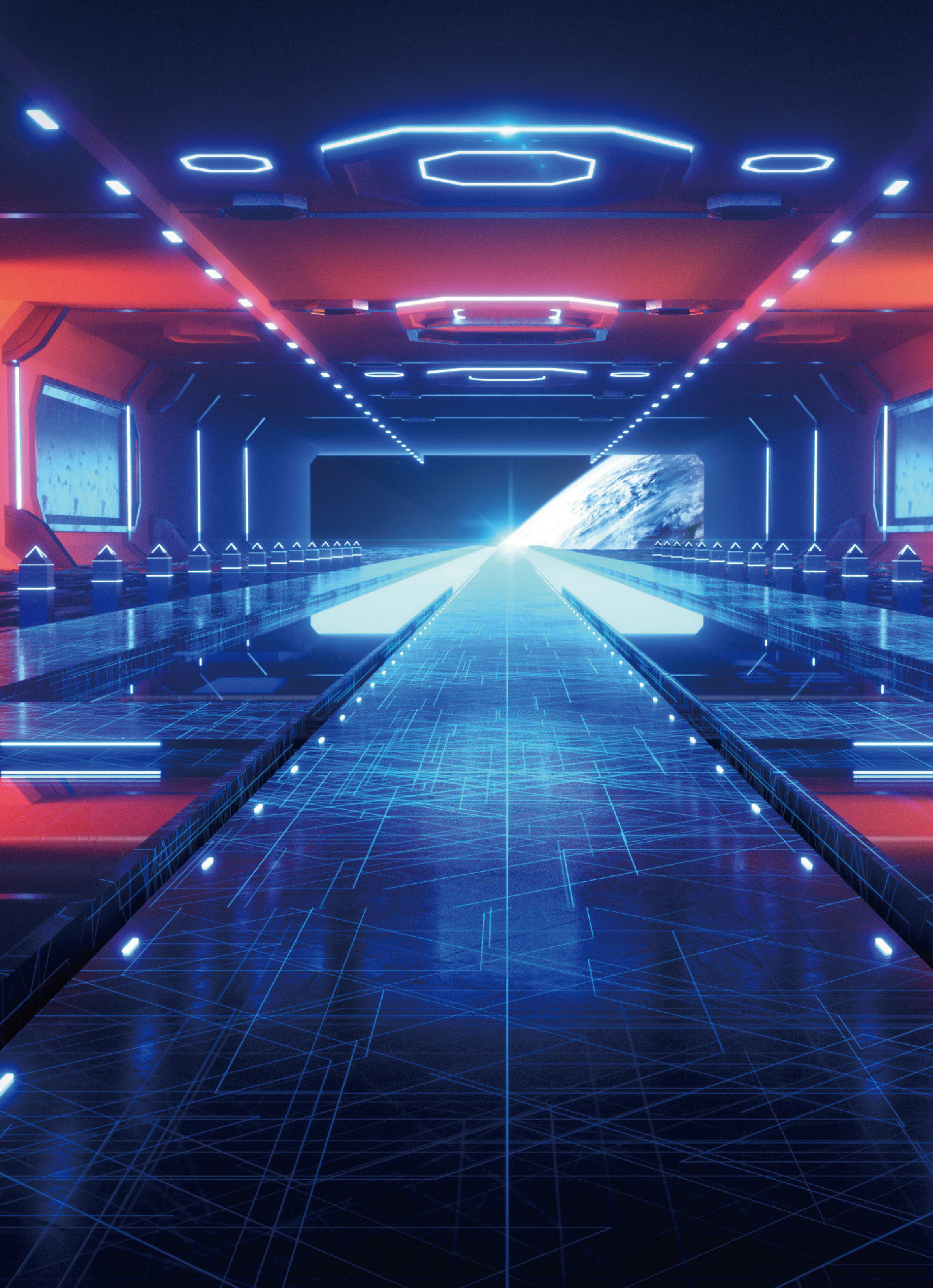


AI21 labs

使用 Amazon EC2 P4d 实例—PyTorch 训练，**1780** 亿参数语言模型 Jurasci-1 Jumbo。

“**Amazon EC2 P4d 实例在 EFA 上提供 400 Gbps 高性能网络**。GPU-GPU 的联网速度直接影响了在扩展数百个 GPU 时，高效扩展和保持成本效益的能力。”

—— Opher Lieber | AI21 Labs Jurassic 技术领导



致 谢

主编人员

宋洪涛

亚马逊云科技 资深产品市场经理

杨佳欢

亚马逊云科技 人工智能与机器学习产品经理

张　洋

亚马逊云科技 产品总监

李　昕

亚马逊云科技 产品市场总监

赫祎诺

亚马逊云科技 人工智能与机器学习产品经理

想要快速在亚马逊云科技平台上玩转 AIGC 动手实验
↓↓↓ 请扫码或点击以下技术博客 ↓↓↓

初级

- 基于 SageMaker Notebook 快速搭建托管的 Stable Diffusion – AI 作画可视化环境
- 在 Amazon SageMaker 使用 HuggingFace Diffusers 快速运行 Stable Diffusion 模型

高级

- 使用 Dreambooth 进行模型微调在 SageMaker 上的优化实践
- 基于 Amazon SageMaker 使用 Grounded-SAM 加速电商广告素材生成
- 在 SageMaker 中进行 LoRA fine tuning 及推理

如果您有任何问题，欢迎拨打亚马逊云科技热线电话

亚马逊云科技海外区域: 1010 0866
亚马逊云科技中国（宁夏）区域-由西云数据运营: 1010 0966
亚马逊云科技中国（北京）区域-由光环新网运营: 1010 0766

- 1 键-申请账号及产品咨询
- 2 键-云创计划及联合创新中心
- 3 键-账号账单问题
- 4 键-备案咨询（仅由（宁夏）区域和（北京）区域热线支持）
- 5 键-培训与认证
- 6 键-市场活动查询
- 7 键-亚马逊云科技合作伙伴网络（仅由海外区域热线支持）
- 8 键-Marketplace 产品咨询-仅由（宁夏）区域热线支持

扫码或点击了解更多亚马逊云科技
AIGC 技术能力与解决方案

亚马逊云科技

亚马逊云科技

免责声明*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。