AWS 决策指南

选择 AWS 机器学习服务

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

选择 AWS 机器学习服务: AWS 决策指南

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或者贬低或诋毁 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

决策指南 1
简介 1
明白 2
考虑一下 3
选择 6
使用 8
Explore 19
资源 19
文档历史记录 21
..... xxii

选择 AWS 机器学习服务

选择合适的机器学习服务和框架来支持你的工作

目的	帮助确定哪些 AWS 机器学习服务最适合您的需求。
上次更新	2024 年 5 月 3 日
承保服务	• Amazon Augmented AI • Amazon CodeGuru • Amazon Comprehend • Amazon DevOps Guru • Amazon Forecast • Amazon Kendra • Amazon Lex • Amazon Personalize • Amazon Polly • Amazon Rekognition • 亚马逊 SageMaker AI • Amazon Textract • Amazon Transcribe • Amazon Translate

简介

最基本的是，机器学习（ML）旨在提供数字工具和服务，用于从数据中学习、识别模式、做出预测，然后根据这些预测采取行动。当今几乎所有的人工智能 (AI) 系统都是使用机器学习创建的。机器学习使用大量数据来创建和验证决策逻辑。这种决策逻辑构成了人工智能模型的基础。

可以应用 AWS 机器学习服务的场景包括：

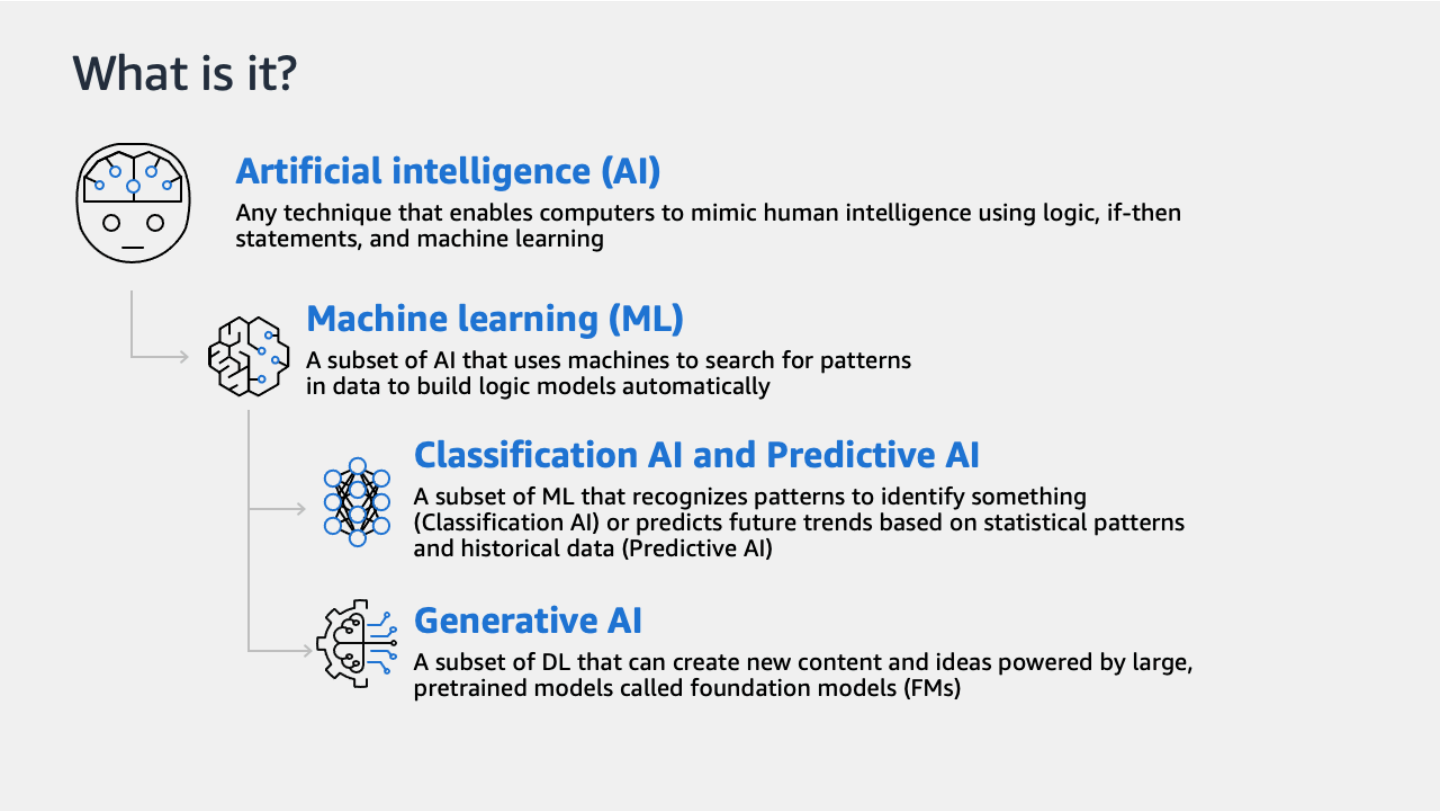
- 特定用例 — AWS 机器学习服务可以为你的 AI 驱动的用例提供广泛的预建算法、模型和解决方案，为常见用例和行业提供支持。您可以选择 23 种预先培训的服务，包括 Amazon Personalize、Amazon Kendra 和 Amazon Monitron。
- 自定义和扩展机器学习 — Amazon SageMaker AI 旨在帮助您构建、训练和部署适用于任何用例的机器学习模型。您可以 AWS 通过 Amazon A SageMaker I 和 Amazon Bedrock 构建自己的开源基础模型或访问开源基础模型。
- 访问专用基础架构 — 如果您需要更高的灵活性和对机器学习工作流程的控制，并且愿意自己管理底层基础设施和资源，请使用提供的机器学习框架和基础架构。 AWS

本决策指南将帮助您提出正确的问题，评估您的标准和业务问题，并确定哪些服务最适合您的需求。

明白

随着组织继续采用人工智能和机器学习技术，理解和选择 AWS 机器学习服务的重要性已成为一项持续的挑战。

AWS 提供了一系列机器学习服务，旨在帮助组织更快速、更轻松地构建、训练和部署机器学习模型。这些服务可用于解决各种业务问题，例如客户流失预测、欺诈检测以及图像和语音识别。



在深入研究 AWS 机器学习服务之前，让我们来看看 AI 和 ML 之间的关系。

- 简而言之，人工智能是一种描述任何可以复制以前需要人类智能的任务的系统的方式。 大多数 AI 用例都在寻找概率结果，即做出与人类判断相似的高度确定性预测或决策。
- 如今，几乎所有的人工智能系统都是使用机器学习创建的。机器学习使用大量数据来创建和验证决策逻辑，即所谓的模型。
- 分类 AI 是机器学习的一个子集，它可以识别模式以识别某物。预测性人工智能是机器学习的一个子集，它根据统计模式和历史数据预测未来的趋势。
- 最后，生成式人工智能是深度学习的一个子集，它可以创造新的内容和想法，例如对话、故事、图像、视频和音乐。生成式人工智能由超大型模型提供支持，这些模型是在庞大的数据库上进行预训练的，称为基础模型或。 FMs Amazon Bedrock 是一项完全托管的服务，它 FMs 为构建和扩展生成式 AI 应用程序提供了多种高性能选择。 Amazon Q D eveloper 和 Amazon Q Busines s 是基于 Generative AI 的助手，适用于特定用例。

本指南主要涵盖分类人工智能和预测人工智能机器学习类别中的服务。

此外，还为高性能机器学习训练和推理 AWS 提供专门的加速硬件。

- Amazon EC2 P5 实例配备了 NVIDIA H100 Tensor Core GPUs，非常适合机器学习中的训练和推理任务。 Amazon EC2 G5 实例配备多达 8 个 NVIDIA A10G Tensor Core GPUs 和第二代 AMD EPYC 处理器，适用于各种图形密集型和机器学习用例。
- AWS Trainium是第二代机器学习加速器，专 AWS 为深度学习 (DL) 训练超过 1000 亿个参数模型而设计。
- AWS Inferentia基于 2 的 Amazon EC2 Inf2 实例旨在在亚马逊以最低的成本 EC2 为您的深度学习和生成式 AI 推理应用程序提供高性能。

考虑一下

在使用 AWS 机器学习服务解决业务问题时，考虑几个关键标准有助于确保成功。以下部分概述了选择机器学习服务时需要考虑的一些关键标准。

Problem definition

问题定义

机器学习生命周期的第一步是构思业务问题。了解您要解决的问题对于选择正确的 AWS 机器学习服务至关重要，因为不同的服务旨在解决不同的问题。确定机器学习是否最适合您的业务问题也很重要。

一旦确定机器学习最合适，就可以从一系列专门构建的 AWS AI 服务（语音、视觉和文档等领域）中进行选择。

如果您需要构建和训练自己的模型， SageMaker Amazon AI 可提供完全托管的基础设施。 AWS 提供了一系列高级机器学习框架和基础架构选择，以满足您需要高度定制和专业化的机器学习模型的情况。 AWS 还提供了一系列流行的基础模型，用于使用生成式 AI 构建新应用程序。

ML algorithm

机器学习算法

为你想要解决的业务问题选择机器学习算法取决于你正在处理的数据类型以及预期的结果。以下信息概述了每个主要的 AWS AI/ML 服务类别如何使您能够使用其算法：

- 专门的 AI 服务：这些服务提供的自定义机器学习算法的能力有限，因为它们是针对特定任务进行了优化的预训练模型。通常，您可以自定义输入数据和某些参数，但无法访问底层机器学习模型，也无法构建自己的模型。
- Amazon SageMaker AI：该服务为机器学习算法提供了最大的灵活性和控制力。您可以使用 SageMaker AI 使用自己的算法和框架来构建自定义模型，也可以使用提供的 AWS预建模型和算法。这允许对机器学习流程进行高度的自定义和控制。
- 较低级别的机器学习框架和基础架构：这些服务为机器学习算法提供了最大的灵活性和控制力。您可以使用这些服务使用它们自己的算法和框架来构建高度自定义的机器学习模型。但是，使用这些服务需要大量的机器学习专业知识，可能并非适用于所有用例。

Security

安全性

如果您需要在 VPC 中使用私有终端节点，则您的选项将根据您使用的 AWS 机器学习服务层而有所不同。这些指令包括：

- 专业 AI 服务：大多数专业的 AI 服务目前不支持中的私有终端节点 VPCs。但是，可以使用 VPC 终端节点访问亚马逊 Rekognition 自定义标签和亚马逊 Comprehend Custom。
- 核心人工智能服务：Amazon Translate、Amazon Transcribe 和 Amazon Comprehend 都支持 VPC 终端节点。

- Amazon SageMaker A SageMaker I：AI 为 VPC 终端节点提供内置支持，允许您将经过训练的模型部署为只能从其 VPC 内部访问的终端节点。
- 较低级别的机器学习框架和基础设施：您可以在 Amazon EC2 实例或 VPC 内的容器中部署模型，从而完全控制联网配置。

Latency

延迟

更高级别的人工智能服务，例如Amazon Rekognition和Amazon Transcribe，旨在处理各种用例，并在速度方面提供高性能。但是，它们可能不符合特定的延迟要求。

如果您使用的是较低级别的机器学习框架和基础设施，我们建议您利用 Amazon A SageMaker I。由于其完全托管的服务和优化的部署选项，此选项通常比构建自定义模型更快。虽然高度优化的自定义模型可能优于 SageMaker 人工智能，但它需要大量的专业知识和资源来构建。

Accuracy

准确性

AWS 机器学习服务的准确性因具体用例和所需的自定义级别而异。更高级别的人工智能服务（例如 Amazon Rekognition）建立在预训练模型之上，这些模型已针对特定任务进行了优化，并且在许多用例中都具有很高的准确性。

在某些情况下，您可以选择使用 Amazon A SageMaker I，它为构建和训练自定义 ML 模型提供了更加灵活和可自定义的平台。通过构建自己的模型，您可能能够实现比预训练模型更高的精度。

您还可以选择使用机器学习框架和基础架构（例如 TensorFlow 和 Apache MXNet）来构建高度自定义的模型，从而为您的特定用例提供尽可能高的准确性。

AWS and responsible AI

AWS 还有负责任的人工智能

AWS 构建基础模型 (FMs)，在开发过程的每个阶段都要考虑负责任的人工智能。在设计、开发、部署和运营过程中，我们会考虑一系列因素，包括：

1. 准确性（摘要与基础文档的匹配程度；传记在事实上是否正确）
2. 公平，（产出是否以同样方式对待人口群体）
3. 知识产权和版权注意事项
4. 适当用法（过滤掉用户对法律建议、医疗诊断或非法活动的请求）

5. 毒性（仇恨言论、亵渎和侮辱）
6. 隐私（保护个人信息和客户提示）

AWS 将解决这些问题的解决方案融入用于获取训练数据的流程、 FMs 自身以及用于预处理用户提示和后处理输出的技术中。

选择

既然您已经知道了评估机器学习服务选项所依据的标准，就可以选择适合您组织需求的 AWS 机器学习服务了。下表突出显示了哪些机器学习服务针对哪些情况进行了优化。使用它来帮助确定最适合您的用例的 AWS 机器学习服务。

类别	你什么时候会用它？	它针对什么进行了优化？	相关 AI/ML 服务或环境
具体用例 这些人工智能服务旨在满足特定需求。它们包括个性化、预测、异常检测、语音转录等。由于它们是作为服务交付的，因此无需任何机器学习专业知识即可将其嵌入到应用程序中。	如果您需要将特定的预先构建的功能集成到应用程序中，而无需大量的自定义设置或机器学习专业知识，请使用所提供的 AI 服务。 AWS 这些服务经过精心设计，易于使用，不需要太多的编码或配置。	这些服务旨在易于使用，不需要太多的编码、配置或机器学习专业知识。	Amazon Augmented AI Amazon CodeGuru Amazon Comprehend Amazon Comprehend Medical Amazon DevOps Guru Amazon Forecast Amazon Kendra Amazon Lex Amazon Personalize Amazon Polly Amazon Rekognition

类别	你什么时候会用它？	它针对什么进行了优化？	相关 AI/ML 服务或环境
			Amazon Textract Amazon Transcribe Amazon Translate
机器学习服务 这些服务可用于开发定制的机器学习模型或工作流程，这些模型或工作流程超出了核心人工智能服务提供的预建功能。	当您需要更多自定义的机器学习模型或超出核心人工智能服务提供的预建功能的工作流程时，请使用这些服务。	这些服务针对构建和训练自定义机器学习模型、在多个实例或 GPU 集群上进行大规模训练、更好地控制机器学习模型部署、实时推理以及构建 end-to-end工作流程进行了优化。	亚马逊 SageMaker AI 亚马逊 SageMaker AI JumpStart SageMaker 人工智能工作室 SageMaker AI 画布 SageMaker AI 工作室实验室 SageMaker AI Ground Truth PyTorch on AWS Apache MxNet Hugging Face TensorFlow on AWS
基础设施 要在生产环境中部署机器学习，您需要经济实惠的基础设施，Amazon 通过 AWS内置芯片实现这种基础设施。	当您想要实现最低的训练模型成本并且需要在云端运行推理时使用。	经过优化，可支持经济高效的机器学习部署。	AWS Trainium AWS Inferentia 和推理2 亚马逊 SageMaker AI HyperPod

类别	你什么时候会用它？	它针对什么进行了优化？	相关 AI/ML 服务或环境
工具和相关服务 这些工具和相关服务旨在帮助您轻松部署机器学习。	这些服务和工具旨在帮助您加速云端的深度学习，提供 Amazon 机器映像、docker 映像和实体分辨率。	经过优化，可帮助您加速云端的深度学习。	AWS Deep Learning AMIs s AWS 深度学习容器 AWS 实体分辨率

使用

既然您已经清楚地了解了在选择 AWS 机器学习服务时需要应用的标准，那么您可以选择针对您的业务需求进行了优化的 AWS AI/ML 服务。

为了探索如何使用和进一步了解您选择的服务，我们提供了三组途径来探索每种服务的工作原理。第一组途径提供了深入的文档、动手教程和资源，供您开始使用亚马逊 Comprehend、Amazon Textract、Amazon Translate、Amazon Lex、Amazon Polly、Amazon Rekognition 和 Amazon Transcribe。

Amazon Comprehend

- 开始使用亚马逊 Comprehend

  使用 Amazon Comprehend 控制台创建和运行异步实体检测任务。

  开始阅读本教程”

- 使用 Amazon Comprehend 分析文本中的见解

  学习如何使用 Amazon Comprehend 来分析文本并从中得出见解。

  开始阅读本教程”

- 亚马逊 Comprehend 定价

  浏览有关亚马逊 Comprehend 定价和示例的信息。

  浏览指南”

Amazon Textract

- 亚马逊 Textract 入门

  了解 Amazon Textract 如何与格式化文本配合使用，以检测彼此靠近的单词和单词行，以及分析文档中的相关文本、表格、键值对和选择元素等项目。

  浏览指南”

- 使用 Amazon Textract 提取文本和结构化数据

  了解如何使用 Amazon Textract 从文档中提取文本和结构化数据。

  开始阅读本教程”

- AWS 力量时刻：Machine Learning

  在本集中深入了解 Amazon Textract AWS Management Console，花点时间阅读并查看代码示例，这些示例将帮助您了解如何充分利用服务。 APIs

  观看视频”

Amazon Translate

- 使用控制台开始使用 Amazon Translate

  Amazon Translate 入门最简单的方法是使用控制台翻译一些文本。学习如何使用控制台翻译多达 10,000 个字符。

  浏览指南”

- 在云端语言之间翻译文本

  在本教程示例中，作为一家国际箱包制造公司的一员，您需要了解客户在用当地市场语言（法语）撰写的评论中对您的产品的评价。

  开始阅读本教程”

- 亚马逊 Translate 定价

探索 Amazon Translate 的定价，包括免费套餐，在 12 个月内每月提供 200 万个字符。

浏览指南”

Amazon Lex

- 亚马逊 Lex V2 开发者指南

  浏览有关 Amazon Lex V2 的入门信息、工作原理和定价信息。

  浏览指南”

- Amazon Lex 简介我们向您介绍 Amazon Lex 对话服务，并向您介绍如何创建机器人并将其部署到不同的聊天服务的示例。

  参加课程”（需要登录）
- 在对话体验中探索生成式 AI

  探索生成式 AI 在对话体验中的应用。

  阅读博客”

Amazon Polly

- 什么是 Amazon Polly？

  浏览云服务的完整概述，该服务可将文本转换为逼真的语音，并可用于开发应用程序，以提高客户参与度和可访问性。

  浏览指南”

- 使用 Amazon Polly 突出显示正在朗读的文本

  我们将向您介绍在说话时突出显示文本的方法，以便为图书、网站、博客和其他数字体验中的音频添加视觉功能。

  阅读博客”

- 在 Amazon Polly 中使用相同的 TTS 语音角色为多种语言的内容创建音频

  我们将解释 Neural Text-to-Speech (NTTS)，并讨论如何使用广泛的可用语音组合，以支持的语言提供一系列不同的说话者。

  阅读博客”

Amazon Rekognition

- 什么是亚马逊 Rekognition？

  探索如何使用此服务将图像和视频分析添加到您的应用程序中。

  浏览指南”

- 动手 Rekognition：自动图像和视频分析

  了解面部识别如何与流媒体视频配合使用，以及代码示例和关键点。

  开始阅读本教程”

- Amazon Rekognition FAQs

  了解 Amazon Rekognition 的基础知识，以及它如何帮助您改进深度学习和直观分析应用程序。

  阅读 FAQs”

Amazon Transcribe

- 什么是 Amazon Transcribe？

  探索使用机器学习将音频转换为文本的 AWS 自动语音识别服务。了解如何将此服务用作独立转录或向任何应用程序添加 speech-to-text功能。

  浏览指南”

- 亚马逊 Transcribe 定价

我们向您介绍 AWS pay-as-you-go 转录，包括自定义语言模型选项和 Amazon Transcribe 免费套餐。

浏览指南”

- 使用 Amazon Transcribe 创建音频脚本

  了解如何使用 Amazon Transcribe 使用真实用例场景创建录制音频文件的文本脚本，以便根据您的需求进行测试。

  开始阅读本教程”

- 构建 Amazon Transcribe 直播应用程序

  学习如何开发一款用于实时录制、转录和翻译实时音频的应用程序，并将结果直接通过电子邮件发送给您。

  浏览指南”

第二组 AI/ML AWS 服务 途径提供了深入的文档、动手教程和资源，让您可以开始使用 Ama SageMaker zon AI 系列中的服务。

SageMaker AI

- 亚马逊 A SageMaker I 是如何运作的

  浏览机器学习概述以及 SageMaker AI 的工作原理。

  浏览指南”

- 亚马逊 A SageMaker I 入门

  了解如何加入亚马逊 A SageMaker I 域名，让你能够访问亚马逊 A SageMaker I Studio 和 RStudio SageMaker 人工智能。

  浏览指南”

- 将 Apache Spark 与亚马逊 A SageMaker I 配合使用

学习如何使用 Apache Spark 进行数据预处理，以及如何使用 SageMaker AI 进行模型训练和托管。

浏览指南”

- 使用 Docker 容器来构建模型

  探索 Amazon SageMaker AI 如何广泛使用 Docker 容器来执行构建和运行时任务。学习如何为其内置算法部署预构建的 Docker 镜像以及用于训练和推理的支持的深度学习框架。

  浏览指南”

- 机器学习框架和语言

  了解如何使用 Amazon SageMaker AI Python 软件开发工具包开始使用 SageMaker 人工智能。

  浏览指南”

SageMaker AI Autopilot

- 为表格数据创建 SageMaker Amazon AI 自动驾驶实验

  学习如何创建 Amazon A SageMaker I Autopilot 实验，以便在表格数据集中探索、预处理和训练各种候选模型。

  浏览指南”

- 自动创建机器学习模型

  学习如何使用 Amazon SageMaker AI Autopilot 自动构建、训练和调整机器学习模型，以及如何部署模型进行预测。

  开始阅读本教程”

- 使用这些示例笔记本探索使用 Amazon SageMaker AI AI Autopilot 进行建模

  探索用于直接营销、客户流失预测的示例笔记本，以及如何将自己的数据处理代码引入 Amazon A SageMaker I Autopilot。

浏览指南”

SageMaker AI Canvas

- 开始使用亚马逊 A SageMaker I Canvas

  了解如何开始使用 SageMaker AI Canvas。

  浏览指南”

- 无需编写代码即可生成机器学习预测

  本教程介绍如何使用 Amazon SageMaker AI Canvas 构建机器学习模型并生成准确的预测，而无需编写任何代码。

  开始阅读本教程”

- 深入了解 SageMaker AI Canvas

  深入了解 SageMaker AI Canvas 及其可视化、无代码机器学习功能。

  阅读博客”

- 使用 Amazon SageMaker AI Canvas 制作你的第一个机器学习模型

  了解如何使用 Amazon SageMaker AI Canvas 创建机器学习模型，根据针对新产品和服务的电子邮件活动来评估客户留存率。

  开始使用实验室”

SageMaker AI Data Wrangler

- 开始使用 Amazon A SageMaker I Data Wrangler

  探索如何设置 SageMaker AI Data Wrangler，然后使用现有示例数据集提供演练。

  浏览指南”

- 使用最少的代码为机器学习准备训练数据

  了解如何使用 Amazon A SageMaker I Data Wrangler 为机器学习准备数据。

  开始阅读本教程”

- SageMaker AI Data Wrangler 深度潜水研讨会

  学习如何在数据集上应用适当的分析类型来检测异常和问题，在数据集转换过程中使用派生的分析 results/insights 来制定补救措施，并使用 AI Data Wrangler 提供的快速建模选项测试正确的转换选择和顺序。 SageMaker

  开始研讨会吧”

SageMaker AI Ground Truth

- 亚马逊 Ground Truth 入门

  探索如何使用控制台创建标签作业、分配公共或私人员工，以及如何将标签作业发送给员工。了解如何监控标签作业的进度。

  浏览指南”

- 为 Machine Learning 标记训练数据

  了解如何在 Amazon A SageMaker I Ground Truth 中设置标签作业，为机器学习模型的训练数据添加注释。

  开始阅读本教程”

- Amazon Ground Truth Plus 入门探索如何完成启动亚马逊 A SageMaker I Ground Truth Plus 项目的必要步骤，查看标签并满足 SageMaker AI Ground Truth Plus 先决条件。

  浏览指南”

- 开始使用 Amazon Ground Trut h Watch 如何通过 SageMaker AI Ground Truth 控制台在几分钟内开始为数据添加标签。

  观看视频”

- Amazon SageMaker AI Ground Truth Plus — 无需代码或内部资源即可创建训练数据集

  了解 Ground Truth Plus，这是一项交钥匙服务，它使用专业的员工队伍快速提供高质量的培训数据集，并将成本降低多达 40%。

  阅读博客”

SageMaker AI JumpStart

- 开始使用 A SageMaker I 进行机器学习 JumpStart

  探索为常见用例设置基础架构的 SageMaker AI JumpStart 解决方案模板，以及用于通过 SageMaker AI 进行机器学习的可执行示例笔记本。

  浏览指南”

- 使用 Amazon A SageMaker I 快速开始您的机器学习项目 JumpStart

  了解如何使用 Amazon AI 提供的预训练模型和预构建的解决方案快速跟踪您的机器学习项目。SageMaker JumpStart然后，您可以通过 Amazon A SageMaker I Studio 笔记本部署所选模型。

  开始阅读本教程”

- 通过这个沉浸式日研讨会 JumpStart ，亲身体验 SageMaker Amazon AI

  了解 Amazon A SageMaker I Data Wrangler、Autopilot 和 Jumpstart 中的低代码机器学习功能如何简化更快地进行实验并将高度精确的模型投入生产。

  开始研讨会吧”

SageMaker AI Pipelines

- 亚马逊 A SageMaker I Pipelines 入门

  了解如何创建管理和部署 SageMaker AI 作业的工作 end-to-end流程。 SageMaker AI Pipelines 集成了 SageMaker AI Python SDK，因此您可以使用基于 Python 的界面构建管道的每个步骤。

  浏览指南”

- 自动执行机器学习工作流程

  学习如何使用亚马逊 A SageMaker I Pipelines、Amazon AI 模型注册表和 Amazon AI Clar SageMaker ify 创建和自动执行 end-to-end机器学习 (ML) 工作流程。 SageMaker

  开始阅读本教程”

- 如何使用 Amazon A SageMaker I Pipelines 创建全自动机器学习工作流程

  了解 Amazon SageMaker AI Pipelines，这是世界上第一款专为每位开发者和数据科学家都可访问而设计的机器学习 CI/CD 服务。 SageMaker AI Pipel CI/CD ines 为机器学习带来了流水线，从而缩短了所需的编码时间。

  观看视频”

SageMaker AI Studio

- 在本地构建和训练机器学习模型

  了解如何在 Amazon A SageMaker I Studio 笔记本中本地构建和训练机器学习模型。

  开始阅读本教程”

- SageMaker AI Studio 与 EMR 研讨会的集成

  学习如何大规模利用分布式处理来准备数据并随后训练机器学习模型。

  开始研讨会吧”

第三组 AI/ML AWS 服务 途径提供了深入的文档、动手教程和资源，供您开始使用 AWS Trainium、AWS Inferentia和 Amazon Titan。

AWS Trainium

- 使用 AWS Trainium 和 Amazon EKS 扩展分布式训练

  了解如何从由 AWS Trainium专门构建的机器学习加速器提供支持的 Amazon EC2 Trn1 实例中受益，该实例经过优化，可为在云中训练深度学习模型提供高性能、经济实惠且可大规模扩展的平台。

阅读博客”

- 的概述 AWS Trainium

  了解 AWS Trainium第二代机器学习 (ML) 加速器，专为 1000 亿多个参数模型的深度学习训练 AWS 而构建。每个 Amazon Elastic Compute Cloud (EC2) Trn1 实例最多可部署 16 个 AWS Trainium 加速器，为云端深度学习 (DL) 训练提供高性能、低成本的解决方案。

  浏览指南”

- 推荐的 Trainium 实例

  探索如何设计 AWS Trainium 实例以为深度学习模型推理工作负载提供高性能和成本效益。

  浏览指南”

AWS Inferentia

- 的概述 AWS Inferentia

  了解加速器是如何设计的 AWS ，以最低的成本为您的深度学习 (DL) 推理应用程序提供高性能。

  浏览指南”

- AWS Inferentia 2 在 AWS Inferentia 1 的基础上构建，吞吐量提高 4 倍，延迟降低 10 倍

  了解 AWS Inferentia 2 针对什么进行了优化，并探索它是如何从头开始设计的，以提供更高的性能，同时降低生成式 AI 推理的成 LLMs 本。

  阅读博客”

- 使用机器学习推理 AWS Inferentia

  学习如何使用运行 Amazon EC2 Inf1 实例的节点创建 Amazon EKS 集群，以及（可选）部署示例应用程序。Amazon EC2 Inf1 实例由 AWS Inferentia 芯片提供支持，芯片由 AWS 芯片定制构建，可在云中提供高性能和最低成本的推理。

  浏览指南”

Amazon Titan

- Amazon Titan 概述

  探索 Amazon Titan FMs 如何对大型数据集进行预训练，使其成为强大的通用模型。了解如何按原样使用它们或私下使用它们，使用您自己的数据针对特定任务对其进行自定义，而无需为大量数据添加注释。

  浏览指南”

Explore

- 架构图

  这些参考架构图显示了正在使用的 AWS AI 和 ML 服务的示例。

  浏览架构图”
- 白皮书

  浏览白皮书以帮助您入门并学习选择和使用 AI/ML 服务的最佳实践。

  浏览白皮书”
- AWS 解决方案

  探索经过审查的解决方案和架构指南，了解 AI 和 ML 服务的常见用例。

  探索解决方案”

资源

基础模型

支持的基础模型包括：

- 人类 Claude
- Cohere 命令和嵌入

- AI21 《侏罗纪实验室》
- Meta Llama
- Mistral AI
- 稳定扩散 XL
- Amazon Titan

使用 Amazon Bedrock，您可以尝试各种基础模型，并使用您的数据对其进行私密自定义。

用例或行业特定服务

- Amazon Comprehend Medical
- Amazon Fraud Detector
- AWS HealthLake
- Amazon Lookout for Equipment
- Amazon Lookout for Metrics
- Amazon Lookout for Vision
- Amazon Monitron
- AWS HealthOmics
- AWS Panorama

相关的博客文章

- 重要的新功能使您可以更轻松地使用 Amazon Bedrock 来构建和扩展生成式 AI 应用程序，并取得令人印象深刻的成果
- AWS 推断并 AWS Trainium 以最低的成本在 Amazon AI 中部署 Llama 3 模型 SageMaker JumpStart
- 在 Ama SageMaker zon AI 上为您的企业量身定制的奖励模式，彻底提高客户满意度
- Amazon Personalize 推出支持更大商品目录且延迟更低的新配方

文档历史记录

下表描述了本决策指南的重要更改。要获取有关本指南更新的通知，您可以订阅 RSS feed。

变更	说明	日期
次要更新	更新了 Amazon Q 和亚马逊最新人工智能和机器学习堆栈的内容。	2024 年 5 月 3 日
初始版本	决策指南的首次发布。	2023 年 7 月 24 日

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。