AWS 決策指南

選擇 AWS 機器學習服務

選擇 AWS 機器學習服務: AWS 決策指南

Table of Contents

選擇 AWS 機器學習服務

選擇正確的 ML 服務和架構來支援您的工作

用途	協助判斷哪些 AWS ML 服務最適合您的需求。
上次更新	2024 年 5 月 3 日
涵蓋的服務	• Amazon 增強版 AI • Amazon CodeGuru • Amazon Comprehend • Amazon DevOps Guru • Amazon Forecast • Amazon Kendra • Amazon Lex • Amazon Personalize • Amazon Polly • Amazon Rekognition • Amazon SageMaker AI • Amazon Textract • Amazon Transcribe • Amazon Translate

簡介

最基本的機器學習 (ML) 旨在提供數位工具和服務，以便從資料中學習、識別模式、進行預測，然後根據這些預測採取行動。現在幾乎所有人工智慧 (AI) 系統都是使用 ML 建立的。ML 使用大量資料來建立和驗證決策邏輯。此決策邏輯構成 AI 模型的基礎。

可能套用 AWS 機器學習服務的案例包括：

- 特定使用案例 — AWS 機器學習服務可以透過針對常見使用案例和產業的各種預先建置演算法、模型和解決方案，支援 AI 支援的使用案例。您可以選擇 23 種預先訓練的服務，包括 Amazon Personalize、Amazon Kendra 和 Amazon Monitron。

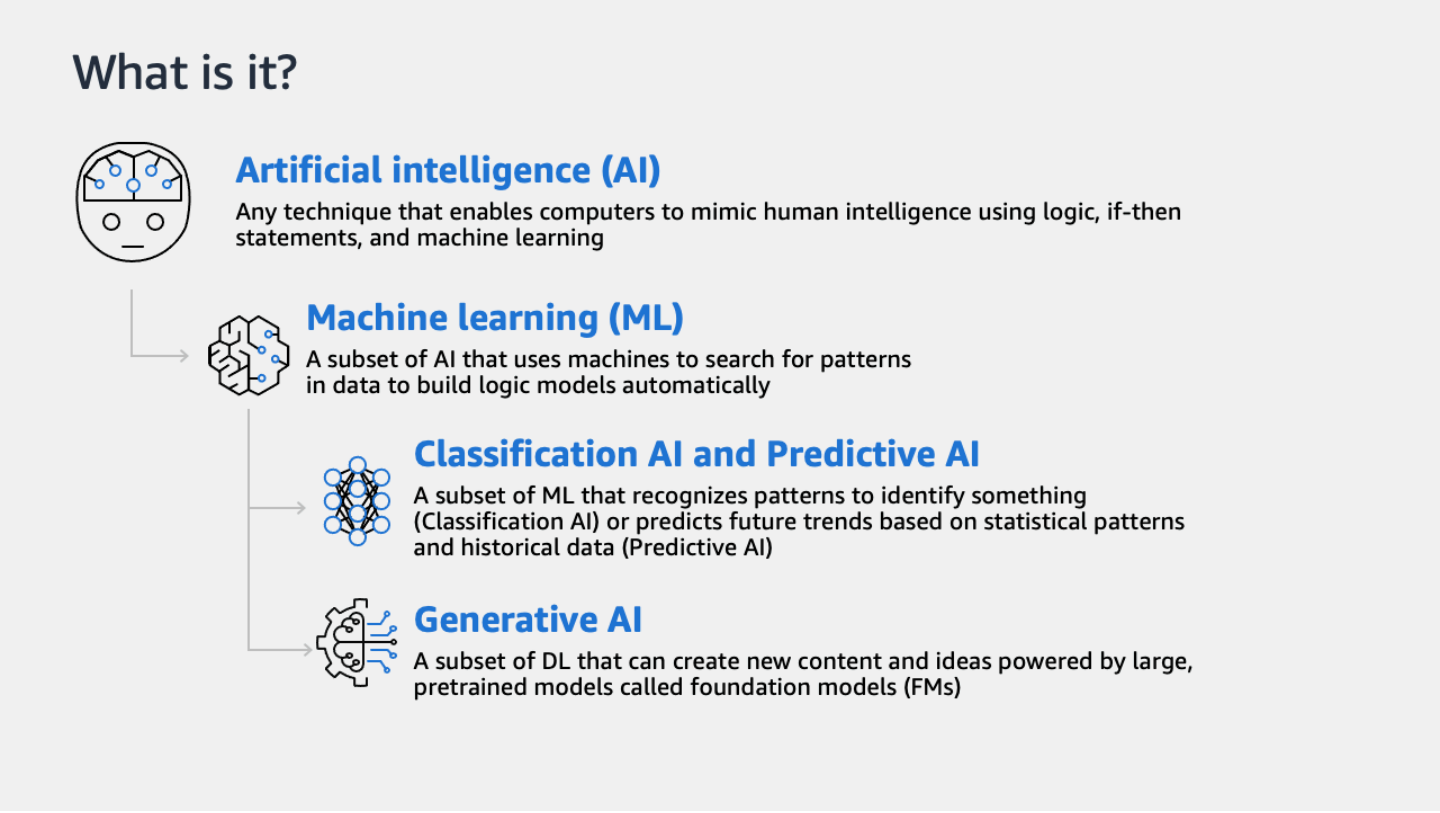
- 自訂和擴展機器學習 — Amazon SageMaker AI 旨在協助您針對任何使用案例建置、訓練和部署 ML 模型。您可以透過 Amazon SageMaker AI 和 Amazon Bedrock 在 AWS 上建置或存取開放原始碼基礎模型。
- 存取專用基礎設施 — 當您需要更大的彈性和對機器學習工作流程的控制 AWS 時，請使用 提供的 ML 架構和基礎設施，並且願意自行管理基礎基礎設施和資源。

此決策指南將協助您提出正確的問題、評估您的條件和業務問題，以及判斷哪些服務最適合您的需求。

了解

隨著組織持續採用 AI 和 ML 技術，了解和選擇 AWS ML 服務的重要性是一項持續的挑戰。

AWS 提供各種 ML 服務，旨在協助組織更快速輕鬆地建置、訓練和部署 ML 模型。這些服務可用於解決各種商業問題，例如客戶流失預測、詐騙偵測，以及影像和語音辨識。



在深入了解 AWS ML 服務之前，讓我們先來看看 AI 與 ML 之間的關係。

- 從高層次而言，人工智慧是一種描述任何系統的方法，這些系統可以複寫先前需要人類智慧的任務。大多數 AI 使用案例都在尋找機率結果：做出高確定性的預測或決策，類似於人類判斷。

- 現在幾乎所有 AI 系統都是使用機器學習建立的。ML 使用大量資料來建立和驗證稱為模型的決策邏輯。
- 分類 AI 是 ML 的子集，可辨識識別某物的模式。預測性 AI 是 ML 的子集，可根據統計模式和歷史資料預測未來趨勢。
- 最後，生成式 AI 是深度學習的子集，可以建立新的內容和想法，例如對話、故事、影像、影片和音樂。生成式 AI 採用非常大型的模型，這些模型在大量資料主體上預先訓練，稱為基礎模型或 FMs。Amazon Bedrock 是一種全受管服務，可提供高效能 FMs 的選擇，用於建置和擴展生成式 AI 應用程式。Amazon Q Developer 和 Amazon Q Business 是適用於特定使用案例的生成式 AI 輔助。

本指南的設計主要涵蓋分類 AI 和預測性 AI 機器學習類別中的服務。

此外， AWS 還提供專門的加速硬體，以實現高效能 ML 訓練和推論。

- Amazon EC2 P5instances 配備 NVIDIA H100 Tensor 核心 GPUs，非常適合機器學習的訓練和推論任務。Amazon EC2 G5instances 具有高達 8 個 NVIDIA A10G Tensor 核心 GPUs 和第二代 AMD EPYC 處理器，適用於各種圖形密集型和機器學習使用案例。
- AWS Trainium是第二代 ML 加速器， AWS 專為 100B+ 參數模型的深度學習 (DL) 訓練而打造。
- AWS Inferentia基於 2 的 Amazon EC2 Inf2 執行個體旨在為您的 DL 和生成式 AI 推論應用程式以最低成本在 Amazon EC2 中提供高效能。

考慮

解決 AWS ML 服務的業務問題時，考量幾個關鍵條件有助於確保成功。下一節概述選擇 ML 服務時要考慮的一些關鍵條件。

Problem definition

問題定義

ML 生命週期的第一步是架構業務問題。了解您嘗試解決的問題對於選擇正確的 AWS ML 服務至關重要，因為不同的服務旨在解決不同的問題。判斷 ML 是否適合您的業務問題也很重要。

確定 ML 最適合之後，您可以從一系列專用 AWS AI 服務 （在語音、視覺和文件等領域） 中進行選擇。

如果您需要建置和訓練自己的模型，Amazon SageMaker AI 會提供全受管基礎設施。 為需要高度自訂和專業 ML 模型的案例 AWS 提供一系列進階 ML 架構和基礎設施選擇。 AWS 也提供廣泛的熱門基礎模型，以使用生成式 AI 建置新的應用程式。

ML algorithm

ML 演算法

針對您嘗試解決的業務問題選擇 ML 演算法取決於您正在使用的資料類型，以及所需的結果。以下資訊概述每個主要 AWS AI/ML 服務類別如何讓您能夠使用其演算法：

- 專用 AI 服務：這些服務提供自訂 ML 演算法的有限能力，因為它們是針對特定任務最佳化的預先訓練模型。您通常可以自訂輸入資料和一些參數，但無法存取基礎 ML 模型或能夠建置自己的模型。
- Amazon SageMaker AI：此服務可為 ML 演算法提供最大的彈性和控制。您可以使用 SageMaker AI 使用您自己的演算法和架構來建置自訂模型，或使用 提供的預先建置模型和演算法 AWS。這允許對 ML 程序進行高度的自訂和控制。
- 低階 ML 架構和基礎設施：這些服務可為 ML 演算法提供最大的彈性和控制。您可以使用這些服務，使用自己的演算法和架構來建置高度自訂的 ML 模型。不過，使用這些服務需要重要的 ML 專業知識，而且並非所有使用案例都可行。

Security

安全性

如果您在 VPC 中需要私有端點，您的選項會根據您使用的 AWS ML 服務層而有所不同。其中包含：

- 專用 AI 服務：大多數專用 AI 服務目前不支援 VPCs中的私有端點。不過，可以使用 VPC 端點存取 Amazon Rekognition 自訂標籤和 Amazon Comprehend Custom。
- 核心 AI 服務：Amazon Translate、Amazon Transcribe 和 Amazon Comprehend 都支援 VPC 端點。
- Amazon SageMaker AI：SageMaker AI 為 VPC 端點提供內建支援，可讓您將訓練模型部署為只能在其 VPC 內存取的端點。
- 低階 ML 架構和基礎設施：您可以在 Amazon EC2 執行個體或 VPC 內的容器中部署模型，以完全控制聯網組態。

Latency

Latency (延遲)

Amazon Rekognition 和 Amazon Transcribe 等高階 AI 服務旨在處理各種使用案例，並在速度方面提供高效能。不過，它們可能不符合特定延遲要求。

如果您使用的是較低層級的 ML 架構和基礎設施，建議您利用 Amazon SageMaker AI。由於全受管服務和最佳化部署選項，此選項通常比建置自訂模型更快。雖然高度最佳化的自訂模型可能會優於 SageMaker AI，但它需要大量的專業知識和資源才能建置。

Accuracy

準確度

AWS ML 服務的準確性會根據所需的特定使用案例和自訂層級而有所不同。Amazon Rekognition 等高階 AI 服務是以預先訓練的模型為基礎，這些模型已針對特定任務進行最佳化，並在許多使用案例中提供高準確度。

在某些情況下，您可以選擇使用 Amazon SageMaker AI，它提供更靈活且可自訂的平台，用於建置和訓練自訂 ML 模型。透過建置您自己的模型，您可以達到比預先訓練模型更高的準確度。

您也可以選擇使用 ML 架構和基礎設施，例如 TensorFlow 和 Apache MXNet，來建置高度自訂的模型，為您的特定使用案例提供最高的準確性。

AWS and responsible AI

AWS 和負責任的 AI

AWS 在開發程序的每個階段，以負責任的 AI 建立基礎模型 (FMs)。在整個設計、開發、部署和操作過程中，我們考慮了各種因素，包括：

1. 準確性（摘要符合基礎文件的程度；傳記是否實際正確）
2. 公平性（輸出是否以類似的方式處理人口統計群組）
3. 智慧財產權和著作權考量
4. 適當的用量（篩選出使用者對法律建議、醫療診斷或非法活動的請求）
5. 毒性（仇恨語音、褻瀆和侮辱）
6. 隱私權（保護個人資訊和客戶提示）

AWS 建置解決方案來解決這些問題，包括用於取得訓練資料的程序、FMs 本身，以及用於預先處理使用者提示和後續處理輸出的技術。

選擇

現在您知道評估 ML 服務選項的條件，您就可以選擇適合您組織需求的 AWS ML 服務。下表重點說明哪些 ML 服務已針對哪些情況進行最佳化。使用它來協助判斷最適合您的使用案例的 AWS ML 服務。

類別	您會何時使用它？	它針對什麼進行最佳化？	相關的 AI/ML 服務或環境
特定使用案例 這些人工智慧服務旨在滿足特定需求。其中包括個人化、預測、異常偵測、語音轉錄等。由於它們以 服務的形式交付，因此可以內嵌到應用程式中，而不需要任何 ML 專業知識。	AWS 當您需要特定預先建置的功能整合到您的應用程式時，請使用 提供的 AI 服務，而不需要廣泛的自訂或機器學習專業知識。這些服務的設計易於使用，不需要太多編碼或組態。	這些服務的設計易於使用，不需要太多編碼、組態或 ML 專業知識。	Amazon 增強版 AI Amazon CodeGuru Amazon Comprehend Amazon Comprehend Medical Amazon DevOps Guru Amazon Forecast Amazon Kendra Amazon Lex Amazon Personalize Amazon Polly Amazon Rekognition Amazon Textract Amazon Transcribe Amazon Translate
ML 服務 這些服務可用來開發自訂機器學習模型或	當您需要比核心 AI 服務提供的預先建置功能更自訂的機器學習	這些服務已針對建置和訓練自訂機器學習模型、在多個執行個	Amazon SageMaker AI

類別	您會何時使用它？	它針對什麼進行最佳化？	相關的 AI/ML 服務或環境
工作流程，超越核心 AI 服務提供的預先建置功能。	模型或工作流程時，請使用這些服務。	體或 GPU 叢集上進行大規模訓練、進一步控制機器學習模型部署、即時推論，以及建置end-to-end工作流程進行最佳化。	Amazon SageMaker AI JumpStart SageMaker AI Studio SageMaker AI Canvas SageMaker AI Studio 實驗室 SageMaker AI Ground Truth 上的 PyTorch AWS Apache MxNet Hugging Face 上的 TensorFlow AWS
基礎設施 若要在生產環境中部署機器學習，您需要符合成本效益的基礎設施，Amazon 可透過AWS建置的晶片啟用這些基礎設施。	當您想要達到訓練模型的最低成本，且需要在雲端執行推論時，請使用 。	最佳化以支援具成本效益的機器學習部署。	AWS Trainium AWS Inferentia 和 Inferentia2 Amazon SageMaker AI HyperPod

類別	您會何時使用它？	它針對什麼進行最佳化？	相關的 AI/ML 服務或環境
工具和相關服務 這些工具和相關服務旨在協助您輕鬆部署機器學習。	這些服務和工具旨在協助您在雲端加速深度學習，提供 Amazon Machine Image、Docker Image 和實體解析度。	專為協助您在雲端加速深度學習而最佳化。	AWS 深度學習 AMIs s AWS 深度學習容器 AWS 實體解析

使用

現在您已清楚了解選擇 AWS ML 服務時需要套用的條件，您可以選擇哪些 AWS AI/ML 服務已針對您的業務需求最佳化 (AI/ML)。

為了探索如何使用和進一步了解您選擇的服務，我們提供了三組路徑來探索每個服務的運作方式。第一組路徑提供深入的文件、實作教學和資源，以開始使用 Amazon Comprehend、Amazon Textract、Amazon Translate、Amazon Lex、Amazon Polly、Amazon Rekognition 和 Amazon Transcribe。

Amazon Comprehend

- 開始使用 Amazon Comprehend

  使用 Amazon Comprehend 主控台來建立和執行非同步實體偵測任務。

  教學課程入門 »

- 使用 Amazon Comprehend 分析文字中的洞見

  了解如何使用 Amazon Comprehend 從文字中分析和衍生洞見。

  教學課程入門 »

- Amazon Comprehend 定價

  探索 Amazon Comprehend 定價和範例的相關資訊。

  探索指南 »

Amazon Textract

- Amazon Textract 入門

  了解 Amazon Textract 如何與格式化文字搭配使用，以偵測彼此靠近的單字和單字行，以及分析文件是否有相關文字、資料表、鍵/值對和選取元素等項目。

  探索指南 »

- 使用 Amazon Textract 擷取文字和結構化資料

  了解如何使用 Amazon Textract 從文件中擷取文字和結構化資料。

  開始使用教學課程 »

- AWS 動力小時：Machine Learning

  深入探討本集的 Amazon Textract、花時間在 中 AWS Management Console，並檢閱程式碼範例，以協助您了解如何充分利用服務 APIs。

  觀看影片 »

Amazon Translate

- 使用主控台開始使用 Amazon Translate

  開始使用 Amazon Translate 的最簡單方法是使用 主控台來翻譯一些文字。了解如何使用主控台翻譯最多 10，000 個字元。

  探索指南 »

- 在雲端語言之間翻譯文字

  在本教學範例中，作為國際包船製造公司的一部分，您需要了解客戶在當地市場語言法文的評論中對您的產品說了什麼。

  開始使用教學課程 »

- Amazon Translate 定價

  探索 Amazon Translate 定價，包括 免費方案 - 提供 12 個月每月 200 萬個字元。

  探索指南 »

Amazon Lex

- Amazon Lex V2 開發人員指南

  探索入門的相關資訊、其運作方式，以及 Amazon Lex V2 的定價資訊。

  探索指南 »

- Amazon Lex 簡介 我們向您介紹 Amazon Lex 對話式服務，並逐步解說示範如何建立機器人並將其部署到不同聊天服務的範例。

  參加課程 »（需要登入）

- 探索對話體驗中的生成式 AI

  探索在對話體驗中使用生成式 AI。

  閱讀部落格 »

Amazon Polly

- 什麼是 Amazon Polly？

  探索雲端服務的完整概觀，將文字轉換為逼真的語音，並可用於開發應用程式，以提高您的客戶參與度和可存取性。

  探索指南 »

- 使用 Amazon Polly 在說出文字時反白文字

  我們向您介紹反白文字的方法，因為它正在說話，以在書籍、網站、部落格和其他數位體驗中為音訊新增視覺功能。

  閱讀部落格 »

- 在 Amazon Polly 中使用相同的 TTS 語音角色，以多種語言建立內容的音訊

  我們解釋神經Text-to-Speech(NTTS)，並討論各種可用的語音組合，以支援的語言提供各種不同的發言者，如何為您服務。

  閱讀部落格 »

Amazon Rekognition

- 什麼是 Amazon Rekognition？

  探索如何使用此服務將影像和影片分析新增至您的應用程式。

  探索指南 »

- 實作 Rekognition：自動化影像和影片分析

  了解臉部辨識如何以自我引導的速度使用串流影片，以及程式碼範例和關鍵點。

  教學課程入門 »

- Amazon Rekognition FAQs

  了解 Amazon Rekognition 的基本概念，以及它如何協助您改善深度學習並視覺化分析應用程式。

  閱讀FAQs」

Amazon Transcribe

- 什麼是 Amazon Transcribe？

  探索使用 ML 將音訊轉換為文字的 AWS 自動語音辨識服務。了解如何使用此服務做為獨立轉錄，或將speech-to-text功能新增至任何應用程式。

  探索指南 »

- Amazon Transcribe 定價

我們向您介紹 AWS pay-as-you-go轉錄，包括自訂語言模型選項和 Amazon Transcribe 免費方案。

探索指南 »

- 使用 Amazon Transcribe 建立音訊文字記錄

  了解如何使用 Amazon Transcribe，使用實際的使用案例案例來建立錄製音訊檔案的文字文字記錄，以根據您的需求進行測試。

  教學課程入門 »

- 建置 Amazon Transcribe 串流應用程式

  了解如何建置應用程式以即時記錄、轉錄和翻譯即時音訊，並將結果透過電子郵件直接傳送給您。

  探索指南 »

第二組 AI/ML AWS 服務 路徑提供深入的文件、實作教學和資源，以開始使用 Amazon SageMaker AI 系列中的服務。

SageMaker AI

- Amazon SageMaker AI 的運作方式

  探索機器學習的概觀，以及 SageMaker AI 的運作方式。

  探索指南 »

- Amazon SageMaker AI 入門

  了解如何加入 Amazon SageMaker AI 網域，讓您存取 Amazon SageMaker AI Studio 和 RStudio on SageMaker AI。

  探索指南 »

- 搭配 Amazon SageMaker AI 使用 Apache Spark

了解如何使用 Apache Spark 預先處理資料，以及使用 SageMaker AI 進行模型訓練和託管。

探索指南 »

- 使用 Docker 容器建置模型

  探索 Amazon SageMaker AI 如何廣泛使用 Docker 容器進行建置和執行期任務。了解如何為內建演算法部署預先建置的 Docker 映像，以及用於訓練和推論的支援深度學習架構。

  探索指南 »

- 機器學習架構和語言

  了解如何使用Amazon SageMaker AI Python SDK 開始使用 Amazon SageMaker AI。

  探索指南 »

SageMaker AI Autopilot

- 為表格式資料建立 Amazon SageMaker AI Autopilot 實驗

  了解如何建立 Amazon SageMaker AI Autopilot 實驗，以探索、預先處理和訓練表格式資料集上的各種模型候選項目。

  探索指南 »

- 自動建立機器學習模型

  了解如何使用 Amazon SageMaker AI Autopilot 自動建置、訓練和調校 ML 模型，並部署模型以進行預測。

  開始使用教學課程 »

- 使用這些範例筆記本探索 Amazon SageMaker AI Autopilot 的建模

  探索用於直接行銷、客戶流失預測，以及如何將自己的資料處理程式碼帶入 Amazon SageMaker AI Autopilot 的範例筆記本。

  探索指南 »

SageMaker AI Canvas

- 開始使用 Amazon SageMaker AI Canvas

  了解如何開始使用 SageMaker AI Canvas。

  探索指南 »

- 產生機器學習預測，無需撰寫程式碼

  本教學課程說明如何使用 Amazon SageMaker AI Canvas 來建置 ML 模型並產生準確的預測，而無需撰寫單行程式碼。

  開始使用教學課程 »

- 深入了解 SageMaker AI Canvas

  探索 SageMaker AI Canvas 及其視覺效果的深入介紹，無程式碼 ML 功能。

  閱讀部落格 »

- 使用 Amazon SageMaker AI Canvas 建立您的第一個 ML 模型

  了解如何根據新產品和服務的電子郵件行銷活動，使用 Amazon SageMaker AI Canvas 建立 ML 模型來評估客戶保留。

  實驗室入門 »

SageMaker AI Data Wrangler

- Amazon SageMaker AI Data Wrangler 入門

  探索如何設定 SageMaker AI Data Wrangler，然後使用現有的範例資料集提供逐步解說。

  探索指南 »

- 以最少的程式碼準備機器學習的訓練資料

  了解如何使用 Amazon SageMaker AI Data Wrangler 為 ML 準備資料。

  教學課程入門 »

- SageMaker AI Data Wrangler 深入探討研討會

  了解如何在資料集上套用適當的分析類型，以偵測異常和問題、使用衍生的結果/洞見在資料集的轉換過程中制定補救措施，並使用 SageMaker AI Data Wrangler 提供的快速建模選項來測試正確的轉換選擇和序列。

  研討會入門 »

SageMaker AI Ground Truth

- Amazon Ground Truth 入門

  探索如何使用 主控台建立標籤工作、指派公有或私有人力資源，以及將標籤工作傳送至您的人力資源。了解如何監控標籤工作的進度。

  探索指南 »

- Machine Learning的標籤訓練資料

  了解如何在 Amazon SageMaker AI Ground Truth 中設定標籤工作，以註釋 ML 模型的訓練資料。

  教學課程入門 »

- Amazon Ground Truth Plus 入門 探索如何完成啟動 Amazon SageMaker AI Ground Truth Plus 專案、檢閱標籤和滿足 SageMaker AI Ground Truth Plus 先決條件的必要步驟。

  探索指南 »

- 開始使用 Amazon Ground Truth 觀看如何透過 SageMaker AI Ground Truth 主控台在幾分鐘內開始標記您的資料。

  觀看影片 »

- Amazon SageMaker AI Ground Truth Plus – 建立不含程式碼或內部資源的訓練資料集

  了解 Ground Truth Plus，這是一種使用專家人力快速交付高品質訓練資料集並降低成本高達 40% 的統包服務。

閱讀部落格 »

SageMaker AI JumpStart

- 開始使用 SageMaker AI JumpStart 進行機器學習

  探索為常見使用案例設定基礎設施的 SageMaker AI JumpStart 解決方案範本，以及使用 SageMaker AI 進行機器學習的可執行範例筆記本。

  探索指南 »

- 使用 Amazon SageMaker AI JumpStart 快速開始使用您的機器學習專案

  了解如何使用 Amazon SageMaker AI JumpStart 提供的預先訓練模型和預先建置解決方案，快速追蹤 ML 專案。然後，您可以透過 Amazon SageMaker AI Studio 筆記本部署選取的模型。

  教學課程入門 »

- 透過此沉浸式日研討會取得 Amazon SageMaker AI JumpStart 實作

  了解Amazon SageMaker AI Data Wrangler、Autopilot 和 Jumpstart 中找到的低程式碼 ML 功能如何讓您更輕鬆地實驗更快，並將高度準確的模型帶入生產環境。

  研討會入門 »

SageMaker AI Pipelines

- Amazon SageMaker AI 管道入門

  了解如何建立end-to-end工作流程來管理和部署 SageMaker AI 任務。SageMaker AI Pipelines 隨附 SageMaker AI Python SDK 整合，因此您可以使用 Python 型界面建置管道的每個步驟。

  探索指南 »

- 自動化機器學習工作流程

  了解如何使用 Amazon SageMaker AI Pipelines、Amazon SageMaker AI Model Registry 和 Amazon SageMaker AI Clarify 建立和自動化end-to-end機器學習 (ML) 工作流程。

開始使用教學課程 »

- 如何使用 Amazon SageMaker AI Pipelines 建立全自動化 ML 工作流程

  了解 Amazon SageMaker AI Pipelines，這是全球第一個專為每位開發人員和資料科學家設計的 ML CI/CD 服務。SageMaker AI 管道將 CI/CD 管道帶入 ML，減少了所需的編碼時間。

  觀看影片 »

SageMaker AI Studio

- 在本機建置和訓練機器學習模型

  了解如何在 Amazon SageMaker AI Studio 筆記本中於本機建置和訓練 ML 模型。

  教學課程入門 »

- SageMaker AI Studio 與 EMR 整合研討會

  了解如何利用大規模分散式處理來準備資料，並隨後訓練機器學習模型。

  研討會入門 »

第三組 AI/ML AWS 服務 路徑提供深入的文件、實作教學和資源，以開始使用 AWS Trainium AWS Inferentia、 和 Amazon Titan。

AWS Trainium

- 使用 AWS Trainium 和 Amazon EKS 擴展分散式訓練

  了解如何從採用 的 Amazon EC2 Trn1 執行個體的一般可用性中獲益 AWS Trainium，此專用 ML 加速器已最佳化，可提供高效能、經濟實惠且可擴展的平台，以在雲端中訓練深度學習模型。

  閱讀部落格 »

- 的概觀 AWS Trainium

了解 AWS 專為 100B+ 參數模型深度學習訓練打造的 AWS Trainium第二代機器學習 (ML) 加速器。每個 Amazon Elastic Compute Cloud (EC2) Trn1 執行個體最多部署 16 AWS Trainium 個加速器，為雲端深度學習 (DL) 訓練提供高效能、低成本的解決方案。

探索指南 »

- 建議的 Trainium 執行個體

  探索 AWS Trainium 執行個體的設計方式，為深度學習模型推論工作負載提供高效能和成本效益。

  探索指南 »

AWS Inferentia

- 概觀 AWS Inferentia

  了解 加速器如何由 設計 AWS ，以最低成本為您的深度學習 (DL) 推論應用程式提供高效能。

  探索指南 »

- AWS Inferentia 2 建置於 AWS Inferentia 1，可提供 4 倍的輸送量和 10 倍的低延遲

  了解針對 最佳化的 What AWS Inferentia 2，並探索它如何從頭開始設計，以提供更高的效能，同時降低 LLMs和生成式 AI 推論的成本。

  閱讀部落格 »

- 使用 的機器學習推論 AWS Inferentia

  了解如何使用執行 Amazon EC2 Inf1 執行個體的節點建立 Amazon EKS 叢集，以及 （選用） 部署範例應用程式。 Inf1 Amazon EC2 Inf1 執行個體採用由 自訂的 AWS Inferentia 晶片， AWS 可在雲端提供高效能和最低成本的推論。

  探索指南 »

Amazon Titan

- Amazon Titan 概觀

  探索如何在大型資料集上預先訓練 Amazon Titan FMs，使其成為強大的一般用途模型。了解如何以原狀或私有方式使用它們，以針對特定任務使用您自己的資料來自訂它們，而無需註釋大量資料。

  探索指南 »

探索

- 架構圖

  這些參考架構圖顯示使用中的 AWS AI 和 ML 服務範例。

  探索架構圖 »
- 白皮書

  探索白皮書，協助您開始使用並了解選擇和使用 AI/ML 服務的最佳實務。

  探索白皮書 »
- AWS 解決方案

  探索 AI 和 ML 服務常見使用案例的審核解決方案和架構指引。

  探索解決方案 »

資源

基礎模型

支援的基礎模型包括：

- Anthropic Claude
- Cohere 命令與內嵌
- AI21 實驗室 Jurassic
- Meta Llama

- 混合式 AI
- 穩定擴散 XL
- Amazon Titan

使用 Amazon Bedrock，您可以實驗各種基礎模型，並使用您的資料私下自訂它們。

使用案例或產業特定的服務

- 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

關聯的部落格文章

- 重要的新功能可讓您更輕鬆地使用 Amazon Bedrock 來建置和擴展生成式 AI 應用程式，並實現令人驚豔的結果
- AWS 推論並提供 AWS Trainium 在 Amazon SageMaker AI JumpStart 中部署 Llama 3 模型的最低成本
- 透過 Amazon SageMaker AI 為您的業務量身打造的獎勵模型，徹底改變客戶滿意度
- Amazon Personalize 推出新的配方，支援具有較低延遲的大型項目目錄

文件歷史記錄

下表說明此決策指南的重要變更。如需有關本指南更新的通知，您可以訂閱 RSS 摘要。

變更	描述	日期
次要更新	已更新 Amazon Q 和 Amazon 最新 AI 和 ML 堆疊的內容。	2024 年 5 月 3 日
初始版本	決策指南的初始版本。	2023 年 7 月 24 日

本文為英文版的機器翻譯版本，如內容有任何歧義或不一致之處，概以英文版為準。