检索增强生成选项和架构 AWS

AWS 规范性指导

AWS 规范性指导: 检索增强生成选项和架构 AWS

Table of Contents

检索增强生成选项和架构 AWS

Amazon Web Services 的 Mithil Shah、Rajeev Muralidhar 和 Natacha Fort

2024 年 10 月（文档历史记录）

生成式 AI 是指 AI 模型的子集，这些模型可以根据简单的文本提示创建新的内容和工件，例如图像、视频、文本和音频。生成式 AI 模型基于大量数据进行训练，这些数据涵盖了广泛的主题和任务。这使他们能够在执行各种任务时表现出非凡的多功能性，即使是那些他们没有接受过明确培训的任务也是如此。由于单个模型能够执行多项任务，因此这些模型通常被称为基础模型 (FMs)。

生成式人工智能模型的显著应用之一是它们在回答问题方面的熟练程度。但是，当使用这些模型根据自定义文档回答问题时，会出现一些特定的挑战。自定义文档可以包括专有信息、内部网站、内部文档、ConfluenceSharePoint页面、页面等。一种选择是使用检索增强生成 (RAG)。使用 RAG，基础模型在生成响应之前会引用其训练数据源（例如您的自定义文档）之外的权威数据源。

本指南描述了不同的生成式 AI 选项，这些选项可用于回答自定义文档中的问题，包括检索增强生成 (RAG) 系统。它还概述了在 Amazon Web Services 上构建 RAG 系统 (AWS)。通过查看 RAG 选项和架构，您可以在 RAG 架构上的完全托管 AWS 服务和自定义 RAG 架构之间进行选择。

目标受众

本指南的目标受众是生成式人工智能架构师和管理人员，他们想要构建 RAG 解决方案，查看可用架构，并了解每种选项的优缺点。

目标

本指南可以帮助您执行以下操作：

- 了解可用于回答自定义文档中问题的生成式 AI 选项
- 在上查看 RAG 系统的架构选项 AWS
- 了解每个 RAG 选项的优缺点
- 为您的环境选择 RAG 架构 AWS

用于查询自定义文档的生成式 AI 选项

组织通常有各种来源的结构化和非结构化数据。本指南重点介绍如何使用生成式 AI 来回答来自非结构化数据的问题。

组织中的非结构化数据可能来自各种来源。这些可能是文本文件 PDFs、内部 wiki、技术文档、面向公众的网站、知识库或其他内容。如果您想要一个可以回答有关非结构化数据问题的基础模型，则可以使用以下选项：

- 使用您的自定义文档和其他训练数据训练新的基础模型
- 使用自定义文档中的数据对现有基础模型进行微调
- 当你提问时，使用情境学习将文档传递给基础模型
- 使用检索增强生成 (RAG) 方法

从头开始训练包含您的自定义数据的新基础模型是一项雄心勃勃的任务。一些公司已经成功地做到了这一点Bloomberg，例如他们的BloombergGPT模型。另一个例子是多模式EXAONE模型LG AI Research，该模型通过使用6000亿件艺术品和2.5亿张高分辨率图像以及文字进行训练。根据《人工智能的成本：你应该建造还是购买基础模型》（LinkedIn），类似的模型的训练MetaLlama 2成本约为480万美元。从头开始训练模型有两个主要先决条件：获得资源（财务、技术、时间）和明确的投资回报。如果这似乎不合适，那么下一个选择是微调现有的基础模型。

微调现有模型包括采用一个模型，例如 Amazon Titan、Mistral 或 Llama 模型，然后根据您的自定义数据调整模型。有多种微调技术，其中大多数只涉及修改几个参数，而不是修改模型中的所有参数。这称为参数效率微调。有两种主要的微调方法：

- 监督式微调使用带标签的数据，并帮助您训练模型以完成一种新的任务。例如，如果您想基于 PDF 表单生成报告，则可能需要通过提供足够的示例来教模型如何做到这一点。
- 无监督微调与任务无关，它会根据您自己的数据调整基础模型。它训练模型以了解文档的上下文。然后，经过微调的模型使用更适合组织自定义的样式来创建内容，例如报告。

但是，对于问答用例，微调可能并不理想。有关更多信息，请参阅本指南中的比较 RAG 和微调。

当你提问时，你可以将基础模型传递给文档，然后使用模型的情境学习从文档中返回答案。此选项适用于对单个文档进行即席查询。但是，此解决方案不适用于查询多个文档或查询系统和应用程序，例如微软 SharePoint 或Atlassian Confluence。

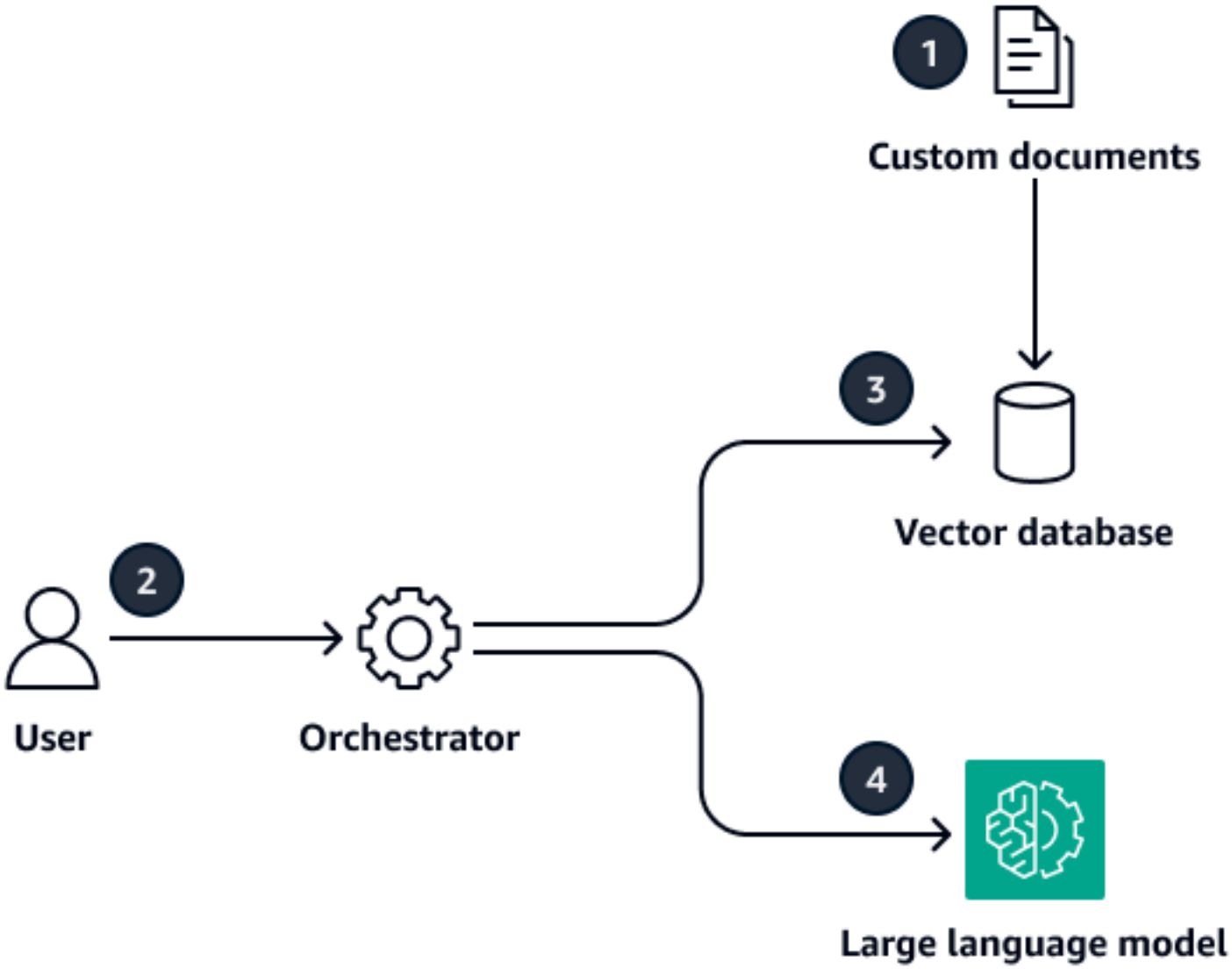
最后一个选择是使用 RAG。使用 RAG，基础模型会在生成响应之前引用您的自定义文档。RAG 将模型的功能扩展到贵组织的内部知识库，无需重新训练模型。这是一种具有成本效益的方法，可以改善模型输出，使其在各种情况下保持相关性、准确性和实用性。

本节中的主题：

- 了解检索增强生成
- 比较检索增强生成和微调
- 检索增强生成用例

了解检索增强生成

检索增强生成 (RAG) 是一种用于使用外部数据（例如公司的内部文档）增强大型语言模型 (LLM) 的技术。这为模型提供了为特定用例生成准确而有用的输出所需的上下文。RAG是一种在企业 LLMs 中使用的实用而有效的方法。下图简要概述了 RAG 方法的工作原理。



从广义上讲，RAG 流程分为四个步骤。第一步只完成一次，其他三个步骤根据需要多次完成：

1. 您可以创建嵌入以将内部文档摄取到矢量数据库中。嵌入是文档中文本的数字表示形式，用于捕捉数据的语义或上下文含义。矢量数据库本质上就是这些嵌入的数据库，它有时被称为向量存储或向量索引。此步骤需要进行数据清理、格式化和分块，但这是一次性的前期活动。
2. 人类用自然语言提交查询。
3. 协调器在矢量数据库中执行相似度搜索并检索相关数据。协调器将检索到的数据（也称为上下文）添加到包含查询的提示中。
4. 协调器将查询和上下文发送给 LLM。LLM 使用其他上下文生成对查询的响应。

从用户的角度来看，RAG 看起来像与任何 LLM 互动。但是，该系统对相关内容的了解要多得多，并根据组织的知识库提供了经过微调的答案。

有关 RAG 方法的工作原理的更多信息，请参阅网站上的 RAG 是什么。 AWS

生产级 RAG 系统的组件

构建生产级 RAG 系统需要仔细考虑 RAG 工作流程的几个不同方面。从概念上讲，无论具体实现如何，生产级 RAG 工作流程都需要以下功能和组件：

- 连接器-这些连接器将不同的企业数据源与矢量数据库连接起来。结构化数据源的示例包括交易数据库和分析数据库。非结构化数据源的示例包括对象存储、代码库和软件即服务 (SaaS) 平台。每个数据源可能需要不同的连接模式、许可证和配置。
- 数据处理-数据有多种形状和形式，例如扫描的图像 PDFs、文档、演示文稿和Microsoft SharePoint 文件。必须使用数据处理技术来提取、处理和准备要编制索引的数据。
- 嵌入-要执行相关性搜索，必须将文档和用户查询转换为兼容的格式。通过使用嵌入语言模型，可以将文档转换为数字表示。这些基本上是基础模型的输入。
- 矢量数据库-矢量数据库是嵌入项、关联文本和元数据的索引。该索引针对搜索和检索进行了优化。
- Retriever — 对于用户查询，检索器从矢量数据库中获取相关上下文，并根据业务需求对响应进行排名。
- 基础模型 — RAG 系统的基础模型通常是 LLM。通过处理上下文和提示，基础模型为用户生成并格式化响应。
- Guardrails — Guardrails旨在确保查询、提示、检索到的上下文以及法学硕士的响应是准确、负责任、合乎道德的，并且没有幻觉和偏见。
- Orchestrat or — 协调器负责安排和管理工作流程。 end-to-end
- 用户体验 — 通常，用户与对话聊天界面进行交互，该界面具有丰富的功能，包括显示聊天记录和收集用户对回复的反馈。
- 身份和用户管理-精细控制用户对应用程序的访问至关重要。在中 AWS 云，策略、角色和权限通常通过 AWS Identity and Access Management (IAM) 进行管理。

显然，规划、开发、发布和管理 RAG 系统还有大量工作要做。完全托管的服务，例如Amazon Bedrock或Amazon Q Business，可以帮助您管理一些无差别的繁重工作。但是，自定义 RAG 架构可以提供对组件的更多控制，例如检索器或矢量数据库。

比较检索增强生成和微调

下表描述了微调和基于 RAG 的方法的优缺点。

方法	优点	缺点
微调	• 如果使用无人监督的方法训练经过微调的模型，那么它就能创建更符合组织风格的内容。 • 根据专有或监管数据进行训练的微调模型可以帮助您的组织遵循内部或行业特定的数据和合规标准。	• 微调可能需要几个小时到几天，具体取决于模型的大小。因此，如果您的自定义文档经常更改，这不是一个好的解决方案。 • 微调需要对低等级适应 (LoRa) 和参数高效微调 (PEFT) 等技术的理解。微调可能需要数据科学家。 • 可能并非所有型号都提供微调。 • 经过微调的模型在响应中不提供对来源的参考。 • 使用经过微调的模型回答问题时，出现幻觉的风险可能会增加。
RAG	• RAG 允许您为自定义文档构建问答系统，无需进行微调。 • RAG 可以在几分钟内整合最新的文档。 • AWS 提供完全托管的 RAG 解决方案。因此，不需要数据科学家或机器学习方面的专业知识。 • 在响应中，RAG 模型提供了对信息源的参考。	• 在汇总整个文档中的信息时，RAG 效果不佳。

方法	优点	缺点
	• 由于RAG使用向量搜索的上下文作为其生成的答案的基础，因此降低了产生幻觉的风险。	

如果您需要构建引用您的自定义文档的问答解决方案，那么我们建议您从基于 RAG 的方法开始。如果您需要模型执行其他任务，例如汇总，请使用微调。

您可以将微调和 RAG 方法组合到一个模型中。在这种情况下，RAG 架构不会改变，但是生成答案的 LLM 也会使用自定义文档进行微调。这结合了两全其美的优势，它可能是您的用例的最佳解决方案。有关如何将监督微调与 RAG 结合起来的更多信息，请参阅中的 RAFT：使语言模型适应特定领域的 RAG 研究。University of California, Berkeley

检索增强生成用例

以下是使用 RAG 方法的常见用例：

- 搜索引擎 — 支持 RAG 的搜索引擎可以在其搜索结果中提供更准确和更具 up-to-date特色的片段。
- 问答系统 — RAG 可以提高问答系统的回答质量。基于检索的模型使用相似度搜索来查找包含答案的相关段落或文档。然后，它会根据这些信息生成简洁而相关的响应。
- 零售或电子商务 — RAG可以通过提供更具相关性和个性化的产品推荐来增强电子商务中的用户体验。通过检索和整合有关用户偏好和产品详细信息的信息，RAG可以为客户生成更准确、更有用的推荐。
- 工业或制造业 — 在制造业中，RAG 可帮助您快速访问关键信息，例如工厂运营情况。它还可以帮助决策流程、故障排除和组织创新。对于在严格监管框架内运营的制造商，RAG可以迅速从内部和外部来源（例如行业标准或监管机构）检索更新的法规和合规标准。
- 医疗保健 — RAG 在医疗保健行业具有潜力，在该行业，获得准确、及时的信息至关重要。通过检索和整合来自外部来源的相关医学知识，RAG 可以在医疗保健应用中提供更准确和更具情境感知能力的响应。这样的应用程序增强了人类临床医生可访问的信息，他们最终会打电话而不是模型。
- 法律 — RAG 可以强有力地应用于法律场景，例如兼并和收购，在这些场景中，复杂的法律文件为查询提供了背景信息。这可以帮助法律专业人士快速应对复杂的监管问题。

完全托管检索增强生成选项已开启 AWS

要管理检索增强生成 (RAG) 工作流程 AWS，您可以使用自定义 RAG 管道或使用其提供的某些完全托管的服务功能。 AWS 由于它们包含基于 RAG 的系统的许多核心组件，因此完全托管的服务可以帮助您管理一些无差别的繁重工作。但是，这些服务提供的定制机会较少。

完全托管 AWS 服务 使用连接器从外部数据源（例如网站、Atlassian Confluence 或 Microsoft）提取数据。 SharePoint支持的数据源因而异 AWS 服务。

本节探讨了以下用于构建 RAG 工作流程的完全托管选项： AWS

- Amazon Bedrock 知识库
- Amazon Q Business
- 亚马逊 A SageMaker I Canvas

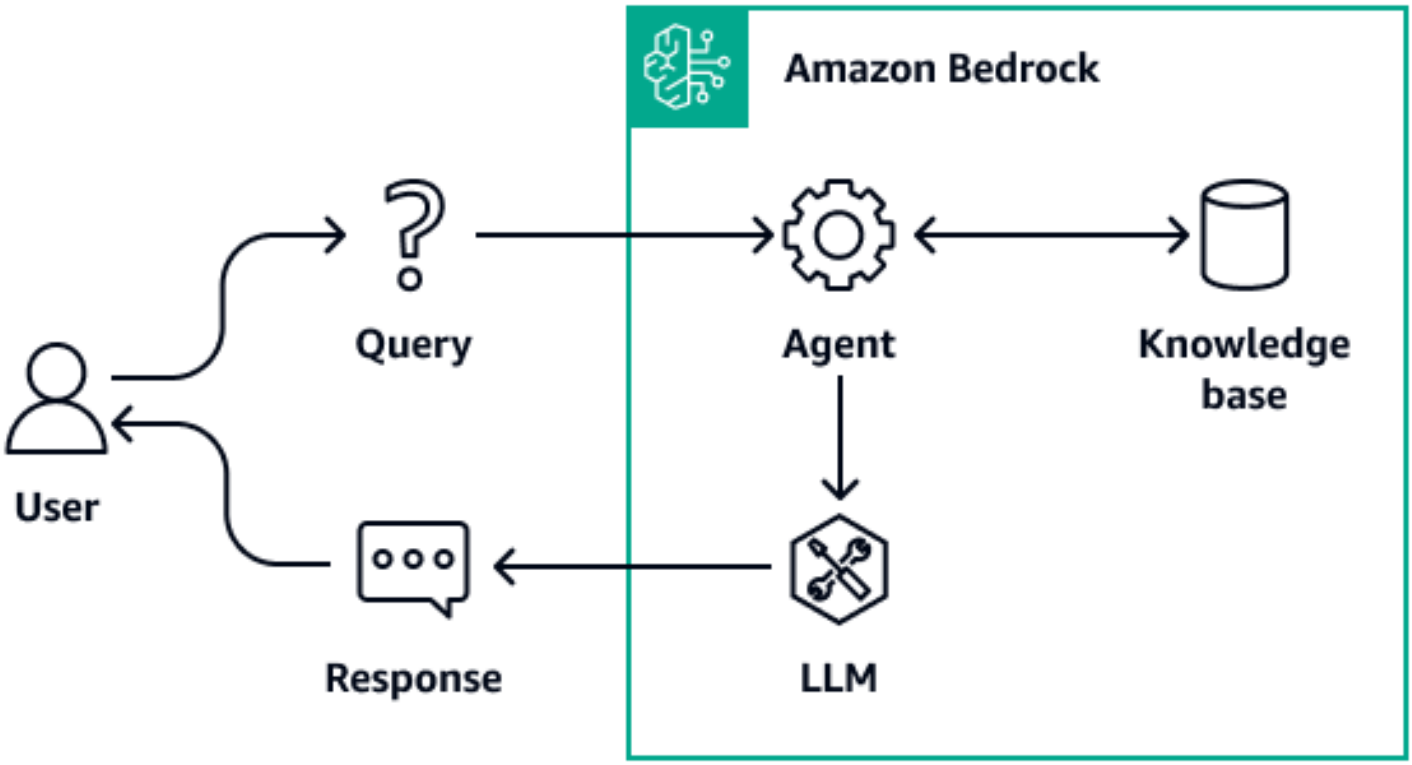
有关如何在这些选项之间进行选择的更多信息，请参阅本指南选择 “检索增强生成” 选项 AWS中的。

Amazon Bedrock 知识库

Amazon Bedrock 是一项完全托管的服务，它通过统一的 API 提供来自领先的人工智能初创公司和亚马逊的高性能基础模型 (FMs) 供您使用。知识库是 Amazon Bedrock 的一项功能，可帮助您实施从摄取到检索和即时增强的整个 RAG 工作流程。无需构建与数据源的自定义集成，也无需管理数据流。内置会话上下文管理，因此您的生成式 AI 应用程序可以轻松支持多回合对话。

在您指定数据位置后，Amazon Bedrock 的知识库会在内部提取文档，将它们分成文本块，将文本转换为嵌入内容，然后将嵌入内容存储在您选择的矢量数据库中。Amazon Bedrock 管理和更新嵌入内容，使矢量数据库与数据保持同步。有关知识库工作原理的更多信息，请参阅 Amazon Bedrock 知识库的工作原理。

如果您向 Amazon Bedrock 代理添加知识库，则代理会根据用户输入的内容识别相应的知识库。代理检索相关信息并将这些信息添加到输入提示中。更新后的提示为模型提供了更多上下文信息以生成响应。为了提高透明度并最大限度地减少幻觉，从知识库中检索到的信息可以追溯到其来源。



Amazon Bedrock APIs 为 RAG 支持以下两个：

- RetrieveAndGenerate— 您可以使用此 API 来查询您的知识库并根据其检索到的信息生成响应。在内部，Amazon Bedrock 将查询转换为嵌入式，查询知识库，使用搜索结果作为上下文信息来增加提示，然后返回 LLM 生成的响应。Amazon Bedrock 还管理对话的短期记忆，以提供更符合情境的结果。
- 检索-您可以使用此 API 来查询您的知识库，其中包含直接从知识库中检索的信息。您可以使用从此 API 返回的信息来处理检索到的文本，评估其相关性，或者开发用于生成响应的单独工作流程。在内部，Amazon Bedrock 会将查询转换为嵌入内容，搜索知识库并返回相关结果。您可以在搜索结果之上构建其他工作流程。例如，您可以使用该LangChain`AmazonKnowledgeBasesRetriever`插件将 RAG 工作流程集成到生成式 AI 应用程序中。

有关架构模式示例和使用 step-by-step说明 APIs，请参阅知识库现在在 Amazon Bedrock 中提供完全托管的 RAG 体验（AWS 博客文章）。有关如何使用 `RetrieveAndGenerate` API 为基于聊天的智能应用程序构建 RAG 工作流程的更多信息，请参阅使用 Amazon Bedrock 知识库构建上下文聊天机器人应用程序（博客文章）。AWS

知识库的数据来源

您可以将您的专有数据连接到知识库。配置数据源连接器后，您可以将数据与知识库同步或保持最新状态，并使您的数据可供查询。Amazon Bedrock 知识库支持与以下数据源的连接：

- 亚马逊简单存储服务 (Amazon S3) S imple Service — 您可以使用控制台或 API 将亚马逊 S3 存储桶连接到亚马逊 Bedrock 知识库。知识库会提取存储桶中的文件并将其编入索引。这种类型的数据源支持以下功能：
  - 文档元数据字段-您可以包含一个单独的文件来指定 Amazon S3 存储桶中文件的元数据。然后，您可以使用这些元数据字段来筛选和提高响应的相关性。
  - 包含或排除过滤器-您可以在抓取时包含或排除某些内容。
  - 增量同步-跟踪内容更改，并且仅抓取自上次同步以来发生更改的内容。
- Confluence— 您可以使用控制台或 API 将Atlassian Confluence实例连接到 Amazon Bedrock 知识库。这种类型的数据源支持以下功能：
  - 自动检测主文档字段-自动检测和搜索元数据字段。您可以使用这些字段进行筛选。
  - 包含或排除内容过滤器-您可以通过在空间、页面标题、博客标题、评论、附件名称或扩展名上使用前缀或正则表达式模式来包含或排除某些内容。
  - 增量同步-跟踪内容更改，并且仅抓取自上次同步以来发生更改的内容。
  - OAuth 2.0 身份验证，使用 Confluence API 令牌进行身份验证-身份验证凭据存储在 AWS Secrets Manager。
- Microsoft SharePoint— 您可以使用控制台或 API 将SharePoint实例连接到知识库。这种类型的数据源支持以下功能：
  - 自动检测主文档字段-自动检测和搜索元数据字段。您可以使用这些字段进行筛选。
  - 包含或排除内容过滤器-您可以通过在主页标题、事件名称和文件名（包括其扩展名）上使用前缀或正则表达式模式来包含或排除某些内容。
  - 增量同步-跟踪内容更改，并且仅抓取自上次同步以来发生更改的内容。
  - OAuth 2.0 身份验证-身份验证凭据存储在 AWS Secrets Manager。
- Salesforce— 您可以使用控制台或 API 将Salesforce实例连接到知识库。这种类型的数据源支持以下功能：
  - 自动检测主文档字段-自动检测和搜索元数据字段。您可以使用这些字段进行筛选。
  - 包含或排除内容过滤器-您可以使用前缀或正则表达式模式包含或排除某些内容。有关您可以应用筛选条件的内容类型列表，请参阅 A ma zon Bedrock 文档中的包含/排除筛选条件。
  - 增量同步-跟踪内容更改，并且仅抓取自上次同步以来发生更改的内容。

  - OAuth 2.0 身份验证-身份验证凭据存储在 AWS Secrets Manager。
- Web Crawler — Amazon Bedrock Web Crawler 连接并抓取您提供的内容。 URLs 支持以下功能：
  - 选择多个 URLs 进行爬行
  - 遵守标准的 robots.txt 指令，例如`Allow`和 `Disallow`
  - 排除与模式 URLs 相匹配的内容
  - 限制爬行速度
  - 在 Amazon 中 CloudWatch，查看已抓取的每个网址的状态

有关可以连接到 Amazon Bedrock 知识库的数据源的更多信息，请参阅为您的知识库创建数据源连接器。

知识库的矢量数据库

在知识库和数据源之间建立连接时，必须配置矢量数据库，也称为矢量存储。矢量数据库是 Amazon Bedrock 存储、更新和管理代表您的数据的嵌入内容的地方。每个数据源都支持不同类型的矢量数据库。要确定哪个矢量数据库可用于您的数据源，请参阅数据源类型。

如果您希望让 Amazon Bedrock 在 Amazon OpenSearch Serverless 中自动为您创建矢量数据库，则可以在创建知识库时选择此选项。但是，您也可以选择建立自己的矢量数据库。如果您设置了自己的矢量数据库，请参阅自己的矢量存储的先决条件以获取知识库。每种类型的矢量数据库都有自己的先决条件。

根据您的数据源类型，Amazon Bedrock 知识库支持以下矢量数据库：

- Amazon OpenSearch 无服务器
- Amazon Aurora PostgreSQL 兼容版
- Pinecone（Pinecone 文档）
- Redis Enterprise Cloud（Redis 文档）
- MongoDB Atlas（MongoDB 文档）

Amazon Q Business

Amazon Q Busines s 是一款完全托管的、基于 Generative AI 的助手，您可以将其配置为根据企业数据回答问题、提供摘要、生成内容和完成任务。它允许最终用户从企业数据源收到带有引文的即时权限感知响应。

主要 功能

Amazon Q Business 的以下功能可以帮助您构建基于 RAG 的生产级生成式 AI 应用程序：

- 内置连接器 — Amazon Q Business 支持 40 多种类型的连接器，例如、Adobe Experience Manager (AEM)SalesforceJira、和的连接器Microsoft SharePoint。有关完整列表，请参阅支持的连接器。如果您需要不支持的连接器，则可以使用亚马逊将数据源中的数据提取 AppFlow到亚马逊简单存储服务 (Amazon S3) Simple Storage Service，然后将 Amazon Q Business 连接到 Amazon S3 存储桶。有关 Amazon AppFlow 支持的数据源的完整列表，请参阅支持的应用程序。
- 内置索引管道 — Amazon Q Business 为向量数据库中的数据编制索引提供了内置管道。您可以使用 AWS Lambda 函数为索引管道添加预处理逻辑。
- 索引选项 — 您可以在 Amazon Q Business 中创建和配置原生索引，并使用 Amazon Q Business 检索器从该索引中提取数据。或者，您可以使用预先配置的 Amazon Kendra 索引作为检索器。有关更多信息，请参阅为 Amazon Q Business 应用程序创建检索器。
- 基础模型 — 亚马逊 Q Business 使用 Amazon Bedrock 支持的基础模型。有关完整列表，请参阅 Amazon Bedrock 中支持的基础模型。
- 插件 — Amazon Q Business 提供了使用插件与目标系统集成的功能，例如自动汇总工单信息和创建工单Jira。配置完成后，插件可以支持读取和写入操作，从而帮助您提高最终用户的工作效率。Amazon Q Business 支持两种类型的插件：内置插件和自定义插件。
- 护栏 — Amazon Q Business 支持全局控制和主题级控制。例如，这些控件可以检测提示中的个人身份信息 (PII)、滥用信息或敏感信息。有关更多信息，请参阅 Amazon Q Business 中的管理员控制和护栏。
- 身份管理 — 借助 Amazon Q Business，您可以管理用户及其对基于 RAG 的生成式 AI 应用程序的访问权限。有关更多信息，请参阅 Amazon Q Business 的身份和访问管理。此外，Amazon Q Business 连接器还会索引与文档本身一起附加到文档的访问控制列表 (ACL) 信息。然后，Amazon Q Business 将其索引的 ACL 信息存储在 Amazon Q Business 用户商店中，以创建用户和群组映射，并根据最终用户对文档的访问权限筛选聊天响应。有关更多信息，请参阅数据源连接器概念。
- 文档扩充-文档扩充功能可帮助您控制将哪些文档和文档属性提取到索引中，以及它们的摄取方式。这可以通过两种方法来实现：
  - 配置基本操作-使用基本操作在数据中添加、更新或删除文档属性。例如，您可以通过选择删除任何与 PII 相关的文档属性来清理 PII 数据。
  - 配置 Lambda 函数-使用预配置的 Lambda 函数对数据执行更自定义、更高级的文档属性操作逻辑。例如，企业数据可能以扫描图像的形式存储。在这种情况下，您可以使用 Lambda 函数对扫描的文档运行光学字符识别 (OCR)，以从中提取文本。然后，在摄取过程中，每个扫描的文档都

作为文本文档处理。最后，在聊天期间，Amazon Q 将在生成响应时考虑从扫描文档中提取的文本数据。

在实施解决方案时，您可以选择将两种文档充实方法结合使用。您可以使用基本操作对数据进行首次解析，然后使用 Lambda 函数进行更复杂的操作。有关更多信息，请参阅 Amazon Q Business 中的文档充实。

- 集成 — 创建 Amazon Q Business 应用程序后，您可以将其集成到其他应用程序中，例如Slack或 Microsoft Teams。例如，请参阅为 Amazon Q Business 部署Slack网关和为 Amazon Q Business 部署Microsoft Teams网关（AWS 博客文章）。

终端用户定制

Amazon Q Business 支持上传可能未存储在贵组织数据源和索引中的文档。上传的文档不会被存储。它们仅可用于上传文档的对话。Amazon Q Business 支持上传特定类型的文档。有关更多信息，请参阅在 Amazon Q Business 中上传文件和聊天。

Amazon Q Business 包含按文档属性筛选功能。管理员和最终用户都可以使用此功能。管理员可以使用属性为最终用户自定义和控制聊天响应。例如，如果数据来源类型是附加到文档的属性，您可以指定仅从特定数据来源生成聊天回复。或者，您可以允许最终用户使用您选择的属性过滤器来限制聊天回复的范围。

最终用户可以在更广泛的 Amazon Q Business 应用程序环境中创建轻量级、专门构建的 Amazon Q 应用程序。Amazon Q 应用程序允许对特定域进行任务自动化，例如为营销团队设计的专用应用程序。

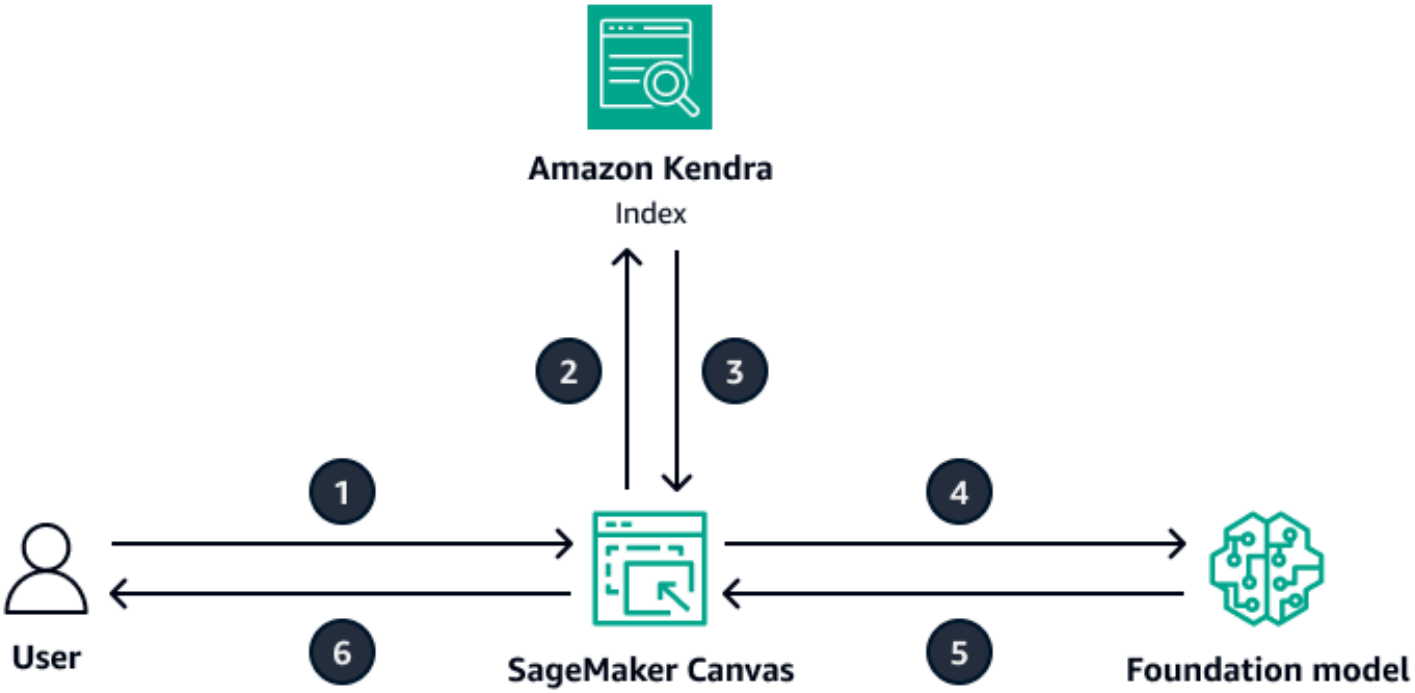
亚马逊 A SageMaker I Canvas

Amazon SageMaker AI Canvas 可帮助您使用机器学习生成预测，而无需编写任何代码。它提供了一个无需代码的可视化界面，使您能够准备数据、构建和部署机器学习模型，从而在统一的环境中简化 end-to-end机器学习生命周期。数据准备、模型开发、偏差检测、可解释性和监控的复杂性被抽象到直观的界面背后。用户无需成为 SageMaker AI 或机器学习运营 (MLOps) 专家，即可使用 AI Canvas SageMaker as 开发、操作和监控模型。

在 A SageMaker I Canvas 中，RAG 功能是通过无代码的文档查询功能提供的。您可以使用 Amazon Kendra 索引作为底层企业搜索来丰富 A SageMaker I Canvas 中的聊天体验。有关更多信息，请参阅通过文档查询从文档中提取信息。

将 SageMaker AI Canvas 连接到 Amazon Kendra 索引需要一次性设置。作为域配置的一部分，云管理员可以选择一个或多个 Kendra 索引，供用户在与 Canvas 交互时查询。 SageMaker 有关如何启用文档查询功能的说明，请参阅 Amazon A SageMaker I Canvas 使用入门。

SageMaker AI Canvas 管理亚马逊 Kendra 与所选基础模型之间的底层通信。有关 SageMaker AI Canvas 支持的基础模型的更多信息，请参阅 A I Canvas 中的生成式 SageMaker AI 基础模型。下图显示了云管理员将 SageMaker AI Canvas 连接到 Amazon Kendra 索引后，文档查询功能的工作原理。



下图显示了如下工作流：

1. 用户在 SageMaker AI Canvas 中开始新的聊天，打开查询文档，选择目标索引，然后提交问题。
2. SageMaker AI Canvas 使用该查询在 Amazon Kendra 索引中搜索相关数据。
3. SageMaker AI Canvas 从亚马逊 Kendra 索引中检索数据及其来源。
4. SageMaker AI Canvas 更新提示以包含从 Amazon Kendra 索引中检索到的上下文，并将提示提交给基础模型。
5. 基础模型使用原始问题和检索到的上下文来生成答案。
6. SageMaker AI Canvas 向用户提供生成的答案。它包括对用于生成响应的数据源（例如文档）的引用。

自定义检索增强生成架构开启 AWS

上一节介绍如何使用完全托管 AWS 服务 的检索增强生成 (RAG)。但是，某些用例需要对系统组件进行更多控制，例如检索器或 LLM（也称为生成器）。例如，您可能需要灵活地选择自己的矢量数据库或访问不支持的数据源。对于这些用例，您可以构建自定义 RAG 架构。

本节包含以下主题：

- 适用于 RAG 工作流程的检索器
- 用于 RAG 工作流程的生成器

有关本节中如何在检索器和生成器选项之间进行选择的更多信息，请参阅本指南选择 “检索增强生成” 选项 AWS中的。

适用于 RAG 工作流程的检索器

本节介绍如何建造寻回犬。您可以使用完全托管的语义搜索解决方案，例如 Amazon Kendra，也可以使用 AWS 矢量数据库构建自定义语义搜索。

在查看检索器选项之前，请务必了解矢量搜索过程的三个步骤：

1. 您可以将需要编制索引的文档分成较小的部分。这称为分块。
2. 您可以使用名为 embedd in g 的过程将每个区块转换为数学向量。然后，为矢量数据库中的每个向量编制索引。您用来为文档编制索引的方法会影响搜索的速度和准确性。索引方法取决于矢量数据库及其提供的配置选项。
3. 您可以使用相同的过程将用户查询转换为向量。检索器在矢量数据库中搜索与用户查询向量相似的向量。相似度是通过使用欧几里得距离、余弦距离或点积等度量来计算的。

本指南介绍如何使用以下服务 AWS 服务 或第三方服务在上构建自定义检索层 AWS：

- Amazon Kendra
- 亚马逊 OpenSearch 服务
- 亚马逊 Aurora PostgreSQL 和 pgvector
- Amazon Neptune Analytics
- Amazon MemoryDB

- Amazon DocumentDB
- Pinecone
- MongoDB Atlas
- Weaviate

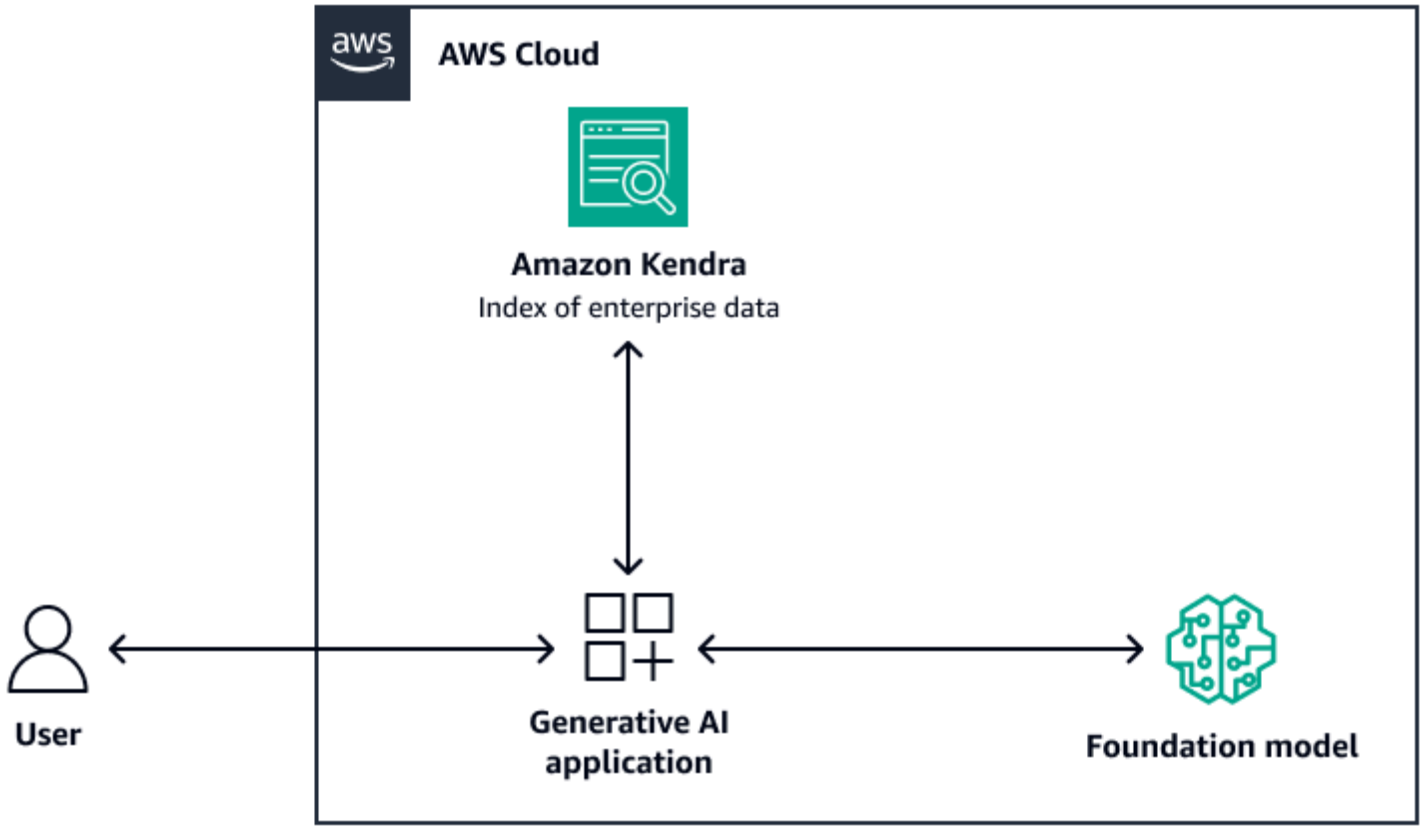
Amazon Kendra

Amazon Kendra 是一项完全托管的智能搜索服务，它使用自然语言处理和高级机器学习算法，从您的数据中返回搜索问题的具体答案。Amazon Kendra 可帮助您直接从多个来源获取文档，并在文档成功同步后对其进行查询。同步过程创建了在摄取的文档上创建矢量搜索所需的必要基础架构。因此，Amazon Kendra 不需要矢量搜索过程的传统三个步骤。初始同步后，您可以使用定义的时间表来处理正在进行的摄取。

以下是使用 Amazon Kendra for RAG 的优势：

- 您不必维护矢量数据库，因为 Amazon Kendra 负责处理整个矢量搜索过程。
- Amazon Kendra 包含用于常用数据源的预建连接器，例如数据库、网站抓取工具、Amazon S3 存储桶、Microsoft SharePoint实例和实例。Atlassian Confluence AWS 合作伙伴开发的连接器可用，例如Box和的连接器GitLab。
- Amazon Kendra 提供访问控制列表 (ACL) 筛选功能，仅返回最终用户有权访问的文档。
- Amazon Kendra 可以根据元数据（例如日期或源存储库）提高响应速度。

下图显示了使用 Amazon Kendra 作为 RAG 系统的检索层的示例架构。有关更多信息，请参阅使用 Amazon Kendra 在企业数据上快速构建高精度的生成式 AI 应用程序LangChain，以及大型语言模型 AWS （博客文章）。



对于基础模型，你可以使用 Amazon Bedrock 或通过 Amazon AI 部署的 LLM。 SageMaker JumpStart您可以使用 AWS Lambda LangChain来协调用户、Amazon Kendra 和 LLM 之间的流程。要构建使用 Amazon Kendra 等的 RAG 系统，请参阅 Amazon Kendra LangChain E LLMs xt ension LangChain s 存储库。 GitHub

亚马逊 OpenSearch 服务

Amazon S OpenSearch ervic e 为 k 最近邻 (k-nn) 搜索提供了内置 ML 算法，以便执行向量搜索。OpenSearch 该服务还为 Amazon EMR Serverless 提供了矢量引擎。您可以使用此矢量引擎来构建具有可扩展和高性能矢量存储和搜索功能的 RAG 系统。有关如何使用无服务器构建 RAG 系统的更多信息，请参阅使用适用 OpenSearch 于 Amazon Serverless 和 Amazon Bedrock Claude 模型的矢量引擎构建可扩展的无服务器 OpenSearch RAG 工作流程（博客文章）。AWS

以下是使用 OpenSearch 服务进行矢量搜索的优势：

- 它提供对矢量数据库的完全控制，包括使用 OpenSearch Serverless 构建可扩展的矢量搜索。
- 它提供了对分块策略的控制。

- 它使用来自非公制空间库 (NMSLIB)、Faiss 和 A pache Lucen e 库的近似最近邻 (ANN) 算法来支持 k-nn 搜索。您可以根据用例更改算法。有关通过 OpenSearch 服务自定义矢量搜索的选项的更多信息，请参阅 Amazon Ser OpenSearch vice 矢量数据库功能说明（AWS 博客文章）。
- OpenSearch Serverless 作为向量索引与 Amazon Bedrock 知识库集成。

亚马逊 Aurora PostgreSQL 和 pgvector

Amazon Aurora PostgreSQL 兼容版是一个完全托管的关系数据库引擎，可帮助您设置、操作和扩展 PostgreSQL 部署。 pgvector 是一个开源 PostgreSQL 扩展，它提供了向量相似度搜索功能。此扩展适用于兼容 Aurora PostgreSQL 和适用于 PostgreSQL 的亚马逊关系数据库服务 (Amazon RDS)。有关如何使用与 Aurora PostgreSQL 兼容和 pgvector 构建基于 RAG 的系统的更多信息，请参阅以下博客文章： AWS

- 使用 SageMaker 亚马逊 AI 和 pgvector 在 PostgreSQL 中构建人工智能驱动的搜索
- 利用 pgvector 和 Amazon Aurora PostgreSQL 进行自然语言处理、聊天机器人和情感分析

以下是使用兼容 pgvector 且兼容 Aurora PostgreSQL 的优点：

- 它支持精确和近似最近邻搜索。它还支持以下相似度量：L2 距离、内积和余弦距离。
- 它支持采用平面压缩的倒置文件 (IVFFlat) 和分层可导航小世界 (HNSW) 索引。
- 您可以将向量搜索与对同一 PostgreSQL 实例中可用的特定域数据的查询相结合。
- 与 Aurora PostgreSQL 兼容，已针对分层缓存进行了优化 I/O 并提供了分层缓存。对于超过可用实例内存的工作负载，pgvector 可以将每秒向量搜索的查询量增加多达 8 倍。

Amazon Neptune Analytics

Amazon Neptune Analytics 是一款内存优化的图形数据库引擎，用于分析。它支持经过优化的图形分析算法、低延迟图形查询和图形遍历中的矢量搜索功能库。它还具有内置的向量相似度搜索功能。它提供了一个端点来创建图形、加载数据、调用查询和执行向量相似度搜索。有关如何构建使用 Neptune Analytics 的基于 RAG 的系统的更多信息，请参阅使用知识图通过亚马逊 Bedrock 和 Amazon Neptune 构建 GraphRag 应用程序（博客文章AWS ）。

以下是使用 Neptune Analytics 的优势：

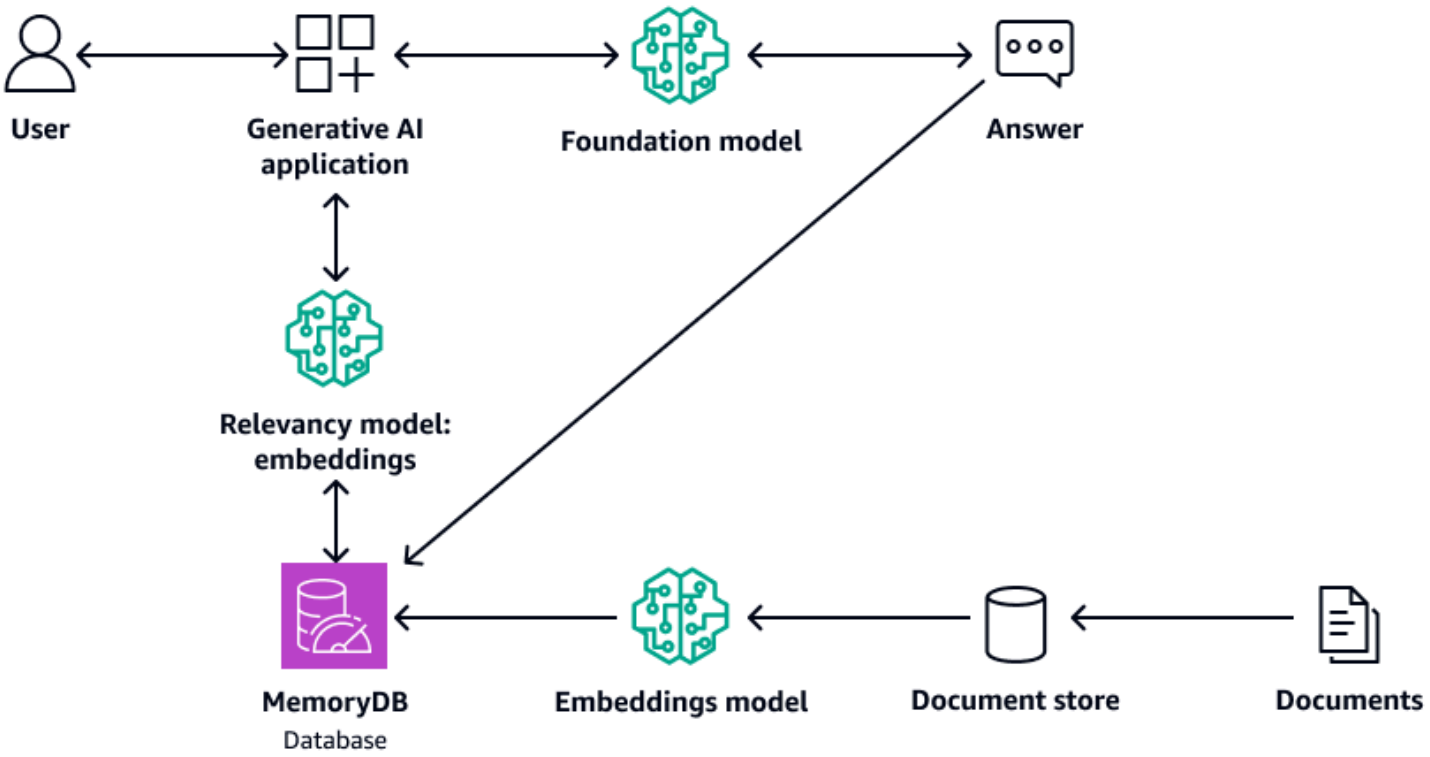
- 您可以在图形查询中存储和搜索嵌入内容。
- 如果您将 Neptune Analytics 与集成LangChain，则此架构支持自然语言图形查询。

- 这种架构将大型图形数据集存储在内存中。

Amazon MemoryDB

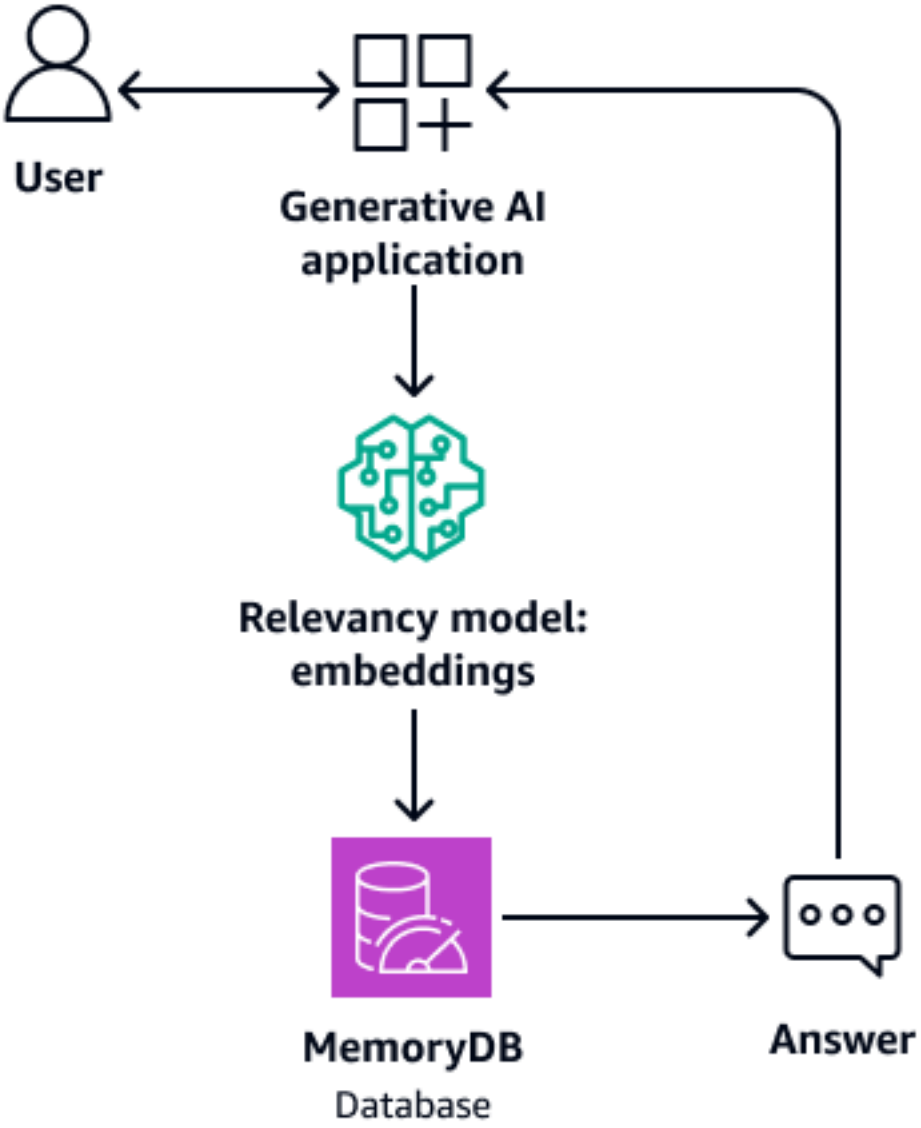
Amazon MemoryDB 是一项耐用的内存数据库服务，可提供超快的性能。您的所有数据都存储在内存中，内存支持微秒读取、个位数毫秒写入延迟和高吞吐量。M@@ emoryDB 的矢量搜索扩展了 MemoryDB 的功能，可以与现有的 MemoryDB 功能结合使用。有关更多信息，请参阅开启的 LLM 和 RAG 存储库中的问题解答。 GitHub

下图显示了使用 MemoryDB 作为矢量数据库的示例架构。



以下是使用 MemoryDB 的优点：

- 它同时支持 Flat 和 HNSW 索引算法。欲了解更多信息，请参阅 Amazon MemoryDB 的矢量搜索现已在新闻博客上正式推出 AWS
- 它也可以用作基础模型的缓冲存储器。这意味着先前回答的问题将从缓冲区中检索，而不是再次进行检索和生成过程。下图显示了此流程。

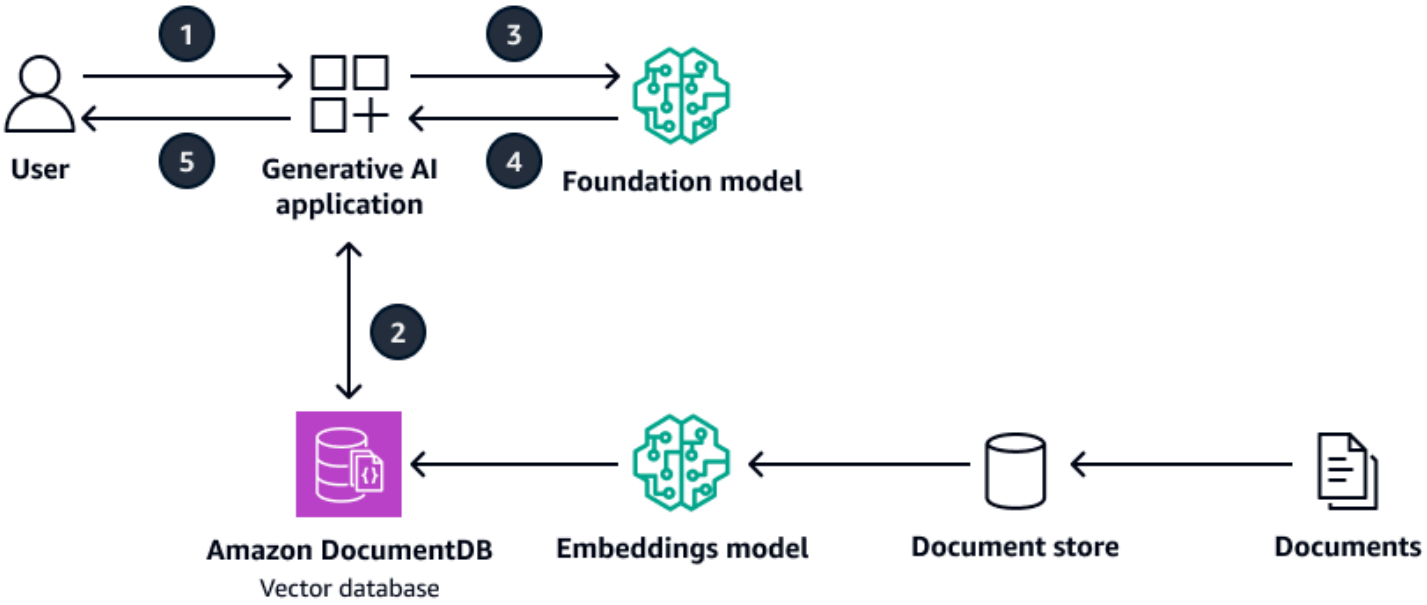


- 由于它使用内存数据库，因此该架构为语义搜索提供了个位数毫秒的查询时间。
- 它在95-99%的召回率下每秒最多提供33,000个查询，在召回率超过99%时每秒提供26,500个查询。欲了解更多信息，请参阅 re AWS : Invent 2023 — 超低延迟向量搜索 Amazon Memory DB 视频。YouTube

Amazon DocumentDB

Amazon DocumentDB（与 MongoDB 兼容）是一种快速、可靠、完全托管的数据库服务。它可以轻松地在云中设置、操作和扩展MongoDB兼容的数据库。Amazon DocumentDB 的矢量搜索将基于 JSON 的文档数据库的灵活性和丰富的查询功能与矢量搜索的强大功能相结合。有关更多信息，请参阅开启的 LLM 和 RAG 存储库中的问题解答。 GitHub

下图显示了使用 Amazon DocumentDB 作为矢量数据库的示例架构。



下图显示了如下工作流：

1. 用户向生成式 AI 应用程序提交查询。
2. 生成式 AI 应用程序在 Amazon DocumentDB 矢量数据库中执行相似度搜索并检索相关的文档摘录。
3. 生成式 AI 应用程序使用检索到的上下文更新用户查询，并将提示提交给目标基础模型。
4. 基础模型使用上下文生成对用户问题的回应并返回响应。
5. 生成式 AI 应用程序将响应返回给用户。

以下是使用 Amazon DocumentDB 的优势：

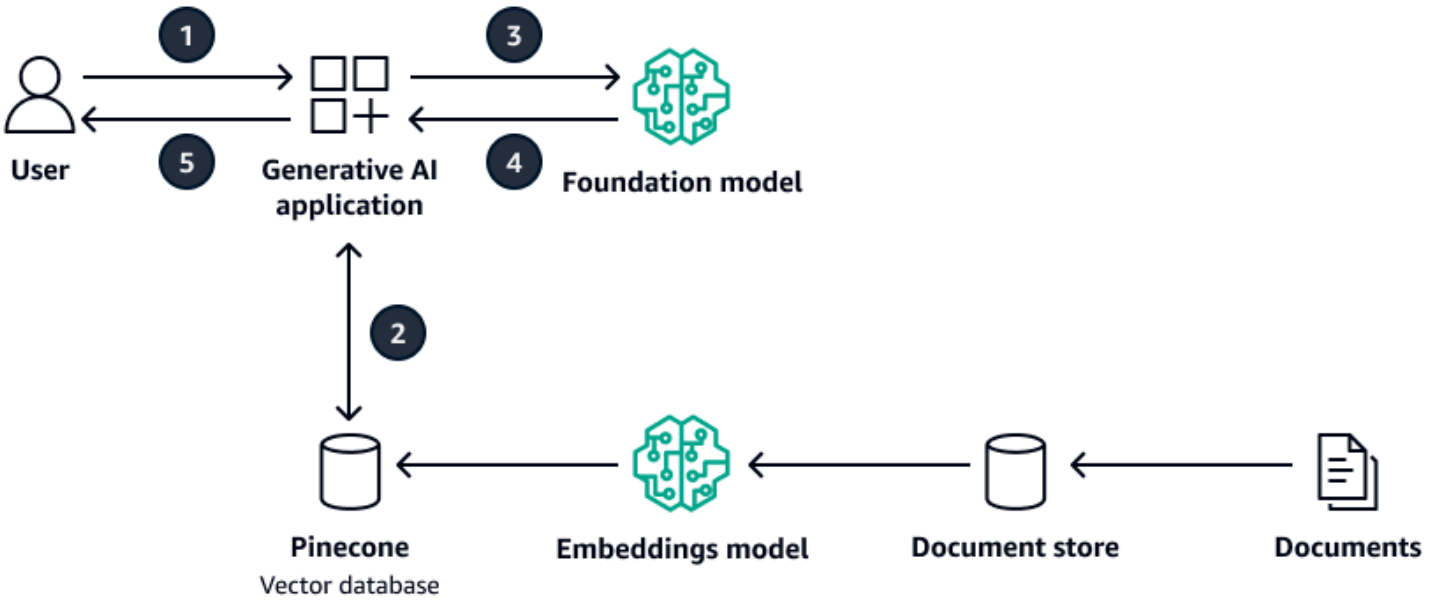
- 它同时支持 HNSW 和 IVFFlat 索引方法。
- 它支持向量数据中多达 2,000 个维度，并支持欧几里得、余弦和点积距离度量。
- 它提供了毫秒级的响应时间。

Pinecone

Pinecone是一个完全托管的矢量数据库，可帮助您将矢量搜索添加到生产应用程序中。可通过以下方式获得AWS Marketplace。计费基于使用量，费用是通过将 pod 价格乘以 pod 数量来计算的。有关如何构建使用的 RAG 系统的更多信息Pinecone，请参阅以下 AWS 博客文章：

- 使用Pinecone矢量数据库和 Amazon AI 的 Llama-2 通过 RAG 缓解幻觉 SageMaker JumpStart
- 使用 Amazon SageMaker AI Studio 使用 Llama 2 构建 RAG 问答解决方案LangChain，并Pinecone进行快速实验

下图显示了Pinecone用作矢量数据库的示例架构。



下图显示了如下工作流：

1. 用户向生成式 AI 应用程序提交查询。
2. 生成式 AI 应用程序在Pinecone矢量数据库中执行相似度搜索并检索相关的文档摘录。
3. 生成式 AI 应用程序使用检索到的上下文更新用户查询，并将提示提交给目标基础模型。
4. 基础模型使用上下文生成对用户问题的回应并返回响应。
5. 生成式 AI 应用程序将响应返回给用户。

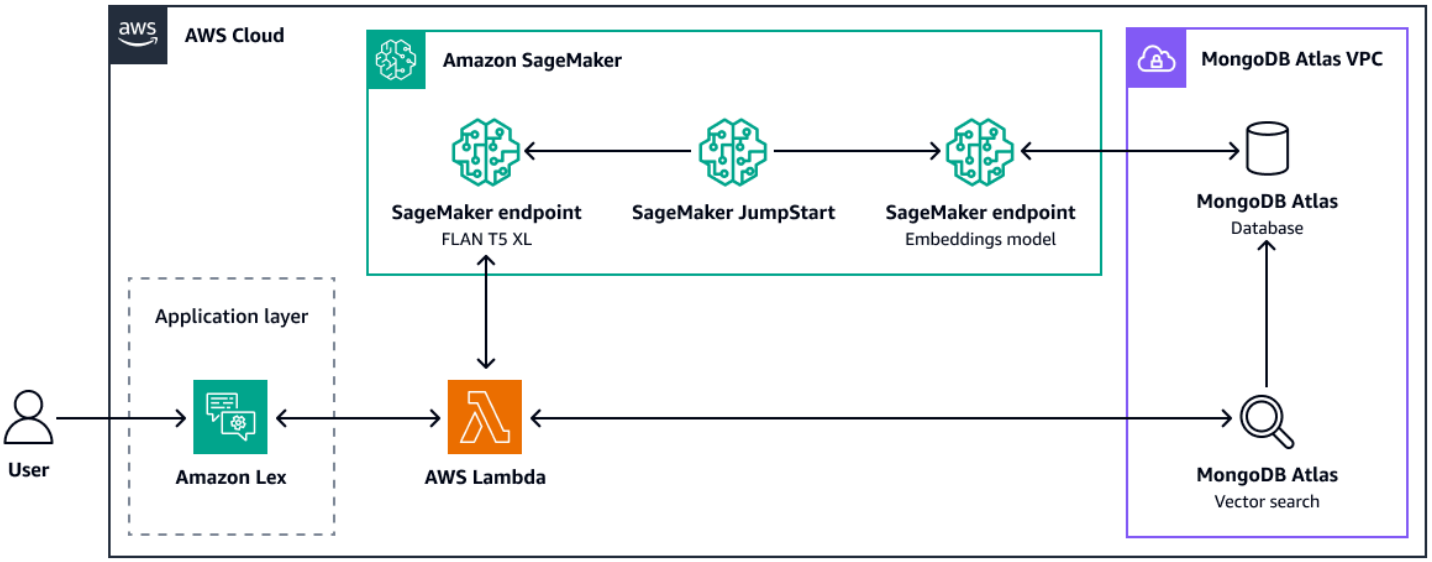
以下是使用的优点Pinecone：

- 它是一个完全托管的矢量数据库，可以省去管理自己的基础设施的开销。
- 它提供了筛选、实时索引更新和关键字提升（混合搜索）等其他功能。

MongoDB Atlas

MongoDB Atlas是一个完全托管的云数据库，可处理在上部署和管理部署的所有复杂性 AWS。您可以使用矢量搜索MongoDB Atlas将矢量嵌入存储在MongoDB数据库中。Amazon Bedrock 知识库MongoDB Atlas支持矢量存储。有关更多信息，请参阅MongoDB文档中的 Amazon Bedrock 知识库集成入门。

有关如何对 RAG 使用MongoDB Atlas矢量搜索的更多信息，请参阅使用 Amazon AI 检索增强生成、LangChain Amazon A SageMaker I JumpStart 和MongoDB Atlas语义搜索（博客文章）。AWS下图显示了此博客文章中详细介绍的解决方案架构。



以下是使用MongoDB Atlas矢量搜索的优点：

- 您可以使用现有的实现MongoDB Atlas来存储和搜索矢量嵌入。
- 您可以使用MongoDB查询 API 来查询矢量嵌入。
- 您可以独立缩放矢量搜索和数据库。
- 向量嵌入存储在源数据（文档）附近，这提高了索引性能。

Weaviate

Weaviate是一个流行的开源、低延迟矢量数据库，支持多模态媒体类型，例如文本和图像。该数据库存储对象和向量，将矢量搜索与结构化过滤相结合。有关使用Weaviate和 Amazon Bedrock 构建 RAG 工作流程的更多信息，请参阅使用 Amazon Bedrock 中的 Cohere 基础模型和Weaviate矢量数据库构建企业级生成人工智能解决方案（博客文章）。 AWS MarketplaceAWS

以下是使用的优点Weaviate：

- 它是开源的，并由强大的社区提供支持。
- 它专为混合搜索（矢量和关键字）而构建。
- 您可以将其部署 AWS 为托管软件即服务 (SaaS) 产品或 Kubernetes 集群。

用于 RAG 工作流程的生成器

大型语言模型 (LLMs) 是基于大量数据进行预训练的非常大的深度学习模型。它们非常灵活。 LLMs 可以执行各种任务，例如回答问题、总结文档、翻译语言和完成句子。它们有可能破坏内容创作以及人们使用搜索引擎和虚拟助手的方式。虽然并不完美，但要 LLMs 表现出根据相对较小的提示或输入数量做出预测的非凡能力。

LLMs 是 RAG 解决方案的关键组成部分。对于自定义 RAG 架构，有两个 AWS 服务 可用作主要选项：

- Amazon Bedrock 是一项完全托管的服务， LLMs 由领先的人工智能公司和亚马逊提供给您通过统一的 API 供您使用。
- Amazon SageMaker AI JumpStart 是一个机器学习中心，提供基础模型、内置算法和预构建的机器学习解决方案。借助 SageMaker AI JumpStart，您可以访问预训练模型，包括基础模型。您也可以使用自己的数据对预训练模型进行微调。

Amazon Bedrock

Amazon Bedrock 提供来自Anthropic、、Stability AI、Meta、CohereAI21 LabsMistral AI、和亚马逊的行业领先机型。有关完整列表，请参阅 Amazon Bedrock 中支持的基础模型。Amazon Bedrock 还允许您使用自己的数据自定义模型。

您可以评估模型性能，以确定哪些模型最适合您的 RAG 用例。您可以测试最新的型号，也可以进行测试，看看哪些功能和特性可以提供最佳结果和最优惠的价格。AnthropicClaude Sonnet 模型是 RAG 应用程序的常见选择，因为它擅长执行各种任务，并且具有高度的可靠性和可预测性。

SageMaker AI JumpStart

SageMaker AI JumpStart 为各种问题类型提供经过预训练的开源模型。在部署之前，您可以逐步训练和微调这些模型。您可以通过 A mazon AI Studio 中的 SageMaker 人工智能 JumpStart 登录页面访问预训练模型、解决方案模板和示例，也可以使用 SageMaker A SageMaker I Python SDK。

SageMaker AI JumpStart 为内容编写、代码生成、问题解答、文案写作、摘要、分类、信息检索等用例提供了 state-of-the-art基础模型。使用 JumpStart 基础模型构建自己的生成式 AI 解决方案，并将自定义解决方案与其他 SageMaker AI 功能集成。有关更多信息，请参阅 Amazon A SageMaker I 入门 JumpStart。

SageMaker AI JumpStart 载入并维护公开可用的基础模型，供您访问、自定义和集成到您的机器学习生命周期中。有关更多信息，请参阅公开可用的基础模型。 SageMaker 人工智能 JumpStart 还包括来自第三方提供商的专有基础模型。有关更多信息，请参阅专有基础模型。

选择 “检索增强生成” 选项 AWS

本指南的完全托管 RAG 选项和自定义 RAG 架构部分描述了构建基于 RAG 的搜索解决方案的各种方法。 AWS本节介绍如何根据您的用例在这些选项之间进行选择。在某些情况下，可能有多个选项起作用。在这种情况下，选择取决于实施的难易程度、组织中可用的技能以及公司的政策和标准。

我们建议您按以下顺序考虑完全托管和自定义 RAG 选项，并选择适合您用例的第一个选项：

1. 使用 Amazon Q Business，除非：
   - 此服务在您中不可用 AWS 区域，您的数据也无法移动到可用的区域
   - 你有特定的理由自定义 RAG 工作流程
   - 你想使用现有的矢量数据库或特定的 LLM
2. 使用 Amazon Bedrock 知识库除非：
   - 您的矢量数据库不受支持
   - 你有特定的理由自定义 RAG 工作流程
3. 将 Amazon Kendra 与你选择的生成器结合使用，除非：
   - 你想选择自己的矢量数据库
   - 你想自定义分块策略
4. 如果你想更好地控制检索器并想要选择自己的矢量数据库：
   - 如果您没有现有的矢量数据库，也不需要低延迟或图形查询，请考虑使用 Amazon OpenSearch 服务。
   - 如果你有现有的PostgreSQL矢量数据库，可以考虑使用 Amazon Aurora Postgre pgvector SQL 和选项。
   - 如果你需要低延迟，可以考虑使用内存选项，例如 Amazon Memor yDB 或 Amazon DocumentDB。
   - 如果您想将矢量搜索与图表查询相结合，可以考虑使用 Amazon Neptune Analytics。
   - 如果您已经在使用第三方矢量数据库，或者从中找到了具体的好处，请考虑PineconeMongoDB Atlas、和Weaviate。
5. 如果你想选择法学硕士：
   - 如果您使用 Amazon Q Business，则无法选择 LLM。
   - 如果您使用 Amazon Bedrock，则可以选择其中一种支持的基础模型。
   - 如果您使用 Amazon Kendra 或自定义矢量数据库，则可以使用本指南中描述的生成器之一，也可以使用自定义 LLM。

> **Note**
>
> 您还可以使用自定义文档对现有 LLM 进行微调，以提高其响应的准确性。有关更多信息，请参阅本指南中的比较 RAG 和微调。

6. 如果你想使用现有的 Amazon A SageMaker I Canvas 实现，或者你想比较不同的 RAG 响应 LLMs，可以考虑使用 Amazon A SageMaker I Canvas。

结论

本指南描述了在上构建检索增强生成 (RAG) 系统的各种选项。 AWS您可以从完全托管的服务开始，例如亚马逊 Q Business 和 Amazon Bedrock 知识库。如果您想更好地控制 RAG 工作流程，可以选择自定义检索器。对于生成器，您可以使用 API 在 Amazon Bedrock 中调用支持的 LLM，也可以使用 Amazon AI 部署自己的 LLM。 SageMaker JumpStart查看选择 RAG 选项中的建议，以确定哪个选项最适合您的用例。为您的用例选择最佳选项后，请使用本指南中提供的参考资料开始构建基于 RAG 的应用程序。

文档历史记录

下表介绍了本指南的一些重要更改。如果您希望收到有关未来更新的通知，可以订阅 RSS 源。

变更	说明	日期
初次发布	—	2024 年 10 月 28 日

AWS 规范性指导词汇表

以下是 AWS 规范性指导提供的策略、指南和模式中的常用术语。若要推荐词条，请使用术语表末尾的提供反馈链接。

数字

7 R

将应用程序迁移到云中的 7 种常见迁移策略。这些策略以 Gartner 于 2011 年确定的 5 R 为基础，包括以下内容：

- 重构/重新架构：充分利用云原生功能来提高敏捷性、性能和可扩展性，以迁移应用程序并修改其架构。这通常涉及到移植操作系统和数据库。示例：将本地 Oracle 数据库迁移到 Amazon Aurora PostgreSQL 兼容版。
- 更换平台：将应用程序迁移到云中，并进行一定程度的优化，以利用云功能。示例：将本地 Oracle 数据库迁移到 AWS 云中的 Amazon Relational Database Service（Amazon RDS）for Oracle。
- 重新购买：转换到其他产品，通常是从传统许可转向 SaaS 模式。示例：将客户关系管理（CRM）系统迁移到 Salesforce.com。
- 重新托管（直接迁移）：将应用程序迁移到云中，无需进行任何更改即可利用云功能。示例：将本地 Oracle 数据库迁移到 AWS 云中 EC2 实例上的 Oracle。
- 重新放置（虚拟机监控器级直接迁移）：将基础设施迁移到云中，无需购买新硬件、重写应用程序或修改现有操作。您将服务器从本地平台迁移到同一平台的云服务中。示例：将Microsoft Hyper-V应用程序迁移到 AWS。
- 保留（重访）：将应用程序保留在源环境中。其中可能包括需要进行重大重构的应用程序，并且您希望将工作推迟到以后，以及您希望保留的遗留应用程序，因为迁移它们没有商业上的理由。
- 停用：停用或删除源环境中不再需要的应用程序。

A

ABAC

请参阅基于属性的访问控制。

抽象服务

请参阅托管服务。

ACID

请参阅原子性、一致性、隔离性、持久性。

主动-主动迁移

一种数据库迁移方法，在这种方法中，源数据库和目标数据库保持同步（通过使用双向复制工具或双写操作），两个数据库都在迁移期间处理来自连接应用程序的事务。这种方法支持小批量、可控的迁移，而不需要一次性割接。它比主动-被动迁移更灵活，但工作量更大。

主动-被动迁移

一种数据库迁移方法，在这种方法中，源数据库和目标数据库保持同步，但在将数据复制到目标数据库时，只有源数据库处理来自连接应用程序的事务。目标数据库在迁移期间不接受任何事务。

聚合函数

一种 SQL 函数，它对一组行进行操作并计算该组的单个返回值。聚合函数的示例包括 SUM 和 MAX。

AI

请参阅人工智能。

AIOps

请参阅人工智能运营。

匿名化

永久删除数据集中个人信息的过程。匿名化可以帮助保护个人隐私。匿名化数据不再被视为个人数据。

反模式

一种用于解决反复出现的问题的常用解决方案，而在这类问题中，此解决方案适得其反、无效或不如替代方案有效。

应用程序控制

一种安全方法，仅允许使用经批准的应用程序，以帮助保护系统免受恶意软件的侵害。

应用程序组合

有关组织使用的每个应用程序的详细信息的集合，包括构建和维护该应用程序的成本及其业务价值。这些信息是产品组合发现和分析过程的关键，有助于识别需要进行迁移、现代化和优化的应用程序并确定其优先级。

人工智能（AI）

计算机科学领域致力于使用计算技术执行通常与人类相关的认知功能，例如学习、解决问题和识别模式。有关更多信息，请参阅什么是人工智能？

人工智能操作 (AIOps)

使用机器学习技术解决运营问题、减少运营事故和人为干预以及提高服务质量的过程。有关如何在 AIOps AWS 迁移策略中使用的更多信息，请参阅操作集成指南。

非对称加密

一种加密算法，使用一对密钥，一个公钥用于加密，一个私钥用于解密。您可以共享公钥，因为它不用于解密，但对私钥的访问应受到严格限制。

原子性、一致性、隔离性、持久性（ACID）

一组软件属性，即使在出现错误、电源故障或其他问题的情况下，也能保证数据库的数据有效性和操作可靠性。

基于属性的访问权限控制（ABAC）

根据用户属性（如部门、工作角色和团队名称）创建精细访问权限的做法。有关更多信息，请参阅 AWS Identity and Access Management (I AM) 文档 AWS中的 AB AC。

权威数据来源

存储主要数据版本的位置，被认为是最可靠的信息源。您可以将数据从权威数据来源复制到其他位置，以便处理或修改数据，例如对数据进行匿名化、编辑或假名化。

可用区

中的一个不同位置 AWS 区域 ，不受其他可用区域故障的影响，并向同一区域中的其他可用区提供低成本、低延迟的网络连接。

AWS 云采用框架 (AWS CAF)

该框架包含指导方针和最佳实践 AWS ，可帮助组织制定高效且有效的计划，以成功迁移到云端。AWS CAF将指导分为六个重点领域，称为视角：业务、人员、治理、平台、安全和运营。业务、人员和治理角度侧重于业务技能和流程；平台、安全和运营角度侧重于技术技能和流程。例如，人

员角度针对的是负责人力资源（HR）、人员配置职能和人员管理的利益相关者。从这个角度来看，AWS CAF 为人员发展、培训和沟通提供了指导，以帮助组织为成功采用云做好准备。有关更多信息，请参阅 AWS CAF 网站和 AWS CAF 白皮书。

AWS 工作负载资格框架 (AWS WQF)

一种评估数据库迁移工作负载、推荐迁移策略和提供工作估算的工具。 AWS WQF 包含在 AWS Schema Conversion Tool (AWS SCT) 中。它用来分析数据库架构和代码对象、应用程序代码、依赖关系和性能特征，并提供评测报告。

B

恶意机器人

一种旨在扰乱或伤害个人或组织的机器人。

BCP

请参阅业务连续性计划。

行为图

一段时间内资源行为和交互的统一交互式视图。您可以使用 Amazon Detective 的行为图来检查失败的登录尝试、可疑的 API 调用和类似的操作。有关更多信息，请参阅 Detective 文档中的行为图中的数据。

大端序系统

一个先存储最高有效字节的系统。另请参阅字节顺序。

二进制分类

一种预测二进制结果（两个可能的类别之一）的过程。例如，您的 ML 模型可能需要预测诸如“该电子邮件是否为垃圾邮件？”或“这个产品是书还是汽车？”之类的问题

bloom 筛选条件

一种概率性、内存高效的数据结构，用于测试元素是否为集合的成员。

蓝/绿部署

一种部署策略，您可以创建两个独立但完全相同的环境。在一个环境中运行当前应用程序版本（蓝色），在另一个环境中运行新应用程序版本（绿色）。此策略可帮助您在影响最小的情况下快速回滚。

自动程序

一种通过互联网运行自动任务并模拟人类活动或交互的软件应用程序。有些机器人是有用或有益的，例如在互联网上索引信息的 Web 爬网程序。还有一些被称为恶意机器人的机器人，其目的是扰乱或伤害个人或组织。

僵尸网络

被恶意软件感染并受单方（称为僵尸网络控制者或僵尸网络操作者）控制的僵尸网络。僵尸网络是最著名的扩展机器人及其影响力的机制。

分支

代码存储库的一个包含区域。在存储库中创建的第一个分支是主分支。您可以从现有分支创建新分支，然后在新分支中开发功能或修复错误。为构建功能而创建的分支通常称为功能分支。当功能可以发布时，将功能分支合并回主分支。有关更多信息，请参阅关于分支（GitHub 文档）。

紧急（break-glass）访问

在特殊情况下，通过批准的流程，用户 AWS 账户 可以快速访问他们通常没有访问权限的内容。有关更多信息，请参阅 AWS Well-Architected Guidance 中的 Implement break-glass procedures 指示器。

棕地策略

您环境中的现有基础设施。在为系统架构采用棕地策略时，您需要围绕当前系统和基础设施的限制来设计架构。如果您正在扩展现有基础设施，则可以将棕地策略和全新策略混合。

缓冲区缓存

存储最常访问的数据的内存区域。

业务能力

企业如何创造价值（例如，销售、客户服务或营销）。微服务架构和开发决策可以由业务能力驱动。有关更多信息，请参阅在 AWS上运行容器化微服务白皮书中的围绕业务能力进行组织部分。

业务连续性计划（BCP）

一项计划，旨在应对大规模迁移等破坏性事件对运营的潜在影响，并使企业能够快速恢复运营。

C

CAF

请参阅 AWS 云采用框架。

金丝雀部署

缓慢而渐进地向最终用户发布版本。当您确信无误后，即可部署新版本，并完全替换当前版本。

CCoE

请参阅云卓越中心。

CDC

请参阅更改数据捕获。

更改数据捕获（CDC）

跟踪数据来源（如数据库表）的更改并记录有关更改的元数据的过程。您可以将 CDC 用于各种目的，例如审计或复制目标系统中的更改以保持同步。

混沌工程

故意引入故障或破坏性事件来测试系统的韧性。您可以使用 AWS Fault Injection Service (AWS FIS) 来执行实验，对您的 AWS 工作负载施加压力并评估其响应。

CI/CD

请参阅持续集成和持续交付。

分类

一种有助于生成预测的分类流程。分类问题的 ML 模型预测离散值。离散值始终彼此不同。例如，一个模型可能需要评估图像中是否有汽车。

客户端加密

在目标 AWS 服务 收到数据之前，对数据进行本地加密。

云卓越中心 (CCoE)

一个多学科团队，负责推动整个组织的云采用工作，包括开发云最佳实践、调动资源、制定迁移时间表、领导组织完成大规模转型。有关更多信息，请参阅 AWS 云 企业战略博客上的 CCoE 帖子。

云计算

通常用于远程数据存储和 IoT 设备管理的云技术。云计算通常连接到边缘计算技术。

云运营模型

在 IT 组织中，一种用于构建、完善和优化一个或多个云环境的运营模型。有关更多信息，请参阅构建您的云运营模型。

云采用阶段

组织迁移到 AWS 云中时通常会经历四个阶段：

- 项目 - 出于概念验证和学习目的，开展一些与云相关的项目
- 基础 — 进行基础投资以扩大云采用率（例如，创建着陆区、定义 CCo E、建立运营模型）
- 迁移 - 迁移单个应用程序
- 重塑 - 优化产品和服务，在云中创新

Stephen Orban在 AWS 云 企业战略博客的博客文章《云优先之旅和采用阶段》中定义了这些阶段。有关它们与 AWS 迁移策略的关系的信息，请参阅迁移准备指南。

CMDB

请参阅配置管理数据库。

代码存储库

通过版本控制过程存储和更新源代码和其他资产（如文档、示例和脚本）的位置。常见的云存储库包括 GitHub 或 Bitbucket Cloud。每个版本的代码都称为一个分支。在微服务结构中，每个存储库都专门用于一个功能。单个 CI/CD 管线可以使用多个存储库。

冷缓存

一种空的、填充不足或包含过时或不相关数据的缓冲区缓存。这会影响性能，因为数据库实例必须从主内存或磁盘读取，这比从缓冲区缓存读取要慢。

冷数据

很少访问的数据，且通常是历史数据。查询此类数据时，通常可以接受慢速查询。将这些数据转移到性能较低且成本更低的存储层或类别可以降低成本。

计算机视觉（CV）

一种 AI 领域，它使用机器学习来分析和提取数字图像和视频等视觉格式中的信息。例如，Amazon SageMaker AI 为 CV 提供了图像处理算法。

配置偏移

对于工作负载而言，一种偏离预期状态的配置更改。这可能会导致工作负载变得不合规，且通常是渐进的，不是故意的。

配置管理数据库（CMDB）

一种存储库，用于存储和管理有关数据库及其 IT 环境的信息，包括硬件和软件组件及其配置。您通常在迁移的产品组合发现和分析阶段使用来自 CMDB 的数据。

合规性包

一系列 AWS Config 规则和补救措施，您可以汇编这些规则和补救措施，以自定义您的合规性和安全性检查。您可以使用 YAML 模板将一致性包作为单个实体部署在 AWS 账户 和区域或整个组织中。有关更多信息，请参阅 AWS Config 文档中的一致性包。

持续集成和持续交付（CI/CD）

自动执行软件发布过程的源代码、构建、测试、暂存和生产阶段的过程。 CI/CD 通常被描述为管道。 CI/CD 可以帮助您实现流程自动化、提高生产力、提高代码质量和更快地交付。有关更多信息，请参阅持续交付的优势。CD 也可以表示持续部署。有关更多信息，请参阅持续交付与持续部署。

CV

请参阅计算机视觉。

D

静态数据

网络中静止的数据，例如存储中的数据。

数据分类

根据网络中数据的关键性和敏感性对其进行识别和分类的过程。它是任何网络安全风险管理策略的关键组成部分，因为它可以帮助您确定对数据的适当保护和保留控制。数据分类是 Well-Architecte AWS d Framework 中安全支柱的一个组成部分。有关详细信息，请参阅数据分类。

数据漂移

生产数据与用来训练机器学习模型的数据之间的有意义差异，或者输入数据随时间推移的有意义变化。数据漂移可能降低机器学习模型预测的整体质量、准确性和公平性。

传输中数据

在网络中主动移动的数据，例如在网络资源之间移动的数据。

数据网格

一种架构框架，可提供分布式、去中心化的数据所有权以及集中式管理和治理。

数据最少化

仅收集并处理绝对必要数据的原则。在中进行数据最小化 AWS 云 可以降低隐私风险、成本和分析碳足迹。

数据边界

AWS 环境中的一组预防性防护措施，可帮助确保只有可信身份才能访问来自预期网络的可信资源。有关更多信息，请参阅在上构建数据边界。 AWS

数据预处理

将原始数据转换为 ML 模型易于解析的格式。预处理数据可能意味着删除某些列或行，并处理缺失、不一致或重复的值。

数据溯源

在数据的整个生命周期跟踪其来源和历史的过程，例如数据如何生成、传输和存储。

数据主体

正在收集和处理其数据的个人。

数据仓库

一种支持商业智能（例如分析）的数据管理系统。数据仓库通常包含大量历史数据，通常用于查询和分析。

数据库定义语言（DDL）

在数据库中创建或修改表和对象结构的语句或命令。

数据库操作语言（DML）

在数据库中修改（插入、更新和删除）信息的语句或命令。

DDL

请参阅数据库定义语言。

深度融合

组合多个深度学习模型进行预测。您可以使用深度融合来获得更准确的预测或估算预测中的不确定性。

深度学习

一个 ML 子字段使用多层人工神经网络来识别输入数据和感兴趣的目标变量之间的映射。

defense-in-depth

一种信息安全方法，经过深思熟虑，在整个计算机网络中分层实施一系列安全机制和控制措施，以保护网络及其中数据的机密性、完整性和可用性。当你采用这种策略时 AWS，你会在 AWS

Organizations 结构的不同层面添加多个控件来帮助保护资源。例如，一种 defense-in-depth方法可以结合多因素身份验证、网络分段和加密。

委派管理员

在中 AWS Organizations，兼容的服务可以注册 AWS 成员帐户来管理组织的帐户并管理该服务的权限。此账户被称为该服务的委托管理员。有关更多信息和兼容服务列表，请参阅 AWS Organizations 文档中使用 AWS Organizations的服务。

部署

使应用程序、新功能或代码修复在目标环境中可用的过程。部署涉及在代码库中实现更改，然后在应用程序的环境中构建和运行该代码库。

开发环境

请参阅环境。

侦测性控制

一种安全控制，在事件发生后进行检测、记录日志和发出提醒。这些控制是第二道防线，提醒您注意绕过现有预防性控制的安全事件。有关更多信息，请参阅在 AWS上实施安全控制中的侦测性控制。

开发价值流映射（DVSM）

用于识别对软件开发生命周期中的速度和质量产生不利影响的限制因素并确定其优先级的流程。DVSM 扩展了最初为精益生产实践设计的价值流映射流程。其重点关注在软件开发过程中创造和转移价值所需的步骤和团队。

数字孪生

真实世界系统的虚拟再现，如建筑物、工厂、工业设备或生产线。数字孪生支持预测性维护、远程监控和生产优化。

维度表

星型架构中的一种较小的表，其中包含事实表中定量数据的数据属性。维度表属性通常是文本字段或行为类似于文本的离散数字。这些属性通常用于查询约束、筛选和结果集标注。

灾难

阻止工作负载或系统在其主要部署位置实现其业务目标的事件。这些事件可能是自然灾害、技术故障或人为操作的结果，例如无意的配置错误或恶意软件攻击。

灾难恢复（DR）

您用来最大程度地减少由灾难造成的停机时间和数据丢失的策略和流程。有关更多信息，请参阅 Well-Architected Fr ame AWS work 中的 “工作负载灾难恢复：云端 AWS 恢复”。

DML

请参阅数据库操作语言。

领域驱动设计

一种开发复杂软件系统的方法，通过将其组件连接到每个组件所服务的不断发展的领域或核心业务目标。Eric Evans 在其著作领域驱动设计：软件核心复杂性应对之道（Boston: Addison-Wesley Professional, 2003）中介绍了这一概念。有关如何将领域驱动设计与 strangler fig 模式结合使用的信息，请参阅使用容器和 Amazon API Gateway 逐步将原有的 Microsoft ASP.NET（ASMX）Web 服务现代化。

DR

请参阅灾难恢复。

偏差检测

跟踪与基准配置的偏差。例如，您可以使用 AWS CloudFormation 来检测系统资源中的偏差，也可以使用 AWS Control Tower 来检测着陆区中可能影响监管要求合规性的变化。

DVSM

请参阅开发价值流映射。

E

EDA

请参阅探索性数据分析。

EDI

请参阅电子数据交换。

边缘计算

该技术可提高位于 IoT 网络边缘的智能设备的计算能力。与云计算比较时，边缘计算可以减少通信延迟并缩短响应时间。

电子数据交换（EDI）

组织之间业务文件的自动交换。有关更多信息，请参阅什么是电子数据交换。

加密

一种将人类可读的纯文本数据转换为加密文字的计算流程。

加密密钥

由加密算法生成的随机位的加密字符串。密钥的长度可能有所不同，而且每个密钥都设计为不可预测且唯一。

字节顺序

字节在计算机内存中的存储顺序。大端序系统先存储最高有效字节。小端序系统先存储最低有效字节。

端点

请参阅服务端点。

端点服务

一种可以在虚拟私有云（VPC）中托管，与其他用户共享的服务。您可以使用其他 AWS 账户 或 AWS Identity and Access Management (IAM) 委托人创建终端节点服务， AWS PrivateLink 并向其授予权限。这些账户或主体可通过创建接口 VPC 端点来私密地连接到您的端点服务。有关更多信息，请参阅 Amazon Virtual Private Cloud（Amazon VPC）文档中的创建端点服务。

企业资源规划（ERP）

一种自动化和管理企业关键业务流程（例如会计、MES 和项目管理）的系统。

信封加密

用另一个加密密钥对加密密钥进行加密的过程。有关更多信息，请参阅 AWS Key Management Service (AWS KMS) 文档中的信封加密。

环境

正在运行的应用程序的实例。以下是云计算中常见的环境类型：

- 开发环境 — 正在运行的应用程序的实例，只有负责维护应用程序的核心团队才能使用。开发环境用于测试更改，然后再将其提升到上层环境。这类环境有时称为测试环境。
- 下层环境 — 应用程序的所有开发环境，比如用于初始构建和测试的环境。

- 生产环境 — 最终用户可以访问的正在运行的应用程序的实例。在 CI/CD 管道中，生产环境是最后一个部署环境。
- 上层环境 — 除核心开发团队以外的用户可以访问的所有环境。这可能包括生产环境、预生产环境和用户验收测试环境。

epic

在敏捷方法学中，有助于组织工作和确定优先级的功能类别。epics 提供了对需求和实施任务的总体描述。例如， AWS CAF 安全史诗包括身份和访问管理、侦探控制、基础设施安全、数据保护和事件响应。有关 AWS 迁移策略中 epics 的更多信息，请参阅计划实施指南。

ERP

请参阅企业资源规划。

探索性数据分析（EDA）

分析数据集以了解其主要特征的过程。您收集或汇总数据，并进行初步调查，以发现模式、检测异常并检查假定情况。EDA 通过计算汇总统计数据和创建数据可视化得以执行。

F

事实表

星型架构中的中心表。它存储有关业务运营的定量数据。通常，事实表包含两种类型的列：包含度量的列和包含维度表外键的列。

快速失效机制

一种使用频繁且增量式的测试来缩短开发生命周期的理念。这是敏捷方法的关键部分。

故障隔离边界

在中 AWS 云，诸如可用区 AWS 区域、控制平面或数据平面之类的边界，它限制了故障的影响并有助于提高工作负载的弹性。有关更多信息，请参阅 AWS 故障隔离边界。

功能分支

请参阅分支。

特征

您用来进行预测的输入数据。例如，在制造环境中，特征可能是定期从生产线捕获的图像。

特征重要性

特征对于模型预测的重要性。这通常表示为数值分数，可以通过各种技术进行计算，例如 Shapley 加法解释（SHAP）和积分梯度。有关更多信息，请参阅使用机器学习模型的可解释性 AWS。

功能转换

为 ML 流程优化数据，包括使用其他来源丰富数据、扩展值或从单个数据字段中提取多组信息。这使得 ML 模型能从数据中获益。例如，如果您将“2021-05-27 00:15:37”日期分解为“2021”、“五月”、“星期四”和“15”，则可以帮助学习与不同数据成分相关的算法学习精细模式。

少样本提示

在要求 LLM 执行类似任务之前，先向其提供少量示例，以演示任务和预期输出。此技术是上下文内学习的一种应用，其中模型可以从提示中嵌入的示例（样本）中学习。对于需要特定格式、推理或领域知识的任务，少样本提示可能非常有效。另请参阅零样本提示。

FGAC

请参阅精细访问控制。

精细访问控制（FGAC）

使用多个条件允许或拒绝访问请求。

快闪迁移

一种数据库迁移方法，通过更改数据捕获使用连续数据复制，在极短的时间内迁移数据，而非使用分阶段方法。目标是将停机时间降至最低。

FM

请参阅基础模型。

基础模型（FM）

一个大型深度学习神经网络，一直在广义和未标记数据的大量数据集上进行训练。 FMs 能够执行各种各样的一般任务，例如理解语言、生成文本和图像以及用自然语言进行对话。有关更多信息，请参阅什么是基础模型。

G

生成式人工智能

AI 模型的一个子集，这些模型已经过大量数据训练，可以使用简单的文本提示来创建新的内容和构件，例如图像、视频、文本和音频。有关更多信息，请参阅什么是生成式人工智能。

地理阻止

请参阅地理限制。

地理限制（地理阻止）

在 Amazon 中 CloudFront，一种阻止特定国家/地区的用户访问内容分发的选项。您可以使用允许列表或阻止列表来指定已批准和已禁止的国家/地区。有关更多信息，请参阅 CloudFront 文档中的限制内容的地理分布。

GitFlow 工作流程

一种方法，在这种方法中，下层和上层环境在源代码存储库中使用不同的分支。Gitflow 工作流程被认为是传统的工作流程，而基于中继的工作流程则是现代的、首选的方法。

黄金映像

系统或软件的快照，用作部署该系统或软件的新实例的模板。例如，在制造业中，黄金映像可用于在多个设备上预调配软件，并有助于提高设备制造操作的速度、可扩展性和生产效率。

全新策略

在新环境中缺少现有基础设施。在对系统架构采用全新策略时，您可以选择所有新技术，而不受对现有基础设施（也称为棕地）兼容性的限制。如果您正在扩展现有基础设施，则可以将棕地策略和全新策略混合。

防护机制

帮助管理各组织单位的资源、策略和合规性的高级规则 (OUs)。预防性防护机制会执行策略以确保符合合规性标准。它们是使用服务控制策略和 IAM 权限边界实现的。侦测性护栏会检测策略违规和合规性问题，并生成提醒以进行修复。它们通过使用 AWS Config、、Amazon、 AWS Security Hub CSPM GuardDuty AWS Trusted Advisor、Amazon Inspector 和自定义 AWS Lambda 支票来实现。

H

HA

请参阅高可用性。

异构数据库迁移

将源数据库迁移到使用不同数据库引擎的目标数据库（例如，从 Oracle 迁移到 Amazon Aurora）。异构迁移通常是重新架构工作的一部分，而转换架构可能是一项复杂的任务。AWS 提供了 AWS SCT 来帮助实现架构转换。

高可用性（HA）

在遇到挑战或灾难时，工作负载无需干预即可连续运行的能力。HA 系统旨在自动进行故障转移、持续提供良好性能，并以最小的性能影响处理不同负载和故障。

历史数据库现代化

一种用于实现运营技术（OT）系统现代化和升级以更好满足制造业需求的方法。历史数据库是一种用于收集和存储工厂中各种来源数据的数据库。

保留数据

从用于训练机器学习模型的数据集中保留的一部分标注的历史数据。通过将模型预测与保留数据进行比较，您可以使用保留数据来评估模型性能。

同构数据库迁移

将源数据库迁移到共享同一数据库引擎的目标数据库（例如，从 Microsoft SQL Server 迁移到 Amazon RDS for SQL Server）。同构迁移通常是更换主机或更换平台工作的一部分。您可以使用本机数据库实用程序来迁移架构。

热数据

经常访问的数据，例如实时数据或近期的转化数据。这些数据通常需要高性能存储层或存储类别才能提供快速的查询响应。

修补程序

针对生产环境中关键问题的紧急修复。由于其紧迫性，修补程序通常是在典型的 DevOps 发布工作流程之外进行的。

hypercare 周期

割接之后，迁移团队立即管理和监控云中迁移的应用程序以解决任何问题的时间段。通常，这个周期持续 1-4 天。在 hypercare 周期结束时，迁移团队通常会将应用程序的责任移交给云运营团队。

我

IaC

请参阅基础设施即代码。

基于身份的策略

附加到一个或多个 IAM 委托人的策略，用于定义他们在 AWS 云 环境中的权限。

空闲应用程序

90 天内平均 CPU 和内存使用率在 5% 到 20% 之间的应用程序。在迁移项目中，通常会停用这些应用程序或将其保留在本地。

IIoT

请参阅工业物联网。

不可变基础设施

一种模型，可为生产工作负载部署新的基础设施，而不是更新、修补或修改现有基础设施。不可变基础设施本质上比可变基础设施更一致、更可靠、更可预测。有关更多信息，请参阅 AWS Well-Architected Framework 中的使用不可变基础设施进行部署最佳实践。

入站（入口）VPC

在 AWS 多账户架构中，一种接受、检查和路由来自应用程序外部的网络连接的 VPC。AWS 安全参考架构建议设置您的网络帐户，包括入站、出站和检查， VPCs 以保护您的应用程序与更广泛的互联网之间的双向接口。

增量迁移

一种割接策略，在这种策略中，您可以将应用程序分成小部分进行迁移，而不是一次性完整割接。例如，您最初可能只将几个微服务或用户迁移到新系统。在确认一切正常后，您可以逐步迁移其他微服务或用户，直到停用遗留系统。这种策略降低了大规模迁移带来的风险。

工业 4.0

该术语由 Klaus Schwab 在 2016 年提出，指的是通过连接、实时数据、自动化、分析和 AI/ML 的进步来实现制造流程的现代化。

基础设施

应用程序环境中包含的所有资源和资产。

基础设施即代码（IaC）

通过一组配置文件预调配和管理应用程序基础设施的过程。IaC 旨在帮助您集中管理基础设施、实现资源标准化和快速扩展，使新环境具有可重复性、可靠性和一致性。

工业物联网 (IIoT)

在工业领域使用联网的传感器和设备，例如制造业、能源、汽车、医疗保健、生命科学和农业。有关更多信息，请参阅制定工业物联网 (IIoT) 数字化转型战略。

检查 VPC

在 AWS 多账户架构中，一种集中式 VPC，用于管理对 VPCs（相同或不同 AWS 区域）、互联网和本地网络之间的网络流量的检查。AWS 安全参考架构建议设置您的网络帐户，包括入站、出站和检查， VPCs 以保护您的应用程序与更广泛的互联网之间的双向接口。

物联网（IoT）

由带有嵌入式传感器或处理器的连接物理对象组成的网络，这些传感器或处理器通过互联网或本地通信网络与其他设备和系统进行通信。有关更多信息，请参阅什么是 IoT？

可解释性

它是机器学习模型的一种特征，描述了人类可以理解模型的预测如何取决于其输入的程度。有关更多信息，请参阅使用机器学习模型的可解释性 AWS。

物联网

请参阅物联网。

IT 信息库（ITIL）

提供 IT 服务并使这些服务符合业务要求的一套最佳实践。ITIL 是 ITSM 的基础。

IT 服务管理（ITSM）

为组织设计、实施、管理和支持 IT 服务的相关活动。有关将云运营与 ITSM 工具集成的信息，请参阅运营集成指南。

ITIL

请参阅 IT 信息库。

ITSM

请参阅 IT 服务管理。

L

基于标签的访问控制（LBAC）

强制访问控制（MAC）的一种实施方式，其中明确为用户和数据本身分配了安全标签值。用户安全标签和数据安全标签之间的交集决定了用户可以看到哪些行和列。

登录区

landing zone 是一个架构精良的多账户 AWS 环境，具有可扩展性和安全性。这是一个起点，您的组织可以从这里放心地在安全和基础设施环境中快速启动和部署工作负载和应用程序。有关登录区的更多信息，请参阅设置安全且可扩展的多账户 AWS 环境。

大语言模型（LLM）

一种基于大量数据进行预训练的深度学习 AI 模型。LLM 可以执行多项任务，例如回答问题、总结文档、将文本翻译成其他语言以及完成句子。有关更多信息，请参阅什么是 LLMs。

大规模迁移

迁移 300 台或更多服务器。

LBAC

请参阅基于标签的访问控制。

最低权限

授予执行任务所需的最低权限的最佳安全实践。有关更多信息，请参阅 IAM 文档中的应用最低权限许可。

直接迁移

请参阅 7 R。

小端序系统

一个先存储最低有效字节的系统。另请参阅字节顺序。

LLM

请参阅大型语言模型。

下层环境

请参阅环境。

M

机器学习（ML）

一种使用算法和技术进行模式识别和学习的人工智能。ML 对记录的数据（例如物联网（IoT）数据）进行分析和学习，以生成基于模式的统计模型。有关更多信息，请参阅机器学习。

主分支

请参阅分支。

恶意软件

旨在危害计算机安全或隐私的软件。恶意软件可能会破坏计算机系统、泄露敏感信息或获得未经授权的访问权限。恶意软件的示例包括病毒、蠕虫、勒索软件、木马、间谍软件和键盘记录器。

托管式服务

AWS 服务 它 AWS 运行基础设施层、操作系统和平台，您可以访问端点来存储和检索数据。Amazon Simple Storage Service（Amazon S3）和 Amazon DynamoDB 就是托管服务的示例。这些服务也称为抽象服务。

制造执行系统（MES）

一种软件系统，用于跟踪、监控、记录和控制在车间将原材料转化为成品的生产过程。

MAP

请参阅迁移加速计划。

机制

一个完整的过程，您可以在其中创建工具，推动工具的采用，然后检查结果以进行调整。机制是一种在运作过程中自我强化和改善的循环。有关更多信息，请参阅在 Well-Architect AWS ed 框架中构建机制。

成员账户

AWS 账户 除属于组织中的管理账户之外的所有账户 AWS Organizations。一个账户一次只能是一个组织的成员。

MES

请参阅制造执行系统。

消息队列遥测传输（MQTT）

一种基于发布/订阅模式的轻量级 machine-to-machine (M2M) 通信协议，适用于资源受限的物联网设备。

微服务

一种小型的独立服务，通过明确的定义进行通信 APIs ，通常由小型的独立团队拥有。例如，保险系统可能包括映射到业务能力（如销售或营销）或子域（如购买、理赔或分析）的微服务。微服务

的好处包括敏捷、灵活扩展、易于部署、可重复使用的代码和恢复能力。有关更多信息，请参阅使用 AWS 无服务器服务集成微服务。

微服务架构

一种使用独立组件构建应用程序的方法，这些组件将每个应用程序进程作为微服务运行。这些微服务使用轻量级通过定义明确的接口进行通信。 APIs该架构中的每个微服务都可以更新、部署和扩展，以满足对应用程序特定功能的需求。有关更多信息，请参阅在上实现微服务。 AWS

迁移加速计划（MAP）

AWS 该计划提供咨询支持、培训和服务，以帮助组织为迁移到云奠定坚实的运营基础，并帮助抵消迁移的初始成本。MAP 提供了一种以系统的方式执行遗留迁移的迁移方法，以及一套用于自动执行和加速常见迁移场景的工具。

大规模迁移

将大部分应用程序组合分波迁移到云中的过程，在每一波中以更快的速度迁移更多应用程序。本阶段使用从早期阶段获得的最佳实践和经验教训，实施由团队、工具和流程组成的迁移工厂，通过自动化和敏捷交付简化工作负载的迁移。这是 AWS 迁移策略的第三阶段。

迁移工厂

跨职能团队，通过自动化、敏捷的方法简化工作负载迁移。迁移工厂团队通常包括运营、业务分析师和所有者、迁移工程师、开发人员和冲刺 DevOps 领域的专业人员。20% 到 50% 的企业应用程序组合由可通过工厂方法优化的重复模式组成。有关更多信息，请参阅本内容集中有关迁移工厂的讨论和云迁移工厂指南。

迁移元数据

有关完成迁移所需的应用程序和服务器的信息。每种迁移模式都需要一套不同的迁移元数据。迁移元数据的示例包括目标子网、安全组和 AWS 账户。

迁移模式

一种可重复的迁移任务，详细列出了迁移策略、迁移目标以及所使用的迁移应用程序或服务。示例：使用 AWS 应用程序迁移服务重新托管向 Amazon EC2 的迁移。

迁移组合评测（MPA）

一种在线工具，提供了用于验证迁移到 AWS 云的业务案例的信息。MPA 提供了详细的组合评测（服务器规模调整、定价、TCO 比较、迁移成本分析）以及迁移计划（应用程序数据分析和数据收集、应用程序分组、迁移优先级排序和波次规划）。所有 AWS 顾问和 APN 合作伙伴顾问均可免费使用 MPA 工具（需要登录）。

迁移准备情况评测（MRA）

使用 AWS CAF 深入了解组织的云就绪状态、确定优势和劣势以及制定行动计划以缩小已发现差距的过程。有关更多信息，请参阅迁移准备指南。MRA 是 AWS 迁移策略的第一阶段。

迁移策略

将工作负载迁移到 AWS 云的方法。有关更多信息，请参见术语表中的 7 R 词条，以及动员您的组织以加快大规模迁移。

ML

请参阅机器学习。

现代化

将过时的（原有的或单体）应用程序及其基础设施转变为云中敏捷、弹性和高度可用的系统，以降低成本、提高效率和利用创新。有关更多信息，请参阅在 AWS 云中实现应用程序现代化的策略。

现代化准备情况评估

一种评估方式，有助于确定组织应用程序的现代化准备情况；确定收益、风险和依赖关系；确定组织能够在多大程度上支持这些应用程序的未来状态。评估结果是目标架构的蓝图、详细说明现代化进程发展阶段和里程碑的路线图以及解决已发现差距的行动计划。有关更多信息，请参阅在 AWS 云中评估应用程序的现代化准备情况。

单体应用程序（单体式）

作为具有紧密耦合进程的单个服务运行的应用程序。单体应用程序有几个缺点。如果某个应用程序功能的需求激增，则必须扩展整个架构。随着代码库的增长，添加或改进单体应用程序的功能也会变得更加复杂。若要解决这些问题，可以使用微服务架构。有关更多信息，请参阅将单体分解为微服务。

MPA

请参阅迁移组合评测。

MQTT

请参阅消息队列遥测传输。

多分类器

一种帮助为多个类别生成预测（预测两个以上结果之一）的过程。例如，ML 模型可能会询问“这个产品是书、汽车还是手机？”或“此客户最感兴趣什么类别的产品？”

可变基础设施

一种用于更新和修改生产工作负载的现有基础设施的模型。为了提高一致性、可靠性和可预测性，Well-Architect AWS ed Framework 建议使用不可变基础设施作为最佳实践。

O

OAC

请参阅来源访问控制。

OAI

请参阅来源访问身份。

OCM

请参阅组织变革管理。

离线迁移

一种迁移方法，在这种方法中，源工作负载会在迁移过程中停止运行。这种方法会延长停机时间，通常用于小型非关键工作负载。

OI

请参阅运营集成。

OLA

请参阅运营级别协议。

在线迁移

一种迁移方法，在这种方法中，源工作负载无需离线即可复制到目标系统。在迁移过程中，连接工作负载的应用程序可以继续运行。这种方法的停机时间为零或最短，通常用于关键生产工作负载。

OPC-UA

请参阅开放流程通信 – 统一架构。

开放流程通信 – 统一架构（OPC-UA）

一种用于工业自动化的 machine-to-machine（M2M）通信协议。OPC-UA 提供了一个包含数据加密、身份验证和授权方案的互操作性标准。

运营级别协议（OLA）

一项协议，阐明了 IT 职能部门承诺相互交付的内容，以支持服务水平协议（SLA）。

运营准备情况审查（ORR）

一份问题核对清单和关联的最佳实践，可帮助您了解、评估、预防或缩小事件和可能的故障的范围。有关更多信息，请参阅 AWS Well-Architected Framework 中的运营准备情况审查（ORR）。

运营技术（OT）

与物理环境配合使用以控制工业运营、设备和基础设施的硬件和软件系统。在制造业中，OT 和信息技术（IT）系统的集成是工业 4.0 转型的关键重点。

运营整合（OI）

在云中实现运营现代化的过程，包括就绪计划、自动化和集成。有关更多信息，请参阅运营整合指南。

组织跟踪

由 AWS CloudTrail 此创建的跟踪记录组织 AWS 账户 中所有人的所有事件 AWS Organizations。该跟踪是在每个 AWS 账户 中创建的，属于组织的一部分，并跟踪每个账户的活动。有关更多信息，请参阅 CloudTrail文档中的为组织创建跟踪。

组织变革管理（OCM）

一个从人员、文化和领导力角度管理重大、颠覆性业务转型的框架。OCM 通过加快变革采用、解决过渡问题以及推动文化和组织变革，帮助组织为新系统和战略做好准备和过渡。在 AWS 迁移策略中，该框架被称为人员加速，因为云采用项目需要变更的速度。有关更多信息，请参阅 OCM 指南。

来源访问控制（OAC）

在中 CloudFront，一个增强的选项，用于限制访问以保护您的亚马逊简单存储服务 (Amazon S3) 内容。OAC 全部支持所有 S3 存储桶 AWS 区域、使用 AWS KMS (SSE-KMS) 进行服务器端加密，以及对 S3 存储桶的动态PUT和DELETE请求。

来源访问身份（OAI）

在中 CloudFront，一个用于限制访问权限以保护您的 Amazon S3 内容的选项。当您使用 OAI 时，CloudFront 会创建一个 Amazon S3 可以对其进行身份验证的委托人。经过身份验证的委托人只能通过特定 CloudFront 分配访问 S3 存储桶中的内容。另请参阅 OAC，其中提供了更精细和增强的访问控制。

ORR

请参阅运营准备情况审查。

OT

请参阅运营技术。

出站（出口）VPC

在 AWS 多账户架构中，一种处理从应用程序内部启动的网络连接的 VPC。AWS 安全参考架构建议设置您的网络帐户，包括入站、出站和检查， VPCs 以保护您的应用程序与更广泛的互联网之间的双向接口。

P

权限边界

附加到 IAM 主体的 IAM 管理策略，用于设置用户或角色可以拥有的最大权限。有关更多信息，请参阅 IAM 文档中的权限边界。

个人身份信息（PII）

直接查看其他相关数据或与之配对时可用于合理推断个人身份的信息。PII 的示例包括姓名、地址和联系信息。

PII

请参阅个人身份信息。

playbook

一套预定义的步骤，用于捕获与迁移相关的工作，例如在云中交付核心运营功能。playbook 可以采用脚本、自动化运行手册的形式，也可以是操作现代化环境所需的流程或步骤的摘要。

PLC

请参阅可编程逻辑控制器。

PLM

请参阅产品生命周期管理。

policy

一个对象，可以定义权限（请参阅基于身份的策略）、指定访问条件（请参阅基于资源的策略）或定义 AWS Organizations 的组织中所有账户的最大权限（请参阅服务控制策略）。

多语言持久性

根据数据访问模式和其他要求，独立选择微服务的数据存储技术。如果您的微服务采用相同的数据存储技术，它们可能会遇到实现难题或性能不佳。如果微服务使用最适合其需求的数据存储，则可以更轻松地实现微服务，并获得更好的性能和可扩展性。

组合评测

一个发现、分析和确定应用程序组合优先级以规划迁移的过程。有关更多信息，请参阅评估迁移准备情况。

谓词

返回 `true` 或 `false` 的查询条件，通常位于 `WHERE` 子句中。

谓词下推

一种数据库查询优化技术，可在传输之前筛选查询中的数据。这将减少从关系数据库检索和处理的数据量，并提高查询性能。

预防性控制

一种安全控制，旨在防止事件发生。这些控制是第一道防线，帮助防止未经授权的访问或对网络的意外更改。有关更多信息，请参阅在 AWS上实施安全控制中的预防性控制。

主体

中 AWS 可以执行操作和访问资源的实体。此实体通常是 IAM 角色的根用户或用户。 AWS 账户有关更多信息，请参阅 IAM 文档中角色术语和概念中的主体。

隐私设计

一种在整个开发过程中都考虑隐私的系统工程方法。

私有托管区

一个容器，其中包含有关您希望 Amazon Route 53 如何响应针对一个或多个 VPCs域名及其子域名的 DNS 查询的信息。有关更多信息，请参阅 Route 53 文档中的私有托管区的使用。

主动控制

一种安全控制，旨在防止部署不合规资源。这些控制会在资源预置之前对其进行扫描。如果资源与控制不兼容，则不会预置它。有关更多信息，请参阅 AWS Control Tower 文档中的控制参考指南，并参见在上实施安全控制中的主动控制 AWS。

产品生命周期管理（PLM）

对产品在其整个生命周期内的数据和流程的管理，从设计、开发和发布，到增长和成熟，再到衰退和淘汰。

生产环境

请参阅环境。

可编程逻辑控制器（PLC）

在制造业中，一种高度可靠、适应性强的计算机，用于监控机器并实现制造过程自动化。

提示串接

使用一个 LLM 提示的输出作为下一个提示的输入，以生成更好的响应。该技术用于将复杂的任务分解为子任务，或者迭代地完善或扩展初步响应。它有助于提高模型响应的准确性和相关性，并允许获得更精细的个性化结果。

假名化

用占位符值替换数据集中个人标识符的过程。假名化可以帮助保护个人隐私。假名化数据仍被视为个人数据。

publish/subscribe (pub/sub)

一种支持微服务间异步通信的模式，可提高可扩展性和响应能力。例如，在基于微服务的 MES 中，微服务可以将事件消息发布到其他微服务可以订阅的频道。系统可以在不更改发布服务的情况下添加新的微服务。

Q

查询计划

一系列用于访问 SQL 关系数据库系统中的数据的步骤，类似于指令。

查询计划回归

当数据库服务优化程序选择的最佳计划不如数据库环境发生特定变化之前时。这可能是由统计数据、约束、环境设置、查询参数绑定更改和数据库引擎更新造成的。

R

RACI 矩阵

请参阅责任、问责、咨询和知情（RACI）。

RAG

请参阅检索增强生成。

勒索软件

一种恶意软件，旨在阻止对计算机系统或数据的访问，直到付款为止。

RASCI 矩阵

请参阅责任、问责、咨询和知情（RACI）。

RCAC

请参阅行列访问控制。

只读副本

用于只读目的的数据库副本。您可以将查询路由到只读副本，以减轻主数据库的负载。

重新架构

请参阅 7 R。

恢复点目标（RPO）

自上一个数据恢复点以来可接受的最长时间。这决定了从上一个恢复点到服务中断之间可接受的数据丢失情况。

恢复时间目标（RTO）

服务中断和服务恢复之间可接受的最大延迟。

重构

请参阅 7 R。

Region

地理区域内的 AWS 资源集合。每一个 AWS 区域 都相互隔离，相互独立，以提供容错、稳定性和弹性。有关更多信息，请参阅指定您的账户可以使用的 AWS 区域。

回归

一种预测数值的 ML 技术。例如，要解决“这套房子的售价是多少？”的问题 ML 模型可以使用线性回归模型，根据房屋的已知事实（如建筑面积）来预测房屋的销售价格。

重新托管

请参阅 7 R。

版本

在部署过程中，推动生产环境变更的行为。

重新放置

请参阅 7 R。

更换平台

请参阅 7 R。

重新购买

请参阅 7 R。

韧性

应用程序抵御中断或从中断中恢复的能力。在 AWS 云中规划韧性时，高可用性和灾难恢复是常见的考虑因素。有关更多信息，请参阅 AWS 云 韧性。

基于资源的策略

一种附加到资源的策略，例如 AmazonS3 存储桶、端点或加密密钥。此类策略指定了允许哪些主体访问、支持的操作以及必须满足的任何其他条件。

责任、问责、咨询和知情（RACI）矩阵

定义参与迁移活动和云运营的所有各方的角色和责任的矩阵。矩阵名称源自矩阵中定义的责任类型：负责（R）、问责（A）、咨询（C）和知情（I）。支持（S）类型是可选的。如果包括支持，则该矩阵称为 RASCI 矩阵，如果将其排除在外，则称为 RACI 矩阵。

响应性控制

一种安全控制，旨在推动对不良事件或偏离安全基线的情况进行修复。有关更多信息，请参阅在AWS上实施安全控制中的响应性控制。

保留

请参阅 7 R。

停用

请参阅 7 R。

检索增强生成（RAG）

一种生成式人工智能技术，其中 LLM 在生成响应之前引用其训练数据来源之外的权威数据来源。例如，RAG 模型可以对组织的知识库或自定义数据执行语义搜索。有关更多信息，请参阅什么是 RAG。

轮换

定期更新密钥以使攻击者更难访问凭证的过程。

行列访问控制（RCAC）

使用已定义访问规则的基本、灵活的 SQL 表达式。RCAC 由行权限和列掩码组成。

RPO

请参阅恢复点目标。

RTO

请参阅恢复时间目标。

运行手册

执行特定任务所需的一套手动或自动程序。它们通常是为了简化重复性操作或高错误率的程序而设计的。

S

SAML 2.0

许多身份提供商 (IdPs) 使用的开放标准。此功能支持联合单点登录 (SSO)，因此用户无需在 IAM 中为组织中的所有人创建用户即可登录 AWS 管理控制台 或调用 AWS API 操作。有关基于 SAML 2.0 的联合身份验证的更多信息，请参阅 IAM 文档中的关于基于 SAML 2.0 的联合身份验证。

SCADA

请参阅监督控制和数据采集。

SCP

请参阅服务控制策略。

机密密钥

在中 AWS Secrets Manager，您以加密形式存储的机密或受限信息，例如密码或用户凭证。它由密钥值及其元数据组成。密钥值可以是二进制、单个字符串或多个字符串。有关更多信息，请参阅 Secrets Manager 文档中的什么是 Amazon Secrets Manager 密钥？。

安全设计

一种在整个开发过程中都考虑安全的系统工程方法。

安全控制

一种技术或管理防护机制，可防止、检测或降低威胁行为体利用安全漏洞的能力。安全控制有以下四种类型：预防性、检测性、响应性和主动性。

安全固化

缩小攻击面，使其更能抵御攻击的过程。这可能包括删除不再需要的资源、实施授予最低权限的最佳安全实践或停用配置文件中不必要的功能等操作。

安全信息和事件管理（SIEM）系统

结合了安全信息管理（SIM）和安全事件管理（SEM）系统的工具和服务。SIEM 系统会收集、监控和分析来自服务器、网络、设备和其他来源的数据，以检测威胁和安全漏洞，并生成警报。

安全响应自动化

一种预定义的程序化操作，旨在自动响应或修复安全事件。这些自动化可作为侦探或响应式安全控制措施，帮助您实施 AWS 安全最佳实践。自动响应操作的示例包括修改 VPC 安全组、修补 Amazon EC2 实例或轮换凭证。

服务器端加密

由接收数据的人在目的地对数据 AWS 服务 进行加密。

服务控制策略（SCP）

一种策略，用于集中控制组织中所有账户的权限 AWS Organizations。 SCPs 定义防护措施或限制管理员可以委托给用户或角色的操作。您可以使用 SCPs 允许列表或拒绝列表来指定允许或禁止哪些服务或操作。有关更多信息，请参阅 AWS Organizations 文档中的服务控制策略。

服务端点

的入口点的 URL AWS 服务。您可以使用端点，通过编程方式连接到目标服务。有关更多信息，请参阅 AWS 一般参考 中的 AWS 服务 端点。

服务水平协议（SLA）

一份协议，阐明了 IT 团队承诺向客户交付的内容，比如服务正常运行时间和性能。

服务水平指示器（SLI）

对服务性能方面的衡量，例如错误率、可用性或吞吐量。

服务水平目标（SLO）

代表服务运行状况的目标指标，由服务水平指示器衡量。

责任共担模式

描述您在云安全与合规方面共同承担 AWS 的责任的模型。 AWS 负责云的安全，而您则负责云中的安全。有关更多信息，请参阅责任共担模式。

SIEM

请参阅安全信息和事件管理系统。

单点故障（SPOF）

应用程序的单个关键组件出现故障，可能会中断系统。

SLA

请参阅服务水平协议。

SLI

请参阅服务水平指示器。

SLO

请参阅服务水平目标。

split-and-seed 模型

一种扩展和加速现代化项目的模式。随着新功能和产品发布的定义，核心团队会拆分以创建新的产品团队。这有助于扩展组织的能力和服务，提高开发人员的工作效率，支持快速创新。有关更多信息，请参阅在 AWS 云中实现应用程序现代化的分阶段方法。

SPOF

请参阅单点故障。

星型架构

一种数据库组织结构，它使用一个大型事实表来存储事务数据或测量数据，并使用一个或多个较小的维度表来存储数据属性。此结构专为在数据仓库中使用或用于商业智能目的而设计。

strangler fig 模式

一种通过逐步重写和替换系统功能直至可以停用原有的系统来实现单体系统现代化的方法。这种模式用无花果藤作为类比，这种藤蔓成长为一棵树，最终战胜并取代了宿主。该模式是由 Martin Fowler 提出的，作为重写单体系统时管理风险的一种方法。有关如何应用此模式的示例，请参阅使用容器和 Amazon API Gateway 逐步将原有的 Microsoft ASP.NET（ASMX）Web 服务现代化。

子网

您的 VPC 内的一个 IP 地址范围。子网必须位于单个可用区中。

监督控制和数据采集（SCADA）

在制造业中，一种使用硬件和软件来监控实物资产和生产操作的系统。

对称加密

一种加密算法，它使用相同的密钥来加密和解密数据。

综合测试

以模拟用户交互的方式测试系统，以检测潜在问题或监控性能。你可以使用 Amazon S CloudWatch ynthetics 来创建这些测试。

系统提示

一种为 LLM 提供上下文、说明或准则以指导其行为的技术。系统提示有助于设置上下文并制定与用户交互的规则。

T

标签

键值对，用作组织资源的元数据。 AWS 标签有助于您管理、识别、组织、搜索和筛选 资源。有关更多信息，请参阅标记您的 AWS 资源。

目标变量

您在监督式 ML 中尝试预测的值。这也被称为结果变量。例如，在制造环境中，目标变量可能是产品缺陷。

任务列表

一种通过运行手册用于跟踪进度的工具。任务列表包含运行手册的概述和要完成的常规任务列表。对于每项常规任务，它包括预计所需时间、所有者和进度。

测试环境

请参阅环境。

训练

为您的 ML 模型提供学习数据。训练数据必须包含正确答案。学习算法在训练数据中查找将输入数据属性映射到目标（您希望预测的答案）的模式。然后输出捕获这些模式的 ML 模型。然后，您可以使用 ML 模型对不知道目标的新数据进行预测。

中转网关

一个网络传输中心，可用于将您的网络 VPCs 和本地网络互连。有关更多信息，请参阅 AWS Transit Gateway 文档中的什么是公交网关。

基于中继的工作流程

一种方法，开发人员在功能分支中本地构建和测试功能，然后将这些更改合并到主分支中。然后，按顺序将主分支构建到开发、预生产和生产环境。

可信访问权限

向您指定的服务授予权限，该服务可代表您在其账户中执行任务。 AWS Organizations 当需要服务相关的角色时，受信任的服务会在每个账户中创建一个角色，为您执行管理任务。有关更多信息，请参阅 AWS Organizations 文档中的AWS Organizations 与其他 AWS 服务一起使用。

优化

更改训练过程的各个方面，以提高 ML 模型的准确性。例如，您可以通过生成标签集、添加标签，并在不同的设置下多次重复这些步骤来优化模型，从而训练 ML 模型。

双披萨团队

一个小 DevOps 团队，你可以用两个披萨来喂食。双披萨团队的规模可确保在软件开发过程中充分协作。

U

不确定性

这一概念指的是不精确、不完整或未知的信息，这些信息可能会破坏预测式 ML 模型的可靠性。不确定性有两种类型：认知不确定性是由有限的、不完整的数据造成的，而偶然不确定性是由数据中固有的噪声和随机性导致的。有关更多信息，请参阅量化深度学习系统中的不确定性指南。

无差别任务

也称为繁重工作，即创建和运行应用程序所必需的工作，但不能为最终用户提供直接价值或竞争优势。无差别任务的示例包括采购、维护和容量规划。

上层环境

请参阅环境。

V

vacuum 操作

一种数据库维护操作，包括在增量更新后进行清理，以回收存储空间并提高性能。

版本控制

跟踪更改的过程和工具，例如存储库中源代码的更改。

VPC 对等连接

两者之间的连接 VPCs ，允许您使用私有 IP 地址路由流量。有关更多信息，请参阅 Amazon VPC 文档中的什么是 VPC 对等连接。

漏洞

损害系统安全的软件缺陷或硬件缺陷。

W

热缓存

一种包含经常访问的当前相关数据的缓冲区缓存。数据库实例可以从缓冲区缓存读取，这比从主内存或磁盘读取要快。

暖数据

不常访问的数据。查询此类数据时，通常可以接受中速查询。

窗口函数

一种对与当前记录有某种关联的一组行执行计算的 SQL 函数。窗口函数对于处理任务很有用，例如计算移动平均值或根据当前行的相对位置访问行的值。

工作负载

一系列资源和代码，它们可以提供商业价值，如面向客户的应用程序或后端过程。

工作流

迁移项目中负责一组特定任务的职能小组。每个工作流都是独立的，但支持项目中的其他工作流。例如，组合工作流负责确定应用程序的优先级、波次规划和收集迁移元数据。组合工作流将这些资产交付给迁移工作流，然后迁移服务器和应用程序。

WORM

请参阅一次写入多次读取。

WQF

请参阅 AWS 工作负载资格鉴定框架。

一次写入多次读取（WORM）

一种存储模型，可一次写入数据并防止数据被删除或修改。授权用户可以根据需要多次读取数据，但无法对其进行更改。此数据存储基础设施被认为不可变。

Z

零日漏洞利用

一种利用零日漏洞的攻击，通常为恶意软件。

零日漏洞

生产系统中不可避免的缺陷或漏洞。威胁主体可能利用这种类型的漏洞攻击系统。开发人员经常因攻击而意识到该漏洞。

零样本提示

为 LLM 提供执行任务的说明，但没有可以帮助指导的示例（样本）。LLM 必须使用预先训练的知识来处理任务。零样本提示的有效性取决于任务的复杂性和提示的质量。另请参阅少样本提示。

僵尸应用程序

平均 CPU 和内存使用率低于 5% 的应用程序。在迁移项目中，通常会停用这些应用程序。

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。