为医疗保健行业创建检索增强一代解决方案 AWS

AWS 规范性指导

AWS 规范性指导: 为医疗保健行业创建检索增强一代解决方案 AWS

Table of Contents

为医疗保健行业创建检索增强一代解决方案 AWS

亚马逊 Web Services，Accenture，以及 Cadiem （贡献者）

2025 年 3 月（文档历史记录）

在大型语言模型 (LLMs) 和生成式人工智能出现之前，在医疗保健行业开发自动化和高精度应用程序是一项艰巨的任务。传统方法严重依赖手动数据输入和分析。分析医学影像和患者记录的复杂性需要大量的人为干预，这通常会导致工作流程分散且效率低下。人工智能技术的进步可帮助您大规模构建超个性化的应用程序。医疗保健应用程序现在可以与医学知识库集成，更准确地解释诊断图像，并使用预测模型预测患者预后。

本指南探讨了如何通过可以 LLMs 用来构建的检索增强生成应用程序来彻底改变医疗保健。 AWS 服务检索增强生成 (RAG) 是一种生成式人工智能技术，其中法学硕士在生成响应之前引用其训练数据源之外的权威数据源。RAG 应用程序将模型的输出建立在现实世界知识的基础上，从而减少幻觉并增加反应相关性。在医疗保健领域，RAG可用于提供准确的 up-to-date医疗信息，确保医疗保健提供者能够获得最新的研究和临床指南。通过将数据转化为切实可行的见解并自动化复杂的流程，这些技术有助于增强患者护理、简化运营并提高医疗保健专业人员的工作效率。

在 Amazon Bedrock 中，您可以对其进行微调 LLMs 并将其与智能代理集成，以创建高级医疗解决方案。该指南重点介绍了 Amazon Serv OpenSearch ic e 和 Amazon Neptune 之间的协同作用，演示了这些服务如何通过增强的搜索相关性和高级多源数据检索来提升 RAG 解决方案。你可以编排全面的 Amazon Bedrock 解决方案，这些解决方案使用亚马逊 Bedrock 代理和 LangChain无缝协调不同数据存储库之间的交互。这种集成展示了组合专业服务以创建更有效、更高效的人工智能驱动系统的力量。

患者护理和工作效率

本指南介绍了患者护理和生产力的两个现实用例：患者数据增强和再入院风险预测。它为大规模实施这些解决方案提供了战略蓝图，为医疗保健组织提供了实现人工智能驱动流程工业化的明确途径。通过这些见解，医疗机构可以使用先进的人工智能技术来创建更高效、更智能的工作流程。

人才管理

本指南还概述了重新培养技能和增强医护人员能力的策略，使他们能够将生成人工智能无缝整合到他们的日常生活中。这可以提高工作效率和患者护理质量。通过为员工提供有效使用高级人工智能工具的技能，医疗保健组织可以最大限度地提高投资回报率并推动患者护理创新。

这个由人工智能驱动的人才管理解决方案包括以下主要功能：

- 智能人才简历解析器 — 通过使用 Amazon Bedrock 中 LLMs 提供的高级功能，该工具可以有效地提取和分析简历中的关键人才技能和属性。该工具可以简化招聘流程。
- 人才知识库 — 这个动态数据库由 Amazon Neptune 提供支持，可提供有关人员配备水平、技能分布和行业趋势的实时见解。这可以帮助您就劳动力管理做出数据驱动的决策。
- 学习推荐引擎 — 这个由人工智能驱动的工具可以识别组织内部的技能差距，并为医务人员推荐个性化的培训计划。该工具可促进持续的专业发展，并帮助您的员工适应不断变化的医疗保健技术。

这些人工智能驱动的功能共同帮助优化员工绩效，通过提高智能和效率来彻底改变人才管理。

机遇与挑战

Amazon Bedrock 可以提供更高的生产力、可扩展性、成本效益和数据驱动的见解。Amazon Bedrock 使医疗保健组织能够在各种用例中 LLMs 高效使用，从内容创建和数据分析到自动决策。本指南提供了克服常见生成式人工智能挑战的方法，例如数据质量问题、基础设施可扩展性、模型性能维护以及从概念验证过渡到生产期间的持续改进要求。

生成式人工智能在医疗保健领域应用的机会

在生成式人工智能应用带来的机遇的推动下，医疗保健行业正准备进行变革性转变。生成式人工智能有可能增强患者护理、简化运营和加速医学研究。通过使用先进的人工智能模型，医疗保健提供者可以自动增强医疗记录。全面的 up-to-date患者病史有助于制定更准确的诊断和治疗计划。人工智能驱动的图像分析，例如解释超声检查和其他医学成像，可以提供快速而精确的见解，从而减少医疗专业人员的工作量并最大限度地降低人为错误的风险。

除了诊断和治疗之外，生成式人工智能还可以在预测分析中发挥关键作用。预测性分析可帮助医疗保健组织预测患者预后并相应地对护理计划进行个性化设置。该技术还可以优化管理流程，从管理患者数据到简化提供者与患者之间的沟通。通过将生成式人工智能解决方案与现有的医疗保健系统相结合，医疗机构可以提高效率，降低成本，并最终提供更高质量的护理。人工智能与医疗保健的整合不仅是一种增强，而且是向更智能、响应更快和以患者为中心的护理的根本转变。

高级图像分析

将 Amazon Bedrock 与数据存储（例如亚马逊 Neptune 和 A OpenSearch mazon Service）相结合，可以帮助您解决医疗保健领域高级图像分析的复杂性。信息检索解决方案可以通过评估诊断图像和解释超声检查来增强疾病发现过程并提高解释的准确性。该解决方案可以将视觉和文字评估数据与医生手动进行患者评估审查相结合。

解决方案工业化面临的挑战

在医疗保健领域实现人工智能解决方案工业化时，需要解决的主要障碍是数据质量和可用性。医疗保健数据通常以分散、不一致的格式存在。确保 AI 模型能够访问干净、结构化和具有代表性的数据，对于在现实场景中保持性能至关重要。由于生产环境，基础架构的可扩展性可能成为一项挑战。这些环境需要处理大量的实时患者数据，同时提供快速的响应时间并遵守数据隐私法规，例如健康保险便携性和责任法案 (HIPAA)。此外，随着新出现的医疗信息和患者数据会随着时间的推移而演变，需要对人工智能模型进行再训练和更新，以保持相关性并提供准确的建议。最后，由于互操作性问题以及需要与当前的

临床工作流程保持一致，将这些 AI 解决方案集成到现有的医疗保健系统中可能很复杂。这种集成需要技术和操作上的改变。

用例：使用增强的患者数据构建医疗智能应用程序

生成式人工智能可以通过增强临床和管理功能来帮助提高患者护理和工作人员的工作效率。人工智能驱动的图像分析（例如解释超声图）可加快诊断过程并提高准确性。它可以提供关键见解，为及时的医疗干预提供支持。

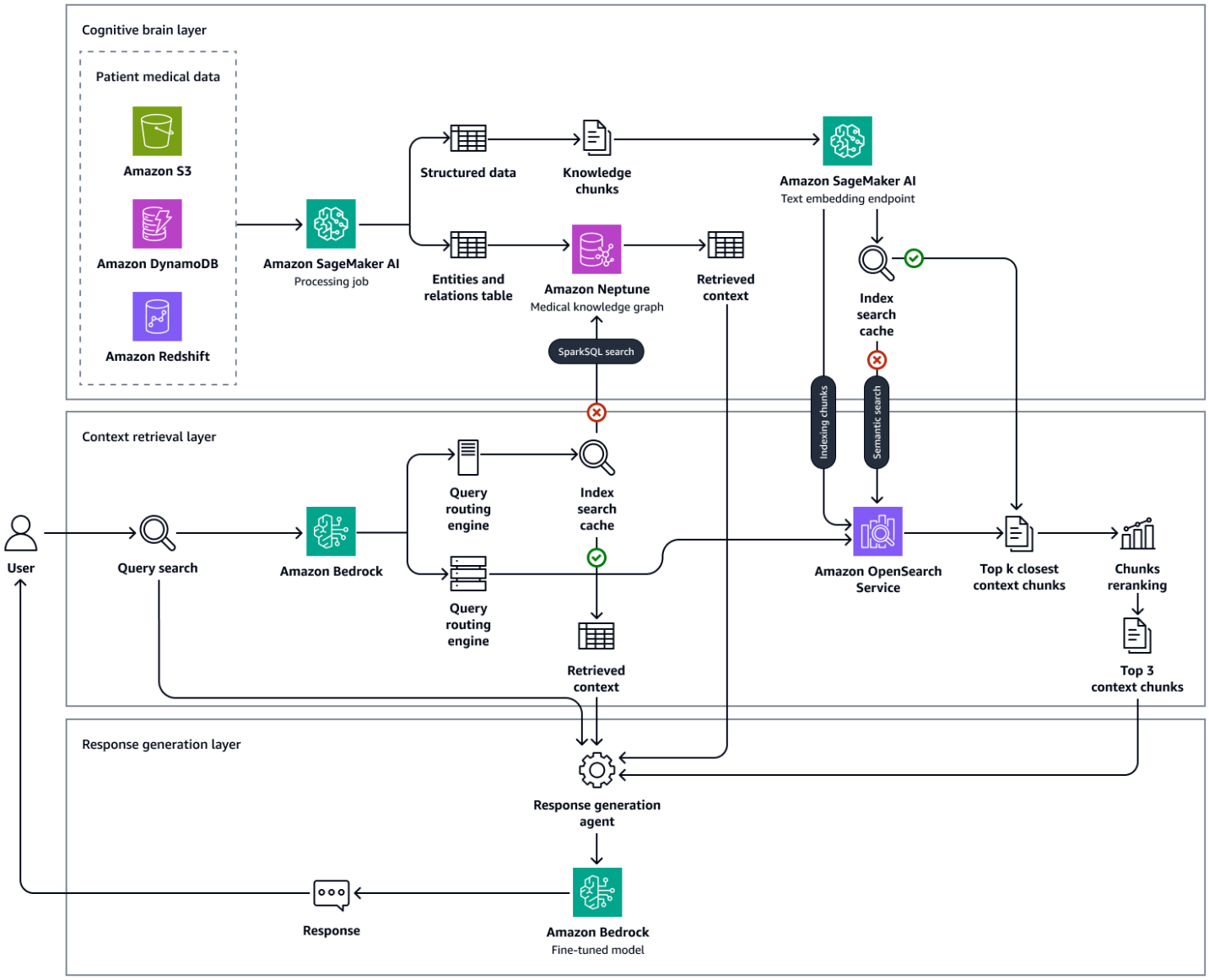
当您将生成式 AI 模型与知识图相结合时，您可以自动按时间顺序组织电子患者记录。这可以帮助您整合来自医患互动、症状、诊断、实验室结果和图像分析的实时数据。这为医生提供了全面的患者数据。这些数据可以帮助医生做出更准确、更及时的医疗决策，从而提高患者疗效和医疗保健提供者的工作效率。

解决方案概述

人工智能可以通过综合患者数据和医学知识来提供有价值的见解，从而增强医生和临床医生的能力。该检索增强生成 (RAG) 解决方案是一个医疗智能引擎，它使用来自数百万次临床互动的一整套患者数据和知识。它利用生成式人工智能的力量来创建基于证据的见解，以改善患者护理。它旨在增强临床工作流程，减少错误并改善患者预后。

该解决方案包括由提供支持的自动图像处理功能。 LLMs此功能减少了医务人员必须花费在手动搜索相似诊断图像和分析诊断结果上的时间。

下图显示了此解决方案 end-to-end-workflow的。它使用亚马逊 Neptune、亚马逊 SageMaker AI、亚马逊 OpenSearch 服务，以及亚马逊 Bedrock 中的基础模型。对于与 Neptune 中的医学知识图谱交互的上下文检索代理，您可以在 Amazon Bedrock 代理和 Amazon Bedrock 代理之间进行选择 LangChain 代理人。



在我们对医学问题样本的实验中，我们观察到，我们使用Neptune中维护的知识图谱、存放临床知识库的 OpenSearch矢量数据库和Amazon Bedrock中的方法 LLMs 得出的最终答案以事实为基础，通过减少误报和提高真阳性，准确得多。该解决方案可以生成有关患者健康状况的循证见解，旨在改善临床工作流程，减少错误并改善患者预后。

构建此解决方案包括以下步骤：

- 步骤 1：发现数据
- 第 2 步：建立医学知识图谱
- 第 3 步：构建上下文检索代理以查询医学知识图谱
- 第 4 步：创建实时描述性数据的知识库

- 第 5 步： LLMs 用于回答医疗问题

步骤 1：发现数据

您可以使用许多开源医疗数据集来支持医疗保健人工智能驱动解决方案的开发。其中一个数据集是 MIMIC-IV 数据集，它是一个公开可用的电子健康记录 (EHR) 数据集，广泛用于医疗保健研究界。MIMIC-IV 包含详细的临床信息，包括患者记录中的自由文本出院记录。您可以使用这些记录来尝试文本求和和实体提取技术。这些技术可以帮助您从非结构化文本中提取医疗信息（例如患者症状、给药和处方治疗）。

您也可以使用一个数据集，该数据集提供带注释、去识别化的患者出院摘要，这些摘要是专门为研究目的精心策划的。出院摘要数据集可以帮助您尝试实体提取，从而使您能够从文本中识别出关键的医疗实体（例如病症、手术和药物）。 第 2 步：建立医学知识图谱本指南介绍了如何使用从 MIMIC-IV 和出院摘要数据集中提取的结构化数据来创建医学知识图表。该医学知识图谱是医疗保健专业人员高级查询和决策支持系统的支柱。

除了基于文本的数据集外，您还可以使用图像数据集。例如，肌肉骨骼 X 光片 (MURA) 数据集，这是一个包含骨骼多视角射线照相图像的综合数据库。使用此类图像数据集通过医学图像解码技术进行诊断评估。这些解码技术对于肌肉骨骼疾病、心血管疾病和骨质疏松症等疾病的早期诊断至关重要。通过微调医学图像数据集上的视觉和语言基础模型，您可以检测诊断图像中的异常。这有助于该系统为临床医生提供早期和准确的诊断见解。通过使用图像和文本数据集，您可以创建人工智能驱动的医疗保健应用程序，该应用程序能够处理文本和图像数据，从而改善患者护理。

第 2 步：建立医学知识图谱

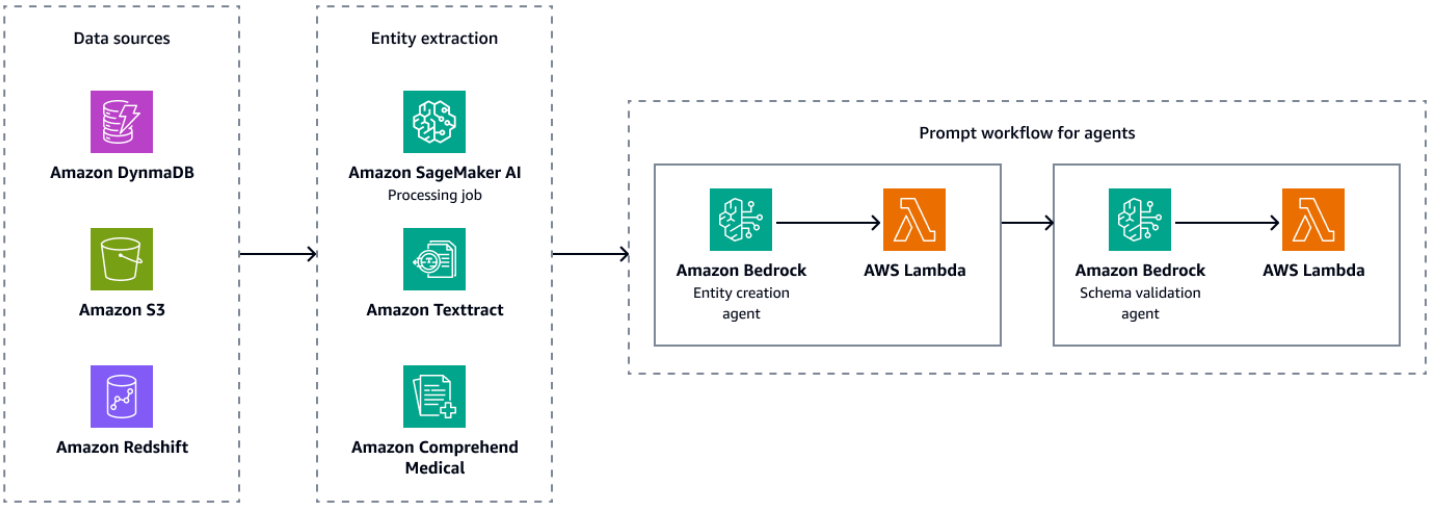
对于任何想要建立基于庞大知识库的决策支持系统的医疗保健组织来说，关键的挑战是找到和提取临床记录、医学期刊、出院摘要和其他数据源中存在的医疗实体。您还需要从这些病历中获取时间关系、受试者和确定性评估，以便有效地使用提取的实体、属性和关系。

第一步是通过对基础模型（例如Amazon Bedrock中的Llama 3）使用少量提示从非结构化医学文本中提取医学概念。Few-shot 提示是指在让 LLM 执行类似任务之前，向其提供少量演示任务和所需输出的示例。使用基于 LLM 的医学实体提取器，您可以解析非结构化医学文本，然后生成医学知识实体的结构化数据表示形式。您还可以存储患者属性以进行下游分析和自动化。实体提取过程包括以下操作：

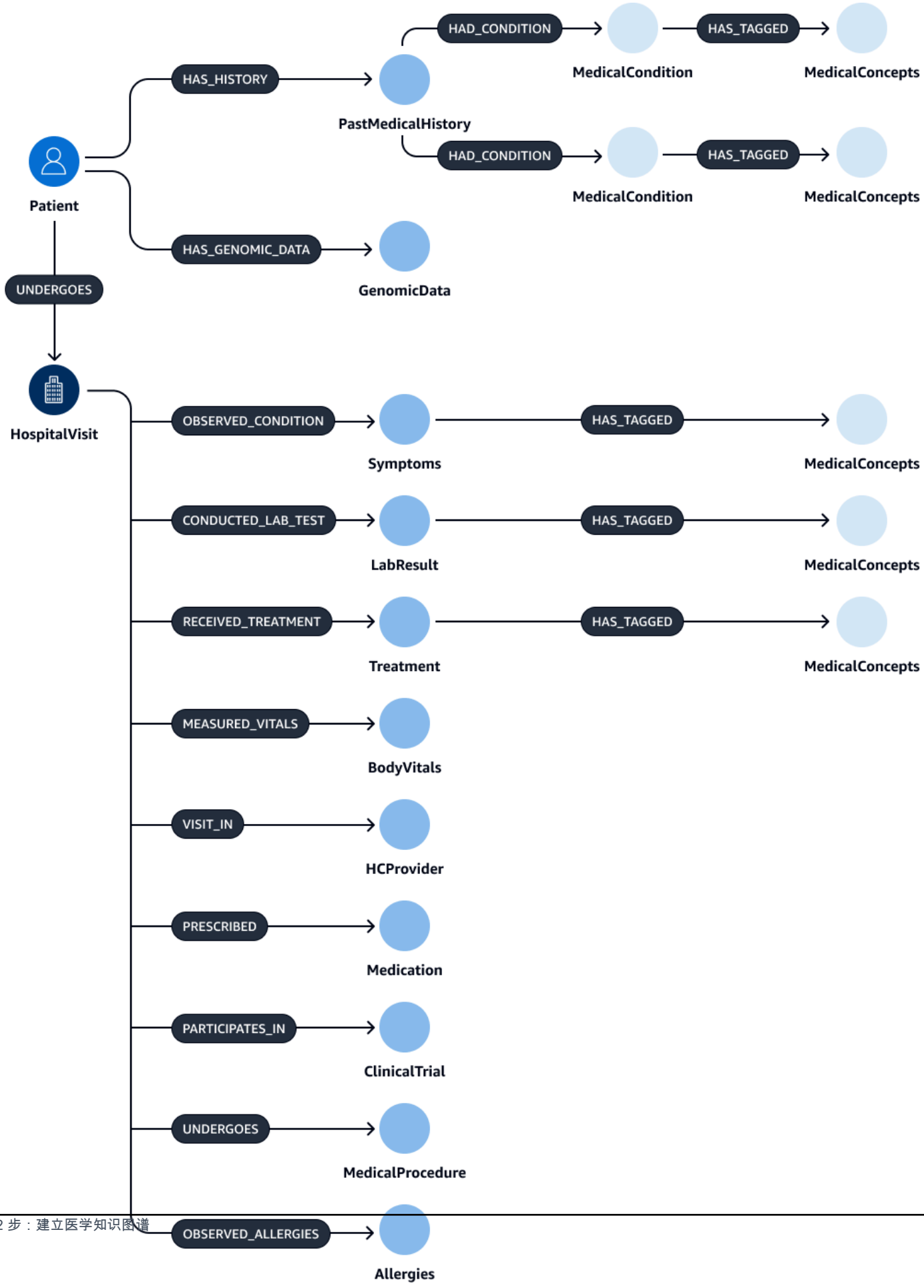
- 提取有关医学概念的信息，例如疾病、药物、医疗器械、剂量、用药频率、用药持续时间、症状、医疗程序及其临床相关属性。
- 捕获功能特征，例如提取的实体、受试者和确定性评估之间的时间关系。

- 扩展标准医学词汇，例如：
  - 来自数据库的概念标识符 (rxCUI) RxNorm
  - 《国际疾病分类》第10次修订版，临床修改 (ICD-10-CM) 中的代码
  - 医学主题标题 (MeSH) 中的术语
  - 来自系统化医学命名法、临床术语（SNOMED CT）的概念
  - 来自统一医学语言系统 (UMLS) 的代码
- 汇总出院记录并从笔录中得出医学见解。

下图显示了创建实体、属性和关系的有效配对组合的实体提取和架构验证步骤。您可以在亚马逊简单存储服务 (Amazon S3) Simple Storage Service 中存储非结构化数据，例如出院摘要或患者记录。您可以在 Amazon Redshift 和 Amazon DynamoDB 中存储结构化数据，例如企业资源规划 (ERP) 数据、电子患者记录和实验室信息系统。您可以构建 Amazon Bedrock 实体创建代理。该代理可以整合服务，例如亚马逊 SageMaker 人工智能数据提取管道、Amazon Textract和Amazon Comprehend Medical，从结构化和非结构化数据源中提取实体、关系和属性。最后，您可以使用 Amazon Bedrock 架构验证代理来确保提取的实体和关系符合预定义的图形架构，并保持节点边缘连接和关联属性的完整性。



提取和验证实体、关系和属性后，可以将它们链接起来创建 subject-object-predicate三元组。您将这些数据提取到 Amazon Neptune 图表数据库中，如下图所示。图形数据库经过优化，可以存储和查询数据项之间的关系。


HAD_CONDITION
HAS_TAGGED
MedicalCondition
MedicalConcepts
HAS_HISTORY
PastMedicalHistory
HAD_CONDITION
HAS_TAGGED
MedicalCondition
MedicalConcepts
Patient
HAS_GENOMIC_DATA
GenomicData
UNDERGOES
HospitalVisit
OBSERVED_CONDITION
HAS_TAGGED
Symptoms
MedicalConcepts
CONDUCTED_LAB_TEST
HAS_TAGGED
LabResult
MedicalConcepts
RECEIVED_TREATMENT
HAS_TAGGED
Treatment
MedicalConcepts
MEASURED_VITALS
BodyVitals
VISIT_IN
HCProvider
PRESCRIBED
Medication
PARTICIPATES_IN
ClinicalTrial
UNDERGOES
MedicalProcedure
OBSERVED_ALLERGIES
Allergies

你可以用这些数据创建一个全面的知识图谱。知识图可帮助您整理和查询各种关联信息。例如，您可以创建一个包含以下主要节点的知识图谱：HospitalVisit、PastMedicalHistory、Symptoms、MedicationMedicalProcedures、和Treatment。

下表列出了您可以从出院通知中提取的实体及其属性。

实体	Attributes
Patient	PatientID ，Name, Age, Gender, Address, ContactInformation
HospitalVisit	VisitDate ，Reason, Notes
HealthcareProvider	ProviderID ，Name, Specialty ，ContactInformation ，Address, AffiliatedInstitution
Symptoms	Description ，RiskFactors
Allergies	AllergyType ，Duration
Medication	MedicationID ，Name, Description ，Dosage, SideEffects ，Manufacturer
PastMedicalHistory	ContinuingMedicines
MedicalCondition	ConditionName ，Severity, Treatment Received ，DoctorinCharge ，HospitalName ，MedicinesFollowed
BodyVitals	HeartRate ，BloodPressure ，RespiratoryRate ，BodyTemperature ，BMI
LabResult	LabResultID ，PatientID ，TestName, Result, Date
ClinicalTrial	TrialID, Name, Description ，Phase, Status, StartDate ，EndDate

实体	Attributes
GenomicData	GenomicDataID , PatientID , SequenceData , VariantInformation
Treatment	TreatmentID , Name, Description , Type, SideEffects
MedicalProcedure	ProcedureID , Name, Description , Risks, Outcomes
MedicalConcepts	UMLSCodes , MedicalVocabularies

下表列出了实体可能具有的关系及其对应的属性。例如，该Patient实体可能通过该[UNDERGOES]关系连接到HospitalVisit实体。这种关系的属性是VisitDate。

主体实体	关系	对象实体	Attributes
Patient	[UNDERGOES]	HospitalVisit	VisitDate
HospitalVisit	[VISIT_IN]	HealthcareProvider	ProviderName , Location, ProviderID , VisitDate
HospitalVisit	[OBSERVED_CONDITION]	Symptoms	Severity, CurrentStatus , VisitDate
HospitalVisit	[RECEIVED_TREATMENT]	Treatment	Duration, Dosage, VisitDate
HospitalVisit	[PRESCRIBED]	Medication	Duration, Dosage, Adherence , VisitDate
Patient	[HAS_HISTORY]	PastMedicalHistory	无

主体实体	关系	对象实体	Attributes
PastMedicalHistory	[HAD_CONDITION]	MedicalCondition	DiagnosisDate，CurrentStatus
HospitalVisit	[PARTICIPATES_IN]	ClinicalTrial	VisitDate，Status, Outcomes
Patient	[HAS_GENOMIC_DATA]	GenomicData	CollectionDate
HospitalVisit	[OBSERVED_ALLERGIES]	Allergies	VisitDate
HospitalVisit	[CONDUCTED_LAB_TEST]	LabResult	VisitDate，AnalysisDate，Interpretation
HospitalVisit	[UNDERGOES]	MedicalProcedure	VisitDate，Outcome
MedicalCondition	[HAS_TAGGED]	MedicalConcepts	无
LabResult	[HAS_TAGGED]	MedicalConcepts	无
Treatment	[HAS_TAGGED]	MedicalConcepts	无
Symptoms	[HAS_TAGGED]	MedicalConcepts	无

第 3 步：构建上下文检索代理以查询医学知识图谱

构建医学图形数据库后，下一步是构建用于图形交互的代理。这些代理会检索医生或临床医生输入的查询的正确和必需的上下文。配置这些从知识图谱中检索上下文的代理有多种选项：

- 亚马逊 Bedrock 代理商
- LangChain 代理

用于图形交互的 Amazon Bedrock 代理

亚马逊 Bedrock 代理可与亚马逊 Neptune 图形数据库无缝协作。您可以通过 Amazon Bedrock 操作组进行高级互动。操作组通过调用一个运行 Neptune OpenCypher 查询的 AWS Lambda 函数来启动该进程。

要查询知识图谱，您可以使用两种不同的方法：直接执行查询或使用上下文嵌入进行查询。这些方法可以独立应用，也可以组合使用，具体取决于您的具体用例和排名标准。通过结合这两种方法，您可以为法学硕士提供更全面的背景信息，从而改善结果。以下是两种查询执行方法：

- 无需嵌入即可@@ 直接执行 Cypher 查询 — Lambda 函数无需任何基于嵌入的搜索即可直接针对 Neptune 执行查询。以下是这种方法的示例：

```
MATCH (p:Patient)-[u:UNDERGOES]->(h:HospitalVisit) WHERE h.Reason = 'Acute Diabetes'
 AND date(u.VisitDate) > date('2024-01-01')
RETURN p.PatientID, p.Name, p.Age, p.Gender, p.Address, p.ContactInformation
```

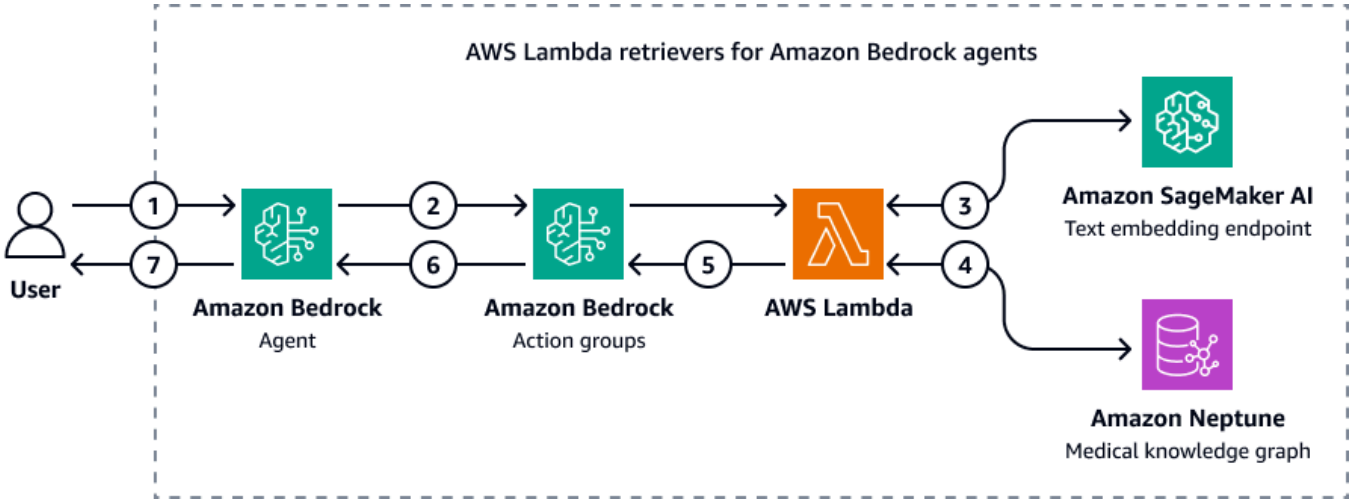
- 使用嵌入式搜索直接执行 Cypher 查询 — Lambda 函数使用嵌入式搜索来增强查询结果。这种方法通过合并嵌入来增强查询执行，嵌入是数据的密集向量表示形式。当查询需要语义相似性或超出精确匹配范围的更广泛理解时，嵌入式特别有用。您可以使用预先训练或自定义训练的模型为每种疾病生成嵌入数据。以下是这种方法的示例：

```
CALL { WITH "Acute Diabetes" AS query_term RETURN search_embedding(query_term) AS
 similar_reasons }

MATCH (p:Patient)-[u:UNDERGOES]->(h:HospitalVisit) WHERE h.Reason IN similar reasons
 AND date(u.VisitDate) > date('2024-01-01')
RETURN p.PatientID, p.Name, p.Age, p.Gender, p.Address, p.ContactInformation
```

  在此示例中，该`search_embedding("Acute Diabetes")`函数检索语义上接近“急性糖尿病”的病症。这有助于查询还能找到患有糖尿病前期或代谢综合征等疾病的患者。

下图显示了亚马逊 Bedrock 代理如何与亚马逊 Neptune 交互以对医学知识图谱执行密码查询。



图表显示了以下工作流：

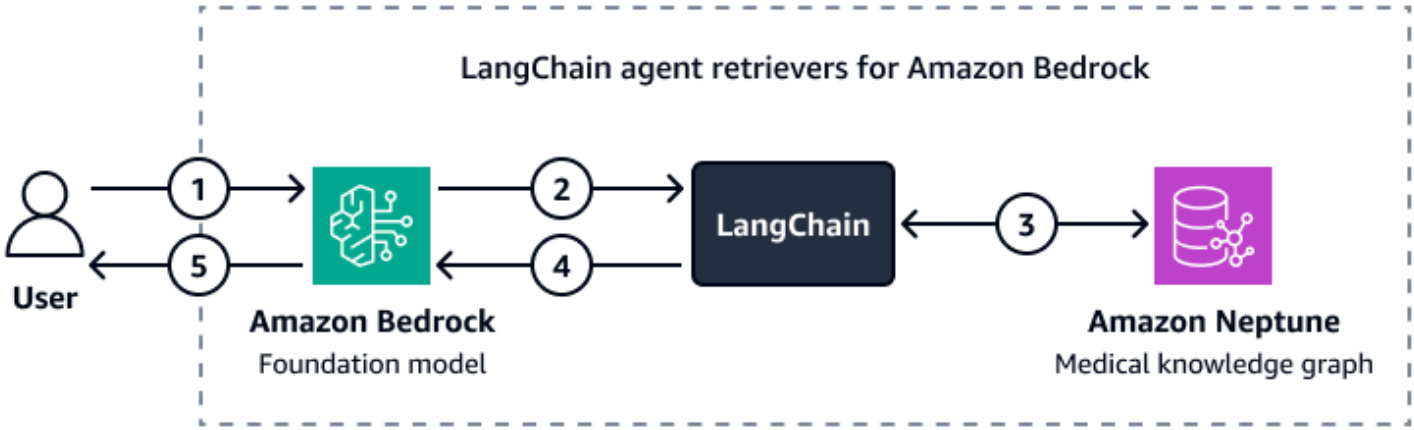
1. 用户向 Amazon Bedrock 代理提交问题。
2. Amazon Bedrock 代理将问题和输入筛选变量传递给 Amazon Bedrock 行动小组。这些操作组包含一个与亚马逊 A SageMaker I 文本嵌入端点和亚马逊 Neptune 医学知识图谱交互的 AWS Lambda 函数。
3. Lambda 函数与 SageMaker AI 文本嵌入端点集成，可在 OpenCypher 查询中执行语义搜索。它使用底层语言将自然语言查询转换为 OpenCypher 查询 LangChain 代理人。
4. Lambda 函数在海王星医学知识图中查询正确的数据集，并接收来自海王星医学知识图的输出。
5. Lambda 函数将 Neptune 的结果返回给亚马逊 Bedrock 行动小组。
6. 亚马逊 Bedrock 行动小组将检索到的上下文发送给亚马逊 Bedrock 代理。
7. Amazon Bedrock 代理使用原始用户查询和从知识图谱中检索到的上下文生成响应。

LangChain 用于图形交互的代理

你可以整合 LangChain 使用 Neptune 可以实现基于图形的查询和检索。这种方法可以通过使用 Neptune 中的图形数据库功能来增强 AI 驱动的工作流程。习俗 LangChain 寻回犬充当中介。Amazon Bedrock 中的基础模型可以通过使用直接的 Cypher 查询和更复杂的图形算法与 Neptune 进行交互。

你可以使用自定义的检索器来细化 LangChain 代理与 Neptune 图算法进行交互。例如，您可以使用少量提示，它可以帮助您根据特定的模式或示例定制基础模型的响应。您还可以应用 LLM 识别的过滤器来完善上下文并提高响应的精度。在与复杂的图形数据交互时，这可以提高整个检索过程的效率和准确性。

下图显示了如何自定义 LangChain 代理精心策划了亚马逊 Bedrock 基础模型和亚马逊 Neptune 医学知识图谱之间的互动。



图表显示了以下工作流：

1. 用户向 Amazon Bedrock 提交问题和 LangChain 代理人。
2. Amazon Bedrock 基础模型使用 Neptune 架构，该架构由 LangChain 代理，为用户的问题生成查询。
3. 这些区域有：LangChain 代理对照 Amazon Neptune 医学知识图谱运行查询。
4. 这些区域有：LangChain 代理将检索到的上下文发送到 Amazon Bedrock 基础模型。
5. Amazon Bedrock 基础模型使用检索到的上下文来生成用户问题的答案。

第 4 步：创建实时描述性数据的知识库

接下来，您将创建一个包含实时、描述性的医患互动笔记、诊断图像评估和实验室分析报告的知识库。该知识库是一个矢量数据库。通过使用矢量数据库，该数据库可以以索引、矢量化的形式存储描述性医学知识，医疗保健提供者可以高效地从庞大的存储库中查询和访问相关信息。这些矢量化表示法可帮助您检索语义上相似的数据。护理提供者可以快速浏览临床记录、医学图像和实验室结果。通过提供即时访问情境相关信息，提高诊断和治疗计划的准确性和速度，从而加快知情决策。

使用 OpenSearch 服务医疗知识库

Amazon S OpenSearch ervic e 可以管理大量的高维医疗数据。它是一项托管服务，可促进高性能搜索和实时分析。它非常适合作为 RAG 应用程序的矢量数据库。 OpenSearch 服务充当后端工具，用于管理大量非结构化或半结构化数据，例如医疗记录、研究文章和临床记录。其高级语义搜索功能可帮助您检索与上下文相关的信息。这使得它在临床决策支持系统、患者查询解决工具和医疗保健知识管理系

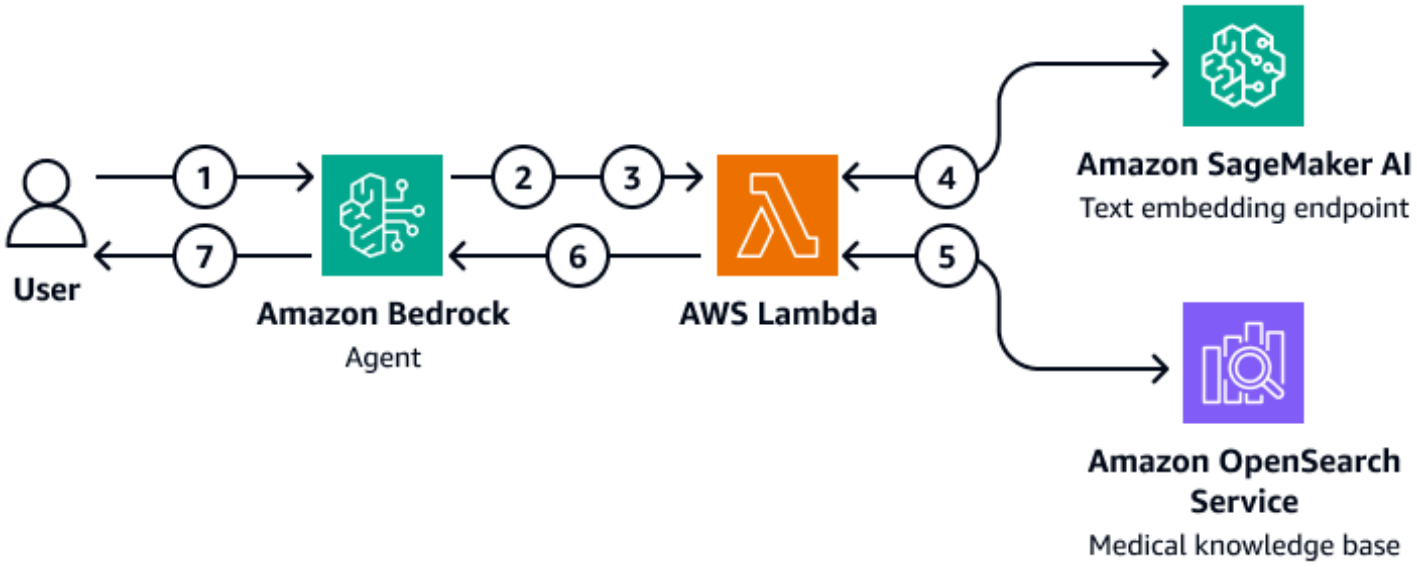
统等应用中特别有用。例如，临床医生可以快速找到与特定症状或治疗方案相匹配的相关患者数据或研究。这可以帮助临床医生根据最 up-to-date相关的信息做出决策。

OpenSearch 服务可以扩展和处理实时数据索引和查询。这使其成为动态医疗保健环境的理想之选，在这种环境中，及时访问准确的信息至关重要。此外，它还具有多模式搜索功能，最适合需要多个输入的搜索，例如医学图像和医生笔记。在为医疗保健应用程序实施 OpenSearch 服务时，必须定义精确的字段和映射，以优化数据索引和检索。字段表示各个数据，例如患者记录、病史和诊断代码。映射定义了如何存储这些字段（以嵌入形式或原始形式）和查询这些字段。对于医疗保健应用程序，必须建立适应各种数据类型的映射，包括结构化数据（例如数值测试结果）、半结构化数据（例如患者记录）和非结构化数据（例如医学图像）

在 S OpenSearch ervice 中，您可以通过精心策划的提示执行全文神经搜索查询，搜索病历、临床记录或研究论文，从而快速找到有关特定症状、治疗或患者病史的相关信息。神经搜索查询使用内置的神经网络模型自动处理输入提示和图像的嵌入。这有助于它理解和捕捉多模态数据中更深层次的语义关系，与其他搜索查询算法（例如 k-nearest Neighbor (k-nn) 搜索）相比，可提供更具上下文感知能力和精确度的搜索结果。

创建 RAG 架构

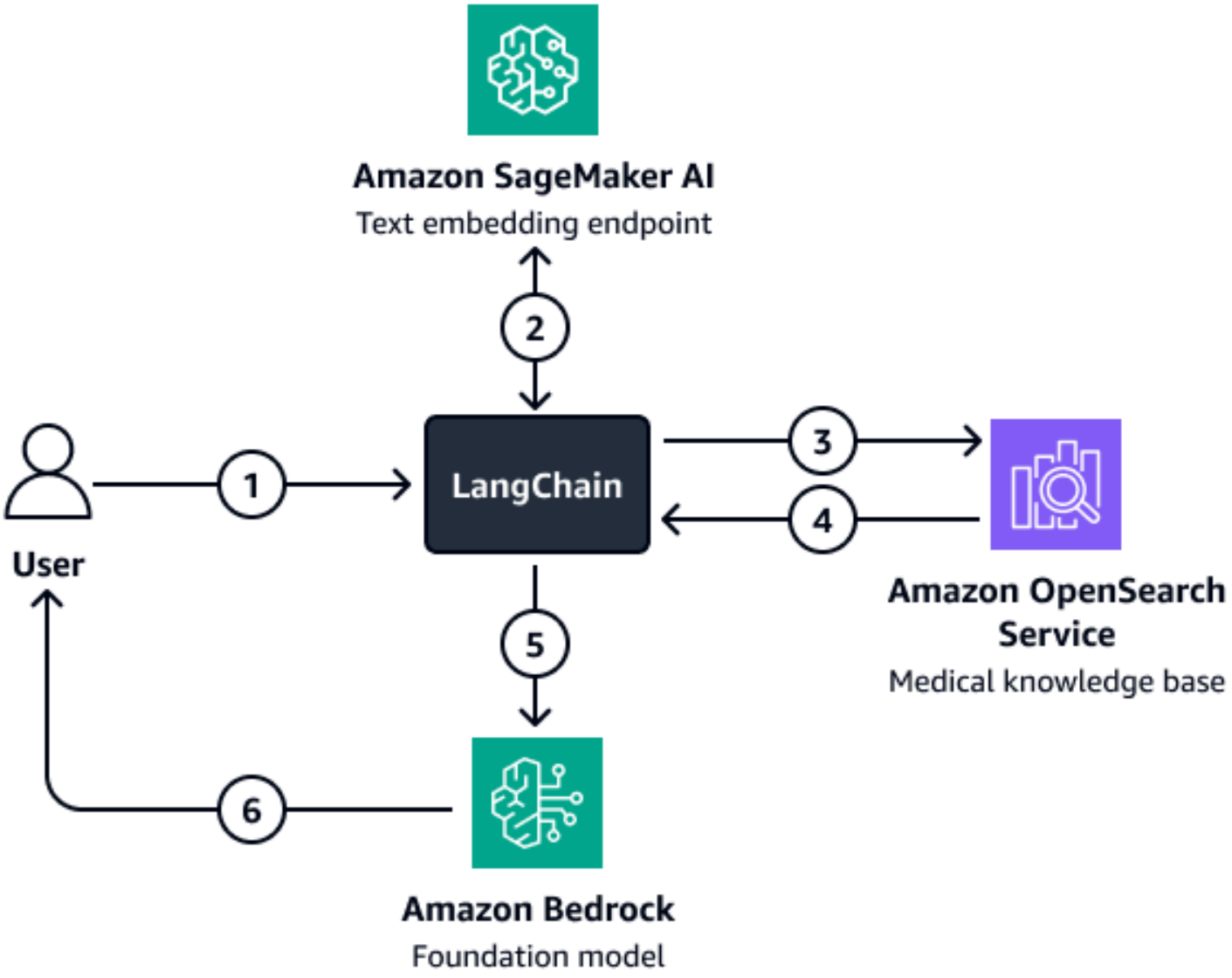
您可以部署自定义 RAG 解决方案，该解决方案使用 Amazon Bedrock 代理在服务中 OpenSearch 查询医学知识库。为此，您需要创建一个可以与 OpenSearch 服务进行交互和查询的 AWS Lambda 函数。Lambda 函数通过访问 A SageMaker I 文本嵌入端点来嵌入用户的输入问题。Amazon Bedrock 代理会将其他查询参数作为输入传递给 Lambda 函数。该函数在 Service 中查询 S OpenSearch ervice 中的医学知识库，返回相关的医学内容。设置 Lambda 函数后，将其作为操作组添加到 Amazon Bedrock 代理中。Amazon Bedrock 代理接收用户的输入，识别必要的变量，将变量和问题传递给 Lambda 函数，然后启动该函数。该函数返回一个上下文，该上下文可帮助基础模型为用户的问题提供更准确的答案。



图表显示了以下工作流：

1. 用户向 Amazon Bedrock 代理提交了一个问题。
2. Amazon Bedrock 代理选择要启动的操作组。
3. Amazon Bedrock 代理启动一个 AWS Lambda 函数并将参数传递给该函数。
4. Lambda 函数启动 Amazon A SageMaker I 文本嵌入模型以嵌入用户问题。
5. Lambda 函数将嵌入的文本以及其他参数和筛选条件传递给亚马逊 OpenSearch 服务。亚马逊 OpenSearch 服务查询医学知识库并将结果返回到 Lambda 函数。
6. Lambda 函数将结果传回给亚马逊 Bedrock 代理。
7. Amazon Bedrock 代理中的基础模型根据结果生成响应，并将响应返回给用户。

对于涉及更复杂筛选的情况，您可以使用自定义 LangChain 寻回犬。通过设置直接加载到的 OpenSearch 服务矢量搜索客户端来创建此检索器 LangChain。 这种架构允许您传递更多变量来创建过滤器参数。设置好寻回犬后，使用 Amazon Bedrock 模型和寻回器设置检索问答链。该链通过将用户输入和潜在的过滤器传递给检索器来协调模型和检索器之间的交互。检索器返回相关的上下文，帮助基础模型回答用户的问题。



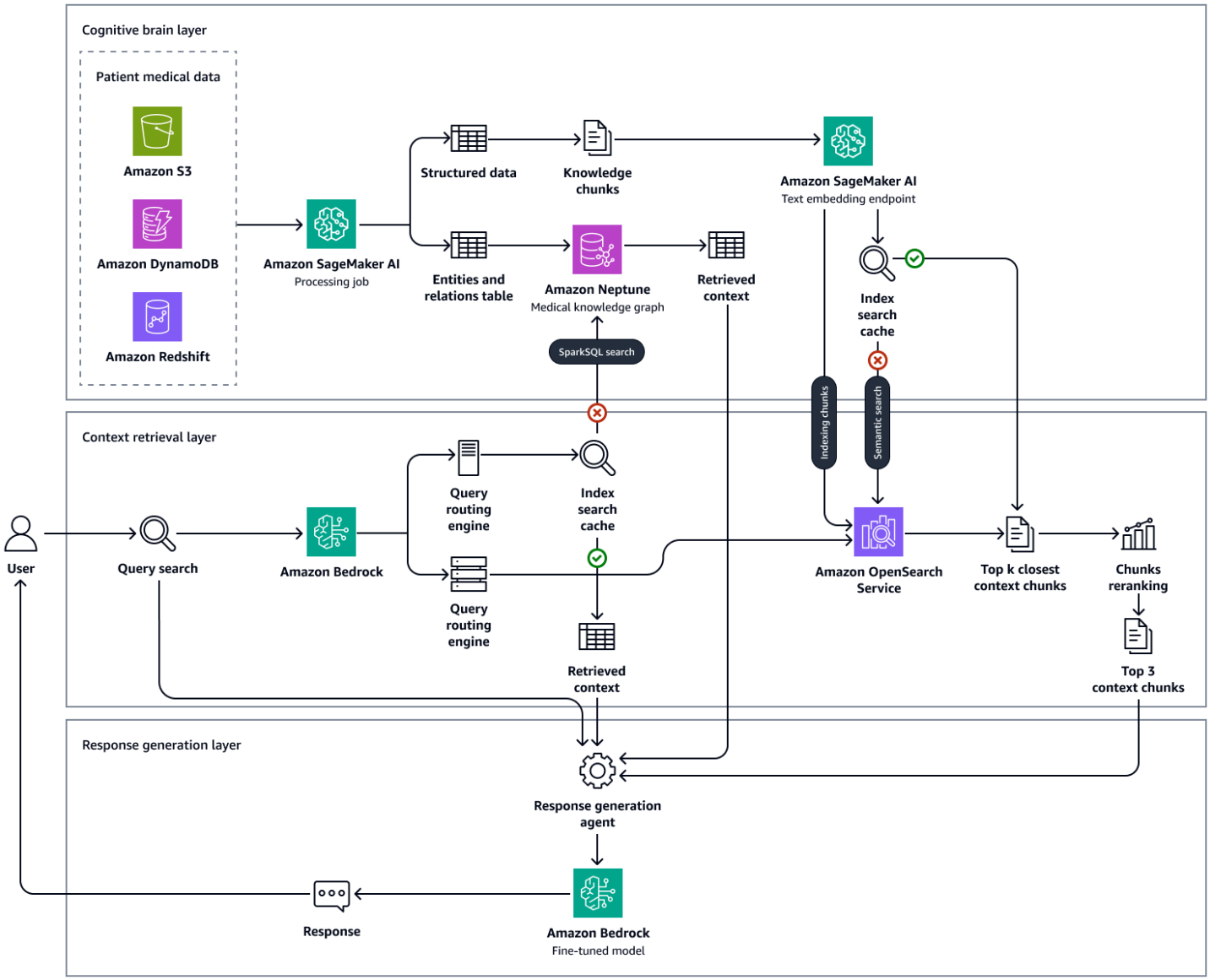
图表显示了以下工作流：

1. 用户向 LangChain 寻回犬代理。
2. 这些区域有：LangChain 检索器代理将问题发送到 Amazon A SageMaker I 文本嵌入端点以嵌入问题。
3. 这些区域有：LangChain 检索器代理将嵌入的文本传递给 Amazon OpenSearch 服务。
4. Amazon OpenSearch 服务会将检索到的文档返回到 LangChain 寻回犬代理。
5. 这些区域有：LangChain 检索器代理将用户问题和检索到的上下文传递给 Amazon Bedrock 基础模型。
6. 基础模型生成响应并将其发送给用户。

第 5 步： LLMs 用于回答医疗问题

前面的步骤可帮助您构建医疗智能应用程序，该应用程序可以获取患者的病历并汇总相关药物和潜在诊断。现在，你构建生成层。该层使用 Amazon Bedrock 中的 LLM（例如 Llama 3）的生成功能来增强应用程序的输出。

当临床医生输入查询时，应用程序的上下文检索层会从知识图中执行检索过程，并返回与患者病史、人口统计、症状、诊断和结果相关的热门记录。它还从矢量数据库中检索实时、描述性的医患互动笔记、诊断图像评估见解、实验室分析报告摘要以及来自大量医学研究和学术书籍的见解。然后，这些检索率最高的结果、临床医生的查询和提示（根据查询的性质量身定制答案）将传递给 Amazon Bedrock 中的基础模型。这是响应生成层。LLM 使用检索到的上下文生成对临床医生查询的响应。下图显示了此解决方案中各步骤 end-to-end的工作流程。



您可以在 Amazon Bedrock 中使用预先训练的基础模型（例如 Llama 3）来处理医疗智能应用程序必须处理的一系列用例。对于给定任务，最有效的法学硕士学位因用例而异。例如，预先训练的模型可能足以总结患者与医生的对话，搜索药物和患者病史，并从内部医疗数据集和科学知识中检索见解。但是，对于其他复杂的用例，例如实时实验室评估、医疗程序建议和患者预后预测，可能需要进行微调的法学硕士。您可以通过在医学领域数据集上训练法学硕士来对其进行微调。特定或复杂的医疗保健和生命科学要求推动了这些微调模型的开发。

有关微调法学硕士学位或选择已接受过医学领域数据培训的现有法学硕士学位的更多信息，请参阅在医疗保健和生命科学用例中使用大型语言模型。

与 Well-Architect AWS ed 框架保持一致

该解决方案与 Well-Ar AWS chitected Framework 的所有六大支柱保持一致，如下所示：

- 卓越运营 — 该架构已分离，可实现高效的监控和更新。Amazon Bedrock 代理并 AWS Lambda 帮助您快速部署和回滚工具。
- 安全 — 此解决方案旨在遵守医疗保健法规，例如 HIPAA。您还可以实施加密、精细访问控制和 Amazon Bedrock 护栏，以帮助保护患者数据。
- 可靠性 — AWS 托管服务，例如亚马逊 OpenSearch 服务和 Amazon Bedrock，为持续的模型交互提供了基础设施。
- 性能效率 — RAG 解决方案使用优化的语义搜索和 Cypher 查询快速检索相关数据，而代理路由器则为用户查询确定最佳路由。
- 成本优化 — Amazon Bedrock 和 RAG 架构中的 pay-per-token模型降低了推理和预训练成本。
- 可持续性-使用无服务器基础设施和 pay-per-token计算可以最大限度地减少资源使用并增强可持续性。

用例：预测患者预后和再入院率

人工智能驱动的预测分析通过预测患者疗效和实现个性化治疗计划来提供更多好处。这可以提高患者的满意度和健康结果。通过将这些 AI 功能与 Amazon Bedrock 和其他技术集成，医疗保健提供商可以显著提高工作效率、降低成本并提高患者护理的整体质量。

您可以将医疗数据（例如患者病史、临床记录、药物和治疗方法）存储在知识图中。通过将对情境的深刻理解 LLMs 与医学知识图谱中的结构化时间数据相结合，医疗保健提供者可以获得对个体患者模式的更多见解。使用预测分析，您可以尽早发现潜在的不依从性或治疗并发症，并生成个性化的再入院倾向评分。

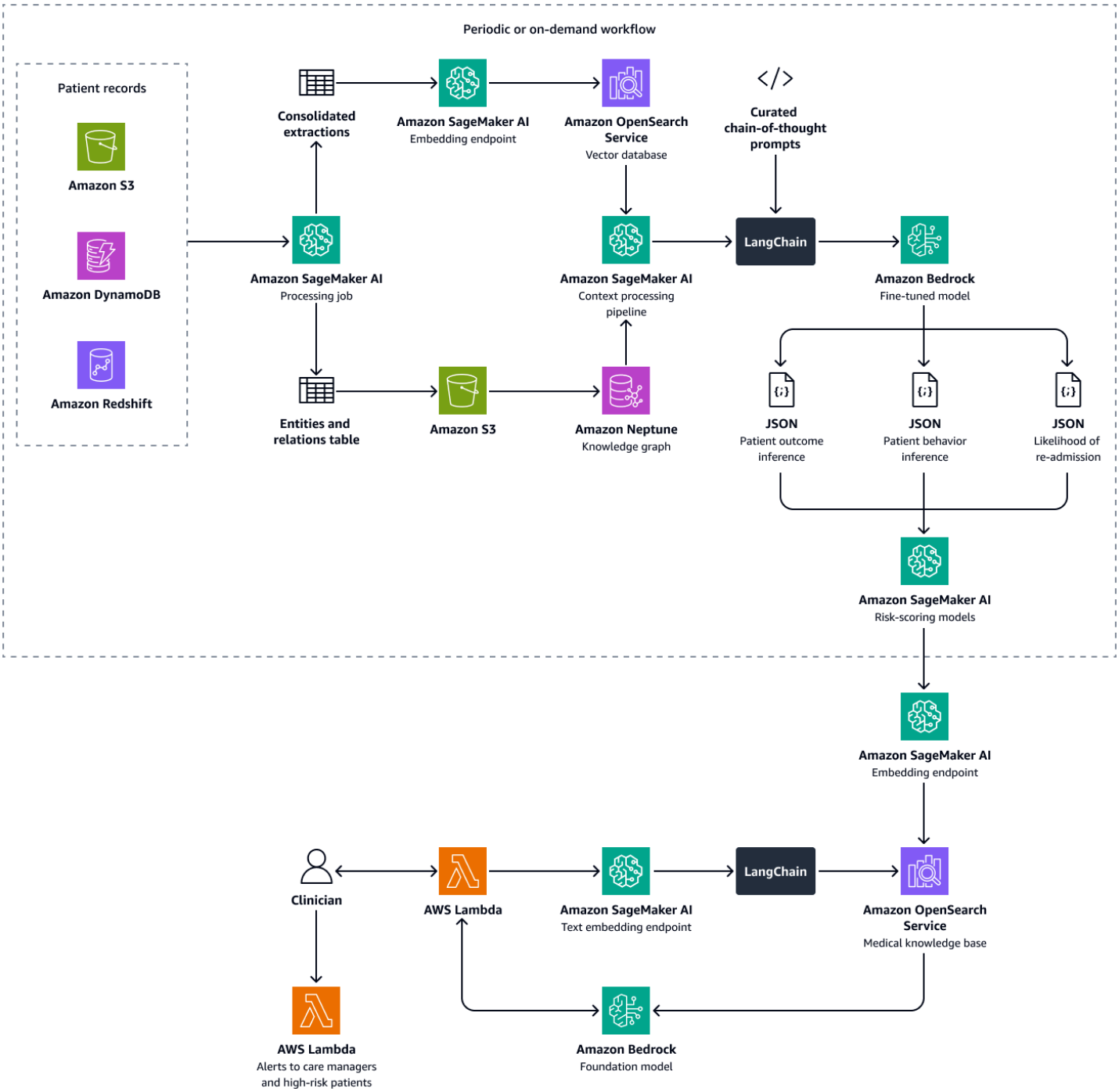
此解决方案可帮助您预测重新入院的可能性。这些预测可以改善患者的预后并降低医疗成本。该解决方案还可以帮助医院临床医生和管理人员将注意力集中在再入院风险较高的患者身上。它还可以通过警报、自助服务和数据驱动的行动，帮助他们主动干预这些患者。

解决方案概述

该解决方案使用多检索器检索增强生成 (RAG) 框架来分析患者数据。它可以预测个别患者再次入院的可能性，并帮助您计算医院级别的再入院倾向评分。该解决方案集成了以下功能：

- 知识图表 — 存储按时间顺序排列的结构化患者数据，例如医院就诊情况、以前的再入院情况、症状、实验室结果、处方治疗和药物依从性历史记录
- 矢量数据库 — 存储非结构化临床数据，例如出院摘要、医生记录以及错过预约或报告的药物副作用的记录
- 经过微调的法学硕士 — 使用知识图谱中的结构化数据和来自矢量数据库的非结构化数据，以得出有关患者行为、治疗依从性和再入院可能性的推断

风险评分模型将法学硕士的推论量化为数字分数。您可以将分数汇总为医院级别的再入院倾向分数。该分数定义了每位患者的风险敞口，您可以定期或根据需要进行计算。所有推断和风险评分都已编制索引并存储在 Amazon S OpenSearch ervice 中，以便护理经理和临床医生可以对其进行检索。通过将对话式 AI 代理与该矢量数据库集成，临床医生和护理经理可以无缝提取个人患者级别、机构范围或医学专业的见解。您还可以根据风险评分设置自动警报，以鼓励主动干预。



构建此解决方案包括以下步骤：

- 第 1 步：使用医学知识图谱预测患者预后
- 第 2 步：预测患者对处方药或治疗的行为
- 第 3 步：预测患者再次入院的可能性
- 第 4 步：计算再入院倾向分数

第 1 步：使用医学知识图谱预测患者预后

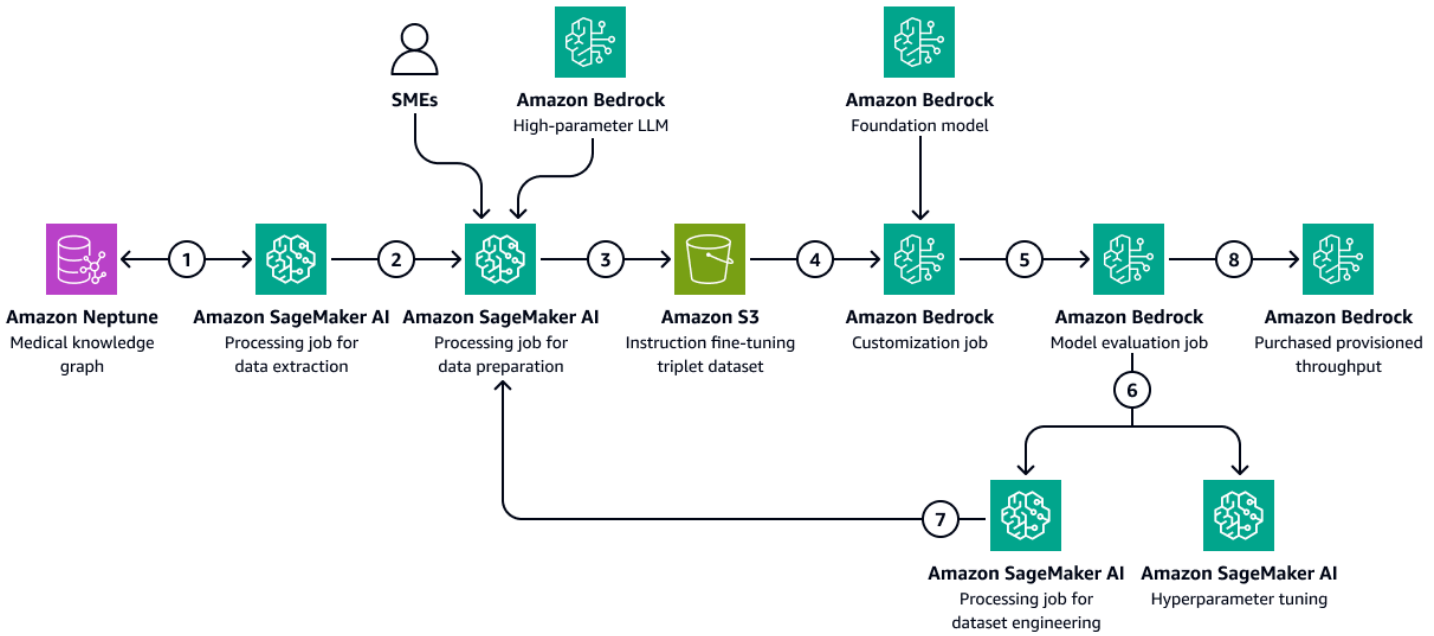
在 Amazon Neptune 中，您可以使用知识图来存储一段时间内有关患者就诊和结果的时间知识。构建和存储知识图谱的最有效方法是使用图模型和图形数据库。图形数据库专为存储和浏览关系而构建。图形数据库可以更轻松地对高度互联的数据进行建模和管理，并具有灵活的架构。

知识图可帮助您执行时间序列分析。以下是图表数据库中用于患者预后的时间预测的关键元素：

- 历史数据 — 患者先前的诊断、持续服药、以前使用的药物和实验室结果
- 患者就诊（按时间顺序）— 就诊日期、症状、观察到的过敏症、临床记录、诊断、手术、治疗、处方药和实验室结果
- 症状和临床参数 — 临床和基于症状的信息，包括严重程度、进展模式和患者对药物的反应

你可以利用医学知识图谱中的见解来微调 Amazon Bedrock 中的法学硕士，例如 Llama 3。你可以使用关于患者在一段时间内对一组药物或治疗的反应的顺序患者数据来微调法学硕士。使用带有标签的数据集，该数据集将一组药物或治疗方法以及患者与诊所的互动数据分为预定义的类别，以表明患者的健康状况。这些类别的例子包括健康状况恶化、改善或稳定进展。当临床医生输入有关患者及其症状的新背景时，经过微调的法学硕士可以使用训练数据集中的模式来预测潜在的患者预后。

下图显示了使用医疗保健专用训练数据集在 Amazon Bedrock 中微调 LLM 所涉及的顺序步骤。这些数据可能包括一段时间内患者的健康状况和对治疗的反应。该训练数据集将帮助模型对患者预后做出广义预测。



图表显示了以下工作流：

1. Amazon SageMaker AI 数据提取任务查询知识图表，以检索按时间顺序排列的数据，了解不同患者在一段时间内对一组药物或治疗的反应。
2. SageMaker 人工智能数据准备工作整合了 Amazon Bedrock LLM 和主题专家的意见 () SMEs。该工作将从知识图谱中检索到的数据分为预定义的类别（例如健康状况恶化、改善或稳定进展），这些类别表明了每位患者的健康状况。
3. 该工作创建了一个微调数据集，其中包括从知识图谱中提取的信息、 chain-of-thought提示和患者结果类别。它会将此训练数据集上传到 Amazon S3 存储桶。
4. Amazon Bedrock 自定义任务使用此训练数据集来微调 LLM。
5. Amazon Bedrock 定制工作集成了训练环境中首选的 Amazon Bedrock 基础模型。它启动微调作业，并使用您配置的训练数据集和训练超参数。
6. Amazon Bedrock 评估工作使用预先设计的模型评估框架对经过微调的模型进行评估。
7. 如果模型需要改进，则在仔细考虑训练数据集后，训练作业将使用更多数据重新运行。如果模型没有显示出性能的增量改进，也可以考虑修改训练超参数。
8. 在模型评估符合业务利益相关者定义的标准后，您可以将经过微调的模型托管到 Amazon Bedrock 预配置的吞吐量。

第 2 步：预测患者对处方药或治疗的行为

Fine-tuned LLMs 可以处理临床记录、出院摘要和其他来自临时医学知识图的患者特定文档。他们可以评估患者是否可能服用处方药或治疗。

此步骤使用中创建的知识图谱第 1 步：使用医学知识图谱预测患者预后。知识图谱包含来自患者档案的数据，包括作为节点的患者历史依从性。它还包括药物或治疗不依从性、药物副作用、药物缺乏途径或成本壁垒，或者给药方案复杂等情况，这些都是这些节点的属性。

Fine-tuned LLMs 可以使用医学知识图谱中过去的处方配送数据和亚马逊 OpenSearch 服务矢量数据库中临床记录的描述性摘要。这些临床记录可能会提及经常错过预约或不遵守治疗。法学硕士可以使用这些注释来预测未来不遵守的可能性。

1. 按如下方式准备输入数据：
   - 结构化数据 — 从医学知识图中提取最近的患者数据，例如最近三次就诊和实验室结果。
   - 非结构化数据 — 从 Amazon S OpenSearch ervice 矢量数据库中检索最近的临床记录。
2. 创建包含患者病史和当前背景的输入提示。以下是提示示例：

```
You are a highly specialized AI model trained in healthcare predictive analytics.
 Your task is to analyze a patient's historical medical records, adherence patterns,
 and clinical context to predict the **likelihood of future non-adherence** to
 prescribed medications or treatments.

### **Patient Details**
- **Patient ID:** {patient_id}
- **Age:** {age}
- **Gender:** {gender}
- **Medical Conditions:** {medical_conditions}
- **Current Medications:** {current_medications}
- **Prescribed Treatments:** {prescribed_treatments}

### **Chronological Medical History**
- **Visit Dates & Symptoms:** {visit_dates_symptoms}
- **Diagnoses & Procedures:** {diagnoses_procedures}
- **Prescribed Medications & Treatments:** {medications_treatments}
- **Past Adherence Patterns:** {historical_adherence}
- **Instances of Non-Adherence:** {past_non_adherence}
- **Side Effects Experienced:** {side_effects}
- **Barriers to Adherence (e.g., Cost, Access, Dosing Complexity):** {barriers}

### **Patient-Specific Insights**
- **Clinical Notes & Discharge Summaries:** {clinical_notes}
- **Missed Appointments & Non-Compliance Patterns:** {missed_appointments}

### **Let's think Step-by-Step to predict the patient behaviour**
1. You should first analyze past adherence trends and patterns of non-adherence.
2. Identify potential barriers, such as financial constraints, medication side
 effects, or complex dosing regimens.
3. Thoroughly examine clinical notes and documented patient behaviors that may hint
 at non-adherence.
4. Correlate adherence history with prescribed treatments and patient conditions.
5. Finally predict the likelihood of non-adherence based on these contextual
 insights.

### **Output Format (JSON)**
Return the prediction in the following structured format:
```json
{
  "patient_id": "{patient_id}",
  "likelihood_of_non_adherence": "{low | moderate | high}",
  "reasoning": "{detailed_explanation_based_on_patient_history}"
```
```

```
}
```

3. 将提示传递给经过微调的 LLM。法学硕士处理提示并预测结果。以下是 LLM 的回复示例：

```
{
  "patient_id": "P12345",
  "likelihood_of_non_adherence": "high",
  "reasoning": "The patient has a history of missed appointments, has reported side
 effects to previous medications. Additionally, clinical notes indicate difficulty
 following complex dosing schedules."
}
```

4. 解析模型的响应以提取预测的结果类别。例如，上一步中示例响应的类别可能是不遵守的可能性很高。
5. （可选）使用模型对数或其他方法来分配置信度分数。对数是属于某个类别或类别的项目的非标准化概率。

第 3 步：预测患者再次入院的可能性

由于医疗管理成本高昂以及对患者健康的影响，重新入院是一个主要问题。计算再入院率是衡量患者护理质量和医疗保健提供者绩效的一种方法。

为了计算再入学率，您定义了一个指标，例如 7 天再入学率。该指标是在出院后七天内返回医院进行计划外就诊的住院患者的百分比。为了预测患者再次入院的机会，经过微调的法学硕士可以使用你在中创建的医学知识图谱中的时间数据。第 1 步：使用医学知识图谱预测患者预后该知识图按时间顺序保存了患者遭遇、手术、药物和症状的记录。这些数据记录包含以下内容：

- 自患者上次出院以来的持续时间
- 患者对过去治疗和药物的反应
- 随着时间的推移，症状或病情的进展

您可以处理这些时间序列事件，通过精心策划的系统提示来预测患者再次入院的可能性。该提示将预测逻辑传递给经过微调的 LLM。

1. 按如下方式准备输入数据：
   - 依从性历史记录 — 从医学知识图中提取药物取药日期、药物补充频率、诊断和用药详情、按时间顺序排列的病史以及其他信息。
   - 行为指标 — 检索并包括有关错过预约和患者报告的副作用的临床记录。

2. 创建包含依从历史记录和行为指标的输入提示。以下是提示示例：

```
You are a highly specialized AI model trained in healthcare predictive analytics.
 Your task is to analyze a patient's historical medical records, clinical events, and
 adherence patterns to predict the **likelihood of hospital readmission** within the
 next few days.

### **Patient Details**
- **Patient ID:** {patient_id}
- **Age:** {age}
- **Gender:** {gender}
- **Primary Diagnoses:** {diagnoses}
- **Current Medications:** {current_medications}
- **Prescribed Treatments:** {prescribed_treatments}

### **Chronological Medical History**
- **Recent Hospital Encounters:** {encounters}
- **Time Since Last Discharge:** {time_since_last_discharge}
- **Previous Readmissions:** {past_readmissions}
- **Recent Lab Results & Vital Signs:** {recent_lab_results}
- **Procedures Performed:** {procedures_performed}
- **Prescribed Medications & Treatments:** {medications_treatments}
- **Past Adherence Patterns:** {historical_adherence}
- **Instances of Non-Adherence:** {past_non_adherence}

### **Patient-Specific Insights**
- **Clinical Notes & Discharge Summaries:** {clinical_notes}
- **Missed Appointments & Non-Compliance Patterns:** {missed_appointments}
- **Patient-Reported Side Effects & Complications:** {side_effects}

### **Reasoning Process – You have to analyze this use case step-by-step.**
1. First assess **time since last discharge** and whether recent hospital encounters
 suggest a pattern of frequent readmissions.
2. Second examine **recent lab results, vital signs, and procedures performed** to
 identify clinical deterioration.
3. Third analyze **adherence history**, checking if past non-adherence to medications
 or treatments correlates with readmissions.
4. Then identify **missed appointments, self-reported side effects, or symptoms
 worsening** from clinical notes.
5. Finally predict the **likelihood of readmission** based on these contextual
 insights.

### **Output Format (JSON)**
```

```
Return the prediction in the following structured format:
```json
{
  "patient_id": "{patient_id}",
  "likelihood_of_readmission": "{low | moderate | high}",
  "reasoning": "{detailed_explanation_based_on_patient_history}"
}
```

3. 将提示传递给经过微调的 LLM。法学硕士处理提示并预测重新录取的可能性和原因。以下是 LLM 的回复示例：

```
{
  "patient_id": "P67890",
  "likelihood_of_readmission": "high",
  "reasoning": "The patient was discharged only 5 days ago, has a history of more
 than two readmissions to hospitals where the patient received treatment. Recent
 lab results indicate abnormal kidney function and high liver enzymes. These factors
 suggest a medium risk of readmission."
}
```

4. 将预测归类为标准化尺度，例如低、中或高。
5. 查看法学硕士学位提供的推理，并确定有助于预测的关键因素。
6. 将定性输出映射到定量分数。例如，非常高可能等于 0.9 的概率。
7. 使用验证数据集根据实际再入学率校准模型输出。

第 4 步：计算再入院倾向分数

接下来，计算每位患者的再入院倾向分数。该分数反映了在前面步骤中进行的三项分析的净影响：潜在的患者预后、患者对药物和治疗的行为以及患者再次入院的可能性。通过将患者层面的再入院倾向分数汇总到专业级别，然后汇总到医院层面，您可以获得临床医生、护理经理和管理人员的见解。再入院倾向评分可帮助您按机构、专业或病情评估整体表现。然后，你可以使用这个分数来实施主动干预。

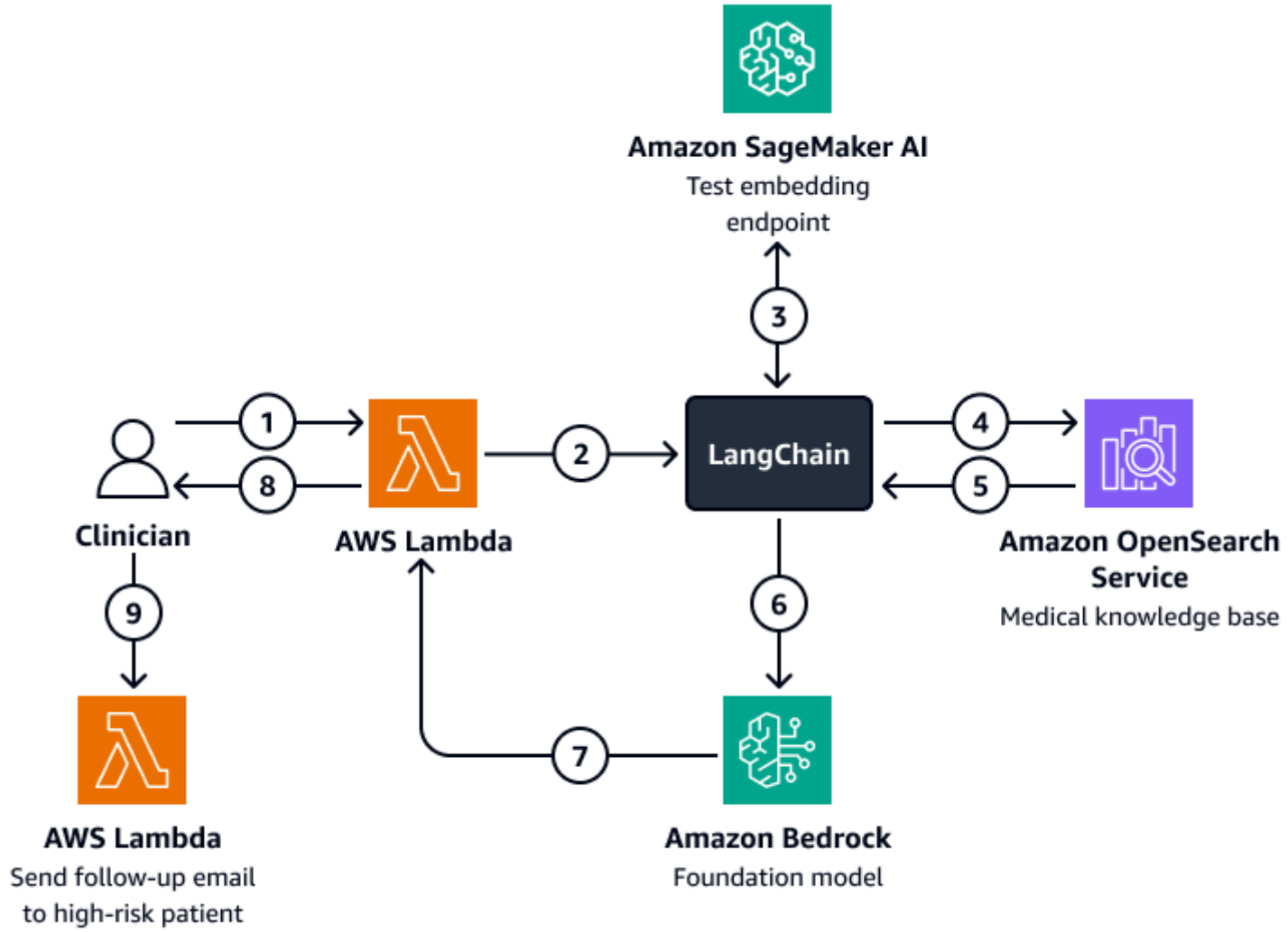
1. 为每个不同的因素（结果预测、依从可能性、重新入院）分配权重。以下是权重示例：
   - 结果预测权重：0.4
   - 依从性预测权重：0.3
   - 再入院可能性权重：0.3
2. 使用以下计算方法计算综合分数：
```

ReadadmissionPropensityScore = (OutcomeScore × OutcomeWeight) +
(AdherenceScore × AdherenceWeight) +
(ReadmissionLikelihoodScore × ReadmissionLikelihoodWeight)

3. 确保所有个人分数都采用相同的等级，例如 0 到 1。
4. 定义操作阈值。例如，分数高于 0.7 会启动警报。

根据上述分析和患者的再入院倾向评分，临床医生或护理经理可以设置警报，根据计算得出的分数对个别患者进行监测。如果超过预定义的阈值，则在达到该阈值时会通知他们。这有助于护理管理人员在为患者制定出院护理计划时积极主动而不是被动。以索引形式将患者的预后、行为和再入院倾向分数保存在 Amazon S OpenSearch ervice 矢量数据库中，以便护理经理可以使用对话式 AI 代理无缝检索这些分数。

下图显示了对话式 AI 代理的工作流程，临床医生或护理经理可以使用该代理来检索有关患者预后、预期行为和再入院倾向的见解。用户可以在患者层面、部门层面或医院层面检索见解。AI 代理会检索这些见解，这些见解以索引形式存储在 Amazon Serv OpenSearch ice 矢量数据库中。该代理使用查询来检索相关数据，并提供量身定制的响应，包括为再次入院风险高的患者建议的措施。根据风险程度，代理人还可以为患者和护理人员设置提醒。



图表显示了以下工作流：

1. 临床医生向对话式 AI 代理提出问题，该代理包含一个 AWS Lambda 功能。
2. Lambda 函数启动一个 LangChain 代理人。
3. 这些区域有：LangChain 代理将用户的问题发送到 Amazon A SageMaker I 文本嵌入端点。端点嵌入了问题。
4. 这些区域有：LangChain 代理将嵌入式问题传递到 Amazon OpenSearch 服务中的医学知识库。
5. Amazon Ser OpenSearch vice 会将与用户查询最相关的具体见解返回给 LangChain 代理人。
6. 这些区域有：LangChain 代理将查询和检索到的上下文从知识库发送到 Amazon Bedrock 基础模型。
7. Amazon Bedrock 基础模型生成响应并将其发送到 Lambda 函数。
8. Lambda 函数将响应返回给临床医生。

9. 临床医生启动 Lambda 函数，向再次入院风险高的患者发送一封后续电子邮件。

与 Well-Architect AWS ed 框架保持一致

用于跟踪患者行为和预测医院再入院率的架构整合了医学知识图表 AWS 服务，并在 LLMs 与 Well-Arch AWS itected Framework 的六大支柱保持一致的同时改善医疗结果：

- 卓越运营 — 该解决方案是一个独立的自动化系统，它使用 Amazon Bedrock 并发出实时 AWS Lambda 警报。
- 安全 — 此解决方案旨在遵守医疗保健法规，例如 HIPAA。您还可以实施加密、精细访问控制和 Amazon Bedrock 护栏，以帮助保护患者数据。
- 可靠性-该架构使用容错、无服务器。 AWS 服务
- 性能效率 — Amazon Service 和经过微调的 OpenSearch 服务 LLMs 可以提供快速而准确的预测。
- 成本优化 — 无服务器技术和 pay-per-inference模型有助于最大限度地降低成本。尽管使用微调的 LLM 可能会产生额外费用，但该模型使用 RAG 方法，可以减少微调过程所需的数据和计算时间。
- 可持续性 — 该架构通过使用无服务器基础架构，最大限度地减少了资源消耗。它还支持高效、可扩展的医疗保健运营。

用例：管理和提高医护人员的技能

实施人才转型和技能提升策略可以帮助员工保持在医疗和医疗保健服务中使用新技术和实践的能力。积极的技能提升计划可确保医疗保健专业人员能够提供高质量的患者护理，优化运营效率并遵守监管标准。此外，人才转型促进了持续学习的文化。这对于适应不断变化的医疗保健格局和应对新出现的公共卫生挑战至关重要。传统的培训方法，例如课堂培训和静态学习模块，可为广大受众提供统一的内容。他们通常缺乏个性化的学习路径，而这对于满足个别从业者的特定需求和熟练程度至关重要。这种one-size-fits-all策略可能导致脱离接触和知识保留率不理想。

因此，医疗保健组织必须采用创新、可扩展和技术驱动的解决方案，以确定每位员工在当前状态和潜在的未来状态下的差距。这些解决方案应推荐高度个性化的学习途径和正确的学习内容集。这有效地让员工为医疗保健的未来做好了准备。

在医疗保健行业，你可以应用生成式人工智能来帮助你了解和提高员工的技能。通过大型语言模型 (LLMs) 和高级检索器的连接，组织可以了解他们目前拥有的技能，并确定将来可能需要的关键技能。这些信息可帮助您通过雇用新员工和提高现有员工的技能来弥合差距。使用 Amazon Bedrock 和知识图表，医疗保健组织可以开发特定领域的应用程序，以促进持续学习和技能发展。

此解决方案提供的知识可帮助您有效地管理人才、优化员工绩效、推动组织成功、识别现有技能和制定人才战略。此解决方案可以帮助您在几周而不是几个月内完成这些任务。

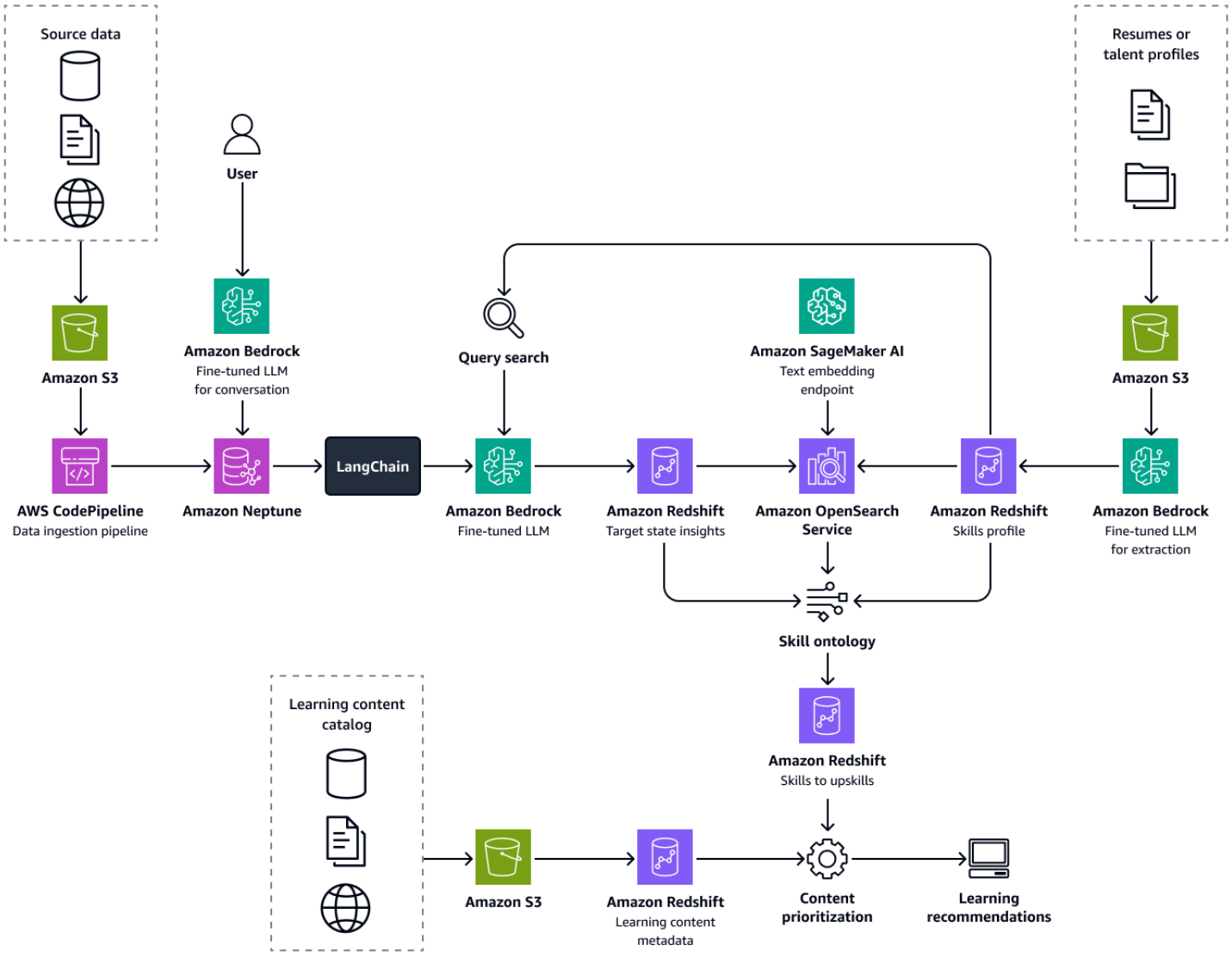
解决方案概述

该解决方案是一个医疗保健人才转型框架，由以下部分组成：

- 智能简历解析器 — 该组件可以读取候选人的简历并精确提取候选人信息，包括技能。智能信息提取解决方案使用 Amazon Bedrock 中经过微调的 Llama 2 模型构建，其专有培训数据集涵盖了 19 多个行业的简历和人才概况。这个基于 LLM 的流程通过自动执行简历的人工审核流程并将最佳候选人与空缺职位进行匹配，节省了数百个小时。
- 知识图表 — 建立在 Amazon Neptune 之上的知识图表，Amazon Neptune 是一个统一的人才信息存储库，包括组织和行业的角色和技能分类，使用技能、角色及其属性、关系和逻辑约束的定义来捕捉医疗保健人才的语义。
- 技能本体论 — 通过本体论算法发现候选人技能与理想的当前状态或未来状态技能（使用知识图进行检索）之间的技能接近性，该算法衡量候选人技能和目标状态技能之间的语义相似性。
- 学习途径和内容 — 此组件是一个学习推荐引擎，可以根据已确定的技能差距，从任何供应商提供的学习材料目录中推荐正确的学习内容。通过分析技能差距并推荐优先的学习内容，为每位候选人确定最佳的技能提升途径，从而使每位候选人在向新职位过渡期间能够实现无缝和持续的专业发展。

这种基于云的自动化解决方案由机器学习服务 LLMs、知识图和检索增强生成 (RAG) 提供支持。它可以扩展到在最短的时间内处理成千上万份简历，创建即时候选人档案，确定他们当前或潜在的未来状态中的差距，然后有效地推荐正确的学习内容来缩小这些差距。

下图显示了框架的 end-to-end流程。该解决方案建立在 Amazon Bedrock LLMs 中经过微调的基础上。它们从 Amazon Neptune 的医疗保健人才知识库中 LLMs 检索数据。数据驱动的算法为每位候选人提供最佳学习途径的建议。



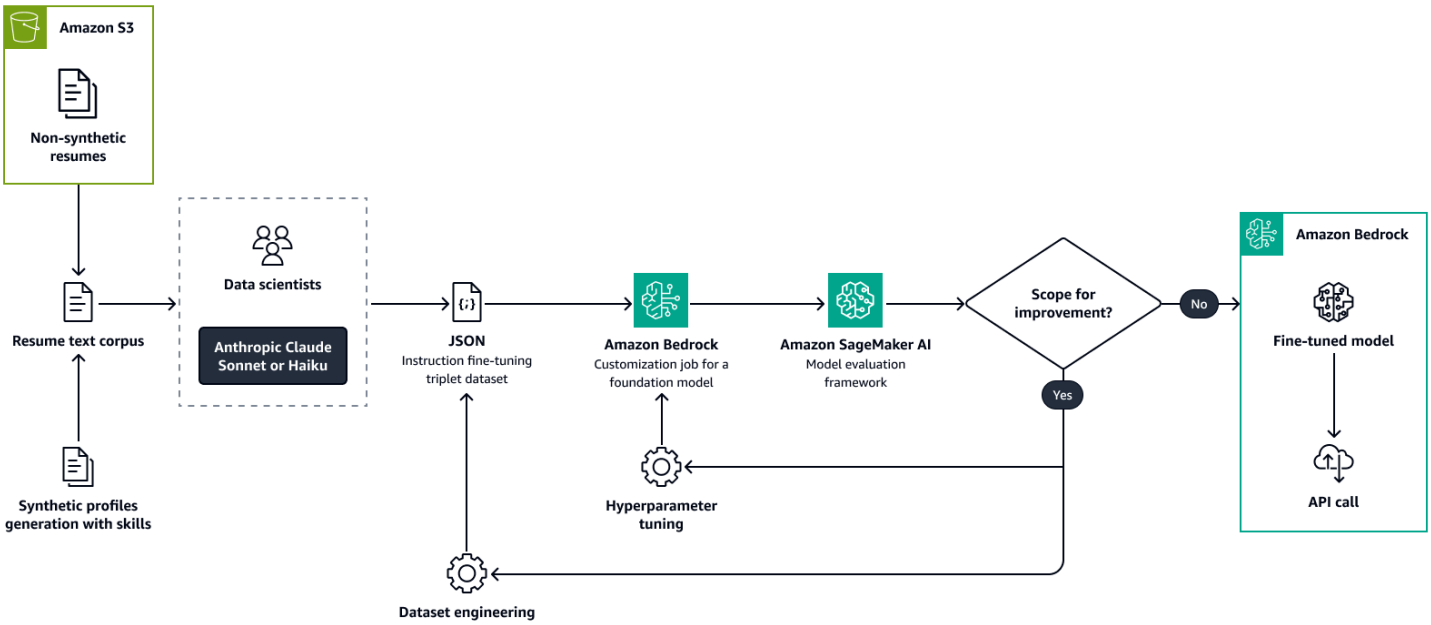
构建此解决方案包括以下步骤：

- 第 1 步：提取人才信息并建立技能档案
- 第 2 步：从知识图谱中发现 role-to-skill相关性
- 第 3 步：找出技能差距并推荐培训

第 1 步：提取人才信息并建立技能档案

首先，您可以使用自定义数据集对 Amazon Bedrock 中的大型语言模型（例如 Llama 2）进行微调。这会根据用例调整法学硕士。在培训期间，您可以准确、一致地从候选人简历或类似人才档案中提取关键人才属性。这些天赋属性包括技能、当前职称、带日期跨度的经验头衔、教育和认证。有关更多信息，请参阅 Amazon Bedrock 文档中的自定义模型以提高其针对您的用例的性能。

下图显示了使用 Amazon Bedrock 微调简历解析模型的过程。真实和综合创建的简历都将传递给法学硕士，以提取关键信息。一组数据科学家根据原始原始文本验证提取的信息。然后，通过使用chain-of-thought提示和原始文本将提取的信息串联起来，以得出训练数据集进行微调。然后，该数据集将传递给 Amazon Bedrock 自定义任务，该任务会对模型进行微调。Amazon SageMaker AI 批处理作业运行模型评估框架，用于评估经过微调的模型。如果模型需要改进，则使用更多数据或不同的超参数再次运行作业。评估达到标准后，您可以通过 Amazon Bedrock 预配置的吞吐量托管自定义模型。



第 2 步：从知识图谱中发现 role-to-skill相关性

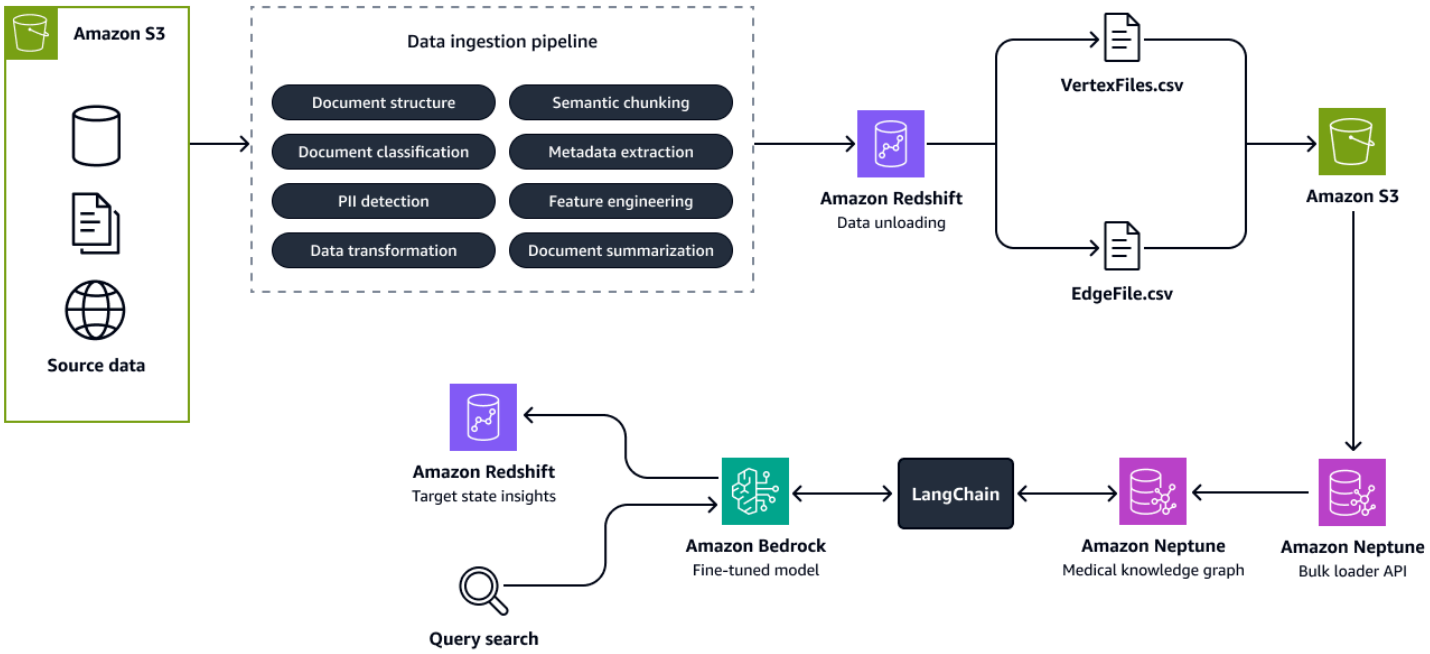
接下来，您将创建一个知识图表，其中概述了您的组织和医疗保健行业其他组织的技能和角色分类法。这个内容丰富的知识库来自于 Amazon Red shift 中的汇总人才和组织数据。您可以从一系列劳动力市场数据提供商以及组织特定的结构化和非结构化数据源收集人才数据，例如企业资源规划 (ERP) 系统、人力资源信息系统 (HRIS)、员工简历、职位描述和人才架构文档。

在 Amazon Neptune 上构建知识图谱。节点代表技能和角色，边缘代表它们之间的关系。使用元数据丰富此图表，包括组织名称、行业、职系、技能类型、角色类型和行业标签等详细信息。

接下来，您将开发一个图形检索增强生成 (Graph RAG) 应用程序。Graph RAG 是一种 RAG 方法，用于从图形数据库中检索数据。以下是 Graph RAG 应用程序的组件：

- 与 Amazon Bedrock 中的 LLM 集成 — 该应用程序使用 Amazon Bedrock 中的 LLM 来理解自然语言和生成查询。用户可以使用自然语言与系统进行交互。这使得非技术利益相关者可以访问它。
- 编排和信息检索 — 使用或 LlamaIndexLangChain协调员，以促进法学硕士和海王星知识图谱之间的整合。他们管理将自然语言查询转换为 OpenCyph er 查询的过程。然后，他们在知识图上运行查询。使用提示工程来指导 LLM 了解构建 OpenCypher 查询的最佳实践。这有助于优化查询以检索相关的子图，该子图包含与所查询的角色和技能有关的所有相关实体和关系。
- 洞察生成 — Amazon Bedrock 中的 LLM 处理检索到的图表数据。它可以生成有关当前状态的详细见解，并预测所查询角色和相关技能的未来状态。

下图显示了根据源数据构建知识图谱的步骤。您将结构化和非结构化源数据传递到数据摄取管道。该管道提取信息并将其转换为与 Amazon Neptune 兼容的 CSV 批量加载格式。批量加载器 API 将存储在 Amazon S3 存储桶中的 CSV 文件上传到 Neptune 知识图表。对于与人才未来状态、相关角色或技能相关的用户查询，Amazon Bedrock 中经过微调的 LLM 会通过以下方式与知识图谱进行交互 LangChain 管弦乐师。协调器从知识图中检索相关上下文，并将响应推送到 Amazon Redshift 中的见解表。这些区域有：LangChain 像 Gr aph QAChain 这样的协调器将来自用户的自然语言查询转换为 OpenCypher 查询，以便查询知识图谱。Amazon Bedrock 经过微调的模型会根据检索到的上下文生成响应。



第 3 步：找出技能差距并推荐培训

在此步骤中，您可以准确计算医疗保健专业人员的当前状态与潜在的未来州角色之间的接近程度。为此，您可以通过将个人的技能组合与工作角色进行比较来进行技能亲和力分析。在 A mazon S OpenSearch ervic e 矢量数据库中，您可以存储技能分类信息和技能元数据，例如技能描述、技能类型和技能集群。使用 Amazon Bedrock 嵌入模型，例如 Amazon Titan 文本嵌入模型，将已识别的关键技能嵌入到向量中。通过向量搜索，您可以检索当前状态技能和目标状态技能的描述，并进行本体分析。该分析提供了当前状态和目标状态技能对之间的接近分数。对于每对，您可以使用计算出的本体论分数来确定技能亲和力的差距。然后，你推荐提升技能的最佳途径，候选人在角色过渡期间可以考虑这个路径。

对于每个角色，推荐正确的学习内容以提高技能或重新培养技能都需要一种系统的方法，首先要创建全面的学习内容目录。该目录存储在 Amazon Redshift 数据库中，汇总了来自不同提供商的内容，并包括元数据，例如内容时长、难度级别和学习模式。下一步是提取每篇内容提供的关键技能，然后将其映射到目标角色所需的个人技能。您可以通过技能接近度分析来分析内容提供的覆盖范围，从而实现此映射。该分析评估了内容所教授的技能与该职位所需技能的密切程度。元数据在为每项技能选择最合适的内容方面起着至关重要的作用，可确保学员获得适合其学习需求的量身定制的推荐。 LLMs 在 Amazon Bedrock 中使用可以从内容元数据中提取技能、执行功能工程和验证内容推荐。这提高了技能提升或技能再培训过程中的准确性和相关性。

与 Well-Architect AWS ed 框架保持一致

该解决方案符合 Well-Ar AWS chitected Framework 的所有六大支柱：

- 卓越运营 — 模块化的自动化管道可增强卓越运营。管道的关键组件是分离和自动化的，因此可以更快地更新模型，更轻松地进行监控。此外，自动训练管道支持更快地发布经过微调的模型。
- 安全 — 该解决方案处理敏感和个人身份信息 (PII)，例如简历和人才档案中的数据。在 AWS Identity and Access Management (IAM) 中，实施精细的访问控制策略，并确保只有经过授权的人员才能访问这些数据。
- 可靠性 — 该解决方案使用的诸如 Neptune AWS 服务、Amazon Bedrock 和 OpenSearch 服务等，即使在需求旺盛的情况下也能提供容错能力、高可用性和不间断地访问见解。
- 性能效率 — LLMs 在 Amazon Bedrock 和 S OpenSearch ervice 矢量数据库中进行了微调，旨在快速准确地处理大型数据集，从而提供及时、个性化的学习建议。
- 成本优化 — 该解决方案使用 RAG 方法，可减少对模型进行持续预训练的需求。系统不会反复微调整个模型，而是仅微调特定的流程，例如从简历中提取信息和构造输出。这可以显著节省成本。通过最大限度地减少资源密集型模型训练的频率和规模以及使用 pay-per-use云服务，医疗保健组织可以在保持高性能的同时优化运营成本。

- 可持续性 — 该解决方案使用可扩展的云原生服务，可动态分配计算资源。这减少了能源消耗和对环境的影响，同时仍然支持大规模的数据密集型人才转型计划。

为医疗保健开发和协调生成式 AI 解决方案

要构建本指南中的解决方案，您必须构建一个 RAG 架构，该架构使用微调功能 LLMs 为医疗保健提供者提供增强的患者数据、临床和诊断见解以及预测的患者预后。这需要整合多种 AWS 服务 工具来创建有凝聚力和高效的工作流程。本节讨论以下内容：

- Amazon Q 开发者版— 使用 Amazon Q Developer 解决开发过程中的工程问题和代码错误。
- 多毛寻回器 RAG 设计— 设计和实现 RAG 解决方案，使用多个检索器为用户的问题获取正确的医疗背景。
- ReAct 代理人— 实现将推理与动态动作相结合的代理。

Amazon Q 开发者版

在构建生成式 AI 解决方案时，可能很难创建 AI 代理和连接关键服务。但是，Amazon Q De veloper 通过提供对高级生成人工智能助手的访问权限来帮助数据科学家和人工智能工程师。Amazon Q 可以快速准确地解决用户问题和代码错误，这可以帮助您优化 LLM 开发流程。Amazon Q 为开发者创建使用 Amazon Bedrock 基础模型的应用程序提供了显著的优势。它可以简化工作流程并提高代码质量。它可以自动生成 Python 脚本和基础设施即代码 (IaC) 配置，从而显著减少开发时间和工作量。通过高级重构功能，Amazon Q 可以提高代码性能，识别安全漏洞，并确保开发人员遵守最佳实践。此外，它还通过提供情境感知建议和解释来促进初学者的学习和采用，使复杂的编码任务更易于访问和高效。

多毛寻回器 RAG 设计

在生成式 AI 应用程序中，多检索器 RAG 管道可以高效地从多个数据源检索信息，以帮助医疗保健提供者和临床医生回答医疗问题。该管道使用不同类型的检索器从不同的知识库中提取相关数据。每只寻回犬都专门获取特定类型的信息，例如患者病史、诊断见解、临床记录或医学研究和学术论文中的内容。

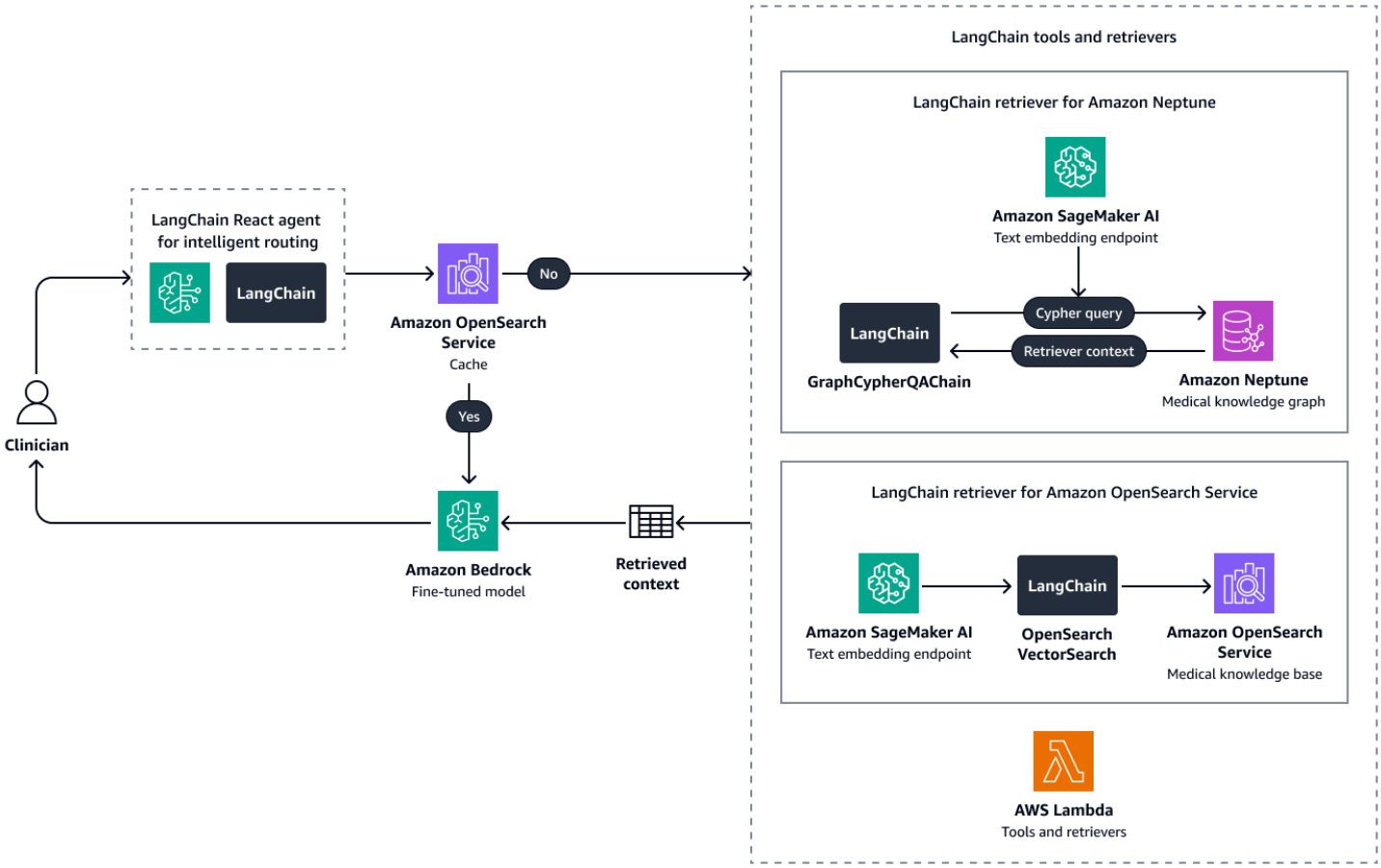
根据数据的性质和特定的应用程序要求来确定哪些正确的后端知识库适合您的用例。Amazon S OpenSearch ervice 矢量数据库非常适合存储大量非结构化或半结构化医疗数据，包括图像诊断评估摘要、出院摘要、临床报告、医学研究和学术文本内容。另一方面，诸如Amazon Neptune之类的图形数据库服务非常适合需要深入探索实体之间时间关系的医疗用例，例如患者、患者病史、医疗保健提供者、药物、症状和治疗。

该管道的一个关键组成部分是用户查询意图预测。这样可以确保系统将查询路由到正确的检索器链。例如，如果临床医生询问患者的治疗史、症状、与医院的互动、再次入院的可能性或潜在的患者预后，则

查询意图预测模块会识别出这种意图。它将请求定向到检索器链，该链可以从医学知识图中获取患者记录或按时间顺序排列的治疗数据。或者，如果问题涉及疾病发现、特定的诊断评估或学术教科书中特定临床程序的细节，则将查询路由到检索器链，该链可以从 OpenSearch 服务向量数据库中获取这些信息。您可以使用的工具调用功能 LangChain 将自定义工具绑定到 Amazon Bedrock LLM，该工具可以将用户问题归类为预定义的意图。

这个多回收器 RAG 系统包括 LangChain 专为管理对特定知识库的访问而设计的代理。您可以使用 … LangChain 编排 Amazon Bedrock LLM、不同的检索器和工具之间的互动。LangChain 包括一个工具调用类，可帮助您创建自定义工具，例如意图分类器、Neptune 的检索器、Service 的检索器或任何其他可以开发的 OpenSearch 用于对用户意图进行分类和以结构化格式访问特定知识库中的数据的工具。然后，你将这些工具提供给全班以创建推理和行动 (ReAct) 代理。 ReAct 代理处理用户问题，计划回答问题的顺序步骤，然后迭代执行可用工具并处理工具响应以最终回答用户查询。

下图显示了专为高效知识检索和智能查询解析而设计的多检索器 RAG 系统的工作原理。A LangChain ReAct 代理分析用户的意图，制定结构化的执行计划，并选择最相关的检索工具。系统会查询先前的问题缓存，并根据关键属性（例如患者 ID、医疗状况和就诊日期）检查是否存在类似的查询。如果找到高度相似的问题，则直接检索相应的答案。否则，代理会执行相应的检索器。为了检索以患者为中心的信息，例如治疗史、症状、医院互动或再次入院的可能性，该系统使用图表检索器。对于诊断评估、临床程序和结构化医学发现，该代理使用矢量数据库检索器。在需要将来自两个数据存储的上下文知识组合起来才能生成全面响应的场景中，系统使用混合检索策略，该策略将知识图谱和矢量数据库的结果整合在一起。



ReAct 代理人

推理和行动 (ReAct) 代理专为多方面的 RAG 应用程序而设计。这些代理提供了推理和动态操作的强大组合，特别是对于涉及 step-by-step逻辑信息检索工作流程的复杂应用程序。有关更多信息，请参阅 ReAct：在语言模型中协同推理和行动。

在医疗和保健领域，来自临床医生或医生的询问通常是多方面的。例如，临床医生可能会问 “对同时患有高血压和2型糖尿病的类似患者进行了哪些治疗？” 在确定了用户的意图（即获取高血压和 2 型糖尿病的治疗方法）之后，AI 代理需要将此查询分为子任务，然后选择最有效的检索策略。在这种情况下，AI 代理应确定最相关的节点（例如患者年龄、性别、病情、治疗和药物），然后在图表中查询这些实体及其属性和关系。 ReAct 代理非常有用，因为它们将 LLM 的推理（逻辑推断）能力与操作（查询外部资源或知识库或与之交互）相结合。

回答用户询问 “对同时患有高血压和2型糖尿病的类似患者进行了哪些治疗？”，以下示例说明了 ReAct 代理的工作原理：

1. 代理推理 — ReAct 代理推断问题涉及检索有关疾病（糖尿病和高血压）的信息。它考虑了患者的年龄、治疗方法、药物和分析期。

2. 代理操作 — 代理使用 OpenCypher 在知识图中查询 2 型糖尿病和高血压特有的治疗方法。它还检索给药的药物、就诊日期、药物的副作用、已知的患者预后以及相似患者（例如相同性别和年龄的患者）的交叉参考数据。
3. 药物@@ 观察 — 从知识图中，该药物检索了最近六个月有关同时患有高血压和2型糖尿病的患者所接受治疗的表格数据。
4. 药物@@ 推理 — 为了对检索到的记录的结果进行排名，代理人可以识别重要属性，例如最近程度、药物的副作用或已知的患者预后。
5. 代理操作-代理根据已识别的属性和通过系统提示传递的预定义逻辑对记录进行重新排名。
6. 响应生成 — Amazon Bedrock 中的 LLM 根据 ReAct 代理准备的上下文生成响应。

评估医疗保健行业的生成式 AI 解决方案

评估您构建的医疗保健 AI 解决方案对于确保它们在现实医疗环境中有效、可靠和可扩展至关重要。使用系统的方法来评估解决方案中每个组件的性能。以下是可用于评估解决方案的方法和指标的摘要。

主题

评估信息的提取

评估信息提取解决方案（例如智能简历解析器和自定义实体提取器）的性能。您可以使用测试数据集来衡量这些解决方案的响应是否一致。如果您没有涵盖多功能医疗保健人才档案和患者病历的数据集，则可以使用法学硕士的推理功能创建自定义测试数据集。例如，您可以使用大型参数模型，例如 Anthropic Claude 模型，以生成测试数据集。

以下是可用于评估信息提取模型的三个关键指标：

- 准确性和完整性 — 这些指标评估输出在多大程度上捕获了地面实况数据中存在的正确和完整的信息。这包括检查提取信息的正确性以及提取的信息中是否存在所有相关细节。
- 相似性和相关性 — 这些指标评估输出和实况数据之间的语义、结构和上下文相似性（相似性），以及输出与地面真相数据的内容、上下文和意图（相关性）一致和解决的程度。
- 调整后的召回率或捕获率 — 这些速率根据经验决定了模型正确识别了地面实况数据中有多少当前值。该费率应包括对模型提取的所有错误值的惩罚。
- 精度分数 — 精度分数可帮助您确定预测中存在多少误报，与真阳性相比有多少误报。例如，您可以使用精度指标来衡量提取的技能熟练度的正确性。

使用多个检索器评估 RAG 解决方案

要评估系统检索相关信息的效果以及它如何有效地使用这些信息生成准确且符合上下文的响应，您可以使用以下指标：

- 响应相关性-衡量生成的响应（使用检索到的上下文）与原始查询的相关性。

- 上下文精度-在检索到的总结果中，评估检索到的与查询相关的文档或片段的比例。上下文精度越高，表明检索机制在选择相关信息方面是有效的。
- 忠诚度-评估生成的响应在检索到的上下文中反映信息的准确程度。换句话说，衡量回复是否符合来源信息。

使用 LLM 评估解决方案

您可以使用一种名为 LLM- 的技术as-a-judge来评估生成式 AI 解决方案中的文本响应。它涉及使用 LLMs 来评估和评估模型输出的性能。该技术利用 Amazon Bedrock 的功能来判断各种属性，例如响应质量、连贯性、依从性、准确性以及对人类偏好或实况数据的完整性。您可以使用 chain-of-thought (CoT) 和少量提示技术进行全面评估。提示指示法学硕士使用评分量规评估生成的响应，提示中的少量样本演示了实际的评估过程。该提示还包括法学硕士评估人员应遵循的指导方针。例如，您可以考虑使用以下一种或多种评估技术，这些技术使用 LLM 来判断生成的响应：

- 成对比 ——向法学硕士评估人员提供一个医学问题以及由你创建的不同迭代版本的 RAG 系统生成的多个答案。提示法学硕士评估人员根据回答质量、连贯性和对原始问题的遵守程度来确定最佳答案。
- 单答分级 — 此技术非常适合需要评估分类准确性的用例，例如患者预后分类、患者行为分类、患者重新入院可能性和风险分类。使用法学硕士评估器单独分析个人分类或分类，并根据事实数据评估其提供的推理。
- 参考文献指导评分 — 为法学硕士评估人员提供一系列需要描述性答案的医学问题。为这些问题创建示例答案，例如参考答案或理想答案。提示法学硕士评估员将法学硕士生成的响应与参考答案或理想答案进行比较，并提示法学硕士评估员根据准确性、完整性、相似性、相关性或其他属性对生成的响应进行评分。此技术可帮助您评估生成的响应是否与定义明确的标准答案或示例性答案一致。

资源

AWS 文档

- 亚马逊 Bedrock 文档
- 亚马逊 Neptune 文档
- 亚马逊 OpenSearch 服务文档
- 为亚马逊 Neptune 应用 AWS Well-Architected 框架（规范性指南）AWS
- Amazon OpenSearch 服务最佳运营实践（OpenSearch 服务文档）
- 使用亚马逊 Comprehend Medical LLMs 以及用于医疗保健和生命科学（规范性指导）AWS

AWS 博客文章

- 使用 Amazon Bedrock 中提供的全新 Amazon Titan Text Premier 模型构建基于 RAG 和代理的生成式 AI 应用程序
- 使用 Amazon Neptune 从数据仓库中构建知识图谱来补充商业情报
- 使用知识图使用亚马逊 Bedrock 和 Amazon Neptune 构建 GraphRag 应用程序

其他资源

- 在肾脏病学中将检索增强生成与大型语言模型相结合：推进实际应用（PubMed 中央，国立医学图书馆）
- 简介 LangChain (LangChain 文档）

贡献者

编写

- Nitu Nivedita，埃森哲数据与人工智能董事总经理——人工智能主管
- Manoj Appully，Cadiem 创始人兼首席技术官
- Conor Folan，埃森哲数据与人工智能顾问
- Deepak Krishna AR，埃森哲数据与人工智能顾问
- Almore Cato，埃森哲数据与人工智能经理
- Soonam Kurian，首席解决方案架构师 AWS

正在审阅

- Sally Lin，埃森哲数据科学高级经理 — 数据与人工智能
- Terry Huang，埃森哲数据科学经理 — 数据与人工智能
- 威廉·洛伦兹，合作伙伴解决方案架构师， AWS

技术写作

- Lilly AbouHarb，高级技术撰稿人， AWS

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	—	2025 年 3 月 14 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：将您的本地 Oracle 数据库迁移到兼容 Amazon Aurora PostgreSQL 的版本。
- 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：在中将您的本地 Oracle 数据库迁移到适用于 Oracle 的亚马逊关系数据库服务 (Amazon RDS) AWS Cloud。
- 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：将您的客户关系管理 (CRM) 系统迁移到 Salesforce.com。
- 更换主机（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：在中的 EC2 实例上将您的本地 Oracle 数据库迁移到 Oracle AWS Cloud。
- 重新定位（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。您可以将服务器从本地平台迁移到同一平台的云服务。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

参见托管服务。

ACID

参见原子性、一致性、隔离性、耐久性。

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。与主动-被动迁移相比，它更灵活，但需要更多的工作。

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

一个 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

参见人工智能。

AIOps

参见人工智能操作。

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

坏机器人

旨在破坏个人或组织或对其造成伤害的机器人。

BCP

参见业务连续性计划。

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。另请参见字节顺序。

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的应用程序版本（蓝色），在另一个环境中运行新的应用程序版本（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互动的软件应用程序。有些机器人是有用或有益的，例如在互联网上索引信息的网络爬虫。其他一些被称为恶意机器人的机器人旨在破坏个人或组织或对其造成伤害。

僵尸网络

被恶意软件感染并受单方（称为机器人牧民或机器人操作员）控制的机器人网络。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）。

破碎的玻璃通道

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 Well -Architected 指南中的 “实施破碎玻璃程序” 指示 AWS 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

参见AWS 云采用框架。

金丝雀部署

向最终用户缓慢而渐进地发布版本。当你有信心时，你可以部署新版本并全部替换当前版本。

CCoE

参见云卓越中心。

CDC

请参阅变更数据捕获。

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弹性。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

查看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 Cloud 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。云计算通常与边缘计算技术相关。

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

组织迁移到以下阶段时通常会经历四个阶段 AWS Cloud：

- 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 Cloud 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

参见配置管理数据库。

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 管道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)

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

配置偏差

对于工作负载，配置会从预期状态发生变化。这可能会导致工作负载变得不合规，而且通常是渐进的，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 is commonly described as a pipeline. 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

参见计算机视觉。

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 Cloud 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

参见数据库定义语言。

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。

委托管理员

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。

后

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

参见环境。

侦测性控制

一种安全控制，在事件发生后进行检测、记录日志和发出警报。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

在星型架构中，一种较小的表，其中包含事实表中有关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)

您用来最大限度地减少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

参见数据库操作语言。

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

参见灾难恢复。

漂移检测

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

参见开发价值流映射。

E

EDA

参见探索性数据分析。

EDI

参见电子数据交换。

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。与云计算相比，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)

组织之间自动交换业务文档。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密文的计算过程。

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

参见服务端点。

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

参见企业资源规划。

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

星形架构中的中心表。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。

失败得很快

一种使用频繁和增量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 Cloud，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

参见分支。

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。

少量提示

在要求法学硕士执行类似任务之前，向其提供少量示例，以演示该任务和所需的输出。这种技术是情境学习的应用，模型可以从提示中嵌入的示例（镜头）中学习。对于需要特定格式、推理或领域知识的任务，Few-shot 提示可能非常有效。另请参见零镜头提示。

FGAC

请参阅精细的访问控制。

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，它使用连续的数据复制，通过更改数据捕获在尽可能短的时间内迁移数据，而不是使用分阶段的方法。目标是将停机时间降至最低。

FM

参见基础模型。

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

人工智能模型的子集，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工件，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 AI。

地理封锁

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，而基于主干的工作流程是现代的首选方法。

金色影像

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配置软件，并有助于提高设备制造运营的速度、可扩展性和生产力。

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

一项高级规则，可帮助管理各组织单位的资源、策略和合规性 (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。侦测性防护机制会检测策略违规和合规性问题，并生成警报以进行修复。它们通过使用 AWS Config、、Amazon、 AWS Security Hub GuardDuty AWS Trusted Advisor、Amazon Inspector 和自定义 AWS Lambda 支票来实现。

H

HA

参见高可用性。

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。

抵制数据

从用于训练机器学习模型的数据集中扣留的一部分带有标签的历史数据。通过将模型预测与抵制数据进行比较，您可以使用抵制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

参见基础设施即代码。

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 Cloud 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

参见工业物联网。

不可变的基础架构

一种为生产工作负载部署新基础架构，而不是更新、修补或修改现有基础架构的模型。不可变基础架构本质上比可变基础架构更一致、更可靠、更可预测。有关更多信息，请参阅 Well-Architected Framework 中的使用不可变基础架构 AWS 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

该术语由克劳斯·施瓦布（Klaus Schwab）于2016年推出，指的是通过连接、实时数据、自动化、分析和人工智能/机器学习的进步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置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。

IoT

参见物联网。

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。

大型语言模型 (LLM)

一种基于大量数据进行预训练的深度学习 AI 模型。法学硕士可以执行多项任务，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

见 7 R。

小端序系统

一个先存储最低有效字节的系统。另请参见字节顺序。

LLM

参见大型语言模型。

下层环境

参见环境。

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

参见分支。

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。恶意软件的示例包括病毒、蠕虫、勒索软件、特洛伊木马、间谍软件和键盘记录器。

托管服务

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。亚马逊简单存储服务 (Amazon S3) Service 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

参见迁移加速计划。

机制

一个完整的过程，在此过程中，您可以创建工具，推动工具的采用，然后检查结果以进行调整。机制是一种在运行过程中自我增强和改进的循环。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

参见制造执行系统。

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 DevOps 人员和冲刺专业人员。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： EC2 使用 AWS 应用程序迁移服务重新托管向 Amazon 的迁移。

迁移组合评测（MPA）

一种在线工具，可提供信息，用于验证迁移到的业务案例。 AWS Cloud 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

用于将工作负载迁移到的方法 AWS Cloud。有关更多信息，请参阅此词汇表中的 7 R 条目和动员组织以加快大规模迁移。

ML

参见机器学习。

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，请参阅中的应用程序现代化策略。 AWS Cloud

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，请参阅中的评估应用程序的现代化准备情况 AWS Cloud。

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

参见迁移组合评估。

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？”

可变基础架构

一种用于更新和修改现有生产工作负载基础架构的模型。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

请参阅源站访问控制。

OAI

参见源访问身份。

OCM

参见组织变更管理。

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

参见运营集成。

OLA

参见运营层协议。

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

参见开放流程通信-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)

一份问题清单和相关的最佳实践，可帮助您理解、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 Well-Architecte AWS d Frame work 中的运营准备情况评估 (ORR)。

操作技术 (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

由此创建的跟踪 AWS CloudTrail ，用于记录组织 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

参见运营准备情况审查。

OT

参见运营技术。

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

查看个人身份信息。

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

参见可编程逻辑控制器。

PLM

参见产品生命周期管理。

policy

一个对象，可以在中定义权限（参见基于身份的策略）、指定访问条件（参见基于资源的策略）或定义组织中所有账户的最大权限 AWS Organizations （参见服务控制策略）。

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。有关更多信息，请参阅在微服务中实现数据持久性。

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`或的查询条件`false`，通常位于子`WHERE`句中。

谓词下推

一种数据库查询优化技术，可在传输前筛选查询中的数据。这减少了必须从关系数据库检索和处理的数据量，并提高了查询性能。

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。

通过设计保护隐私

一种在整个开发过程中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

一种安全控制措施，旨在防止部署不合规的资源。这些控件会在资源配置之前对其进行扫描。如果资源与控件不兼容，则不会对其进行配置。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)

在产品的整个生命周期中，从设计、开发和上市，到成长和成熟，再到衰落和移除，对产品进行数据和流程的管理。

生产环境

参见环境。

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。

提示链接

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，以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

一系列步骤，例如指令，用于访问 SQL 关系数据库系统中的数据。

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

参见 “负责任、负责、咨询、知情”（RACI）。

RAG

请参见检索增强生成。

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

参见 “负责任、负责、咨询、知情”（RACI）。

RCAC

请参阅行和列访问控制。

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。

重新架构师

见 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

见 7 R。

区域

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

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

见 7 R。

版本

在部署过程中，推动生产环境变更的行为。

搬迁

见 7 R。

更换平台

见 7 R。

回购

见 7 R。

故障恢复能力

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

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

见 7 R。

退休

见 7 R。

检索增强生成（RAG）

一种生成式人工智能技术，其中法学硕士在生成响应之前引用其训练数据源之外的权威数据源。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据的过程。

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

参见恢复点目标。

RTO

参见恢复时间目标。

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 Management Console 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

参见监督控制和数据采集。

SCP

参见服务控制政策。

secret

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 Secret s Manager 密钥中有什么？ 在 Secrets Manager 文档中。

安全性源于设计

一种在整个开发过程中考虑安全性的系统工程方法。

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。安全控制主要有四种类型：预防性、侦测、响应式和主动式。

安全加固

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

一种预定义和编程的操作，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 实例或轮换证书。

服务器端加密

在目的地对数据进行加密，由接收方 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。

服务级别指示器 (SLI)

对服务性能方面的衡量，例如其错误率、可用性或吞吐量。

服务级别目标 (SLO)

代表服务运行状况的目标指标，由服务级别指标衡量。

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

参见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

参见服务级别协议。

SLI

参见服务级别指标。

SLO

参见服务级别目标。

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，请参阅中的分阶段实现应用程序现代化的方法。 AWS Cloud

恶作剧

参见单点故障。

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交易数据或测量数据，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。

监控和数据采集 (SCADA)

在制造业中，一种使用硬件和软件来监控有形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。

系统提示符

一种向法学硕士提供上下文、说明或指导方针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

tags

键值对，充当用于组织资源的元数据。 AWS 标签可帮助您管理、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

参见环境。

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

参见环境。

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

一个 SQL 函数，用于对一组以某种方式与当前记录相关的行进行计算。窗口函数对于处理任务很有用，例如计算移动平均线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。

蠕虫

参见一次写入，多读。

WQF

参见AWS 工作负载资格框架。

一次写入，多次读取 (WORM)

一种存储模型，它可以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，但他们无法对其进行更改。这种数据存储基础架构被认为是不可变的。

Z

零日漏洞利用

一种利用未修补漏洞的攻击，通常是恶意软件。

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。

零镜头提示

向法学硕士提供执行任务的说明，但没有示例（镜头）可以帮助指导任务。法学硕士必须使用其预先训练的知识来处理任务。零镜头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 few-shot 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。