构建可扩展的 AI 驱动应用程序与云数据库：架构、最佳实践和性能考虑

Santosh Bhupathi Sr. Solutions Architect bhupathi.santosh@gmail.com

Abstract

人工智能驱动的应用程序的快速采用要求高性能、可扩展和高效的云数据库解决方案，因为传统架构往往难以应对需要实时数据访问、向量搜索和低延迟查询的人工智能工作负载。

本文探讨了如何通过利用专门为人工智能设计的技术（如向量数据库（pgvector）、图数据库（AWS Neptune）、NoSQL 存储（Amazon DocumentDB, DynamoDB）以及关系云数据库（Aurora MySQL 和 PostgreSQL））使云原生数据库支持人工智能驱动的应用程序。

它介绍了将人工智能工作负载与云数据库集成的架构模式，包括带有大型语言模型（LLM）的检索增强生成（RAG）[1]、实时数据管道、由 AI 驱动的查询优化以及基于嵌入的搜索。

通过评估性能基准测试、可扩展性考虑因素和成本效益策略来指导设计人工智能支持的应用程序。

来自医疗保健、金融和客户体验等行业的真实案例研究说明了企业如何利用云数据库增强人工智能能力，同时确保符合企业和监管标准的安全性、治理和合规性。

通过对人工智能与云数据库集成的全面分析，本文为研究人员、架构师和企业提供了一份实用指南，帮助他们构建下一代优化性能、可扩展性和云计算环境成本效益的人工智能应用程序。

介绍

随着人工智能的采用加速，企业需要能够高效处理复杂人工智能工作负载的云数据库 [2]。传统数据库通常难以应对由人工智能生成的数据的巨大体量、多样性和速度，因此有必要采用针对人工智能驱动应用程序优化的云原生架构。本文探讨了组织如何利用可扩展、高性能的云数据库增强人工智能能力，利用先进的存储、检索和处理机制。特别是，向量数据库在人工智能应用中发挥着关键作用，通过实现语义搜索、相似性匹配以及高效检索高维数据，显著提升了人工智能模型的性能。我们考察了关键架构、性能考虑因素以及实时数据流和人工智能驱动查询增强等人工智能优化的作用，以提高整体效率。通过将向量能力集成到云数据库中，企业可以简化人工智能工作负载，减少延迟，并实现更智能的数据检索，最终推动各行各业的创新。

增强 AI 应用的检索增强生成（RAG）：解决幻觉问题以实现可靠的 AI 响应

人工智能驱动的应用程序利用先进的机器学习（ML）和深度学习（DL）等人工智能模型来实现流程自动化、增强决策能力和提供个性化体验。这些应用程序广泛应用于医疗保健、金融和电子商务等行业，实现了智能聊天机器人、预测分析和实时推荐等功能。然而，尽管具有这些能力，AI 模型特别是大型语言模型（LLMs）面临着一个被称为幻觉的关键挑战。这种情况发生在模型生成错误的、误导性的或凭空捏造的信息时，而这些信息并未基于事实数据。由于 LLMs 是

根据概率预测单词而不是检索现实世界知识，因此它们有时会产生听起来合理但实际上不准确的回答，在高风险应用中引发了可靠性方面的担忧。

处理幻觉问题的 RAG 方法

为了解决幻觉问题并提高 AI 生成响应的准确性，检索增强生成（RAG）作为一种稳健的解决方案应运而生。RAG 通过在生成响应之前加入外部知识检索过程来提升大规模语言模型的功能。该模型不是仅仅依赖预训练的信息，而是动态地从结构化和非结构化的来源中检索相关数据，如向量数据库、文档仓库和企业知识库。这种方法确保了 AI 响应的事实准确性与语境相关性，减少了错误信息并增加了对 AI 驱动应用程序的信任。通过集成基于检索的机制，RAG 使 AI 系统能够提供可验证的实时洞察，使其在企业和任务关键型用例中更加可靠。

人工智能驱动的检索增强生成工作流

RAG 工作流程确保 AI 生成的响应基于相关检索到的数据，减少幻觉并提高准确性。该工作流程的关键步骤如下：

1. **数据摄入与预处理**系统处理来自各种来源的结构化和非结构化数据，如企业数据库、文档和 API。原始数据被划分为可管理的数据块，确保最佳检索性能。这些文档片段随后通过嵌入模型（例如，Amazon Titan Text Embeddings V2，OpenAI 的嵌入模型）生成向量表示。这些嵌入捕获了文本的语义精髓，使得能够进行有意义的相似性搜索 [3]。将嵌入及其对应的文档片段存储在优化快速和高效检索的向量数据库中。

2. **用户查询处理与语义搜索**当用户提交查询时，该查询会通过相同的嵌入模型进行转换，将其转化为向量表示。然后，向量数据库使用查询嵌入作为搜索向量执行语义相似性搜索。系统通过比较它们的向量接近度检索出最相关的前 k 个文档片段，确保检索到的内容与用户的意图紧密一致。此检索步骤有助于 AI 模型将其响应基于事实和领域特定的信息。

3. **上下文增强与响应生成**检索到的文档片段与用户的原始查询相结合，创建了一个丰富化的提示，确保 AI 模型能够访问最相关的外部知识。这个增强后的提示随后被输入到部署在云 AI 平台（如 Amazon Bedrock）上的基础模型（例如 Claude 3、GPT-4 或 Cohere 模型）中。基础模型综合生成响应，利用其预训练的知识和检索到的数据。这种混合方法产生的响应不仅上下文准确，而且针对特定领域进行了定制。

4. **响应交付与反馈循环** AI 生成的响应通过应用程序界面返回给用户。同时，日志和反馈机制捕获用户交互、响应质量和提供的任何更正。此反馈用于迭代模型改进，提高搜索准确性，并持续优化提示工程技巧。

通过将检索机制与 AI 生成的响应相结合，RAG 工作流增强了基于 AI 的应用程序的可靠性和事实准确性。这种方法在各行各业中广泛应用，从医疗保健和金融到客户服务和知识管理，确保 AI 系统生成的信息丰富、语境丰富的响应，而不仅仅依赖于预训练的知识。

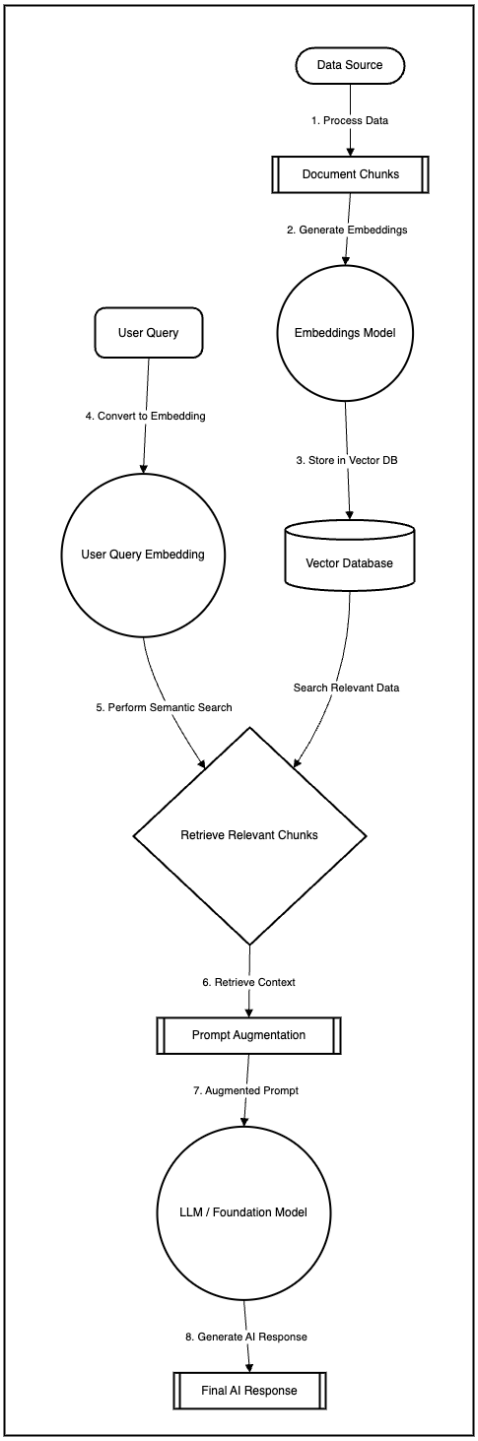

Figure 1: 生成 AI 工作流

理解向量搜索和高效存储向量

什么是向量搜索？向量搜索是一种高级检索技术，它基于语义相似性而非精确关键词匹配来识别数据。与依赖关系型或关键词查询的传统数据库不同，向量搜索利用高维向量表示来揭示数据点之间的模式和关系。这使其在 AI 驱动的应用程序中特别有价值，例如语义搜索、图像识别、推荐系统和自然语言处理（NLP），其中检索出上下文相关的结果比精确字符串匹配更为重要。传

统的基于关键词的搜索难以捕捉词、图像和结构化数据之间更深层次的关系。通过利用向量搜索，AI 驱动的系统可以有效识别概念上的相似性，增强推荐引擎、语义搜索和生成式 AI 等应用，提供更加准确和有意义的结果。

向量数据库如何工作？向量数据库存储和检索高维向量嵌入，这些嵌入捕捉了不同类型数据的语义含义，包括文本、图像和音频。这些数据库高效地对向量进行索引，并使用近似最近邻（ANN）搜索算法，如分层可导航的小世界图（HNSW）、倒排文件索引（IVF）和乘积量化（PQ），以在大规模数据上执行相似性搜索。当用户查询数据库时，系统将输入转换为向量，在向量空间中搜索最近的匹配项，并返回最相关的结果。

什么是向量表示？向量是数据的数值表示，捕捉其在高维空间中的含义。这些向量嵌入使用各种技术生成，包括从文本、图像和其他非结构化数据源中提取语义关系的机器学习模型如 Word2Vec、BERT 和 CLIP。数据哈希方法，如 SimHash 和 MinHash，通过将数据转换为一致的向量表示来提供轻量级替代方案，从而加快相似性计算速度。此外，数据索引技术有助于高效地构建和扩展向量搜索。

让我们创建一个基于文本的示例，捕捉将文本数据转换为向量嵌入的本质，使用短语“编程很有趣”和“调试具有挑战性”。

这里是可视化表示：

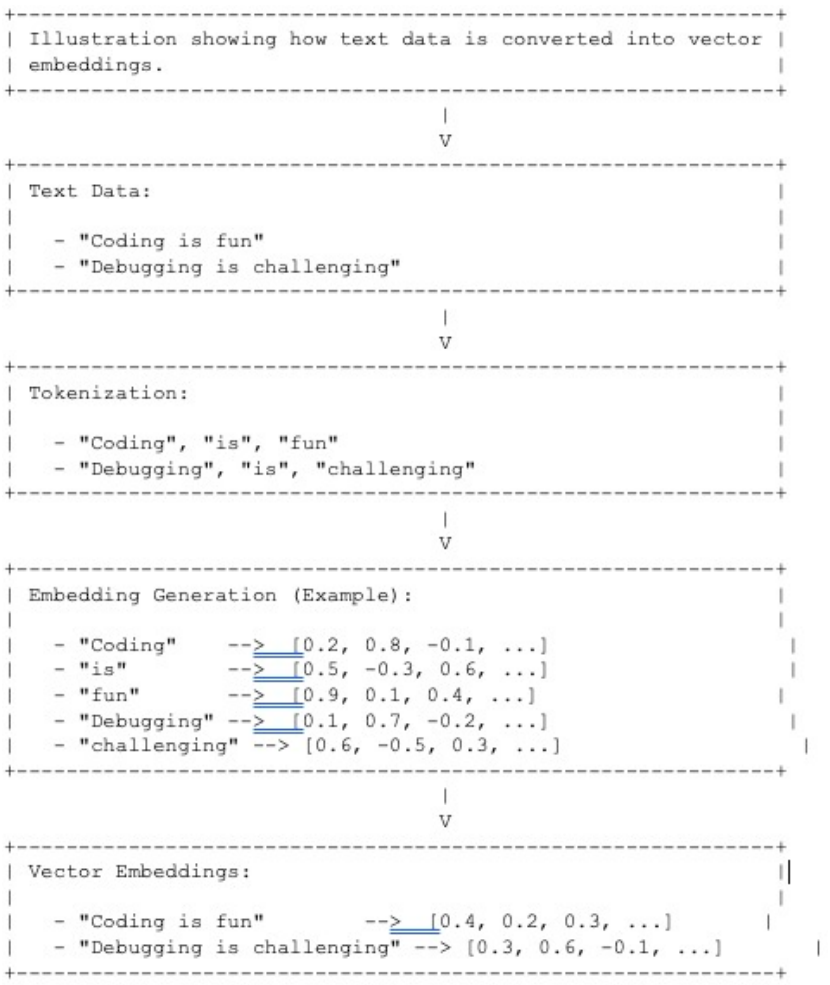

Figure 2: 文本到向量嵌入

- **文本数据：**输入短语，“编程很有趣”和“调试很有挑战性。”
- **分词：**将短语分解成单个词或标记的过程。
- **嵌入生成：**提供了将单词映射到向量的简化示例。[...] 表示在实际场景中这些向量会有更多的维度。
- **向量嵌入：**这显示了最终的输出，其中每个短语都由一个向量表示。再次，[···] 表示这些向量具有更多的维度。

如何生成向量？

向量生成的过程涉及将原始数据（文本、图像、视频等）转换为可以高效比较的数值格式。例如，在基于文本的应用程序中，诸如“牛奶杯”这样的短语会通过嵌入模型如 BERT、Word2Vec 或 Amazon Titan Embeddings 被转化为一个数值向量。这些嵌入捕捉了根据上下文使用情况得出的单词之间的关系。类似地，图像可以通过基于 CNN 的模型（如 CLIP）进行矢量化，这将对象和视觉特征表示为多维向量。可以根据数据类型及其预期用途采用不同的方法生成向量。一些最常见的方法包括：

- **机器学习模型：**诸如 Word2Vec、BERT（用于文本）、CLIP（用于图像和文本）以及 OpenAI 的嵌入模型生成捕捉数据语义意义的向量。这些模型在大型数据集上进行训练，以理解词关联性、上下文和嵌入。
- **数据哈希：**像 SimHash 和 MinHash 这样的技术将数据转换成紧凑的数值表示。哈希通过降低维度同时保留相似性来确保更快的检索。
- **数据索引：**从文本、图像和视频中提取特征，进行归一化处理，并将这些特征组合成向量以实现结构化的存储和检索。
- 通过将多种数据来源向量化，组织可以以统一的格式存储信息，从而使搜索、分析和检索相关信息变得更加容易。

如何高效存储向量？

高效存储向量需要优化的索引、压缩和检索技术。一些最佳的向量存储实践包括：

- **使用向量数据库：**像 FAISS、Milvus、Vespa 和 pgvector（PostgreSQL 扩展）这样的数据库旨在处理具有高效相似性搜索功能的高维向量数据。
- **高效索引：**实现分层可导航小世界（HNSW）图或倒排文件索引（IVF）可以快速进行最近邻搜索，而无需扫描整个数据集。
- **降维：**如主成分分析（PCA）和自动编码器等技术有助于在保留关键信息的同时减少存储占用。
- 混合存储解决方案：结合向量存储与元数据过滤（例如类别、时间戳）确保结果不仅在语义上相关，而且在上下文中也合适。

通过实现这些技术，向量数据库能够支持快速且可扩展的 AI 驱动应用程序，例如语义搜索、推荐引擎和实时聊天机器人，这些应用需要高效检索上下文信息。

云计算驱动的 AI：智能应用的高级数据库技术和架构模式

人工智能驱动的应用程序需要高性能、可扩展和适应性强的数据库解决方案，这些方案能够处理大量结构化和非结构化数据，并与人工智能和机器学习模型无缝集成。现代云数据库已经超越了传统的存储和查询功能，集成了向量搜索、知识图谱、实时分析和语义检索等由 AI 驱动的功能。然而，AI 应用程序不仅仅需要向量能力——它们还需要强大的扩展性、高可用性和高效的写操作来适应不断增长的数据量和实时处理需求。云数据库在这些方面表现出色，提供了弹性、自动扩展和分布式架构，以低延迟处理高吞吐量的工作负载。通过利用增强型人工智能的云数据库，组织可以解锁强大的洞察力，优化决策，并构建能够动态扩展以满足现代 AI 驱动系统不断变化需求的智能应用程序。

构建 AI 驱动应用程序的关键特性

开发人工智能驱动的应用程序需要利用前沿的数据库技术和架构，以增强数据检索、处理和推理能力。以下是能够显著提高人工智能应用程序效率、可扩展性和智能性的几个重要特性。

1. 1 向量能力向量搜索使 AI 应用程序能够执行基于相似性的查询，而不是依赖传统的关键词匹配。这特别适用于推荐系统、图像和视频识别、语义搜索以及自然语言理解。具有向量功能的云数据库（例如 PostgreSQL 的 pgvector、FAISS 和 Milvus）允许 AI 模型通过在高维向量空间中进行搜索来高效地检索相关信息。这些能力增强了检索增强生成 (RAG)，使大型语言模型 (LLM) 能够提供更符合上下文的准确响应。

2. 基于图的检索增强生成（GraphRAG）通过整合图数据库来改进传统的 RAG 模型，以提高知识检索的效果。与仅依赖向量搜索的传统 RAG 不同，GraphRAG 利用实体和概念之间的关系，使其非常适合需要深层次语境理解的应用场景，例如法律研究、生物医学信息检索和企业知识图谱。通过集成像 AWS Neptune 或 Neo4j 这样的图数据库，AI 应用可以根据语义关系检索信息，确保提供更相关且结构化的响应。

3. 实时 AI 推理与流数据处理对于需要即时决策的 AI 应用——如欺诈检测、推荐引擎或预测分析——实时推理和流数据管道至关重要。具备内置流处理能力的云数据库，例如将 Apache Kafka 与 Amazon DynamoDB 或 Amazon Redshift 集成，允许连续数据摄取以及近乎瞬时的 AI 模型执行。这确保了由 AI 驱动的应用程序保持响应性、适应性和应对动态工作负载的能力。

4. AI 驱动的查询优化与自动化索引传统查询优化依赖于基于规则的调整，但由 AI 支持的数据库使用机器学习模型自适应地优化查询、索引和执行计划。由 AI 驱动的查询引擎，如 Amazon Aurora 和 Google BigQuery 中的那些，分析历史查询性能并动态调整索引策略以提高执行速度。这减少了操作开销，同时确保了 AI 工作负载下的最优数据库性能。

5. 混合 AI 搜索与多模态数据集成人工智能应用日益需要在多种模式之间进行搜索，包括文本、图像、音频和视频。混合搜索结合了基于向量的语义搜索和传统的结构化查询，提供了更精

确且上下文相关的结果。多模态数据库，如带有向量搜索的 OpenSearch 或 Azure 认知搜索，使 AI 系统能够高效处理和检索多种数据格式的信息。

通过整合这些功能，AI 驱动的应用程序可以实现更高的准确性、更好的用户体验和改进的可扩展性，同时利用云数据库进行无缝的数据存储、检索和处理。

采用专为 AI 应用设计的数据库：对话、情境和语义上下文

人工智能应用需要超越传统存储和检索能力的数据库来支持实时分析、向量搜索、知识图谱和语义理解。通过采用专门构建的数据库，组织可以优化性能、可扩展性和由 AI 驱动的洞察力，同时减少运营开销。人工智能应用程序主要依赖于对话上下文（聊天历史和交互）、情境上下文（结构化数据如用户配置文件和交易历史）以及语义上下文（用于相似性搜索的高维嵌入）。以下我们探讨各种专门构建的数据库及其在人工智能应用中的集成 [4]。

极光 MySQL 在 AI 应用中的情感分析显著性 Aurora MySQL 提供内置的机器学习（ML）集成，允许应用程序执行诸如情感分析等基于 AI 的任务，而无需单独的 ML 基础设施。这通过使开发人员能够直接从 SQL 查询中运行 ML 模型来简化 AI 的应用，从而减少操作复杂性。

人工智能集成示例——情感分析在 AI 驱动的对话应用程序中，理解用户情感对于提供个性化响应至关重要。通过利用 Amazon Aurora MySQL 内置的 ML 函数，开发人员可以集成预训练的 Amazon Comprehend 模型进行情感分析，而无需构建自定义的 ML 管道。

架构模式与工作流

1. **用户输入处理：**客户与聊天机器人或人工智能支持助手进行互动。
2. **数据存储：**对话历史存储在 Aurora MySQL 中。
3. **情感分析集成：**该 AI 应用程序调用一个 SQL 函数，使用 Amazon Comprehend 分析用户消息的情绪。
4. **响应优化：**基于情感评分（正面、负面、中性），应用程序会动态调整响应以提高客户参与度。

2. 用于 AI 驱动个性化推荐的 Aurora PostgreSQL 显著性 Aurora PostgreSQL，支持 pgvector 扩展，使 AI 应用程序能够高效执行向量搜索。它广泛用于推荐系统、语义搜索以及用于 GenAI 应用的检索增强生成（RAG）模型。

人工智能集成示例——电子商务个性化推荐在电子商务平台上，个性化产品推荐可以显著提高用户参与度和转化率。通过将用户偏好以向量嵌入的形式存储在配备 pgvector 的 Aurora PostgreSQL 中，AI 模型可以根据之前的购买记录和浏览历史检索类似的产品推荐。

架构模式与工作流

1. **用户交互：**一位客户在电子商务网站上搜索一款产品。
2. **向量存储：**产品元数据和用户偏好作为向量嵌入存储在 Aurora PostgreSQL 中。
3. **相似性搜索：**AI 模型使用 pgvector 相似性搜索检索前 K 个最近的产品。
4. **个性化推荐：**系统实时生成人工智能驱动的推荐，提升购物体验。

3. 亚马逊 Neptune 用于图 RAG 和人工智能驱动的知识图谱显著性 Amazon Neptune 提供了支持图数据库功能，这些功能为需要 GraphRAG（基于图的检索增强生成）[5]、欺诈检测、实体关系和语义推理的人工智能应用程序提供动力。它在金融、医疗保健和企业知识图谱等领域特别有用。

人工智能集成示例 – 基于图的 AI 在金融欺诈检测中的应用金融机构可以利用 Neptune 的图形数据库通过分析交易关系来识别欺诈模式。基于图嵌入训练的人工智能模型能够检测异常并标记可疑活动。

架构模式与工作流

1. **数据收集：**金融交易持续存储在海王星中。
2. 图形关系映射：AI 模型构建了一个知识图谱，映射客户行为和实体连接。
3. **人工智能驱动的欺诈检测：**使用 GraphRAG 和 Neptune Analytics，AI 算法可以检测异常情况，例如意外的大额交易或不寻常的账户活动。
4. **告警生成：**如果检测到欺诈行为，系统将标记交易以便进一步审查。

4. 适用于 AI 驱动的向量搜索与知识检索的 Amazon DocumentDB 显著性 Amazon DocumentDB 是一个支持向量搜索的 NoSQL 文档数据库，使其成为基于 AI 的知识检索系统、聊天机器人和智能搜索引擎的理想选择。

人工智能集成示例 – 基于 AI 的医疗诊断搜索在医疗应用中，基于 AI 的临床决策支持系统可以使用 DocumentDB 的向量搜索根据患者症状检索相关的医学案例研究。

架构模式与工作流

1. **医疗数据存储：**电子健康记录（EHRs）存储在 Amazon DocumentDB 中。
2. **向量化：**AI 将医疗案例和症状转换为向量嵌入。
3. **语义搜索：**当医生输入患者症状时，DocumentDB 的向量搜索会检索相关的病例以供诊断。
4. **人工智能驱动的推荐：**系统根据历史患者数据提供临床建议。

5. 亚马逊 MemoryDB 用于低延迟 AI 向量搜索显著性对于需要实时 AI 推理的应用程序，Amazon MemoryDB 为实时聊天机器人、游戏推荐和预测分析等 AI 工作负载提供了具有超低延迟的内存向量搜索解决方案。

人工智能集成示例 – 带有即时知识检索的 AI 聊天机器人 MemoryDB 使 AI 聊天机器人能够即时检索信息，提升对话式 AI 的表现和响应时间。

架构模式与工作流

1. **用户查询：**用户在一个 AI 聊天机器人中提问。
2. **内存数据库中的向量搜索：**查询被转换成向量嵌入，并在 MemoryDB 中搜索最接近的匹配项。
3. **实时响应：**聊天机器人从 MemoryDB 检索相关回复并提供即时答案。
4. **持续学习：**MemoryDB 动态更新其知识库以添加新数据。

结论

人工智能驱动的应用程序的快速发展凸显了对可扩展、高性能和专门构建的云数据库的需求。传统的数据库虽然在结构化数据管理方面有效，但往往难以满足现代 AI 工作负载的要求，后者需要实时处理、向量搜索能力、语义检索以及基于图的知识表示。通过利用专门为如 Aurora PostgreSQL 配以 pgvector 用于向量搜索、Amazon Neptune 用于 GraphRAG 和 Amazon DocumentDB 用于增强文档处理的人工智能而构建的数据库，企业可以优化 AI 应用程序，提高其准确性、效率和可扩展性。本文探讨了将 AI 与云数据库集成的关键架构模式，包括检索增强生成（RAG）、实时数据流以及基于人工智能的查询优化。我们还强调了性能基准测试、成本考虑因素及最佳实践，以确保在云环境中高效部署 AI。随着组织继续通过 AI 进行创新，采用针对特定用例定制的云原生数据库解决方案对于提高运营效率和释放新的由 AI 驱动的能力至关重要。通过将 AI 应用程序与正确的数据库技术相结合，企业可以增强数据驱动决策、改善用户体验，并加速医疗保健、金融和电子商务等行业中的数字转型。

References

[1] M. Lewis, J. Perez, P. Stenetorp, and S. Riedel, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *arXiv preprint*, vol. arXiv:2005.11401, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

[2] C. Guo, J. Sun, W. Chen, Z. Chen, G. Hu, Y. Zhang, and L. Zhang, "Accelerating ai workloads with cloud databases: A survey on cloud-native ai architectures and optimizations," *IEEE Transactions on Cloud Computing*, 2020.

[3] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, 2019. [Online]. Available: https://arxiv.org/abs/1702.08734

[4] Z. Sun, Y. Zhang, J. Zhang, Y. Li, and H. Peng, "Graph neural networks for ai-powered knowledge graphs and their role in ai-driven applications," *ACM Computing Surveys*, 2021.

[1–4]