封面



标题

# Oracle Database 23ai新特性：AI向量搜索

作者：王敏桥

摘要

Oracle Database 23ai的新特性AI向量搜索，允许用户基于数据的语义或含义进行搜索，通过SQL语句来生成、存储、索引和查询向量嵌入以及其他业务数据。该功能支持高性能的向量索引以实现快速的近似搜索，还提供了新的SQL运算符和语法，支持完整的生成式AI管道，包括使用业务数据增强大型语言模型（RAG）。AI向量搜索与业务数据无缝结合，无需将数据移动到单独的向量数据库，并利用Oracle数据库的企业级安全性、可扩展性和分区功能来增强应用程序。AI向量搜索为Oracle数据库带来了新的应用程序开发可能性。

正文

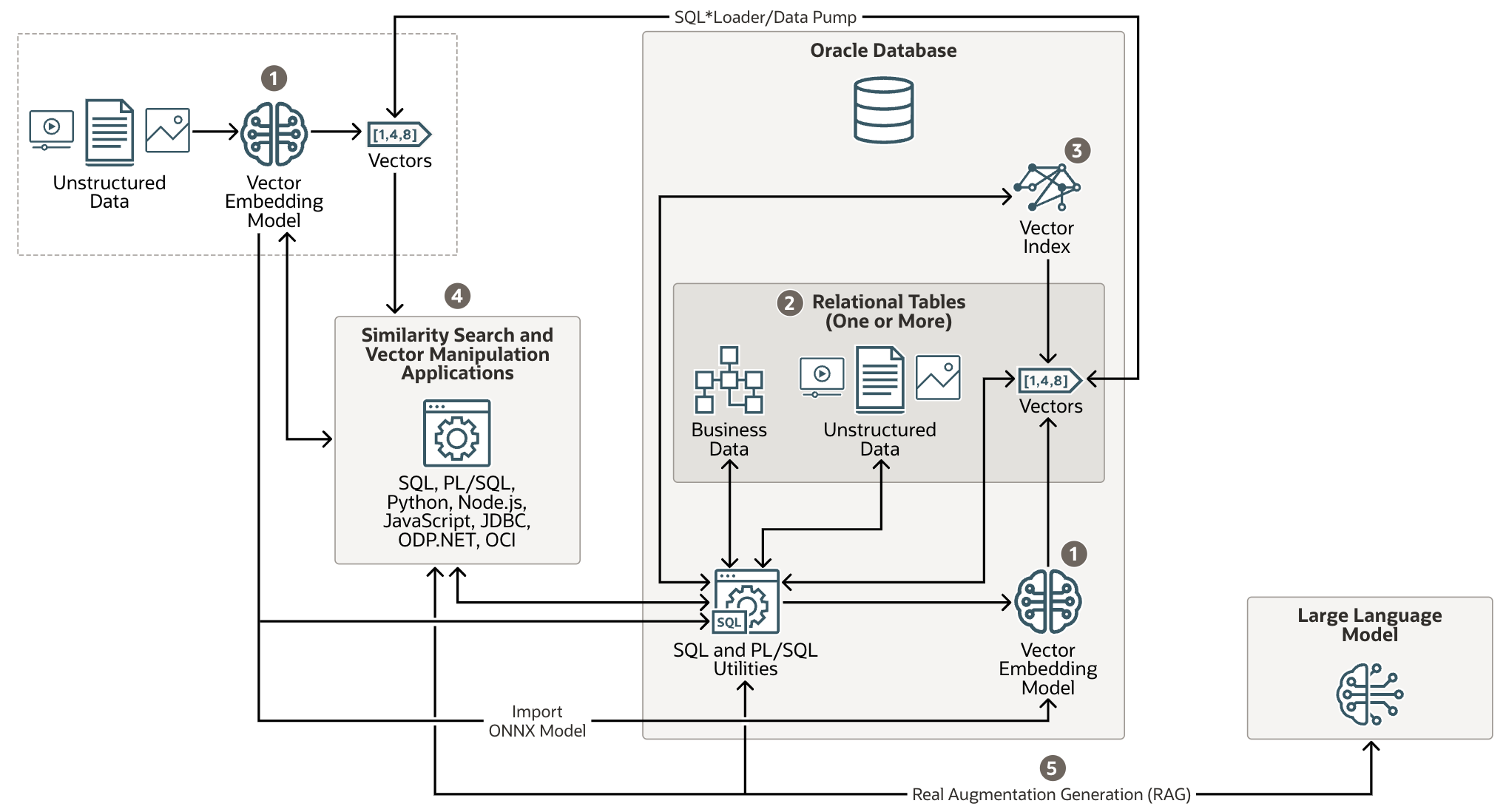
Oracle AI向量搜索（AI Vector Search）是在Oracle Database 23ai中提供的一种新功能，它除了支持数据库传统上的属性值或关键字等数据值搜索之外，允许用户基于数据的语义或含义进行搜索。

向量或向量嵌入是人工智能应用中使用的一种流行的数据结构。向量是由深度学习模型根据不同数据类型（例如图像、文档、视频等）生成的数字列表，对数据的语义进行编码。

Oracle AI向量搜索允许用户利用 SQL 的全部功能来生成、存储、索引和查询向量嵌入以及其他业务数据。在搜索文档时，向量搜索通常比基于关键字的搜索更有效，因为向量搜索基于单词背后的含义和上下文，而不是单词本身。

Oracle AI向量搜索允许将语义文档搜索与结构化文档属性的搜索结合起来。例如，在技术文章数据库中，诸如“查找有关为企业微调大型语言模型（LLM）案例的文章，这些文章在过去 5 年中由某个作者和某个出版商在某些国家/地区出版”，需要搜索文章文本以及文章属性，这些属性可能存在于一个或多个表中。

## Oracle AI向量搜索的技术架构



Oracle AI向量搜索包括一组强大的功能，以支持业务应用场景的语义搜索。这些功能包括：

1. 新的 SQL 嵌入函数可从非结构化数据生成向量嵌入（可以调用外部向量嵌入服务，也可以将向量嵌入模型加载到数据库中）
2. 新的一级**VECTOR数据类型**用于存储向量
3. 新的高性能向量**索引**用于快速近似搜索
4. 新的 SQL 运算符和语法，可轻松表达业务查询中的相似性搜索
5. 支持**完整的生成式人工智能管道**，包括预处理和向量化数据，以及使用业务数据增强LLM

VECTOR 数据类型完全集成在 SQL 和 PL/SQL 中，并支持多种客户端和编程语言，并在 python-oracledb、node-oracledb、JDBC 和 ODP.NET 驱动程序中具有原生绑定功能。这种全面的支持提供了跨多个开发环境的无缝向量搜索功能。

Oracle AI 向量搜索还完全集成到流行的第三方生成式 AI 框架（例如 LangChain）中。

Oracle AI向量搜索包含原生API，可对 LLM API 进行 REST 调用，以生成内容（如文本）或摘要以及其它生成式AI管道中使用的操作。这些新功能可以无缝支持检索增强生成(RAG)，这是一种突破性的生成式AI技术，通过私有业务数据增强大型语言模型(LLM)，来对业务数据上的自然语言问题提供准确的响应。

## Oracle AI向量搜索的优势

Oracle数据库是企业运营和企业数据的主要存储库。企业应用程序通常需要同时搜索业务数据和非结构化数据。例如，零售网站可以基于自然语言产品描述和目标产品图像以及其他过滤条件（例如：价格、商店位置、制造商和当前可用性）进行搜索。这种搜索需要同时搜索非结构化目录数据（产品描述和图像）、结构化目录数据（价格、商店位置和制造商）以及实时交易数据（例如：当前库存）。

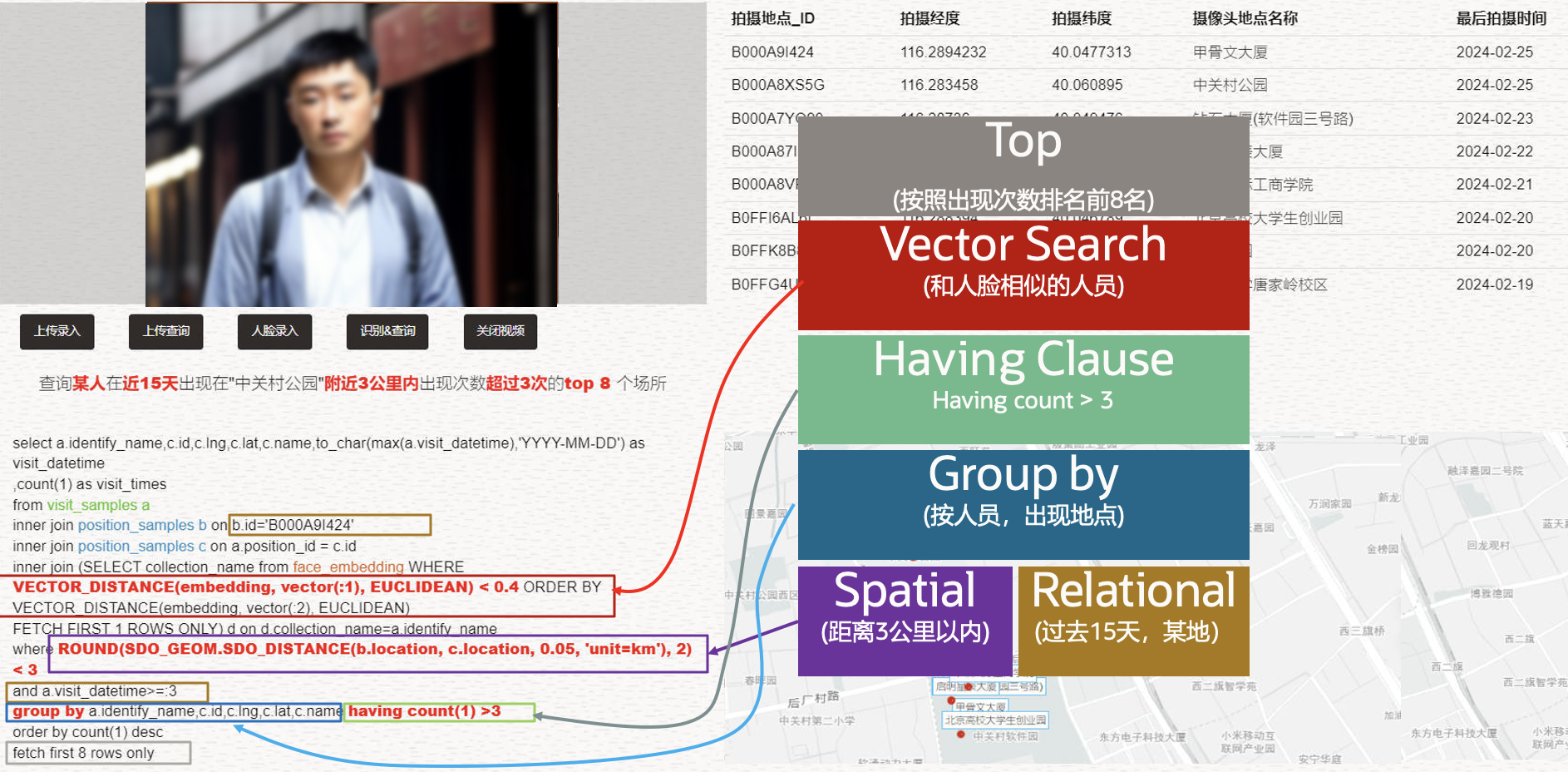
Oracle数据库的融合能力和Oracle AI向量搜索的结合提供了多种独特的优势。

#### 业务数据与AI向量数据的无缝结合

这是Oracle AI向量搜索的一个关键优势，因为它允许用户在现有的Oracle数据库中运行AI驱动的向量相似性搜索，而不是将业务数据移动到单独的向量数据库。避免数据移动可以降低复杂性，提高安全性，并支持对当前数据的搜索。

Oracle AI向量搜索通过使用简单、直观的 SQL 和融合数据库的全部功能（JSON、图形、文本、关系、空间等）将复杂的业务数据搜索与 AI 向量相似性搜索相结合，实现比大多数专用向量数据库更强大的搜索。 所有这些都可以在一个查询中实现。

向量搜索与业务数据结合举例：



#### 利用业务数据增强检索增强生成

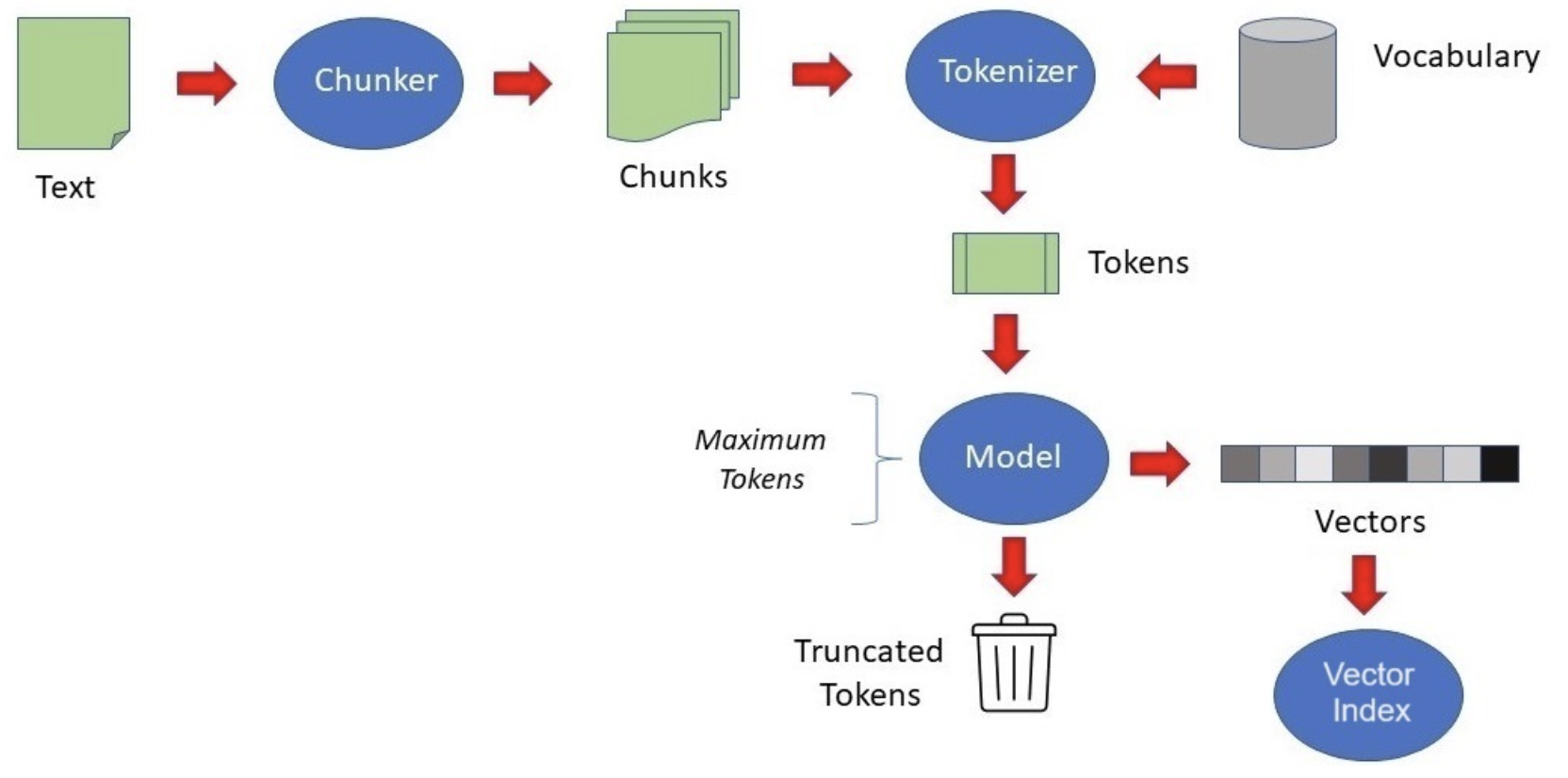
向量数据库通过为LLM提供特定场景的私有上下文，从而改善与LLM的交互，以获得更准确的答案。这是一个当前流行的工作流程，称为检索增强生成(RAG)。

Oracle AI向量搜索充分利用业务数据的全部力量，进一步细化 LLM 交互，利用安全过滤器、业务指标和业务规则等业务标准，为企业提供超复杂的RAG。

#### 支持业务数据的完整生成式AI管道

Oracle AI向量搜索支持原生数据库API来执行生成式AI管道的所有端到端步骤，使开发人员更容易直接在Oracle数据库中使用业务数据构建新一代AI应用程序。

从文本到向量的管道示例：



#### AI向量搜索与完整机器学习套件的独特组合

Oracle数据库提供全套数据库内机器学习算法以及 AI 向量的相似性搜索。这种组合使Oracle数据库能够处理非常广泛的AI使用案例，涉及机器学习操作（决策、预判、分类、预测等）以及AI向量搜索的强大功能。例如，可以轻松地将推理和分类与 AI 向量搜索结合在同一个 SQL 查询中。

#### 经过验证的企业级可扩展性、容错性和安全性

Oracle数据库是领先的业务数据存储库，业务数据和语义搜索的结合是企业实施人工智能解决方案所必需的。 AI 向量搜索内置于 Oracle 数据库中并利用：

* 分区、RAC、分片和 Exadata 可实现经过验证的工业级可扩展性。
* 极致的高可用性和灾难恢复技术，如Data Guard、Golden Gate、Flashback、RMAN、ZDLRA等。
* Oracle Advanced Security 提供前沿的安全性，包括透明数据加密、Key Vault、Audit Vault、虚拟专用数据库等功能。

## 应用场景

Oracle AI向量搜索可将语义和业务数据搜索相结合，从而快速、安全地提供更准确的答案。通过在Oracle数据库中添加AI向量搜索，用户可以快速、轻松地获得人工智能的优势，而无需牺牲安全性、数据完整性或性能。

应用场景包括：

* **对话AI或聊天机器人**：构建AI驱动的数字助手
* **相似性搜索**：匹配客户与产品
* **基于内容的过滤**：实现个性化推荐，根据图片查找零售商品
* **自然语言处理**：文本分类和聚类SQL生成
* **数据分析**：异常检测、模式识别
* **计算机视觉**：面部识别、生物识别、目标检测
* **生物医学研究**：基因/DNA相似性研究，分子结构搜索
* **地理信息系统**：空间分析、地图渲染
* **工业应用**：质量控制、预测性维护、机械故障

## 总结

Oracle AI向量搜索与Oracle数据库结合使用，使得一类新的应用程序成为可能，通过使用由现有业务数据增强的 LLM 进行语义搜索，可以改变传统业务流程。

* 新的SQL运算符和语法，可以轻松结合业务数据的关联搜索与非结构化数据的语义搜索
* 使用Oracle数据库进行AI驱动的向量相似性搜索 - 无需将业务数据移动到单独的向量数据库
* 轻松利用检索增强生成(RAG)来增强LLM对企业特定内容的响应
* 利用Oracle数据库的企业级安全性、可扩展性和分区功能增强应用程序

## 更多信息

要了解更多关于在Oracle数据库中使用Oracle AI向量搜索的信息，请参见：

* 功能：[AI向量搜索](https://www.oracle.com/database/ai-vector-search/)
* 文档：[AI向量搜索用户手册](https://docs.oracle.com/en/database/oracle/oracle-database/23/vecse)
* 博客：[在Oracle Database 23ai 中全面推出 AI 矢量搜索](https://blogs.oracle.com/database/post/oracle-announces-general-availability-of-ai-vector-search-in-oracle-database-23ai)
* 新闻：[甲骨文人工智能创新方案入选 2024 中关村论坛《百项新技术新产品榜单》](https://mp.weixin.qq.com/s/WCpBP9xnvJxFys2L9jD9uQ)