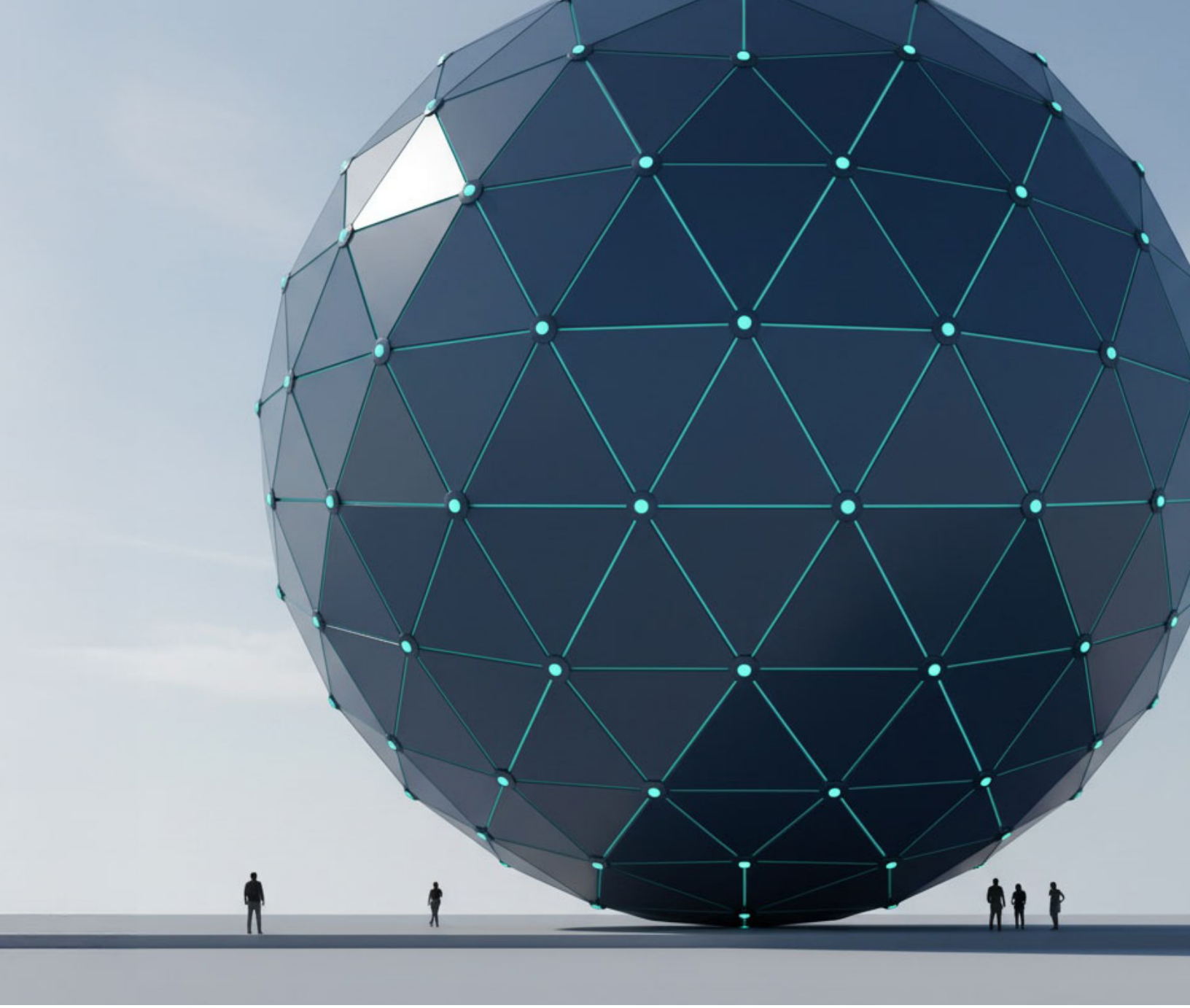


elastic

# 适用于气隙环境的高级矢量搜索

公共部门正处于生成式人工智能（GenAI）采用的关键时刻。组织受到推动，以推出更大规模的 AI 计划，这一推动源于全国范围的战略以及利益相关者的期望。与此同时，GenAI 和搜索已经紧密交织在一起。用户转向交互式 GenAI 以快速获取他们查询的答案，而不是使用搜索引擎提供的需要筛选的链接列表。

## 在公共部门满足这些期望会带来一些挑战：

1. 有 GenAI 使用的选择，但开发人员却不够。组织需要战略性地投入资源，以产生最有效的结果。

2. 公共部门组织需要应对大量数据。他们需要一个可扩展性、快速且高效的知识库作为 AI 的基础。

3. 数据并不总是为 AI 准备好的。关键信息通常以 PDF、图表和视频等格式存储，很难从中提取相关信息。

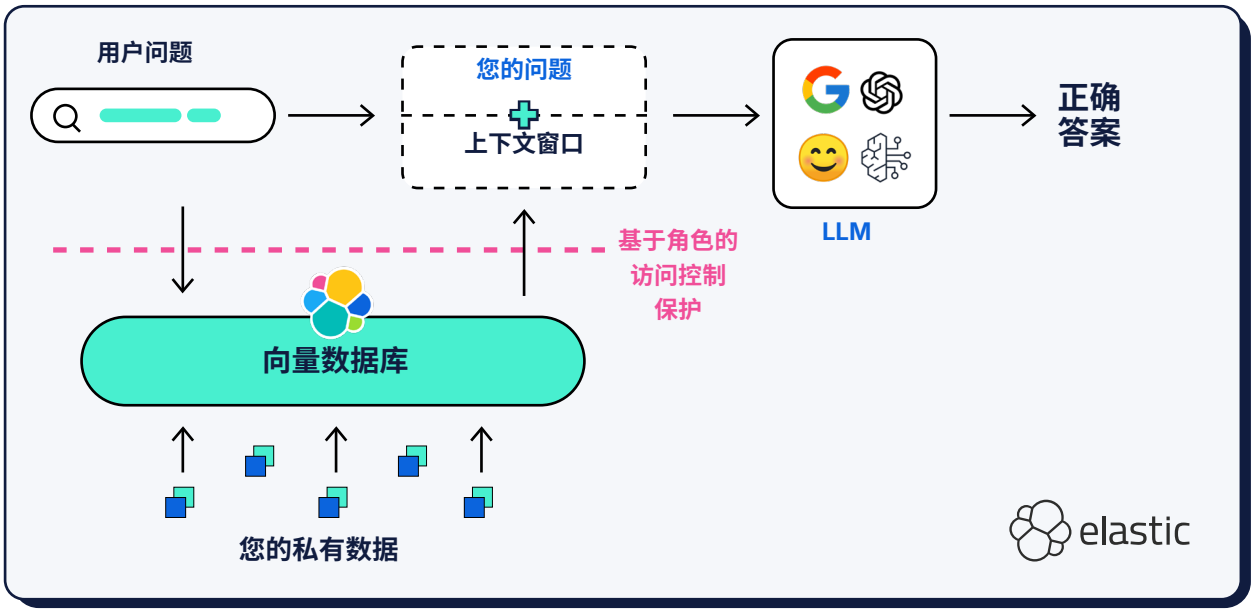
国防和情报机构还面临在气隙环境或离线设备套件中运行系统的额外挑战。

## RAG：一种用于 AI 的高效数据检索模型

许多公共部门机构发现，很难针对其利益相关者正在进行的查询类型对大型语言模型（LLM）进行训练或微调。他们希望将自己的专有数据开箱即用地与 AI 结合使用——即使这些数据很复杂。

 © 2025 Elasticsearch B.V. 版权所有。 

**Retrieval-Augmented Generation (RAG)** 是一种可扩展性解决方案，解决了围绕有限时间和资源以及如何扩展海量异构数据所面临的更大挑战。在 Elastic 的 RAG 模型中，在将用户的查询提交给 LLM 之前，它首先会查询我们的向量数据库，该数据库可以通过向量搜索从组织的专有数据中提取相关上下文。这一步，或称为上下文层，显著减少了 LLM 依赖其输出的数据范围。Agentic AI 进一步改变了此模型的性质，因为查询会与位于用户和 RAG 之间的 AI 代理进行交互。

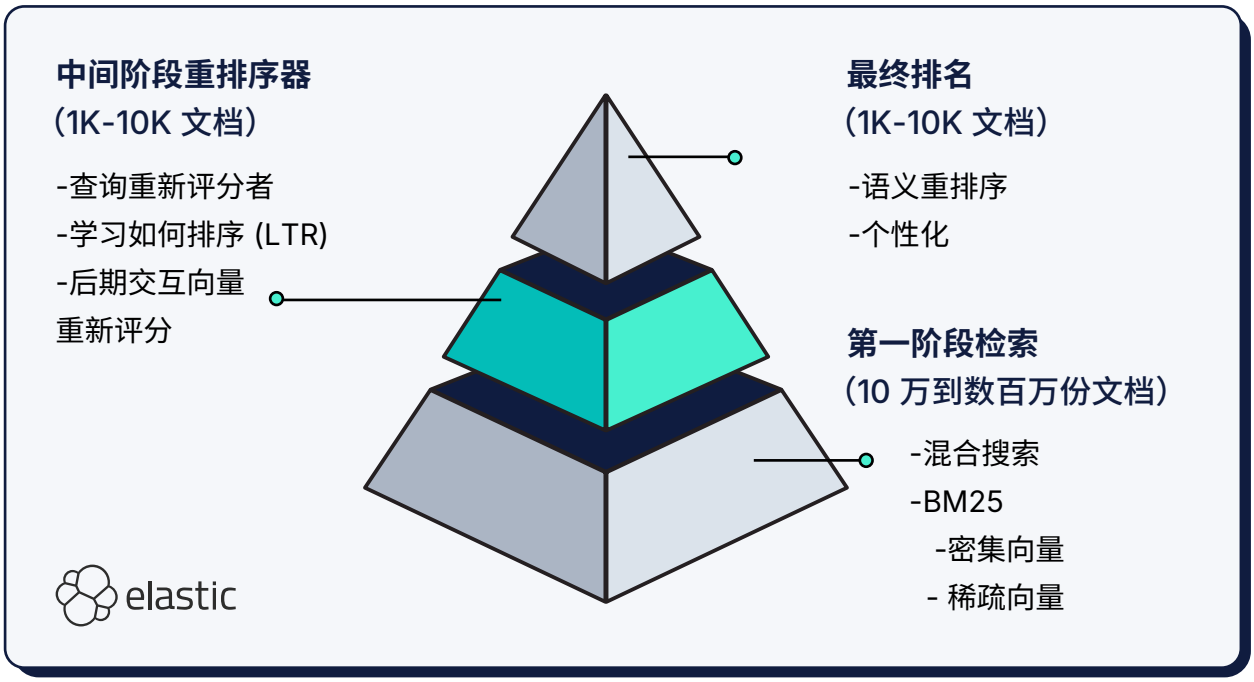



# 针对复杂数据类型的多阶段检索和多模式搜索

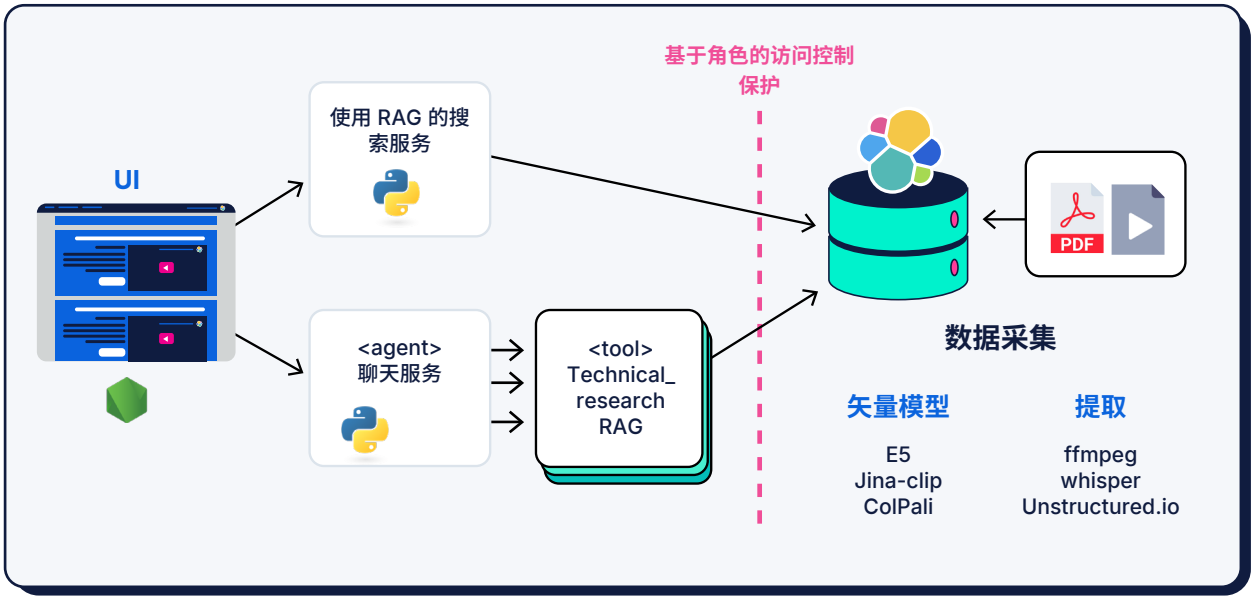
为了帮助公共部门组织使其复杂数据能够以 RAG 方法用于 GenAI，Elastic 利用**多阶段检索**和**多模态搜索**。

多阶段检索可将 RAG 过程中的复杂查询分解为较小的任务，以便提供最终答案。这种方法是基于近似结果和具体结果需要不同的技术。值得注意的是，Elastic 是首个完全支持气隙隔离环境下语义重排的向量数据库。

 © 2025 Elasticsearch B.V. 版权所有。 



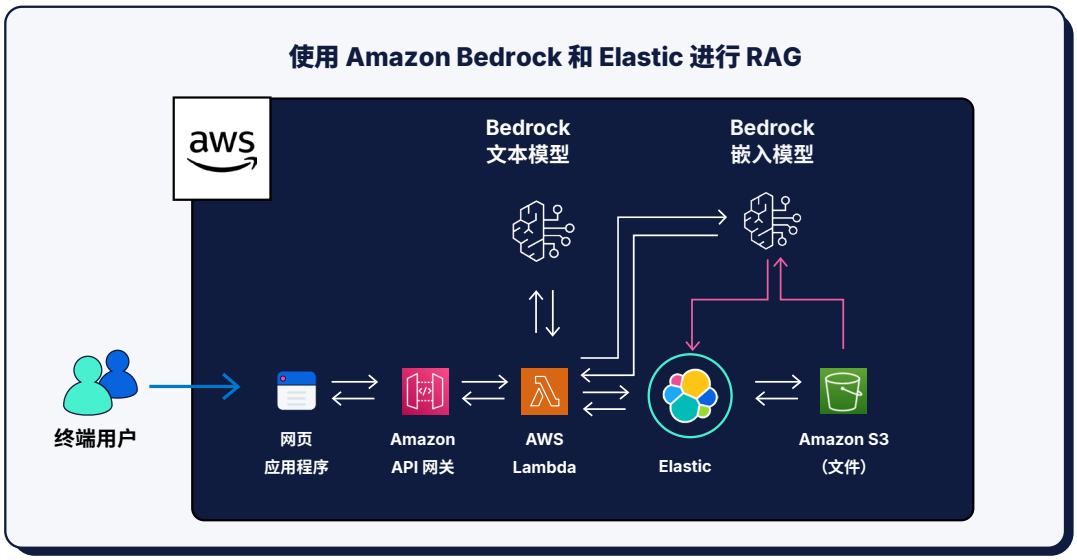

同样，随着组织寻求更直观、更强大的方式来检索不同数据类型的信息（例如带有表格的 PDF 或带有音频和视觉组件的视频），多模态搜索用例正在迅速获得关注。多模态搜索结合了文本、图像、音频和视频输入，最终使AI应用程序能够在正确的上下文中更快地回答问题。



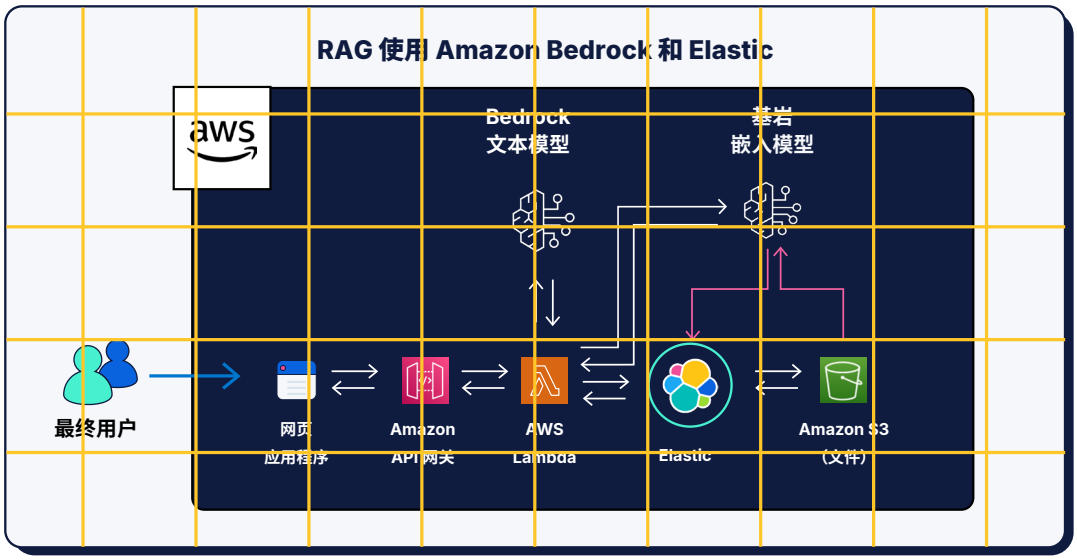

使用上述结构，组织可以找到嵌入在 120 分钟的 .MOV 文件深处的视觉信息，或者 1997 年 PDF 第 63 页图表中的短语。这种功能可确保复杂的专有数据可以全面搜索，并可供生成式 AI 工具在其输出中使用。Elastic 还支持子图像搜索，从而确保即使某些内容占据视频整体图像或帧的相对较小一部分，它仍然可以根据与用户查询的总体相关性在搜索结果中排名靠前。无论是在音轨上讨论的内容，而不是在屏幕上讨论的内容，反之亦然，如果它与用户的问题相关，Elastic 仍然可以在搜索结果中找到该内容并在搜索结果中排名靠前。

 © 2025 Elasticsearch B.V. 版权所有。 

# 借助 RAG 进行高级图像搜索




**ColPali 多向量**




让我们具体看个示例，说明 RAG 方法如何从图像文件中提取相关信息。在此例中，假设我们要搜索“使用 Amazon Bedrock 的 Elastic 架构图”。为了从上面的第一张幻灯片中获取信息，我们可能需要进行一些 OCR 和文本提取，并创建一些流程来帮助保留数据所提取的路径。这可能会变得复杂，尤其是每份文档都会有所不同。当搜索包含多栏排版和 8 页以上表格的长政府 PDF 时，要求就更高了。ColPali 能够帮助解决这个问题，因为它在图像上方创建了密集矢量网格。

从上方图片中我们可以看出，ColPali 已经识别出图片中 Amazon Bedrock 和 Elastic 以文本和徽标的形式出现的某些区域。Elastic 的徽标和参考信息实际上很小，但由于结合了整个图

 © 2025 Elasticsearch B.V. 版权所有。 

片的密集向量和利用 ColPali 排名向量对图片进行重新排名，我们的查询仍然能够找到并向我们展示此幻灯片。Elastic 已将这些功能内置到产品中以简化此过程，并使开发人员更容易创建满足其组织需求的应用程序。

# Elastic 是公共部门开放、可扩展的矢量搜索和 RAG 的解决方案

- 可互操作，支持任意 LLM
- RAG 方法，无需对 LLM 进行训练或微调
- 文档完善的 API 驱动架构，可完全自定义并与应用栈中的工具相集成
- 集成的基于角色和基于属性的安全控制（RBAC 和 ABAC）
- 全球下载量最大的向量数据库
- 推理 API 可在任何环境中注册 AI 服务
- 两种开箱即用的矢量模型——用于密集编码的 E5 和用于稀疏编码的 ELSER——以及使用您选择的任何其他嵌入模型的能力
- 矢量搜索功能针对排名质量、计算资源和减少查询延迟进行了优化
- 更好的二进制量化 (BBQ) 来压缩向量，使其搜索和处理速度更快
- 基于高级检索模型和矢量搜索能力，为智能体 AI 奠定可扩展性基础

 © 2025 Elasticsearch B.V. 版权所有。 

elastic

# 为未来做好准备

联系我们的公共部门技术团队，观看向量搜索如何与您的数据配合工作的演示：
elastic.co/cn/contact/public-sector。

