



使用 NVIDIA NIM 大规模部署 RAG workflow



课程概述

检索增强生成 (RAG) workflow 正在革新企业运营。然而，大多数现有教程仅停留在概念验证阶段，无法扩展。本课程旨在弥合这一差距，重点关注构建可扩展、生产就绪的 RAG workflow，由 NVIDIA NIM 和 Kubernetes 提供支持。学员将获得使用 NIM Operator 部署、监控和扩展 RAG workflow 的实践经验，并学习基础设施优化、性能监控和处理高流量的最佳实践。

本课程首先使用 NVIDIA API 目录构建一个简单的 RAG 管道。参与者将使用 Docker Compose 在本地环境中部署和测试各个组件。熟悉基础知识后，重点将转向使用 NIM Operator 在 Kubernetes 集群中部署 NIM，如 LLM、NeMo Retriever 文本嵌入和 NeMo Retriever 文本重排序。这将包括管理 NIM 的部署、监控和可扩展性。基于这些部署，课程将涵盖使用已部署的 NIM 构建生产级 RAG workflow，并探索 NVIDIA 的 PDF 提取蓝图 (blueprint)，学习如何将其集成到 RAG workflow 中。

为确保线上效率，课程将介绍 Prometheus 和 Grafana，用于监控 workflow 性能、集群健康状况和资源利用率。可扩展性将通过使用 Kubernetes 水平 Pod 自动缩放器 (HPA) 结合 NIM Operator，基于自定义指标动态扩展 NIM 来解决。将创建自定义仪表板以可视化关键指标并解释有关性能的洞察。

学习目标

参加本次培训，您将学到：

- > 使用 API 入口构建简单的 RAG workflow，并使用 Docker Compose 在本地部署
- > 使用 NIM Operator 在 Kubernetes 集群中部署各种 NIM 微服务
- > 将 NIM 组合成一个完整的生产级 RAG workflow，并集成高级数据提取 workflow
- > 使用 NIM Operator 扩展 NIM 以处理流量高峰
- > 为各种智能体 workflow (包括 PDF 提取) 创建、部署和扩展 RAG workflow

课程概要	
课程时长	8 小时，课后 6 个月内可以继续访问课件和实验 (实验资源用量有限额)
课程模式	讲师授课，每位学员使用完全配置的云端实验环境进行实战练习
预备知识	> 熟悉 LLM 应用 > 熟悉 RAG workflow > 熟悉 Kubernetes > 熟悉 Helm
使用的工具、库和框架	Python, NVIDIA NIMs, Kubernetes, Helm, Grafana, Prometheus
课程评测方式	部署企业级 RAG workflow
培训证书	成功完成本课程和测试后，学员将获得 NVIDIA DLI 培训证书，证明在相关领域的技能，为职业发展提供证明
课程语言	中文
学习此课程的硬件要求	您需要一台能够上网的笔记本电脑或台式机，且能够运行最新版 Chrome 或 Firefox 浏览器；我们为您提供在云端实验环境的专用访问权限
课程价格	微信添加 DLI 小助手 (微信号 NVIDIALearn)，沟通培训需求

课程大纲

课程介绍

(15 分钟)

- > 讲师介绍
- > 登录 DLI 学习平台

生成式 AI 应用的生产部署

(45分钟)

- > 调查当前生成式 AI 应用的各种能力
- > 回顾企业级生成式 AI 应用的众多组成部分，包括 RAG 工作流
- > 了解企业在从简单到企业级生成式 AI 应用转型时面临的挑战
- > 学习 NIM 在部署 LLM 和其它生成式 AI 应用组件方面的能力
- > 讨论用于大规模提高 LLM 应用推理性能的技巧和技术

企业级生成式 AI DevOps 的核心概念

(45 分钟)

- > 调查当前可用于企业级容器化应用部署的 DevOps 工具
- > 了解 Kubernetes 容器编排平台的价值

部署简单的 RAG 应用

(45 分钟)

- > 学习如何通过 API 调用访问远程 LLM 和 RAG 服务
- > 回顾在开发 RAG 应用时使用的核心 LangChain 编程模式
- > 使用 API 托管服务构建一个简单的 RAG 应用，并使用 Docker Compose 进行部署

休息 (60 分钟)

Kubernetes 核心概念

(60 分钟)

- > 学习使用 Kubernetes 集群所需的核心理念和技术
- > 熟悉本课程为您提供的交互式多节点 Kubernetes 集群编程环境
- > 交互式地使用 kubectl 在集群中部署、管理和监控基于容器的应用程序

部署自托管 RAG 应用

(60 分钟)

- > 在集群中部署和协调各种容器化微服务，以支持基于集群的 RAG 应用
- > 了解并使用 NIM 操作符来管理和扩展各种 RAG 微服务
- > 为集群上的各种模型缓存配置存储
- > 在集群上部署 LLM、文本嵌入和文本检索服务

监控 GPU 利用率

(45 分钟)

- > 使用 NVIDIA Data Center GPU Manager (DCGM)在 Kubernetes 集群中监控和管理 GPU 资源
- > 在集群上部署 Prometheus 和 Grafana 以更好地监控和可视化集群资源
- > 使用 DCGM Exporter 导出 GPU 利用率指标给 Prometheus，这些指标可以通过 Grafana 进行可视化

休息 (15 分钟)

自动扩展 NIM

(45 分钟)

- > 使用 Prometheus 服务监控器从集群上运行的 NIM中提取自定义指标
- > 创建 HorizontalPodAutoscalers 以根据自定义指标自动扩展集群的服务
- > 使用 Locust 执行负载测试，测试并观察自动水平自动扩展

构建多模态 RAG 流水线

(45 分钟)

- > 学习如何从多模态 PDF 文档中分离不同的模态，如文本、图形和表格
- > 练习各种提取文本的分块策略
- > 从 PDF 文档中提取表格
- > 从 PDF 文档中提取图像 / 表格
- > 使用 NV-YOLOX 模型识别 PDF 页面元素
- > 对 ChipNemo 技术论文执行端到端的多模态数据提取

使用生成式 AI 表示提取的模态

(30 分钟)

- > 为提取的元素 (如文本、图形、图表和表格) 创建详细的、有上下文的描述
- > 使用 VLM 执行图像转换
- > 使用最先进的上下文感知图表元素检测 (CHART) 模型来检测图表基本元素的类别，包括绘图元素
- > 在 ChipNemo 技术论文的端到端示例中结合使用 LLM 和 VLM

课程大纲

多模态嵌入、存储和检索

(30 分钟)

- > 将所有提取的模态转换为通用格式,以便在通用 RAG 流水线中使用
- > 构建端到端的多模态 RAG 流水线

评估测试和总结

(60 分钟)

- > 回顾所学要点
- > 使用全部所学部署并规模化企业级的 RAG 流水线进行评测,完成评估并获得证书
- > 填写培训调查表

下一步

学习更多 DLI 相关课程:

- > 利用提示工程构建大语言模型 (LLM) 应用
- > 构建基于 Transformer 的自然语言处理应用
- > 构建基于大语言模型 (LLM) 的应用
- > 构建大语言模型 RAG 智能体
- > 高效定制大语言模型 (LLM)
- > 模型并行 —— 构建和部署大型神经网络
- > 构建基于扩散模型的生成式 AI 应用

为何选择 NVIDIA 深度学习培训中心 (DLI) 的实战培训?

- > 学习 NVIDIA 与技术专家和行业领导者合作开发的课程,获取全球同步、技术领先和现实可用的专业开发技能和经验。
- > 学习使用行业通用、标准的软件、工具和框架进行端到端的应用开发,能够在广泛的行业中构建基于深度学习、加速计算、图形与仿真和数据科学的应用。
- > 系统化地学习理论,并使用云端完全配置的实验环境同步边练,高效提升实战开发能力。
- > 获得 NVIDIA 全球开发者培训证书,加持专业培训认证,助力职业发展。

准备好开始学习了吗?

查询更多 DLI 课程,请访问 nvidia.cn/training

如有疑问,请通过微信联系 DLI 小助手 (微信号 NVIDIALearn)。

© 2025 NVIDIA Corporation。保留所有权利。NVIDIA、NVIDIA 徽标、Clara、CUDA、DGX、DGX SuperPOD、Index 和 Triton 均为 NVIDIA Corporation 在美国和其他国家 / 地区的商标和 / 或注册商标。其他公司和产品名称可能是其各自关联公司的商标。其他所有商标均为其各自所有者的财产。3760505。2025 年 3 月

