Datasheet

在生产环境大规模部署 RAG 工作流

课程概述

检索增强生成 (RAG) 工作流正在彻底改变企业运营。然而，大多数现有教程仅停留在概念验证阶段，在扩展时常常力不从心。本课程旨在填补这一空白，重点讲解如何构建可扩展、可部署的生产级 RAG 工作流，利用 NVIDIA NIM 和 Kubernetes 实现。参与者将通过动手实践，学习如何使用 NIM Operator 部署、监控和扩展 RAG 工作流，并掌握基础设施优化、性能监控和高并发处理的最佳实践。

课程将从使用 NVIDIA API 目录构建一个简单的 RAG 工作流开始。参与者将在本地环境中使用 Docker Compose 部署和测试各个组件。在掌握基础知识后，重点将转向使用 NIM Operator 在 Kubernetes 集群中部署 NIM，如大语言模型 (LLM)、NeMo 文本嵌入和 NeMo 文本重排序服务。这一部分将涵盖对 NIM 微服务的部署、监控和扩展管理。在此基础上，课程将介绍如何基于这些部署构建生产级的 RAG 工作流，并探索 NVIDIA 提供的 PDF 数据摄取蓝图，学习如何将其集成进 RAG 工作流中。

为了实现运维效率，课程还将介绍如何使用 Prometheus 和 Grafana 监控工作流性能、集群健康状况和资源使用情况。通过结合 NIM Operator 和 Kubernetes 水平 Pod 自动伸缩器 (HPA)，课程将讲解如何基于自定义指标动态扩展 NIM。参与者还将创建自定义仪表盘，用于可视化关键指标并洞察性能瓶颈。

学习目标

参加本次培训，您将学到：

- 使用 API 接口构建本地 Docker Compose 部署的 RAG 工作流
- 使用 NIM Operator 在 Kubernetes 集群中部署多种 NIM
- 整合多个 NIM 构建生产级 RAG 工作流，并集成高级数据摄取流程
- 使用 Prometheus 和 Grafana 监控 RAG 工作流和 Kubernetes 集群
- 利用 NIM Operator 扩展 NIM 应对高并发流量
- 创建、部署并扩展适用于多种智能体工作流 (包括 PDF 摄取) 的 RAG 工作流

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	**> 开发基于提示工程的大语言模型 (LLM) 应用** **> 熟悉 RAG 工作流原理** **> 熟悉 Kubernetes 使用** **> 熟悉 Helm 工具**
课程涵盖主题与技术	部署企业级基于大语言模型的智能体和 RAG 应用： > 企业级生成式 AI 应用现状 > NVIDIA NIM > 企业级 RAG 应用的组件与架构 > 大规模推理的考量与优化 > 使用 Kubernetes、Helm 与 NVIDIA RAG Operator 部署、管理和扩展 RAG 服务

课程概要	
	> 使用 Prometheus 和 Grafana 实现集群行为与性能的可视化 > 部署和扩展多模态 RAG 应用的技巧
使用的工具、库和框架	Python，**NVIDIA NIM**，**Kubernetes**，**Helm**，**Grafana**，**Prometheus**
课程评测方式	部署并扩展一个企业级 RAG 工作流
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
生产级生成式 AI 应用部署 (45分钟)	> 了解当前生成式 AI 应用的广泛能力与应用场景 > 回顾企业级生成式 AI 应用的核心组成部分，包括 RAG 工作流 > 理解企业在从原型到生产部署中面临的挑战 > 了解 NIM 在部署 LLM 和生成式 AI 组件中的能力 > 探讨提升大语言模型推理性能的技术手段
企业级生成式 AI DevOps 的核心概念 (45 分钟)	> 了解当前容器化部署的主流 DevOps 工具 > 学习 Kubernetes 容器编排平台的价值
简单 RAG 应用部署 (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 集群所需的核心概念与技巧 > 熟悉课程提供的交互式多节点 Kubernetes 环境 > 通过 `kubectl` 进行容器应用的部署、管理和监控实践
自托管 RAG 应用部署 (60 分钟)	> 在集群中部署并协调多个容器化微服务构建 RAG 应用 > 使用 NIM Operator 管理和扩展多个 RAG 微服务 > 配置集群存储以支持模型缓存 > 部署 LLM、文本嵌入和文本检索服务
GPU 利用率监控 (45 分钟)	> 使用 NVIDIA DCGM 监控和管理 Kubernetes 集群中的 GPU 资源 > 在集群中部署 Prometheus 和 Grafana 以监控和可视化资源使用 > 使用 DCGM Exporter 将 GPU 指标导出到 Prometheus 并通过 Grafana 展示
休息 (15 分钟)	
NIM自动扩展 (45 分钟)	> 使用 Prometheus 提取集群中 NIM 服务的自定义指标 > 创建 HorizontalPodAutoscaler 实现基于指标的服务自动扩展 > 通过 Locust 进行负载测试，验证和观察自动扩展行为

课程大纲	
构建多模态 RAG 工作流 (45 分钟)	> 从多模态 PDF 文档中提取文本、图像和表格等元素 > 练习文本分块的不同策略 > 执行 PDF 表格与图像的提取任务 > 使用 NV-YOLOX 模型识别 PDF 页面元素 > 在 ChipNemo 技术文档上进行端到端多模态数据提取实践
使用生成式 AI 处理多模态内容 (30 分钟)	> 为提取出的文本、图像、图表、表格生成上下文描述 > 使用视觉语言模型 (VLM) 进行图像转换 > 使用 CHART 模型识别图表元素的类别 > 结合使用 LLM 与 VLM 完成 ChipNemo 技术文档的多模态示例分析
多模态嵌入、存储与检索 (30 分钟)	> 将所有模态转换为统一格式以集成进通用 RAG 工作流 > 构建完整的多模态 RAG 工作流
评估测试和总结 (30 分钟)	> 回顾所学要点 > 运用本课程所学内容，部署并扩展一个企业级 RAG 工作流 > 获取 NVIDIA 培训证书
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** > **基于 Transformer 的自然语言处理入门** > **构建基于大语言模型 (LLM) 的应用** > **构建大语言模型 RAG 智能体** > **模型并行 —— 构建和部署大型神经网络**

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生成式 AI、代理式 AI、大语言模型、加速计算、深度学习、数据科学、机器人开发的应用。
- 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 培训证书，并进一步考取 NVIDIA 认证，助力职业发展，强化团队关键技能。

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)。