Red Hat OpenShift AI

面向混合云中 AI 模型的开放平台

亮点

在久经验证的应用平台基础上实施和扩展 AI 推理和代理式 AI。

为数据科学家、数据工程师、应用开发人员和 DevOps 团队提供一致的用户体验，全面提高各个团队的 AI/ML 运维效率。

通过在本地、云端或边缘环境中构建、训练、部署和监控 AI/ML 工作负载，尽享混合云带来的灵活性优势。

拥抱智能应用与生成式 AI

人工智能（AI）、机器学习（ML）和深度学习（DL）正深刻影响着各行各业的应用现代化进程。创新需求以及从数据中挖掘战略价值和新见解的需求不断增长，这推动 AI 赋能的云原生应用、MLOps 和 GenAIOps 方法的应用范围不断扩大。在这个充满变革的新世界中，挑战错综复杂——从模型投入生产时成本的急剧攀升、复杂的定制需求、僵化的部署限制，到为跟上创新步伐所需的运维工作，不一而足。企业亟需能够降低推理成本、简化扩展与监控流程、并适应持续变革的解决方案。

红帽® AI 可加速混合云环境中企业 AI 解决方案的开发和部署。它作为一个用于管理整个 AI/ML 生命周期的综合性平台，提供 MLOps 和 GenAI Ops 功能。红帽 AI 特别关注四大关键支柱：

- 通过快速、灵活且高效的推理提升效率；
- 简化模型与数据的连接体验；
- 加速代理式 AI 创新；
- 确保在混合云中扩展 AI 时的灵活性和一致性。

红帽 OpenShift® AI 基于领先的混合云应用平台红帽 OpenShift 构建，是红帽 AI 产品组合中的旗舰产品。该 AI 平台为 AI 工程师、数据科学家和开发人员提供了强大的 AI/ML 基础，可用于大规模构建和部署生成式和预测性模型以及 AI 赋能的应用。企业组织可以在一个通用平台内试用各种工具、开展协作并加快产品上市速度。红帽 OpenShift AI 将数据科学家和开发人员需要的自助服务环境与企业 IT 所需的可靠性融为一体。

迅速开发、训练、测试和部署

红帽 OpenShift AI 是一个灵活、可扩展的 MLOps 平台，采用开源技术构建，为团队提供可靠且运维一致的功能，用于实验、模型服务以及交付创新应用。OpenShift AI 可加速交付 AI 赋能的应用，帮助企业组织以更快的速度和更强的控制力，从早期试点阶段迈向稳健的运维部署阶段。

该平台提供集成的用户界面（UI）体验，以及用于构建、训练、调优、部署和监控预测性 AI 与生成式 AI 模型的工具。您可以将模型部署到混合云环境，重点关注为主权 AI 和私有 AI 提供受控和受保护的部署环境。这种方法可确保敏感数据和 AI 模型保留在指定的地理或组织边界内，从而满足严格的监管和合规要求。

红帽官方微博　红帽官方微信

Red Hat
OpenShift AI

生成式 AI 分析机构预测

“AI 预计将成为 2026 年推动数字基础架构预算的一个非常重要的因素，因为各企业组织正努力使工作负载和数据需求与混合基础架构选择相匹配。90% 的决策者认为，AI 将成为影响其 2026 年数字基础架构预算和技术选择的重要因素。”[1]

简化 AI 采用

作为红帽 OpenShift 的附加组件，OpenShift AI 提供了一个平台，旨在通过将开源社区与强大的 AI 生态系统相结合，提高 AI 采用率并增强人们对 AI 计划的信任度。这为企业组织选择合适的 AI/ML 技术提供了更大的灵活性和自由度。用户可以构建自己的预测模型，或者采用外部生成 AI 模型，然后使用平台中提供的多个模型服务器之一，运用检索增强生成（RAG）技术来增强模型。该平台支持快速访问经过优化和验证的第三方模型，如 Llama、Mistral、DeepSeek 和 Granite，这些模型可在 vLLM 上高效运行，并可在 Hugging Face 上的红帽 AI 存储库中获取。此目录允许用户探索这些模型并添加自己的模型。

提高团队之间的运维一致性

红帽 OpenShift AI 提供一致的用户体验，使数据科学家、AI 工程师、开发人员和 DevOps 团队能够高效协作，及时交付 AI 解决方案。它提供对协作式工作流、图形处理单元（GPU）加速和简化运维的自助服务访问权限，从而在混合云环境和网络边缘大规模一致地交付 AI 解决方案。

IT 运维团队能够在久经考验的平台上，受益于简化的配置和自动化程度更高的工作流。该平台可轻松实现扩展或缩减，同时提供更完善的治理和安全防护机制。

提升混合云灵活性

红帽 OpenShift AI 支持在各种环境（云、本地数据中心或物理隔离环境）中训练、部署和监控 AI/ML 工作负载，以满足监管、安全和数据方面的要求。该平台与来自英伟达、AMD 和英特尔等供应商的多种 AI 加速器兼容。这种能力可以进一步扩展，以创建 GPU 即服务环境，使企业组织能够集中管理、分区和调度 GPU 资源，同时还能对资源使用情况进行详细的观测。

生成式 AI 和代理式 AI

对于生成式 AI 项目，OpenShift AI 通过 AI hub（开发人员预览版）等组件提供专门的用户体验。AI hub 是面向平台工程师设计的信息面板，它整合了目录、注册表和模型部署，以设置和部署模型及 MCP 服务器。生成式 AI 工作室（开发人员预览版）提供 AI 资产端点和 AI Playground，供 AI 工程师和应用开发人员访问、实验、比较、评估和测试已部署的模型及 MCP 服务器。

OpenShift AI 通过提供统一的 API 层和灵活、可扩展的基础来加速代理式 AI 的发展。Llama Stack API 和 MCP（技术预览版）支持包括 Llama Stack API 的企业级实施，为各种 AI 功能提供单一标准化入口点。

其他工具包括 LLM 评估（LM Eval）和 LLM 基准测试，以协助现实场景中的推理部署。LLM 压缩器提供的算法,可缩小企业组织自定义模型的规模，所用方法与红帽在 Hugging Face 上的红帽 AI 存储库中创建经过验证和优化的模型时所采用的方法类似。

1 *IDC 技术供应商。“AI Requirements Fuel Demand for On-Premises Infrastructure Deployments and Interoperability with Public Clouds, 2025”（AI 需求推动本地基础架构部署及与公共云互操作性的需求增长（2025）），文档编号 US53418426，2025 年 10 月。（需要客户端登录）*

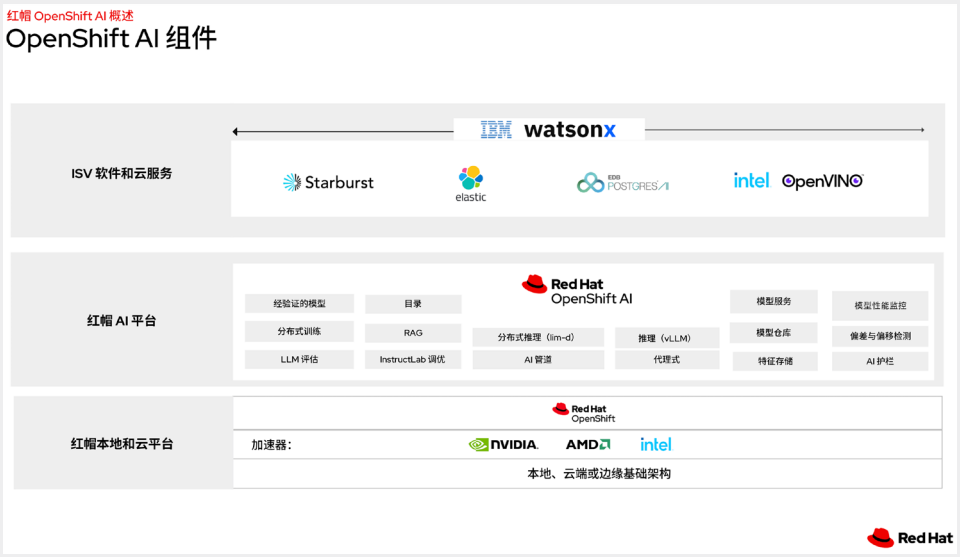

图 1. 红帽 OpenShift AI 和可选的合作伙伴组件

以下是红帽 OpenShift AI 提供的其他几个核心工具和功能，它们共同奠定了坚实的基础：

- **模型构建和定制。**数据科学家可以在 JupyterLab 用户界面中开展探索性数据科学工作，该界面提供了开箱即用、安全构建的 Notebook 镜像，其中包含常用的 Python 库。对于生成式 AI 项目，OpenShift AI 支持检索增强生成（RAG）和分布式 InstructLab 训练，提供模型对齐工具，以便更高效地为生成式 AI 模型贡献技能和知识。
- **模型服务。**红帽 OpenShift AI 提供多种框架，并使用 KServe 作为模型服务的核心引擎，以简化将预测性机器学习或基础模型部署到生产环境的过程。对于需要最高可扩展性的 LLM，OpenShift AI 通过 vLLM 运行时提供并行化服务。llm-d 提供了一个框架，通过将管道拆分为模块化服务来优化 LLM 推理，该框架支持智能自动扩展和高效的请求路由。
- **AI 管道。**红帽 OpenShift AI 提供了一个管道组件，可让您将 AI 任务编排到管道中，并使用图形前端构建管道。企业组织可以将数据准备、模型构建和模型部署等流程串联起来。
- **模型监控。**红帽 OpenShift AI 帮助运维用户监控模型服务器和已部署模型的运维和性能指标。用户可以使用开箱即用的可视化功能查看性能和运维指标，也可以将相关数据集成到其他可观测性服务中。
- **分布式工作负载。**分布式工作负载让团队可以加速数据处理以及模型训练、调优和部署等工作。此功能支持作业执行的优先级排序与分发，并实现节点的最优利用。先进的 GPU 支持能帮助处理基础模型的工作负载需求。

- **AI 防护、偏差和偏移检测。**红帽 OpenShift AI 提供一套工具，不仅可以帮助数据科学家和 AI 工程师基于训练数据评估模型是否公平无偏，还能在实际部署环境中持续监控模型的公平性。AI 护栏提供了一个可自定义的框架来实施关键的安全控制，有助于确保模型在生产环境中使用时透明、公平且可靠。对于已部署的 ML 模型，偏差检测工具会分析其输入数据的分布情况，以识别模型推理所用的实时数据是否与模型训练所用的数据发生了显著偏离。
- **目录和注册表。**红帽 OpenShift AI 提供内部模型目录和精选目录，供平台工程师发现、比较和评估优化的生成式 AI 模型。它还提供一个中央注册表，帮助数据科学家和 AI 工程师共享、修改、部署和跟踪预测性和生成式 AI 模型、元数据和模型工件。
- **特征存储。**管理 ML 模型所需的干净、定义明确的数据特征，提升模型性能，加速工作流。

完整 AI 生命周期所需的工具

红帽 OpenShift 提供一系列服务和软件，帮助企业组织成功训练和部署模型，并将模型投入生产（请参见图 2）。

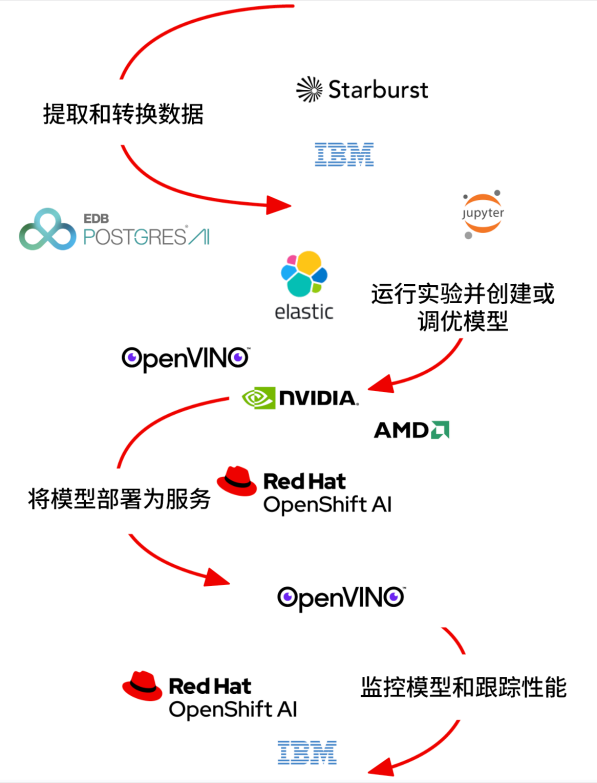

图 2. 典型的模型实施生命周期

Red Hat OpenShift AI

为方便使用，红帽 OpenShift AI 控制面板中还设计了一个用于发现和访问所有应用和文档的中心位置。“智能启动”教程直接集成在控制面板中，为常用组件和合作伙伴集成软件提供了最佳实践指南，从而帮助数据科学家更快速地学习并上手操作。以下各部分将介绍与红帽 OpenShift AI 集成的技术合作伙伴工具。部分工具可能需要从相应的技术合作伙伴处单独获取许可。

Starburst

Starburst 能够加速分析，帮助您的团队快速、轻松地利用您的数据来改善业务运作方式。无论是作为自助式产品还是全托管服务交付，Starburst 都能实现数据访问大众化，为数据消费者提供全方位的业务洞察。Starburst 基于出色的大规模并行处理（MPP）结构化查询语言（SQL）引擎——开源 Trino（以前称为 PrestoSQL）构建，由 Trino 专家构建和运维，让您能够自由查询任意位置的各种数据集，而无需移动数据。

Starburst 与红帽 OpenShift 提供的可扩展云存储和计算服务集成，从而形成一种稳定、注重安全、高效且具有成本效益的方式，用于查询所有企业数据。优势包括：

- **自动化。**Starburst 和红帽 OpenShift Operator 提供集群的自动配置、自动调优和自动管理。
- **高可用性和逐步缩减。**红帽 OpenShift 负载平衡器可以使 Trino 协调器等服务保持始终开启的状态。
- **弹性扩展。**红帽 OpenShift 可以根据查询负载自动扩展 Trino 工作集群。

HPE 机器学习数据管理软件

企业组织需要能够支持从笔记本电脑实验到重要企业部署等各种场景的数据管理解决方案。HPE 机器学习数据管理软件（之前称为 Pachyderm）支持数据科学团队构建和扩展由数据驱动的容器化 ML 管道，并通过自动化数据版本控制保障数据沿袭。HPE 机器学习数据管理软件专为解决现实世界的数据科学问题而设计，它提供的数据基础使团队能够以可再现的方式自动化和扩展其 ML 生命周期。HPE 机器学习数据管理软件的用例涵盖非结构化数据、数据仓库、自然语言处理、视频和图像提取、转换和加载（ETL）、金融服务和生命科学，可提供：

- 自动化数据版本控制：为团队提供了一种高性能的方式来跟踪数据更改。
- 由数据驱动的容器化管道：可加快数据处理速度，同时降低计算成本。
- 不可变的数据沿袭：为 ML 生命周期中的所有活动和资产提供固定记录。
- 控制台：可直观展示有向无环图（DAG），以帮助进行调试并实现可再现性。
- Jupyter Notebook 支持 JupyterLab Mount Extension，用户可在可视化界面中直接操作版本化数据。
- 企业级管理：提供强大的工具，用于在企业组织内不同团队间大规模部署和管理 HPE 机器学习数据管理软件。

Red Hat OpenShift AI

英伟达加速 AI 解决方案的部署

随着 AI/ML 应用对于业务成功越来越重要，企业组织需要能够处理复杂工作负载、优化硬件使用率并提供可扩展性的平台。可扩展的数据处理、数据分析、ML 训练和推理都是高度资源密集型计算任务。英伟达软件可以利用 GPU 的并行处理能力来全方位加速端到端数据科学进程。

在红帽 OpenShift 环境中，英伟达 NIM 能够增强英伟达 GPU 的管理和性能，使 AI 应用能够充分利用英伟达 AI 软件和硬件的潜力。英伟达 NIM 与红帽 OpenShift AI 集成，可实现更合理的资源分配、更高的效率，以及更具成效的 AI 工作负载执行。

英特尔 OpenVINO 工具包

英特尔 OpenVINO 工具包可加速高性能深度学习推理应用在英特尔平台上的开发和部署。借助该工具包，您可以以虚拟方式采用、优化和调优神经网络模型，并使用 OpenVINO 开发工具生态系统运行全面的 AI 推理。

- **模型。**软件开发人员可以灵活地使用自己的深度学习模型。为了加快产品上市速度，他们还可以使用英特尔与 Hugging Face 合作为 OpenVINO 工具包提供的预训练和预优化模型。OpenVINO 支持 Pytorch、ONNX、TensorFlow 和其他热门的模型格式。
- **优化。**OpenVINO 工具包提供多种模型转换方式，以提高便捷性和性能，帮助软件开发人员更快、更高效地执行 AI 模型。开发人员可以跳过模型转换，直接使用 PyTorch、ONNX、TensorFlow、TensorFlow Lite、JAX 或 PaddlePaddle 格式运行推理。通过转换为 OpenVINO IR，可获得最佳性能，使用 OpenVINO 神经网络压缩框架中提供的权重压缩和量化功能，可进一步优化性能。这些功能还可以减少存储和运行时占用空间。
- **部署。**OpenVINO 运行时推理引擎是一个应用编程接口（API），旨在集成到您的应用中以加速推理过程。它采用“一次编写、随处部署”的方式，支持在多种英特尔硬件上高效执行推理任务，包括中央处理单元（CPU）、GPU、NPU 和 FPGA。OpenVINO GenAI 扩展库简化了生成式 AI 工作负载的部署，在许多情况下，只需三到五行代码即可实现。OpenVINO 模型服务器为代理式和模型服务场景提供了多种功能，进一步减少了开发工作量。

EDB

EDB Postgres AI 是一个功能强大的智能平台，旨在处理事务、分析和 AI 工作负载，无论数据位于本地还是任何云环境中，都能提供无与伦比的灵活性。作为企业 Postgres 数据库解决方案的全球领导者，EDB 提供开放的企业级主权数据和 AI 平台，有助于将 AI 项目投入生产的速度提升高达三倍。通过与红帽 OpenShift AI 集成，EDB Postgres AI 允许用户为检索增强生成（RAG）构建强大的 AI 知识库，将 AI 数据、模型和应用统一到一个可在任何地方部署的全堆栈主权 AI 平台。将核心运维数据转化为 AI 就绪型资产，这样可将效率提升高达 30%，并简化私有数据（包括非结构化数据）的使用，使模型输出以企业组织的知识库为基础。

Elastic

Elastic Search AI 平台（基于 ELK Stack [2] 构建）融合了搜索的精准性与 AI 的智能性，使用户能够更快地进行原型设计并与 LLM 集成，同时利用生成式 AI 构建可扩展且经济高效的应用。Elastic Search AI 平台允许用户构建具有变革性的检索增强生成（RAG）应用，主动解决可观测性问题，并应对复杂的安全威胁。Elasticsearch 可部署在您的应用所在的位置：本地环境、您选择的云提供商，或物理隔离的环境中。

Elastic 通过一个简单的 API 调用，与来自红帽 OpenShift AI、Hugging Face、Cohere、OpenAI 等生态系统的嵌入模型集成。这种方式支持对 RAG 工作负载的混合推理进行清晰高效的代码管理，并具备以下特性：

- 数据分块、连接器和网页爬虫功能，支持将多样化数据集导入搜索层。
- 基于内置 ML 模型 Elastic Learned Sparse EncodeR（ELSER）和 E5 嵌入模型实现的语义搜索，支持多语言向量搜索。
- 文档级和字段级安全防护，实施与您企业组织基于角色的访问权限控制（RBAC）相对应的权限和授权。

使用 Elastic Search AI 平台，您将成为全球开发人员社区的一员，在这里您可以轻松获取灵感与支持。您可以通过 Slack、我们的讨论论坛或社交媒体加入 Elastic 社区。

结论

借助红帽 OpenShift AI，企业组织能够开展实验、促进协作，并最终加速其 AI 赋能应用的开发进程。数据科学家和 AI 工程师可以使用红帽 OpenShift AI 灵活地跨混合云构建和部署模型。IT 运维和平台工程师可借助 MLOps 和 GenAIOps 功能，更快地将模型部署到生产环境中。为开发人员、AI 工程师和数据科学家提供的自助服务（包括 GPU 访问权限），可在企业 IT 部门已在使用且充分信赖的应用平台上推动创新。红帽 OpenShift AI 持续提供一个全面、可信且一致的平台，在高效推理、代理式 AI 和可扩展的混合云运维方面提供独特的差异化优势，这一切都得益于其强大的合作伙伴生态系统。

了解更多

立即开始访问红帽 OpenShift AI。

关于红帽

红帽通过一流的支持、培训和咨询服务，帮助客户跨环境实现标准化、开发云原生应用，并实现复杂环境的集成、自动化、安全防护和管理。

2 *ELK 堆栈由 Elasticsearch、Kibana、Beats 和 Logstash 组成。*

销售及技术支持

800 810 2100
400 890 2100

红帽北京办公地址

北京市朝阳区东大桥路 9 号侨福芳草地大厦 A 座 8 层　邮编: 100020
8610 6533 9300

红帽官方微博　红帽官方微信

cn.redhat.com
#2876428_1125