

CNCF AI 工作组

云本土人工智能

Authors

Adel Zaalouk
Alex Jones
Andrey Velichkevich
Boris Kurktchiev
Cassandra Chin Cathy
Zhang Claudia Misale
Huamin Chen

Joel Roberts Kai -
Hsun Chen Malini
Bhandaru Michael
Yao
Nikhita Raghunath 彼
得潘
Rajas Kakodkar
Rasik Pandey

Ricardo Aravena
Ronald Petty Ryan
Taylor Saad Sheikh
Shawn Wilson Tom
Thorley Victor Lu

已发布

3 月 20, 2024 (第 1 版)

执行摘要

云原生（CN）和人工智能（AI）是当今最关键的技术趋势。Cloud Native¹ 技术为运行应用程序提供了可扩展且可靠的平台。鉴于 AI 和机器学习（ML）的最新进展，它作为主要的云工作负载正在稳步上升。虽然 CN 技术很容易支持 AI / ML 工作负载的某些方面，但挑战和差距仍然存在，为创新和更好地适应提供了机会。

本文简要概述了最先进的 AI / ML 技术，其次是 CN 技术提供的内容，在讨论不断发展的解决方案之前涵盖了下一个挑战和差距。本文将为工程师和业务人员提供知识，以了解不断变化的云原生人工智能（CNAI）生态系统及其机遇。

我们建议阅读路径取决于读者的背景和兴趣。假设暴露于微服务² 和 CN 技术³，如 Kuberetes (K8s)。对于那些没有工程 AI 系统经验的人，我们建议从头到尾阅读。对于那些在 AI / ML 采用或交付过程中走得更远的人，根据他们的用户角色⁴，我们建议深入到与他们正在努力解决或有兴趣解决的挑战相关的部分。我们还分享社会在这方面需要投资的地方。

内容表

云原生 (CN) 04 的出现
人工智能进化 (AI) 05

建议 19

结束注释 28



云人工智能 (CNAI) 简介

在我们进入 CNAI 之前，将 Cloud Native 和 AI 技术结合在一起，让我们简要地研究一下每种技术的演变。

云原生的出现

自 2013 年以来广为人知，⁵ 随着容器技术从 LXC6 到 Docker7 再到 Kuberetes（K8s）的兴起，Cloud Native（CN）一词越来越受欢迎。⁸ 如今，Cloud Native 更广泛地成为使用微服务设计模式构建的平衡系统的理想目标，该模式可促进模块化设计和开发，具有高度的可重用性。这也有助于部署性、可扩展性和弹性。

云原生计算基金会定义⁹云原生为：

Cloud Native 技术使组织能够在现代动态环境（如公共云、私有云和混合云）中构建和运行可扩展的应用程序。容器、服务网格、微服务、不可变基础设施和声明性 API 就是这种方法的例证。

这些技术使松散耦合的系统具有弹性，可管理和可观察性。结合强大的自动化，它们使工程师能够以最少的工作量频繁且可预测地进行高影响的更改。

云原生计算基金会寻求通过培育和维持一个开源、供应商中立的项目生态系统来推动这一范例的采用。我们将最先进的模式民主化，使每个人都能获得这些创新。

云原生人工智能是云原生的一个不断发展的扩展。

Cloud Native Artificial Intelligence (CNAI) 是指使用 Cloud Native 原理构建和部署 AI 应用程序和工作负载的方法和模式。启用可重复且可扩展的 AI 工作流，¹⁰ 可让 AI 从业者专注于自己的领域。

Kuberetes 已经发展成为事实上的云操作系统，包括私有、公共和混合云产品。它实现了一个分布式编排器，用于处理多种类型的网络、存储和计算资源。此外，K8s 提供了一个接口，使 DevOps¹⁰ 的最佳实践，如 GitOps.¹¹ 每个云服务提供商（CSP）¹² 都有一些 Kuberetes 服务的味道，便于访问基础设施和一系列支持服务来运行各种工作负载，包括 AI / ML。

人工智能的进化

人工智能，最早在 1956 年被称为一个术语，¹² 是机器模拟人类智能的能力。几十年来，它已被用于语音识别，机器翻译，图像处理，游戏等应用，甚至是作为危险玩家的出色表现。¹³ 但是，由于人工神经网络和深度学习的创新，人工智能最近在 midshare 中爆发了，主要应用于自然语言理解。AI 有两种主要分类：判别性和生成性。



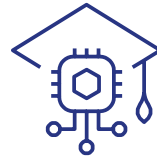
判别式 AI 寻求学习决策边界或分类，将知识捕获为“模型”，用于预测新数据。例如，将电子邮件分类为垃圾邮件，区分猫和狗的图像等等。判别 AI 通常用于已知所需输出的任务（例如 **Procedre**，通过监督学习，一种机器学习的形式）。人工智能擅长序列预测，例如，通过分析大量现有文本，包括我们的个人写作风格，以高概率猜测我们接下来要输入的内容。



生成 AI 学习数据中的潜在结构或表示。它可以使用这些结构或表示来合成新数据，例如创建故事，音乐和视觉艺术来自单词提示。生成性 AI 用于所需输出未知或“正确”输出定义不明确的任务。使用生成性 AI，AI 已经超越了人类认为的创造性，原创性和崇高性。让我们仔细看看 AI 的一些惊人突破。



卷积神经网络¹⁴（CNN） 最初是在 1980 年代开发的，但直到 21 世纪初才被广泛使用。近年来，由于它们能够从图像的大型数据集中进行学习，并在各种图像处理任务（例如对象检测，图像分类和分割）上表现良好，CNN 变得越来越受欢迎。



变压器 由多伦多大学和谷歌的研究人员于 2017 年开发。变形金刚使用一种称为缩放点积注意力的专门机制，该机制使它们充满了类似记忆的结构。¹⁵ 基于变形金刚的模型对于自然语言处理任务非常有效，例如回答问题，总结文本和翻译。因此，它们在大多数大型语言模型（LLM）中至关重要。最著名的 LLM 是 GPT，该模型为流行的 ChatGPT 服务提供动力。

LLM 是在海量数据集上训练的。除了能够针对具有额外数据的专业领域进行微调之外，它们还采取可能很长的提示序列来生成上下文敏感的响应，无论是时事，医学，法律还是其他。用于微调的新技术，例如来自人类反馈的强化学习（RLHF）和直接偏好优化（DPO），已经被开发出来，以使 LLM 更具吸引力。

研究和创新使最终用户的交互比以往任何时候都更快，更有创造力，更准确。与数据科学和软件的创新一样重要的是基础设施的发展 *模型推理*（从 AI 模型计算结果的过程）和 *模型训练*（从数据构建 AI 模型的过程）。使用 AI 加速器技术，人工智能从业者可以更快地迭代，以在几天和几周内提供更高质量的模型，而不是几个月。此外，数据科学家和统计学家采用的几种传统技术正在重新评估，以利用 CN 系统的功能。

云原生与人工智能的融合

正如上一节所述，人工智能是一个更广泛的概念，旨在创建可以执行类似于人类任务的系统。机器学习是一种基于数据进行学习并做出明智预测和决策的方法。它可以被认为是另一种形式的自动化，涉及使用算法来学习和改进，而无需显式编程。最后，数据科学作为一个多学科领域，融合了统计学，数学和计算机科学的技术来制定。

广泛的活动，从数据分析和解释到机器学习算法的应用。

从广义上讲，我们可以将 AI，ML 和数据科学的应用程序分为两大类：**预测性 AI** 和 **生成 AI**。预测性 AI 旨在预测和分析现有模式或结果（例如，分类，聚类，回归，对象检测等）。相比之下，生成 AI 旨在生成新的和原始的内容（例如，LLM，RAG¹⁷ 等）。因此，支持预测性和生成 AI 的算法和技术可能会有很大差异。

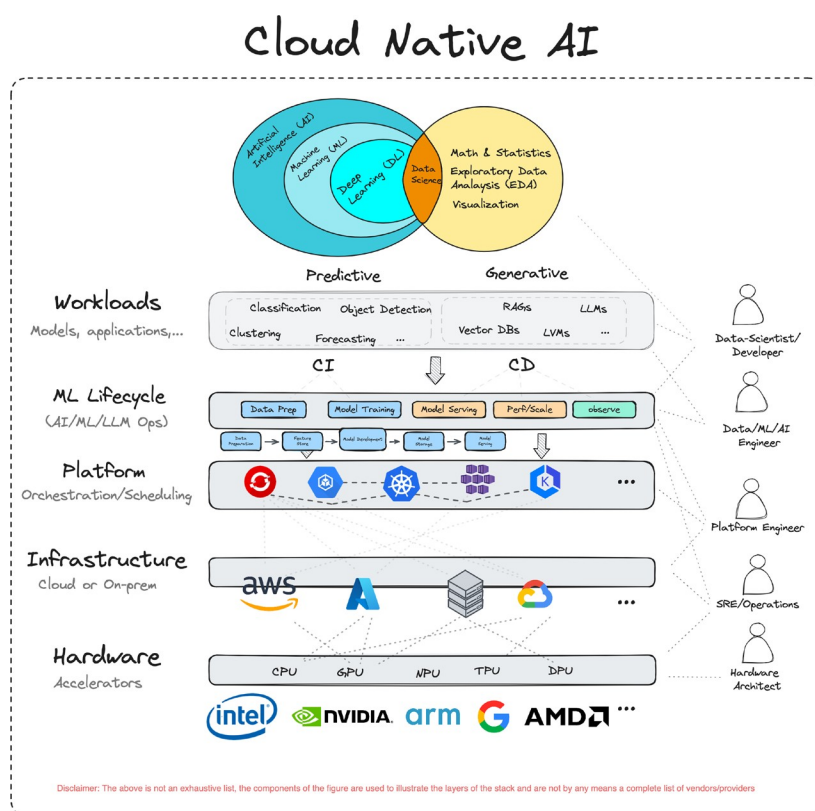


图 1 云原生 AI

以下是预测和生成 AI 在计算、网络和存储方面有不同需求的一组示例：

挑战 / 需求	生成 AI	预测性 AI
计算型 Power	非常高。 需要专门的硬件。	中等到高。 通用硬件就足够了。
数据量和多样性	用于培训的大量、多样化的数据集。	预测的具体历史数据。
模型训练和微调	使用专业计算进行复杂的迭代训练。	适度的训练。
可扩展性和弹性	高度可扩展和弹性的基础设施（可变和密集的计算需求）	可扩展性是必要的，但要求较低的弹性。 批处理或事件驱动的任务。
存储和吞吐量	具有出色吞吐量的高性能存储。数据类型多样。 需要高吞吐量和低延迟的数据访问。	高效存储，吞吐量适中。它更侧重于数据分析，而不是数据生成；数据主要是结构化的。
联网	用于数据传输和模型同步 (例如, 在分布式训练期间) 的高带宽和低延迟。	数据访问的一致可靠连接。

在接下来的部分中，我们将探讨如何满足这两种形式所产生的需求，随之而来的挑战，以及在面对这些挑战时可能提出的建议。

什么是云原生人工智能？

云原生人工智能允许构建实用的系统来部署、运行和扩展 AI 工作负载。CNAI 解决方案解决了 AI 应用科学家、开发人员和部署人员在云基础设施上开发、部署、运行、扩展和监控 AI 工作负载时面临的挑战。通过利用底层云基础设施的计算 (例如 Procdre，CPU 和 GPU)，网络和存储功能，以及提供隔离和受控共享机制，可加速 AI 应用程序性能并降低成本。

图 2（下图）在工具和技术之间映射了这些启用机制。

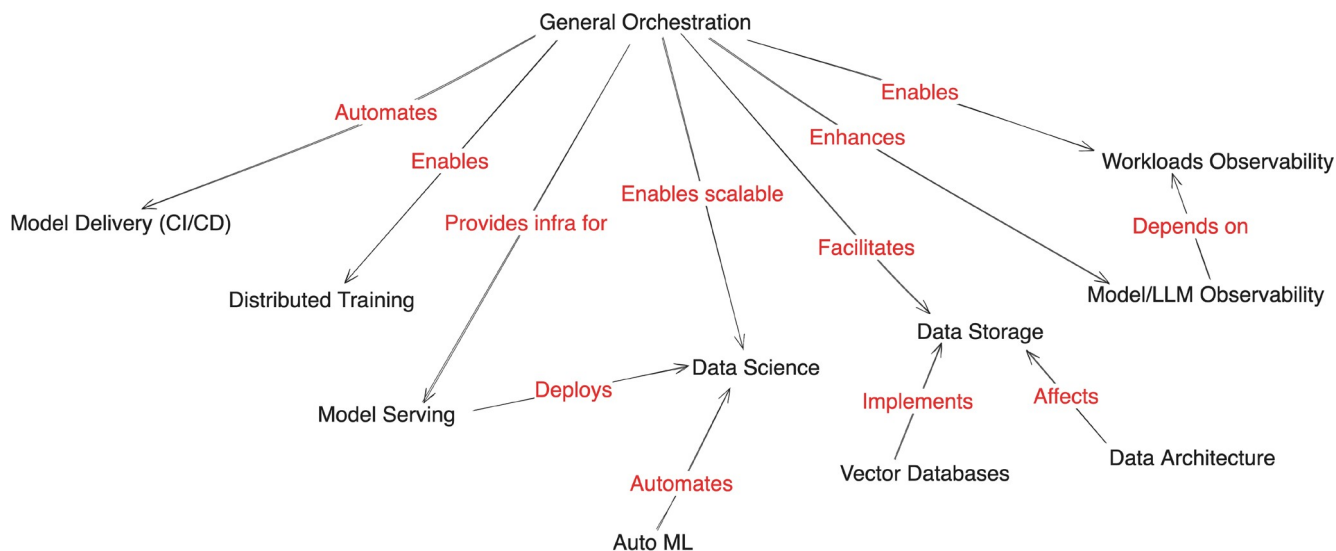


图 2
启用工具和技术 18

在云原生基础设施上运行 AI

云服务提供商和 / 或 AI 公司发布的媒体文章强调了 Cloud Native for AI 的价值。

OPENAI 将 Kubernetes 扩展到 7,500 个节点

拥抱的脸 拥抱脸与 Microsoft 合作在 Azure 上启动拥抱脸模型目录

云原生人工智能是云原生的一个不断发展的扩展。

Kubernetes 是一个可用于部署和管理容器的编排平台，容器是轻量级、可移植、自包含的软件单元，AI 模型可以打包成容器然后部署到 K8s 集群。容器化对于 AI 模型尤其重要，因为不同的模型通常需要不同且通常相互冲突的依赖关系。在容器中隔离这些依赖关系可以在模型部署中提供更大的灵活性。CN 工具允许 AI 模型的高效和可扩展部署，并不断努力为 AI 工作负载定制这些模型。

Kubernetes Scheduler²¹ 继续发展，^{22 23} 特别是为了更好地集成和支持共享图形处理单元 (GPU)，这些图形处理单元在加速 AI 工作负载方面变得非常流行。除了支持共享 GPU 和处理多租户的应用程序之外，还在努力支持利用 Kubernetes 之外的远程资源池。

需要高质量的数据来训练和测试 AI 模型，以获得卓越的推理。云原生基础设施可以通过各种方法访问数据，例如数据湖和仓库。许多云提供商提供块、对象和文件存储系统，非常适合提供低成本、可扩展的存储。例如，模型的大小可以达到千兆字节。在训练阶段，每次拉取模型的检查点都会导致网络 and 存储带宽的严重负载。将模型视为容器化的工件为在 OCI²⁴ 注册表中托管它们打开了大门，并启用了缓存。它进一步允许应用。软件供应链模型的最佳实践，例如工件签名，验证，证明和数据来源。此外，容器化模型 / 工件促进了 WebAssembly (WASM) 二进制文件的捆绑。WASM 是一种独立于平台的高效 CN 推理方法。

为什么选择云原生人工智能？

凭借其弹性，始终在线的基础架构，云允许企业，初创公司和开发人员快速原型，提供新服务，扩展解决方案等等。它还通过资源共享实现了成本效益。普通用户不再需要担心订购硬件或处理空间、电源、网络连接、冷却、软件许可和安装等物流问题。人工智能也有类似的担忧——快速原型设计、访问存储、网络 and 计算资源，以解决小型和大规模的训练和推理任务。

使用 AI 改进云原生系统

无论是打包为可观察性工具还是利用 LLM 功能进行日志的自然语言处理（NLP），AI 驱动的方案 / 项目都在进入运营商和最终用户的手中，以提高他们的生产力并使他们的生活更轻松。一个这样的开源云原生计算基金会（CNCF）项目是 K8sGPT，它利用 LLM 的模式识别和语言功能，如 Bedrock，Cohere 等，以帮助 K8s 运营商日常工作。更重要的是，CN 和 AI 的共生为新的和不可见的机会打开了生态系统。例如，我们预计能够操作和管理复杂系统的技术含量较低的用户将会增加。

云人工智能的挑战

重要的是要注意，CNAI 的挑战在不同的角色之间会有所不同。26 而且，尽管 Cloud Native 的灵活，可扩展的平台非常适合 AI 工作负载，但 AI 的规模和延迟需求带来了挑战，并暴露了 CN 技术中的差距，同时也带来了机会。我们在端到端 ML 流水线的背景下梳理这些内容。27 在文献中也称为 MLOps。28 传统的时间和空间，并行性和同步权衡的问题都存在，暴露了易于使用的差距。总而言之，ML 生命周期如下所示：

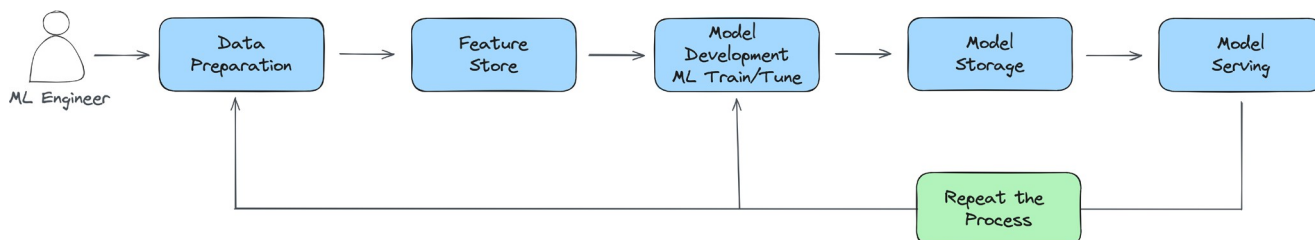


图 3 ML 生命周期

典型的 ML 管道包括：

- 数据准备（收集、清洗 / 预处理、特征工程）
- 模型训练（模型选择、架构、超参数调优）
- CI / CD, 模型注册表（存储）
- 模型服务
- 可观察性（使用负载、模型漂移、安全性）

训练、相似性搜索和模型大小（特别是 LLM）中涉及的数据量，每个驱动器内存和性能方面的考虑因素。虽然 CN 处理 CPU 的访问控制和调度，但具有充分共享的 GPU 分配仍在不断发展。ML 训练阶段涉及搜索，需要跟踪中间模型的性能，以确定要保留哪些模型以及如何进一步调整模型参数以获得更高的准确性。考虑到处理数据的敏感性和模型的内在价值，安全性更为重要。可观察性对于检测模型漂移、使用负载等至关重要。让我们更深入地探讨每个管道阶段的挑战。鼓励读者考虑与其领域相关的其他挑战，并添加到对话中。

数据准备

作为 AI / ML 管道的第一阶段，数据准备可能会带来各种挑战。这些可以大致分为三大类：管理大数据大小，确保开发和部署期间的数据同步以及遵守数据治理策略。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/908051117112006055>