KubeCon 2024: open-source AI tools are becoming crucial enablers of cloud-native AI

March 2024

Joseph Attwood

Analysys Mason attended KubeCon + CloudNativeCon Europe 2024 in Paris, France in March 2024. This annual Cloud Native Computing Foundation (CNCF) event brings together open-source contributors and end users to share how cloud-native technologies can be applied. This year, cloud-native AI (that is using Kubernetes (K8s) to run AI workloads) was a headline issue. The CNCF AI Working Group kicked off discussions by releasing a whitepaper on the topic the day before the conference during the co-located Cloud Native AI Day event. The main focus at KubeCon was on AI inference workloads, and the quotes "inference is the new web app" and "if inference is the new web app, K8s is the new web server" were repeated frequently throughout the conference.

The open-source community is working to ensure that K8s meets the requirements of generative AI workloads

Many large enterprises are already using K8s to run their large language model (LLM) inference workloads; OpenAI has been doing so since 2016, for example. K8s's scalability, portability and self-healing makes it a good fit for these workloads, but it was not designed to support AI workloads; that it can do so is testament to its flexibility. This means that deploying AI models on K8s is not always straightforward, and less experienced developers and/or smaller enterprises often struggle with the scale and latency requirements of LLM workloads.

The landscape for open-source initiatives that enable cloud-native AI has evolved rapidly in the past few years and is expected to continue to do so in the future. For example, the KServe project provides a highly scalable inference platform for deploying traditional/generative AI inference workloads on K8s. The Kubeflow project provides a toolkit that supports developers to run training and inference workloads on K8s, and its Pipelines component helps developers to create and automate ML workflows. However, many of these projects still require updates to improve feature functionality and ease-of-use. For example, one presenter mentioned that Kubeflow is currently "horrible to set up". Cloud-native AI also has other pain points that have not yet been suitably addressed.

Automated workflows, hardware agnosticism and resource sharing are all crucial requirements of cloud-native AI

Deploying AI inference workloads on K8s is often a slow process because it involves numerous disjointed steps that require manual intervention. Developers must download model weights, create and host container images, provision GPU infrastructure, tune deployment parameters to the underlying hardware, troubleshoot deployment hardware and set up the inference server as an endpoint service. Developers will need tools that can streamline this process as much as possible as the number of models that they deploy and manage grows. Kubernetes AI Toolchain Operator (Kaito) is a good example of how this can be done. Kaito helps to automate the deployment of inference models on K8s clusters in Azure Kubernetes Service (AKS); this includes selecting the optimal

infrastructure size. However, enterprises will need tools that can also deploy models on-premises and/or in other public cloud environments.

Hardware agnosticism is another key requirement for cloud-native AI; AI platforms must provide an abstraction layer with a unified set of APIs that work across multiple processor architectures. This enables developers to deploy AI models without needing to understand the ins and outs of underlying hardware. Hardware agnosticism also enables AI workloads to be more portable because developers do not need to manually reoptimise deployment parameters for different deployment environments. The Unified Acceleration (UXL) Foundation is leading work in enabling open accelerated compute.

Optimising GPU resource usage was another important discussion topic at KubeCon because GPUs are expensive to purchase and run and their procurement has long lead times. Presenters shared that there are long periods of relatively low usage for most LLMs, thereby resulting in it being necessary for multiple models to share GPU resources. The allocation of GPU resources will need to be an automated process that is not the responsibility of developers; resources should be allocated automatically at deployment and allocations should be scaled automatically over time as inference demands change.

Other concerns raised during the conference included simplifying how developers can run multi-cloud, multi-cluster AI workloads and extending observability frameworks to support and better understand the semantics of LLM applications and their underlying GPUs. Furthermore, developers will need AI platforms that can deploy inferencing models on CPUs as well as GPUs and that can deploy AI models that use serverless architecture (by using WebAssembly, for example).

Vendors should use open-source components in their AI platforms to address the challenges of cloud-native AI

Not all enterprises will have the capacity to stitch together the various components offered by the open-source community. A reference architecture that could guide platform engineers on the best combination of open-source tools and how to fit them together would go some way to simplifying this. However, many enterprises prefer to rely on vendors' proprietary AI platforms that take the burden of integration and support away from in-house teams.

Vendors will need to extend their existing MLOps platforms to better support running AI workloads on K8s to meet the demand for cloud-native AI. When doing so, vendors should aim to incorporate open-source components where feasible; their value-add proposition should be to pre-integrate these components together, make them production-grade and provide support services. Without the commonality provided by open-source solutions, platforms offered by different vendors will be highly opinionated and it will be difficult for enterprises to switch between platforms without reconstructing their AI/ML workflows. In addition, KubeCon demonstrated that there is enormous momentum behind developing open-source tools that enable cloud-native AI. Vendors that do not use open-source components will lose access to the innovation capabilities of the open-source community, 12 000 members of which turned up in Paris.