Ecco
Groupe

Artificial intelligence is not flash in the pan — it is here to stay. Gartner says more than 80% of enterprises will have used some form of generative AI APIs or applications by 2026. If you plan to be among those 80%, then you must determine the best way to train and deploy it, on premises or in the cloud.

AI training requires specialized hardware that is very, very expensive compared to standard server equipment. It starts at the mid-six figures and can run into the several-million-dollar range. And that hardware cannot be repurposed for other uses such as databases.

In addition to purchasing and maintaining the AI hardware, there is the model on which your AI application is based. Training is the difficult part of AI and the most process intensive. Training can take weeks or even months, depending on the size of the data set. That could be months you don't have.

]

So, you have the option of acquiring the hardware and doing it yourself or turning to an AI-as-a-service provider. AIaaS is the latest in the as-a-service market, specifically oriented around artificial intelligence initiatives. It is typically provided by major cloud service providers (CSPs), but smaller vendors are entering the market as well. Manyalso,aS providers offer not only the hardware for lease, but also prebuilt models, which can shave months off the time to deploy.

AI was already on the radar for many companies, but when ChatGPT and generative AI (genAI) exploded onto the scene in late 2022, businesses felt a greater sense of urgency to adopt it, says Mike Gualtieri, vice president and principal analyst with Forrester Research.

"With generative AI, they can just use a model that's prebuilt, and that's largely what they're doing. So, they don't need to buy their own infrastructure. Many of them are thinking they're going to fine-tune an open-source model," he said.

How AIaaS works

AIaaS provides customers with cloud-based access for integrating and using AI capabilities in their projects or applications without needing to build and maintain their own AI infrastructure. It also offers prebuilt and pretrained models for basic uses like chatbots so the customer doesn't have to go through the process of training their own.

"Basically, AI-as-a-service enables you to accelerate your application engineering and delivery of AI technologies in your enterprise," said Chirag Dekate, vice president and analyst for AI infrastructures and supercomputing at Gartner.

AI-as-a-service offers three entry points: the application level, the model engineering level, and the custom model development level, he said. If you're a relatively low-maturity enterprise and want to get started in genAI, you can leverage it at the application layer. Or if you want to manage your own models, you can do them deeper down the stack.

AIaaS providers offer data preparation, since it is often unstructured, as well as the training of models provided by the customer or the option to use pre-built AI models they provide. These models, trained on massive data sets, can perform various tasks like image recognition, data analysis, natural language processing, and predictive analytics.

You access the services through APIs or user interfaces. This allows you to easily integrate the AI functionality into your own applications or platforms, often with minimal programming required.

Most AI-as-a-service providers offer a pay-as-you-go model, either through metered use or a flat rate. It is much pricier than your traditional IaaS/PaaS scenario. Nvidia, for example, charges a flat rate of $37,000 per month to use its DGX Cloud service.

Is AIaaS right for your organization?

AIaaS requirements?

AIaaS pros and cons?

Leading AIaaS vendors?

For answers to these and other questions on AI please contact us at

info@eccoamerica.com

by Andy Patriz