The Enabling Technologies Making AI on the Battlefield a Possibility

By Ryan Schradin
Managing Editor at
GovDataDownload

With AI and ML poised to revolutionize how the military makes decisions, we wanted to learn more about the technologies that are making it possible and the enabling tools and solutions that will bring AI to the battlefield. To find out more about the cloud solutions and products available to the military that are opening the door to further AI adoption at the edge, we sat down with Dan Holmay, the High-Performance Computing and AI Solutions Leader at NetApp.

GovDataDownload (GDD): What role does the cloud play in artificial intelligence and machine learning initiatives? How is the cloud an "enabling technology" for these solutions?

Dan Holmay: "The cloud" is overused terminology, but it really is just a place where a lot of projects will get started or spun up in a sandbox. The cloud is a great place to experiment and get started. Cloud providers give users a few tools

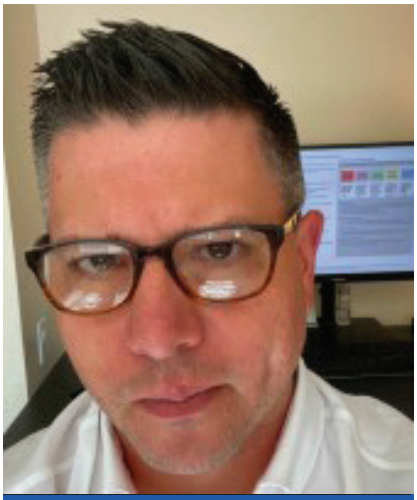
that they can get familiar with, and they also give users access to GPU capabilities. That's important for users that don't have a data center that has a large multi-node, multi-GPU cluster.

So, when it comes to AI programs and projects, the cloud is an easy way to get started and accelerate a project. However, there is a downside. While it's easy to begin projects, it can be difficult to go beyond that proof of concept and prototype. We've found that even large, very sophisticated private companies build a use case, develop a model, and train it up, but then struggle to get it to deployment.

Exposing those models to an organization's data and getting it ready for a production deployment means integrating pilots into existing systems that are typically, in many cases, on-prem. This is where we see many AI projects stall out. The model is developed and needs to be integrated with on-prem systems, but the data is in the cloud and can't be pulled out without an egress charge.

GDD: What cloud advancements, tools, and capabilities are making AI/ML initiatives possible?

Dan Holmay: A big gap at the very beginning of any project is data labeling and data conditioning. Thankfully, there are some very specific tools out there – like the ones built into some of our offerings at NetApp – that work very nicely with some of the metadata tagging and conditioning tools that are available to users of Azure or SageMaker in AWS.



Dan Holmay
High-Performance Computing and AI Solutions Leader at NetApp

What is often overlooked is that a model with "five nines" accuracy needs a very large volume of data to produce the outcomes its developers are looking for. In many cases, these users have data but not necessarily all of the information about that data that the model needs to be as accurate as possible.

"What is often overlooked is that a model with "five nines" accuracy needs a very large volume of data to produce the outcomes its developers are looking for. In many cases, these users have data but not necessarily all of the information about that data that the model needs to be as accurate as possible."

Dan Holmay

This is why data labeling is so important – although it sounds completely mundane.

In fact, many cloud providers have introduced data labeling capabilities either through their own tool sets or through integrations with other companies. Having an embedded capability allows the customer to more quickly and easily label data, allowing them to get their models up and running, and trained much more quickly.

Large-scale automation of solutions like these

is what will fuel the adoption of AI in the ”real world.”

GDD: What hardware advancements are enabling the AI/ML applications and solutions that the DoD is looking to implement?

Dan Holmay: As an example let’s say a model is built in the cloud. When the developer wants to begin training that model they need to have access to massive amounts of data and a large number of GPUs. Moreover, they need a curated set of tools to accelerate their work. Which is what NetApp worked hand-in-hand with NVIDIA to build out.

> **“As an example let’s say a model is built in the cloud. When the developer wants to begin training that model they need to have access to massive amounts of data and a large number of GPUs.”**
>
> Dan Holmay

The ONTAP AI DGX Pod is a validated reference architecture that allows users to leverage the power of multiple DGX nodes and high-performant flash data storage to reduce their model training time from weeks and months down to days or hours.

The capabilities that NetApp has built into ONTAP AI enable users to pull in the data they have stored across multiple disparate, disconnected buckets for training in a curated environment and automated tools. For example, let’s say the DoD is looking to implement predictive maintenance on a Humvee fleet. They can crunch all Humvee sensor and maintenance data from multiple disparate sources leveraging ONTAP AI to build and train that model quickly.

This accelerates what users can do on a project or experiment basis. It allows them to push terabytes or petabytes of data into an on-prem GPU cluster to train models quickly. By accelerating the training of AI models, we can enable government and military developers to experiment, fail fast, and establish beginning points to get models off the ground.

The ONTAP AI data fabric that connects the cloud to on-prem frees data scientists from having their models stuck in the cloud and allows them to access silos of the data for training. This is helping military and government users get past one of the single largest speedbumps or barriers to pushing AI programs into deployment.

GDD: How are NetApp and NVIDIA working collaboratively on behalf of the DoD? What real-world use cases will this technology enable?

Dan Holmay: One of the things that I’m very proud of, and something that we were able to discuss during a panel at AFCEA West, was a proof of concept that we did with Nvidia and a company called ProtopiaAI to account for model and data security at inference. In this proof of concept, we built out a deployed AI data center and ingest point at the edge (in-theater) for the U.S. Marine Corps at the Rim of the Pacific (RIMPAC) Exercise in Hawaii.

This solution was intended to provide increased security and insight into an evacuation command and control scenario. It provided real-time facial recognition, which in itself is not new, but it also used something called ProtopiaAI StainedGlass transformation models to allow inference to be done securely at the edge. This type of model training and inference on obfuscated data effectively made the data worthless to an adversary but still enabled models to inference and alert if the person passing through an evacuation command and control triage zone was a good guy or bad guy.

We've actually presented this to a number of different other agencies, including Army Intel and the Air Force. They're very interested in the ability to provide real-time security insight for a warfighter on the ground. But they're also interested in the larger capability that we demonstrated to do inference in near real-time and then push that data and insights back to command and control in a single pane of glass that might be in the cloud or somewhere stateside.

That capability was built on an Nvidia GPU box, a NetApp A250, and a 5G cradle point. That system had a small form factor and was incredibly mobile. It also provided a capability out of the box that a Marine could set up without any IT background in less than 15 minutes.

As we go forward, we're going to continue to work with NVIDIA to deliver more solutions with this same fast setup and decision-making insight. We want to put these highly automated solutions into the hands of a Marine or other warfighter wherever they are which will enable them to better accomplish their mission and meet their objectives.

"We've found that even large, very sophisticated private companies build a use case, develop a model, and train it up, but then struggle to get it to deployment."

Dan Holmay