MeriTalk spectro cloud HITACHI Inspire the Next

Advancing AI at the Edge to Transform Missions

Infusing artificial intelligence (AI) and machine learning (ML) into data and applications at the edge has the potential to transform operations at Federal civilian agencies and in national defense. But edge environments can be challenging to manage and come with their own set of security risks.

MeriTalk recently sat down with Cornelia Davis, technology fellow and vice president of product at Spectro Cloud; Tommy Scherer, principal architect at Spectro Cloud Government; and Pragyansmita Nayak, chief data scientist at Hitachi Vantara Federal, to discuss accelerating edge innovation and securely managing edge computing environments.

MeriTalk: More than nine in 10 Federal leaders say that edge solutions are very or extremely important to meeting their agency's mission needs. What are the key challenges that Federal agencies face as they deploy and manage edge environments at scale?

Davis: I see four key challenges: bespoke environments, scale, sporadic connectivity, and security. Bespoke environments require a lot of manual labor and customization. The solution there is Kubernetes, which has become the operating system in any environment. It helps mitigate the software-related challenges of managing and scaling edge applications and application delivery.

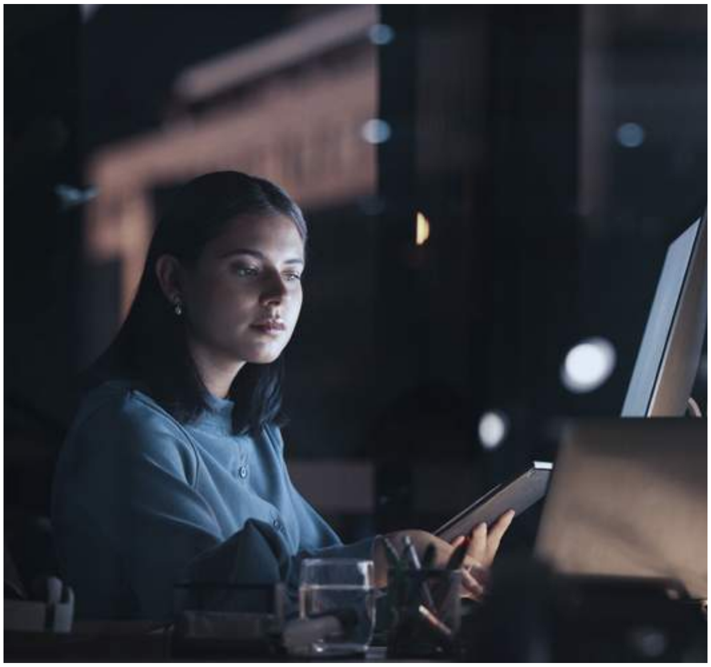
As for scale, the government might be managing hundreds, thousands, or tens of thousands of environments. The answer there is automation and centralized management. And sporadic connectivity – military applications are often in areas with no internet connectivity, but they still need to work. Application management is centralized, but control must be distributed. And lastly, devices are in the field, perhaps in hostile territory. They need to be secure.

Nayak: The key challenge, as Cornelia rightfully pointed out, is the distributed nature of working and the decentralized infrastructure. Data is generated in different formats, on different types of devices, and even on different versions of the device. You're getting your computation and your data closer to where it is needed, but data lineage and data provenance become increasingly difficult. Connectivity is another data-centric challenge. In the future, if you want to tap into the data created at the edge, it needs to come back to the cloud. Lack of reliable communications is a hindrance.

MeriTalk: Gartner predicts that by 2027, deep learning will be included in more than 65 percent of edge use cases. How is AI impacting use cases for Federal civilian agencies and the military?

Scherer: Let's look at a surveillance use case as an example. On a surveillance flight, there are only a few frames of interesting data. If you're able to pre-process that data using AI/ML algorithms, tag that data so you have the provenance, and then eventually move that into a data lake, that's where I see AI/ML at the edge really being useful, especially when you have intermittent connectivity. Another interesting use case is passive sonar in submarines. Huge amounts of data are collected. You want to be able to tag the important data, like the sounds of adversaries' vessels, and bring that back to the core or cloud at the end of the cruise, but that could be 12 months later, so you also need to be able to train your model with that data locally.

Nayak: Advanced, lightweight algorithms and models, which are becoming smaller and smaller in size and getting more efficient, making a number of evolved multi-modal use cases possible. Deep learning is extremely powerful when working with unstructured data such as audio, video, images, and when combined with structured data from IT applications, sensors, and databases, a larger data space can be navigated in various dimensions. Object detection and classification for vehicle, vessel, or aircraft maintenance saves time and money for the military branches and increases their agility and responsiveness. Document analysis for topic modeling and metadata management for search and indexing improves access to more content than was possible before, and generative AI-based applications create more user friendly, conversational interfaces for performant knowledge-driven work environments.



MeriTalk: The White House executive order (EO) on AI calls for developing standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy. How does the SENA framework for securing edge computing environments align with this?

Davis: At the highest level, the EO addresses two main priorities: making sure that the environment stays secure and hasn't been tampered with, and making sure that the AI models are good. Secure edge native architecture (SENA) addresses the first priority. It proposes a reference architecture that addresses primary security concerns at the edge – from software vulnerabilities to hardware tampering to data protection.

Nayak: Data protection applies to AI models as well. Tracking the training data, the process of formulating AI models, and data and model changes are critically important because edge computing often involves real-time data measurements that can trigger actions in the mission space.

Tracking data and models ensures that bad actors can't change a model and cause a negative influence on a system. As long as data and models are being tracked and protected, you have trustworthy and secure AI systems.

MeriTalk: How has the shift to open development accelerated agency and industry transformation around edge computing and edge AI?

Davis: Standardization definitely accelerates innovation in general, and open source has become the best means for developing standards. So, when Kubernetes came along, we finally had a standard for IT systems – it democratized infrastructure management. This open-source project has been a huge accelerant, creating a massive explosion in the number of contributors providing solutions.

The other accelerator, of course, is ChatGPT because its API has become a de facto standard for large language model (LLM) AI inference, with many other projects providing implementations. Before ChatGPT, LLMs had been around for at least a decade, but they were in bespoke research labs. When OpenAI made the API available, backed by their inferencing engine and AI model, everyone could use it. This led to an explosion in the number of implementations that provided or embedded LLM capabilities. Kubernetes and ChatGPT democratized innovation.

Nayak: When I think of open development, I also think of open architecture and how products should work in an interoperable manner. With edge computing, there are bound to be multiple products and multiple organizations all working together – and each one has its own specific data standards. Being able to interoperably work with each other's data and make API calls to get more information when required is a huge part of any architecture going forward. Open architecture also enables data discovery and analytics on connected data. This is extremely important for the future.

MeriTalk: What can Federal agencies do to address specific requirements and challenges throughout the lifecycle of edge infrastructure and AI software stacks?

Scherer: At the edge, everything's constantly changing, and we face challenges around the disconnected nature of some environments.
But updates must take place without bureaucracy slowing them down. That's where continuous authority to operate (cATO) comes into play.

With cATO, security updates and new versions of applications can be moved into production quickly without taking devices offline.

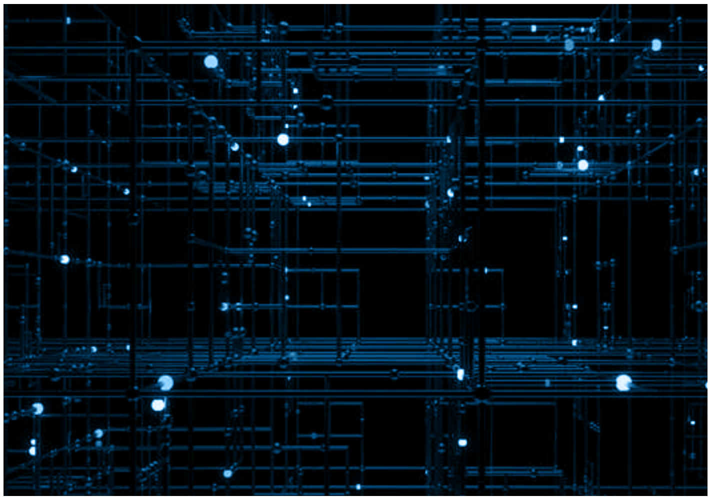
Nayak: In dynamic operational environments, it's important to understand the lifecycle associated with all this data that is being generated, processed, and manipulated. Key stakeholders need to understand that there is a much bigger purpose for the data than one use case.

MeriTalk: How are Spectro Cloud and Hitachi Vantara Federal working together to enable AI at the edge?

Davis: At Spectro Cloud, our Palette management platform enables operations teams to define, deploy, manage, and secure the underlying Kubernetes-based infrastructure that Hitachi's data platform and the agency's AI models run on.

Palette provides a single interface to model the full stack of everything that goes into a cluster, deploy at scale to many different hardware platforms, and conduct a full suite of lifecycle operations on deployed infrastructure, from monitoring to configuration changes, patches and upgrades, and troubleshooting.

We have the mechanics to ensure that we're constantly updating all of the layers from the operating system up, to support AI workloads and securely run AI models at the edge.



Nayak: Spectro Cloud provides the backbone, the infrastructure that runs our solutions on steroids. In AI/ML solutions, data management, data quality, and data discovery are all extremely important – that's what the Hitachi Vantara Pentaho solution provides. Pentaho enables model operations and ML operations to develop data-driven solutions, and Spectro Cloud makes all of this faster and more effective.

MeriTalk: How do Spectro Cloud solutions enable operating in environments that not only have strict requirements for security and compliance, but also face the challenge of constrained resources?

Davis: Many lightweight AI models are available today that can run on smaller and smaller devices, and lightweight distributions of Kubernetes can run on small form factor devices. But not every single device is the same.

In Palette, we have cluster profiles, which are like a model or blueprint defining everything that needs to be deployed as part of a Kubernetes cluster in the field, including Pentaho for data provenance, data, and models for workload cluster deployments.

We use this to securely bootstrap or spin up edge devices, even small form factor devices using different architectures, without lengthy configuration processes. All of those cluster profiles are managed centrally, but they're all controlled and executed out at the edge, so that if you're disconnected, operations can continue.

Palette also has native support for Kairos, which is a distribution-agnostic open-source project that enables upgrades to a system in an immutable infrastructure fashion. If there is a problem with an upgrade, it does not compromise what was already working before the update and can revert to the previous version, so a critical device is not taken offline.

As you can tell, I'm very excited about our partnership with Hitachi Vantara Federal and working with Federal organizations to advance AI at the edge.

> **One of the things that's really important in managing edge environments is the ability to take a very specific hardware configuration and turn it into something that can be managed in a very standardized way.**

For more information on how Spectro Cloud helps government agencies achieve their missions, visit:

spectrocloud.com/solutions/government