Feeding the Insatiable Data Needs of AI and Advanced Analytics

Federal agencies are ramping up use of AI, machine learning and advanced analytics to supercharge the accuracy of weather forecasts, improve patient triage, sniff out malicious threats, root out fraud, and a host of other applications. But sometimes, according to some data storage and cloud experts, the data behind those efforts isn't easily accessible, which leads to slower results and higher costs.

Agencies embracing AI are getting valuable results, but they could realize them much faster, for less money, especially in instances where the data is stored in one or even multiple public clouds.

While there's nothing wrong with public cloud data storage—it's ideal for backing up data as an insurance policy as well as many other applications—it isn't efficient or cost-effective for situations that involve staging millions or billions of records quickly. For example, a large medical diagnostic AI study using 100 million patient records can take days to stage data into an active state, due to bandwidth and latency challenges of public cloud resources.

The government is full steam ahead with the adoption of AI, machine learning and data analytics, so agencies need to become more strategic at storing all the related data. A recent report from the National Security Commission on AI (NSCAI) recommends that the federal government invest $32 billion per year in AI by 2026.

Recalling millions of records from a public cloud also can be prohibitively expensive, doubling or tripling project costs due to cloud egress charges alone. For example, a 9.6 petabyte public cloud would incur about three times the cost in storage as the same amount of data stored in a private cloud.

It comes down to the fact that agencies' heavy reliance on the public cloud may not be the best solution for all data, according to experts from data storage vendor Quantum and cloud service provider ViON. When dealing with large data sets that need to be accessed quickly, a better approach may be a hybrid one, with some or all data stored in an on premise cloud, they said. Robert Renzoni, a director at Quantum, explained the hybrid approach provides the low latency and fast access that AI and machine learning workflows need.

That's the route National Institute of Health's Department of Radiology and Imaging Sciences took to store its massive library of medical images, according to Quantum, which supplied a storage platform for the agency. The department currently has about 750 TB, which is growing by about 10 percent each month. By moving its growing MRI image archive to more modern, high-speed shared storage, its

PRODUCED BY: FCW | SPONSORED BY: ViON Building your enterprise solutions Quantum.

250 scientists have fast access to data for research. Those scientists also can be sure that the department's images will always be stored, protected and available when needed.

A different way of thinking about storage

To produce effective results, projects that use AI and advanced analytics require access to massive stores of data—not only current and historical data (much of it unstructured), but data from multiple locations, including edge devices. It is also very likely that much of that same data will be needed again in another capacity, for another project.

"When you're finished analyzing the data and producing results, that doesn't mean the data will never be used again," said Ray McKay, a solution architect at ViON, a cloud service provider and designer of mission-critical IT infrastructure solutions. "It's got to be accessible for the next AI project. Data now has a shelf life of forever."

The fact that data needs to be retained forever and that multiple types of data from multiple locations have to be easily and quickly accessible—requires a new way of thinking about and managing data. It makes the traditional mindset of simply putting all data into an archive or relying exclusively on public cloud resources to store archived data obsolete.

Instead of storing data based on when it was last accessed and relying solely on the public cloud, it's more effective to base data storage and access on the application itself, and on how the agency expects to use the data over time.

The new data paradigm views all data as either "active" or "temporarily inactive". This method makes all data quickly available when needed, which is especially useful when data access is unpredictable and semi-frequent. For example, if the Center for Disease Control and Prevention is conducting research on air pollution's effects on people's lungs, researchers need access to billions of records. Researchers may start with one set of 10 billion records, which move from "inactive" to "active" temporarily, while another set of 10 billion records sits waiting (currently inactive).

One of the most effective ways to move toward this new way of viewing data is by combining public cloud storage with an on-premise private cloud specifically designed to store temporarily inactive data. The private cloud connects to storage nodes at the edge, close to device sensors, and uses an as-a-service model, according to McKay.

"It's about finding a cloud solution that allows you to access all of the data you need, and that gives you the flexibility to stage your data based on your application needs quickly," McKay explained. The solution should be optimized for the amount of storage you typically need to stage, how fast you need to have the data available and ready to process, and how fast the storage can operate. For example, predicting where a tsunami will next occur after a recent earthquake leaves no time for staging, while a medical research project that starts next week provides plenty of time to stage your data, according to McKay.

The unpredictability of when data may be needed, and the speed with which it may be needed, is the reason agencies choose the service-based access method, he said. With this approach, scalability and latency are no longer barriers. Storage can be added or subtracted quickly, and data can be accessed and ready to use much more quickly. This allows agencies to respond faster to unanticipated events.

In addition, the as-a-service method allows agencies to use operating funds rather than capex budget. Costs also are lower because agencies pay only for capacity they use and avoid all egress charges levied by public cloud providers.

This method also can improve security. With a private cloud infrastructure, agencies know exactly where their data is at all times, and can secure it their own way. This type of solution also provides archiving redundancy for data durability and resilience, as well as non-disruptive upgrades and policy-based data integrity checks.

Preparing for whatever comes next

While viewing storage as "active" and "temporarily inactive" is a new model and a different way of thinking for most agencies, it is the best path toward more effective operations, faster time to value from data, flexibility to respond to unanticipated events, and more manageable costs, according to the Quantum and ViON experts. Most importantly, it allows agencies to take advantage of the benefits of AI, machine learning and advanced analytics—something federal oversight organizations are urging agencies to embrace, they said.

Most agencies today already use multiple clouds for their storage needs, and are very familiar with both public and private clouds. That expertise will be invaluable in moving to the next stage. The key, said David Kushner, a senior vice president at ViON, is finding a way to integrate new technologies and processes smoothly.

"It's about being able to optimize and balance your cloud spend and your infrastructure spend and increasing visibility while taking advantage of the most up-to-date technology via an as-a-service model so you know you're operating as efficiently and cost effectively as possible," he said.