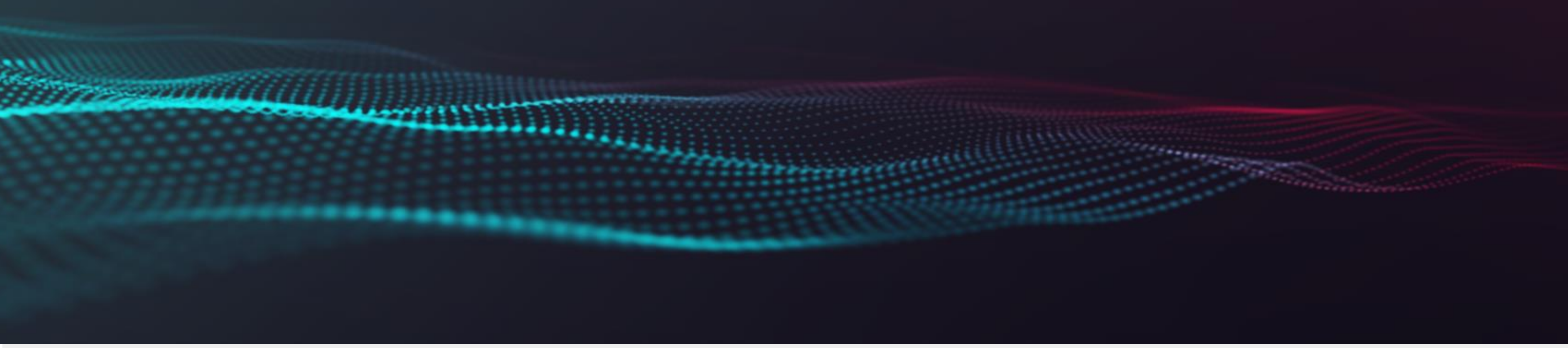
Enterprise Strategy Group | Getting to the bigger truth.™



ESG WHITE PAPER

Enabling AI Adoption at Scale with ViON and NVIDIA

Operationalizing AI with AI Infrastructure-as-a-Service

By Mike Leone, ESG Senior Analyst

September 2021

This ESG White Paper was commissioned by ViON and NVIDIA and is distributed under license from ESG.

Contents

Introduction

Adoption of AI across Federal agencies continues to gain steam. In fact, today 73% of organizations (inclusive of both public and private sectors) currently have AI projects in flight that utilize specialized infrastructure to handle their AI workloads.[1] Agencies want smarter and faster ways to gain value from their data, and AI is proven to deliver. Whether improving operational efficiency through intelligent automation and decision making, rapidly allocating compute and storage resources to effectively meet real-time workload demand, discovering insights from distributed data sets, reducing process backlog during a certain time of year, or enabling troops to deploy in times of crisis or battle safely and securely, the benefits of AI to federal agencies seem all but limitless. Hundreds of applications and use cases continue to emerge that are already yielding impressive results, including:

- Developing autonomous ground vehicles to improve the safety of soldiers and other personnel.
- Maximizing government fleet and asset uptime by using predictive maintenance from sensor data and natural language processing to analyze technician records.
- Increasing drug safety through the application of deep learning on package labeling.
- Detecting anomalies in radiology and pathology images with clinical-grade accuracy.
- Addressing environmental disasters by appropriately directing first responders in scenarios like wildfires.
- Delivering real-time conversational speed transcription and translation services for international government affairs.
- Eliminating congestion and reducing carbon emissions in smart cities through intelligent video analytics and traffic management.

While AI is delivering eye-opening improvements across agencies, a myriad of challenges continue to cause roadblocks, delays, and outright failures in achieving success. Between legacy infrastructure shortcomings, inability to cost-effectively scale AI in the cloud, delays caused by the movement of large data sets, and the ongoing skills gaps experienced throughout the entire AI lifecycle, agencies are looking for help. With the understanding that data gravity is a very real challenge, agencies are increasingly exploring a hybrid approach that brings compute to where the data resides. And for on-premises environments, whether in a data center, a bunker, or a base installation, the expectation from agencies is clear: a cloud-like experience that enables the delivery of resources in an as-a-Service model that promotes agility, speed, scale, and cost savings.

Federal AI Challenges

While systemic challenges remain ever-present across most federal agencies (i.e., ethics, policies, workforce impact, governance, etc.), infrastructure readiness is proving to be a significant issue for the wider federal space. Between skills gaps within the AI lifecycle, inadequate processing power and storage capacity in the infrastructure stack, tight budgets, and aggressive timelines, organizations are facing an uphill battle in achieving AI success.

[1] Source: ESG Master Survey Results, *Supporting AI/ML Initiatives with a Modern Infrastructure Stack*, May 2021. All ESG research references in this white paper have been taken from this master survey results set unless otherwise noted.

Skills Gaps

Nearly **1 in 3** government agencies have a problematic IT skill shortage in the area of AI.

As organizations look to ramp up AI initiatives, the first roadblock they may encounter comes in the form of existing skills gaps. While data scientist shortages continue to grab the headlines, whether due to a lack of someone on staff in that role or the data scientist on staff simply being overburdened with tasks outside of their core skillset, IT is increasingly being viewed as a problematic area. In fact, ESG research shows that nearly 1 in 3 government agencies have a problematic IT skills shortage in AI.[2] This is particularly troublesome, since 45% of organizations say IT outright owns the final decision for AI infrastructure purchases. While agencies continue their AI journeys despite AI skills gaps in IT, short-term success is quickly overcome by long-term impact that severely hinders the ability to deliver AI at scale. Higher CapEx and OpEx costs, infrastructure bottlenecks, inability to scale, and significant delays in the ability to access the right resources for a particular use case are just a few in a long list of negative impacts to agencies that hastily pursue AI without addressing AI skills gaps, particularly in IT. To address these gaps, organizations are looking for help, and one way they're finding that help is by embracing external third parties. In fact, 36% of organizations are turning to external third parties such as AI-focused solutions providers to address skills gaps or help with the selection, implementation, and/or management of infrastructure supporting AI initiatives.

Infrastructure Stack

AI infrastructure requires several optimized and tightly integrated components across both software and hardware to satisfy AI workload requirements. But for many federal agencies that are ramping up AI initiatives, this means the need for an infrastructure revamp. Existing technology (and previous investments) simply cannot keep up with the performance, scale, and diversity of AI workloads. In fact, when ESG asked organizations about the parts of their existing infrastructure stacks they believe to be their organization's weakest links in delivering an effective AI/ML environment, the top response was resource sharing (26%). Integrated development environment, processing (both GPU and CPU), and storage followed. Components stitched together in a DIY manner or even general-purpose GPUs in the cloud are struggling to deliver, especially with diverse AI workloads requiring customization in the way hardware is deployed and allocated. Another interesting component to the infrastructure stack challenge is the diversity of stakeholders that may require access to the system. These can vary from a data scientist or data engineer to an application developer or someone in IT responsible for resource allocation or maintenance. Availability of not only the system, but also the tools, technologies, and underlying data create several bottlenecks, all of which impact the time to value.

Time to Value

The need to procure the right infrastructure based on AI requirements may be the single greatest IT challenge experienced across federal agencies today. In some agencies, procuring the AI infrastructure can take months, with some agencies taking more than a year to procure the right technology. By the time it gets deployed, it's likely outdated or inefficient compared to the latest and greatest technology on the market. While significant progress has been made in the time it takes to see value from AI initiatives, ESG research shows that the average number of months an organization takes to see value from their AI initiatives is just over 8 months.

[2] Source: ESG Master Survey Results, *2021 Technology Spending Intentions Survey*, December 2020.

Data Gravity's Impact on AI Infrastructure Deployments

Organizations with AI in production are **2x likelier** to be training their models in on-premises environments.

Virtually all agencies need help in assembling the right infrastructure to support AI workloads at scale. Agencies are attempting to rationalize the public cloud due to the effortless access to GPU-backed resources through an as-a-Service delivery model, but agencies are reaching a tipping point. As agencies iterate on their AI models in the cloud, the growing complexity, requirements for increased compute cycles, and exponential data growth introduce escalating costs for tight-budgeted agencies. In fact, ESG research shows that while organizations that have yet to deploy AI into production are likelier to leverage the cloud to support most phases of the AI lifecycle (outside of deployment), those with AI in production are 2x likelier to be training their models in on-premises environments.

A key driver of this movement is the notion that federal agencies are losing the fight against data gravity. Data gravity is the ability of a large data set to attract applications, processing power, services, and other data. The force of gravity, in this context, can be thought of as the way these other entities are drawn to data relative to its mass and are driving the repatriation of AI workloads. ESG research shows that, of all enterprise workloads inclusive of AI, 57% of IT organizations have repatriated workloads from the public cloud back to on-premises environments, with at least 1 in 5 organizations citing reasons such as an inability to meet scalability/elasticity expectations, poor or unpredictable performance, and high cost.[3]

57% of IT organizations have **repatriated workloads** from the public cloud back to on-premises environments.

Data gravity is forcing federal agencies to consider other deployment options as new AI use cases emerge and increasingly land outside of the public cloud altogether. Use cases like the Air Force training small, unmanned aircraft systems in different types of virtual environments that adequately prepare them to transition into a real-world environment may introduce issues associated with security requirements that prevent data from being moved altogether. Other use cases that may also introduce these issues include using real-time location data to find deployed troops in certain locations with limited connectivity or accessibility and a defense department developing AI-powered trainers for flight simulation tests that can accurately pinpoint the moment a student fails by monitoring real-time biometric data that requires too large a data set to move to the cloud. ESG research shows that, of all the stages of the AI lifecycle affected by data movement, training is considered the area where data movement is deemed most important. Agencies are embracing the idea of AI at the edge by conducting training where their data lands, meaning moving compute to where data is generated or stored to offset the inefficiencies, time delays, and increased costs of moving TBs of data into a different environment to make faster progress on AI projects.

Agency Priorities to Enable AI Success with a Hybrid Cloud Architecture

The public cloud will continue to enable all organizations to ramp up AI initiatives quickly, especially for early adopters that don't have immediate access to a modern infrastructure stack. However, when the AI cloud tipping point is reached due to model complexity, scale, or data gravity dictating a need for on-premises infrastructure, the rapid availability of right-sized resources in on-premises environments will be essential. This is driving the need for a hybrid AI architecture to set agencies up for success. In fact, when ESG asked organizations about the most important considerations of infrastructure solutions

[3] Source: ESG Research Report, *2021 Data Infrastructure Trends*, TBD.

used to support their AI initiatives throughout the AI lifecycle, the top responses were hybrid cloud capabilities, followed by attributes to deliver fast performance, ensure data security/governance, and improve operational efficiency through faster deployment/provisioning, better hardware/infrastructure utilization, simplified resource management, and effective model management.

Nearly **1 in 3** organizations state their greatest barrier to AI success is the **need for better IT infrastructure capabilities.**

To achieve this hybrid AI nirvana, agencies are looking for new solutions that enable them to embrace data gravity, bring the compute and software stack to where the massive data sets reside, and eliminate the barrier to getting deterministic performance of on-premises AI systems, with the simplicity and ease of the public cloud. Agencies understand that they may not currently have an optimized data center to support the scalability and performance demands of AI workloads. ESG research highlights this challenge, with nearly 1 in 3 organizations stating that one of their greatest barriers to AI success is the need for better IT infrastructure capabilities.[4] A turnkey AI data center in a box wrapped with an as-a-Service delivery model can enable the consumption of the right number of resources at the right time. Agencies are looking for a building-block approach that provides access to a fully optimized platform that supports the end-to-end AI lifecycle, from development to deployment, and supports all AI workloads, including analytics, training, and inference.

The as-a-Service delivery model is becoming essential to offset operational burdens commonly experienced by agency IT staff tasked with resource delivery. This approach will empower IT to consolidate operational AI silos, simplify capacity planning, and ensure resources are optimally delivered based on workload requirements. In fact, ESG research shows that maximizing hardware/infrastructure utilization is one of the most important considerations when selecting an infrastructure solution to support its AI initiatives. Agencies want peace of mind knowing that a single power-hungry training workload will no longer starve other AI workloads or that compute resources will no longer be overprovisioned to support inferencing.

The **top challenge** or barrier experienced with AI projects is the **cost of IT infrastructure.**

Agencies recognize and ESG research validates that in the past, the biggest challenges or barriers experienced with AI/ML projects was the cost of IT infrastructure. This is a big reason why nearly 1 in 3 organizations view cost as the most important consideration when selecting the infrastructure for AI/ML.[5] To offset upfront capital traditionally required for comprehensive and powerful AI infrastructure platforms, a pay-as-you-go cost model will be prioritized to give agencies budget relief. Instead of IT spending time and money procuring, deploying, managing, upgrading, and scaling AI infrastructure, pay-as-you-go enables IT to simply consume resources based on demand. The infrastructure resources are then managed by a separate entity, such as a third-party technology vendor or solutions providers. This will not only enable AI projects to scale in a cost-optimized way, but it will also enable the right sizing of AI experimentation, in other words, minimizing waste of resource consumption for an AI project that fails to meet a hypothesis. For AI, failing fast on a potential use case or idea is proving nearly as valuable as a successful AI project. This enables AI project leaders to iterate faster throughout the experimentation phase while not consuming valuable resources that could be allocated more appropriately to more successful projects.

[4] Source: ESG Master Survey Results, *Artificial Intelligence and Machine Learning: Gauging the Value of Infrastructure*, March 2019.
[5] Ibid.

ViON and NVIDIA

For 40 years, ViON has been solving data center and data management challenges. Through partnerships with market-leading infrastructure and technology providers like NVIDIA, ViON is a leader in helping federal agencies modernize their infrastructures to accelerate IT transformation efforts. With the rise of AI, ViON and NVIDIA are working together to enable agencies to accelerate their AI journeys and capitalize on opportunities to rapidly innovate using next-generation technology. ViON delivers a comprehensive portfolio of AI-centric tools and technologies that help agencies tackle the toughest computing challenges by significantly accelerating the delivery of AI, data analytics, and HPC at any scale across any location, from a top-secret data center to the ruggedized edge.

Understanding that acquisition and procurement regularly plague government agencies and keep them from living on the bleeding edge, ViON's as-a-Service model helps expedite acquisitions of new technology, reduce procurement complexity, and remove financial barriers to leveraging a modern AI infrastructure platform. By bringing the power of a cloud operating model to on-premises environments, ViON's as-a-Service offering provides agencies with agility in attaining new technology though a flexible financial model. The results for agencies include significantly reducing time to POC, alleviating operational burdens on IT, and lowering the cost of entry within and across agencies as they scale the use of AI. Agencies can start small in experimentation sandbox development environments and scale to production environments with peace of mind knowing that resources are right-sized based on the stage of the project, the workload, or any other requirements.

AI Infrastructure-as-a-Service

In partnership with NVIDIA, ViON is delivering AI Infrastructure-as-a-Service to federal agencies. Foundational to the offering is an NVIDIA DGX™ A100 system, featuring eight NVIDIA GPUs and two second generation AMD EPYC™ processors, providing agencies a full-stack solution that is purpose-built for the unique demands of AI workloads, from analytics and experimentation, to training and inference. The NVIDIA DGX is optimized at every layer for maximum performance, including components like GPUs, CPUs, and RAM to data science libraries and deep learning framework integration. Being a full-stack, optimized solution, agencies experience productivity gains virtually everywhere. AI researchers and innovators don't have to waste time integrating, troubleshooting, and supporting hardware and software. Data scientists can confidently utilize resources across their end-to-end workflows, from development to training at scale. And by wrapping everything in an as-a-Service model with ViON, procurement, deployment, optimization, allocation, and maintenance of this powerful AI data center are offloaded from an already overburdened IT staff to ViON and NVIDIA's experts. This enables agencies to get started with AI in a controlled environment in just hours instead of months, boosting productivity and accelerating innovation with AI. Agencies can bypass the impact of data gravity by bringing compute to where the data lives, eliminating escalating I/O costs and achieving a desired and affordable compute cost model that significantly lowers the barrier to AI through the delivery of right-sized resources across all AI workloads.

The Bigger Truth

Federal agencies are the greatest producers and collectors of data in the world. They must be able to gain insight and value from that continuously growing data to be at the forefront of innovation. As programs and potential AI use cases place more demands on data center technology (and those responsible for managing, optimizing, and allocating technology resources), the increased complexity and workload requirements are forcing infrastructure change. Agencies need unprecedented computing power, dynamic scalable storage capacity, and high-speed connectivity. They need immediate access to tightly integrated, next-generation infrastructure components. They need to be able to scale rapidly as experimentation moves to POC and eventually to production. And they need all of this reliably delivered in diverse locations, whether at the edge, the core, or in a hybrid cloud environment. Federal IT leaders are being tasked with evaluating and ultimately deciding on the right AI infrastructure solution and delivery model to satisfy the diverse set of use cases being thrown at them. ESG recommends exploring the benefits of a hybrid cloud approach that offers the best of all worlds.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides market intelligence and actionable insight to the global IT community.

www.esg-global.com contact@esg-global.com 508.482.0188