FEDERAL NEWS NETWORK

EXPERT EDITION

Unleashing data insights to drive government innovation

Insights from

- Army
- Homeland Security Department
- Army Corps of Engineers

BROUGHT TO YOU BY
neo4j

neo4j

Better Serve Citizens With Connected Data

In today's interconnected world, the relationships between data points hold critical insights for improving government services, optimizing resources, and enhancing cybersecurity.

Government agencies often struggle to capture this critical context at scale.

Why? Because traditional data architectures aren't designed for today's densely connected data. Relational databases strip out data connections, forcing IT teams to recreate them manually through increasingly complex and painful JOINs.

Enter Neo4j, the world's leading graph database. A graph database stores data as nodes and relationships, which preserves the full context of data connections. You can then:

- Uncover hidden patterns for deeper understanding and innovation
- Gain instant visibility across all relevant data without complex JOINs
- Adapt quickly to new data sources and evolving requirements

Across use cases — from supply chain optimization and fraud detection to predictive analytics and enhanced citizen services — Neo4j empowers government organizations to solve their toughest challenges by unlocking the power of connected data.

The impact is clear:

- The U.S. Army optimizes equipment maintenance, reducing costs and boosting readiness
- NASA saves millions in R&D using a knowledge graph
- MITRE strengthens cyber threat detection by mapping vulnerabilities in context

Now, you can experience the transformative power of graph technology, firsthand.

Get started today by visiting **neo4j.com/govermment**.

TABLE OF CONTENTS

With data, ‘why’ matters as much as ‘how’

“Large language models are really hot right now. And everyone wants them for a variety of purposes.”

Those two phrases, uttered by the Army Corps of Engineers’ Cody Salter, aren’t terribly shocking. But what might be is that Salter and other engineers and data scientists at USACE’s Engineer R&D Center are frenetically working to give those “everyones” in the Army the LLMs they seek so they can take advantage of data faster, better and smarter.

This ebook is not about artificial intelligence, though you’ll find mention of AI in each of its four articles. The focus here is on the data, on problem solving and on results. Sometimes that will involve AI, machine learning and LLMs, and sometimes it won’t.

And that’s a bit of a tough idea to hold onto when you can barely go half a day without reading, seeing or hearing something about AI. But the end goal that drives an agency to gather, cull and analyze data always should take precedence over how any of that happens, pointed out the Homeland Security Department’s Alexandria Phounsavath.

She shared that the DHS Science & Technology Directorate’s work is “all about first defining what the problem is and understanding what the mission end users are trying to achieve. And AI/ML might come into that. But there are others. Maybe it’s a graph approach that you take — or maybe something a lot simpler — before you get to fancy AI/ML or large language models.”

So as you read the articles in this ebook, we hope that idea stays with you. It’s not about the shiny tools; it’s about achieving the shiny results.

Now, dig in!

Vanessa Roberts
Editor, Custom Content
Federal News Network

DHS researches data solutions for cybersecurity, privacy and AI

BY DAISY THORNTON

"A different way of thinking about security."

That's how Alexandria Phounsavath, director of the Data Analytics Technology Center at the Homeland Security Department, describes her team's data autonomy project.

It compliments other approaches to cybersecurity, like zero trust, "but this one is data-centric," Phounsavath said on Federal Monthly Insights — Unleashing Data Insights to Drive Government Innovation. "The way we think about data autonomy is we've developed a concept in architecture actually, and it has components."

> **To make this data autonomy concept realized, it's going to require new products and services in the future, and a lot of thinking — from all our partners in academia and industry, other government labs — to solve this problem.**
>
> — Alexandria Phounsavath, Director of the Data Analytics Technology Center, DHS

She explained that those components are essentially like fine print attached to the data, allowing a data owner to place controls on it. For example, the data may only be available for five days before it deletes itself. Or the fine print might communicate the best way to view the data, be that in tabular format or a graph of some sort. It can also include a billing element that could, for example, allow the data to be used for research for free but require companies to pay a fee.

"It's actually putting data in charge. And I'll have to say that this is an open-ended problem. We're actually doing some prototypes right now, some of those main components like I described, not the whole thing," Phounsavath said.

"But it's open-ended in the sense that to make this data autonomy concept realized, it's going to require new products and services in the future, and a lot of thinking — from all our partners in academia and industry, other government labs — to solve this problem. So I think it's a very interesting space. It requires a lot of thought, but it's really, really exciting."

That's not the only data project the Science and Technology Directorate, of which the center is part, has going on. Phounsavath said DHS S&T also is working on data privacy–enhancing technologies. Those include homomorphic encryption, secure multiparty computation and both anonymized and synthetic data.

> **Artificial intelligence and machine learning are all the rage. … It's not where we jump right away. It's all about first defining what the problem is and understanding what the mission end users are trying to achieve.**
>
> — DHS' Alexandria Phounsavath

Many agencies are interested in these technologies and are participating in privacy R&D, she said. The goal is that improvements to privacy-enhancing technologies will improve access to data, which will lead to greater benefits from that data.

DHS research: training AI with synthetic data

Some of those technologies have other applications as well. For example, Phounsavath said S&T's Transportation Security Lab is trying to figure out how to effectively use synthetic data to train artificial intelligence and machine learning–driven threat detection for both people and baggage at airports.

"Today, the problem is that you have explosives and they're dangerous to handle, and data collection is expensive, and the way that they collect data today is TSL actually hires people to kind of dress up, pack bags and walk through machines," she said. "And creating simulants for those explosives is also expensive, so we're talking about a dearth of diversified data. So we're looking at ways that TSA and TSL can use synthetic data to help improve those algorithms that detect threats."

The question S&T is working on in this use case is how to use synthetic data for this purpose. Among other things, they're trying to determine what data artifacts matter and whether the images have to be perfect. For example, when an X-ray hits metal, it creates scatter because the beams become visible. Does that scatter need to be included in the synthetic data to properly train the algorithms?

But Phounsavath also said AI needn't to be the go-to solution every time.

"Yes, artificial intelligence and machine learning are all the rage. But even before it became the rage, you had statistics, you had all these other approaches that people do, and AI and ML, it's not where we jump right away," she said. "It's all about first defining what the problem is and understanding what the mission end users are trying to achieve. And AI/ML might come into that. But there are others. Maybe it's a graph approach that you take — or maybe something a lot simpler — before you get to fancy AI/ML or large language models."

Listen to the full conversation between Federal News Network's Eric White and DHS' Alexandria Phounsavath on data research projects underway within the Science and Technology Directorate

Army shines spotlight on building data literacy, community

BY DAISY THORNTON

The Army is trying to build a data community across its enterprise. But that's no small task for such a large organization with so many disparate, specialized components.

That's why it held its first Army Data Summit in the spring at Army Forces Command in Fort Liberty, North Carolina.

One subject the summit focused on was how to spread data literacy across the Army. And according to Army Chief Data and Analytics Officer David Markowitz, it's easier to train an Army specialist like a logistician, intelligence officer or artilleryman on data practices than it is to take a data specialist and embed them in specialized service components.

> **My challenge as the Army CDAO is to create a community space where we can work together.**
>
> — David Markowitz, Chief Data and Analytics Officer, Army

"My challenge as the Army CDAO is to create a community space where we can work together and work as a team, but still folks have their local identity so they understand their local issues, their mission issues, and they can bring their mission issues to the forefront," Markowitz said on [Federal Insights Month — Unleashing Data Insights to Drive Government Innovation](). "But as a community, we can help each other, so the intel officer can help the logistician who can help the finance person with techniques. And so that is our challenge, to have the best of both worlds, and our first Army summit was to get at that issue of building that community space."

The key differentiator, he said, is that local specialists know the questions that need to be answered by the data to plan better and make better decisions. They also know the processes that generate the data.

To further support that dynamic, the summit also resulted in the creation of what Markowitz called command chief data officers, who stand at the junction between command and the end user to translate between the two what data is useful where. Because most local data isn't valuable to the rest of the enterprise, but end users may need access to some command-level data.

"So, we're trying to make sure that marketplace between consumer and supplier is greased and we've got the right folks to build on it," he said on the Federal Drive with Tom Temin.

That's also facilitating a task that Markowitz said he's concerned with: determining whether or not he's meeting end users' needs for data. One thing he said he looks at to determine that is the rate of new analytics products being created. That number is rising at a rate of around 600 to 700 new products a week, he said.

Some of those are altered versions of old products, but others are new. And he said they're getting 12,000 to 13,000 requests a day for data exports through application programming interfaces.

Tapping data to feed Army AI applications

All that data is also being used to feed artificial intelligence applications, which is something Markowitz said the Army is still trying to come to grips with. He said it's got a long history with certain types — ones directly associated with lethality decisions, like independent munitions and smart mines; and platform-specific applications, like self-navigation on unmanned vehicles — but newer, commercially available models like generative AI have varying degrees of applicability to the Army's mission.

> **What we are currently exploring is how applicable [generative AI] tools are? Can they easily be lifted, or do they need a lot of self-training? And what's hard for us to grasp right now is the full cost model.**
>
> — Army's David Markowitz

"What we are currently exploring is how applicable those tools are? Can they easily be lifted, or do they need a lot of self-training? And what's hard for us to grasp right now is the full cost model. It's easy to get a platform that's available, but all of a sudden it's like an in-store app," he said. "You start purchasing content and understanding cost control. You can really blow your own budget in a short amount of time and not realize it. So, how do you put that type of visibility on not only quality of the answer but also have cost control?"

One answer to that is to put responsibility for an AI application's output on the user, he said. But the big question is whether an app is worth the cost, labor and training.

"Most of the time, we're finding it's a nuanced discussion between how much we want to invest in training a tool versus people and labor, and what's the right mix between the two," Markowitz said. "It's often not one or the other. It's how do you get the best tooling with the best people?"

Listen to the full conversation between the Federal Drive's Tom Temin and the Army's David Markowitz on the service's push to expand data literacy

Army Corps of Engineers data scientists enable AI, analytics across Army

BY DAISY THORNTON

As large language models help federal employees conduct back-of-house business processes faster and better, agencies are trying to figure how best to incorporate generative artificial intelligence into their operations. The Army is no different, which is why computer and data scientists at the Army Corps of Engineers are working to enable those AI capabilities across the Army.

"Large language models are really hot right now. And everyone wants them for a variety of purposes. So we have a group of people who are collecting regulations or guidelines, best practices within certain communities in our organization, and putting them within an LLM, optimizing that LLM using those regulations and then being able to ask that LLM questions to help teach or inform bodies of people," Cody Salter, a research mechanical engineer at USACE's Engineer Research and Development Center, said on Federal Monthly Insights — Unleashing Data Insights to Drive Government Innovation.

"That's really an incredibly common task that we see ourselves doing right now for a variety of different collaborators and customers, because that technology in and of itself, that LLM, is just really hot and widely publicized right now. So people are really able to latch on to that. They understand what it is and what it does and how it might be leveraged within their space, so we do a lot of that."

LaKenya Walker, a computer scientist at ERDC, added there are a number of use cases the center is working on. On the civil work side of USACE, document summarization is a common ask. So is knowledge generation — using generative AI (genAI) to populate databases to support training and upskilling newer employees. That's particularly important as agencies struggle with the loss of institutional knowledge as a significant amount of the federal workforce ages toward retirement.

> **Generally, we have been in the business of providing capabilities at the front end of the data science lifecycle, if you will: the data infrastructure, getting data in a format that's conducive to doing analytics.**
>
> — LaKenya Walker,
> Computer Scientist, USACE

Using AI to supplement tasks across the Army

Meanwhile, on the military side, Walker said she sees a significant amount of interest in information summarization. Military systems capture an increasingly large amount of data, too much for human experts to understand and parse — so the Army is looking to AI to supplement them in that task.

Walker said there are two methods for preparing an LLM to work in such a specific domain like the Army. First, they can use domain-specific data to fine-tune an LLM's results, but that's more laborious because all of the data must be generated. Or second, they can layer retrieved augmented generation (RAG) over the LLM to query specific knowledge or vector databases.

"So in the RAG, what you're doing is actually taking your source data, embedding that data, or vectorizing the data, and storing it in a vector database," Walker said on the Federal Drive with Tom Temin. "So instead of your LLM using its training data that's inside of its parameters, it looks to your vector database with your domain-specific embeddings in it to make a judgment call — to give information back."

If the Army wants, for instance, to build a predictive model for the failure of a specific part on a vehicle, it would feed maintenance records into an embedding model, which vectorizes the data. Then, when someone queries the LLM, the query goes first to the RAG module, which pulls the relevant vectorized data. It adds that context to the query, then feeds it to the LLM for output.

Providing capabilities on the front end

Walker said genAI capabilities also allow users to manage the output from the LLM so it's presented in the format that's most useful or comprehensible to them. That could mean tailoring the output to a less technical audience, avoiding the use of highly specific jargon. Walker said the Army tends to particularly favor output in the form of web-based tools like dashboards.

But ERDC doesn't involve itself in the output or visualizations that often. Instead, Walker and Salter said they focus more on getting their internal customers set up with the tools and infrastructure they need to do it themselves.

"Generally, we're not the ones at the latter end of that lifecycle pipeline where people are doing the visualizations. We allow users from different organizations coming in to do their analytics and perform their own type of visualization development," Walker said.

> **We really think that data problems have an agnostic nature to them, like a data problem is a data problem is a data problem, no matter what the kind of domain or application it is.**
>
> — Cody Salter, Research Mechanical Engineer, USACE

One challenge they run into is that their customers often need multiple types of data to reach the results they need. That can mean translating outdated code from legacy applications into formats that are proprietary or support analytics in general. Or it can mean repurposing older data based on the use case.

"The data source is really dependent upon the need, like the analytic need. So when we sit with a customer or a collaborator, we always want to start with a really basic understanding of 'What is the problem that you're trying to solve?' " Salter said. "And that is what's going to then kind of dictate the various data sources that might be required to answer that problem."

Working on the front end, on the infrastructure and data formats, also allows ERDC to be more versatile. Salter said skills, tools and methods transfer across specific domains and disciplines.

"We really think that data problems have an agnostic nature to them, like a data problem is a data problem is a data problem, no matter what the kind of domain or application it is," he said.

Listen to the full conversation between the Federal Drive's Tom Temin and the Army Corps of Engineers' LaKenya Walker and Cody Salter on efforts to take advantage of large language models Army-wide

How to add graph power to your data analytics

BY TOM TEMIN

To get the most insight out of your data, you need to add a little math. Math in the form of graphing technology, that is. Graphing not in the sense of visualizations — although that's part of it — but graphing in the sense of discovering the relationships among data elements such that you can unearth hidden insights.

John Bender, regional vice president for U.S. federal at Neo4j, says data users must understand the difference between relational and graphing database technology.

"A relational database uses tables and then it uses joins to bring the data together," Bender said on Federal Monthly Insights — Unleashing Data Insights to Drive Government Innovation. "Whereas a graph looks at the data and the relationships as equal. So we can query off anything, any relationship or any node in the graph."

For how this benefits users, Bender used the example of fraud detection, something top of mind for many federal IT practitioners. By better understanding the network of entities revealed by a graphing database, evidence of questionable activity will be more visible.

"If someone's doing money laundering," he said, "you're trying to understand the network and what kind of activities they are doing for the evasion." Known commonalities of various types of fraud will form clusters under the right algorithmic queries, he said.

> **A graph looks at the data and the relationships as equal. So we can query off anything, any relationship or any node in the graph.**
>
> — John Bender, Regional Vice President for U.S. Federal, Neo4j

Bender said that the more sources of data brought into the graph database, more — and more clearly seen — commonalities will emerge.

"You're going to find as you start bringing in other sources, they have commonality in the 'nouns' or in the nodes, so new relationships start showing up," he said.

Technically, an agency would retain the databases tied to specific applications, even while contributing them to the graph database.

"We're not replacing the silos," Bender said. "They're purpose built for those applications." Instead, the silos of data are combined in a data lake, "and that's where the graph goes."

An army of data points

In one instance, the Army is using Neo4j to better understand and manage its supply chains. In moving around platforms made of hundreds of thousands of parts, "they had a hard time making sure that they had the right amounts of parts," Bender said, "because they didn't have deep understanding of what their need was." Logisticians would order months in advance and stockpile items, just in case.

Now, he said, the Army has a graph database with 8 million nodes and 21 million relationships to help supply officers model conditions, run what-if scenarios and generally improve readiness.

Law enforcement and health care benefits management also make good use cases for graph technology, Bender said.

One federal agency is using it to improve the claims process in a couple of ways.

First, "they're using the graph to find a way to make it better for the citizens, using graph to understand their journey so the agency can understand ways to improve it," he said. "They're also analyzing claims and looking for anomalies in how people are submitting the claims and looking for patterns of fraud."

In another use case, a contractor built a digital twin of an agency network. By graphing the relationships among network nodes and simulating intrusions, the client can improve cybersecurity.

You can "start to look for the combination of events and relationships and similarities to other events," Bender said. "If this event happened, I can look at the other events that had certain settings. I can take that and use it as a search parameter to other events, looking for other things that match."

In another context, Bender said, NASA has built a lessons learned graph database, ingesting decades of documentation of projects and experiments. It will help the agency avoid duplication of prior efforts while also enriching current and future ones, he said.

> **As agencies start to learn more, as they become more accomplished with their graphs, they start asking more and more connected questions.**
>
> — Neoj4's John Bender

And because graph database technology scales so easily, Bender said, "you can continually add more nodes as you go. You don't have to stop and rebuild the whole graph. So as agencies start to learn more, as they become more accomplished with their graphs, they start asking more and more connected questions."

Bender said Neo4j is architected for speed, with the graph intelligence able to run from RAM. The product can run on premise or in the cloud, or in a combination of both.

Wherever it runs, Bender noted, graph technology, coupled with retrieval augmented generation, can greatly refine results when applied to publicly trained artificial intelligence large language models. It helps the user obtain an output using specific data, rather than everything that might have trained the LLM.

"We can actually take a look at a generative AI app before the actual query or the question," Bender said. The resulting answer "will come through the [agency's own] graph using the inside data."

With genAI, he said, "the whole goal now is to make information more accurate."

Watch or listen to the full conversation between the Federal Drive's Tom Temin and Neo4j's John Bender on how graph technology can help speed predictive analytics