

The

Intersection

of

DATA STRATEGY

and

Modernization

INSIDE:

- Leveraging AI and LLM at Agencies 3
- Infographic: Prioritizing Data Modernization at the CDC 9
- How Health Agencies Use Data to Improve Patient Care 15

SPONSORED BY

 **Progress[®] MarkLogic[®]**

From the writer's desk



Ross Gianfortune, Managing Editor

Data Modernization Requires New Methods

The information age is driven by data, from the data feeding generative artificial intelligence to the vast troves of citizen data guiding decisions. Across agencies, across levels of government and with non-governmental partners, officials need to embrace new ways to manage data.

In the case of public health, improving care depends on electronic health records being shareable and actionable across different organizations. With initiatives underway to move to a common electronic health record, data

modernization and governance take on new importance. In other cases, the use of artificial intelligence is becoming more widespread at agencies, making data classification, interoperability and security at the forefront of officials' concerns.

The amount of data that people produce increases each day and agencies need to harness that data to better fulfill their missions. New approaches to security, management and application will be critical to data modernization in government. ✨

Table of Contents



Jayla Whitfield,
Staff Writer/
Researcher



Sarah Sybert,
Senior
Researcher

ARTICLE

Government Agencies Leverage AI Tools and Large Language Models

Generative AI tools are powerful, officials note, but can come with a series of challenges.

BY JAYLA WHITFIELD

INFOGRAPHIC

How the CDC is Prioritizing Data Modernization

Launched in 2020, the CDC's Data Modernization Initiative is an effort to bring together health data across public and private partners. The DMI prioritizes five pillars at the center of modernizing core data across the federal and state public health landscape.

PARTNER INTERVIEW

Centering Data to Make it Usable and Actionable

Modernization efforts can depend on the ways organizations implement governance and emerging technology.

Matthieu Jonglez, VP Technology, Application & Data Platform – Progress | MarkLogic

ARTICLE

Health Leaders Target Data Exchange to Improve Patient Care

Health IT leaders from SSA, FEHRM and VA discuss challenges and next steps for cross-government health data interoperability.

BY SARAH SYBERT

Government Agencies Leverage AI Tools and Large Language Models

Generative AI tools are powerful, officials note, but come with a series of challenges.

BY JAYLA WHITFIELD

Artificial intelligence (AI) and machine-learning technologies are expanding into government applications as agencies begin to use the powerful resources to fulfill their missions.

While large language models (LLMs) like ChatGPT, OpenAI and Meta can be valuable, agencies officials said they are unable to use the models at scale due to security issues.

“The models are so large that bringing those models inside of our secure environment is probably cost prohibitive and also the company has not given us access to the models themselves,” said Lt. Col. Joseph Chapa, chief data and AI officer at the Department of the Air Force, at an industry event.

Chapa said that large language models are not currently certified for the service’s level of data so there are several restrictions in place.

“Everything we do with the systems has to be completely unclassified and



non-sensitive level,” Chapa said.

But Chapa said the department is hopeful that smaller open-source models will provide tools that are secure for classified data.

In the meantime, agencies are also working internally to combat security issues and create a secure environment for data. Last year, for instance, the U.S. Coast Guard created a climate data branch that focuses on building new infrastructure to run AI and LLM applications.

“What I find most exciting right now is we are a little late to the game, but we also didnt have to learn all the lessons that the early adopters have learned,” said Jonathan White, the Cloud and Data Branch Chief at the Coast Guard, during the summit.

White said large language models are only a slice of the AI pie and the agency is extremely “AI-curious” because it understands how resourceful the technology will be. (ctd.)

Jonathan White

Cloud and Data
Branch Chief,
Coast Guard



“With a limited workforce, it is extremely hard to place people in the right spots all the time, then we have a lot of policy and regulations and rules and everything else that goes along with those missions. Wouldn’t it be great if we could have a tool that can elicit an answer to your question that relates to those policies and regulations?” White said. “If they’re in the middle of an inspection or middle of warning [and] they need to ask a question, wouldn’t it be great to have a digital assistant to surface the information that they need to conduct that inspection.”

However, experts said they can’t rely entirely on the technology because there must be a human in the loop to avoid errors.

“This is no different than how we conduct business in general. When we create a new project, we have to have reviewers, when you do any kind of research you have research reviewers,” said David Na, a data scientist at NASA. “We need reviewers for the content that get generated by these large language models.”

As government agencies find value in large language models, Chapa said that they are aware that challenges may persist. “I know people are using these models on their own — at some point there will be some slips or some leaks,” he noted.

But Chapa said the best use of text generation models is when the basic information is already known, so it is easy to validate the facts.

“It’s going to be tempting for analysts to want to use the tool to gain insights about the world. And I’m not saying we shouldn’t do that. I’m saying we should recognize that if we’re going to use these tools in that capacity, we have to build institutional fact-checking layers on top of that,” Chapa said.

Despite the challenges that large language models may create, agency officials said they are excited about the enhanced search experience, accessibility of data and the other possibilities AI tools will provide. 🌟

“With a limited workforce, it is extremely hard to place people in the right spots all the time...Wouldn’t it be great if we could have a tool that can elicit an answer to your question that relates to those policies and regulations?”

— Jonathan White, Cloud and Data Branch Chief, Coast Guard

How the CDC is Prioritizing Data Modernization

Launched in 2020, the CDC's Data Modernization Initiative is a \$1 billion-plus effort to bring together health data across public and private partners. The DMI prioritizes five pillars at the center of modernizing core data across the federal and state public health landscape



PARTNER INTERVIEW

 Progress[®] MarkLogic[®]

Centering Data to Make it Usable and Actionable

Modernization efforts can depend on how organizations implement governance and emerging technology.

 **What is most important in making your data future proof and reusable for a future mission environment?**

Jonglez You need to prevent the monolithic legacy structure of the past. You need to include data as-is, acknowledging that not all data are rows and columns and that the need for a more dynamic environment now exists. Those include knowledge models and linked data. You also know that data is required to be conditioned and not locked down.

You must take a data-centric view: What is the data? How will it be used and operationalized? Don't be constrained by how it is laid out today, as this is often the consequence of years of application-centric architecture.

Adopt a data-centric mindset. This is not going to replace your applications, they are there for a good reason, and the data they hold and the way they hold it is critical to the success of those



Matthieu Jonglez
VP Technology,
Application & Data Platform,
Progress|MarkLogic


“By linking together data from several of those application silos and joining them together with semantic and operational models, you can turn data into knowledge.”

**— Matthieu Jonglez,
VP Technology,
Application & Data Platform,
Progress | MarkLogic**

applications, but let your data thrive outside of those silos.

Centralizing those existing silos may mean handling completely different types of data (documents, maps, temporal data, etc.), which explains why they are separate: specialization. Only a true multi-model data platform can solve this data architecture challenge.

By linking together data from several of those application silos and joining them together with semantic and operational models, you can turn data into knowledge.

 **What are the steps being taken to help decide what is most important about data?**

Jonglez This is a very loaded question. Data can only be important if you understand enough about it: its provenance and lineage, its bias, and more generally, the context of its creation, its distribution and its inclusion in the data set.

Providing traceability of data collected in the field or data coming from partners can be extremely challenging. It is critical to model the bias and assumptions that data may contain to ensure they are considered later, and we correct for that bias in decision processes. This not only requires all the best data governance practices, but also requires semantic and ontology modelling to capture the organization interpretation of the context, and deep enrichment and harmonization of the data, so you can start tracking the entities and objects described in the data.

This is also where data fabric and knowledge graphs come in, to capture the rich interconnections between those objects and entities. Decisions are driven by the interpretation of those relationships, rarely by the data itself.

Is the progress around data modernization going in the right direction? Is that getting organizations closer to AI implementation?

Jonglez Now that you have a rich knowledge graph or an operational data hub, this forms the natural source of knowledge for machine learning and AI initiatives. The data is enriched along multiple dimensions. Those can drive similarity search to provide large language models (LLM) and other AI tech with the data clusters needed to operate.

LLMs and ML/AI technologies more generally must be controlled — even more so in a highly regulated and sensitive environment — to ensure they are used in legal and ethical ways. The MarkLogic security model defines what the ML algorithms can and cannot see. Provenance, lineage and capitalizing on the metadata management strategy can trace back any LLM answer back to the facts they are rooted in.

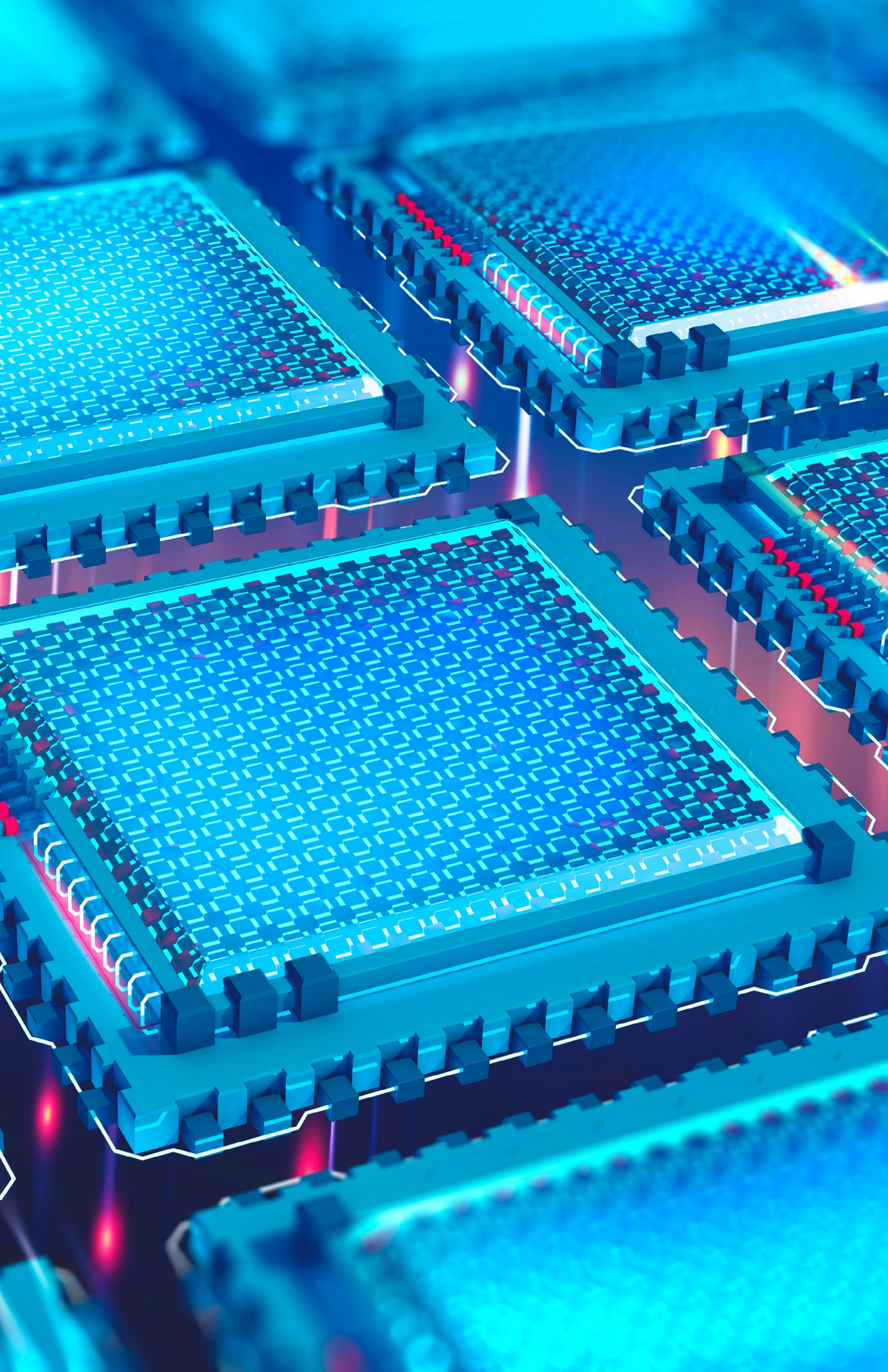
How do you ensure reliability of data through data stewardship?

Jonglez Data needs to be presented as a single source of truth. The goal of modernization is to ensure minimal impact to those who know and those who use the data. With proper federated stewardship and a reliable set of tools we obtain reliable data and knowledge and minimize both risks and costs.

Often in ML and AI projects, the temptation is to hope that the algorithm will do all the data handling and data preparation. The reality couldn't be more different. Data quality, data enrichment and all the goodness of mature data governance and knowledge management practices — metadata management, provenance, lineage tracking — are what gives ML and AI projects a chance to deliver quality results.

Data stewards play a critical role. Instead of each project trying to interpret the app data in its own way and risking misinterpreting and only using partial






information, the data stewards are ensuring (through modelling, validating classification, etc.) that the data held in the graph is a true, accurate and traceable representation of the source data and that they are responsible for ensuring that this stays true over time.

What are the challenges associated with modernizing data?

Jonglez Scale. Why are we building a modern data architecture? Because we are building new digital services and applications, and the current app-centric data architectures are at a breaking point in terms of costs, in terms of data quality, in terms of risks associated with incorrect data interpretation and partial data landscapes are hiding critical information.

Machine learning and generative AI provide fantastic new avenues to serve data and orchestrate the data-human interaction, but the quality — and all the attributes that make data reliable — are essential to ensure the conclusions are explainable and to ensure that they can be cross-checked against a single source of truth to keep the hallucinations in check. 

Enable Your Data Strategy for Mission Success



Simplify Complex Data & Achieve Data Agility

Get a faster, trusted way to unlock value from complex data – and achieve data agility to meet mission requirements now and in the future. Our unified data platform combines a multi-model database and semantic AI technology to provide a comprehensive data layer with strong security, integration, and scale.



LEARN HOW WE SUPPORT THE DOD & IC AT
MarkLogic.com/national-security

Health Leaders Target Data Exchange to Improve Patient Care

Health IT leaders from SSA, FEHRM and VA discuss challenges and next steps for cross-government health data interoperability.

BY SARAH SYBERT

Data interoperability is a key component to effectively implementing solutions such as the joint federal electronic health record and making faster, more accurate decisions to improve patient care. The Federal Electronic Health Record Modernization (FEHRM) Office, Department of Veterans Affairs, Social Security Administration (SSA) and the Centers for Medicare and Medicaid Services (CMS) are looking at new ways to standardize data and processes through key components including governance and data management.

This is especially important with the massive cross-government initiative to move to a single, common electronic health record.

“We’re trying to make sure that we do deliver, and we work toward some standardized practices within the department,” VA Deputy CIO of the Electronic Health Record Modernization Integration Office (EHRM-IO) Laura Prietula said on a panel at the 2023 HIMSS conference. “How do we bring it back to our core so that we’re not delivering and deploying 172 different versions of the product again? ... That way it will make it a repeatable process, and people can now expect what it is that you’re going to get at your medical center.”

VA is building a longitudinal platform to enable the agency to consume, exchange and use data at any time in the point of care.

“We want to make sure that your data is accessible, usable, computable, and it helps and does something,” Amanda Cournoyer, VA’s director of data and



interoperability at EHRM-IO, said at the conference. “We’re moving to all standardized. We want FHIR APIs, we want open access, we want to be able to talk and exchange.”

CMS is improving interoperability by increasing and advancing health care data exchange functionality to better inform decision-making for patients and providers, support and improve patient care, and reduce the administrative burden on providers and payers.



**Amanda
Cournoyer**
Director, Data and
Interoperability EHRM-IO, VA

“CMS is continuing to advance our goals to make health care data flow more freely and securely among payers, providers and patients by continuing to lay foundation to foster a more connected, modern health care system,” a CMS spokesperson told GovCIO Media & Research. “We are working toward this goal.”

Standardizing data into a computable format is key to enabling data-sharing and using other emerging technology tools such as natural language processing (NLP).

“Just having the data is not all the same. It’s not created equal,” FEHRM’s Acting Technical Director and Solutions Integration Director Lance Scott said at the HIMSS conference.

SSA has around 1 million pending disability cases at the initial claims level. The agency’s Assistant Deputy Commissioner in the Office of Analytics, Review and Oversight Joe Lopez is investigating better, faster ways to gather and analyze data to improve decision-making on disability cases.

“Health information technology is one of our solutions that we need to leverage to make sure that we make decisions as quickly as possible... we have to look for every efficiency when we have health IT, so we can quickly request the medical evidence and receive it back within minutes or seconds,” Lopez said. “The Office of Analytics is looking to gather all the data that SSA has to make data-driven decisions. You can’t make those decisions if the data is stored in all kinds of different locations, where you don’t have access.”

SSA only receives around 14% of its structured evidence electronically. SSA is leaning on leadership buy-in and partnerships to collaborate with EHR agencies and receive information quickly.

“We as an agency have to ensure that our leadership buys into this idea that we have to do everything that we can to work with our partners across the electronic health record organizations to receive that evidence electronically, so that we can use automation to go through it, to sift through it and call out the information that’s important,” Lopez said.

“We want to make sure that your data is accessible, usable, computable, and it helps and does something...We’re moving to all standardized. We want FHIR APIs, we want open access, we want to be able to talk and exchange.”

**— Amanda Cournoyer, Director
of Data and interoperability
EHRM-IO, VA**

Looking ahead, health agencies are focusing on improving patient experience through new technology tools. The FEHRM recently launched a new pilot focusing on NLP. Scott said the Longitudinal Natural Language Processing Pilot will enable users to conduct an NLP search on the data, so when the patient comes into the clinician’s office, the clinician already has all the documents related to that issue ready.

“It makes the patient experience that much better,” Scott said. ✨