How to Take a Page from DoD's Data Strategy

RESEARCH BRIEF

MarkLogic®

govloop

Introduction

Agencies wanting to become more data-centric might look to the Defense Department (DoD) for ideas.

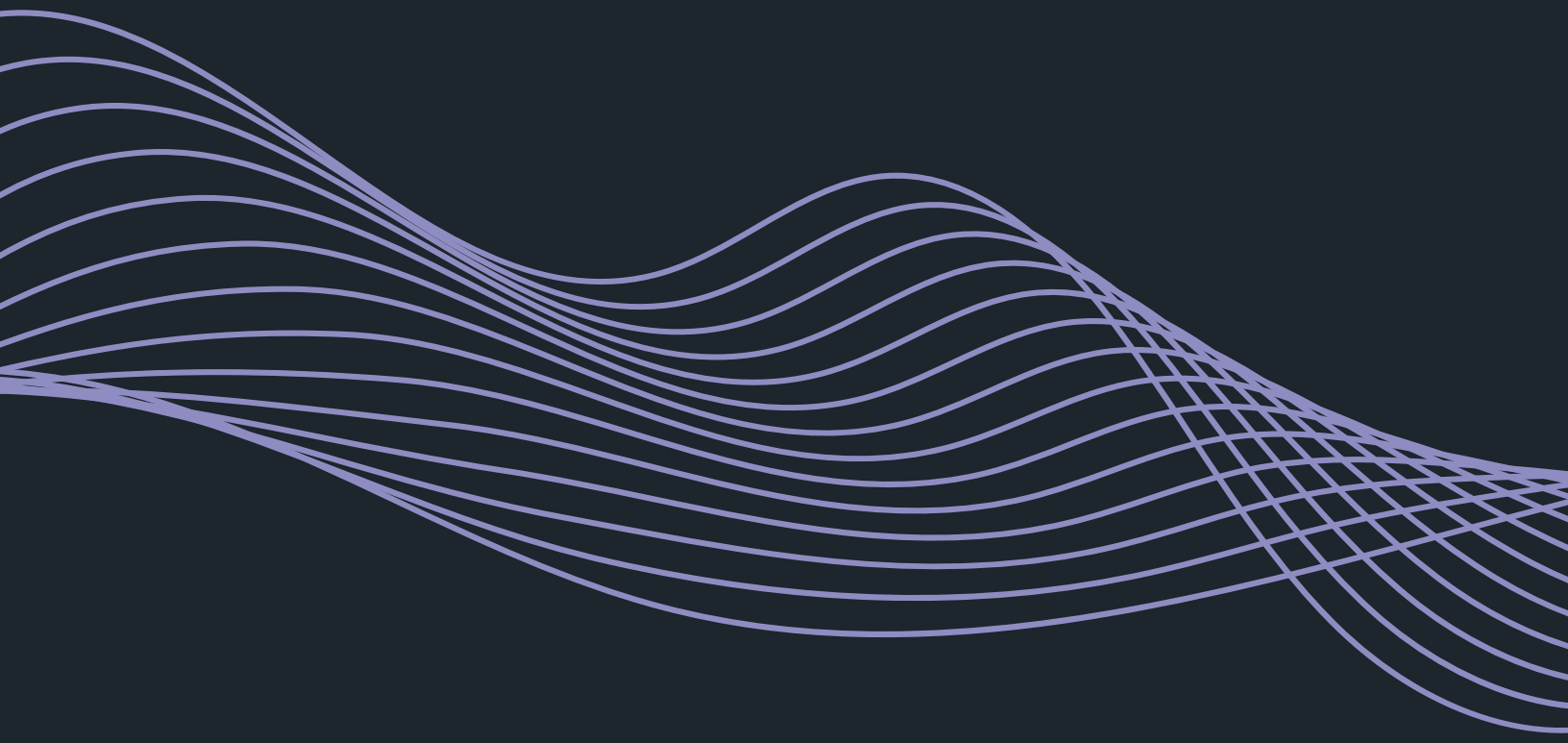
While the federal government has its own initiative in the Federal Data Strategy and many states also have developed plans and made progress, DoD is further along than many. It's putting words into action, using the guidance in its recently published DoD Data Strategy. The document outlines the department's essential capabilities, focus areas, goals, guiding principles and vision necessary to transform DoD into a data-centric enterprise.

Its goals represent those of all agencies: improving agency management, gleaning better insights, making more informed decisions and improving service to constituents. In practical terms, that means making data:

- ***Visible*** so consumers can locate what they need
- ***Accessible*** so consumers can retrieve it
- ***Understandable*** so consumers can recognize the content, context and applicability
- ***Linked*** so consumers can exploit data elements through innate relationships
- ***Trustworthy*** so consumers can be confident in all aspects of data for decision-making
- ***Interoperable*** so consumers have common representation and comprehension of data
- ***Secure*** so consumers know that data is safe from unauthorized use and manipulation

To learn more about how agencies can use the DoD Data Strategy as a blueprint, GovLoop teamed with MarkLogic, a data management software company with deep experience in the public sector. This report explores how agencies are revamping their enterprise data strategies and what lessons they can take from DoD's approach.

Please note: Some charts do not add to 100 because of rounding.



Treating Data as a Strategic Asset

DoD is committed to managing its data as a critical part of its overall mission. By not treating it as a separate commodity, the department expects to make faster, better-informed decisions. A recent survey from MarkLogic and GovLoop found that most federal and state agencies have similar goals, with nearly all agreeing that they must use data in a way that brings both immediate and lasting advantage to their respective agencies' missions (see Figure 1).

In reworking the department's data priorities, the DoD Data Strategy emphasizes the value of collective data stewardship, which assigns data stewards, custodians or even Chief Data Officers (CDOs) to be accountable for data throughout its life cycle. Data stewards are responsible for overseeing datasets, and they manage policy related to their datasets, what systems have access to the data, and how the data is tracked and accounted for.

This is a growing but still somewhat untapped area for other government agencies; although most considered it a priority, about 25% aren't focused on it today (see Figure 2). But there are exceptions: More federal than local agencies have CDOs, along with the occasional state agency.

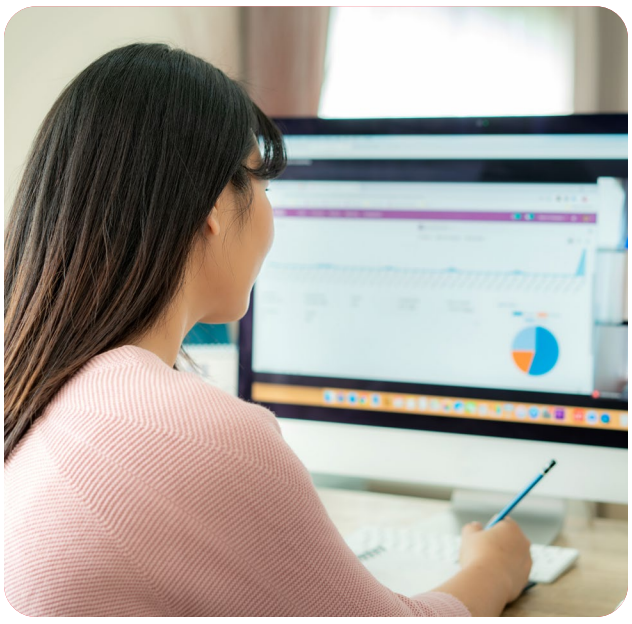


Figure 1: **Data is a strategic asset. Data must be leveraged in a way that brings both immediate and lasting advantage to the agency's mission.**

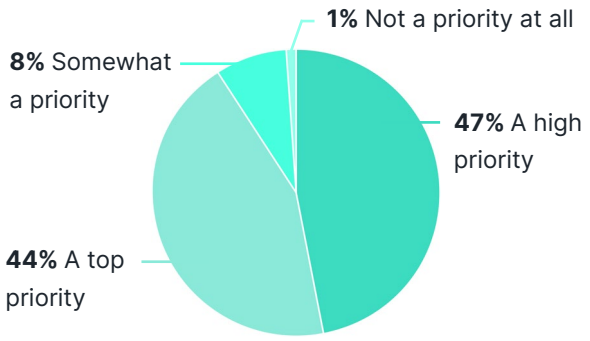
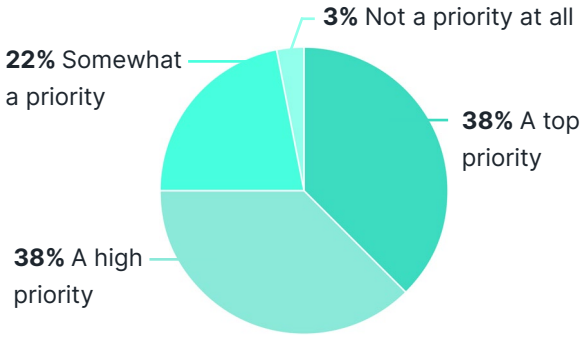


Figure 2: **Collective Data Stewardship: An agency must assign data stewards, data custodians, and a set of functional data managers to achieve accountability throughout the entire data lifecycle.**



Putting Data First

Over the years, agencies have tried to put data first, with varying degrees of success. For example, many have invested in technology such as data lakes to centralize data and improve access. Although data lakes have some benefits, the data in them is often raw, stored in its native format, uncategorized or ranked by prioritization. This creates significant challenges when it comes time to extract value from the data.

One way to become more data-driven is by adopting modern approaches and technology for integrating and managing data. A platform that can ingest data from any source, master and enrich the data, and index it for query and search is a good start. That system can also store metadata — data about the data — alongside the data itself.

"With the metadata stored as a first-class citizen with the data itself, you'll have a lot of important information — where it is, where it came from, who has touched it, etc. — to give you the context you need," explained Kim Kok, Vice President of Sales for Public Sector at MarkLogic.

Data context and flexibility are critical. Different users or departments might need to look at data in different ways, and those needs will evolve. For example, users might need to incorporate data from back-office and mission functions, which they historically have treated very differently. This approach also gives credentialed users the flexibility to change data without shutting down or redoing an entire schema. Credentialed users can change the data in place, the type of metadata being collected on the data or even the data policies.

Case in Point

With flexibility and effectiveness in mind, the Centers for Medicare and Medicaid (CMS) chose this approach for its HealthCare.gov site. CMS, a part of the Health and Human Services Department, understood that it needed to put data front and center to build a secure, effective technology platform to help enroll millions of Americans in new health care plans.

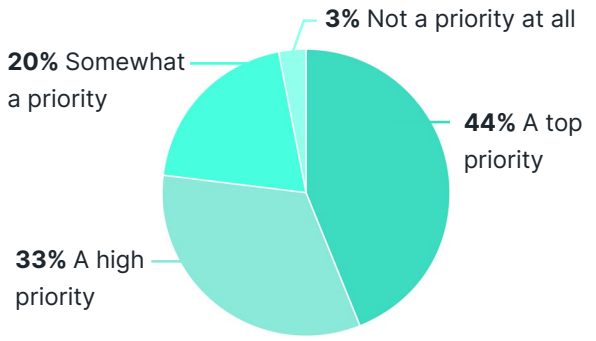
The data was voluminous, complex and included multiple data sources, such as insurance companies, Internal Revenue Service records and state-based legacy systems. The system had to be accurate, fast, scalable and secure. The project, which is the largest personal data integration project in the government's history, met its goals by using an enterprise data hub. The data hub ingests data as-is, while also accommodating any changes or additions to data as they are made, in addition to changes in policies or regulations. Today, it supports 160,000 concurrent users with 99.9% availability.

"Data is very complicated today, and it definitely helps to have a data hub where all data, along with contextual and historical data, all lives alongside each other," Kok said. "There is no better way to get a 360-degree view of your data and know that it's secure, available and usable."

Access and Availability for All

Data collection and organization are important steps in becoming more data-centric. For data to truly be useful, it should be accessible to everyone who needs access to it. Without access to the right data, agency leaders and employees won't have the information and perspective they need to make effective decisions. Respondents to the GovLoop survey agreed, with more than 77% considering it important for agency data to be available to users as needed (see Figure 3).

Figure 3: **Enterprise-Wide Data Access and Availability: Agency data must be made available for use by all authorized individuals and non-person entities through appropriate mechanisms.**



Getting the right information to the right people might seem easy, but privacy and governance, along with security concerns and policies, can complicate the issue. These complexities are probably why the DoD Data Strategy put it this way: "DoD data must be made available for use by all authorized individuals and non-person entities through appropriate mechanisms."

"The goal is for people to have access to the data they need to do their job when they need it, but at the same time, you need to make sure that only people with the right credentials can access the data and prevent people without permission from doing so," Kok said.

The first step in ensuring safe accessibility is by harmonizing data from different sources into a data hub that serves as an authoritative data repository for cross-functional discovery. An enterprise data hub can serve as a real-time, data-centric interchange supporting analysis, discovery and operations throughout the data life cycle.

Agencies can take several steps to ensure that only users with the right credentials can access the data, including:

- Ensuring that any platform you use incorporates technologies such as role-based access control, where each user is assigned appropriate roles, each associated with specific privileges and permissions. Privileges govern the creation of documents and execution of functions, while permissions indicate what users can do with a document, such as update, read, insert or execute.
- Ensuring that your platform conducts frequent security checks, which verify the necessary credentials before granting the requested action.
- Using Attribute-Based Access Control and Policy-Based Access Control, which can further restrict access. Restrictions are based on attributes such as time of day and location, with policy information stored in document metadata, or with labels representing high or low levels of trust.

These methods are particularly helpful when multiple stakeholders are involved. For example, some federal agencies must share data with state and local agencies in areas such as health care and first response. Although sharing is critical, it's equally important to ensure that only authorized personnel can access specific datasets.

Case in Point

Secure access was a big concern for the U.S. Air Force Research Laboratory. It had a long list of access challenges for its multiple exabytes of data, which had to be securely accessible and shareable by 700 scientists and engineers, along with thousands of external collaborators. The data was spread across silos and inaccessible without using specific software required for each silo. To solve the problem, the lab built the HyperThought data management platform based on enterprise data hub technology. Today, collaborators can securely discover and share information, add new data sources, change data models, and increase data volumes.

"Data access is a tricky thing, and the right technology, along with good policies, can go a long way toward ensuring that the right users get the right data at the right time," Kok said.

Agencies across the spectrum understand the need for good data governance. The GovLoop survey found that most prioritize the need for the frameworks, metrics, oversight, policies, principles, and processes required to effectively manage data at all levels (see figures 4 and 5).

Data governance is a priority, but it's not an easy thing to do well. One way to help ensure effective data governance is by using a data hub platform with built-in governance tools, which helps simplify implementation and ensure compliance even as environments and data change.

Figure 4: **Governance: Agency data governance provides the principles, policies, processes, frameworks, tools, metrics, and oversight required to effectively manage data at all levels, from creation to disposition.**

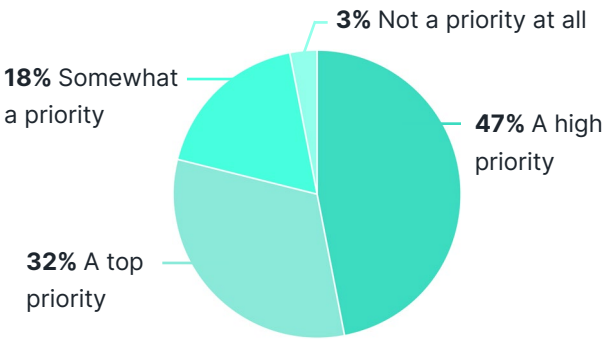
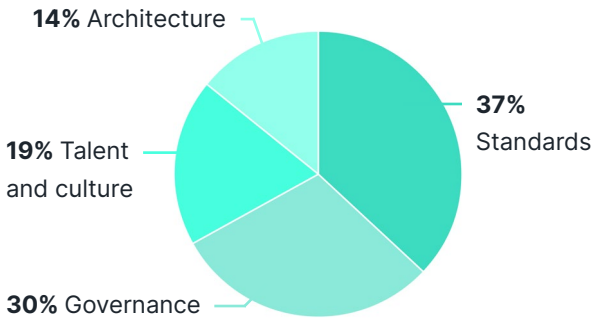


Figure 5: **Of the four capabilities highlighted in the preceding questions, what is your #1 priority for 2021?**



As Always, Security Comes First

Security is, and always will be, the top priority for any government data initiative. Making data secure was the top priority for respondents to the GovLoop survey, over other important priorities, such as making data accessible, interoperable, linked, trustworthy and understandable (see Figure 6).

Figure 6: **What are the goals for your agency in 2021? Respondents ranked the options in order of priority. Scores reflect cumulative distribution of rankings.**

The DoD Data Strategy also puts security front and center, emphasizing that data must be secure while at rest, in motion and in use. It sets a high bar, with eight objectives to make data truly secure. They include:

- Granular privilege management
- Data stewardship
- Approved standards for security markings, handling restrictions and records management
- Use of classification and control markings and content and record retention rules
- Data loss prevention technology
- Limiting access/sharing to authorized users
- Use of access and handling restriction metadata
- Ability to fully audit the access, disposition and use of data

A key part of security and governance is trust, which includes knowing the source of data (provenance) and how it has been processed along the way (lineage), both of which should be captured in metadata. Kok said that the challenges with traditional tools are that metadata is often lost in complex code used to extract, transform and load data; only accessible to advanced technical users; or unable to be tracked at all.

Case in Point

Security was top of mind for the Defense Technical Information Center (DTIC), an agency responsible for national security. It needed to ensure that its 30 sources of unstructured and structured data were discoverable, enriched, secure and shareable. Using an enterprise data hub, DTIC implemented a three-pronged security approach: securing application programming interfaces to ensure the validity and qualifications of the person requesting access, securing data at the document and element levels, and encrypting the data.

Data Collection

Data security and trust start at the collection stage. The DoD Data Strategy emphasizes the need for agencies to enable the electronic collection of data, as opposed to manual entry, at the point of creation and maintain the pedigree of that data at all times. As noted earlier, 85% of survey respondents agree with that strategy (see Figure 7).

Collecting data and keeping it secure applies to all data, whether manmade, machine-to-machine, artificial intelligence or algorithmic. Tracking which algorithms have been applied to a dataset, whether there was a human in the loop, and whether the resulting data was an output from a trusted machine or process can help users make secure decisions about the data and the system. Another key is minimizing the number of touchpoints, Kok said.

The agency architecture, enabled by enterprise cloud and other technologies, also must be as secure as possible. It should allow stakeholders to change the way they use and protect data faster than adversaries can adapt (see Figure 8).

That requires an agile and flexible way of integrating data — one that allows users to quickly pull data together in new ways for new purposes, while ensuring that the data is kept secure. It can be enabled by a data integration and management platform that includes a secure, modern database; a full stack of security capabilities; and technology that passes federal security hurdles, such as the Common Criteria for Information Security Evaluation.

A full stack of data security capabilities should include:

- Fine-grained security at the data, metadata and relationship levels, including element- and property-level security
- Attribute-, policy- and role-based access control
- Encryption in motion and at rest, including support for external key management systems
- Redaction
- Auditing
- The ability for data platforms to integrate with commonly used tools such as Lightweight Directory Access Protocol (LDAP), Kerberos, and Security Assertion Markup Language (SAML)

Figure 7: **Data Collection: The agency must enable electronic collection of data at the point of creation and maintain the pedigree of that data at all times.**

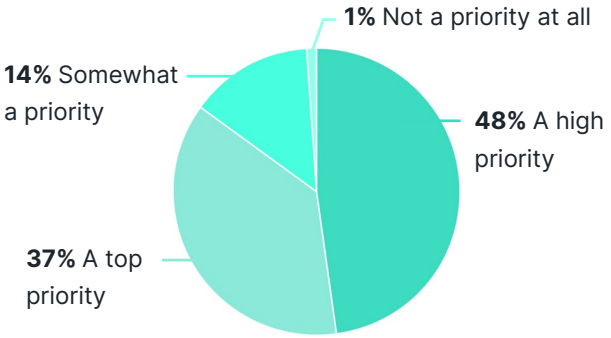
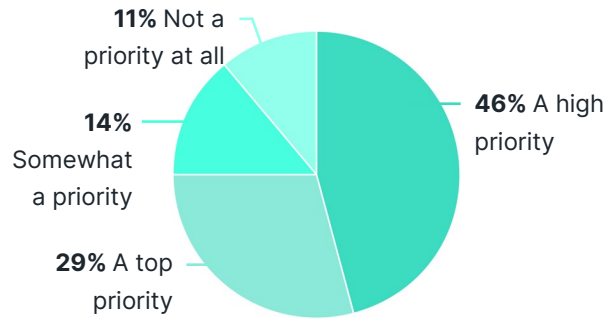


Figure 8: **Architecture: The agency architecture, enabled by enterprise cloud and other technologies, must allow pivoting on data more rapidly than adversaries are able to adapt.**



"Using a centralized data platform to govern and secure the data ensures that data is secured at its core," Kok said. "That makes securing applications easier and faster, and allows the work of data governance to happen in one place. One change in data policy at the database level can be automatically applied to 100 applications."

Kim Kok
VP of Sales, Public Sector
MarkLogic

How MarkLogic Helps

Legacy approaches for integrating and managing data are too complex and data is stuck in silos. MarkLogic provides a unified data platform that unlocks value from all of your enterprise data with more agility and security than ever before.

The MarkLogic Data Hub platform is a unified solution that brings all of your multi-structured data together and curates it for both transactional and analytical purposes. It works by ingesting data and metadata as-is from any source, indexing it for immediate query and search, and curating it through a process of harmonization, mastering and enrichment. Many public-sector agencies at all levels use MarkLogic to simplify data integration, create a 360-degree view of data and promote secure data sharing.

The platform delivers data agility and reduces architectural complexity and costs by bringing together a comprehensive set of capabilities: access, analytics, curation, data integration, management, mastering, search, security, semantics and storage. MarkLogic Data Hub is powered by MarkLogic Server, a multi-model, NoSQL database that meets enterprise requirements for scalability, security and transactional consistency.

To learn more, visit the MarkLogic Public Sector solutions page.

Conclusion

The DoD Data Strategy is a good blueprint for agencies that want to become more data-centric and treat data as a strategic asset. The strategy lays out what it takes to make the most of data, at speed and scale, for operational advantage and increased efficiency.

Although the strategy stresses the importance of cultural and process changes, most of its guidance requires the right technology to achieve goals related to data access, availability, collection, compliance and security. Public-sector respondents to the recent GovLoop survey largely agree with DoD's focus and priorities.

Achieving these goals requires a 360-degree view of all data and ensuring that all data is usable, current and accurate. It requires standardizing the approach and technology for integrating and managing data — ideally, one that can enrich, index, master and normalize all data and metadata.

The guidance calls for comprehensive and targeted security that ensures only credentialed users can access, edit, delete or add to data. Incorporating technologies such as role-based access control, along with a full stack of security capabilities and processes for providing effective data governance, is the best way to ensure data safety and accessibility. A flexible data platform that provides a comprehensive view, tracks all changes and applies security and governance is a step in the right direction to ensuring effective data governance.

About MarkLogic

Data integration is one of the most complex problems IT challenges and our mission is to simplify it. The MarkLogic Data Hub is a highly differentiated data platform that eliminates friction at every step of the data integration process, enabling organizations to gain agility, lower IT costs, and safely share their data.

Organizations around the world trust MarkLogic to handle their mission-critical data, including top banks, pharmaceutical companies, publishers, U.S. government agencies, and many more. These organizations rely on MarkLogic to discover new medicines, run the world's financial systems, prevent terrorism, and much more.

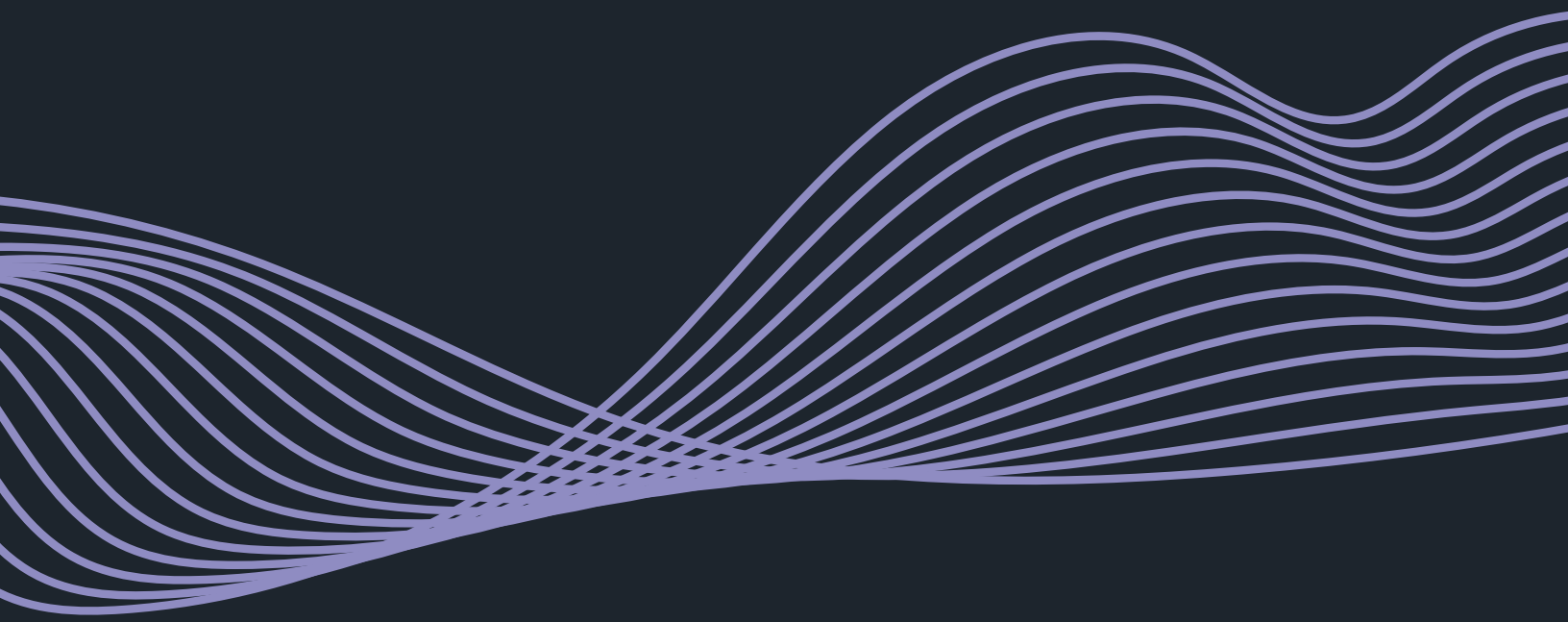
www.marklogic.com | @MarkLogic

About GovLoop

GovLoop's mission is to inspire public sector professionals by serving as the knowledge network for government. GovLoop connects more than 300,000 members, fostering cross-government collaboration, solving common problems and advancing government careers. GovLoop is headquartered in Washington, D.C., with a team of dedicated professionals who share a commitment to the public sector.

For more information about this report, please reach out to info@govloop.com.

www.govloop.com | @GovLoop



1152 15th St. NW Suite 800
Washington, DC 20005

P (202) 407-7421 | F (202) 407-7501
www.govloop.com
@GovLoop