

A Data-Centric Approach: A Priority For IT Modernization

In government IT today, there are few more important initiatives than IT modernization. It's serious business, bolstered by large IT modernization budgets and supplemented by an additional \$1 billion from the Technology Modernization Fund. In the center of modernizing is Data as DOD has rolled out their Data Strategy as part of the National Defense Strategy.

There are many approaches to modernization, including gradual replacement of technology, moving more workloads to the cloud, adopting more digital processes. Yet many of these projects won't make it to the finish line without data as the centerpiece. Data is the center of everything: analytics, artificial intelligence, and collaboration. All of this requires data sharing, which can help accelerate modernization.

“By starting with data, you’re starting in the center of the challenge,” explained Bill Washburn, Federal Chief Program Officer, at MarkLogic, a Silicon Valley technology company that works extensively with government agencies and systems integrators on modern data solutions. “It’s a good place to start when you’re looking to solve big problems that involve complex data.”

Despite the evidence that the data-centric approach works, many agencies, along with the suppliers and systems integrators that provide solutions for them, continue to focus more at the user level. While it’s important to understand user demands, requirements and challenges, it’s a mistake to make that the centerpiece of IT modernization.

“If you don’t start with the data, you lose focus on the strategy. You won’t be able to really understand how users will adapt to all of the data they have, and you’ll miss critical items,” Washburn said. “You’ll lose focus on what is really needed to address the challenge. It becomes a risky approach.”

Government at all levels are emphasizing that a data-centric approach is the way to go. The [Federal Data Strategy](#) and [DoD Data Strategy](#), as well as the data strategies from many states, make this clear. These strategies essentially provide a roadmap for getting there, along with technology insertion targets and capability baselines. They also help ensure appropriate compliance, security and data governance. It’s an ideal way to start the conversation of moving toward a data-focused future, both for agencies and organizations working with government, like vendors and systems integrators.

One of the best ways to ensure a data-centric approach to IT modernization is by gathering all data into a single, secure, searchable repository that can be shared across the department, agency or municipality. This approach using a “data hub” provides access across the enterprise via a vast array of tools users may choose from. With this approach, agencies can better analyze the data to root out fraud, waste and abuse; improve cybersecurity and logistics; and power research projects.

That’s the approach the U.S. Marine Corps took to IT modernization of several Logistics Systems. One of the biggest roadblocks was trying to find a way to integrate aging, siloed systems that still relied on their original hardware and software. It soon became clear that the only effective path forward would be to first migrate and stabilize the data from those legacy systems, and then consolidate it by creating a



“By starting with data, you’re starting in the center of the challenge.”

—Bill Washburn, Federal Chief Program Officer, at MarkLogic

unified data hub. The project, which took only a few months, enabled the USMC to better share data and improve overall efficiencies.

Other agencies at all levels of government also are beginning to understand the importance of data-centric policies and programs. The Food & Drug Administration (FDA), for example, uses a complex knowledge management system for its Center for Tobacco Products (CTP).

As part of its mission, the FDA CTP funds and uses scientific research to better understand tobacco products, how they cause death and disease, and how to best reduce the harm from these products. CTP wanted one database to hold all their documents from different data sources (both internal and external) and types—and enable those documents to be curated for use by multiple disciplines of users with disparate needs. CTP is using MarkLogic as the data platform powering CIRDS (CTP Integrated Research and Data System).

Taking the steps now to create a data-focused infrastructure is not only the best way to approach IT modernization, but the best way to prepare for whatever comes next. Recently appointed Federal CIO Clare Martorana agrees.

“Much of the data we have in government is siloed in systems, and is oftentimes not accessible, even across an agency, let alone across the entire federal enterprise. We must continue to reduce these barriers, while still focusing on the privacy and security that our customers expect,” she said at the ACT-IAC’s Emerging Technology and Innovation virtual conference. “We can make the most of government data by making it easily shareable, adopting common standards, and improving data management.”

DATA: FUEL FOR THE AI FIRE

Argonne National Laboratory is spearheading a project that will allow x-ray experiments or electron microscopy studies to be conducted virtually. The U.S. Citizenship and Immigration Service uses a chatbot named Emma to answer questions on immigration. The DoD's Joint Artificial Intelligence Center is working with the U.S. Transportation Command and Defense Logistics Agency to improve supply chain and inventory management.

On the state and local side, according to news reports, the Atlanta Fire Rescue Department has built predictive analytics software that can recognize buildings with a higher likelihood of fire incidents. The New York City Department of Social Services created an online portal to help streamline the process of providing Supplemental Nutrition Assistance Program (SNAP) resources for about 70,000 people per month.

“The data problem is only going to get more complicated, not less.”

—Bill Washburn, Federal Chief Program Officer, at MarkLogic

These are just a few of the hundreds of projects throughout government that rely heavily on artificial intelligence, machine learning and data analytics. Using these technologies effectively can identify trends, understand patterns and improve insights, leading to better and faster decision-making.

They all do different things, but, at their core, the success of these projects hinges on one thing: comprehensive, accurate data. It also means being able to understand, trust, share and protect that data. Much of that data already exists in agency systems, but it often resides in disconnected data lakes, swamps and other repositories.

Data quality is critical to the effectiveness of analytics, Artificial Intelligence (AI) and machine learning (ML), and should be part of any data process associated with these activities. After all, the results are only as good as the data being used. Ensuring good data quality requires some prep work, which includes collecting, labelling, cleaning and organizing data to understand its lineage and value.

Once the prep work is done, the next step is integrating the data into a single, secure, searchable repository shared across the department, agency or municipality. This approach gives agencies the source material they need to best use advanced

tools. This allows agencies to organize and present data in a way that AI and machine learning can understand, with helping ensure the correct sources and connections.

An effective data hub approach also provides other database analytic and AI capabilities, such as targeting, alerting, location or movement data, finding anomalies, and positioning and understanding objects. Consolidating data, either in the cloud or within an organization, also helps with data visibility across the board. Finally, it helps provide validation and checks to help understand what a particular analyst or system is doing with the data.

“Having all of the right data in one place, all adaptable for the tooling you have available to you, gives you knowledge of what that data is enabling and what can be extracted or done with the data,” said Bill Washburn, Federal Chief Program Officer at MarkLogic.

Agencies across the board are beginning to understand the importance of data quality and consolidation for higher-level analytics and AI. The DoD's Joint Artificial Intelligence Center (JAIC), for example, is currently working on its Data Readiness for Artificial Intelligence Development services contract, which focuses on helping Defense agencies prepare data for AI applications.

Readying data for advanced tasks that include modern technologies like AI, ML and advanced analytics requires this type of data-centric approach, and JAIC knows that.

They are on the right track, Washburn said. “Their request really puts data in the center of things. They know it's the best way to ready data for the challenges of AI and ML; make sure the data is secure, enriched and integrated; and get the value they need from it,” he said.

Using the data hub approach works well in many other areas as well. Take the example of healthcare, with its massive number of data sources and regulations about what can be shared, and who it can be shared with. This type of challenging environment is tailor-made for the data-centric approach. Think about HealthCare.gov, a data hub with information supplied by government agencies, insurance providers, and citizens looking for insurance coverages in a secure, collaborative way. The peace of mind for stakeholders who know that their data is only shared when needed and when approved by them is only possible because of HealthCare.gov's data-centric approach.

“The data problem is only going to get more complicated, not less,” Washburn added. “That means more analysis, more AI, and more machine learning. Getting it right simply requires better data preparedness, quality and consolidation.”

DATA GOVERNANCE: A TOUGH NUT TO CRACK

Many agencies today have a Chief Data Officer (CDO), whose main responsibility is to ensure that government data is inventoried, well-organized, high quality, and effectively and ethically managed.

For many CDOs and the agencies they represent, the end game is preparing data so it can be used for data-driven insights across their agency. For the most part, they have made progress. One recent [report](#) found that nearly three-quarters of federal CDOs have implemented some type of data governance board, one of the requirements of the Federal Data Strategy's 2020 Action Plan.

Despite these positive steps, good data governance remains a tough goal, requiring attention to details. For example, how can you ensure that the data is accurate and current? What is the best way to determine who has access, why they should have access, or decide whether they can share what they have access to with others? How can you ensure secure data sharing?

These are ongoing concerns for every agency, yet there are ways to address these issues. One way is with comprehensive data technologies and strategies that provide granular controls. One example of a granular control is a security profile that might allow for data visibility but prevent the ability to write or edit. Another is allowing different levels of administrative privileges and user access. All types of compliance must be granular, stipulating exactly what is accessible by who, and when. That's especially true when it comes to highly sensitive

data sets that contain financial or medical data.

Traceability is also important. With all data in one data hub, it's much easier to keep track of data and determine which data was used in which models. This type of data lineage verifies accuracy and ethics, as well as the ability to check data against data models.

Ensuring that the information itself is accurate and current is another challenge. Data, of course, is only as good as its source and the information about the source, its history and its purpose. That information that provides information on the data itself, called metadata, should accompany the data wherever it travels.

Tracking the provenance and lineage of the data and its metadata can help improve data accuracy.

Data security, of course, is the most important part of the equation. With a data hub approach, security can be attached to the data instead of being attached to the application or through access. Some platforms also provide other security capabilities like masking and access control, which can help secure data environments. With the right security approach, it's easier to know exactly who has access, why they have access, what they have access to, what level of access they have, and what mission requires access to the data.

"Security isn't only about access; it's also about the ability to share, what to share and what not to share," said Bill Washburn, Federal Chief Program Officer at MarkLogic. "When you have security built into your data, you don't have to point to your data as the flaw."

Embedded security can secure down to the finer elements of an electronic document to ensure the very essence of what can be shared and identifying what needs to be protected. It's especially critical in situations where more than one agency is involved in a project. It allows more movement of data from multiple systems, even if some of the data requires higher security. This is a requirement of many government systems, especially in the Defense Department and Intelligence Community.

The other part of the data equation is the ability to create policies around data that limit and allow sharing and protection as required by the agency ensuring the system can adhere to and enforce these policies. It should also include a way to create policies for how data is updated and changed. These are critical for every agency, and the ability to create these policies should be a minimum for any vendor or systems integrator. "It's not a grey area," Washburn said.



"Security isn't only about access; it's also about the ability to share, what to share and what not to share."

—Bill Washburn, Federal Chief Program Officer, at MarkLogic

HOW TO AVOID COMMON MISTAKES WHEN ADOPTING A DATA-CENTRIC APPROACH

While moving to a data-centric approach can be more effective, productive and secure than older data management methods, it's not without its challenges. MarkLogic Federal Chief Program Officer Bill Washburn, who has extensive experience with customers taking this approach, provides some advice on things to avoid:

Ignoring data quality. Not all data is as clean as it should be, and data quality is a challenge all agencies face. When quality is suspect, so are the results of projects that include that data. While some issues of data quality are unavoidable, others are preventable. Using a system that combines the Extract/Transform/Load (ETL) can help, because part of the ETL process checks for and corrects errors of the incoming data. Better yet, replace ETL and its long development effort and clumsy process with a NoSQL multi-model database like MarkLogic. That allows the system to harmonize and enrich the data as well as the associated metadata, to improve data quality and usability.

Using data sets that are too small for good decision-making: If you don't have enough data, you can't make good decisions. While it's not always possible to enlarge data sets to acceptable levels, it is possible to make the data you have as effective as possible. You can do this by using good data stewardship and data ethics and maintaining data currency.

Using the wrong tools for the job. There are an overwhelming number of options for integrating, managing and analyzing data. If you use the wrong tool, it can be like using a hammer when a wrench would be a better fit. While it's tempting to simply add more tools to the arsenal, that approach can create confusion and increase expenses. Instead, take the time to define your needs and make the right decisions.

Undervaluing your data. It's fairly common to undervalue data, relying instead on the capabilities of the systems processing that data. Without cultivating, managing and continually improving data, though, your outputs won't be accurate or current. You can avoid this by putting data front and center at all times in the enterprise.


Assuming that everything is tracking accurately. It's not only access that requires both trust and verification but data requires it as the treasure in the enterprise. Don't assume that all the data you need for a project is accessible to everyone, ensure those who need access have it and those who don't do not. Adopting a data-centric approach and validating your

data sets and needed access before starting a project is an important risk mitigator.

Relying on too many dashboards. Everybody wants a dashboard, and users tend to customize those dashboards for their own needs. But too many dashboards can result in too many different views, which can cause confusion about the data that is being viewed. The solution is converging multiple dashboards, essentially providing "a single pane of glass" that is a fit for everyone in the organization based on their needs.

Using low quality or incomplete data. There are plenty of sources of low quality or incomplete data, like data transported over a poor communications line or data from an unreliable third-party source with pieces missing. Using this kind of data to make decisions can lead to bad outcomes. Data management systems should be able to identify faulty or flawed data, alert the user and be able to refine it and improve it while alerting the user with metadata tracking.

Relying solely on open source for a proprietary solution. While open source technology starts out free, it often ends up far from free as it requires customization (using thousands of development manhours) into a solution. These types of proprietary solutions often lock agencies into long drawn out development and maintenance. Instead, choose a Commercial-off-the-shelf (COTS) solution, which can get government agencies 70-90 percent or more of the way to their needed solution out of the box. The COTS approach is much less expensive and can often also be customized to get agencies to where they need to be through the last mile of development. The result can be delivery of the required solution within a year rather than five or even ten years using open-source.



"Data is a precious thing and will last longer than the systems themselves."

—Tim Berners-Lee,
Inventor of the
WorldWide Web

DATA: A LONG-LASTING COMMODITY

Tim Berners-Lee, inventor of the World Wide Web, was fond of saying that “Data is a precious thing and will last longer than the systems themselves.” In the years since he made that remark, it has proven true over and over again.

Every organization in the world today depends on data, and that data is only growing over time. For government agencies, that’s definitely true; one [report](#) found that the public sector expects data growth at a factor of 3.5 over the next five years.

It’s what agencies do with all of that data that counts. Instead of thinking of massive data growth as a burden, agencies should think of it as an opportunity—for insights, progress and improvements.

Gaining those benefits means finding a way to gather, manage, share and secure all of that data. Using the data hub approach is an effective way to do that. The MarkLogic Data Hub platform, for example, keeps track of all data, wrapping it in granular controls for data governance and security purposes, and making it available for transactions and analysis. It does this by ingesting data from a variety of sources, indexing it immediately, enriching, harmonizing, and mastering it. It’s powered by a multi-model NoSQL database and meets agency requirements for scalability, security and consistency. Today, nine of the federal government’s 15 major agencies use it in some capacity.

When starting on the journey of better managing data, MarkLogic Federal Chief Program Officer Bill Washburn recommends having a plan, understanding the requirements, and sharing those requirements with industry well ahead of time.

“Let industry know what you’ve done so far, and what you envision in the future,” he advised. “Then listen to industry, and don’t close off possibilities you haven’t considered or how a capability in the commercial world can be applied inside your government world. They can be very similar and not just used to assure product capability but applicability as well.”

When considering a start-from-scratch approach, Washburn said it’s important to understand that it could take hundreds or thousands of person-hours to catch up to a com-

mercial product. In contrast, a commercially available off-the-shelf product is likely to satisfy up to 90 percent of requirements out of the box.

Once implemented, the data hub will take some care and feeding. It’s definitely not a “set it and forget it” solution but it will reduce your manpower requirements.

“Part of the data-centric approach is keeping on top of things,” he said. “You still have to lift the lid every once in a while and check. So make sure that your chosen solution gives you that ability—things like provenance and data lineage, which allow you to compare your original data to all of the maturations that have occurred through processing and transactions.” Give your data the best solution you can to ensure your enterprise success.

Conclusion

What MarkLogic has shown is that the success of modernization efforts is greatly increased when taking a data-centric approach. The goal of modernization, after all, is to provide better data—whether it be for deep search and complex query, building and delivering applications, or delivering insights to power analytics and machine learning.

Risk is ever-present in any modernization effort. Challenges will persist and solutions will need to be adapted. MarkLogic is the only product available to be agile and adaptable while reducing risks associated with the next generation of challenges with data.

Think first about the direction your enterprise is headed and prioritizing the importance data plays. Then choose wisely to accelerate your modernization—ensuring you select a platform that can best manage complex data. With MarkLogic as your data-centric solution, whether in the cloud or on-premises, you can modernize your data approach and ensure you stay modern.

Finally, don’t get caught up in the open source craze thinking it is low cost when really it is high risk—a cost no one wants to pay.