Where the US Government Modernization Journey is Headed

A Survey on the State of the US Federal IT Environment

Discovery Report

August 2022

Commissioned by

databricks

451 Research

S&P Global
Market Intelligence

About this paper

A Discovery paper is a study based on primary research survey data that assesses the market dynamics of a key enterprise technology segment through the lens of the "on the ground" experience and opinions of real practitioners — what they are doing, and why they are doing it.

About the Author

James Curtis

Senior Research Analyst, Data, AI & Analytics

James Curtis is a Senior Research Analyst for the Data, AI & Analytics Channel at 451 Research, a part of S&P Global Market Intelligence. He has had experience covering the BI reporting and analytics sector and currently covers Hadoop, NoSQL and related analytic and operational database technologies.

James has over 20 years' experience in the IT and technology industry, serving in a number of senior roles in marketing and communications, touching a broad range of technologies. At iQor, he served as a VP for an upstart analytics group, overseeing marketing for custom, advanced analytic solutions. He also worked at Netezza and later at IBM, where he was a senior product marketing manager with responsibility for Hadoop and big data products. In addition, James has worked at Hewlett-Packard managing global programs and as a case editor at Harvard Business School.

James holds a bachelor's degree in English from Utah State University, a master's degree in writing from Northeastern University in Boston, and an MBA from Texas A&M University.

Introduction

Despite significant advancements in cloud computing technology and infrastructure along with modern technologies for storing, processing, managing and analyzing data, the U.S. federal government has yet to fully harness these technologies. While good progress has been made, there is still work to be done, according to a recent survey conducted by 451 Research.

This report, based on a recent survey fielded among U.S. federal government IT personnel, many of whom either make or influence IT purchasing decisions or are responsible for technology deployment, examines the general state of the U.S. federal government IT environment, looking specifically at the expected benefits and challenges of deploying analytics-based systems. Further, this report looks at the data management landscape within the U.S. federal government, including data challenges and the efforts to manage and govern data, which can directly impact the delivery of analytical insights.

Key Findings

- Public cloud adoption – for any kind of workload – is high for U.S. federal agencies, which aligns well with the government's Cloud Smart strategy.
- Data lakes are commonly deployed environments for analytical workloads, and government agencies should expect their data lakes to evolve while providing flexibility for a variety of workloads.
- Data lake pain points consist of security, privacy and data quality issues, as well as the ability to ingest data, access and integration with existing systems.
- Machine learning is the primary use case for data lakes, followed by analytics and operational and data engineering workloads.
- Data management pain points consist of security, privacy, and the volume and variety of data.
- Security is a top priority and mandate for all federal agencies.

The Current State

The U.S. federal government is three years into a 10-year initiative known as the Federal Data Strategy to help accelerate the use of data to better the public interest while instituting security, privacy and confidentiality protections for that data. In addition, the U.S. federal government is well into a multi-year cloud computing strategy known as Cloud Smart. This cloud strategy is meant to drive cloud adoption within federal agencies, though the Cloud Smart strategy does not mandate specific technologies, products or approaches. Which technologies to implement, how to manage data, and strategies for managing these systems must still meet the requisite federal government standards and align with the larger context of cloud adoption and the overarching goal to follow the Cloud Smart directive.

With this context, it would be disingenuous to recommend a specific course of action without an understanding of the current state of the U.S. federal government IT environment with regard to data, data management, analytic systems and overall alignment of the Cloud Smart strategy – all of which we address in the following sections.

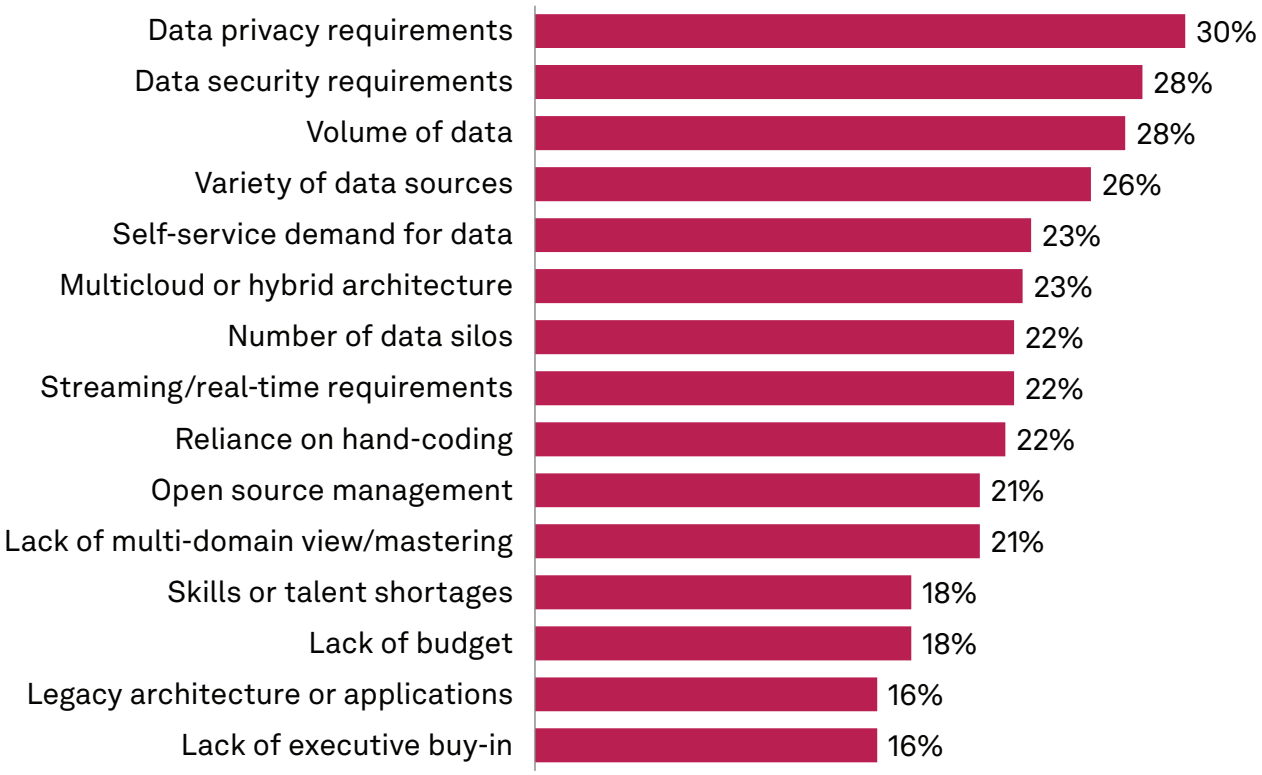
The Role of Data

Data lies at the heart of the U.S. federal government. Results from the survey suggest that not only does the U.S. government have considerable data under management, but it also values that data to make strategic decisions. For instance, nearly 57% of respondents have 100TB-1PB of data under management, with 41% of those respondents reporting they have 500TB-1PB of data under management.

Achieving a Unified View of the Data

This data is also spread out, according to survey results. For example, 41% of respondents have 11-50 data silos. Another 27% have 51-100 data silos, and 16% cited 101-500 silos. Almost a third (30%) of respondents cited data privacy as the number one concern, followed closely by data security data requirements at 28% (see Figure 1). Volume of data, considered part of data management practices, was also cited by 28% of respondents.

Figure 1: Challenges Facing a Unified View of Data



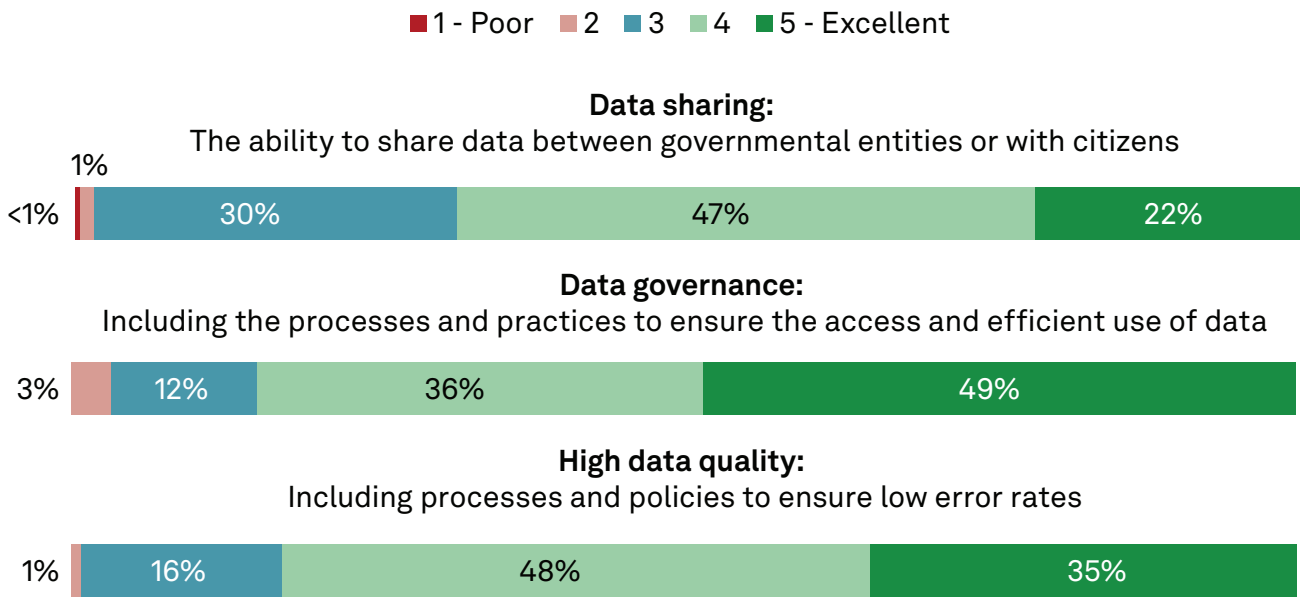
Q. What are the biggest challenges facing your agency as it tries to create a more unified view of data?
Base: All respondents (n=250)
Source: 451 Research custom survey 2022

The Benefits of Good Data Management

Data management – which addresses data sharing, data governance, data quality, data privacy and similar data-specific management operations – remains top of mind for federal agencies. Data management is the natural next step in driving value from data, and for good reason. Data lies at the heart of the systems that federal agencies maintain.

Despite some of the inherent complexities and challenges of managing data, respondents are optimistic about their agency's success with data management. When asked specifically to rate their agency's current practices for data sharing (sharing data between government entities and/or citizens), data governance (ensuring access to and efficient use of data), and data quality (ensuring low data error rates), respondents generally provided favorable responses. Nearly half of respondents gave data governance the highest rating of "excellent," while data sharing and high data quality each received near excellent marks from about half of respondents (see Figure 2).

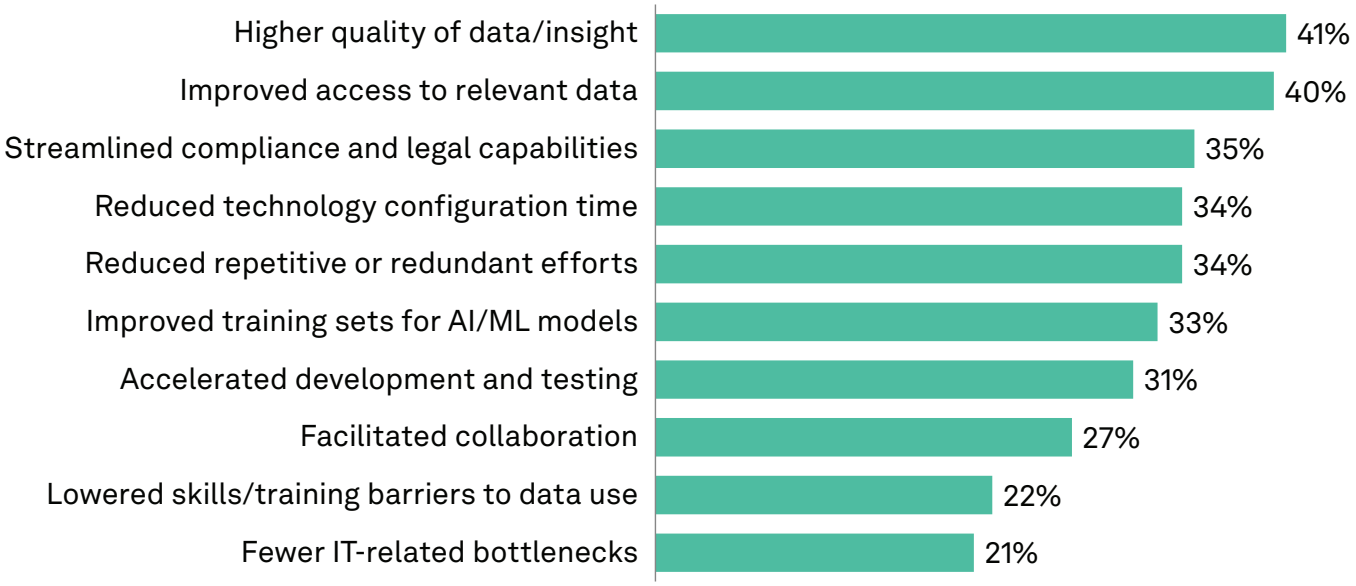
Figure 2: Rating of Data Management Practices



Q: How would you rate your agency on each of the following attributes? Please use a 1-5 scale where 1 is poor and 5 is excellent.
Base: All respondents (n=250)
Source: 451 Research custom survey 2022

A follow-up question on data governance adds further insight (see Figure 3). When respondents were asked how data governance initiatives add value, 41% reported that higher quality of data was the leading indicator, followed by improved access to relevant data (40%). In essence, the federal government greatly values easy access to high-quality data. While not a guarantee, easy access to high-quality data can lead to greater trust in the data, which can potentially lead to better and deeper insights when that data is used for analysis.

Figure 3: Added Value from Data Governance Initiatives



Q. How have data governance initiatives added value to your agency?
Base: All respondents (n=250)
Source: 451 Research custom survey 2022

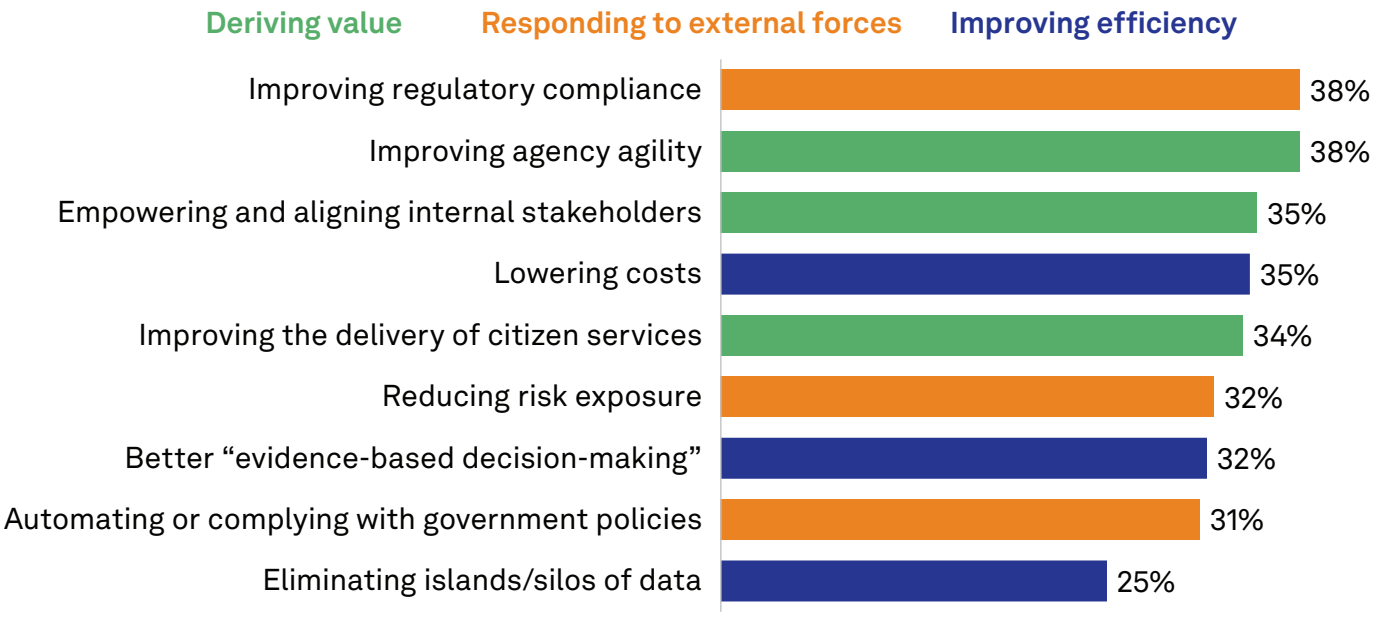
Analytics and the Adoption of Data Lakes

The federal government widely deploys data lakes. These systems provide an environment for running data analytics (e.g., business intelligence, visualization), data science initiatives (e.g., predictive analytics, machine learning), data engineering workloads (e.g., ETL) and operational applications (e.g., ERP, CRM, HR, sales force management applications). More than 50% of government respondents have deployed data lakes, while 48% indicated that data lakes are either in pilot or in a planning phase and expected to be deployed over the next 12 months.

Benefits of a Data Lake

With such high deployment figures, there are great expectations for these data lakes. Respondents expect data lakes to most benefit them by improving regulatory compliance and agility, both at 38% (see Figure 4). Following that, respondents cited lower costs and empowering and aligning internal stakeholders, both at 35%.

Figure 4: Benefits From a Data Lake Environment

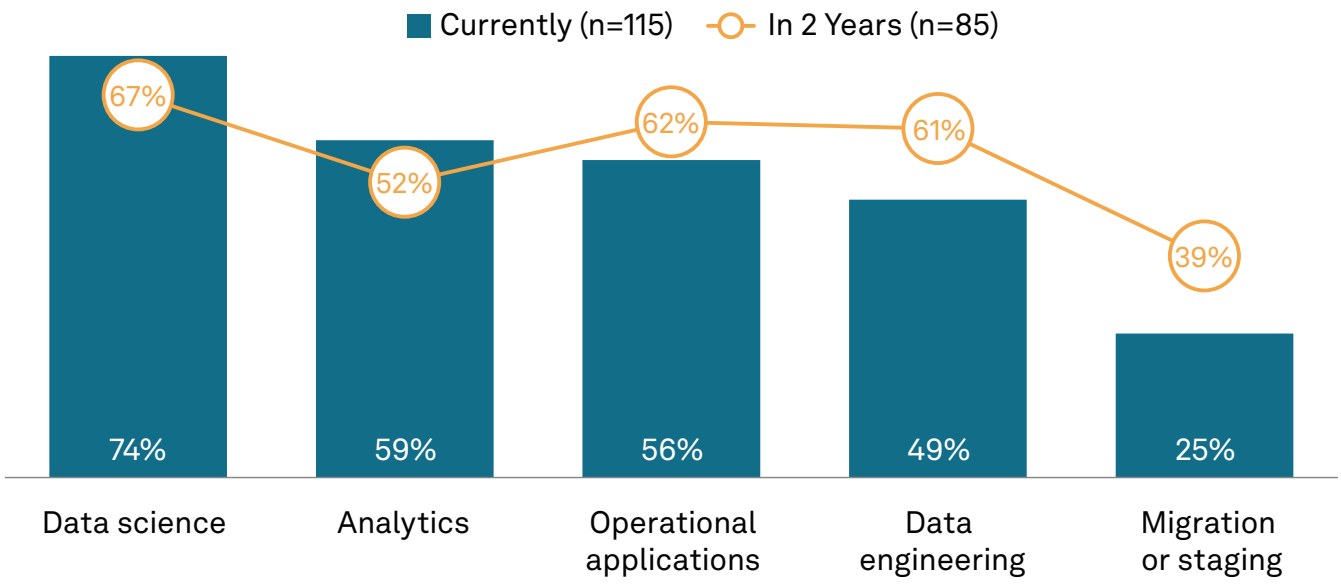


Q. What are the most significant benefits that your agency sees (or expects to see) from its data lake environment?
Base: Respondents whose agency has a data lake currently deployed (n=214)
Source: 451 Research custom survey 2022

Improving regulatory compliance and agility are noteworthy because they address some of the broader needs of many federal agencies. While 38% of respondents cited improving compliance, 32% of respondents cited reducing security vulnerabilities. Compliance and security are interrelated and often considered two sides of the same coin – it would be nearly impossible to have one without the other. Compliance ensures that government agencies abide by and are compliant with policies and regulations, thus protecting agencies from unforeseen risks. Conversely, risk management protects agencies from falling out of compliance.

Many federal respondents identify agility as the primary benefit of data lakes because it inherently implies a certain level of system flexibility. Government agencies want their data lake environments to be adaptable to change when needed. In a separate question, government respondents identified data science workloads (74%) as the primary workload for data lake deployments, followed by analytics (59%) and operational workloads (56%) (see Figure 5).

Figure 5: Data Lake Workloads Now and in Two Years



Q. Which of the following use cases/workloads run (or will run) on your data lake environment in the public cloud? Currently and in two years? Please select all that apply for each column.

Base: Respondents whose agency has a data lake currently deployed

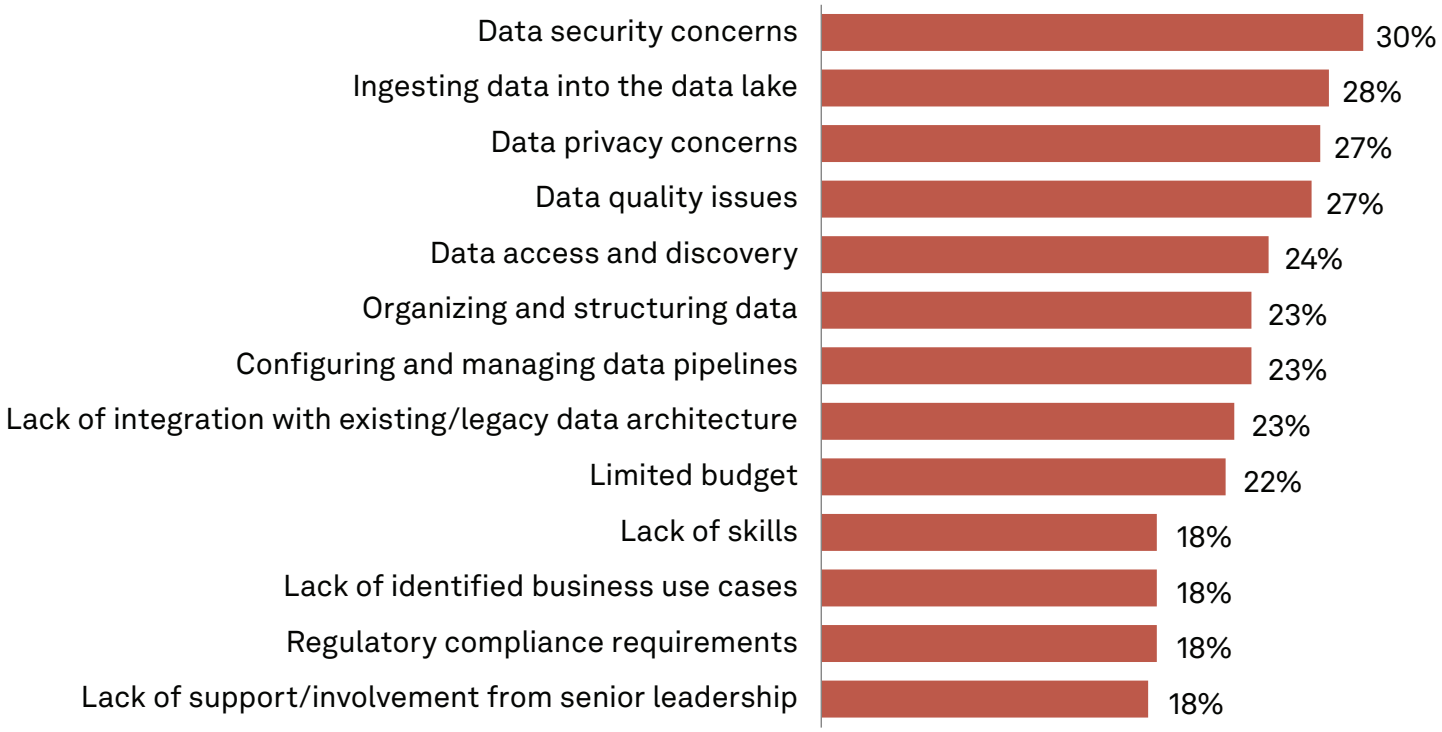
Source: 451 Research custom survey 2022

However, respondents expect there to be some changes to the workloads that will be carried out on their data lakes. Most respondents continued to choose data science (67%) as their expected primary workload in two years, followed by operational workloads (62%) and data engineering (61%). Analytics workloads dropped to 52%. The readjusting of workloads further supports respondents' need for data lake agility. Government agencies should expect their data lakes to evolve while providing flexibility for a variety of workloads.

Challenges of a Data Lake

Data lakes can be complicated beasts and require significant resources and specialized skills to manage. Topping the list of data lake pain points are data security concerns, including encryption and permission controls; 30% of respondents identified this as their number one concern (see Figure 6). Closely related are privacy concerns, which 27% of respondents cited as important. Data ingestion (28%) and data quality issues (27%) are also high on the list of challenges.

Figure 6: Challenges With Data Lakes



Q. What are the most significant challenges your agency faces (or expects to face) when it comes to generating insight from its data lake environment?

Base: Respondents whose agency has a data lake currently deployed (n=213)

Source: 451 Research custom survey 2022

Cloud Adoption

The Cloud Smart strategy to drive cloud adoption in federal agencies is based on three pillars: *security, procurement and workforce*. Procurement and workforce are outside the scope of this survey, but security shows up as a common theme in this survey and appears regularly in participant responses.

The Public Cloud

While the Cloud Smart strategy does not mandate a standard cloud environment (public or private) for federal agencies, the survey reveals that 78% of respondents are running at least some data workloads in a public cloud environment. Another 20% of respondents are planning to deploy workloads to the public cloud within 12 months.

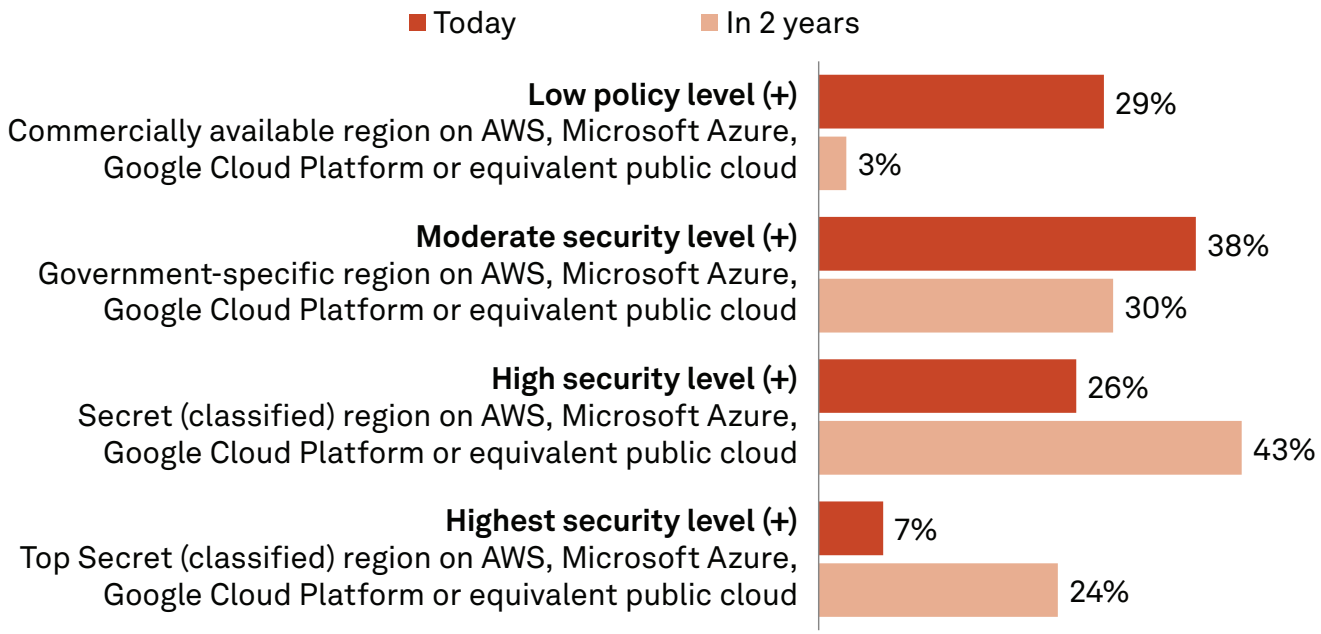
Data lakes likewise see healthy adoption among respondents. Survey results reveal that **54% of respondents are deploying their data lakes in a public cloud environment**, while the remaining **46% are deploying in a private cloud environment**.

Security for Cloud Priorities

Security and privacy concerns are top challenges, and particularly so when deploying to a cloud environment. Additional survey results show that security is not only an immediate concern but will remain so. For example, respondents who indicated that their agency's data lake resides in the public cloud were asked about the level of security for these deployments. The results vary somewhat depending on the specific federal government agency (there are deployments with different levels of security), but overall, they follow a similar pattern.

Respondents who work in defense and intelligence divisions reported that their data lake deployment will require more security in the future, not less. More than a third (38%) of respondents reported current deployments in a government-specific region for a moderate level of security, whereas only 30% of respondents plan to deploy with moderate security (see Figure 8) in two years. In contrast, a growing proportion of respondents will deploy to a high security level (26% currently; 43% in two years) and the highest security level (7% currently; 24%) in two years.

Figure 7: Defense and Intelligence Divisions: Cloud Service Level for Data Lakes in the Public Cloud



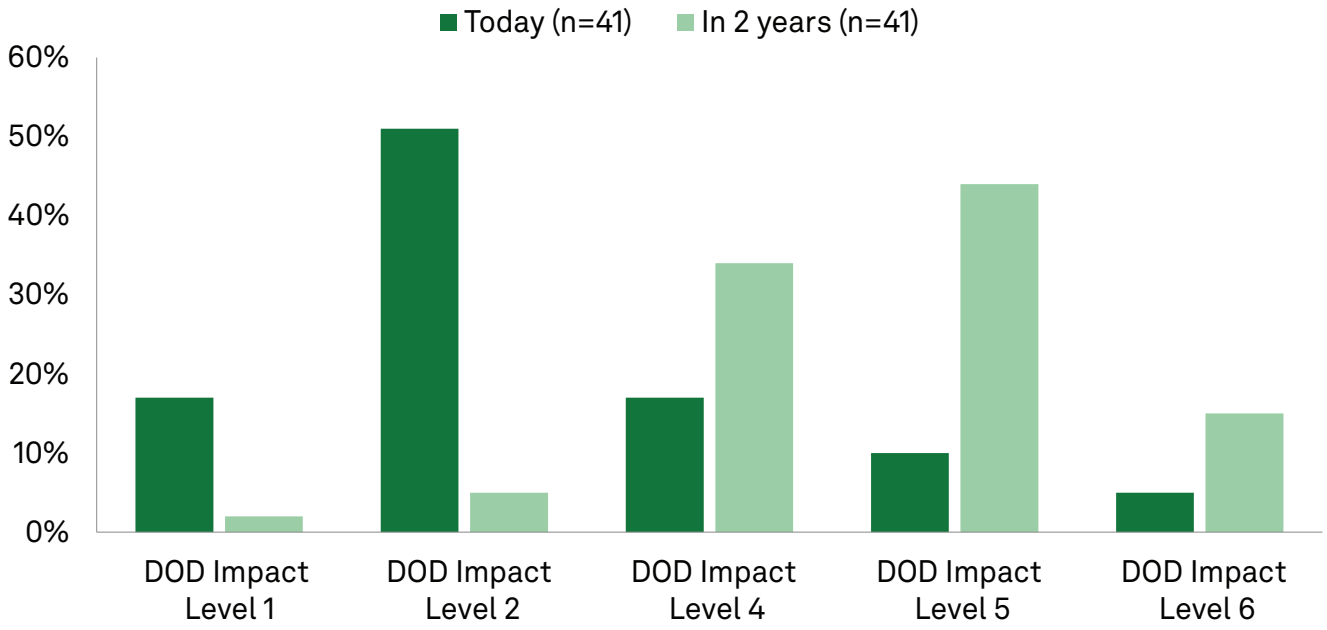
Q. You indicated that your agency currently deploys a data lake in the public cloud. What cloud service level are these systems currently deployed? How will this change over the next two years?

Base: Defense and intelligence division respondents whose agency deploys a data lake in the public cloud (n=107)

Source: 451 Research custom survey 2022

Personnel in civil divisions that follow the Department of Defense (DOD) security-level requirements expressed the same desire for increased security. DOD Impact Level 2 has the largest deployment, with 51% of respondents, but is expected to drop significantly, with only 5% of respondents planning to deploy to this level in two years. Deployments at DOD Impact Level 4 (currently 17%) and Level 5 (currently 10%) are projected to increase in the next two years to 34% and 44% of deployments, respectively.

Figure 8: Civil Division: Cloud Service Level for Data Lakes in the Public Cloud



Q. You indicated that your agency currently deploys a data lake in the public cloud. What cloud service level are these systems currently deployed? How will this change over the next two years?

Base: Civil division respondents whose agency deploys a data lake in the public cloud (n=41)

Source: 451 Research custom survey 2022

Key Observations

A broader look at the survey results reveals several noteworthy observations and allows us to identify a few key trends.

The Value of Data and Data-Driven Insights

Just as the U.S. federal government values cloud adoption, it also greatly values data and will continue to do so well into the future. When asked how their agency will value data in the next 12 months, 41% of respondents said they would value it significantly more, and 51% said it would be somewhat more important. In a separate question, 28% of respondents noted that nearly all strategic decisions are data-driven, and 61% noted that most strategic decisions are data-driven. With 92% of respondents claiming that data will be valued more in the foreseeable future, we might expect to see a nearly identical percentage of organizations using data for all strategic decisions. The disconnect – valuing data versus using data for strategic business decisions – could be attributed to the challenges in managing data, including data volumes, siloed data, lack of data quality and the ability to share data, as well as implement governance policies.

Data Lakes and Integrated Data Management

Data lakes are commonly deployed within the federal government. The systems are primarily used for data science workloads, followed by analytical workloads (business intelligence, visualization), operational workloads (ERP, CRP, HR, etc.) and data engineering (ETL). The results suggest some of these workloads may change over time, but that is expected given governmental changes. Recall that respondents identified data agility and the ability to address regulatory compliance as the primary benefits of a data lake (see Figure 4). Recall also that security concerns are the number one challenge cited, followed by ingesting data into the data lake, data quality and data privacy concerns (see Figure 6). When respondents cite data quality challenges, it directly relates to a poorly implemented data management strategy. It cannot be emphasized enough that the right data management strategy directly impacts the entire data lifecycle from ingestion through management and analysis to insights.

Security and the Entire Data Environment

A theme that runs through the survey data is concern for security and then the challenges enforcing it related to maintaining the federal government's security standard. Security concerns – such as management of encryption and permissions – represent the top challenge for generating insights from a data lake. Maintaining data security and privacy requirements are the top challenges for creating a unified view of the data. Moreover, most respondents cited the use of moderate security level (government-specific region) or DOD Impact Level 2 as the primary deployment security level for their data lakes. In two years, most respondents plan to increase to a higher security level. This trend reveals that security adherence will become increasingly stronger. While the private sector values security standards, the federal government makes security its primary area of focus.

Conclusion: The Future State

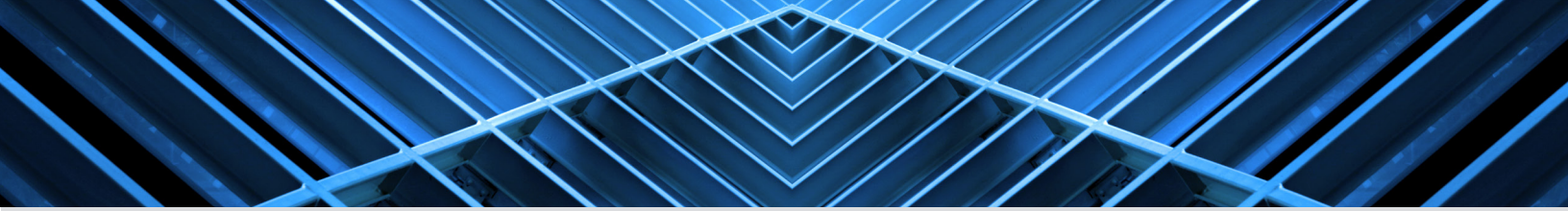
While accelerating the use of data to support data-driven decision-making may seem challenging, it is necessary. The drive to the cloud is also necessary for the federal government. That means that any new systems and processes – along with how data is stored, managed, governed, analyzed and secured – need to operate in a cloud environment.

But it is more than just a cloud adoption journey. We assert that the future for the federal government lies in a deliberate focus on data and especially the management of that data. The federal government will need to put into place the necessary systems and processes to collect, store, refine, manage and analyze its data to ensure high data quality across multiple cloud environments.

As such, we believe the following actions are worth considering:

- **Ingesting high-quality data**. A reliable data environment begins with a strong data ingestion strategy. Respondents cited data ingestion as a top challenge for deriving insights from a data lake (see Figure 6). A faulty data ingestion strategy, however, can create a ripple of challenges. For instance, it can impact data quality, which then impacts the potential quality of analytical workloads as well as data used for machine learning model development. A data ingestion strategy needs to enable the volume of data collected, which may come from streaming or other means, as well as the ability to accommodate a variety of data sources and types, and it needs to do it all with very low latency.
- **Centralizing the data**. Federal agencies already have a significant number of data silos (68% claim 11-100 data silos) with that number expected to rise in the next two years. Consolidating data from various systems can provide several benefits. One benefit is increased data governance outcomes, which respondents agree brought high data quality/insight and improved access (see Figure 3). It can also address cost concerns (see Figure 4) that collectively contribute to overall security and privacy challenges, as consolidation would mean fewer systems for which to maintain security policies.
- **Structured schema with data flexibility**. Survey respondents have validated that data lakes are necessary, in part due to the flexibility they offer. It is important for the government to deploy data management systems that ingest not only structured data but also unstructured data. This flexibility is critical to derive business intelligence and gain insights from the trove of data, and it is imperative to actively move from raw structured and unstructured data to data that adheres to strict schema after transformation and curation. For the data quality and reliability challenges cited by the respondents, it is important that a critical principle of data management, such as ACID compliance, be available to the agencies. ACID (atomicity, consistency, isolation and durability) is a set of principles that ensure database transactions are processed reliably.

Many of the challenges – around analytics, data management and security – identified by respondents are fixable problems, and there are a slew of new and existing innovations geared toward a cloud computing model to address these challenges. The federal government is tasked with providing a service to its public constituents, and as such, all strategic decisions, actions and implementation efforts need to support this goal.



content provided by: databricks

Government agencies looking to adopt many of the Cloud Smart strategies highlighted in this paper likely recognize the need to move toward a secure and open platform that allows them to bring all their data together to support strategic decision-making. While the advent of data lakes addressed some of the limitations of traditional data warehouses, the results of this survey illustrate that challenges remain in the areas of data quality and reliability as well as data readiness for ML and AI. The Federal Data Strategy and the DoD Data Decrees underscore the need for agencies to prioritize solutions that are open source and open standard, and that provide consistent security and governance across all data types in every environment.

Given the benefits and drawbacks of both data warehouses and data lakes, a new paradigm is necessary. Federal agencies need a solution that combines the best elements of data lakes (semi-structured and unstructured data, ML and advanced analytics) and data warehouses (SQL only, structured data only) to deliver the reliability, strong data governance and performance of data warehouses with the openness, flexibility and machine learning support of data lakes. This new paradigm is called the Lakehouse.

The Databricks Lakehouse for Public Sector enables government agencies to harness the full power of data and analytics to solve strategic challenges and make smarter decisions that improve the safety and quality of life for all citizens.

To evaluate how the Databricks Lakehouse for Public Sector can help alleviate problems your agency might be facing, we invite you to download a trial of the Databricks Lakehouse platform or schedule a 30-minute briefing with your Databricks representative to learn more about how best to leverage your agency's cloud investments for data and AI. For more details about the use cases we support and the work we're doing to drive mission value for our Federal government customers, please visit our Federal webpage.

CONTACTS

The Americas
+1 877 863 1306
market.intelligence@spglobal.com

Europe, Middle East & Africa
+44 20 7176 1234
market.intelligence@spglobal.com

Asia-Pacific
+852 2533 3565
market.intelligence@spglobal.com

www.spglobal.com/marketintelligence