EBOOK

Driving Innovation and Transformation in the Federal Government With Data + AI

Empowering the federal government to efficiently deliver on mission objectives and better serve citizens

databricks

Contents

State of the union: Data and AI in the federal government

For the private sector, the growth, maturation and application of data analytics and artificial intelligence (AI) have driven innovation. This has resulted in solutions that have helped to improve efficiencies in everything from optimizing supply chains to accelerating drug development to creating personalized customer experiences and much more. Unfortunately, the federal government and many of its agencies are just beginning to take advantage of the benefits that data, analytics and AI can deliver. This inability to innovate is largely due to aging technology investments, resulting in a sprawl of legacy systems siloed by agencies and departments.

Additionally, the government is one of the largest employers in the world, which introduces significant complexity, operational inefficiencies and a lack of transparency that limit the ability of its agencies to leverage the data at their disposal for even basic analytics – let alone advanced data analytic techniques, such as machine learning.

Recognizing the opportunity for data and AI

The opportunity for the federal government to leverage data analytics and AI cannot be overstated. With access to some of the largest current and historical data sets available to the United States — and with vast personnel resources and some of the best private sector use cases and applications of AI available in the world — the federal government has the ability to transform the efficiency and effectiveness of many of its agencies.

In fact, the federal government plans to spend $4.3 billion in artificial intelligence research and development across agencies in fiscal year 2023, according to a recent report from Bloomberg Government. These priorities are validated by a recent Gartner study of government CIOs across all levels (including state and local), confirming that the top game-changing technologies are AI, data analytics and the cloud.

And as an indication of the potential impact, a recent study by Deloitte shows the government can save upward of $3 billion annually on the low end to more than $41 billion annually on the high end from data-driven automation and AI.

Sources:

- Gartner Survey Finds Government CIOs to Focus Technology Investments on Data Analytics and Cybersecurity in 2019
- Administration Projects Agencies Will Spend $1 Billion on Artificial Intelligence Next Year

> Investment in AI to automate repetitive tasks can improve efficiencies across government agencies, which could save **96.7 million federal hours annually,** with a potential savings of **$3.3 billion.**
>
> WILLIAM EGGERS, PETER VIECHNICKI AND DAVID SCHATSKY
>
> Deloitte Insights

An increased focus on cloud, analytics and AI = operational efficiency

	1. AI/ML 2. Data Analytics 3. Cloud	
$1B	**TOP PRIORITIES**	**$41B+**
Data and AI Research and Development Initiative	Government CIOs' top game-changing technologies	Estimated government savings from data-driven automation
U.S. Government	Gartner	Deloitte.

Fortunately, the President's Management Agenda (PMA) has recognized the need to modernize their existing infrastructure, federate data for easier access and build more advanced data analytics capabilities by establishing mandates for modernization, data openness and the progression of AI innovations.

This will put agencies in a better position to leverage the scale of the cloud and democratize secure access to data in order to enable downstream business intelligence and AI use cases.

IT Modernization Act

Allows agencies to invest in modern technology solutions to improve service to the public, secure sensitive systems and data, and save taxpayer dollars.

Federal Data Strategy

A 10-year vision for how the federal government will accelerate the use of data to achieve its mission, serve the public and steward resources, while protecting security, privacy and confidentiality.

AI Executive Order

Makes AI a top research and development priority for federal agencies, provides a shared ethics framework for developing and using AI, and expands job rotation programs to increase the number of AI experts at agencies.

The end result will be transformative innovation that can not only improve the operational efficiencies of each agency, but also support the delivery of actionable insights in real time for more informed decision-making. This benefits citizens in the form of better services, stronger national security and smarter resource management.

Top data and AI use cases in the government

Across the federal government, data and AI is providing the insights and predictive capabilities to thwart cyberattacks and national threats, provide better social services more efficiently, and improve the delivery and quality of healthcare services.

HOMELAND SECURITY

Detect and prevent criminal activities and national threats with real-time analytics and data-driven decision-making.

- Customs and border protection
- Immigration and citizenship
- Counter-terrorism
- Federal emergency aid management

DEFENSE

Apply the power of predictive analytics to geospatial, IoT and surveillance data to improve operations and protect the nation.

- Logistics
- Predictive maintenance
- Surveillance and reconnaissance
- Law enforcement and readiness

HEALTHCARE

Improve the delivery and quality of healthcare services for citizens with powerful analytics and a 360° view of patients.

- Patient 360
- Population health
- Supply chain optimization
- Insurance management
- Genomics
- Drug discovery and delivery

ENERGY

Improve energy management with data insights that ensure energy resiliency and sustainability.

- Security of energy infrastructure
- Smarter energy management
- Energy exploration
- Electrical grid reliability

COMMERCE

Proactively detect anomalies with machine learning to mitigate risk and prevent fraudulent activity.

- Tax fraud and collection
- Process and operations management
- Grants management
- Customer 360

INTELLIGENCE COMMUNITY

Leverage real-time insights to make informed decisions that can impact the safety of our citizens and the world.

- Threat detection
- Neutralize cyberattacks
- Intelligence surveillance and reconnaissance
- Social media analytics

Challenges to innovation

The opportunity to drive innovation throughout the federal government is massive and has implications for every U.S. citizen. But there are several critical barriers preventing agencies from making the progress needed to realize the value of their data and delivering those innovations.

The complexities and impact of legacy data warehouses and marts

Multiple federal agencies are burdened with a legacy IT infrastructure that is being left behind by the technological advancements seen in the private sector. This infrastructure is traditionally built with on-premises data warehouses and data marts that are highly complex to maintain, costly to scale as compute is coupled with storage, limited from a data science perspective, and they lack support for the growing volumes of unstructured data. This inhibits data-driven innovation and blocks the use of AI, leaving agencies to search for data science tools to fill the gaps.

Infrastructure also becomes harder and more expensive to maintain as it ages. Over time, these environments become more complex due to their need for specialized patches and updates that keep these systems available while doing nothing to solve the issues of poor interoperability, ever-decreasing processing speeds, and an inability to scale – all of which are critically necessary to support today's more data-intensive use cases. For example, systems at the departments of Education, Health and Human Services, Treasury, and Social Security are over 40 years old.[1] This is causing pain in a variety of areas.

Maintaining these systems requires a massive investment of both time and money compared to modern cloud-based systems. For the technical teams that are tasked with trying to integrate any of these legacy systems with third-party tooling or services, this often requires significant customization and, even then, there is still a chance that the final integration won't be successful. These systems also keep personnel from spending their energy and resources on emerging technologies such as AI.

> “Ten of the existing legacy systems most in need of modernization cost about **$337 million a year** to operate and maintain.
>
> THE GOVERNMENT ACCOUNTABILITY OFFICE,
> INFORMATION TECHNOLOGY REPORT TO CONGRESS, JUNE 2019

And data reliability is a big concern. Replication of data occurs across data marts as various teams try to access and explore it, creating data management and governance challenges. Without a single source of truth, teams struggle with data inconsistencies, which can result in inaccurate analysis and model performance that is only compounded over time.

Thankfully, there are initiatives in place, such as the Data Center and Cloud Optimization Initiative Program Management Office (DCCOI PMO), which are investing in modernizing IT infrastructure for federal agencies.[2]

[1] Agencies Need to Develop Modernization Plans for Critical Legacy Systems

[2] IT Modernization

Data is critical … and complicated

Data is both the greatest asset and one of the greatest challenges that federal agencies must learn to manage. While the volume and usefulness of data collected by federal agencies are not in question, much of it is locked in legacy source systems, comes in diverse structured and unstructured formats, and is subject to a variety of governance models.

Not only is this data siloed and very difficult to integrate, but the data volumes collected by federal agencies are massive. At Health and Human Services, for example, or the Department of Veterans Affairs, healthcare data sets will be sized by population and include electronic health records, clinical data, imaging and more. For the Department of Defense and the Department of Homeland Security, data includes everything from mapping, satellite imagery and intelligence data to payroll and human resources data. The Social Security Administration and Internal Revenue Service manage personal data for every single citizen in the United States.

Combining these various forms of data from disparate legacy systems that are not integrated — and doing it across different government agencies and departments — can be slow and error prone, hindering downstream analytics and actionable insights. The teams that are responsible for this are faced with not only integrating these data sources, but also managing the entire ETL workflow in order to enable the application of basic analytics, let alone machine learning and AI.

Data silos hamper any data-driven advancements

In any data-driven organization, the need to have trusted, timely and efficient access to data is critical. For the data teams responsible for driving the digital transformation of federal agencies, the challenges they face are myriad.

We have already seen how existing, legacy infrastructure, as well as the integration of fragmented data sources, will strain data engineering teams trying to deliver high-quality data at scale. Their challenge includes developing the right data pipelines that will take the massive volumes of raw data coming from fragmented sources into one centralized location with clean, secure and compliant data for agency decision-makers.

Data scientists and analysts alike must have the right toolset to collaboratively investigate, extract and report meaningful insights from this data. Unfortunately, data silos extend to organizational silos, which make collaboration inside an agency as well as between agencies very difficult. With different groups of data teams leveraging their own coding and analytical tools, communicating insights and working across teams — let alone across agencies — is almost impossible. This lack of collaboration can drastically limit the capabilities of any data analytics or AI initiatives — from the deployment of shared business intelligence (BI) reports and dashboards for data investigation and decision-making to the training of machine learning models to automate processes and make predictions. Compounding these challenges is an overall lack of data science expertise and skills within federal agencies. As a result, even with access to their data, without intuitive tooling it's very difficult to deliver advanced analytic use cases with ML and AI.

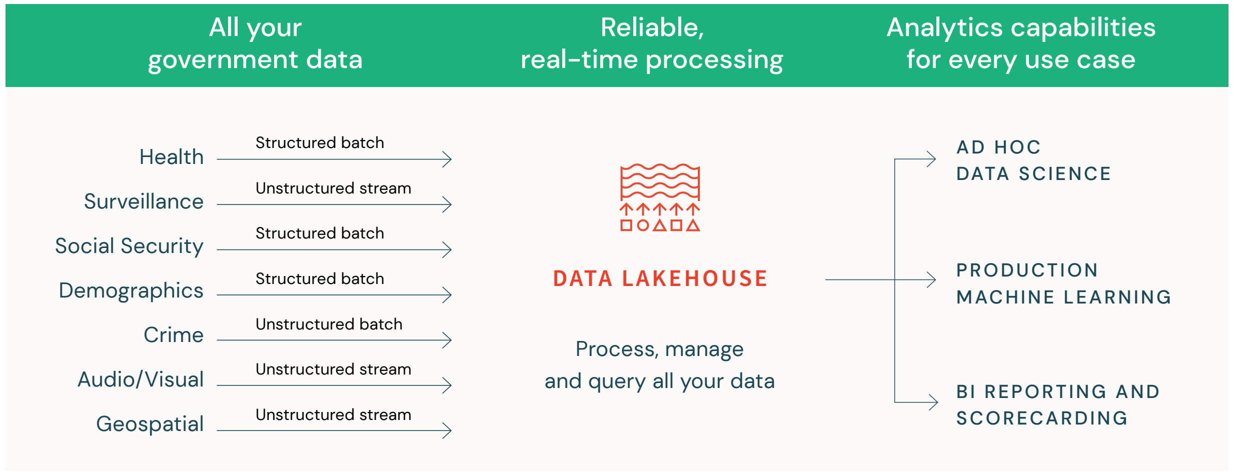
Organizational silos also impact the effectiveness of data analysts, who are responsible for analyzing and reporting insights from the data to better inform subject-matter experts or policy — and decision-makers. Without a data platform that eliminates these silos and enables visualization of and reporting on shared data, data analysts will be limited in how they are able to drive the organizational and policy agendas of their respective agencies.

THE DATABRICKS LAKEHOUSE PLATFORM:

Modernizing the federal government to achieve mission objectives

Databricks provides federal agencies with a Lakehouse Platform that combines the best of data warehouses and data lakes — to store and manage all your data for all your analytics workloads. Databricks federates all data and democratizes access for downstream use cases, empowering federal agencies to unlock the full potential of their data to deliver on their mission objectives and better serve citizens.

Lakehouse offers a single solution for all major data workloads, whether structured or unstructured, and supports use cases from streaming analytics to BI, data science and AI.



The Databricks Lakehouse Platform has three unique characteristics that address head-on the biggest challenges that federal agencies are facing:

1. It offers simplicity with regard to data management, in that the Databricks Lakehouse is architected to support all of an agency's data workloads on one common platform
2. It is built on open standards so that any existing investments in tooling or resources can remain effective
3. And it's collaborative, enabling agency data engineers, analysts and data scientists to work together much more easily

Federal agencies that are powering impactful innovations with Databricks Lakehouse

Using predictive analytics for better passenger safety and experience

U.S. Citizenship and Immigration Services

Enabling operational efficiencies through process automation to streamline the path to citizenship

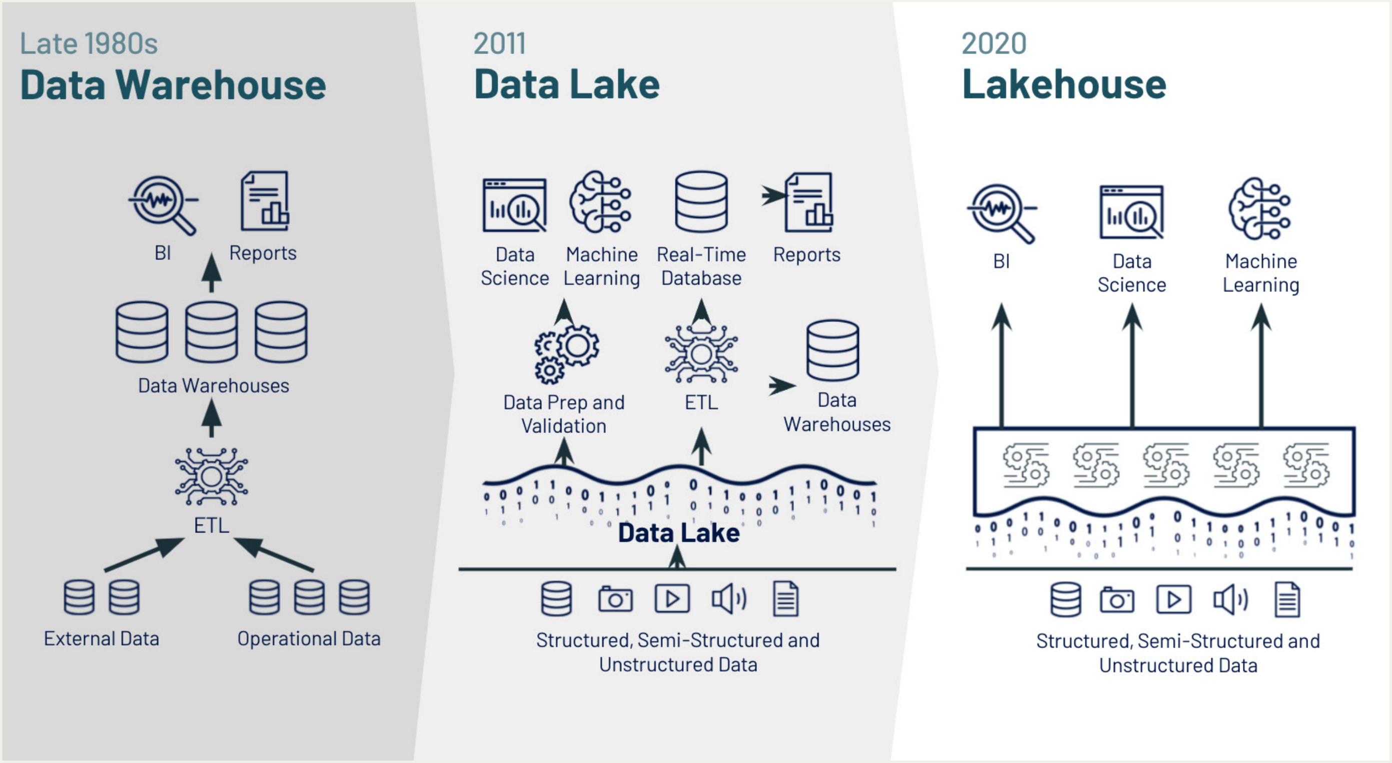
CMS CENTERS FOR MEDICARE & MEDICAID SERVICES

Leveraging advanced analytics to improve outcomes for patients through Medicare and Medicaid services

Managing federal data with a unified approach

Databricks enables aggregation and processing of massive collections of diverse and sensitive agency data that currently exists in silos, both structured and unstructured. As we've seen, for many agencies this would be incredibly difficult with the infrastructure challenges they are experiencing. The Databricks Lakehouse leverages Delta Lake to unify the very large and diverse amounts of data that government agencies are working with. Delta Lake is an open format, centralized data storage layer that delivers reliability, security and performance — for both streaming and batch operations.

By providing a unified data foundation for business intelligence, data science and machine learning, federal agencies can add reliability, performance and quality to existing data lakes while simplifying data engineering and infrastructure management with automation to simplify the development and management of data pipelines.



The Lakehouse Platform combines the best elements of data lakes and data warehouses — delivering the data management and performance typically found in data warehouses with the low-cost, flexible object stores offered by data lakes

Break down the institutional silos limiting collaboration

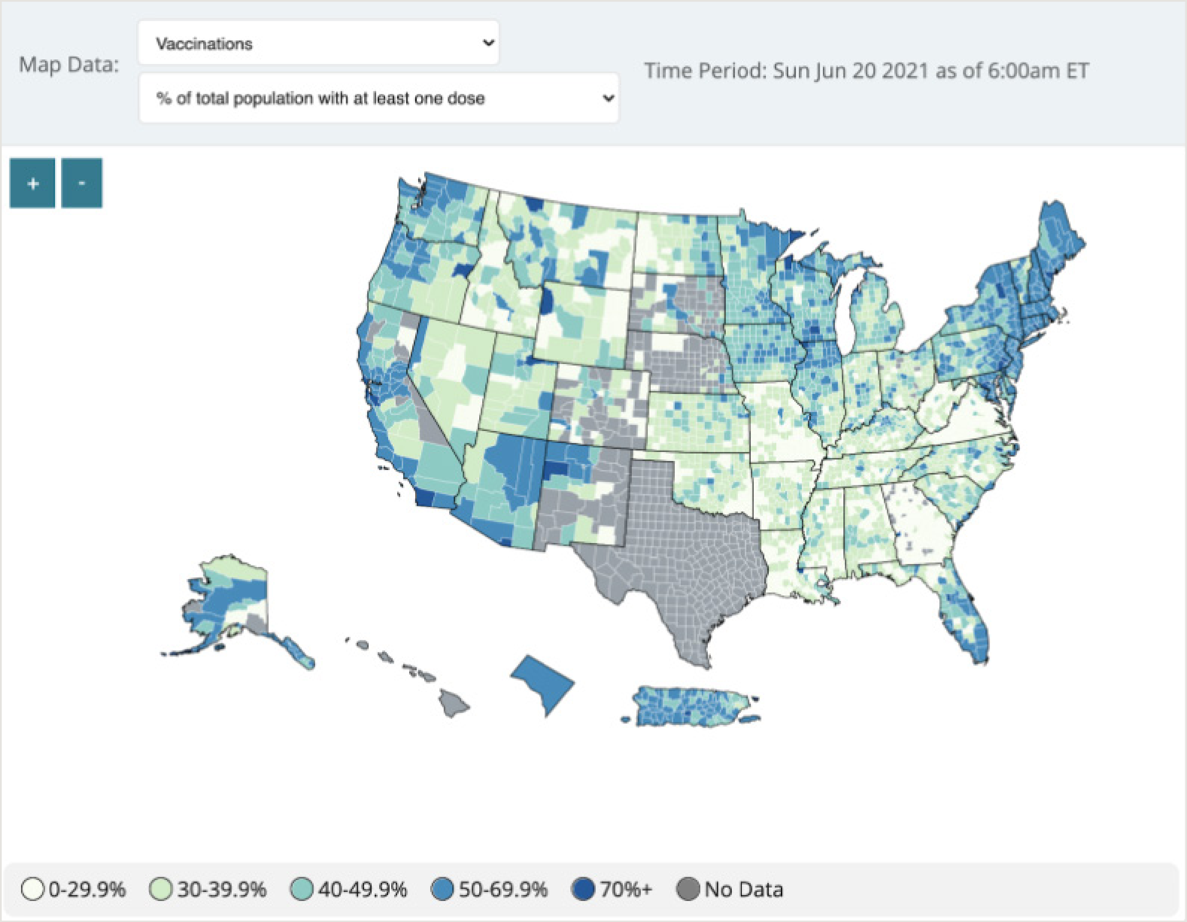
Foster collaboration at every step with the latest machine learning tools that allow everyone to work and build value together — from data scientists to researchers to business decision-makers. Close the glaring skills gap within these government organizations by providing tooling that simplifies the ML lifecycle and empowers the data teams that do not have the data science expertise to still be productive with their data through integrating BI tools and SQL analytics capabilities.

Empower data scientists with an intuitive and interactive workspace where they can easily collaborate on data, share models and code, and manage the entire machine learning lifecycle in one place. Databricks notebooks natively support Python, R, SQL and Scala so practitioners can work together with the languages and libraries of their choice.

Deliver on mission objectives with powerful analytics across agencies

The Databricks Lakehouse Platform includes a business intelligence capability — Databricks SQL. Databricks SQL allows data analysts and users to query and run reports against all of an agency's unified data. Databricks SQL integrates with BI tools, like Tableau and Microsoft Power BI, and complements any existing BI tools with a SQL-native interface, allowing data analysts and data scientists to query data directly within Databricks.

Additionally, with Databricks SQL, the data team can turn insights from real-world data into powerful visualizations designed for machine learning. Visualizations can then be turned into interactive dashboards to share insights with peers across agencies, policymakers, regulators and decision-makers.


Map Data: Vaccinations
% of total population with at least one dose
Time Period: Sun Jun 20 2021 as of 6:00am ET

Easily create visualizations and share dashboards via integrations with BI tools, like Tableau and Microsoft Power BI

Ensure data security and compliance at scale

Databricks is fully aware of the sensitivity of the data that many of our federal agencies are responsible for. From national security and defense data to individual health and financial information to national infrastructure and energy data — all of it is critical. Data is protected at every level of the platform through deep integration with fine-grained, cloud-provider access control mechanisms. The Databricks Lakehouse is a massively secure and scalable multicloud platform running millions of machines every day. It is independently audited and compliant with FedRAMP security assessment protocols on the Azure cloud and can provide a HIPAA-compliant deployment on both AWS and Azure clouds.

The platform's administration capabilities include tools to manage user access, control spend, audit usage, and analyze activity across every workspace, all while seamlessly enforcing user and data governance, at any scale.

With complete AWS accreditation, Databricks runs across all major networks including GovCloud, SC2S, C2S and commercial; all networks, including public, NIPR, SIPR and JWICS; and ATOs, including FISMA, IL5, IL6, ICD 503 INT-A and INT-B.

CUSTOMER STORY: U.S. CITIZENSHIP AND IMMIGRATION SERVICES

Streamlining the path to citizenship with data

U.S. Citizenship and Immigration Services

24x faster
query performance

10 minutes
to process tables with 120 million rows

40 million
applications processed

The U.S. Citizenship and Immigration Services (USCIS) gains actionable insights from dashboards via Tableau to better understand how to streamline operations and more quickly process immigration and employment applications as well as petitions. Today, their data analyst team has over 6,000 Tableau dashboards running — all powered by Databricks.

The U.S. Citizenship and Immigration Services is the government agency that oversees lawful immigration to the United States. Over the last decade, the volume of immigration- and citizenship-related applications has skyrocketed across naturalizations, green cards, employment authorizations and other categories. With millions of applications and petitions flooding the USCIS, processing delays were reaching crisis levels — with overall case processing times increasing 91% since FY2014.

Processing delays fueled by on-premises, legacy architecture

Core to these issues was an on-premises, legacy architecture that was complex, slow and costly to scale. By migrating to AWS and Databricks, USCIS adopted a unified approach to data analytics with more big data processing power and the federation of data across dozens of disparate sources. This has unlocked operational efficiencies and new opportunities for their entire data organization to drive business intelligence and fuel ML innovations designed to streamline application and petition processes.

Removing complexities with a fully managed cloud platform

Since migrating to the cloud and integrating Databricks into their data analytics workflows, USCIS has been able to make smarter decisions that help streamline processes and leverage ML to reduce application processing times. These newfound efficiencies and capabilities have allowed them to scale their data footprint from about 30 data sources to 75 without issue.

Databricks provided USCIS with significant impact where it mattered most — faster processing speeds that enabled data analysts to deliver timely reports to decision-makers — and that freed up data scientists to build ML models to help improve operations. Leveraging the efficiencies of the cloud and Delta Lake, they were able to easily provision a 26-node cluster within minutes and ingest tables with 120 million rows into S3 in under 10 minutes. Prior to Databricks, performing the same processes would have taken somewhere between two and three hours.

A new era of data-driven innovation improves operations

USCIS now has the ability to understand their data more quickly, which has unlocked new opportunities for innovation. With Databricks, they are able to run queries in 19 minutes, something that used to take an entire day — a 24x performance gain. This means they are spending far less time troubleshooting and more time creating value.

> “We discovered Databricks, and the light bulb really clicked for us on what we needed to do moving forward to stay relevant.
>
> SHAWN BENJAMIN
> CHIEF OF DATA AND BUSINESS INTELLIGENCE, USCIS

Conclusion

Enabling federal agencies to take advantage of data analytics and AI will help them execute their missions both effectively and efficiently. The Databricks Lakehouse Platform will unify data, analytics and AI workloads, making agencies data-driven and giving policymakers access to deeper, more meaningful insights for decision-making. It will also eliminate data silos and increase communication and collaboration across agencies to ensure the best results for all citizens.

About Databricks

Databricks is the data and AI company. More than 5,000 organizations worldwide — including Comcast, Condé Nast, H&M, and over 40% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark™, Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems.

Get started with a free trial of Databricks and start building data applications today

START YOUR FREE TRIAL

To learn more, visit us at: **dbricks.co/federal**

databricks