CLOUDERA

WHITE PAPER

Accelerate your Mission with Data in Motion

Highlights from a June 2022 Roundtable

Every level of government is awash in data, and the global pandemic did little to slow its growth. Agencies are now struggling to figure out how to effectively ingest, route and analyze the data (regardless of source) to achieve its full potential.

To address this problem, the Advanced Technology Academic Research Center (ATARC) recently hosted a roundtable in partnership with **Cloudera** on the concept of "data-in-motion." Government thought leaders gathered to discuss their data challenges and how they're harnessing and deploying next-generation data platforms and other new technologies that simplify data acquisition and delivery.

Data Overload

Digital storage was quite expensive as recently as two decades ago. Most stored data was limited in scope and utilized only for a specific task, application, or need. But advancement in hard drives including solid-state technology changed that, dropping the average price per gigabyte of storage from $7.70 in the early 2000s to just pennies 15 years later.

Average Cost of Data Per GB

early 2000s	$7.70
2017	$0.02

Source: USC Marshall

Widespread broadband availability made it possible to store seemingly unlimited amounts of data off-site as cloud storage took off. At the same time, organizations upgraded their servers with ever larger drives. There were no longer hard limits on the amount of data that could be stored. All this data had value, so everyone – including the Federal Government – began storing everything and anything with the belief that "we'll worry about it later." Data management was an afterthought.

While private sector deals with massive amounts of data on its own, the volume of data in government is many factors larger. Agencies like the Social Security Administration or the IRS manage petabytes of data on hundreds of millions of people. Roundtable participants spoke frankly about some of these challenges. Still, they saw opportunities in next-generation technologies like Artificial Intelligence (AI) and Machine Learning (ML) to make sense of the massive amount of data most agencies swim in.

Improved Decision Making

Most participants pointed to AI and ML as two of the most significant developments in the data space for large organizations like the Federal Government. AI/ML is beneficial for looking for correlation in a large data set, and the speed at which it operates – far faster than any human or analytics software – makes quick (and informed) decision-making possible.

Participants shared being able to speed up analytics processes from hours to just seconds. Some noted that their AI/ML deployments helped find relationships that would have otherwise remained hidden.

Perhaps the most important feature of understanding the vast amount of data agencies have gathered over the past several decades is **demonstrating value**. As Government agencies, we must answer to Congress and in turn, our constituents.

The Rise of AI

AI generated data of all data produced:

2021	1%
2025	10%

2024 1/3 of all large businesses will use decision intelligence

Source: Gartner

One participant noted that this would work to Government's advantage during the appropriations process. With AI/ML-backed analyses in hand, agencies can better demonstrate value for projects they are seeking funding on. In the end, no agency's data efforts matter if they cannot demonstrate value to those who hold the purse strings.

An Enterprise Data Strategy for the Federal Government

Participants next turned their attention to discussing enterprise data strategy and how to replicate it within the Federal Government. While new entrants to the Federal workforce are increasingly data savvy, many agencies have a workforce with many long-tenured employees.

With that in mind, agencies won't just be able to roll out new data management strategies and expect them to be used, much less have employees understand how to use them right away. During the discussion, recommendations emerged around four main areas: governance, platform, community, and coherency.

Governance

Frequently, collected data gets stored without much consideration to its contents. Often this could involve personally identifiable information – an easy target for cybercriminals, although efforts must also be made to prevent data misuse from within.

From the usability aspect of governance, participants recommended that agencies establish standards on data structuring with a keen eye toward future interoperability. This way, the data is both easy to find and use.

Platform

After setting boundaries on how data is categorized, stored, and accessed, the next step is building a data hosting platform. All participants agreed that this platform should offer interoperability with the capability to easily integrate with other applications.

Security is paramount, especially so with the many Federal cybersecurity mandates to comply with, participants urged. While it may slow deployment (and possibly anger stakeholders), up front time investment on platform security compliance is worth it. The Zero Trust mandates are less than two years away, so any new platforms should have those principles "baked in" from the get-go.

Community

Governance and platform come hand-in-hand with education on these tools. A typical Federal employee has limited data experience beyond old school spreadsheets. Some participants noted that they started "data universities" to teach their workforce to better utilize available next-generation data tools.

Other activities like "hackathons" can help solve complex problems. External partners from private sector bring additional insights from dealing with similar data management issues on a smaller scale.

Data Ethics

Data ethics is the study of ethical principles that apply to the use of data. It includes privacy, data accuracy, data security, and data sharing.

Coherency

Participants also discussed the concept of 'coherency.' A platform that is difficult to use or doesn't make sense for agency operations, is a waste of resources. While plenty of third-party data tools can help make sense of large data sets, it is crucial to first define the data problems to be solved.

Throughout the roundtable, participants repeatedly stressed the importance of demonstrating the value of modern data technologies, both from an internal and external perspective. Once people understand the tools and how it fits mission objectives, alignment occurs. From there, operationalizing demonstrates value, which spurs further adoption.

Data Ethics and Equity

When employing tools like AI/ML, important questions arise around data ethics and equity. Federal agencies deal with a significant amount of personally identifiable information. Data tools must not infringe on constituents' privacy, and limits must be placed to prevent misuse.

For some agencies, this has brought about efforts to minimize data movement. Others incorporate data ethics courses into their workforce training programs. But ethics is not the only point of worry.

Tools like AI/ML do not think for themselves. Massive amounts of information get ingested along with the inaccuracies, errors, and biases of the humans that entered it. Data equity ensures that marginalized communities are not put at a disadvantage due to biases in the data itself.

It is equally important that agencies use data in a manner that protects the privacy of the constituents it affects, and their employees clearly understand what is appropriate use of agency tools. Also, as agencies roll out next-generation technologies, it is critical to eliminate any underlying human biases. AI and ML have not reached the maturity to remove these biases.

Data Equity

Data equity is the principle that data should be collected, analyzed, interpreted, and distributed fairly and equitably. It considers the needs of different groups and ensures that everyone has access to the data they need to make informed decisions.

How Cloudera Can Help

Cloudera partners with Federal institutions to support data security and governance mandates, modernize data architectures, and meet the Zero Trust mandate related to data flow. Read more about Cloudera's Data-in-Motion Philosophy and other Public Sector Solutions.