FEDERAL BIG DATA SUMMIT

DECEMBER 5, 2017 | MARRIOTT METRO CENTER | WASHINGTON, DC

On behalf of the Advanced Technology Academic Research Center, I am proud to announce the release of a White Paper documenting the MITRE-ATARC Big Data Collaboration Symposium held on December 5, 2017 in Washington, D.C. in conjunction with the ATARC Federal Big Data Summit.

I would like to take this opportunity to recognize the following session leads for their contributions:

MITRE Chair: Christine Harvey

Challenge Area 1: Emerging Technologies in Big Data & Analytics

Government Lead: Marc Wine, VA
MITRE Lead: Dr. Haleh Vafaie

Challenge Area 2: Big Data & Security Technologies

Government Lead: Renata Spinks, IRS
Industry Lead: Shaunak Ashtaputre, Karsun Solutions
MITRE Lead: Anne Tall

Challenge Area 3: Big Data Landscape

Government Lead: Joshua Gustin, FAA
Industry Lead: Jonathan Janos, MapR
MITRE Lead: Anuja Verma

Challenge Area 4: Leveraging Big Data for Mission Success

Industry Lead: Frank A. Romano, Hortonworks
MITRE Lead: Ronald Campbell

Challenge Area 5: Big Data and Healthcare

Government Lead: Yvonne Cole, DoD-VA IPO
Industry Lead: Rashmi Mathur, IBM
MITRE Lead: Richard Eng

Below is a list of government, academic and industry members who participated in these dialogue sessions:

Challenge Area 1: Emerging Technologies in Big Data & Analytics

Mark Brady, DOJ; Gwendolen Brown, DHS ICE; Anil Chaudhry, DHS; Aaron Dent, ASRC Federal; David Dinh, DLA; John Griffith, MITRE; Teferi Hagos, VA; Chien-chih Lin, VA; Jim Maas, DOE; Raj Sood, Snowflake Computing

Challenge Area 2: Big Data & Security Technologies

Michele Cohen, NNSA; Yolanda Darricarrere, DOC; David DeVries, Dataguise; Rick Eulo, IDC; Robert Fleming, GSA; Debbie Granberry, ASRC Federal; George McKay, GSA; Kelly Miller, NRO; Grace Navas, FRB; Victor Pimenthal, GSA; Thomas Reaves, EPA; Gaurav Seth, DoD-VA IPO; Kamran Shah, USCG; Ricardo Thoroughgood, DoD

Challenge Area 3: Big Data Landscape

Holly Carr, SEC; Bianca Depusoir, DHS ICE; Renee Fulton, USDA; Jessica Glace, MITRE; Ralph Gronlund, DoD; Jack Wainwright, IDC

Challenge Area 4: Leveraging Big Data for Mission Success

Ann Balough, DHS ICE; Sara Bauer, BLS; John Broderick, BLM; Guy Francois, DoD-VA IPO; Rich Gallagher, GSA; Alison Kinn Bennett, EPA; Robyne McRey, NSF; Shakeel Mohammed, HHS CMS; Kathy Rondon, R2C; Nick Tran, U.S. Navy; Nancy Zhou, NIH

Challenge Area 5: Big Data and Healthcare

Dan Bolon, Datawatch; Mimi Boussouf, DoD-VA IPO; Dr. Robert Carmack, DoD-VA IPO; Michelle Casagni, MITRE; Swati Kulkarni, FDA; James Miller, FCC; Chris Muir, HHS; Kim Osborne, U.S. House of Representatives; Dr. Greg Pappas, FDA; Jeff Plum, NIH; Dr. Joseph Ronzio, VA; Peter Rubacky, MapR; Heideh Shadmand, DoD-VA IPO; JT Sison, Dataguise; Christine Wang, FDA

Thank you to everyone who contributed to the MITRE-ATARC Big Data Collaboration Symposium. Without your knowledge and insight, this White Paper would not be possible.

Sincerely,

Tom Suder
President, Advanced Technology Academic Research Center (ATARC)
Host organization of the ATARC Federal Big Data Summit

FEDERAL SUMMITS

DECEMBER 2017
FEDERAL BIG DATA SUMMIT REPORT*

February 9, 2018

Christine Harvey, Justin Brunelle, Ronald Campbell,
Richard Eng, Anne Tall, Dr. Haleh Vafaie, Anuja Verma
The MITRE Corporation

Tim Harvey and Tom Suder
The Advanced Technology Academic Research Center

Contents

EXECUTIVE SUMMARY

The most recent installment of the Federal Big Data Summit, held on December 5, 2017, included five MITRE-ATARC (Advanced Technology Academic Research Center) Collaboration Sessions. These collaboration sessions allowed industry, academic, government, and MITRE representatives the opportunity to collaborate and discuss challenges the government faces in big data. The goal of these sessions is to create a forum to exchange ideas and develop recommendations to further the adoption and advancement of big data techniques and best practices within the government.

Participants representing government, industry, and academia addressed five challenge areas in big data: Emerging Technologies in Big Data & Analytics, Big Data and Security Technologies, Big Data Landscape, Leveraging Big Data for Mission Success, and Big Data & Healthcare. This white paper summarizes the discussions in the collaboration sessions and presents recommendations for government and academia while identifying orthogonal points between challenge areas.

As an outcome of these collaboration sessions, conference participants indicate high interest in the potential costs savings/ease of use/better information sharing. The imperative for public/private cloud technology adoption to meet big data processing requirements was identified as a goal by several organizations. However, security issues will require more research and development before full adoption and exploitation of these capabilities can be established. These areas all require further research and exploration.

The collaboration sessions identified detailed, actionable recommendations for the government and academia which are summarized below:

- Data sharing is one of the most important aspects in the modern field of big data. Organizations need to have policies and governance in place to securely share data within the agency, across agencies, and with the private sector when necessary.

- Data security continues to be one of the most important factors in working with big data. Agencies should have solid cybersecurity protection and should have security measures in place to protect data while still adapting to emerging technologies in data science.

- Another recurring theme in the big data sessions is the need for education, not just at the research level, but across the entire organization. Informed and knowledgeable senior executives are necessary to forming organization-level data strategies and to enable progress.

- Inter-agency communication and collaboration is recognized as a valuable way to bring knowledge to bear across organizations and to move the entire government forward quickly and effectively in regards to big data.
- When organizations have firm security policies, and knowledgeable executives, the process of developing a business case and performing valuable analytics is simplified.

1 INTRODUCTION

During the most recent Federal Big Data Summit, held on December 5, 2017, five MITRE-ATARC (Advanced Technology Academic Research Center) Collaboration Sessions gave representatives of industry, academia, government, and MITRE the opportunity to discuss challenges the government faces in big data. Experts who may not otherwise meet or interact used these sessions to identify challenges, best practices, recommendations, success stories, and requirements to advance the state of big data technologies and research in the government.

The MITRE Corporation is a not-for-profit company that operates multiple Federally Funded Research and Development Centers (FFRDCs). ATARC is a non-profit organization that leverages academia to bridge between government and corporate participation in technology. MITRE worked in partnership with ATARC to host these collaborative sessions as part of the Federal Big Data Summit. The invited collaboration session participants across government, industry, and academia worked together to address challenge areas in big data, as well as identify courses of action to be taken to enable government and industry collaboration with academic institutions. Academic participants used the discussions as a way to help guide research efforts, curricula development, and to help produce graduates ready to join the workforce and advance the state of big data research and work in the government.

This white paper is a summary of the results of the collaboration sessions and identifies suggestions and recommendations for government, industry, and academia while identifying cross-cutting issues between the challenge areas.

2 COLLABORATION SESSION OVERVIEW

Each of the five MITRE-ATARC collaboration sessions consisted of a focused and moderated discussion of current problems, gaps in work programs, potential solutions, and ways forward. At this summit, sessions addressed:

- Emerging Technologies in Big Data & Analytics
- Big Data & Security Technologies
- Big Data Landscape
- Leveraging Big Data for Mission Success
- Big Data & Healthcare

This section outlines the goals, themes, and findings of each of the collaboration sessions.

2.1 Emerging Technologies in Big Data & Analytics

The session on Emerging Technologies in Big Data & Analytics facilitated a discussion on developing technologies and their potential impact on government and industry. It also allowed various discussions on how to best leverage these technologies to have the most impact.

The session included discussions of the following:

- How can organizations within the government securely and effectively share data?
- What cybersecurity assurance and privacy practices need to be in place?
- How can organizations make the most of their data? What best practices and lessons learned exist across government organizations for using and learning from data?
- While keeping data secure and managing data sharing, how can organizations implement data transparency practices?

2.1.1 Challenges

- Organizations face many challenges when it comes to data sharing. Government agencies need to determine how to balance risks and rewards, whether to use open or closed models, and how to share knowledge effectively.
- Regarding cybersecurity assurance and privacy, organizations struggle to protect systems and protect from data leaks, zero-day vulnerabilities, and various other threats. Commitment from top leadership is necessary to ensure systems are secure and the proper policies and funding are in place. A Subject Matter Expert (SME) is needed within agencies to champion new technologies.
- Organizational structure should be developed to easily enable the use of data analytics and the ability to learn from that analysis. Data scientists and researchers should be able to build on existing systems, reuse past analysis, and perform testing and evaluation in an interoperable environment. SMEs are necessary to make sure organizations understand the problem, are watching and emulating academic research and, that the solutions developed match the problem.

- Another challenge for emerging technologies in big data is implementing proper data transparency. Organizations need to recognize when biases in data and technologies occur and be able to mitigate the effects.

2.1.2 Discussion Summary

The Emerging Technologies in Big Data & Analytics session focused on how to best leverage new technologies and avoid possible pitfalls. The majority of the discussion focused on cybersecurity and privacy assurance for emerging technologies. Many agencies are pursuing emerging big data technologies and analytics solutions to improve their operations. However, they do not always consider the emerging cybersecurity threats that could impact these systems after implementation.

There are several ways to help prevent cyber attacks. Cryptography can help in avoiding data leaks, but is not a full solution. One way to avoid zero-day vulnerability is to implement Generative Adversarial Networks (GANs) to identify vulnerabilities in the new technologies before adversaries try to infiltrate into the system. By embedding GANs into cybersecurity, the system can constantly test the system's results and regularly try to improve protection against adversaries. Other approaches include devising preventive measures to avoid malicious data from being included in the modeling training data, have a well-defined model life cycle with roll-back plans for reverting to the previous version, and producing secure government policies to prevent tampering of policies.

When deciding on which new and emerging technologies to adopt and implement, agencies should have SMEs involved in the decision-making process. When implementing these technologies, agencies should develop and implement plans for sharing the data and analytics while making the data transparent to allow reuse, easy modification, and reproducible results. The data sharing plan should motivate users to share their data, provide adequate documentation for understanding the data, and maintain the lineage and pedigree of the data to assure trust in the source of the data. Data transparency is defined as both data availability and openness. In an ideal world, the data used is shareable and the analytics developed are open.

Many agencies are pursuing emerging big data technologies and analytics solutions to improve their operations. When developing solutions using existing systems and adapting it to the problem, make sure that the existing systems are up-to-date and the solutions match the problem. Organizations should also adopt a phased implementation approach which includes a verification of the results are each stage. The data used for training, testing, and validation of the analytics must be properly sampled to avoid biases. The results should be

validated by SMEs at all stages of development.

2.1.3 Important Findings

- To avoid zero-day vulnerabilities, organizations can embed GANs into their cybersecurity system to identify vulnerabilities in the new technologies and protect against adversaries.
- SMEs within organizations are necessary to ensure the right technologies are being implemented to address the right problems.
- Agencies need to avoid developing biased analytics and models by proper sampling of the data.

2.2 Big Data & Security Technologies

The Big Data & Security Technologies session discussed the tools and technologies used to secure agency data and maintain the crucial privacy information of U.S. citizens. There are significant challenges to meet Federal government requirements to properly secure the vast amount of data that is transmitted and received by its many agencies. Session participants shared their unique agency perspective, critical concerns, and processes that are being applied to secure big data.

The session included discussions of the following:

- At what stages of the data lifecycle - from data capture, through data process, analytics, and information reporting - is the understanding of security requirements defined and technical solutions applied?
- Are the current data security governance and regulatory requirements being translated into actions to protect big data sets?
- Standards for big data security and protection are emerging from organizations such as the National Institute for Standards and Technology (NIST). Are there experiences or lessons learned from their applications?
- To what extent have organizations developed an awareness of solutions and applied technologies and products?

- Is executive leadership engaged in clarifying and communicating big data security protection requirements and empowering the institution of appropriate security measures?

2.2.1 Challenges

- Organizations need to develop and share best practices and approaches to data governance. The challenge of data ownership needs to be addressed in a manner that facilitates ensuring integrity for historical, legacy data, and data that evolves as it is processed. Data stewardship in the era of big data is critical in the face of data breaches and responses to other disasters.
- A lack of understanding by leadership and data stewards on proven techniques for protecting large data sets in a manner that is risk adverse and fully transparent was a noted in the session as a repeated challenge.
- The government needs a single security architecture or framework to enable cross-agency, public-private big data sharing. This solution should recognize that there are increased sensitivities with data as it is combined, linked, and privacy-preserving management strategies are applied. This was recommended in the recent Federal Cloud Summit [4].
- Although standards for big data security are emerging, such as the draft NIST-1500 series of documents, there is a lack of common understanding and consistent implementation of these standards [3].
- As additional access to large data sets is granted, the ability to join and infer sensitive details is increased. A key challenge is developing the proper standards and technologies for the data aggregation and inference issues associated with big data.
- Careful selection, implementation, and agile management (periodic, incremental updates) of security technologies for big data is a key challenge. Classes of tools that were identified included: data tagging, location/time/access permission enforcement controls, auditing and monitoring key sensitive data in real-time, including policy enforcement and pattern recognition.
- Evolving from a silo mentality of data ownership to an empowered Chief Data Officer (CDO) managed data sharing approach is a challenge, especially across government organizations, with the public sector, and in compliance with Federal laws and policies.

The challenge is to achieve a just-in-time data availability and delivery, with full security for data access and minimal / no manual intervention.

2.2.2 Discussion Summary

Chief Information Security Officers (CISOs) and security / Information Assurance (IA) organizations are being asked to get involved with securing big data systems. These systems apply data processing technologies through fundamentally different approaches to handle the volume, variety, velocity (3Vs) of big data from what has historically been used in traditional relational database management systems. The definition of big data processing is sometimes inconsistent or imprecise, further confusing the key performance parameters that need to be accommodated by the system security services.

However, as the CISOs and IA team are becoming involved with these big data systems, significant security and privacy concerns are clearly apparent. As large data sets are merged and moved within organizations, between agencies and externally to private sector organizations, the security protection concerns are further increased. In the face of Freedom of Information Act (FOIA) and other data sharing obligations, the consideration and development of security solutions is imperative.

During the discussion, several participants indicated and provided examples of personal privacy concerns associated with the increasing amount of data collected and used on company websites for purposes such as targeted advertising. In considering the unease with this sense of loss in personal privacy, remarks were made on the magnitude of how this would be amplified if government datasets containing vast amounts of personal or sensitive data were inadvertently or intentionally disclosed.

Current governance and regulatory guidelines exist, however, there is limited experience in applying these rules, particularly to nascent big data projects. Not only is there a need to protect sensitive data (individual and linked/combined data sets), the algorithms applied to analyze large data sets may be sensitive or proprietary. The collaboration session participants, as well as the overall community, is grappling with not only the approach to securing these big data sets, but also understanding and defining the required security for foundational components in big data processing systems.

Th NIST draft Special Publication 1500 series was identified as active work conference participants are tracking to apply to their big data projects. This consensus-based industry, academia, and government effort is defining reference architectures and frameworks which are vendor-neutral, technology-independent, and infrastructure-independent. This effort is viewed as an important component in understanding and managing risk and underpinning

security accreditation decisions as required by law, such as the Federal Information Security Management Act (FISMA) [1] and the Presidential Policy Directive-21 (PPD-21) [2]. The NIST Big Data Public Working group provides a forum to work through the relationships of data sets and the resulting security implications (e.g., records, demographics, financial data, etc.).

CISO and IA teams are starting to examine security technologies that can be appropriately implemented in an agile manner, where there is a collaborative effort to monitor the implementation of security services to meet requirements through an incremental process. The use of cloud services adds a dimension to the approach to securing data that is different from having structured, on-premise data. Classes of security tools discussed included: (1) data tagging, (2) role, location, temporal based access controls, and (3) real-time detection, protection, audit and monitoring of designated sensitive data, including policy/pattern based recognition. Some participants had very had very limited exposure to these types of tools, and the availability of examples through open source was briefly discussed. All agreed that executive leadership awareness of these types of capabilities and potential applications to use cases are needed. Security tools that are largely cloud platform agnostic is an important consideration.

Systems scale out to support big data, which further drives the need to automate security audits. This could be achieved using tools such as large security log datasets analytics that identify anomalous behavior and threat pattern detection.

Centralizing metadata (i.e., information about the big data sets) and challenges to automate metadata management were also brought up during the discussion. While technologies exist, there needs to be a substantial maturity of the governance process and role of data stewards to enable true value to be realized through consistent implementation

One participant highlighted the initial efforts within their organization to develop and adopt across-organization security framework and architecture for big data. However, across participants, silos of data ownership was acknowledged as more common. To realize big data effectiveness and efficiency, these silos need to be broken down. The role of a CDO needs substantial empowerment to centralize enterprise data and to enable redefining data ownership and sharing. Just-in-time data availability and delivery is critical and should be a fundamental principle in providing full security for such data access with minimal / no manual intervention.

2.2.3 Important Findings

- Security mechanisms must be in place to detect and prevent users from making mistakes. Even with extensive training, users are still going to make mistakes. No security

training program or security measure is error-proof and evolving attack strategies are increasingly difficult to thwart. Therefore, the training and protection measures need to evolve and improve over time.

- Senior executives need to understand the balance between the requirements and the value in data sharing versus the risk associated with exposing sensitive information.
- Integrated cross-agency approaches to big data security based upon a common data architecture and framework need to be further developed. Harmonization and consistent application of security services, such as where and how data encryption is applied, is starting to be discussed. These protections and data sharing agreements need to extend into the private sector, such as when the Department of Veterans Affairs (VA) needs to share healthcare data with private hospitals.
- Technologies for securing big data, either through cloud service providers, commercial products, or open source solutions, are recognized; however, organizations have not yet fully realized how to utilize them to achieve the required data set security.
- An empowered CDO is needed to address data ownership, including specifying what enterprise data must be centralized and shared in the context of big data. Derivative data classification, such as increased sensitivity associated with the aggregation of data sets, is not well understood by the community. Privacy Preserving Data (PPD) algorithms need to be vetted by data owners.

2.3 Big Data Landscape

The Big Data Landscape session brought together government and industry participants to discuss the landscape of big data and analytics for federal government agencies - at the present time and into the future. In this intimate setting, representation from federal agencies was limited and consisted of two distinct perspectives regarding mission oriented big data and analytic solutions. Sponsors and influencers, as well as current and/or future users, participated in this session.

The session included discussions on the following themes and topics:

- What is the current status of the federal government in regards to big data?
- What budgetary constraints and challenges exist and how does this impact enterprise thinking?

- How can big data concepts be applied to very tactical needs?
- How do practitioners demonstrate the value of big data within the government?
- A cultural shift is needed within organizations to maximize the impact of big data.

2.3.1 Challenges

- The current challenge in big data is learning how to leverage data to find solutions within the government. The value of big data can be demonstrated by providing operational context.
- The security, volume, and variety (temporal, linguistic, geographical) of the data all need to be addressed to work effectively with big data. Organizations need to architecturally design and establish a platform with broad uses across the agency.
- To enable greater sharing, data-security policies need to be consistent across agencies, yet still satisfy agency-specific needs.
- Upcoming fiscal budget cuts make it hard to continue working on the IT enterprise initiative and demonstrate what problem needs to be solved.
- The culture surrounding the use and applications of big data need to be addressed so that the strategic focus aligns with big data outcomes.

2.3.2 Discussion Summary

The session began with a discussion of the current state of the Big Data Landscape. Per the discussion, 5-10 years ago the big question in big data was: "How do we make data available?" However, now practitioners talk less about data accessibility and focus more on how to extract value from and put operational context to big data.

The discussion then moved on to data lakes, mission-oriented analytics, and fiscal challenges. Different agencies having different issues: making the data available to do forensic detailed analysis, working with a lack of expertise on big data, and coordinating with ALternative EXperts (ALEX) where organizations work with data scientist from universities. The data architecture is a complicated issue to overcome. How do organizations move from the legacy ways of collecting and releasing the data to the concept of data lakes? How can they deal with the feeling of loss of control over the data? Providing authorization to the right people within the organization, as well as outside of the organization, seems difficult as

organizations move towards data lakes or data-centric architectures. Organizations need to ensure they are investing in the right technologies to address their needs.

Participants in the session discussed how to put big data into action within their agencies. At the Federal Aviation Administration (FAA), the platform that supports the air traffic controller is highly tactical and supports a lot of automation. n performing traffic management, before the airplane takes off a lot of analysis is done to communicate the impact of weather, route changes and more. This is all presented in a consumable form for the controllers. Operationalizing the big data to be used with proper functional context is a long and difficult process.

Additional participants discussed their agency's publication of white papers on rule impact. The data is shared with the public at a very high level since there is a need to protect the stakeholders.

Data received from multiple sources comes with countless challenges - unstructured data, data in varied formats, and data with foreign language components. Data from different sources is difficult to merge for meaningful analysis. The data quality is also an essential piece of the data landscape and a huge challenge for organizations. Identifying risks and the overall cybersecurity of the systems is a priority in big data as well as the international regulations regarding the data. Enterprise data management, content management, interpretation of data, predictive analysis and high-frequency analysis are key areas that need to be understood and implemented.

The conversation quickly turned towards declining fiscal budgets and how organizations need to prioritize tasks that fit the mission. Organizations need to focus on aligning their strategic plan to meet the data management needs. There is need to strategize, align, and communicate a multi-year plan. There must be an IT plan that considers end users and the impact while giving a better public view.

Traditionally, IT was built over years in silos in agencies. An enterprise approach is the new standard approach to IT. Organizations strive to get value from combining multiple sources, connecting the dots, and capturing data changes. However, this is a shift in the culture and mindset of organizations. One of the reasons for silos is to restrict data access. The publish/subscribe method is used to stream data and access control is built at that level. Automation in data delivery is not an issue, but there are other sets of problems that arise from distributing all the data. Organizations need to ensure their policies are strong enough to handle this change. Along with investing in real-time systems and analysis, the other important components are intelligence and the ability to create systems that work concurrently. Collaboration across agencies to accelerate innovation in a budget austere

environment is a challenge.

Ideally, if the big data community can define the problems upfront and prioritize by return on investment (ROI) there might be the option to tack on more funding. This solution does not work for all organizations. Culturally, people in government are trained to think in terms of cost, schedule, and performance. Now, organizations are being told to risks and think in an enterprise capacity. There are two approaches to big data analytics: bottom-up and top-down. The bottom-up approach is an enterprise strategy: solve the problem that demonstrates value. The top-down is a new multi-year IT strategic plan.

Additionally, organizations should recognize that people on the ground know what the problems are. A massive culture shift is needed in the middle management band to keep up with the big data movement. Culture change in the middle is where that needs to be communicated. Moving forward with technology, it is not enough to simply crunch the numbers. Someone also needs to tell that success story to the middle management band. Technical and IT practitioners are not always the best at telling the story and tying in quantitative and qualitative benefits. Problem statements do not usually include computers and data and organizations need to determine how to build business cases around these solutions.

2.3.3 Important Findings

- The availability of data is no longer the sole issue in Big Data. The current and future challenges lie in establishing a strategic approach, implementing the right infrastructure and tools for the mission, and demonstrating success in solving real-world mission problems.
- Data governance is a key enabler to effectively integrate and analyze vast data sets. This requires enterprise-level thinking and a combination of policy, process, people, and technology. This includes a strong commitment from leadership.
- Data-centric agencies are increasingly blurring the lines between business/domain experts and technical/analytic experts-with positive results.
- A cultural shift is needed to expand decision-makers' willingness to take calculated risks and think in enterprise terms. Additionally, these decision-makers should focus narrowly on the scope, schedule, cost, and performance of any one organization or program.
- Agencies with interests in big data analytic solutions that do not have a strong technical focus and resources are outsourcing or partnering with industry, organizations, or

universities.

- Agencies that implement both top-down and bottom-up approaches to addressing mission-oriented big data challenges should focus on communication and obtain buy-in at all levels to increase the likelihood of success.

2.4 Leveraging Big Data for Mission Success

The Leveraging Big Data for Mission Success session discussed the mission context for big data by examining several audience-inspired use cases which focused on identifying the business need(s) for data analytics; identifying data sources and data locations; identifying data access restrictions; educating the workforce on the value of big data analytics; and developing the skills, knowledge and resources required to perform the analysis. The discussion included scenarios from the Bureau of Land Management (BLM), The US Immigration and Customs Enforcement (ICE) Curricular Practical Training (CPT) program, the Environmental Protection Agency (EPA), and a general scenario discussion related to health care. While the unique business needs for each scenario were discussed, several common themes emerged from the discussion.

The session included discussions of the following:

- What is the business need for the analysis? How can the data be combined to "tell the story"? For example, given a specific business need, how can the relevant data be combined to validate the business need and support decision making? How can big data analytics be used to improve operational process efficiency and effectiveness?

- What are the available sources of data? In some cases, the required data is available within the organization (i.e., internal data sources). In many cases, data is required from both internal and external sources. Note that many government organizations are both data consumers and data providers. External sources may include other government organizations, non-governmental organizations, commercial organizations (including vendors and consultants), and academic research institutions.

- What methods and policies are needed to facilitate access to restricted data? For example, restricted data sets (e.g. law enforcement sensitive (LES) data) require special access controls.

- What methods can be used to convey the value of big data analytics and improve organizational awareness?

- How can barriers to data sharing be minimized? How can organizations develop and adopt a comprehensive taxonomy, with the associated metadata, to facilitate internal and external data sharing? When data is shared between organizations, how can the privacy, security, and integrity of the data be preserved?
- How do organizations get started with data analytics?

2.4.1 Challenges

- Much of the available data is unstructured, making it difficult to combine data for analysis.
- Data sharing appears to be a problem both between internal groups and with external organizations.
- Education is lacking. There is a lack of understanding of the benefits big data analysis can provide to organizations. This ranges from the executive level to the level of scientists and researchers within the organization. Knowledgeable and skilled personnel are needed to distinguish between good (useful) information and nonsense. Similarly, skilled personnel are needed to correctly interpret the analysis results. Skilled personnel includes personnel with detailed knowledge of the mission and personnel skilled in the use of the various data analysis techniques (e.g., data scientist). There is currently a challenge in identifying, hiring, and retaining personnel with these skills.
- There appear to be limited communications between data analysis personnel within government organizations. For example, how often do Chief Information Officers (CIOs) and CDOs get together to share information?
- Data stewards are often viewed as hoarding and being overly protective of data, preventing the sharing of information.
- A shared vocabulary is missing in many organizations. As a result, it is difficult to combine data for analysis and share data with both internal groups and external organizations.

2.4.2 Discussion Summary

Key themes that continued to come up during the discussion were developing a business case; identifying sources of data; developing a shared lexicon along with relevant metadata;

developing in-house skills for performing data analysis; and creating policies and procedures that facilitate data sharing.

Starting with a good business case was determined to be a high priority since it shows the value in performing the data analysis activities. It was also pointed out that leadership support is key. A "champion" is needed to be a strong advocate and provide the support and resources needed to accomplish the data analysis objectives.

Identifying sources of data is also a necessary activity for organizations. In many organizations, data are stored in multiple disparate repositories, which complicates the data analysis process. Furthermore, the required data may include internal and external data sources. Identifying the sources and gaining access to the required data elements should be a part of the planning process. Once the data sources are identified the data may be collected into a single repository. Data cleansing and pre-processing may be required to ensure the data is formatted and structured to facilitate the data analysis.

When collecting data from multiple sources there is a chance that the metadata will vary. For example, one organization may refer to a telephone number as "cell". Another organization may refer to the same telephone number as "mobile". A shared lexicon would be helpful for identifying the data elements required for the analysis. Developing a shared lexicon at the enterprise level would provide a reference resource for the entire organization as well as the relevant external stakeholders.

Since the data analysis is focused on improving the mission, skilled personnel are needed who are knowledgeable of the mission and skilled in the various data analysis techniques. This is key if the data analysis results are intended to support the organization's mission success elements. For example, the mission success elements for one of the organizations discussed include environmental procurement measures and public health impact measures. Government SMEs who understand procurement rules and environmental policies are needed to define the problem, define the relationship between the mission success elements and the analysis, provide input to the data analysis, and help interpret the analysis results.

Knowledgeable data science personnel are required to identify, procure, and operate the tools required to perform the analysis. Data science personnel also assist in relating the analysis results to the mission success elements. One recommended method for working with SMEs and data science personnel is to develop use cases. Use cases can be used to help define the problem, relate the problem to the mission, determine the data requirements, define the data analysis methods, determine the resources required to perform the analysis, and provide the context for interpreting the analysis results.

The key roles discussed for the analysis team include the analysis team champion; project

leader; data analyst; data steward; data scientist; a representative from the end user community (i.e., consumer of the analysis results); and a technical lead. Some of these roles may be combined based on the organization's needs and the analysis objectives.

Finally, updated policies and procedures are needed which facilitate data sharing (including the sharing of sensitive data) within organizations and with external organizations. Part of the problem is that existing policies and procedures were established without big data analytics and data sharing in mind. Revised policies and procedures are needed to encourage data sharing while providing the necessary data security, protecting privacy rights, and preserving data integrity.

2.4.3 Important Findings

- A good place for organizations to start with big data is to develop a business case for employing big data analytics. The business case should include the relevant organization mission elements, the purpose of the analysis, and the benefits the analysis will provide the organization and the stakeholders. The business case may also include use cases that help to explain the purpose and utility of the analysis.
- Leadership is key. Organizations need to identify a champion who is knowledgeable of the organization and the mission, has a vision for how the analysis results will benefit the organization, is highly motivated, and is able to provide the resources needed to implement and perform the analysis.
- Communication is very important. Government agencies need to continuously educate organization leaders and stakeholders on the benefits of the data analysis activities.
- Manage expectations. Prioritize analysis activities and communicate the priorities with leaders and stakeholders.
- Start with a small problem that will have a big impact when solved. Develop a 6-month proof-of-concept (PoC) that can be used to show the results and benefits of the analysis activities.
- When defining the data required for the analysis, identify the data sources, develop and standardize the metadata and, document this information in a data dictionary. Also, ensure leaders and stakeholders agree on the information provided in the data dictionary.

- Develop a data sharing culture to realize the benefits of working with others within the organization.
- Develop more inter-agency working groups to facilitate the sharing of lessons learned (both successes and failures). Learn from others who are currently reaping the benefits from big data analytics results. This includes learning about the tools that are used to provide a data repository, perform the analysis, and present the results (e.g., visualization tools).
- Establish the governance required (including streamlined policies) to facilitate data sharing while ensuring data privacy, security, and integrity.

2.5 Big Data & Healthcare

The Big Data & Healthcare session discussed the implementations and future solutions of big data in government healthcare organization. big data is characterized as having high volume, velocity, variety, variability, veracity, and value. The Department of Defense (DoD), the VA, and the private sector continue to generate large amounts of data related to healthcare. This data can be mined to provide actionable insight to improve service to service members and veterans alike regardless of the point of care. The flow of shared data and shared analytics needs to cross the three constituents efficiently and economically with the goal of providing improved patient outcomes. Even within each constituents' area, there will remain some variations in technology, data, and analytics platforms, standards, and capabilities.

Meeting emerging data needs requires new technologies and analytics. The Defense Health as well as Veterans Affairs not only work with their respective architecture pillar but also with multiple external vendors to provide quality care for their respective patient population. Shared analytics is one way of leveraging standardization among big data users.

Sharing analytics by "prescription" may enable better privacy and security via mobile health platforms. Discussions are underway regarding standards and adherence to regulations put forth by the regulatory body of government.

The session included discussions of the following:

- What are the practical strategies for minimizing and removing impediments to shared data and shared analytics?
- What emerging trends will allow us to extract maximum value from the big data that each constituent group possesses?

- What concrete steps can we take to realize the benefits of shared data and shared analytics?

2.5.1 Challenges

- Organizations struggle to develop strategies to minimize or remove impediments to sharing data and analytics across involved parties.
- Government organizations need to study emerging trends to extract maximum value from big data.
- Concrete steps are needed to realize benefits of shared data and analytics.

2.5.2 Discussion Summary

This session was heavily focused on minimizing and removing impediments to data sharing and moving government agencies forward in the realm of big data. Strategies for minimizing or removing impediments to sharing data and analytics proffered included using standard data taxonomies and implementing data governance. Selecting and using a single data standard was mentioned as a strategy to reduce the complexity of having to support multiple standards. Participants pointed out that data governance was required to better curate data and to simplify data sharing and analytics. It was noted that if all patients eventually to take ownership of their own data that it might eliminate the need for organizations to curate the Personally Identifiable Information (PII) and Protected Health Information (PHI) data. In other words, the patient brings their data to the healthcare provider.

Potential solutions to removing data privacy and protection concerns included: shared analytics which focuses on pushing analytics to individuals and allowing the data to remain at the source (to mask and remove PII) and share only the analytic results. Educating stakeholder on HIPPA compliance was suggested to mitigate their reluctance to share data because of uncertainty about HIPPA compliance. Data silos hinder data sharing and many of the participants agreed that eliminating data silos was a potential solution. The group discussed the value of having an exemplar of a successful cultural shift to sharing data and data analytics for others to emulate. Removing political impediments by providing a compelling and easy to comprehend business case for sharing data is necessary. Providing training on data sharing and data analytics and analytic tools would help change the culture to share data. Some participants discussed moving toward free market-driven solutions and away from government-led solutions to change the culture for sharing. It was noted that health

information exchanges move away from charging for data to charging for doing the analytics. Transitioning to applications that are data agnostic would encourage data sharing. Currently, data is tied to applications.

The discussion of emerging trends to extract maximum value from big data included several themes. Structured data capture, re-engineering workflows, and one-time collection were discussed as ways to reduce the burden of data collection. Some participants noted that an emerging trend in healthcare is to increase the use of drop downs and box clicks would allow clinicians to collect lots of details and then simply write 2-3 sentences to complete patient notes. Technology trends mentioned included the hosting of Hadoop data registries, injection of technology to determine presence or absence of conditions, wearable health monitors to capture data, graph databases to improve search performance and the ability to identify relationships in the data, and the use of artificial intelligence and machine learning to uncover hidden data trends. The participants suggested that younger people will likely want to own their health data. This means a change in quality and integrity of data analytics because the patient controls their data and can decide what to share and with whom.

Concrete steps to realize benefits of shared data and share analytics were discussed, as well. Frequent themes included the need for standard data models, analytic models, technical reference models, and analytic registries. Providing appropriate benchmarks to show how providers stack up against their peers would help stakeholders make better decisions about acquiring big data and data sharing solutions. Participants indicated that use of Public/Private Partnerships would help to realize the benefits of shared data and shared analytics. The need for a more holistic approach to data and big data solutions was discussed at length. The main argument was to stop relying on applications that use proprietary data schemas and to move toward those that are data schema agnostic. Organizations should have specialized applications based on practice (e.g., surgery, pediatrics), but should not specialize data for an application. Another step for organizations discussed was the need to provide funding to support adoption and transition to standards and interoperability. Participants indicated that empowering a Chief X Officer (CXO) to lead the shared data and shared analytics change is required. Since VA patient records need to be retained 70 years since the last encounter, a "long-term view" of data and systems is needed. It was suggested that a long-term view is defined as 125 plus years. Another proposal discussed was for the government to create middle-ware solutions with hooks to proprietary solutions rather than being dependent on proprietary applications and data standards developed by vendors. The longevity of interoperability was also raised as an issue. For example, what will happen after the Cerner 10-year contract ends? Will DoD and VA be locked into Cerner's proprietary data model and need

to pay to get the data back out? Since health analytics focuses on finding relationships, it was suggested that graph databases might help organizations realize the potential benefit of shared data. Additional steps to realizing benefits of shared data and shared analytics might come from looking at what the industry is doing (e.g., sensor data, finance, oil companies, casinos, Disney) and adopting relevant lessons learned for healthcare. It was acknowledged that the healthcare data set is very different and complex.

2.5.3 Important Findings

- Strategies for minimizing or removing impediments to shared data and shared analytics require:
  - Explaining and demonstrating the business value of data sharing and shared analytics.
  - Arriving at agreed on data taxonomy, standards, and processes.
  - Changing the culture to encourage data sharing and shared analytics via change management and training.
- Emerging trends that will allow us to extract the maximum value from big data that each constituent group possesses include:
  - Training and educating employees on how to use data and demonstrate value
  - Utilizing artificial intelligence and machine learning techniques to gain insight from data.
  - Developing structured data capture processes and restructuring workflows.
  - Changing the culture to encourage data sharing and shared analytics via change management and training.
  - Patient ownership, responsibility, and stewardship of their own healthcare data in the future.
- Organizations need to take concrete steps to realize the benefits of shared data and analytics by taking the long-term view of solutions (125+ years), implementing Chief "X" Officer roles, taking a holistic and application-agnostic approach to data, and by adopting and using new emerging technologies,

3 SUMMIT RECOMMENDATIONS

Many common themes were apparent across several of the challenge areas: data governance is highly valuable and touches all aspects of big data; data sharing standards need to be put in place to allow for easy collaboration; researchers and senior executives alike need to understand the requirements, value, and impact of big data; security is key to protecting data and trusting big data systems; and recognizing the mission and understanding the value of the mission can help drive analysis and provide the biggest impact.

All of the collaboration sessions noted the importance of data governance; this is valuable and necessary across all government agencies. A full data governance strategy requires policy, process, people, and technology. Organizations need to have the right policies in place to protect the data, ensure data access, and ensure data is being used in appropriate ways. In order for organizations to be highly effective and efficient with big data, they need to have quick processes in place to streamline the policies in place. Without knowledgeable data scientists with access to the right tools, mission impact is much harder to obtain.

The value of education and knowledge at multiple levels of organizations is needed for organizations to progress. Senior-level executives need to understand the value of big data as well as understand the balance of data sharing and security. Empowered CDOs should address ownership and should continue to work against the data silo mentality. Big data is a new aspect of companies and organizations need to start building this into their long-term strategy. Without executive buy-in, organizations will begin to lag behind progress in big data.

Not only does corporate leadership need to have an understanding of big data, but researchers within the organization and middle-level management need to be educated on big data. Employees need to know how to gather value from big data and this value and the practices need to be able to move up the organization's management chain. In addition to research staff and data scientists, organizations should employ SMEs who are highly experienced in big data technologies, tools, and applications.

We have moved past simply collecting data, we now need to make sure we're solving the right problems and the right people have access to this data. There is often fear in organizations when it comes to data sharing and agencies need to find a balance in the risk of data sharing vs. the reward of mission impact and value derived from the data. Data sharing requires easily understandable data governance policies and processes in place to make sure the right people have access to the right data. Organizations need to move away from building small data silos within departments and should work towards developing an

organization-wide, or even an inter-agency, data strategy. Executive buy-in is important to demonstrate the business value of shared data.

Government agencies need to continue to work together to form cross-agency approaches to big data. This is especially important for data security and data sharing. Many organizations have an interest in similar data sets and can derive mission value from the same data sets. Enabling cross-agency data sharing brings value to organizations. Additionally, government organizations should participate in multi-agency working groups to facilitate sharing lessons learned, tools, analysis, and results.

Security is one of the most important aspects of big data. Strong data governance regarding security policies and processes should be in place to protect data, researchers, and organizations. It is important for agencies to adopt cutting-edge tools and technologies to perform research, while also protecting the data. Organizations also need to have strong policies in place to protect data even when users make mistakes.

Finally, an important topic discussed across all collaboration sessions was how to use data to help government organizations address their mission. Agencies should begin with a clear business case for the data research and analytics. The possibilities with big data analytics are endless and organizations need to make sure they prioritize their research projects. When a research project has a clear goal, and the right data policies and procedures are in place, it should be straightforward to accomplish mission success using big data.

4 CONCLUSIONS

The December 2017 Federal Big Data Summit reviewed many challenges facing the federal government's adoption of big data technologies and techniques. These challenges spanned multiple collaboration areas and were widely discusses by all groups, as well as during the morning's panel sessions. Specifically, enabling data sharing within and external to organizations, providing and implementing clear governance and policies, educating the entire workforce on the value of big data, and keeping up with cybersecurity needs remain challenges. Developing policies and easy avenues for collaboration, implementing SME roles within organizations, and top-to-bottom security implementations can help to mitigate these identified challenges in the future.

While the December 2017 ATARC Federal Big Data Summit highlighted areas of continued challenges and barriers to progress, the Summit also cited notable advances in mitigating these perennial challenges. Organizations have a more complete understanding of big data as an overall concept, now questions move on to how to best use the data to address mission

needs, how to protect the data, and how to share data for greater impact. Agencies are moving forward with much of the technical progress, but still need guidance on how to effectively implement programs to allow for quick development and collaboration across agencies.

From the recommendations made in the collaboration sessions, government practitioners (at all levels of government) should participate in special interest groups or working groups to increase collaboration; continue to influence standards development within the discipline, and continue to partner with academia to leverage cross-cutting research and to help train the government workforce. These activities will further mitigate the perennial big data adoption challenges cited by the participating big data practitioners

ACKNOWLEDGMENTS

The authors of this paper would like to thank The Advanced Technology Academic Research Center and The MITRE Corporation for their support and organization of the summit.

We would also like to thank the session leads and participants that helped make the collaborations and discussions possible.

REFERENCES

[1] Federal Information Security Management Act of 2002, Tit. III, E-Government Act of 2002, Pub. L. No. 107-296 (Tit. X), 116 Stat. 2259; Pub. L. No. 107-347 (Tit. III), 116 Stat. 2946. 44 U.S.C. Ch. 35, Subchapters II and III, codified at 40 U.S.C.11331, 15 U.S.C. 278g-3 & 4 (full-text).

[2] United States., Obama, B., & United States. (2015). Presidential Policy Directive 21 Implementation: An Interagency Security Committee White Paper. Washington: U.S. G.P.O.

[3] National Institute of Standards and Technology. (2017). Big Data Information. Gaithersburg, MD: U.S.

[4] J. F. Brunelle, S. Anand, G. Barmine, M. Spina, K. Warren, A. Winston, M. Javid, A. Kemmer, C. Kim, S. Masoud, T. Harvey, and T. Suder. August 2017 ATARC federal cloud & data center summit report. Technical report, The MITRE Corporation; The Advanced Technology Academic Research Center, 2016.