FEDERAL SUMMITS

DECEMBER 2015
FEDERAL BIG DATA SUMMIT REPORT[*]

January 20, 2016

Christine Harvey, Dr. Laila Moretto, Bob Natale,
Dr. Haleh Vafaie, Irina Vayndiner, Nicholas Hamisevicz
The MITRE Corporation[†]

Tim Harvey and Tom Suder
The Advanced Technology Academic Research Center

January 20, 2016

[*] APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. CASE NUMBER 16-0155. ©2016 THE MITRE CORPORATION. ALL RIGHTS RESERVED.

[†] The authors' affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the authors.

Contents

EXECUTIVE SUMMARY

The most recent installment of the Federal Big Data Summit, held on December 8, 2015, included four 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 research and technologies. 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 four challenge areas in big data: Governance in Big Data, Big Data Integration and Management, Architecting Systems for Big Data, and Data Science and Scientists.

This white paper summarizes the discussions in the collaboration sessions and presents recommendations for government and academia while identifying orthogonal points between challenge areas. The sessions identified detailed actionable recommendations for the government and academia which are summarized below:

- Flexible data architectures and standards are necessary for the government to operate in the fast-paced technical environment. Government agencies need to work on becoming more agile in developing and piloting programs to keep up with the progress of industry.

- Inter-agency collaboration is becoming increasingly more possible and valuable with big data and other technological advances. Many roadblocks are still in place that bar data sharing and collaboration. Standards must be put in place to allow for simplified collaboration between, and even within, agencies.

- Every agency needs to have a solid and concrete understanding of the needs and requirements of big data. All involved parties in big data programs should understand the governance requirements of the efforts in place.

- Data science is an exciting field with incredible amounts of progress and research in recent years. The federal government needs to recognize this and prioritize hiring and education in this specialty.

1 INTRODUCTION

During the most recent Federal Big Data Summit, held on December 8, 2015, four 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 would 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 work force 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 four 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:

- Governance in Big Data
- Big Data Integration and Management
- Architecting Systems for Big Data
- Data Science and Scientists

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

2.1 Governance in Big Data

The Governance in Big Data session discussed the unique challenges and benefits in attempting to impose regulations, provenance, and governance.

The session included discussions of the following:

- What is governance?
- What do organizations need to do in order to maintain proper data governance?
- What is the appropriate level of data provenance within and across organizations?
- How does data governance and stewardship apply to and impact the security of the data?

2.1.1 Challenges

- There is a lack of understanding of basic governance, such as definitions, framework, processes, etc.
- The word “governance" causes cultural barriers. Understanding the value of governance and how it contributes to organizational success in big data access, analytics, and security is inadequate.
- There is no existing agreed upon standard for provenance and governance across agencies.

2.1.2 Discussion Summary

First and foremost, the collaboration session participants generally agreed that governance is important to their agencies and across the federal sector. Participants also agreed that despite this priority, governance standards are lacking across the federal agencies. Even at agencies with governance programs, participants reported that these agencies are small; governance efforts are either led by a small team or a single individual. The participants quickly recognized that the definition of governance varies across agencies. Each participant identified a different definition of governance and all noted that there is not one agreed upon definition across agencies. The definitions provided for governance covered themes such as:

- Rules for managing the assets and their applications
- Principles, policies, processes, people
- Understanding the data assets, data security, and achieving consensus

Participants cited various scenarios where governance is considered a compliance program; as a result of this belief many programs have a negative perception of governance. Participants agreed that governance causes cultural barriers and this impedes the agencies from working together towards better solutions. Many attendees remarked they prefer to avoid the use of the word “governance” due to the negative implications. Participants advocated a consensus approach whereby governance could be practiced from the point of view of “collaboration” as opposed to a “compliance” point of view.

The discussion also covered the organizational maintenance of proper data governance. The participants agreed there must be confidence in governance processes and leveraging what others have done (e.g., Data Governance Institute). A reliable catalog of data and metadata would increase this confidence. Another aspect of maintaining proper data governance was attributed to aligning resources with requirements and understanding the drivers of the organization.

In regards to provenance, participants agreed that maintaining provenance in metadata is critical to ensuring data are used correctly, especially for discovery in big data and cloud environments. However, participants were unsure how enterprise level provenance could be established for big data. Participants did not know what would be the appropriate level of provenance within their agency as well as across agencies. Session attendees agreed that for data provenance to be managed effectively across organizations, agencies would need to mutually understand their objectives and processes when collecting and using data. Additionally, context, standardization, and requirements are equally important for agencies to effectively manage provenance.

Participants noted that data governance and stewardship apply to and impact the security of the data. Most agreed that 100% security is difficult to achieve. Session attendees did acknowledge that breaches will happen and it is critical to prepare for these occurrences in order to successfully recover. Participants discussed various security strategies such as divide and conquer, key rotation, data masking, etc.

Two major issues that touched every challenge discussed in the session were budget and security. Session attendees agreed that security should be considered as early as possible in any data governance and provenance program. Additionally, participants felt strongly that the budget for projects generally do not include funding to implement governance and provenance. Without designated funds, agencies will not have the full benefit of these critical areas.

2.1.3 Important Findings

- Agencies should adopt a concrete definition of governance and develop a framework for governance. Clarification of this process will provide a common foundation and a solid starting point for further work.
- Standardized metadata for big data provenance is necessary in order to easily provide information to interested users.
- Budget and security are critical infrastructure components for the success of governance and provenance. These two concerns must be considered early and often.

2.2 Big Data Integration and Management

The Big Data Integration and Management session discussed the difficulties of combining data from a variety of sources in an efficient manner.

The session included discussions of the following:

- What are the best practices for integrating data from various agencies? What agencies can provide success stories?
- How does privacy and security affect the ability to integrate disparate systems and how can these difficulties be mitigated?
- What are the best tools and resources for data management?

2.2.1 Challenges

- Despite a desire to share data, organizations may face political, legal, and cultural challenges; therefore they need to have strong leadership to drive change.
- Historical data can be difficult to integrate if there is an incomplete understanding of the data transformations performed at the time the data was initially implemented. Without sufficient documentation, proper integration can be difficult to impossible.
- The quality and standards of data from different data sources, even within agencies, varies.

- Unstructured data, such as PDF files and images, are harder to integrate. This type of data is also more difficult to clean, especially when data origins are external to the agency or there is insufficient documentation.

2.2.2 Discussion Summary

Session participants agreed that federal agencies have a wealth of experience integrating multiple data sources of various types and origins. Discussion during this session reviewed how to handle disparate data sources. Participants debated whether it is preferable for agencies to integrate data sets or to keep separate sets for various uses and agencies. Both of the possibilities were explored and discussed by the participants.

Trust in the data cleaning process is also a concern for agencies. Participants reported that users will discard data sources if they fail to produce a correct result when the user knows the result. This false reporting leads to a lack of trust between users and the data sources.

While the concept of data sharing is embraced in most government agencies, implementation across agencies has taken much longer. Data sharing may requires changing policies on data control or even the culture of data governance. Openness and interoperability often clash with privacy and security and security tends to trump attempts at data sharing. New or expanded initiatives can be hampered by budgets and cost-cutting efforts. If an agency has historical or legacy data, time, resources, and funding is required to move the data into a "modern" data store.

When inter-agency sharing was discussed by the participants, the first item addressed was the type of data to be exchanged. There must be a clear understanding of what data is being shared and the goals the sharing is meant to accomplish. Some participants expressed a desire to have a data exchange protocol in place to address data interoperability issues. These standards should specify acceptable uses for the data as well as data history. Processing steps for all data should be clearly explained and shared with all involved parties for a clear understanding of what assumptions and data transformation were involved in the process.

Access to data was an additional noted area of concern. In many situations, users are either given complete access or are completely blocked from accessing the data. Some participants felt it would be better to have partial access to data rather than none at all. The government agencies should provide guidance on partial or limited use access for newer data acquisitions.

Most participants acknowledged the need for metadata to describe the history, or provenance, of the data, especially when dealing with integration. The metadata should define the rules for handling and using the data, as well as data provenance. One of the main concerns expressed by session participants was the changing of data formats during data delivery or transformation, therefore breaking the existing process. It was noted that the inherent nature of data transformation processes makes it difficult to respond rapidly to sudden changes in data format.

Agencies are starting to recognize that changes need to happen rapidly. Often in an agency, there is a single IT group and all others in the organization are their customers. Customers need the support of the IT group to get to access and utilize the data. The “Next Gen” agencies are realizing they are able to accomplish much of the reporting themselves without reaching out to IT, therefore self-service is becoming more prominent.

2.2.3 Important Findings

- Many participants had concerns that integrating historical data could lead to possible misuse due to the incomplete understanding of the legacy of that data.
- In industry, success in data integration happens through quick pilots, which is much more difficult to accomplish in the government environment.

2.3 Architecting Systems for Big Data

The Architecting Systems for Big Data session sought to outline recommendations and best practices for building big data processing systems that perform effectively, efficiently, and affordably in operational mission contexts.

The session included discussions of the following:

- What are the best practices for system architecture in big data?
- How are big data solutions best architected for operational relevance – mission effectiveness, performance, and scalability?
- How can cloud service delivery models (Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), Machine Learning as a Service (MLaaS)) be used in critical-mission big data solutions?
- How can legacy data analytics applications augment modern big data solutions as well as vice versa?

2.3.1 Challenges

- Big data is generally distinguished by some combination of exceptional volume, velocity, variety, and variability that requires non-traditional data management and processing capabilities and capacities that, in turn, often require updated and specialized technical skills and tools.

- Big data sources and collections must often be used in conjunction with more traditional data sources and collections for which appropriate tools have been developed and optimized over long periods of time. In these cases, specialized big data tools are potentially not as effective or efficient.

- Budgets are generally tight; large investments over significant periods of time must be planned carefully and executed adroitly; opportunity costs must be understood and duly considered; business/mission value objectives must be clear and compelling; success projections must be calibrated against delivered results.

- Compute, storage, and networking infrastructure models continue to evolve via virtualization through cloud computing to software-defined everything. This evolutionary trend generally improves optionality and may lower costs. The (effectively) constant change dynamic introduces investment risk relative to mission delivery time frames.

2.3.2 Discussion Summary

On the topic of best practices for architecting big data solutions, the session discussion reflected the emphasis on mission value and improved decision making that marked nearly all of the panel sessions from the morning portion of the Summit. The participants spoke first about the need to define the big data architecture's purpose, scope, and granularity relative to specific business or mission value. The session attendees discussed the need to architect for the particular combination of volume, velocity, variety, and variability characteristics represented by the applicable data sources to be used by the application. They noted that the architecture must encompass hardware, software, and services; people, processes, and technologies; and span the complete data lifecycle from collection and storage, through governance and management, to business/mission value delivery.

Participants observed that model-based (as opposed to document-based) architectures can enhance value over the big data implementation lifespan and data preparation should be optimized to meet specific business goals or mission objectives.

The discussion identified a variety of pitfalls or “causes of failure” in big data projects, including:

- Not piloting first.
- Not using agile methods.
- Not understanding the available data.
- The lack of an empowered champion.
- Not using the federal enterprise architecture.
- Re-inventing the wheel.
- The lack of data stewards (or data stewardship).
- The lack of engaged business representation.
- Lack of experience and requisite technical skills.
- Lack of “shared pain” (across the entire stakeholder space).

The session participants recognized that many of these factors are not unique to the big data domain and that multiple factors often occur together and may even be mutually reinforcing.

Data location considerations (local or remote, centralized or distributed) and associated performance characteristics are important and are sometimes predetermined constraints and sometimes negotiable trade-space options. Similarly, the matter of whether data should be normalized or otherwise tuned in-place for specific application needs or stored in raw formats to allow for application flexibility must be resolved and supported by the architecture. These considerations and characteristics can impact system and application scalability, performance, and usability.

On the topic of architecting for operational relevance, the session participants noted that system availability must be commensurate with mission requirements and performance characteristics must support required decision pace and timeliness. It was observed that intelligent data staging techniques can be used to enhance performance for timely decision making. The probable importance of data sharing agreements for many big data applications was called out, along with awareness of the associated security and privacy concerns that must be addressed.

Concerning the role of cloud services for big data solution architectures, the general consensus was that utilization of IaaS facilities for storage and compute aspects of big data applications is well-established. On the other hand, PaaS offerings need to mature to provide enterprise-grade data collection, curation, and analytics solutions. Recent commercial and open source MLaaS announcements suggest future opportunities on this highly specialized front. Early and planned SaaS implementations of data analytics services are promising in the near- to mid-term.

Lastly, the session addressed the role of legacy data, tools, and applications in the new age of big data. The overall consensus was that legacy data - defined here as any data not qualifying as big data per the notional “four Vs” characteristics - can (and probably should) be used as a baseline and/or building block for big data applications. Indeed, in many cases, such legacy data can be essential for a complete picture and, therefore, accurate results and maximum mission effectiveness. At the same time, the group concluded its discussion with the observation that big data analytics might very well reshape longstanding operational and policy norms.

2.3.3 Important Findings

- Big data architecture for operational systems in the federal domain must be responsive to business goals and mission objectives above all else.
- Big data architecture must encompass all relevant aspects in a consistent manner to enable solutions that are optimally effective, efficient, and affordable.
- No “one size fits all” big data architecture exists.
- Cloud services offer a variety of alternatives for big data solution implementation.
- Optimal results from big data initiatives often entail integration of legacy data sources.
- Sustained success from big data initiatives - enabled, in part, by appropriate architectural foundations - may reshape longstanding operational and policy norms.

2.4 Data Science and Scientists

The Data Science and Scientists session focused on understanding and reviewing the current market and needs from data scientists as well as to define what constitutes a data science team.

The session included discussions of the following:

- What tools and techniques are still needed for data science?
- What consists of a data science education? What courses, training, and experience do students and employees need to become successful data scientists?
- Identify recommendations to counter the workforce shortage among big data and analytics professionals.

2.4.1 Challenges

- Managers as well as senior leaders need to recognize that tools used over the past twenty years are not necessairly the best tools for current research. Accommodations must be made to update these tools as needed.
- A data science curriculum needs to include technical as well as non-technical courses.
- The relationship between universities and government agencies needs to be strengthened and provide should internship opportunities to students enrolled in data science programs.
- There is no classification in the General Schedule (GS) system of positions specifically for data scientists.
- Positions and roles requiring data scientists lack the proper skills and staffing.

2.4.2 Discussion Summary

The most common overall theme of the participants' discussion focused on educating data scientists. An increasing number of positions in the government require data science training and expertise. Students looking to fill these roles should pursue degree programs with technical as well as non-technical courses. Such a curriculum should include courses on: exploring and analyzing data, data storage and retrieval, programming and algorithms, statistics, machine learning, data visualization and communication, and databases (SQL and NoSQL). In addition, students will also benefit from a background and coursework in communications, facilitation, law, ethics, and policy.

Government organizations should be willing to invest in creating data science curricula for university degree programs and certifications. Government agencies and academic institutions should also work together to strengthen their relationships in order to offer internships to students enrolled in data science programs.

To overcome the intermediary shortage of data scientists in the workforce, agencies must work to identify the smaller skill sets that current staff can learn. Government organizations can provide training or otherwise support staff to learn these skills in online programs or in university certificate programs. In addition to specifically trained data scientists, agencies should hire interdisciplinary staff to compensate for the lack of data scientists. Individuals with a Ph.D. in statistics, economics, and computer science often have the necessary skills to work in data science.

Additional recommendations for the shortage of data scientists include developing pathway programs, creating temporary positions, creating internships specifically for data scientists, and use of the United States digital service to attract talents. Session participants recognized that government agencies need to be proactive in adopting data science technologies and developing or acquiring skills related to data science.

Another concern discussed in the session was the lack of leadership awareness that data scientists need up-to-date tools in order to access, harness, and analyze big data. Senior leadership should be familiar with data science topics, in order to forecast and plan for acquiring new tools. To reduce the attrition rate of skilled, talented data scientists in the government, agencies need to improve working conditions by providing top of the line hardware and focus on a high quality of life and work/life balance. Agencies should demonstrate how employees can make an impact.

Finally, the participants in the session discussed the need for government organizations to recognize the importance and uniqueness of data scientists in the workplace, especially those with the skills and knowledge to effectively manage big data. Contracting vehicles should be put in place to hire data scientists. The final recommendation of the summit participants is for the Office of Personnel Management (OPM) to reclassify GS positions with a data science category.

2.4.3 Important Findings

- Senior leadership must understand that a “data scientist” may not be simply an individual position, but an interdisciplinary combination or team of personnel.

- Federal agencies need to specifically address the shortage of data sciences by collaborating with universities to provide educational programs and training to students and staff.

- The OPM must be encouraged to reclassify GS positions to add a data science category.

3 Summit Recommendations

Several common themes recurred across all or many of the challenge areas. Participants noted four topics as having particular importance: flexible architecture and standards are necessary to operate in the fast-paced technical environment, data sharing standards need to be put in place to allow for easy collaboration, a solid and concrete understanding of the needs and requirements of big data are also important for all involved parties, and data science needs to be recognized as a blossoming field and supported by the federal government.

The need for flexible architecture was mentioned in several of the summit collaboration sessions. It is important for big data architectures to encompass all relevant, necessary aspects of the system. This must be performed in a consistent manner that is optimally effective, efficient, and affordable. Session participants agreed that there is no “one size fits all” solution for big data architectures. Even though there is no general solution, best practices, and similar problem spaces can leverage common solutions. Cloud services also offer a variety of alternatives for solution implementation and provide a wide range of flexible options.

Big data architecture must be responsive to business goals and mission objectives above all else. These systems need to be able to effectively and efficiently address the agencies’ top priorities. In order to accomplish these goals, architectures need to be flexible enough to handle rapid changes. The data industry is extremely fast paced and agencies need to be able to keep up and have policies and planned budgets in place for rapid change and piloting new systems.

Collaboration sessions also focused on the disappointing progress in regards to cross agency and even inter-agency data sharing. Data sharing is often a complex topic due to the lack of regulations or the lack of flexibility in regulations. Data openness and sharing directly conflicts with privacy and security. Compromises and proper security provisions can be put in place to ensure collaboration is possible in a secure manner. With the proper policies in place to protect privacy and security, agencies will be able to collaborate with greater ease. Data sharing should be recognized in the initial planning for projects and architecture. When the possibility of data sharing is considered early in the process it is easier to account and plan for the necessary budgeting and security measures that should be taken.

The third general topic considered across collaboration sessions at the summit directly relates to flexible architecture and data sharing. Agencies need to have a solid and concrete understanding of the needs and requirements of big data integration. This understanding needs to include the specific needs for governance, managing metadata, and shaping a valuable workforce.

In order for agencies to have data governance programs or policies, they first need to fully understand data governance. The summit collaboration sessions found that there was no concrete, agreed-upon definition of data governance. The first step to effectively implementing data governance is to clearly define the concept. Defining governance and developing a framework for governance will provide a common foundation and solid starting point for any future work.

Metadata is another important aspect to understanding big data. Metadata is an crucial artifact that gives a deeper value and history to the data in use by agencies. Standardizing metadata for data provenance is necessary to use and share the data in an effective and efficient manner. Without proper history and descriptions for the data, it is difficult for agencies to interpret the real value.

Understanding the needs and requirements of big data also means understanding the workforce needed to bring value to these efforts. Senior leadership at federal organizations must understand that a “data scientist” may be more than an individual position, and depending on the scope of the work, may entail and an interdisciplinary combination or team of personnel. Federal agencies need to specifically address the shortage of data sciences by collaborating with universities to provide educational programs and training to students and staff. Education programs across federal agencies or within agencies with a large need will help build a larger understanding of the problems data scientists are trying to solve. In order to realize change in this area, it is strongly recommended that the OPM reclassify GS positions to add a data science category.

4 Conclusions

The December 2015 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, designing flexible architectures, data sharing, effective governance, and building a workforce of data scientists remain difficulties to overcome. Developing policies for collaboration, recognizing the need for governance and for data scientists, and reshaping policies for flexible development can help to mitigate these identified challenges.

While the December 2015 Federal Big Data Summit highlighted areas of continued challenges and barriers to progress, the Summit also cited notable advances in mitigating these perennial challenges. There is now a more concrete understanding of big data as an overall concept. Now, questions move to definite other aspects such as governance and proper sharing protocols. 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. Sustained success from big data initiatives may reshape longstanding operational and policy norms.

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. Increased collaboration should take place to continue the current work towards improved adoption and knowledge of big data. ATARC has developed the Innovation Labs which work to facilitate these collaborations. The Big Data Innovation Lab presented for the first time during a panel session at the December 2015 Summit. Continuing work towards collaboration and understanding is crucial to move forward with the successful implementation and adaption of big data.

ACKNOWLEDGMENTS

The authors of this paper would like to thank ATARC and The MITRE Corporation for their support and organization of the summit. Special thanks are in order to Justin Brunelle and Patrick Benito for their contributions in reviewing the paper.

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