Federal Big Data Summit: Summary and Way Forward

Abstract

The Federal Big Data Summit took place on June 19th and 20th at the Ronald Reagan Center in Washington, DC. The Summit began with MITRE-Advanced Technology Academic Research Center (ATARC) Collaboration Sessions that allowed industry, academic, government, and MITRE representatives the opportunity to collaborate and discuss the government's challenge areas in big data. The goal of the collaboration sessions is to create a forum for an exchange of ideas and a way to create recommendations to further the adoption and advancement of big data within the Government. The goal of the collaboration sessions is to create a forum for an exchange of ideas and a way to create recommendations to further the adoption and advancement of big data within 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. MITRE worked in partnership with the 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 hire-ready graduates to advance the state of big data in the government.

Several recommendations were made as a result of the exchange of ideas in the collaboration sessions. This white paper summarizes these results, as well as identifies recommendations for government and academia while identifying orthogonal points between challenge areas. It also recommends an increase in cross-government and academic collaboration to share best practices and address cross-cutting challenges.

Collaboration Session Outcomes

Each MITRE Collaboration Session was a focused and moderated discussion between government and industry representatives about a big data challenge area.

The challenge areas for each session were as follows:

- Big Data in Healthcare
- Big Data Applications and Analytics.
- Big Data Solutions for Privacy Protection

Participants discussed current problems, gaps in work programs, potential solutions, and ways forward for each of the challenge areas. This section outlines the goals, outcomes, and summaries of each of the four collaboration sessions.

Big Data in Healthcare

The Big Data in Healthcare Symposium facilitated discussion on big data's impact on healthcare. Specifically, it focused on how big data pertains to Government run healthcare organizations such as the Center for Disease Control and Prevention (CDC) and National Institute of Health (NIH). The targeted topics for this sessions where data encryption, health care security, sharing information across domains, and using data clouds to speed health care analytics.

The goals of this session included:

- Identify big data encrypting challenges and practices
- Health Care security
    - Access Control (e.g., Personal Identifiable Information (PII))
- Identify best practices for sharing healthcare information across domains
- Identify using big data technologies to speed up health care analytics

The session discussions identified the following needs:

- Data Anonymization
    - Need to ensure privacy right protection and maintain data integrity
    - Need to ensure rules of anonymization of data are adequate when data is correlated
    - Need to ensure privacy and policy are on the same page as technology
- Open Government Initiative
    - Need to make data available to novel users
    - Need to promote research into using standards that promote reuse
    - Need to promote and actively engage in releasing medical data (releasing data has already shown to be useful in identifying quality control issues and biases)
    - Releasing medical data increases risk of incorrect conclusions by non-SMEs, how can this be mitigated?
- Public Generated Data
    - Need a standard definition for a healthy person
    - Need the ability to rapidly subsume publicly generated medical data
    - Data on healthy people is useful in research and analysis, so how can the medical community get more healthy people to share medical data?
- Public Leverage-Able Datasets
    - Need to increase the amount of publicly available data sets and data transparency (currently dwarfed in volume by non-public data)
    - Need to solidify a way to measure the ROI of releasing data
    - Need to help drive data to decision
    - Need to create technology and or process that help streamline the releasing of publically available data
- Standards
    - Need for businesses and academia to foster the creation of standards

    - Need for standards that fill any gaps left by Health Level Seven International (HL7)[1]
    - Need to encourage vendors to adopt HL7 for their products
    - Need for help with gaps in paper to digital transition
    - Need for general ingest that allows for consuming medical data that has different formats
    - Many physicians are still using paper for medical record; how can they be incentivize to digitize their patients data?
- Human Side of Data Science
    - Need a way to measure ROI in releasing data (seems to be high but cannot solidify that claim)
    - Need a way to prioritize the releasing of data (i.e., what data is high value?)
    - Need to identify metrics that will help with the releasing of data
    - Need to understand the demographic of the audience for the data (e.g., computer science, statistics, and domain expertise)
    - Need a way to share data, and collaborate without data sharing agreements
- Academic Intervention
    - Need for sharable patient identifiers that do not compromise data anonymization
    - Need for parameters of identification that are resistant to attacks (e.g., rainbow tables)
    - Need to help facilitate data analysis techniques and technologies to help drive data to decision
- Cloud Realities/Dependencies
    - Need for a more efficient way to move big data in the cloud
    - Need for SMEs in data cloud technologies
    - Need for third party comparison between different data cloud technologies
    - Need for diverse teams to tackle different aspects of big data challenges(e.g., digitizing physicians notes for early cancer detection require a set of disciplines)

Session Summary:

The Big Data in Healthcare Session focused on the challenges big data brings to securing healthcare data, sharing data, and effectively utilizing the data. Although the term big data seems fairly new, having a massive amount of data in healthcare is nothing new. What is relatively new are the technologies that facilitate the ability to efficiently anatomize vast amounts of data for analytics. As healthcare data is subsumed, anonymization becomes increasing vulnerable to misuse due to the heightened risk of not being able to foresee all correlation that can be made by querying the dataset. This is further exacerbated by the current pattern of privacy being a step behind policy and policy being a step behind technology. Despite these challenges, complying with the Open Government Initiative has not only been achievable but fruitful. There seems to be a high ROI associated with releasing healthcare data, however, gathering metrics to solidify its value is difficult. The government has a large number of datasets that pertain to healthcare and publicizing any given dataset can be relatively expensive. Thus, the lack of informative metrics hinders not only the rate of sharing data but also the ability to prioritize publicizing datasets release based on ROI.

Sharing and reuse of information from multiple healthcare datasets through standards will lower the cost and improving the efficiency, quality, and safety in healthcare. Standards such as HL7 provide a comprehensive framework for the exchange, integration, sharing, and retrieval of electronic health information, but implementation challenges remain. Not having standards at a lower level increases the

[1] http://www.hl7.org/

difficulty of sharing data. Another obstacle associated with sharing data is transmission time over a network. When dealing with terabytes or even petabytes of information shipping hard drives in the mail is still the fastest way. This makes big data clouds difficult to move once established.

The dialogue within the session indicated that it is often difficult for practitioners in business and academia to get involved due to legal, technical, and privacy issues. However, one area that could be used is encrypted sharable patient identifiers. This would allow a dataset to be released to the public while minimizing the risk of protected information being accessed.

Big Data Analytics and Applications

The Big Data Analytics and Applications Session aimed to facilitate discussions on applying big data in the Government and preparing for the continued growth in importance of big data. The targeted topics included disconnected environments, interoperability between data providers, parallel processing (e.g., MapReduce), and moving from data to decision in an optimal fashion.

The goals of this session included:

- Identifying techniques for big data processing and use
- Governance, sharing, and utilization of Big Data in the Government
- Opportunities for leveraging academia and the open-source community
- The use of cloud computing to include:
    - Stack construction
    - Data source and stack provider interoperability
    - Cost and acquisition of cloud resources
- Identifying areas in need of change at the policy level

The session discussions identified the following needs:

- Need to understand when data becomes “big”
    - At the agency level
    - At the task level
    - At the Government level
    - Using effective use cases
- Need to understand when cloud computing is necessary and appropriate
    - What cloud services or stacks match what use cases?
- Need to understand what talent (i.e., engineers) the Government needs to utilize big data
- How does the Government identify useful data?
    - Need metadata to make big data smaller
    - Need consistent access to the important data while hiding unimportant data
- Need the ability to select cloud computing tools despite rapid change
- Need to inform leadership and policy makers of big data challenges
    - Need a way to hire or train a specialized and increasingly scarce workforce
    - Need a more sufficient operational and effectively appropriated budget
    - Need to influence policy – mandates do not match use cases

The summary of the session is below:

The Big Data Analytics and Applications Session focused on the challenges the government is facing when applying big data to mission tasks. Big data terminology is increasingly overloaded and can have multiple meanings dependent upon context. All data is not big data, but big data must be recognized when encountered. For this reason, a classification and more formal definition of big data should be developed for the government. The government also needs use cases to demonstrate the impact, utility, and handling of big data in certain situations. Use cases will make determining methods of handling big data easier and more realistic. It is also easier to communicate to leadership and policy makers the impact of policy on the use cases and applications of big data. Industry has successfully leveraged big data for specific tasks (e.g., measuring the growth of the stock market in the financial sector) and the government should apply industry's successful tactics to the broader tasks of the government in appropriate use cases. A feedback loop should be put into place to allow big data practitioners a method to influence policy and decision making that impacts the use and implementation of big data principles. Finally, the session identified that the big data workforce is small and expensive. To mitigate the expense and train qualified big data workers, the government should leverage and guide academic research to help academics solve direct government challenges and increase the graduates that can be hired into government big data positions.

Big Data Solutions for Privacy Protection

The Big Data Solutions for Privacy Protection Session facilitated discussions on challenges big data brings to protected data. The targeted topics included access control, cloud security, cloud compliance, and data encryption.

The goals of this session included:

- Discuss current technologies that are applicable to this space
    - Accumulo
    - Sentry
- Discuss access control solutions such as:
    - Attribute Based Access Control (ABAC)
    - Identity Access Management (IdAM)
    - Role-Based Access Control (RBAC)
- Identify big data security practices
- Identify challenges with big data architecture compliance (e.g., Federal Information Security Management Act (FISMA))
- Identify big data encryption techniques and practices

The session discussions identified the following needs:

- Need to address increase complexities of data privacy attributed to big data
- Need for privacy risk mitigation when combining data sets
- Need for privacy analysis to be a part of cost/benefit analysis
- Need to better quantify privacy concerns for data collectors
- Need to address privacy risk for open data and publicly released data

The summary of the session is below:

The Big Data Solutions for Privacy Protections Session was a discussion focused on the impact big data has on privacy protection. Privacy and data protection is not solely a big data problem. However, having large amounts of data from a wide variety of sources adds more complexity to an already complex problem. Just as in the Big Data Analytics and Applications Session, the question "What is meant by 'big data?'" is not well defined – "Big Data" is a relatively new term even though having large amounts of data is not a recent development. Recent technological developments in handling large amount of data are opening new possibilities in gathering and anatomizing large amounts of data. These new possibilities are not only increasing the instances of large data stores but also their capabilities both of which present complex challenges when protecting data privacy. Aside from the technical issues there are additional complexities to governance and getting data providers "onboard" since huge data stores usually aggregate from multiple sources. This results in an increased risk of aggregated search results unintendedly divulging information. Virtually no complex system is perfect and therefor a cost/benefit analysis needs to be done to determine if sharing the respective data makes sense from a business prospective. The more secure the system the less the chance of an unintentional security leak but also the more that system will cost. Furthermore, there is the added issue with who is collecting the data. If a user of the system is collecting data from multiple sources themselves they can be aggregating information or sharing information with an unauthorized party.

Summary and Recommendations

Across the collaboration sessions there were commonalities that resonated. First was the lack of an accepted standard definition of "big data". During each session there was at least one conversation that defined a version of "big data". Second is the additional complexity associated with aggregating data from multiple data providers. Third is the need to understand the many technologies in the big data space and how they compare to each other. Fourth and finally, there is a need in Government for internal expertise on big data technologies and best practices.

Create a Standard Definition for Big Data

Big data is a commonly used term with no concrete definition. Defining big data has become a conundrum in computer science that often leads to cross talk and misunderstandings. During the sessions, the term big data was often given with context or dynamically defined. Academia may be able to help create a generally accepted definition for big data. Jonathan Stuart Ward and Adam Barker from the University of St Andrews in Scotland are taking steps towards this end[2]. Their work has resulted in defining big data as follows:

> "Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning."[2]

[2] http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/

Even if the definition from Ward and Barker needs work, the spirit of their effort is something that can be further pursued my others in academia.

Create Sharable Identifiers

The ability to relatively quickly aggregate data from similar or disparate sources and then sharing that data with third parties has become increasingly common, largely due to big data technologies. Since big data technologies are designed to handle large amounts of data, they are often designed to scale horizontally (i.e., across multiple distributed systems or nodes). The ability to scale horizontally makes housing large datasets more practical and cost effective vice scaling vertically (i.e., within a single system or node). Furthermore, big data technologies have a strong focus on facilitating the platform that allows for decisions to be made from large amounts of data. This combination is helping drive a data sharing doctrine. The push toward sharing data adds a layer of complexity to large datasets in the form of more data warehouses and more data aggregation. The issues caused by data aggregation such as security and governance may be difficult for academia to address since they may not be privy to essential information needed to contribute. However, if academia was to focus on a more general technology to help in this space, it may be very useful. For example, as pointed out in the Big Data in Healthcare Session, if academia was to develop a way to share data using encrypted sharable identifiers it could be a significant help to the domain of big data in health care. By encrypting personally identifiable information, but still allowing the data to be shared, the risk will be reduced when aggregating data or making data available to a third parties.

Create Third Party Technical Use Cases and Comparisons of Big Data Technologies

There is a push in Government for data modernization and moving towards a big data paradigm. Software in this space focuses on many different aspects of big data and there is no software that fits every problem set. Often a combination of different software technologies are needed for example a solution for querying data in HDFS may involve Impala[3], Spark[4], Mahout[5] and MAPR[6]. This issue is further complicated by the fact that many of the technologies provide a similar approach. For example, HBase[7], Accumulo[8], and Cassandra[9], are NoSQL databases that sit on top of HDFS. However, depending on the user's problem, one may prefer one technology over another. Understanding what available software best fits a specific problem is something that is very useful but is often lacking. Academia can help in this area by focusing research on utilizing big data technologies to solve complex problems and publishing practical comparisons on big data software.

Drive Curricula That are Geared Towards Big Data

The lack of big data SMEs in government is a concern. Moving to and effectively utilizing big data is extremely difficult with a lack of SMEs. No matter how beneficial a technology might be, if the right

[3] http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/impala.html
[4] https://spark.apache.org/
[5] https://mahout.apache.org/
[6] https://www.mapr.com/
[7] http://hbase.apache.org/
[8] https://accumulo.apache.org/
[9] http://cassandra.apache.org/

people are not there to utilize it properly, desired results may be very difficult to achieve. Driving curricula that have a strong focus on big data is an area where academia can help resolve this issue.

Big data is revolutionizing how large data sets are viewed and handled, and has an increasingly growing community. The dialogue in the sessions were fascinating because it helped give an insight to how government is handling the big data revolution. Each session not only gave an insight on what challenges government is experiencing but also how government would like to utilize big data.

Daniel Ruiz
The MITRE Corporation
daruiz@mitre.org

Tom Suder
ATARC
tsuder@mobilegovt.com