

# Data Driven Decision Making in Utah Government: Assessment for the Use of Big Data

*Data Driven Decision Making Project*

*Volume 1: Policy and Governance  
September 2015*



**EXECUTIVE SUMMARY**

|   |            |
|---|------------|
| <b>1 INTRODUCTION</b>   | <b>1</b>   |
| <b>2 POLICY AND LEGAL ISSUES FOR DATA SHARING BETWEEN AGENCIES</b>                              | <b>3</b>   |
| <b>2.1 Cross Agency Data Sharing</b>  | <b>5</b>   |
| 2.1.1 Encouraging Data Sharing Across Agencies  | 6          |
| <b>2.2 How Big Data is Different</b>  | <b>18</b>  |
| 2.2.1 Big Data Architecture and Associated Vulnerabilities                                      | 19         |
| 2.2.2 Risks associated with Big Data Technologies   | 20         |
| <b>2.3 Using Data Derived from Multiple Contexts</b>  | <b>21</b>  |
| <b>2.4 Privacy and Security</b>   | <b>22</b>  |
| 2.4.1 Privacy   | 23         |
| 2.4.2 Security  | 28         |
| 2.4.3 Security/Cloud Hosting Security Requirements  | 28         |
| 2.4.4 Security for Big Data in a Cloud Environment  | 30         |
| <b>3 LIABILITY/RISKS</b>  | <b>32</b>  |
| <b>4 EXAMPLES FROM OTHER STATES</b>   | <b>35</b>  |
| <b>4.1 Indiana Management and Performance Hub (MPH)</b>   | <b>35</b>  |
| <b>4.2 North Carolina Government Data Analytics Center (GDAC)</b>                               | <b>35</b>  |
| <b>4.3 Michigan Enterprise Data Warehouse (EDW) and Enterprise Information Management (EIM)</b> | <b>37</b>  |
| <b>5 MAKING DATA ACCESSIBLE/SHARED BY DEFAULT</b>   | <b>38</b>  |
| <b>APPENDIX A: LEGISLATION, EXECUTIVE ORDERS, EXECUTIVE DIRECTIVES FROM OTHER STATES</b>        |            |
| <b>A-1: Indiana Big Data Legislation</b>  | <b>A-1</b> |
| <b>A-2: North Carolina Big Data Legislation</b>   | <b>A-4</b> |

|   |             |
|---|-------------|
| <b>A-3: Michigan Big Data Legislation</b>                                     | <b>A-14</b> |
| <b>A-4: New Jersey Legislation</b>  | <b>A-17</b> |
| <b>APPENDIX B: FIPPS PRIVACY PRINCIPLES, RISKS, AND MITIGATION STRATEGIES</b> | <b>B-1</b>  |
| <b>APPENDIX C: BIG DATA GLOSSARY</b>  | <b>C-1</b>  |
| <b>APPENDIX D: ACRONYM LIST</b>   | <b>D-1</b>  |
| <b>APPENDIX E: 17 STEPS TO IMPLEMENT A PUBLIC SECTOR BIG DATA PROJECT</b>     | <b>E-1</b>  |
| <b>ANNOTATED BIBLIOGRAPHY</b>   |             |

## Executive Summary

Data Driven Decision Making involves integrating disparate data sources to form a common pool of data, applying combination of statistical and optimization techniques to uncover hidden insights, and use it to take informed decisions. Government organizations can utilize predictive analytics to become more efficient and effective in their delivery of services by creating a holistic view of individual citizens, thereby ensuring government programs and services address the overall needs of its citizens. DDD can also help improve accountability and transparency which are also being demanded by citizens of governments across the world. Through the effective use of DDD technologies and techniques, the public sector will be able to make decisions that are based on facts rather than assumptions, politics, and myths.<sup>1</sup>

Big data has been widely adopted in the Private Sector to assist business in understanding their customers and determining the most effective means of targeting them or making the customer interface more effective. Big Data has made less headway in the Public Sector, but, as identified in a McKinsey global Institute Study “Big data: The next frontier for innovation, competition, and productivity” Big Data is now relevant for leaders across every sector, and consumers of products and services stand to benefit from its application. The ability to store, aggregate, and combine data and then use the results to perform deep analyses has become ever more accessible. The study notes that sectors such as government faces higher hurdles because of a lack of data-driven mind-set.<sup>2</sup> As shown in Figure 1, capturing value in the government sector is in the bottom quintile. While there is a lot of data it is not necessarily available or accessible due to issues with sharing, formatting, and restrictions due to policy and or legal issues. Also, government agencies need work to enhance the data-driven mindset.

---

<sup>1</sup>[https://www.thegovernmentsummit.org/EventFolder/KnowledgeHubFolder/KnowledgeHub\\_14/achieving\\_excellence\\_via\\_data-driven\\_decision\\_making\\_in\\_government\\_eng.pdf](https://www.thegovernmentsummit.org/EventFolder/KnowledgeHubFolder/KnowledgeHub_14/achieving_excellence_via_data-driven_decision_making_in_government_eng.pdf)

<sup>2</sup>[http://www.mckinsey.com/insights/business\\_technology/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

Relative Ease of Capturing Value Potential

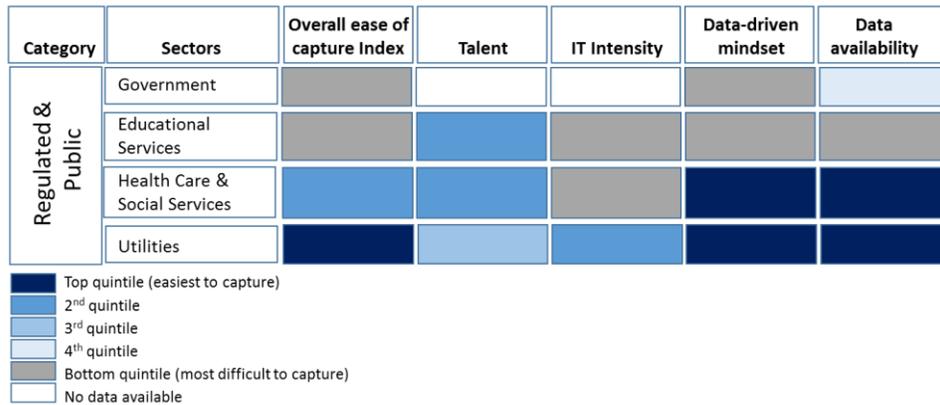


Figure 1. Relative Ease of Capturing Value Potential from Big Data in Government<sup>3</sup>

However, as illustrated in Figure 2, government offers the highest potential for capturing value from the use of Big Data.

Value Potential of Using Big Data

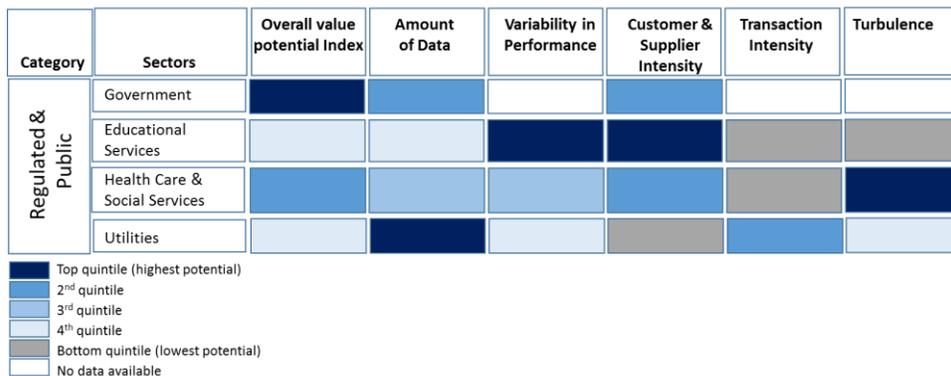


Figure 2. Value Potential from Using Big Data in Government<sup>4</sup>

Policy makers must consider the choices they should make in order to help individual agencies capture value out of using big data. The major areas where policy can play a role are:

1. Building human capital (Improving the supply of graduates knowledgeable about big data)
2. Aligning incentives to ensure access to data

<sup>3</sup> [http://www.mckinsey.com/insights/business\\_technology/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

<sup>4</sup> [http://www.mckinsey.com/insights/business\\_technology/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

3. Addressing privacy and security concerns
4. Establishing intellectual property frameworks
5. Overcoming technological barriers to data
6. Promoting information and communication technology infrastructure.<sup>5</sup>

To realize benefits of data driven decision making, it is necessary to integrate and share data across the Enterprise. The state of Utah has established an extremely effective Open Data program making data that can be shared open and shared by default. These same principles can be extended to data sharing to establish the basis for making Utah truly a truly data driven Enterprise. While there is a solid basis of data sharing, integration and business intelligence in the state of Utah this process needs to be extended across the Enterprise to be truly effective and deliver results.

This paper presents the following recommendations to facilitate the necessary data sharing, governance and security:

- **Executive Sponsorship:** Identify an Executive Sponsor as a champion for the Big Data and analytics program effort. Use legislation to establish the program, provide structure, funding, appropriate authorities, facilities, tools, and direct agency participation.
- **Representational Oversight:** Establish a big data oversight board with executive participation from multiple agencies to coordinate decisions, set priorities, coordinate involvement, resolve issues of data security and access, determine data ownership, and approve data sharing issues.
- **Legal Counsel:** Engage with the Attorney General's office in establishment of legal counsel to assist agencies in interpreting restrictions regarding data sharing to enable appropriate and legal sharing while protecting privacy, civil rights and civil liberties
- **Chief Data Officer:** Establish a Chief Data Officer position to establish and enforce a data strategy for the state of Utah.
- **Data Ownership:** Assign data owners. Implement a data ownership policy identifying the role and associated responsibilities, processes, procedures. Also assign information or process owners to ensure appropriate control over products of analytics.
- **Data Stewards:** Identify Data Stewards in each agency or program providing data sets
- **Enterprise Data Lake:** Utilize a Data Lake within DTS to provide the Enterprise the greatest agility in leveraging the collected and combined data as a true

---

<sup>5</sup> [http://www.mckinsey.com/insights/business\\_technology/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

Enterprise asset. The establishing legislation should also establish direction of the Data Lake and provide appropriate authorities as well as responsibility for management and security of the Data and Data Lake infrastructure.

- **Interface Strategy:** Establish an interface strategy to replicate data sets into the Data Lake, provide policies and procedures to manage interfaces, such as guidance for interface strategy and design documentation, interface reconciliation controls, and audit logs.
- **Standardized Partnership Agreements:** Develop a standard Partnership Agreement template and streamlined process to facilitate data ingest. Oversee execution of Partnership Agreements with each agency or Program providing data sets.
- **Standardized Data Sharing Agreements:** Establish a standard template and process for Data Sharing/Data Use Agreements between Agencies or Programs to detail utilization of data, restrictions, and other expectations or requirements for analytics projects. A possible model for data sharing agreements exists in the Utah Digital Spatial Data Sharing and Integration Project where use of a statewide agreement eliminates the necessity of developing multiple agreements between the individual participating agencies for the purpose of sharing data. This approach decreases the duplication of effort, promotes the exchange of information, and fosters communication between agencies in Utah.
- **Access Controls:** Establish and implement effective user access controls to manage and control access to data, detect/prevent inappropriate access to and modification of data (insert, update, and delete production data in the environment) within the Data Lake. These controls should enable authenticated users with access to ‘see’ and access data that they are authorized to use in accordance with rules established by the data owners and documented in data sharing agreements.
- **Enterprise Big Data Scheme:** Establish an Enterprise Big Data Scheme at the outset to address security and privacy issues specific to Big Data such as heterogeneous components, protection for data at rest and in motion (to include streaming data and sensor streams).
- **Update Privacy and Consent Statements:** Review, revise, and/or expand upon consent and privacy statements at point of collection to inform citizens of expended data use. Ensure that mechanisms are in place for redress and corrections.
- **Configuration and Change Management:** Establish a robust configuration and change management plan. Provide for oversight and tight control over IT resources and spending. Ensure that even basic maintenance of operations and

allocating additional resources are incorporated in the decision-making process. The security team must be aware of any changes being performed as part of the system lifecycle with big data platforms capable of utilizing cloud services.

- **Risk Mitigation Plan:** Develop a Risk Mitigation Plan. Identify potential risk areas and specific risks. Identify mitigation strategies and specific activities to minimize risks or their impact if realized. Establish a Risk Management Board and processes to regularly meet and review possible emergent risks and address appropriate mitigation actions.
- **Data Management and Governance Plan:** Implement a Data Management Plan and Data Governance to provide establish, implement, and oversee processes to ensure secure and appropriate access to and use of data.
- **Communication Plan:** Implement a Communication Plan to provide transparency and inform and engage stakeholders on plans, progress, and problems. The Communication Plan should include a Crisis Communications Plan for addressing data breaches or other risks that are realized.

## 1 Introduction

This document is one in a series of five documents, describing an assessment of the policy and legal issues related to data sharing between agencies, using information derived from multiple data contexts, and the feasibility of making data open by default. This Task 1 Policy and Governance paper describes lessons learned, best practices, and suggestions for policy to enable data sharing and fusion to implement Big Data and analytics and make data open by default. The companion documents in this study are Task 2 People Skills and Collaborations, Task 3 Technology Roadmap, Task 4 Business Case, and Task 5 Data Science and Value.

Utah state government deals with many of society's biggest problems. As the state's population continues to grow, these challenges become more complex and the public expects government to respond to these issues as effectively as possible<sup>6</sup>. Data driven decision making based on big data and analytics offers better results through statistical analysis. On Ted Talks Anne Milgram discussed how she brought the 'Moneyball' concept applying statistical analysis or big data analytics to criminal justice in New Jersey to move away from 'gut instinct' to informed understanding of who was being arrested and charged<sup>7</sup>. In Camden, New Jersey data driven decision making helped lower the murder rate by 41%. The 'Moneyball' approach or data driven decision making has been demonstrated as extremely effective in multiple public sector contexts.

Data driven decision making in government requires access to vast amounts of data from multiple agencies. Big Data infrastructure, analytics tools, and data scientists are needed to facilitate pulling actionable information from that data. Several Utah state agencies (Department of Workforce Services (DWS) – eFind sharing employment, labor, and economic data), Department of Human Services (DHS), Department of Health (DOH), Department of Corrections (DOC), Utah Department of Transportation (UDOT) (transportation data), Utah State Tax Commission (revenue data) have individually turned to data and analytics as a way to reduce cost and improve service delivery. State of Utah Enterprise Big Data analytics facilitated by the Department of Technology Services (DTS) is the next evolution to accomplish these objectives by creating an Enterprise data

---

<sup>6</sup> Data Driven Decision-Making in Utah, Business Case p 1

<sup>7</sup>

[http://www.ted.com/talks/anne\\_milgram\\_why\\_smart\\_statistics\\_are\\_the\\_key\\_to\\_fighting\\_crime/transcript?language=en](http://www.ted.com/talks/anne_milgram_why_smart_statistics_are_the_key_to_fighting_crime/transcript?language=en)

lake supporting unprecedented data sharing across agencies to enable analytics and ultimately data driven decision making in Utah.

Figure 3 illustrates the intersection of Big Data, Open Government and Open Data.

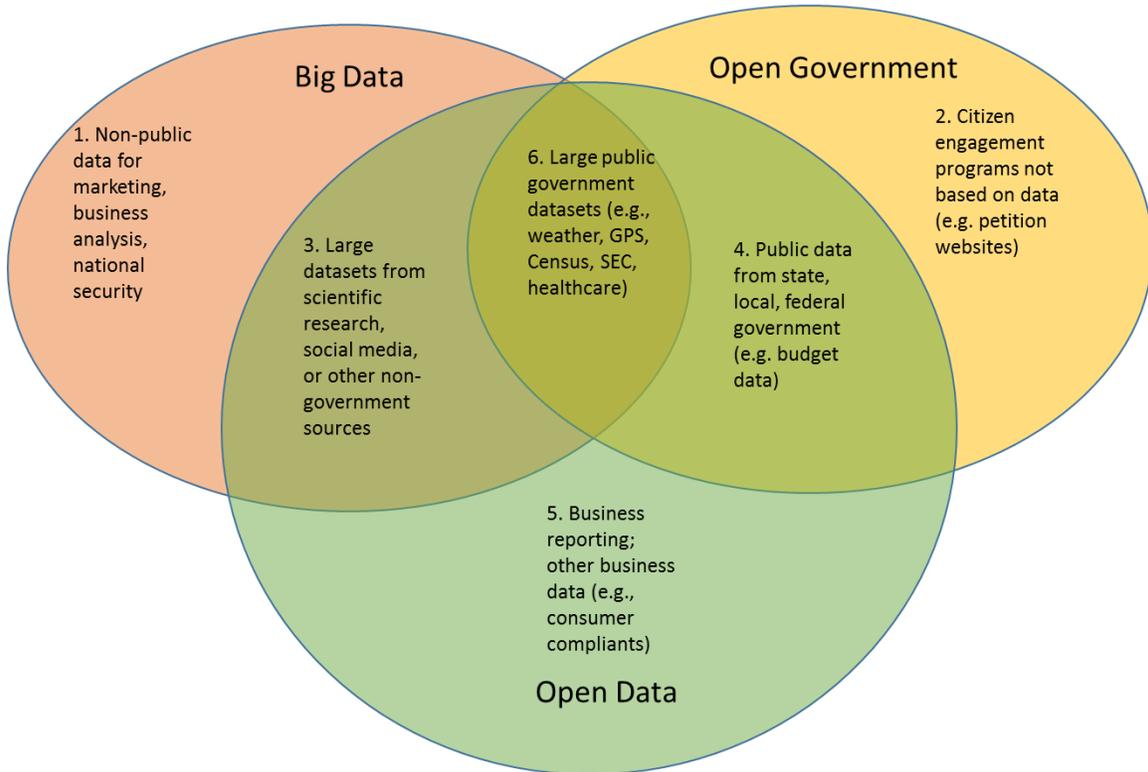


Figure 3. Intersection of Big Data, Open Government and Open Data

To derive benefits from data collected across the Enterprise, it will be necessary for many or all state agencies to share data. To successfully use big data analytics it will be necessary to combine multiple agency data sets. Also, we explore the feasibility of making data accessible and shared by default<sup>8</sup> subject to legal limitations for sharing personal and private information, across agencies. The state of Utah Big Data infrastructure may be hosted on premise in a Private Cloud, in a Public Cloud environment, or may be a hybrid solution.

<sup>8</sup> The concept of moving from a closed provisioning of finished information in response to requests to making the data itself accessible and using smart data and smart authentication to provide to each user access only that data they have the authority to see and use.

This move toward Big Data and analytics along with data accessibility will necessitate implementation of new data management governance and management processes as well as policy changes to address implications of cross agency data sharing, data fusion, and privacy and security issues. Possible Cloud hosting for the state of Utah raises additional policy issues.

This paper addresses the following Task Objectives identified in the Statement of Work (SOW):

- Objective 1: Policy and legal issues for Data sharing between agencies
- Objective 2: Policy and legal issues for Using information derived from multiple data contexts
- Objective 6: Feasibility of making data accessible and sharable by default.

For purposes of this assessment, Data Sharing can be viewed as:

- Data Ingest: Each agency must agree to ‘share’ or allow its data to be replicated into the Enterprise Data Lake managed by Partnership Agreements executed between DTS and each participating agency.
- Data Analytics/Using information derived from multiple sources: Participating agencies establish permissions and restrictions for access to and use of their data through data sharing agreements.
- As with Open Data and Open Government, the state should assess what data can be easily shared and made accessible across the Enterprise by default as an Enterprise Asset.

## 2 Policy and Legal Issues for Data Sharing Between Agencies

The McKinsey Group points out that access to data will need to broaden to capture the full potential for value creation. They find that often incentives are misaligned so that stakeholders want to keep the information to themselves. However, in order to fully capture the value that can be enabled by big data, the barriers to accessing data will have to be overcome<sup>9</sup>. Beyond establishing the capability, obtaining and cultivating the appropriate resources, this is an area where executive leadership and legislation can help. Data sharing between agencies has long been an issue of significant public sector attention due to both technical and cultural barriers. New programs and technologies have demonstrated the value of sharing data and lowered the technological barriers significantly. Data sharing across agencies in the Big Data context does not necessarily introduce new challenges, but does introduce new complexities. These deal primarily

---

<sup>9</sup> [http://www.mckinsey.com/insights/business\\_technology/big\\_data\\_the\\_next\\_frontier\\_for\\_innovation](http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation)

with the volume and level of data sharing, the sensitivity of individual data sets and increased value and sensitivity of combined data sets.

A literature review identifies a trend toward legislating agency participation and data sharing as well as committing to making appropriate government data open by default. Both Public and Private Sector entities are opting to establish Chief Data Officers (CDO) and Data Stewards in conjunction with Data Management and Data Governance Policy. This approach recognizes the value of the Enterprise's data as an asset and provides for appropriate oversight, management, and protection of the data itself. Several states (North Carolina, Michigan) have retained legal counsel to work with individual agencies specifically to enable data sharing through review and interpretation of specific policies that might be interpreted to prevent sharing of specific data sets.

Within the state of Utah the DWS, DHS, DOH, DOC, UDOT, and Utah State Tax Commission have already initiated analytics programs and successfully share data across federal, state, local and other boundaries to streamline services to the citizens. The state of Utah will be able to build upon the lessons learned and best practices from these programs. An excellent model for data sharing agreements exists in the Utah Digital Spatial Data Sharing and Integration Project<sup>10</sup> where use of a statewide agreement eliminates the necessity of developing multiple agreements between the individual participating agencies for the purpose of sharing data. This approach decreases the duplication of effort, promotes the exchange of information, and fosters communication between agencies in Utah.

The states of North Carolina, Michigan, and Indiana are all in various stages of Big Data and analytics programs and offer valuable insights and lessons learned.

Figure 4 illustrates the how data governance applies to facilitate data sharing for Big Data from: Discovery of where sensitive data resides, how it can be used and who may access it; Definition and Classification of sensitive data across the Enterprise along with requirements specific to its protection; Application of privacy and security policies; and, Measurement and Monitoring of compliance.

---

<sup>10</sup> <https://www.fgdc.gov/grants/2009CAP/InterimFinalReports/088-09-5-UT-AppendixD-DataSharingMOU.pdf>

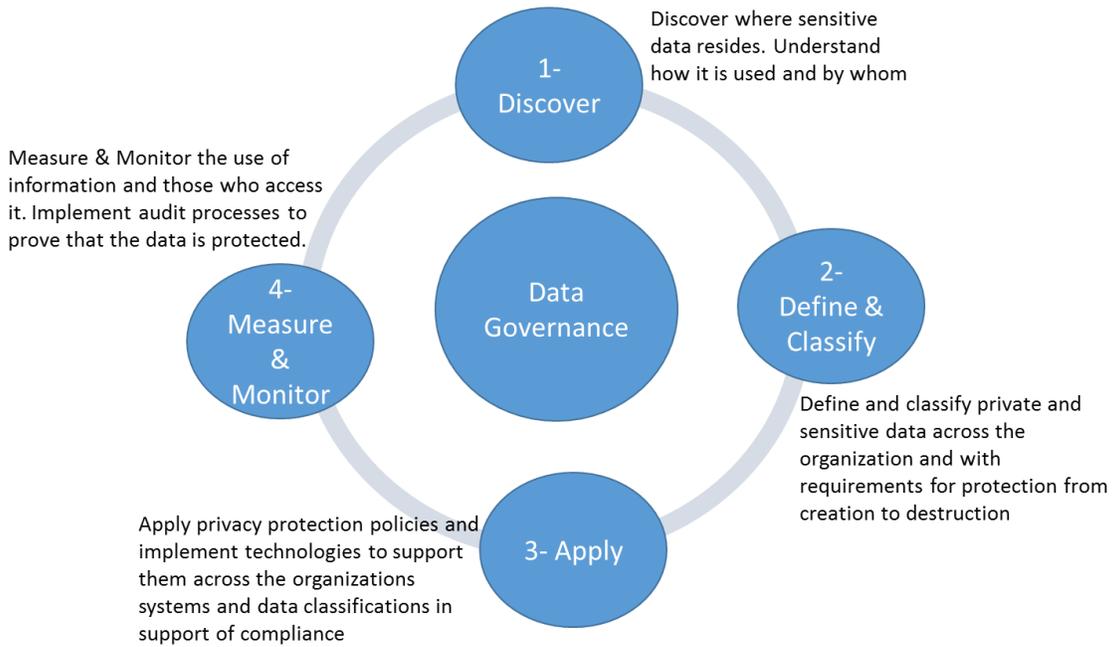


Figure 4. Enterprise Data Governance for Big Data

**2.1 Cross Agency Data Sharing**

Studies over the past five years (including two conducted by the Government Accountability Office (GAO)), reviews and lessons learned by federal, state, and local entities undertaking big data efforts indicate challenges or impediments to data sharing across agencies. Issues fall into two broad categories: those related to federal privacy; and those related to information security requirements and organizational and implementation issues. As is evident in **Table 1** below, privacy and security issues are the underlying drivers impacting or restricting agency’s willingness or ability to share data.

*Table 1. Challenges Identified by State and Local Governments to Sharing Data*

| Category of Challenges                    | Specific Challenges to Data Sharing   |
|---|---|
| Federal Privacy and Security Requirements | Federal privacy requirements that govern data sharing are inconsistent across multiple programs         |
|   | Agencies may be overly cautious and interpret federal privacy requirements more narrowly than necessary |
|   | Confusion or misperceptions around what agencies are or are not allowed to share                        |
|   | Agencies are not always sure when client consent is required to share data                              |
|   | Agencies are hesitant to use clients’ Social Security numbers to match data across systems              |

|  |  |
|--|--|
|  | Federal privacy requirements about sharing data with third parties (e.g., non-profit service providers) are overly restrictive |
|  | Agencies may not always be aware of the capacity of technology to protect personal information                                 |
|  | Security standards for sharing and storing data are inconsistent   |
| <b>Organizational and Implementation</b> | Agencies are concerned about the accuracy of data from other agencies  |
|  | Agencies may not trust that other agencies will sufficiently protect shared data   |
|  | Data sharing agreements between agencies are cumbersome to establish   |
|  | Agencies tend to adopt data sharing agreements that are too specific and do not allow for flexibility                          |
|  | Past practice has created a mindset or culture that agencies should not share data   |
|  | Public perception regarding sharing personal information deters agencies from sharing data                                     |
|  | Confusion or misperceptions around what agencies are or are not allowed to share   |
|  | Agencies do not provide sufficient training to workers on allowable sharing  |

**2.1.1 Encouraging Data Sharing Across Agencies**

This section provides discussion on organizational and implementation issues as well as culture and governance to encourage, support, and facilitate cross agency data sharing for Big Data and analytics. It also provides recommendations based on lessons learned and best practices from other Public Sector Big Data and analytics programs.

While dealing with privacy and security issues is a concern in sharing data, as has been demonstrated in North Carolina, Michigan, and Indiana, it can be overcome through various means. Building a culture of data sharing and ensuring the policies, procedures, processes, and technology are in place to facilitate that sharing needs to be addressed.

**2.1.1.1 Executive Sponsorship and Establishing Policy**

The state of Utah Big Data and analytics program should have an Executive sponsor. Based on experiences in other states (See Appendix A), it is recommended that Utah initiate the program through legislation which provides the necessary authorities for program initiation, management, and control. It also facilitates or directs agency participation and data sharing.<sup>11</sup> This provides clear direction for establishing the

---

<sup>11</sup> Executive Directive No. 2013-1, Data and Information Sharing, Management and Governance; Governor Rick Snyder, November 1, 2013  
 State of Indiana Executive Order 14-06; Establishing the Governor’s Management and Performance Hub  
 General Assembly of North Carolina Session 2013 Session Law 2013-360 Senate Bill 402  
 S582 ScaAa (2R) Designates New Jersey Big Data Alliance as State's advanced cyber infrastructure consortium

program and mandate for data sharing as well as providing the necessary and appropriate authorities to ensure that the system and data are appropriately administered and protected.

- North Carolina Government Data Analytics Center (GDAC) was established via legislation.
- The Indiana Management and Performance Hub (MPH) were established via Executive Order.
- Michigan established their Enterprise Information Management (EIM) program via Executive Directive.

Figure 5 illustrates the key components of a successful Big Data strategy that include: Establishing processes, best practices and techniques; Governance; Identification of the Authoritative Data Source; Data Stewardship; Data and Information or Process Ownership; and Data Security.

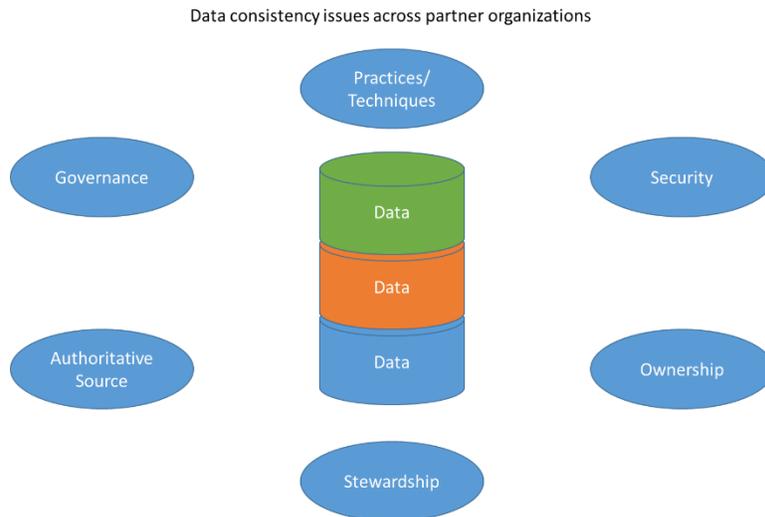


Figure 5. Key Components of a Successful Big Data Strategy

**Recommendation:** Identify an Executive Sponsor as a champion for the Big Data and analytics program effort. Use legislation to establish the program, provide structure, funding, appropriate authorities, facilities, tools, and direct agency participation.

**2.1.1.2 Program Governance and Oversight Board**

Defining and establishing a clear representative program governance structure has been demonstrated as being very effective in facilitating agency participation and enabling data sharing across agencies. As identified in the paper ‘Data Driven Decision-Making in Utah’ governance will be critical to the success of the state of Utah Big Data and

analytics program: “To address issues that will arise with the more effective use of data, the State needs a way to coordinate involvement, resolve issues of data security and access, determine data ownership, and approve data sharing issues. Starting with the Planning Phase, there are several best practices and lessons learned from both private and public sector Big Data efforts where new policies are warranted or existing policies should be expanded.

Beyond facilitating sharing, governance provides crucial oversight functions ensuring compliance and appropriate attention to agency and/or data set specific policy, privacy, and security concerns. Findings from both federal, state, and local experiences (including the US Department of Homeland Security (DHS), North Carolina, Michigan, Indiana, the Office for Science and Technology Policy (OSTP), Center for Data Innovation, etc.) have indicated that while government’s primary concern regarding Big Data is related to safeguarding data, privacy, civil rights and civil liberties, there is sufficient evidence to indicate that policies are needed to encourage and govern data sharing among agencies to enable and realize the public benefit of Big Data.

**Recommendation:** Establish a big data oversight board with executive participation from multiple agencies to coordinate decisions, set priorities, coordinate involvement, resolve issues of data security and access, determine data ownership, and approve data sharing issues.

#### ***2.1.1.3 Facilitating Data Sharing Through Legal Counsel***

Beyond establishing a representative governance structure, North Carolina and Indiana have both recognized the need for having legal counsel to review cases for data sharing. This alleviates cases where agencies’ interpretation of their responsibilities and restrictions in protecting the data are more restrictive than necessary. While information sharing and shared data have become more commonplace, in examining data sharing in Big Data programs in Michigan, Utah, Alleghany County and New York stakeholders report some reluctance to share. Stakeholders in these jurisdictions told GAO analysts that often agencies do not believe that they can legally share data. This is most commonly a factor related to requests to share data sets that include health data or data regarding children. Stakeholders told GAO that they felt it was beneficial to specify what data can and cannot legally be shared; how the specified data can be shared to mitigate risks and restrictions; who can and cannot legally have access to the data based on their role or function; and, for what purposes the data can and cannot be legally utilized. North Carolina has engaged legal consul to assist or facilitate interpretation of specific agency regulations and policies that could inhibit data sharing.

**Recommendation:** Recommend that DTS engage with the Attorney General’s office in establishment of legal counsel to assist agencies in interpreting restrictions regarding data

sharing to enable appropriate and legal sharing while protecting privacy, civil rights and civil liberties.

#### **2.1.1.4 Chief Data Officer**

In addition to a big data oversight board and or governance structure, several public sector Enterprises pursuing Big Data Analytics programs have created a Chief Data Officer (CDO) position. As the Enterprise data is leveraged and thus assumes greater significance and value, many organizations have found it beneficial to provide data specific oversight following the model of Chief Information Officers who have responsibility for information systems. The CDO is responsible for enterprise-wide governance and utilization of information as an asset. The CDO has responsibility for determining what kinds of information the enterprise will choose to capture, retain and exploit and for what purposes. This role includes defining strategic priorities for the enterprise in the area of data systems and opportunities, identifying new business opportunities pertaining to data, and generally representing data as a strategic asset at the executive table. Michigan has appointed a CDO.<sup>12</sup>

Best practices identified include the following:

- Establish Goals and Objectives for Big Data and Open Data
- Establish Strategy for Enterprise Data, Big Data and Open Data Strategy
- Executive Directive or Order establishing appropriate authorities, program leadership, governance structure, budget, participation, and resources
- Identify, prioritize, plan, execute, and control big data analytics projects

**Recommendation:** Establish a Chief Data Officer position to establish and enforce a data strategy for the state of Utah.

#### **2.1.1.5 Data Owners and Process Owners**

It is necessary to identify data owners (owners of the raw data) as well as for the outputs of Big Data processes. Data ownership will be distinct from information ownership. Data ownership and stewardship is typically assigned to the agency that collected (or originated) the data. Data owners are the subject matter experts that identify specific mission area, legal, or other restrictions regarding access and use of the data. Data owners

---

<sup>12</sup> Larissa Moss from Sid Adelman & Associates lays out a very comprehensive approach to the Chief Data Officer requirements, roles, responsibilities, qualifications and duties in <http://www.cutter.com/content-and-analysis/resource-centers/business-intelligence/sample-our-research/biar1302.html>

retain responsibility for providing the Fair Information Practice Principles (FIPPS) mechanisms that enable the principles of:

- Participation through: Periodic refresh the data ingested into the Data Repository; Providing citizens with the same access and redress opportunities in the Data Repository they would have in the original Information Technology (IT) system; Providing data governance; Providing a process and mechanism to verify the data accuracy with the original IT system in cases where action or decisions impacting individuals will result from use of the Personally Identifiable Information (PII) Use Limitation through: Specifying requirements to restrict access to PII within a particular data set based on the user's specified purpose<sup>13</sup>

Data owners have formal accountability for business responsibilities ensuring effective control and use of data assets. The data owner has the decision-making authority to resolve issues that crop up. Further, the data owners are responsible for inventory and configuration management control of their data. The data owner will be responsible for data sharing agreements specifying data access, data use, data restrictions, etc. for their data.

**Recommendation:** Assign data owners. Implement a data ownership policy identifying the role and associated responsibilities, processes, procedures. Also assign information or process owners to ensure appropriate control over products of analytics.

#### **2.1.1.6 Agency Data Stewards**

Agency Data Stewards have been identified as playing a valuable role in ensuring data integrity. A data steward is responsible for the management of data elements – both the content and metadata. They incorporate processes, policies, guidelines and responsibilities for administering an agency's data in compliance with policy and/or regulatory obligations. The data steward will provide for periodic refresh of data and facilitate compliance with agency specific guidelines, policies, laws, and restrictions as they impact access to and use of agency shared data sets.

**Recommendation:** Identify Data Stewards in each agency or program providing data sets.

#### **2.1.1.7 Enterprise Data Lake vs. Enterprise Data Warehouse**

Table 2 provides an overview of the benefits of a Data Lake and a comparison of Data

---

<sup>13</sup> Privacy Impact Assessment Update DHS Data Framework, DHS/ALL/PIA-046(a), August 29, 2014 pages 10-16

Warehouses to Data Lakes.

Table 2. Data Lake Advantages and Comparison with Data Warehouse<sup>14</sup>

| Data Warehouse  | Data Lake  |
|---|--|
| <b>1. Data Lakes Retain All Data</b>  |  |
| <ul style="list-style-type: none"> <li>• During the development of a data warehouse, a considerable amount of time is spent analyzing data sources, understanding business processes and profiling data. The result is a highly structured data model designed for reporting.</li> <li>• A large part of this process includes making decisions about what data to include and to not include in the warehouse.</li> <li>• Generally, if data isn't used to answer specific questions or in a defined report, it may be excluded from the warehouse. This is usually done to simplify the data model and also to conserve space on expensive disk storage that is used to make the data warehouse performant</li> </ul> | <ul style="list-style-type: none"> <li>• The data lake retains ALL data. Not just data that is in use today but data that may be used and even data that may never be used just because it MIGHT be used someday. Data is also kept for all time so that we can go back in time to any point to do analysis.</li> <li>• This approach becomes possible because the hardware for a data lake usually differs greatly from that used for a data warehouse.</li> <li>• Commodity, off-the-shelf servers combined with cheap storage makes scaling a data lake to terabytes and petabytes fairly economical</li> </ul> |
| <b>2. Data Lakes Support All Data Types</b>   |  |
| <p>Data warehouses generally consist of data extracted from transactional systems and consist of quantitative metrics and the attributes that describe them. Non-traditional data sources such as web server logs, sensor data, social network activity, text and images are largely ignored. New uses for these data types continue to be found but consuming and storing them can be expensive and difficult</p>  | <p>The data lake approach embraces these non-traditional data types. In the data lake, we keep all data regardless of source and structure. We keep it in its raw form and we only transform it when we're ready to use it. This approach is known as "Schema on Read" vs. the "Schema on Write" approach used in the data warehouse</p>   |
| <b>3. Data Lakes Support All Users</b>  |  |
| <p>In most organizations, 80% or more of users are "operational". They want to get their reports, see their key performance metrics or slice the same set of data in a spreadsheet every day. The data warehouse is usually ideal for these users because it is well structured, easy to use and understand and it is purpose-built to answer their questions.</p>  | <p>The data lake approach supports all of these users equally well. The data scientists can go to the lake and work with the very large and varied data sets they need while other users make use of more structured views of the data provided for their use.</p>   |

<sup>14</sup> Derived from a post by Chris Campbell <http://www.blue-granite.com/blog/bid/402596/Top-Five-Differences-between-Data-Lakes-and-Data-Warehouses>

|  |   |
|--|---|
| <p>The next 10% or so, do more analysis on the data. They use the data warehouse as a source but often go back to source systems to get data that is not included in the warehouse and sometimes bring in data from outside the organization. Their favorite tool is the spreadsheet and they create new reports that are often distributed throughout the organization. The data warehouse is their go-to source for data but they often go beyond its bounds</p> <p>Finally, the last few percent of users do deep analysis. They may create totally new data sources based on research. They mash up many different types of data and come up with entirely new questions to be answered. These users may use the data warehouse but often ignore it as they are usually charged with going beyond its capabilities. These users include the <a href="#">Data Scientists</a> and they may use advanced analytic tools and capabilities like statistical analysis and predictive modeling.</p> |   |
| <p><b>4. Data Lakes Adapt Easily to Changes</b></p>  |   |
| <p>One of the chief complaints about data warehouses is how long it takes to change them. Considerable time is spent up front during development getting the warehouse’s structure right. A good warehouse design can adapt to change but because of the complexity of the data loading process and the work done to make analysis and reporting easy, these changes will necessarily consume some developer resources and take some time.</p> <p>Many business questions can’t wait for the data warehouse team to adapt their system to answer them. The ever increasing need for faster answers is what has given rise to the concept of self-service business intelligence.</p>  | <p>In the data lake on the other hand, since all data is stored in its raw form and is always accessible to someone who needs to use it, users are empowered to go beyond the structure of the warehouse to explore data in novel ways and answer their questions at their pace.</p> <p>If the result of an exploration is shown to be useful and there is a desire to repeat it, then a more formal schema can be applied to it and automation and reusability can be developed to help extend the results to a broader audience. If it is determined that the result is not useful, it can be discarded and no changes to the data structures have been made and no development resources have been consumed.</p> |

Because data lakes contain all data and data types, because it enables users to access data before it has been transformed, cleansed and structured it enables users to get to their results faster than the traditional data warehouse approach. However, this early access to the data comes at a price. The work typically done by the data warehouse development team may not be done for some or all of the data sources required to do an analysis. This leaves users in the driver’s seat to explore and use the data as they see fit but the first tier

of business users may not want to do that work. In the data lake, these operational report consumers will make use of more structured views of the data in the data lake that resemble what they have always had before in the data warehouse. The difference is that these views exist primarily as metadata that sits over the data in the lake rather than physically rigid tables that require a developer to change.<sup>15</sup>

While Big Data technologies can provide for access to data where it resides within various agency systems, the move to a Data Lake appears to offer the most effective approach to facilitating cross agency sharing based on centralized common standards and governance structures.

Michigan, Indiana, and North Carolina programs all utilize Enterprise Data Warehouses (EDW). The IT Agency managing the big data environment must have direction and authority over governance. As demonstrated in Michigan's experience, establishing and managing the enterprise data warehouse without legislation or Executive Directive to establish requirements and appropriate authorities allowed poor governance and security oversight. The 2013 Executive Directive established appropriate authorities for the Department of Technology, Management and Budget (DTMB) removing previous ambiguities in its responsibilities for securing and protecting data.

**Recommendation:** Utilize a Data Lake within DTS to provide the Enterprise the greatest agility in leveraging the collected and combined data as a true Enterprise asset. The establishing legislation should also establish direction of the Data Lake and provide appropriate authorities as well as responsibility for management and security of the Data and Data Lake infrastructure.

#### **2.1.1.8 Interface Management**

In accordance with GAO Federal Information System Controls Audit Manual (FISCAM)<sup>16</sup>: An Interface Strategy should be developed to keep data synchronized between a source system and a target system. The interface Strategy should contain the following elements:

- An explanation of each interface
- The interface method chosen
- The data fields being interfaced

---

<sup>15</sup> From Chris Campbell <http://www.blue-granite.com/blog/bid/402596/Top-Five-Differences-between-Data-Lakes-and-Data-Warehouses>

<sup>16</sup> <http://www.gao.gov/assets/80/77142.pdf>

- The controls to reasonably ensure that the data is interfaced completely and accurately
- Timing requirements
- Security requirements.
- Interface design documentation, such as mapping tables that describe how data is transformed between a data source and destination, validations and edits, roles and responsibilities for the interface process, and error correction and communication methods, should also be developed for each interface
- Interface Reconciliation Controls between a source system, such as the use of control totals, record counts, hash totals, or batch run totals, would help ensure the complete and accurate transfer of data.

**Recommendation:** Establish an interface strategy to replicate data sets into the Data Lake, provide policies and procedures to manage interfaces, such as guidance for interface strategy and design documentation, interface reconciliation controls, and audit logs.

#### ***2.1.1.9 Partnership Agreements/Interconnection Security Agreements***

Obtaining the data, or access to data, is the first step to obtaining value from the data. Making the data available, accessible, interoperable, and useful requires significant and diligent work. These agreements will be necessary to extract the relevant data sets, transform and load data sets into the data warehouse. The partnership agreements will establish the roles and responsibilities of all relevant parties as they relate to protection of and management (including data integrity, access, use) of the data within the data warehouse. This will also establish data ownership and stewardship. Agencies that initially collect the data should be the data owners. Metadata should identify data owners, data source and rules specific to the data set. Best practices and compliance with the National Institute for Standards and Technology (NIST) Special Publication (SP) 800-35<sup>17</sup> guidelines indicate the value of having each participating agency execute a partnership agreement with the DTS to:

- Define roles and responsibilities of those charged with governance
- Provide clear guidance to help prevent data from being misused
- Help mitigate miscommunication of roles and responsibilities between data providers and recipients
- Specify the services being provided and make all parties aware of their roles, responsibilities, and performance expectations for IT services provided

---

<sup>17</sup> <http://csrc.nist.gov/publications/nistpubs/800-35/NIST-SP800-35.pdf>

The U.S. Department of Homeland Security (US DHS) has been a leader in adopting cross-agency information sharing and has implemented a data lake – the DHS Data Framework – to make the data they collect easy to share and increase its value to the Enterprise. Data sets from across the US DHS Enterprise are copied into a Data Lake where they exist independent of the IT systems on which they were originally collected. In this, they have identified the need for new policies to govern the data and ensure appropriate use and viewing.

The US DHS Data Framework Program effort yields many good insights regarding the privacy and security risks and need for new policies. The Program defines four elements for controlling data<sup>18</sup>:

- **User attributes** to identify characteristics about the user requesting access such as organization, clearance, and training
- **Data tags** that label the data based on the type of data involved, the authoritative system from which the data originated, and when it was ingested into the Framework
- **Context** that combines what type of search and analysis can be conducted (function), with the purpose for which data can be used (authorized purpose)
- **Dynamic access control policies** that evaluate user attributes, data tags, and context to grant or deny access to US DHS data in the repository based on legal authorities and appropriate policies of the Department and/or Components.

US DHS lessons learned include:

- Establish a scalable big data architecture and governance process to evaluate integration of new data, new missions, new users, and new analytical tools
- Follow an incremental development approach to allow opportunity to ensure new capabilities comply with established legal and policy requirements to protect privacy, civil rights, and civil liberties
- Incorporate the ability for redress and the ability to periodically refresh and update data to establish a long term operational utility that protects privacy, civil rights and civil liberties
- Engage stakeholders: mission operators, system administrators and data stewards
- Promote transparency to help the public understand how the Enterprise is using its data

---

<sup>18</sup> Privacy Impact Assessment Update DHS Data Framework, DHS/ALL/PIA-046(a), August 29, 2014 pages 10-16

**Recommendation:** Develop a standard Partnership Agreement template and streamlined process to facilitate data ingest. Oversee execution of Partnership Agreements with each agency or Program providing data sets.

#### **2.1.1.10 Data Sharing Agreements**

While Partnership Agreements will provide the protection and structure needed to facilitate agency participation in the Big Data program by sharing data sets to the Data Lake, Data Sharing Agreements are needed to manage the access to and sharing of specific data sets needed for analytics projects:

- Must include information related to securing shared data
- How long data can be retained after termination of the agreement
- Authority to conduct audits
- Restrictions on disclosure of information
- Security requirements over transferred data
- Method of data transfer
- Notification requirements if the data transfer method changes or an error in shared data is identified
- Responsibilities for completeness, accuracy, and timeliness of shared data.

A key risk mitigation strategy incorporated in the US DHS Data Framework Program approach is to maintain the authoritative data with the data owner IT system and make the data owner Agency responsible for identifying and communicating the applicable legal and policy protections needed as well as the necessary safeguarding measures including restriction on access or use of the data as well as specifying other rules related to the data – such as destruction. In this way, the data owners who are the experts in the laws and policies governing their data retain responsibility and DTS need only ensure compliance with these requirements.

**Recommendation:** Establish a standard template and process for Data Sharing/Data Use Agreements between Agencies or Programs to detail utilization of data, restrictions, and other expectations or requirements for analytics projects. A possible model for data sharing agreements exists in the Utah Digital Spatial Data Sharing and Integration Project<sup>19</sup> where use of a statewide agreement eliminates the necessity of developing multiple agreements between the individual participating agencies for the purpose of

---

<sup>19</sup> <https://www.fgdc.gov/grants/2009CAP/InterimFinalReports/088-09-5-UT-AppendixD-DataSharingMOU.pdf>

sharing data. This approach decreases the duplication of effort, promotes the exchange of information, and fosters communication between agencies in Utah.

#### **2.1.1.11 Temporary Privileged Account Access Controls**

Establish and implement effective access controls over temporary privileged accounts. Temporary Privileged accounts must have appropriate authorization, and should only be granted to appropriate users for valid business purposes. Temporary Privileged account use must be monitored. Ensure use of these accounts is appropriate and that no unauthorized changes be made to the data, database structure, or database configuration.

**Recommendation:** Design effective policies and procedures to authorize and monitor temporary privileged accounts. Establish automated, centralized process to facilitate authorization and review of temporary privileged accounts. Develop process for authorizing privileged account access for State employees and third party contractors and vendors.

#### **2.1.1.12 Access Control**

Agencies must fully establish and implement effective user access controls. Detect/prevent inappropriate access to and modification of data (inserts, update, and delete production data in the environment). Policies should be established to allow access to be managed, controlled, and periodically reviewed to ensure user access is based on current job responsibilities:

- Designate effective access request forms
- Access request forms should include user access rights
- Modification of data in the Environment should be controlled and monitored using temporary privileged access process
- State agencies must document approval of access granted to users<sup>20, 21</sup> Helps ensure that only appropriate individuals have access to the environment

---

<sup>20</sup> At its simplest, the Data Sharing Agreement should define the users by function or role and establish any restrictions. These are also captured in the Partnership Agreement with the Data Lake Manager. The expectation is that the Data Lake Manager provides audit and enforces compliance restrictions. Most approaches to facilitate this use a two-pronged approach whereby data's metadata and user's metadata sync to establish what data is 'visible' or available/accessible to any given user.

<sup>21</sup> In the Michigan EDW an early priority was setting governance rules for clients, who now number 10,000 individuals from 21 different state agencies. The state has developed secure access controls which determine who can see specific rows and columns of data, and protects confidential data.

- PUBLIC account should be restricted from having unnecessary access rights to system tables containing metadata and user data. Any rights granted to PUBLIC are automatically inherited by all users.
- Establish and implement controls for periodic review of user access rights to ensure that user access level remains appropriate for their job responsibilities
- State agencies should design effective policies and procedures governing the granting and periodic review of user access rights

**Recommendation:** Establish and implement effective user access controls to manage and control access to data, detect/prevent inappropriate access to and modification of data (insert, update, and delete production data in the environment) within the Data Lake. These controls should enable authenticated users with access to ‘see’ and access data that they are authorized to use in accordance with rules established by the data owners and documented in data sharing agreements.

## 2.2 How Big Data Technology is Different

Ways in which Security and Privacy in Big Data projects differ from traditional implementations:<sup>22</sup>

- Big Data projects often encompass heterogeneous components in which a single security scheme has not been designed from the outset.
- Most security and privacy methods have been designed for batch or online transaction processing systems. Big Data projects increasingly involve one or more streamed data sources that are used in conjunction with data at rest, creating unique security and privacy scenarios.
- The use of multiple Big Data sources not originally intended to be used together can compromise privacy, security, or both. Approaches to de-identify or anonymize PII that were satisfactory prior to Big Data may no longer be adequate, while alternative approaches to protecting privacy are made feasible. Although de-identification techniques can apply to data from single sources as well, the prospect of unanticipated multiple datasets exacerbates the risk of compromising privacy.
- An increased reliance on sensor streams, such as those anticipated with the Internet of Things (IoT); (e.g., smart medical devices, smart cities, smart homes)

---

<sup>22</sup> NIST Special Publication 1500-4, NIST Big Data Interoperability Framework: Volume 4, Security and Privacy, August 14, 2015

can create vulnerabilities that were more easily managed before amassed to Big Data scale.

- Certain types of data thought to be too big for analysis, such as geospatial and video imaging, will become commodity Big Data sources. These uses were not anticipated and/or may not have implemented security and privacy measures.
- Issues of veracity, context, provenance, and jurisdiction are greatly magnified in Big Data. Multiple organizations, stakeholders, legal entities, governments, and an increasing number of citizens will find data about themselves included in Big Data analytics.
- Volatility is significant because Big Data scenarios envision that data is permanent by default. Security is a fast-moving field with multiple attack vectors and countermeasures. Data may be preserved beyond the lifetime of the security measures designed to protect it.

**Recommendation:** Establish an Enterprise Big Data Scheme at the outset to address security and privacy issues specific to Big Data such as heterogeneous components, protection for data at rest and in motion (to include streaming data and sensor streams).

**Recommendation:** Review, revise, and/or expand upon consent and privacy statements at point of collection to inform citizens of expended data use. Ensure that mechanisms are in place for redress and corrections.

**Recommendation:** Include data preservation and destruction information and responsibility for monitoring, implementation, and verification in Partnership Agreements, Data Sharing Agreements as well as in the metadata.

### 2.2.1 Big Data Architecture and Associated Vulnerabilities

Some of the fundamental differences in Big Data architecture are as follows<sup>23</sup>:

- **Distributed Architecture:** Big data architecture is highly distributed on the scale of 1000s of data and processing nodes. Data is horizontally partitioned, replicated and distributed among multiple data nodes available.
- **Real-Time, Stream and Continuous Computations:** Performing computation real-time and continuously.

---

<sup>23</sup> <http://www.ivizsecurity.com/blog/penetration-testing/top-5-big-data-vulnerability-classes/>

- **Ad-hoc Queries:** Big data enables Knowledge Workers to create and execute data analyzing queries on the fly.
- **Parallel and Powerful Programming Language:** The computations performed in Big Data are much more complex, highly parallel and computationally intensive than traditional SQL / Procedural Language extension to SQL (PL/SQL) queries.
- **Move the code:** In Big Data, it is easy to move the code, rather than data.
- **Non-Relational Data:** The data stored in Big Data is non-relational.
- **Auto-tiering:** In Big Data, hottest data blocks are tiered into higher performance media, while the coldest data is sent to lower cost high capacity drives. As a result, it is extremely difficult to know precisely where the data is exactly located among the available data nodes.
- **Variety of Input Data Sources:** Big Data requires collecting data from many sources such as logs, end to point devices, social media etc.

### 2.2.2 Risks Associated with Big Data Technologies

The following examples represent some of the complexities with Big Data that are non-traditional causes for concern from both a security perspective and an IT governance perspective:<sup>24</sup>

- **Database structure:** Hadoop and other next-generation databases are designed for unstructured data.
- **Scalability:** Big Data technologies are often designed to “scale out,” or cluster. Instead of having a single large database server, an agency may have 500 smaller systems operating together as a cluster. Some of these systems could be virtual, some physical, and some in the cloud.
- **Configuration management:** Traditionally, the Federal Information Security Act (FISMA) (through Federal Information Processing Standards (FIPS)-200) has required agencies to develop robust configuration management plans, develop configuration and change management boards, and ensure that security impact analysis is performed as part of system changes. When working with big data, mature and robust configuration and change management is a must.
- **Cost:** Since new nodes could be spun up in almost any cloud provider’s environment, or even on additional desktops within an agency, tight control over IT resources and spending must be in place.

---

<sup>24</sup> <http://gcn.com/Articles/2013/07/29/ISC2-big-data.aspx?Page=2>

- **Operations:** Who is responsible for patching? Who is responsible for vulnerability scanning? What happens if the software has vulnerability and there is no vendor to contact for support?

**Recommendation:** Establish a robust configuration and change management plan. Provide for oversight and tight control over IT resources and spending. Ensure that even basic maintenance of operations and allocating additional resources are incorporated in the decision-making process. The security team must be aware of any changes being performed as part of the system lifecycle with big data platforms capable of utilizing cloud services.

#### **2.2.2.1 Insecure Data Storage and Communication**

There are multiple challenges related to data storage and communication in Big Data:

- Big Data implementations typically include open source code, with the potential for unrecognized back doors and default credentials
- The attack surface of the nodes in a cluster may not have been reviewed and servers adequately hardened
- User authentication and access to data from multiple locations may not be sufficiently controlled
- Regulatory requirements may not be fulfilled, with access to logs and audit trails problematic
- There is significant opportunity for malicious data input and inadequate data validation. Data is stored at various Distributed Data Nodes. Authentication, authorization and Encryption of data is a challenge at each node.
- Auto-tiering: Auto partitioning and moving of data can save sensitive data on a lower cost and less sensitive medium.
- Real Time analytics and Continuous computation requires low latency with respect to queries and hence encryption and decryption may provide additional overhead in terms of performance.
- Secure communication among nodes, middleware and end users
- Transactional logs of big data should be protected same as data

### **2.3 Using Data Derived from Multiple Contexts**

In public sector Big Data and analytics programs it is anticipated that agency data sets will be shared across the Enterprise using data in previously unanticipated ways.

Previously information or data sharing may have been negotiated between two or several agencies for a specified purpose. Combining of multiple data sets or data streams from multiple sources changes the relationship between the citizen or individual and the agency collecting the data. Because the Enterprise, and not specific agency, is entering the relationship without the implicit or explicit consent of the citizen this establishes a new non-linear relationship of greater complexity.

While combining multiple data sets and using data derived from multiple contexts (such as purchased data sets, private data sets, social media, etc.) has been proven extremely useful it also raises significant concerns regarding privacy and security. Using data from multiple contexts and data sets typically involves utilizing data for purposes other than identified when it was collected. The state will need to review, revise, and expand privacy statements, disclosure statements, and use policies in alignment with the FIPPS of Transparency, Individual Participation, Purpose Specification, Minimization, and Use Limitation. Figure 6 illustrates the intersections of Personal, Proprietary, Open, and Big Data.

**Recommendation:** Review, update and expand Privacy Policies and Consent statements where data is collected from citizens.

**Recommendation:** Provide a mechanism for review, and correction of data.

**Recommendation:** Provide a process for periodic refresh of data by the originating systems (data owner) to capture updates.

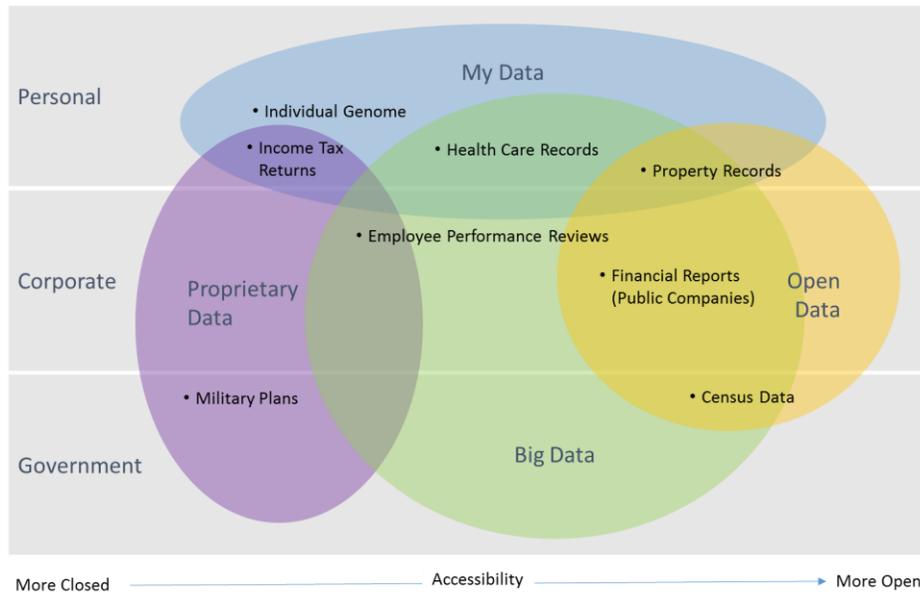


Figure 6. Intersections of Personal, Proprietary, Open, and Big Data

## 2.4 Privacy and Security

With public sector adoption of Big Data, analytics, and cloud hosting there has been significant attention to the need to expand upon, refine, and reinterpret Security and Privacy requirements. Privacy, Security and processes need to be balanced to enable what O'Reilly refers to as 'Democratization of Data'. Government needs to provide for efficient use of government data across the Enterprise to enable new and provide improved services and service delivery to citizens, provide transparency, and prevent fraud. As most public sector entities working in Big Data and analytics have identified, there is a real need to plan for and adopt a mechanism for interpretation of legal and policy requirements and restrictions.

Because of the volume and variety of data that will be collocated, it is imperative that IT security be appropriate to the most sensitive data that will be stored, processed, transmitted, or displayed. Because data will be utilized in conjunction with other data in previously unplanned relationships, it is possible that sensitivity levels will be greater for the aggregated/collocated data than for any single data set on its own.

Big Data has necessitates paradigm shifts in the understanding and enforcement of security and privacy requirements. Diverse datasets are becoming easier to access and increasingly contain personal content. A new set of emerging issues must be addressed, including balancing privacy and utility, enabling analytics and governance on encrypted data, and reconciling authentication and anonymity. Security and privacy measures are becoming ever more important with the increase of Big Data generation and utilization and increasingly public nature of data storage and availability. (Public Cloud)<sup>25</sup>

Security and privacy measures for Big Data involve a different approach than traditional systems. Big Data is increasingly stored on public cloud infrastructure built by employing various hardware, operating systems, and analytical software. Traditional security approaches usually addressed small-scale systems holding static data on firewalled and semi-isolated networks. The surge in streaming cloud technology necessitates extremely rapid responses to security issues and threats.<sup>26</sup>

---

<sup>25</sup> **NIST Special Publication 1500-4, NIST Big Data Interoperability Framework: Volume 4, Security and Privacy, August 14, 2015**

<sup>26</sup> Big Data Working Group, "Expanded Top Ten Big Data Security and Privacy Challenges," *Cloud Security Alliance*, April 2013, [https://downloads.cloudsecurityalliance.org/initiatives/bdvwg/Expanded\\_Top\\_Ten\\_Big\\_Data\\_Security\\_and\\_Privacy\\_Challenges.pdf](https://downloads.cloudsecurityalliance.org/initiatives/bdvwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf)

### 2.4.1 Privacy

This Section examines privacy issues impacting cross agency data sharing for Big Data. Privacy concerns top the policy and legal issues for data sharing and combining data from multiple sources in a Big Data context. The need to protect privacy, civil rights, and civil liberties is not new. The sheer quantities of PII now held in one repository makes it an attractive target that necessitates a greater degree of attention. Because of the fusion of multiple data sets across agencies revealing ever more detailed information about individual citizens and new uses of data collected for other purposes, much attention is focused on the need to expand on existing privacy laws and policies to protect citizens.

The collection, storage, manipulation and retention of massive amounts of data have resulted in serious security and privacy considerations. Various regulations are being proposed to handle Big Data so that the privacy of the individuals is not violated. For example, even if personally identifiable information is removed from the data, when data is combined with other data, an individual can be identified. This is essentially the inference and aggregation problem that data security researchers have been exploring for the past four decades. This problem is exacerbated with the management of Big Data as different sources of data now exist that are related to various individuals.

**NIST Special Publication 800-53 Revision 4 Privacy Controls:** NIST has issued an updated and Revised Special Publication 800-53 (NIST SP 800-53 Revision 4) Security and Privacy Controls for Federal Information Systems and Organizations. Appendix J identifies new controls related to Privacy. Now there are greater implications with respect to controlling the integrity of an individual's information, and with ensuring that an individual's information is available on demand. The challenging landscape requires federal organizations to expand their view of privacy, in order to meet citizen expectations of privacy that go beyond information security<sup>27</sup>.

Protecting the privacy of individuals and their PII that is collected, used, maintained, shared, and disposed of by programs and information systems, is a fundamental responsibility of federal organizations. Privacy also involves each individual's right to decide when and whether to share personal information, how much information to share, and the particular circumstances under which that information can be shared.<sup>28</sup> Privacy controls are the administrative, technical, and physical safeguards employed within organizations to protect and ensure the proper handling of PII. The privacy controls in this appendix are based on the FIPPs<sup>121</sup> embodied in the Privacy Act of 1974, Section

---

<sup>27</sup> NIST Special Publication 800-53 Revision 4 Security and Privacy Controls for Federal Information Systems and Organizations Appendix J pJ-1

<sup>28</sup> NIST Special Publication 800-53 Revision 4 Security and Privacy Controls for Federal Information Systems and Organizations Appendix J pJ-1

208 of the E-Government Act of 2002, and Office of Management and Budget (OMB) policies. There are eight privacy control families, each aligning with one of the FIPPs. These privacy control families are as follows:

1. Authority and Purpose (AP)
2. Accountability, Audit, and Risk Management (AR)
3. Data Quality and Integrity (DI)
4. Data Minimization and Retention (DM)
5. Individual Participation and Redress (IP)
6. Security (SE)
7. Transparency (TR)
8. Use Limitation (UL)

**FIPPs:** To comply with FIPPS, public sector entities undertaking Big Data programs should review, update and expand Privacy Policies and statements where data is collected from citizens. Table 3 provides an overview of organizational data management responsibilities to comply with FIPPS.

Table 3. Organizational Data Management Responsibilities to Comply with FIPPS

| FIPPS Principle                                | Data Management Responsibilities   |
|--|--|
| <b>Principle of Transparency</b>               | <i>Be transparent and provide notice to the individual regarding its collection, use, dissemination, and maintenance of PII. Technologies or systems using PII must be described in a System of Record Notice (SORN) and Privacy Impact Assessment (PIA), as appropriate. There should be no system the existence of which is a secret</i> |
| <b>Principle of Individual Participation</b>   | <i>Involve the individual in the process of using PII. To the extent practical, seek individual consent for the collection, use, dissemination, and maintenance of PII and should provide mechanisms for appropriate access, correction, and redress regarding use of PII</i>  |
| <b>Principle of Purpose Specification</b>      | <i>should specifically articulate the authority which permits the collection of PII and specifically articulate the purpose or purposes for which the PII is intended to be used</i>   |
| <b>Principle of Data Minimization</b>          | <i>Should only collect PII that is directly relevant and necessary to accomplish the specified purpose(s) and only retain PII for as long as is necessary to fulfill the specified purpose(s). PII should be disposed of in accordance with records disposition schedules</i>  |
| <b>Principle of Use Limitation</b>             | <i>Should use PII solely for the purpose(s) specified in the notice. Sharing PII should be for a purpose compatible with the purpose for which the PII was collected</i>   |
| <b>Principle of Data Quality and Integrity</b> | <i>should, to the extent practical, ensure that PII is accurate, relevant, timely, and complete within the context of each use of the PII</i>  |
| <b>Principle of Security</b>                   | <i>should protect PII (in all forms) through appropriate security safeguards against</i>   |

| FIPPS Principle                                 | Data Management Responsibilities   |
|---|--|
|   | <i>risks such as loss, unauthorized access or use, destruction, modification, or unintended or inappropriate disclosure</i>  |
| <b>Principle of Accountability and Auditing</b> | <i>should be accountable for complying with these principles, providing training to all employees and contractors who use PII, and should audit the actual use of PII to demonstrate compliance with these principles and all applicable privacy protection requirements</i> |

**NIST Special Publication 1500-4 NIST Big Data Interoperability Framework:** Volume 4, Security and Privacy explores security and privacy topics with respect to Big Data. This volume considers new aspects of security and privacy with respect to Big Data, reviews security and privacy use cases, proposes security and privacy taxonomies, presents details of the Security and Privacy Fabric of the NIST Big Data Reference Architecture (NBDRA), and begins mapping the security and privacy use cases to the NBDRA.

**Big Data Impact on Key Privacy Legislation:** As evidenced by the White House Reports “Big Data and Privacy: A Technological Perspective”, May 2014; and “Big Data: Seizing Opportunities, Preserving Values”, May 2014 and the release of the “NIST Big Data Interoperability Framework Volume 4: Security and Privacy” August 2015, the jury is still out on how Privacy and Security policies need to be changed to accommodate the impact of Big Data and to protect the privacy, civil rights, and civil liberties of Americans. This is an active topic worldwide. Table 4 below addresses several of the more troublesome areas for information sharing (health information, child-related data, and PCI data) and complexities for compliance in data sharing in a Big Data and analytics context.

Table 4. Specific Privacy Policies and Complexities the Present to Data Sharing

| Privacy Policy   | Data Sharing Complexities  |
|--|--|
| <ul style="list-style-type: none"> <li>● Family Educational Rights and Privacy Act of 1974 (FERPA)</li> <li>● Children’s Online Privacy Protection Act of 1998 (COPPA) Department of Education, Protecting Student Privacy While Using Online Educational Services: Requirements and Best Practices, Feb 2014: Schools and districts can enter into agreements with third parties involving student data only so long</li> </ul> | <p>Schools may disclose, without consent, "directory" information such as a student’s name, address, telephone number, date and place of birth, honors and awards, and dates of attendance. However, schools must tell parents and eligible students about directory information and allow parents and eligible students a reasonable amount of time to request that the school not disclose directory information about them. Schools must notify parents and eligible students annually of their rights under FERPA.</p> <p>FERPA requires that whenever a school shares student data, it “must retain ‘direct control’ over that information.” But FERPA does a poor job of ensuring that schools remain in control. Mandating specific</p> |

| Privacy Policy  | Data Sharing Complexities  |
|---|--|
| <p>as requirements under the Family Educational Rights and Privacy Act and Protection of Pupil Rights Amendment are met</p> | <p>contractual requirements is a first step toward retaining such control.</p> <p>Big Data can provide unprecedented insight into how students are learning and what educational techniques are effective. It has been recommended that Congress “modernize the privacy regulatory framework under the FERPA and COPPA to: 1) protect students against their data being shared or used inappropriately, especially when that data is gathered in an educational context, and 2) ensure that innovation in educational technology, including new approaches and business models, have ample opportunity to flourish.”<sup>29</sup></p>  |
| <ul style="list-style-type: none"> <li>Health Insurance Portability and Accountability Act (HIPAA) of 1996</li> </ul>       | <p>Protects the privacy of individually identifiable health information; the HIPAA Security Rule, which sets national standards for the security of electronic protected health information; the HIPAA Breach Notification Rule, which requires covered entities and business associates to provide notification following a breach of unsecured protected health information; and the confidentiality provisions of the Patient Safety Rule, which protect identifiable information being used to analyze patient safety events and improve patient safety. HIPAA has a set of required contractual elements before data can be shared.</p> <p>Much of the “Big Data” discussion is outside of the context of health care, BUT there is a wide variety of health care information (both HIPAA regulated and not) that is being scrutinized in the context of Big Data and there is a growing range of “Big Data” activities being conducted by healthcare entities, both in and out of HIPAA.</p> |
| <ul style="list-style-type: none"> <li>Health Information Technology for Economic and Clinical Health (HITECH)</li> </ul>   | <p>This legislation anticipates a massive expansion in the exchange of electronic protected health information (ePHI): It widens the scope of privacy and security protections available under HIPAA; it increases the potential legal liability for non-compliance; and it provides for more enforcement.</p> <p>Imposes data breach notification requirements for unauthorized uses and disclosures of "unsecured PHI." Under the <a href="#">HITECH Act</a> "unsecured PHI" essentially means "unencrypted PHI."</p>  |
| <ul style="list-style-type: none"> <li>Health Information Trust Alliance (HITRUST)</li> </ul>                               | <p>Ensure that Information security becomes a core pillar of, rather than obstacle to, the broad adoption of health information systems and exchanges. Challenges in information security include:</p> <ul style="list-style-type: none"> <li>Redundant and inconsistent requirements and standards for healthcare organizations</li> <li>Inconsistent adoption of minimum controls</li> </ul>   |

<sup>29</sup> <http://safegov.org/2014/5/6/big-data-and-our-children%E2%80%99s-future-on-reforming-ferpa>

| Privacy Policy  | Data Sharing Complexities  |
|---|--|
|   | <ul style="list-style-type: none"> <li>• Inability to implement security in medical devices and healthcare applications</li> <li>• Rapidly changing business, technology and regulatory environment</li> <li>• Ineffective and inefficient internal compliance management processes</li> <li>• Inconsistent business partner requirements and compliance expectations</li> <li>• Increasing scrutiny from regulators, auditors, underwriters, customers, and business partners</li> <li>• Growing risk and liability associated with information security</li> </ul> <p>For population health management to be successful, there needs to be a blend of predictive analytics, chronic care management, a timely feedback mechanism and measurable outcomes. HITRUST has created a new framework based around de-identification of sensitive patient information. The framework provides guidance on use of de-identification in a simplified and streamlined way through standards and controls that also adhere to HIPAA's privacy rules, as well as HITRUST CSF. .</p> |
| <ul style="list-style-type: none"> <li>• Payment Card Industry (PCI)</li> </ul> | <p>The PCI Data Security Standard specifies twelve requirements for compliance, organized into six logically related groups called "control objectives".</p> <p>As the requirements of an increasing variety of risk, conduct, transparency, and technology standards grow to Exabyte scale, agencies are struggling with new compliance challenges, particularly as they relate to balancing data administration cost and complexity, data-intensive operations, and the value of insights that can be gained from such large, rich datasets.</p> <p>The most straightforward way to comply with the PCI DSS requirement to protect stored cardholder information is to encrypt all data-at-rest and manage the encryption keys away from the protected data.</p>   |

**2.4.2 Security**

This section provides a general overview of Security Requirements. Big Data is a powerful new tool subject to the same legal, regulatory and policy considerations as existing information technology. Big Data expands the boundaries of existing information security responsibilities and introduces significant new risks and challenges. With compilation of multiple data sets in a single Enterprise Data Warehouse information

classification becomes even more critical. Critical security requirements for Big Data and analytics include: Proper authentication; Access Control; File System Integrity; Data Validation; Operating System Hardening.

### 2.4.3 Security/Cloud Hosting Security Requirements

The State of Utah DTS IT Security Policy follows the NIST 800 controls. It should be noted that current IT Security Policy references NIST SP800-53 Rev 3 which has been superseded by Revision 4. To be in compliance with current security guidelines, the State will need to update its policies to respond to the updated and expanded controls set forth in NIST SP 800-53 Rev. 4. For the most part, existing IT Security policies (assuming projected compliance with NIST 800-53 Rev 4) should be appropriate to meet the IT security requirements for new physical infrastructure supporting Big Data Analytics effort.

**Recommendation:** Update DTS IT Security Plans to be in compliance with and reference the controls identified in NIST SP 800-53 Rev 4.

#### 2.4.3.1 Security Issues

As reported in Government Computer News (GCN), agencies often approach big data as if it were an expansion of or significant increase in their database capability. However, Big Data encompasses new tools, technologies, and deployment and operational methods. From an information security perspective, big data can mean “big exposure” to risk if approached solely from a traditional IT perspective.<sup>30</sup> Certain aspects of big data include traditional IT approaches with traditional challenges that do not require an entirely new perspective. Many agencies already have the foundation laid for developing an approach to Big Data security. That foundation includes mature processes for cloud computing, continuous monitoring and FISMA compliance.

As agencies optimize their continuous monitoring capabilities, they can utilize existing tools that support big data, including vulnerability management and patching services. While these capabilities are all necessary first steps to approaching big data security, a new perspective is required when considering the differences between big data and the large data processing and storage of the past.<sup>31</sup> Security teams will need to rely largely on an array of operational and managerial techniques — including segmentation and robust, auditable access controls — to help ensure big data does not become “big exposure.” Security teams must look at big data from a holistic perspective of protecting the

---

<sup>30</sup> <http://gcn.com/articles/2013/07/29/isc2-big-data.aspx>

<sup>31</sup> <http://gcn.com/articles/2013/07/29/isc2-big-data.aspx>

infrastructure and operating system, applying as much automation and existing policy as possible.

Security teams will need to become more integrated and involved in the lives of data scientists and business units to understand how they are operating and where they need support. While big data is new to many agencies, the principles in protecting information and bringing mature management to an operation often is not. Agencies should leverage their existing operational and managerial controls to protect new technologies while automated tools are developed to add further rigor, maturity and automation.<sup>32</sup>

**Multilevel protection of data processing nodes** means implementing security controls at the application, operating system and network level while keeping a bird's eye on the entire system using actionable intelligence to deter any malicious activity, emerging threats and vulnerabilities<sup>33</sup> Techniques such as attribute based encryption may be necessary to protect sensitive data and apply access controls (being attributes of the data itself, rather than the environment in which it is stored).

**Recommendation:** If, Utah wishes to share information they must conform to FISMA, the Federal Risk and Authorization Program (FedRAMP) and Control Objectives for Information and Related Technology (COBIT), plus the ones listed in 2.1.2.1 for Cloud.

**Recommendation:** Implement segmentation and robust, auditable access controls.

#### 2.4.4 Security for Big Data in a Cloud Environment

Processing and storing private information in the cloud means organizations won't always know where their data resides, yet they still need to comply with privacy laws and be able to demonstrate this compliance. These problems are confounded by traditional information protection methods, which may be difficult to apply or ineffective in the cloud. New IT security guideline specific to use of Cloud has recently been issued by the International Organization for Standardization and International Electrotechnical Commission (ISO/IEC) ISO/IEC 27018:2014 establishes commonly accepted control objectives, controls and guidelines for implementing measures to protect PII in accordance with the privacy principles in ISO/IEC 29100 for the public cloud computing environment and specifies guidelines based on ISO/IEC 27002, taking into consideration the regulatory requirements for the protection of PII which might be applicable within the context of the information security risk environment(s) of a provider of public cloud services.

---

<sup>32</sup> <http://gcn.com/Articles/2013/07/29/ISC2-big-data.aspx?Page=2>

<sup>33</sup> <http://www.elementalsecurity.com/bigdata/>

Cloud computing may seem risky because you cannot secure its perimeter—where are a cloud’s boundaries? In addition, many agencies must comply with specific regulatory statutes, such as the HIPAA, the Sarbanes–Oxley Act of 2002 (SOX), and FISMA. Yet your organization can move forward even while security standards are being defined. NIST likens the adoption of cloud computing to that of wireless technology. Agencies learned how to protect their wireless data as they moved forward—and they will do the same with cloud computing. It comes down to this: Federal, state, and local agencies vary in their security and regulatory compliance needs, and you know your needs best. You must look carefully at how well cloud providers protect key functions and sensitive data.

**The recommended Cloud Security checklist** includes the following:

- **Integration.** Look for integration points with security and identity management technologies you already have, such as Active Directory, and controls for role-based access and entity-level applications.
- **Privacy.** Make sure a cloud service includes data encryption, effective data anonymization, and mobile location privacy. In federal agencies, your contract with the service provider should include provisions for complying with the Privacy Act of 1974.<sup>iv</sup>
- **Identity and access.** When you place your resources in a shared cloud infrastructure, the provider must have a means of preventing inadvertent access. How can identities federate across different services and from your internal environment to the cloud? How are the databases protected for access?
- **Compliance.** What certifications does your provider possess? How do you handle dispute resolution and liability issues? What industry or government standards do you comply with? Are there clearly defined metrics for the cloud service to be monitored? How are e-discovery and criminal compliance requests handled? What are the processes to move into the cloud and back?
- **Service integrity.** How is the software protected from corruption (malicious or accidental)? How does your provider ensure the security of the written code? How do they do threat modeling? What is the hiring process for the personnel doing administrative operations? What levels of access do they have?
- **Jurisdiction.** The location of a cloud provider’s operations can affect the privacy laws that apply to the data it hosts. Does your data need to reside within your legal jurisdiction? Federal records management and disposal laws may limit the ability of agencies to store official records in the cloud.
- **Information protection.** Who owns your data? Can it be encrypted? Who has access to encryption keys? Where is the backup located, and do you have an on-premise

backup? How is the backup purged? What requirements do you have with regard to the physical location of your data?

### 3 Liability/Risks

The average cost to the government of a data breach has been estimated at \$5.5 million or \$194 per individual record. Data breaches and cyber-attacks also affect citizens directly.

Big data projects are complex undertakings at best. This is especially true in the public sector, where such projects often require large infrastructure changes, program designs and agreements across agencies and departments. This section addresses potential areas of risk for the state in undertaking a Big Data and analytics program. Figure 7 illustrates individual versus organizational risks and potential mitigation strategies.

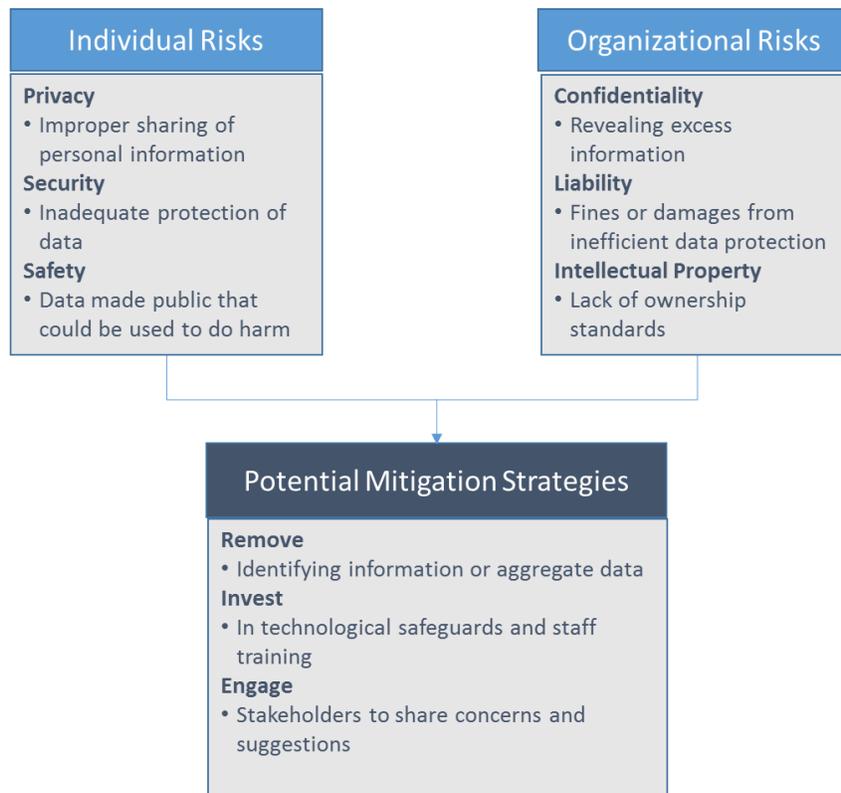


Figure 7. Individual versus organizational risks and potential mitigation strategies

The unauthorized use or misuse of personally identifiable information can impact an individual’s ability to get a job, secure a loan, pay for education, obtain insurance, defend

against identity theft, or benefit from public programs. Citizens need to know that they can trust public organizations with their personal information.<sup>3435</sup>

The citizens stand to benefit from the outcomes of a Big Data and analytics program. At the same time, as with any Big Data effort, it also creates potential risks to individual citizens whose privacy, civil rights, or civil liberties may be impacted negatively by either Enterprise program management or program output. It is critical that the Enterprise carefully plan the effort including appropriate IT and information security and incorporate sound data governance and mitigation strategies. As illustrated in Figure 7, mitigation strategies to protect citizens and limit the liability of the Enterprise include:

- Limitation and/or removal of personally identifiable information from aggregate data wherever possible. As has been widely discussed in Big Data forums, once multiple data sets are aggregated individuals may be fairly readily identified. Likewise, while anonymization techniques are in broad use, there are de-anonymization tools. A Data Strategy and Data Management/Data Governance Plan should identify an Enterprise approach to protecting PII and strictly limiting its use.
- Establishing a secure IT infrastructure using well-trained staff following strict processes is the best defense. It is critical that staff be provided with appropriate tools, knowledge, and skills needed to protect the systems and the data they contain.
- Transparency or stakeholder engagement has proven valuable to both public sector and private sector Big Data efforts. Providing stakeholders with a clear understanding of the undertaking, its anticipated benefits, as well as possible risks and how the Enterprise will mitigate or manage risk helps alleviate concerns.

In an article entitled '17 Steps to Implement a Public Sector Big Data Project'<sup>36</sup> the author states that: "Public sector databases contain citizens' data, making them valuable targets. You must assesses the potential impact of compromised data and develop a risk mitigation plan with processes for reducing the risks. It is important to consider who has

---

<sup>34</sup> [https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/white-paper/safeguarding-sensitive-data-state-local-governments\\_white-paper\\_2382.pdf](https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/white-paper/safeguarding-sensitive-data-state-local-governments_white-paper_2382.pdf)

<sup>35</sup> [https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/white-paper/safeguarding-sensitive-data-state-local-governments\\_white-paper\\_2382.pdf](https://www.informatica.com/content/dam/informatica-com/global/amer/us/collateral/white-paper/safeguarding-sensitive-data-state-local-governments_white-paper_2382.pdf)

<sup>36</sup> <http://www.cio.com/article/2368491/big-data/144854-17-Steps-to-Implement-a-Public-Sector-Big-Data-Project.html>

access to data, how much sensitive information is returned when database queries are made and what the physical security surrounding server rooms is. You should also develop a communications plan alongside the risk mitigation plan to ensure that messages are accurate and advance the goals of your agency or program. The communications plan should include dealing with press, academia and other agencies.”

Other risks include:

- Failure to capitalize on Big Data analytics within the Enterprise to identify and mitigate risks. Data mining can enable a much more efficient audit process by allowing auditors to focus more quickly on critical areas.
- There are potentially significant legal, compliance, and ethical risks associated with unfettered exploitation of Enterprise data. It is important to establish Governance with a clear understanding of the need, role, and potential of big data to lead the effort or respond to the potential challenges.
- Bad Analytics: There is a risk of misinterpreting patterns or assuming causal links. Very large volumes of data involving many variables have a high probability of displaying bogus patterns or correlations, thereby establishing relationships between variables by the sheer volume of sample data, where such relationships do not exist. Often as a result of such erroneous analytics, managers are misled into making wrong decisions.

**Recommendation:** Develop a Risk Mitigation Plan. Identify potential risk areas and specific risks. Identify mitigation strategies and specific activities to minimize risks or their impact if realized. Establish a Risk Management Board and processes to regularly meet and review possible emergent risks and address appropriate mitigation actions.

**Recommendation:** Implement a Data Management Plan and Data Governance to provide establish, implement, and oversee processes to ensure secure and appropriate access to and use of data.

**Recommendation:** Implement a Communication Plan to provide transparency and inform and engage stakeholders on plans, progress, and problems. The Communication Plan should include a Crisis Communications Plan for addressing data breaches or other risks that are realized.

**Recommendation:** Establish Key Risk Indicators within the system. Build in the capability to utilize analytics for security monitoring, system situational awareness, and audit.

## 4 Examples from Other States

A recent report by the National Association of State Chief Information Officers describes state governments as “enormous data generation engines.” State officials who treat data as an asset, analyzing it to discover new patterns, correlations and insights can “gain a competitive advantage.”<sup>37</sup> Federal, state and local agencies across the country are embracing Big Data and analytics to improve outcomes. Within the past five years, several states have embarked on Enterprise level Big Data Programs. Of note are the Indiana State Enterprise Management Performance Hub (MPH); North Carolina Government Data Analytics Center (GDAC); and Michigan State Enterprise Information Management Program and Integrated Data Warehouse (IDW).

### 4.1 Indiana Management and Performance Hub (MPH)

Indiana’s MPH consists of three parts: The initiative itself and system for sharing data among agencies established by [Executive Order 14-06](#)<sup>38</sup>; The MPH transparency website/portal<sup>39</sup>; and, the MPH Technology Center<sup>40</sup> which connects to and builds on the transparency portal. MPH is coordinated by the Office of Management and Budget and the Indiana Office of Technology (IOT). Mission: Develop an industry leading comprehensive enterprise wide data driven management system. Vision: Indiana will have the most effective, efficient, and transparent state government in the country. “The MPH effort has executive sponsorship, with the governor and his OMB as its driver. Second, MPH is a collaborative effort across all state agencies, coordinated by OMB and IOT and drawing on dedicated contributors with diverse job descriptions. And third, the pool of government data is both large and largely in the right place.”<sup>41</sup> All state agencies are be required to provide data, system access or other requested resources to the project. The hub's top priority is to tackle Indiana’s infant mortality problem.

### 4.2 North Carolina Government Data Analytics Center (GDAC)

The vision for the GDAC is to transform existing data assets into an information utility for the State’s policy and operational leaders for their use in making program investment decisions, managing resources, and improving financial programs, budgets, and results. A key function of the GDAC is the management of data sharing and integration initiatives,

---

<sup>37</sup> <http://www.ncsl.org/research/telecommunications-and-information-technology/big-data-big-benefits.aspx>

<sup>38</sup> [http://www.in.gov/gov/files/Executive\\_Order\\_14-06.pdf](http://www.in.gov/gov/files/Executive_Order_14-06.pdf)

<sup>39</sup> [www.in.gov/mph](http://www.in.gov/mph)

<sup>40</sup> <http://www.govtech.com/data/Indiana-Management-and-Performance-Hub-Takes-Transparency-to-the-Next-Level.html>

<sup>41</sup> <http://www.governing.com/blogs/bfc/gov-indiana-managment-performance-hub-data-analytics.html>

including identifying opportunities where data sharing and integration can generate greater efficiencies and improved service delivery by State agencies, institutions and departments. The GDAC manages enterprise program activities as well as the development and support of analytics projects and systems including the North Carolina Financial Accountability and Compliance Technology System (NC FACTS) fraud, waste and improper payment detection project, the Criminal Justice Law Enforcement Automated Data Services (CJLEADS) criminal justice system, and state reporting and analytics efforts. It consists of three program areas:

- GDAC Program Management and Business Services: Provides business services and execution and expansion of new and existing data sharing agreements
- CJLEADS Operations and GDAC Technical Environment
- GDAC Solution Development: Data Integration identify key sources of data in performing the extract, transformation, load and quality analysis of that data to support enterprise analytics; Fraud and Compliance Alerts – NC FACTS will leverage the SAS Fraud Framework not only for fraud, waste and improper payment detection but also to support other areas of compliance analytics and alerts such as worker’s compensation coverage compliance; Reporting and Analytics – solution development will leverage SAS reporting and analytics tools to support business needs for program management metrics and analysis such as the State Health Plan of North Carolina’s analytics repository

Session Law 2012-142<sup>42</sup>, HB 950, expanded the State’s current data integration and business intelligence initiatives by creating the OSC Government Business Intelligence Competency Center (GBICC) to manage the State’s enterprise data integration and business analytics efforts. Session Law 2013-360<sup>43</sup>, SB 402, amended Article 9 of Chapter 143B and codified the data integration and business intelligence, changed the name of the program to the GDAC, and authorized recurring administrative appropriations. Session Law 2013-360, SB 402, also directed the transfer of the GDAC program to the Office of the State Chief Information Officer effective July 1, 2104.<sup>44</sup>

---

<sup>42</sup> <http://www.ncleg.net/Sessions/2011/Bills/House/PDF/H950v7.pdf>

<sup>43</sup>

[http://www.ncleg.net/fiscalresearch/budget\\_legislation/budget\\_legislation\\_pdfs/2013/conference/S402v7.pdf](http://www.ncleg.net/fiscalresearch/budget_legislation/budget_legislation_pdfs/2013/conference/S402v7.pdf)

<sup>44</sup> [http://gdac.nc.gov/documents/GDAC\\_Legis\\_Report\\_Oct\\_2013.pdf](http://gdac.nc.gov/documents/GDAC_Legis_Report_Oct_2013.pdf)

### 4.3 Michigan Enterprise Data Warehouse (EDW) and Enterprise Information Management (EIM)

The Michigan EDW is a centralized repository of historical data that is used to support State agencies' decision-making and business processes. DTMB, in conjunction with State agencies, extracts data from source systems, transforms it into the proper format, and loads it into EDW. State agencies use analytical tools to query the data stored on the EDW to generate State and federal reports, project State revenues, perform trend analyses, and detect fraud.<sup>45</sup> The EDW consists of over 9,600 production tables containing 121.5 billion rows of data. Much of the data is sensitive or confidential.

Executive Directive 2013-1<sup>46</sup> Data and Information Sharing, Management and Governance established the Enterprise Information Management (EIM) program to improve upon the sharing and management of data across all executive branch agencies. The Director of the DTMB is responsible for the establishment and implementation of EIM. The Directive requires participation and engagement by all Executive Branch departments and agencies to establish new and improved protocols for data and information sharing, management, and governance. The EIM program includes a cross-agency data sharing protocol, a Michigan Information Management Governance Board, an information management. There is an implementation plan for each state department, and a five-year Michigan Statewide Data and Analytics Plan "All State departments and agencies must work in partnership with DTMB to establish procedures and protocols for cross-departmental and jurisdictional data sharing and processing. The Michigan Information Management Governance Board (MIMGB) is the primary governing body, to be chaired by a representative from the Governor's Legal Counsel. The MIMGB will have membership representation from Directors or Chief Deputy Directors of all Executive Branch departments and agencies. The MIMGB responsibilities are to adopt, support, and provide advice regarding all activities related to achieving the goals of the EIM program.

In its September 2014 Digital Strategy, the State of Michigan DTMB identified its Target State for 2018<sup>47</sup>:

- Michigan will be the first state in the country to operationalize enterprise-wide data governance and truly manage its data as an asset;

---

45

[https://www.michigan.gov/documents/mdch/Enterprise\\_Data\\_Warehouse\\_465692\\_7.pdf](https://www.michigan.gov/documents/mdch/Enterprise_Data_Warehouse_465692_7.pdf)

<sup>46</sup> [http://www.michigan.gov/documents/snyder/ED\\_2013-1\\_439597\\_7.pdf](http://www.michigan.gov/documents/snyder/ED_2013-1_439597_7.pdf)

<sup>47</sup> [http://www.michigan.gov/dtmb/0,5552,7-150-56345\\_56351-336646--,00.html](http://www.michigan.gov/dtmb/0,5552,7-150-56345_56351-336646--,00.html)

- Michigan will aggressively pursue a share first, open data policy to drive unprecedented government transparency and citizen engagement;
- By 2015 Michigan will have identified all Master Data across state agencies; all departments of state government will have identified and empowered a functional Chief Data Steward, and the time and resources expended on data sharing will have been reduced by 50%. Governance will enable data and analytics driven policymaking and service delivery by Michigan's departments of state government, driving true value for money government.

## 5 Making Data Accessible/Shared by Default

Open data policies differ by state, but most have some common elements. Key among these are requirements that data be open by default. State governments and agencies publish all information, such as public records, expenditure information, and legislative records, as a matter of course, unless there is an overruling justification against it, such as confidentiality and privacy reasons. Many open data policies also require that data be released in a non-proprietary, machine-readable format. Machine readability is crucial for ensuring that businesses, non-profits, and others can easily process and repurpose public data sets. Typically, policies also specify that data be made available to the public for any purpose, and often at no cost<sup>48</sup>.

Figure 8 illustrates the spectrum of open data including accessibility, machine readability, cost, and rights from completely open to completely closed.

---

<sup>48</sup> <http://www.datainnovation.org/2014/08/state-open-data-policies-and-portals/>

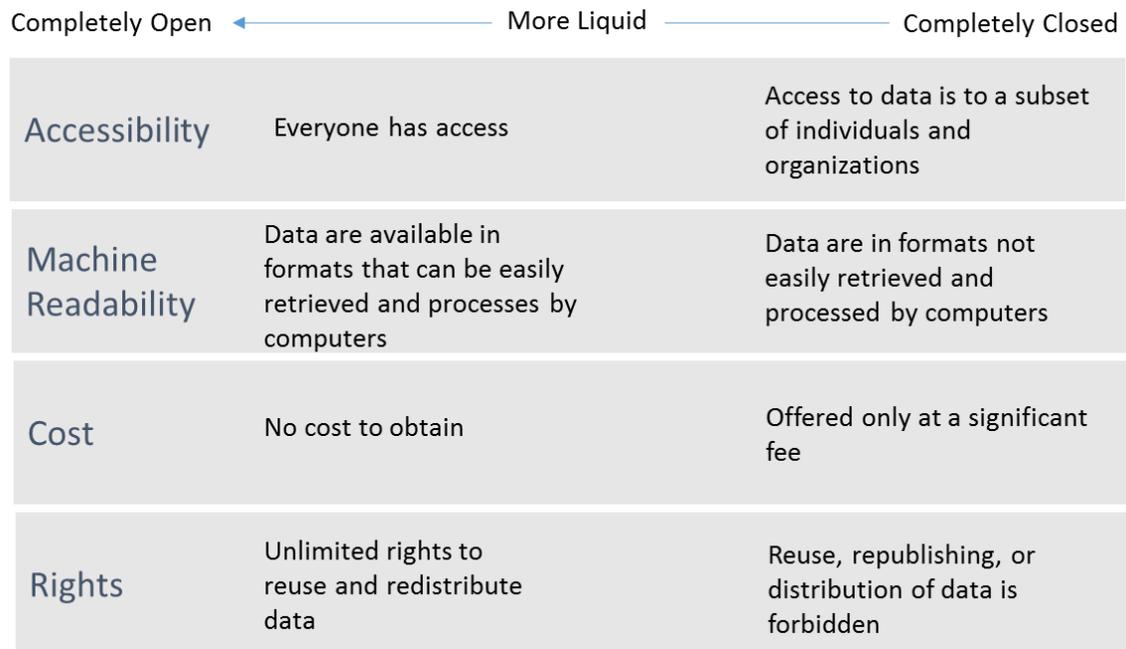


Figure 8. Spectrum of open data

The Center for Data Innovation has ranked Utah in the top six of 10 states with Open Data policies in the US awarding it 8 of a possible 8 points.<sup>49</sup> Open Data Policies were ranked based on the following criteria:

- Presence of an Open Data Policy
- Quality of Open Data Policy: If the policy applies broadly to all government data with extra points if the policy specifies that only certain types of government data must be provided, e.g., only spending information and if a state's open data policy specified machine-readability.
- Presence of an Open Data Portal
- Quality of Open data Portal: Points were awarded for machine readable data sets (assessed by identifying whether more than 50% of the data files on a portal could be downloaded in CSV, JSON, KML, or other such file formats).

<sup>49</sup> Legislation: Utah Code Ann. §§ 63A-3-401 - 406

## Appendix A: Legislation, Executive Orders, Executive Directives from other States

### A-1: Indiana Big Data Legislation

**STATE OF INDIANA<sup>50</sup>  
EXECUTIVE DEPARTMENT  
INDIANAPOLIS**

**EXECUTIVE ORDER 14-06**

**FOR: ESTABLISHING THE GOVERNOR'S MANAGEMENT AND PERFORMANCE HUB**

TO ALL WHOM THESE PRESENTS MAY COME, GREETINGS.

WHEREAS, Hoosiers can benefit from a comprehensive and coordinated effort by state agencies to share data and improve and strengthen services, maximize the utilization of available resources, and ensure that state services are available for all Hoosiers;

WHEREAS, the Indiana Office of Technology (IOT), established under Ind. Code § 4-23.I-2-1, stores data and has the responsibility to ensure the protection of data in compliance with all applicable laws;

WHEREAS, state agencies, as defined at Ind. Code § 4-12-I-2(d), administer Indiana taxpayer and federal funds in the name of and on behalf of the State of Indiana; operate on property or in buildings owned, maintained, or leased by the State of Indiana; use vehicles and equipment owned by the State of Indiana; manage and provide benefits to state employees; enter into contracts on behalf of the State of Indiana; and spend, use, and commit other resources and assets owned by the State of Indiana;

WHEREAS, centralized data sharing, correlation, and analysis capacity will enable the state to achieve efficiencies in the administration of state programs and

---

<sup>50</sup> [http://in.gov/gov/files/Executive\\_Order\\_14-06.pdf](http://in.gov/gov/files/Executive_Order_14-06.pdf)

services and to more efficiently address public health, public safety, and quality of life issues;

WHEREAS, with any data collection or use of data, state government must protect individual privacy, transparency of government operations, and public safety;

WHEREAS, the Office of Management and Budget (OMB), established under Ind. Code § 4-3-22-3, has been given the statutory authority under Ind. Code § 4-3-22-1 to devote adequate resources to:

- (1) Gather and coordinate data in a timely manner.
- (2) Perform comprehensive and detailed budgeting analysis.
- (3) Put in place comprehensive and effective budgeting practices.
- (4) Coordinate all functions related to budgeting and controlling spending in state government.
- (5) Perform comprehensive and detailed financial analysis.
- (6) Perform comprehensive financial oversight.
- (7) Ensure that effective financial management policies are implemented throughout state government.
- (8) Perform comprehensive and detailed performance analysis.
- (9) Ascertain whether the burdens imposed by laws and rules are justified by their benefits using a rigorous cost benefit analysis.
- (10) Measure the performance of government activities;

WHEREAS, Ind. Code § 4-3-22-15 provides that “all state agencies (as defined in Ind. Code § 4-12-1-2(d) shall, in addition to complying with all statutory duties applicable to state purchasing, be accountable to the OMB for adherence to policies, procedures, and spending controls established by the OMB and approved by the governor.”;

WHEREAS, the OMB is exercising this statutory responsibility through the creation of the Governor’s Management and Performance Hub (MPH).

NOW, THEREFORE, I Michael R. Pence, by virtue of the authority vested in me as the Governor of the State of Indiana, do hereby order that:

2. The OMB shall create MPH as a tool for continuous process improvement for the State of Indiana.

3. IOT shall as directed by OMB work with the agencies with respect to the MPH's data needs and technical requirements to IOT.
4. The OMB shall provide recommendation to the Governor on:
  - a. Opportunities to use data collected by state agencies to drive innovation and efficiency across state agencies;
  - b. Improvements to information technology systems, practices, and procedures to enhance the security of data retained by state agencies; and
  - c. Opportunities to enhance the transparency of state government.
5. The OMB and IOT shall collaborate with private and public sector industry experts to ensure the MPH utilizes the best practices in data analytics and security.
6. All state agencies as defined in Ind. Code Ind. Code § 4-12-1-2(d), shall participate in the MPH by providing data, information, system access, or other resources to IOT and OMB upon request.
7. To the extent data requested by OMB and IOT is maintained as confidential under state or federal law; all agencies shall identify the data as confidential. If the transmission of the data to OMB or IOT is specifically prohibited by state or federal law, agencies shall work with the OMB to identify if any edits, deletions, or additional protections can be made to comply with state and federal laws allowing data to be provided to the OMB. Agencies shall provide the data to OMB with plans and procedures for ensuring data shared with the OMB continues to be protected in accordance with such laws. Agencies shall coordinate with the OMB in the development of data sharing agreements and shall execute such agreements to facilitate OMB's receipt and the use of any sensitive data.

State Seal

IN TESTIMONY  
WHEREOF, I Michael R.  
Pence, have hereunto set  
my hand and caused to be  
affixed, the Great Seal of  
the State of Indiana on  
this seventeenth day of  
March, 2014.

Signature  
Michael R. Pence  
Governor of Indiana

ATTEST: Signature  
Connie Lawson  
Secretary of State



**A-2: North Carolina Big Data Legislation**

**GENERAL ASSEMBLY OF NORTH CAROLINA<sup>51</sup>  
SESSION 2011**

**SESSION LAW 2012-142  
HOUSE BILL 950**

**AN ACT TO MODIFY THE CURRENT OPERATIONS AND CAPITAL IMPROVEMENTS  
APPROPRIATIONS ACT OF 2011 AND FOR OTHER PURPOSES.**

The General Assembly of North Carolina enacts:

**PART I. INTRODUCTION AND TITLE OF ACT**

**INTRODUCTION**

**SECTION 1.1.** The appropriations made in this act are for maximum amounts necessary to provide the services and accomplish the purposes described in the budget. Savings shall be effected where the total amounts appropriated are not required to perform these services and accomplish these purposes and, except as allowed by the State Budget Act, or this act, the savings shall revert to the appropriate fund at the end of each fiscal year as provided in G.S. 143C-1-2(b).

**TITLE OF ACT**

**SECTION 1.2.** This act shall be known as "The Current Operations and Capital Improvements Appropriations Act of 2012."

**PART VI-A. INFORMATION TECHNOLOGY  
ENHANCE ENTERPRISE-LEVEL BUSINESS INTELLIGENCE TO INCREASE  
EFFICIENCY IN STATE GOVERNMENT**

**SECTION 6A.7A.(a)** Creation of Initiative. –

- (1) Creation. – The enterprise-level business intelligence initiative (initiative) is established in the Office of State Controller. The purpose of the initiative is to support the effective and efficient development of State agency business intelligence capability in a coordinated manner and reduce unnecessary information silos and technological barriers. The initiative is not intended to

---

<sup>51</sup> <http://www.ncleg.net/Sessions/2011/Bills/House/PDF/H950v7.pdf>

replace transactional systems, but is instead intended to leverage the data from those systems for enterprise-level State business intelligence.

The initiative shall include a comprehensive evaluation of existing data analytics projects and plans in order to identify data integration and business intelligence opportunities that will generate greater efficiencies in, and improved service delivery by, State agencies. The Office of State Controller may partner with current vendors and providers to assist in the initiative. However, to limit the cost to the State, the Office of the State Controller shall use current licensing agreements wherever feasible.

- (2) Application to State government. – The initiative shall include all State agencies, departments, and institutions, including The University of North Carolina.
- (3) Governance. – The State Controller shall lead the initiative established pursuant to this section. The Chief Justice of the North Carolina Supreme Court and the Legislative Services Commission each shall designate an officer or agency to advise and assist the State Controller with respect to implementation of the initiative in their respective branches of government. The judicial and legislative branches shall fully cooperate in the initiative mandated by this section in the same manner as is required of State agencies.

**SECTION 6A.7A.(b) Government Business Intelligence Competency Center. –**

- (1) GBICC established. – There is established in the Office of the State Controller the Government Business Intelligence Competency Center (GBICC). GBICC shall assume the work, purpose, and resources of the current data integration effort in the Office of the State Controller and shall otherwise advise and assist the State Controller in the management of the initiative. The State Controller shall make any organizational changes necessary to maximize the effectiveness and efficiency of GBICC.
- (2) Powers and duties of the GBICC. – The State Controller shall, through the GBICC, do all of the following:
  - a. Continue and coordinate ongoing enterprise data integration efforts, including:
    - 1. The deployment, support, technology improvements, and expansion for CJLEADS.
    - 2. The pilot and subsequent phase initiative for NC FACTS.
    - 3. Individual-level student data and workforce data from all levels of education and the State workforce.
    - 4. Other capabilities developed as part of the initiative.

- b. Identify technologies currently used in North Carolina that have the capability to support the initiative.
- c. Identify other technologies, especially those with unique capabilities, that could support the State's business intelligence effort.
- d. Compare capabilities and costs across State agencies.
- e. Ensure implementation is properly supported across State agencies.
- f. Ensure that data integration and sharing is performed in a manner that preserves data privacy and security in transferring, storing, and accessing data, as appropriate.
- g. Immediately seek any waivers and enter into any written agreements that may be required by State or federal law to effectuate data sharing and to carry out the purposes of this section.
- h. Coordinate data requirements and usage for State business intelligence applications in a manner that (i) limits impacts on participating State agencies as those agencies provide data and business knowledge expertise and (ii) assists in defining business rules so the data can be properly used.
- i. Recommend the most cost-effective and reliable long-term hosting solution for enterprise-level State business intelligence as well as data integration, notwithstanding Section 6A.2(f) of S.L. 2011-145.

**SECTION 6A.7A.(c)** Implementation of the Enterprise-Level Business Intelligence Initiative. –

- (1) Phases of the initiative. – The initiative shall commence no later than August 1, 2012, and shall be phased in accordance with this subsection. The initiative shall cycle through these phases on an ongoing basis:
  - a. Phase I requirements. – In the first phase, the State Controller through GBICC shall:
    1. Inventory existing State agency business intelligence projects, both completed and under development.
    2. Develop a plan of action that does all of the following:
      - I. Defines the program requirements, objectives, and end state of the initiative.
      - II. Prioritizes projects and stages of implementation in a detailed plan and benchmarked timeline.
      - III. Includes the effective coordination of all of the State's current data integration initiatives.

- IV. Utilizes a common approach that establishes standards for business intelligence initiatives for all State agencies and prevents the development of projects that do not meet the established standards.
  - V. Determines costs associated with the development effort and identifies potential sources of funding.
  - VI. Includes a privacy framework for business intelligence consisting of adequate access controls and end user security requirements.
  - VII. Estimates expected savings.
  3. Inventory existing external data sources that are purchased by State agencies to determine whether consolidation of licenses is appropriate for the enterprise.
  4. Determine whether current, ongoing projects support the enterprise-level objectives.
  5. Determine whether current applications are scalable, or are applicable for multiple State agencies, or both.
  - b. Phase II requirements. – In the second phase, the State Controller through the GBICC shall:
    1. Identify redundancies and determine which projects should be discontinued.
    2. Determine where gaps exist in current or potential capabilities.
  - c. Phase III requirements. – In the third phase:
    1. The State Controller through GBICC shall incorporate or consolidate existing projects, as appropriate.
    2. The State Controller shall, notwithstanding G.S. 147-33.76 or any rules adopted pursuant thereto, eliminate redundant business intelligence projects, applications, software, and licensing.
    3. State Controller through GBICC shall complete all necessary steps to ensure data integration in a manner that adequately protects privacy.
- (2) Commencement of projects. – Subject to the availability of funds, and subsequent to the submission of the written report required by sub-subdivision a. of subdivision (1) of subsection (e) of this section, the State Controller shall begin projects to carry out the purposes of this section no

later than November 1, 2012. The State Controller may also expand existing data integration or business intelligence contracts with current data integration efforts, as appropriate, in order to implement the plan required by this section in accordance with the schedule established and the priorities developed during Phase I of the initiative, and may use public-private partnerships as appropriate to implement the plan.

**SECTION 6A.7A.(d) Funding. –**

- (1) Allocation. – Of the funds appropriated from the General Fund to the General Assembly for the 2011-2013 fiscal biennium, the sum of five million dollars (\$5,000,000) shall be used to fund the initiative established by this section. The Office of the State Controller shall use up to seven hundred fifty thousand dollars (\$750,000) to cover the cost of administering the initiative.
- (2) Federal funds. – The Office of State Controller, with the support of the Office of State Budget and Management, shall identify and make all efforts to secure any matching funds or other resources to assist in funding this initiative.
- (3) Use of savings. – Savings resulting from the cancellation of projects, software, and licensing, as well as any other savings from the initiative, shall be returned to the General Fund and shall remain unexpended and unencumbered until appropriated by the General Assembly in a subsequent fiscal year. It is the intent of the General Assembly that expansion of the initiative in subsequent fiscal years be funded with these savings and that the General Assembly appropriate funds for projects in accordance with the priorities identified by the Office of the State Controller in Phase I of the initiative.

**SECTION 6A.7A.(e) Reporting. –**

- (1) Routine reports. – The Office of the State Controller shall submit and present the following reports:
  - a. By no later than October 1, 2012, a written report on the implementation of Phase I of the initiative and the plan developed as part of that phase to the Chairs of the House of Representatives Appropriations and Senate Base Budget/Appropriations Committees, to the Joint Legislative Oversight Committee on Information Technology, and to the Fiscal Research Division of the General Assembly. The State Controller shall submit this report prior to implementing any improvements, expending funding for expansion of existing business intelligence efforts, or establishing other projects as a result of its evaluations.
  - b. By February 1, 2013, and quarterly thereafter, a written report detailing progress on, and identifying any issues associated with, State business intelligence efforts.

- (2) Extraordinary reports. – The Office of the State Controller shall report the following information as needed:
- a. Any failure of a State agency to provide information requested pursuant to this section. The failure shall be reported to the Joint Legislative Committee on Information Technology and to the Chairs of the House of Representatives Appropriations and Senate Base Budget/Appropriations Committees.
  - b. Any additional information to the Joint Legislative Commission on Governmental Operations and the Joint Legislative Oversight Committee on Information Technology that is requested by those entities.

**SECTION 6A.7A.(f) Duties of State Agencies. –**

- (1) Duties of State agencies. – The head of each State agency shall do all of the following:
- a. Grant the Office of the State Controller access to all information required to develop and support State business intelligence applications pursuant to this section. The State Controller and the GBICC shall take all necessary actions and precautions, including training, certifications, background checks, and governance policy and procedure, to ensure the security, integrity, and privacy of the data in accordance with State and federal law and as may be required by contract.
  - b. Provide complete information on the State agency's information technology, operational, and security requirements.
  - c. Provide information on all of the State agency's information technology activities relevant to the State business intelligence effort.
  - d. Forecast the State agency's projected future business intelligence information technology needs and capabilities.
  - e. Ensure that the State agency's future information technology initiatives coordinate efforts with the GBICC to include planning and development of data interfaces to incorporate data into the initiative and to ensure the ability to leverage analytics capabilities.
  - f. Provide technical and business resources to participate in the initiative by providing, upon request and in a timely and responsive manner, complete and accurate data, business rules and policies, and support.

- g. Identify potential resources for deploying business intelligence in their respective State agencies and as part of the enterprise-level effort.
- h. Immediately seek any waivers and enter into any written agreements that may be required by State or federal law to effectuate data sharing and to carry out the purposes of this section, as appropriate.

**SECTION 6A.7A.(g) Miscellaneous Provisions. –**

- (1) Status with respect to certain information. – The State Controller and the GBICC shall be deemed to be all of the following for the purposes of this section:
  - a. With respect to criminal information, and to the extent allowed by federal law, a criminal justice agency (CJA), as defined under Criminal Justice Information Services (CJIS) Security Policy. The State CJIS Systems Agency (CSA) shall ensure that CJLEADS receives access to federal criminal information deemed to be essential in managing CJLEADS to support criminal justice professionals.
  - b. With respect to health information covered under the Health Insurance Portability and Accountability Act of 1996 (HIPAA), as amended, and to the extent allowed by federal law:
    - 1. A business associate with access to protected health information acting on behalf of the State's covered entities in support of data integration, analysis, and business intelligence.
    - 2. Authorized to access and view individually identifiable health information, provided that the access is essential to the enterprise fraud, waste, and improper payment detection program or required for future initiatives having specific definable need for the data.
    - c. Authorized to access all State and federal data, including revenue and labor information, deemed to be essential to the enterprise fraud, waste, and improper payment detection program or future initiatives having specific definable need for the data.
    - d. Authorized to develop agreements with the federal government to access data deemed to be essential to the enterprise fraud, waste, and improper payment detection program or future initiatives having specific definable need for such data.
- (2) Release of information. – The following limitations apply to (i) the release of information compiled as part of the initiative, (ii) data from State agencies

that is incorporated into the initiative, and (iii) data released as part of the implementation of the initiative:

- a. Information compiled as part of the initiative. – Notwithstanding the provisions of Chapter 132 of the General Statutes, information compiled by the State Controller and the GBICC related to the initiative may be released as a public record only if the State Controller, in that officer's sole discretion, finds that the release of information is in the best interest of the general public and is not in violation of law or contract.
- b. Data from State agencies. – Any data that is not classified as a public record under G.S. 132-1 shall not be deemed a public record when incorporated into the data resources comprising the initiative. To maintain confidentiality requirements attached to the information provided to the State Controller and GBICC, each source agency providing data shall be the sole custodian of the data for the purpose of any request for inspection or copies of the data under Chapter 132 of the General Statutes.
- c. Data released as part of implementation. – Information released to persons engaged in implementing the State's business intelligence strategy under this section that is used for purposes other than official State business is not a public record pursuant to Chapter 132 of the General Statutes.

**SECTION 6A.7A.(h)** G.S. 75-66(d) reads as rewritten:

"(d) Nothing in this section shall:

- (1) Limit the requirements or obligations under any other section of this Article, including, but not limited to, G.S. 75-62 and G.S. 75-65.
- (2) Apply to the collection, use, or release of personal information for a purpose permitted, authorized, or required by any federal, State, or local law, regulation, or ordinance.
- (3) Apply to data integration efforts to implement the State's business intelligence strategy as provided by law or under contract."

**STATE PRIVATE CLOUD**

**SECTION 6A.9.(a)** Findings. – The General Assembly finds that:

- (1) The wide distribution of information technology facilities across multiple locations causes infrastructure and operational inefficiencies.
- (2) Infrastructure as a service, also known as cloud computing, has the potential to increase efficiency and enhance operations by reducing

information technology costs and accelerating the provision of services.

- (3) The creation of a secure and flexible State private cloud is in the best interest of the people of this State.

**SECTION 6A.9.(b) Plan Required.** – The State Chief Information Officer shall create a plan for the development and implementation of a State-owned, State-hosted infrastructure as a service, or private cloud, project to be operated and managed by the State.

**SECTION 6A.9.(c) Components of the Plan.** – The State private cloud plan created pursuant to this section shall include:

- (1) Requirements for:
- a. The State to have complete control and ownership of all components of the private cloud, including hardware, software, network infrastructure, security, and data.
  - b. All components of the private cloud to be maintained at State-owned, State-operated facilities.
  - c. The private cloud to fully comply with all legislative, regulatory, policy, and security requirements that apply to State agencies and entities conducting business with the State.
  - d. The State's existing information technology infrastructure to be used to support the private cloud.
  - e. Documentation of any redundancy built into the infrastructure to support requirements for increased availability and disaster recovery.
  - f. A service-centric approach to computing resources. Users of computing resources shall be able to efficiently access powerful, predefined computing environments based on their requirements.
  - g. A self-service ability to provision and deprovision, as requested by users, while maintaining high levels of security.
  - h. A fully functional, efficient, fair system to bill State agencies for private cloud usage. This requirement includes mechanisms to capture usage data and enable chargeback integration within the billing system.
  - i. A plan to manage infrastructure resources that can be scaled in response to State agency requirements.
  - j. An inventory of all potential resources, both public and private, available to support the development, implementation,

operation, and management of the private cloud, and the costs and benefits associated with each.

- (2) A detailed timeline, documentation of agency requirements, identification and resolution of security issues, and an assessment of the impact on any ongoing projects or current applications.
- (3) Identification of costs associated with developing the private cloud.
- (4) Identification and documentation of private cloud management and monitoring tools to facilitate the maintenance of complete control of private cloud resources; automate provisioning, deprovisioning, and scheduling; and maintain system capacity.
- (5) Identification of ways to improve the private cloud's supporting infrastructure.
- (6) Identification of potential sources of savings to support development, implementation, and maintenance of the State private cloud.

**SECTION 6A.9.(d) Funding and Implementation.** – No funds from any source shall be used for the development and implementation of a private cloud without specific authorization by the General Assembly appropriating funds for this purpose.

**SECTION 6A.9.(e) Report.** – The State Chief Information Officer shall report the plan created pursuant to this section to the Joint Legislative Oversight Committee on Information Technology no later than January 1, 2013.

**SECTION 6A.9.(f) Access by Private Vendors.** – If the State Chief Information Officer provides to a potential vendor any information or access to State facilities in connection with or anticipation of the private cloud project described in this section, the State Chief Information Officer shall provide the same information or access to all potential vendors. The State Chief Information Officer shall certify the Officer's compliance with this subsection to the General Assembly.

**A-3: Michigan Big Data Legislation****EXECUTIVE DIRECTIVE<sup>52</sup>  
No. 2013- 1**

DATE: November 1, 2013

TO: All Executive Branch Departments and Agencies

FROM: Governor Rick Snyder (signed)

RE: Data and Information Sharing, Management and Governance

To continue the process of reinventing state government, we must improve upon the sharing and management of data across all executive branch agencies. Data and information are valued assets that require effective and secure management. It is my goal to establish an environment where improved sharing and management of data will enhance services to citizens. This can only be accomplished by establishing an Enterprise Information Management (EIM) program.

EIM will improve analysis and reporting for the state and it will make our operations more efficient. I envision a state government that allows a single sign-on for citizens and businesses to access all of their state account information. We must improve upon the data available on our Open Michigan website. By implementing EIM, the state can improve service delivery and transparency in a number of our priority areas, including public safety, education, healthcare and economic growth.

Section 1, Article 5 of the Michigan Constitution vests the executive power of the state of Michigan in the Governor. Section 8, Article 5 of the Michigan Constitution places each principal department under the supervision of the Governor. Pursuant to these provisions of the Michigan Constitution, I direct the following:

The Director of the Department of Technology, Management and Budget (DTMB) shall establish and implement an EIM program requiring participation and engagement by all Executive Branch departments and agencies to establish new and improved protocols for data and information sharing, management, and governance.

---

<sup>52</sup> [https://www.michigan.gov/documents/snyder/ED\\_2013-1\\_439597\\_7.pdf](https://www.michigan.gov/documents/snyder/ED_2013-1_439597_7.pdf)

The EIM program shall include a cross-agency data sharing protocol, a Michigan Information Management Governance Board, an information management implementation plan for each state department, and a five-year Michigan Statewide Data and Analytics Plan.

All state departments and agencies must work in partnership with DTMB to establish the procedures and protocols for cross-departmental and jurisdictional data sharing and processing. I would like to create a "share first" environment for data sharing while taking all possible measures to ensure personal privacy and protect personal information in a secure manner.

The Director of DTMB shall create and establish the Michigan Information Management Governance Board (MIMGB) as the primary governing body for the state EIM program, to be chaired by a representative from the Governor's Legal Counsel. The MIMGB will have membership representation from Directors or Chief Deputy Directors of all Executive Branch departments and agencies. The MIMGB responsibilities are to adopt, support, and provide advice regarding all activities related to achieving the goals of the EIM program.

Each Department Director shall create and establish a Department Information Management Governance Board (DIMGB) to provide an operational support structure for and to coordinate with the MIMGB. The DIMGB shall be chaired by the Department Director or Chief Deputy Director and will have membership representation from all Bureau and/or Division administrators that have responsibility over business data and information management systems. The DIMGB responsibilities are to advise, adopt, and support all activities related to achieving the goals of the EIM program within each respective department.

Each department shall establish a Chief Data Steward responsible for establishing and implementing EIM within the department. The Chief Data Steward will provide administrative support to the chair of the DIMGB, and serve on working group(s) of the MIMGB. The Chief Data Steward shall not serve as the representative on the MIMGB.

The MIMGB shall direct the development of the Michigan Statewide Information and Analytics Plan, focused on long-term statewide information management and analytics goals. The plan shall include the establishment of a centralized information management and analytics service center and be fully integrated with state agency plans and with DTMB's Information and Communication Technology (ICT) Assessment Roadmap. The plan shall also incorporate an EIM strategy for successful crossboundary collaboration with external partners of state departments and agencies.

The process of data governance in the state will be open, transparent, timely, and will require cooperation and trust. My expectation is that all state departments and agencies will work together with DTMB to ensure that the EIM program is successful. Citizens and other stakeholders deserve the improvements that can be achieved from an effective EIM program, whereby data is effectively governed and managed.

cc: Department Directors and State Agency Heads

**A-4: New Jersey Legislation****CHAPTER 33<sup>53</sup>**

**AN ACT** designating the New Jersey Big Data Alliance as the State’s advanced cyberinfrastructure consortium, and supplementing Title 52 of the Revised Statutes.

**BE IT ENACTED** by the Senate and General Assembly of the State of New Jersey:

C.52:17C-3.4 New Jersey Big Data Alliance designated as State’s advanced cyberinfrastructure consortium; definitions.

1. a. The New Jersey Big Data Alliance is designated as the State’s advanced cyberinfrastructure consortium. The purpose of the consortium shall be to encourage State government, academia, and industry to address, in a strategic and coordinated manner, the significant and immediate challenges posed by the proliferation of big data sources and the resultant deluge of digital data. Major initiatives may include, but not be limited to: (1) encouraging the creation of joint education programs, including the establishment of a common curriculum for data sciences and the creation of coordinated certificates, workforce training, and outreach programs; (2) promoting inter-university research collaborations; (3) catalyzing interaction with national and international data consortiums, such as the National Consortium for Data Science; (4) organizing events that promote big data education and collaborating across State government, academia, and industry; (5) collaborating with the Rutgers Discovery Informatics Institute and the Office of Information Technology to develop an advanced cyberinfrastructure plan for the State; and (6) developing a shared data cloud that integrates data infrastructure, hosted data, and data analytics.

b. The New Jersey Big Data Alliance shall consist of the following members: Rutgers, the State University; Princeton University; New Jersey Institute of Technology; Rowan University; the Richard Stockton College of New Jersey; Kean University; Montclair State University; and the Stevens Institute of Technology. The New Jersey Big Data Alliance shall determine the appropriate size of its membership and admit future members as the alliance deems appropriate.

c. As used in this section:

“Big data” means high volume information assets, high velocity information assets, high variety information assets, or all three, that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.

---

<sup>53</sup> [http://www.njleg.state.nj.us/2014/Bills/PL14/33\\_.PDF](http://www.njleg.state.nj.us/2014/Bills/PL14/33_.PDF)

The term “cyberinfrastructure” includes, but is not limited to, data networks, computational facilities, computing resources, large data sets, specialized software applications, information technology usage improvements, and the human expertise necessary to develop and manage these resources.

2. This act shall take effect immediately.

Approved August 15, 2014.

## Appendix B: FIPPS Privacy Principles, Risks, and Mitigation Strategies<sup>54</sup>

| Principle                                    | Risk  | Possible Mitigation Approach   |
|--|---|--|
| <b>Principle of Transparency</b>             | Individuals may not be aware their PII is being compared against other information in a big data project                              | <ul style="list-style-type: none"> <li>• Ensure that System of Records Notices provide notice to the public that the information may be compared against other data sets and be subject to analysis</li> <li>• Pursue ways to provide transparency outside of the traditional privacy documentation process because of the privacy sensitivities surrounding big data technology and use</li> </ul>  |
| <b>Principle of Individual Participation</b> | Individual will not be able to receive appropriate access, correction, and redress regarding use of PII                               | <ul style="list-style-type: none"> <li>• Provide for the ability to refresh the data that is ingested into the Data Repository</li> <li>• Develop a process to provide an individual with the same access and redress opportunities in the Data Repository that he or she would have in the original IT system</li> <li>• Data governance structure and process</li> </ul>   |
|  | Changes made to PII in the underlying IT system as a result of correction and redress will not be replicated into the Data Repository | <ul style="list-style-type: none"> <li>• Establish a process to refresh the data provided from the original IT system to the Data Repository with refresh timelines based on operational need, available resources, and technical capabilities</li> <li>• In cases where action or decisions impacting individuals will result from use of the PII, establish a process to verify the data accuracy with the original IT system</li> </ul> |

<sup>54</sup> Source: Privacy Impact Assessment Update DHS Data Framework, DHS/ALL/PIA-046(a), August 29, 2014 pages

|  |   |  |
|--|---|--|
| <p><b>Principle of Purpose Specification</b></p> | <p>Data will be included in the Data Repository and Analytics efforts for a purpose other than the purpose for which is was collected in the original IT system</p> | <ul style="list-style-type: none"> <li>•</li> </ul>  |
| <p><b>Principle of Data Minimization</b></p>     | <p>More data sets will be included in the Data Repository than those which is necessary to fulfill the purposes authorized</p>                                      | <ul style="list-style-type: none"> <li>• Evaluate each data set to determine whether its use is directly relevant and necessary to accomplish the purposes authorized</li> </ul>   |
|  | <p>Analytics program will encourage replication of data sets across the Enterprise, proliferating data</p>  | <ul style="list-style-type: none"> <li>• Establish a goal of the Data Repository to reduce the number of copies of data sets across the Enterprise.</li> <li>• An Enterprise-wide big data solution, will actually reduce the number of copies of data sets in the long-term.</li> <li>• Eventually, some data aggregation systems may be decommissioned as their capabilities are replicated and centralized within the Data Repository.</li> <li>• Data Repository must successfully replicate the capabilities of other systems and build operator support</li> </ul> |
|  | <p>Data will be retained in the Data Repository for longer than is allowed in the original IT system</p>  | <ul style="list-style-type: none"> <li>• Establish policy that the retention period for the original IT system will also apply when that information is ingested into the Data Repository.</li> <li>• Tag data with the time that it was ingested into the original IT systems so that the information can be deleted when the retention period ends</li> </ul>  |

|  |  |   |
|--|--|---|
| <b>Principle of Use Limitation</b>             | Data Repository/Analytics users will access more PII than is necessary to accomplish their specified purpose   | <ul style="list-style-type: none"> <li>• Establish policy to restrict access to PII within a particular data set based on the user’s specified purpose.</li> <li>• Tag elements from each data set as belonging to one of three categories—core biographic, extended biographic, and encounter information—and users are only able to access the categories that are necessary to perform their function.</li> <li>• Minimize data access according to specified purpose</li> </ul> |
|  | Users will use the data for purposes other than those authorized   | <ul style="list-style-type: none"> <li>• Establish policy-based controls to ensure that a user is only able to access information that is permitted for a particular purpose and function</li> </ul>  |
|  | Elements of data access and control may be insufficiently developed or incorrectly implemented and will fail to limit the use of the data to the purposes authorized | <ul style="list-style-type: none"> <li>•</li> </ul>   |
|  | Enterprise will share PII for a purpose that is not compatible with the purpose for which the PII was collected  | <ul style="list-style-type: none"> <li>•</li> </ul>   |
| <b>Principle of Data Quality and Integrity</b> | PII transferred outside of the original IT system and into the Data Repository will not be accurate, relevant, timely, or complete                                   | <ul style="list-style-type: none"> <li>• Establish a process to refresh the data provided from the original IT system to the Data Repository, so that updates or corrections are replicated from the original IT system into the Repository</li> <li>• Provide training to users to understand the risk associated with data latency (due to limited refresh capabilities)</li> <li>• Establish policy and process for users to verify</li> </ul>                                   |

|   |   |   |
|---|---|---|
|   |   | information at the source system before completing any final analysis or using the information operationally  |
| <b>Principle of Security</b>                    | Data Repository and Analytics infrastructure systems will not have appropriate security safeguards                                  | <ul style="list-style-type: none"> <li>● Follow the requirements for information assurance and security and the development of sensitive systems and handling of sensitive information</li> <li>● Require that the Data Repository and Analytics system have system security plans and the Chief Information Security Officer’s approval for Authority to Operate</li> <li>● Require that information will be encrypted and safeguarded during transport and storage</li> <li>● Limit access to pre-approved users whose access to data, data sets and query tools will be determined based on their authenticated attributes and their predetermined functions and purposes</li> </ul> |
| <b>Principle of Accountability and Auditing</b> | Use of PII will not be auditable to demonstrate compliance with these principles and all applicable privacy protection requirements | <ul style="list-style-type: none"> <li>● Establish requirements and policy to ensure that the Data Repository and Analytics system incorporate audit capabilities adequate to support an audit of whether PII was accessed properly and that the dynamic access controls could sufficiently limit the data that is viewed to the users who are permitted to view it</li> <li>● Audit logs should contain the user name and the query performed, but not the responses provided back</li> </ul>  |
|   | Enterprise will not perform reviews of the audit logs to determine compliance with policies   | <ul style="list-style-type: none"> <li>●</li> </ul>   |

## Appendix C: Big Data Glossary<sup>55</sup>

This glossary is intended to be an authoritative explanation of the meaning of technical terms, for all users of data.gov.uk. Users are encouraged to improve it by suggesting a better way of explaining the definitions, and by adding new definitions.

A

### AGGREGATED DATA

A combination of unit records created with the objective that individual details are not disclosed.

### ANONYMISATION

The process of adapting data so that individual people or businesses cannot be identified.

### APPLICATION PROGRAMMING INTERFACE (API)

A specification intended to be used as an interface by software components to communicate with each other. An API may include specifications for routines, data structures, object classes, and variables.

### ATTRIBUTION LICENCE

A license that requires that the original source of the licensed material is cited (attributed).

### AUTHORITATIVE

Able to be trusted as being accurate or true; reliable: e.g. "clear, authoritative information".

### AUTHORITATIVE DATA SOURCE

A recognized or official data production source with a designated mission statement or source/product to publish reliable and accurate data for subsequent use by customers. An authoritative data source may be the functional combination of multiple, separate data sources.

B

### BIG DATA

A loose term, not formally defined, for high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing, that can give enhanced insight and decision making.

### BIG DATA ANALYTICS

The process of examining and interrogating big data assets to derive insights of value for decision making.

---

<sup>55</sup> <https://data.gov.uk/glossary>

## C

**COMMERCIAL USE/RE-USE**

Use that is intended for or directed toward commercial advantage or private monetary compensation. For the purposes of the UK Government Licensing Framework, 'private monetary compensation' does not include the exchange of the **Information** for other copyrighted works by means of digital file-sharing or otherwise provided there is no payment of any monetary compensation in connection with the exchange of the **Information**.

**COMPILED DATABASE RIGHT**

The legal protection provided by EC and UK law to a collection of databases (which have been compiled from a number of different sources and normalised to facilitate cross searching).

**CONTENT**

The collection of information stored for a purpose in a file, folder or electronic message

**COPYRIGHT**

The protection given to literary works (including books and articles, and also databases and computer programs) which are recorded in writing. Electronic works may be recorded in analogue or digital form. Databases are protected by copyright where the selection and arrangement of the contents of the database are the author's own intellectual creation. Other aspects of a database may be protected by database right.

**CORE-REFERENCE DATA**

**Authoritative** or definitive data necessary to use other information, produced by the public sector as a service in itself due to its high importance and value. Usually including a field that may be used as a database key, or locational coordinates that may not be changed.

**COSTS - FIXED**

Costs which do not vary with the level of activity in the short run.

**COSTS - FULL**

The total cost of all the resources used in providing a good or service in any accounting period (usually one year). This will include all direct and indirect costs of producing the output (both cash and non-cash costs), including a full proportional share of overhead costs and any selling and distribution costs, insurance, depreciation, and the cost of capital, and any selling and distribution costs, insurance, depreciation, and the cost of capital, including any appropriate adjustment for expected cost increases.

**COSTS - MARGINAL**

The incremental cost of providing one further unit of a good or service.

**CREATIVE COMMONS**

A non-profit US organisation that enables the sharing and use of creativity and knowledge through free legal tools.

**CROWN COPYRIGHT**

Crown copyright covers material created by civil servants, ministers and government departments and agencies. It is legally defined under section 163 of the **Copyright, Designs and Patents Act 1988** as works made by officers or services of the Crown in the course of their duties. **Copyright** can also come into Crown ownership by means of an assignment or transfer of the copyright from the legal owner of the copyright to the Crown.

**D****DATA (CAN BE SINGULAR OR PLURAL IN COMMON USAGE)**

The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media. (The terms data, information and knowledge are frequently used for overlapping concepts. The main difference is in the level of abstraction being considered. Data is a broad term, embracing others, but is often the lowest level of abstraction, information is the next level and, finally, knowledge is the highest level.) See Raw data, **Derived data**, **Metadata**.

**DATA DISCOVERY**

The process of finding out what data exists and how it can be accessed.

**DATA SHARING**

The transfer, by agreement, of data collected for a specific purpose between two or more parties

**DATABASE RIGHTS**

An intellectual property right which applies to databases defined by the **Copyright and Rights in Databases Regulations 1997** as 'a collection of independent works or materials arranged in a systematic or methodical way and that are individually accessible by electronic or other means'. **Database rights** apply only to the collection of works, not to the individual works contained within it. Database right protection lasts for 15 years from when the database was completed but the 15 year period will restart if the database is altered significantly.

**DATASET**

A collection of data, usually presented in tabular form, presented either electronically or in other formats.

**DE-ANONYMISATION**

The technical process of attempting to determine the identity of a person or individual to whom a pseudonymised dataset relates.

**DEFINITIVE**

Of recognised authority or excellence

**DELEGATIONS OF AUTHORITY**

Authority granted by the Controller of Her Majesty's Stationery Office to Crown bodies

enabling them to license the re-use of information which they produce. Crown bodies with complete delegations to license information include trading funds, however some departments have partial delegations to license the use of particular information. All Crown bodies with delegations of authority are subject to the supervision of the Information Fair Trader Scheme.

#### DERIVED DATA

A data element or dataset adapted from other data sources using a mathematical, logical, or other type of transformation, e.g. arithmetic formula, composition, aggregation. See Value-added data.

#### DIGITAL RIGHTS MANAGEMENT

A class of access control technologies that are used by hardware manufacturers, publishers, copyright holders and individuals with the intent to limit the use of digital content and devices after sale.

#### DISCLOSIVE

Data is potentially disclosive if, despite the removal of obvious identifiers, characteristics of this dataset in isolation or in conjunction with other datasets might lead to identification of the individual to whom a record belongs.

#### DOCUMENT

Any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording).

### F

#### FREE AT POINT OF USE

Where there is no charge or fee to the end-user for the use or re-use of information.

#### FREEMIUM

A business model by which a product or service (typically a digital offering such as software, media, games or web services) is provided free of charge, but a premium is charged for advanced features or functionality.

### G

#### GEOSPATIAL DATA

Also known as spatial data or geographic information, it is the data that represents the geographic location of natural and man-made features on Earth. Spatial data is usually stored as coordinates of points, lines and areas and may include their topological relationship and attributes.

### I

#### INFORMATION

Interpretation and analysis of data that when presented in context represents added value, message or meaning. See Data.

**INFORMATION ASSET REGISTERS (IAR)**

Registers specifically set up to capture and organise metadata about the vast quantities of information held by government departments and agencies. A comprehensive IAR includes databases, old sets of files, recent electronic files, collections of statistics, research and so forth.

**INFORMATION FAIR TRADER SCHEME (IFTS)**

A scheme to set and assess standards for public sector bodies in allowing the re-use of their information. Any public sector body may apply to become IFTS accredited. However, all Crown bodies that hold a delegation of authority from the Controller of HMSO must become IFTS accredited. IFTS measures members' performance against the six principles of maximisation, simplicity, transparency, fairness, challenge and innovation. It considers both the commercial re-use of public sector information and non-commercial citizen access to information.

**INFORMATION PROVIDER**

The person, creator or organisation providing the information for re-use under the Open Government Licence or the **Non-Commercial Government Licence**.

**INTELLECTUAL PROPERTY (RIGHTS)**

A set of property rights that grant the right to protect the created materials. Intellectual property rights comprise trade marks, patents, registered designs copyright and database rights.

**L****LICENCE (NOUN)**

A legal document giving permission to use information

**LICENSE (VERB)**

The act of giving a formal licence (usually written) authorisation.

**LINKED DATA**

The technical term used to describe the best practice of exposing, sharing and connecting items of data on the semantic web using unique resource identifiers (URIs) and resource description framework (RDF). Not to be confused with data linking.

**M****METADATA**

Data that describes or defines other data. Anything that users need to know to make proper and correct use of the real data, in terms of reading, processing, interpreting, analysing and presenting the information. Thus metadata includes file descriptions, codebooks, processing details, sample designs, fieldwork reports, conceptual motivations, etc., in other words, anything that might influence the way in which the information is used.

**MODELLED DATA**

**Information** created by mathematical representation of data relationships; sometimes used to simulate environments that are difficult to observe reliably or consistently.

#### **MOSAIC/JIGSAW EFFECT**

The technical process of combining anonymised data with auxiliary data in order to attempt to reconstruct identifiers linking data to the individual it relates to.

## **N**

#### **NON-COMMERCIAL GOVERNMENT LICENCE SEARCH FOR TERM**

The **Non-Commercial Government Licence** offers a legal solution to enable the provision and use of public sector information under a common set of terms and conditions at no charge for Non-Commercial use only. It enables any public sector information holder to make their information available for use and re-use under its terms. The main requirement for re-users is to attribute the information provider and source.

#### **NON-COMMERCIAL USE**

Use that is not intended for or directed toward commercial advantage or private monetary compensation. For the purposes of the UK Government Licensing Framework, 'private monetary compensation' does not include the exchange of the **Information** for other copyrighted works by means of digital file-sharing or otherwise provided there is no payment of any monetary compensation in connection with the exchange of the **Information**.

## **O**

#### **ONTOLOGY**

Formal representation of knowledge as a set of concepts within a domain, and the relationships among those concepts.

#### **OPEN ACCESS (ACADEMIC)**

Provision of free access to peer-reviewed academic publications.

#### **OPEN DATA**

Data is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness.

#### **OPEN GOVERNMENT LICENCE (OGL)**

The Open Government Licence offers a legal solution to enable the provision and use of public sector information under a common set of terms and conditions. It enables any public sector information holder to make their information available for use and re-use under its terms. The main requirement for re-users is to attribute the **Information** Provider and source.

## **P**

#### **PERSONAL DATA**

Data which relate to a living individual who can be identified – (a) from those data, or (b)

from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller, and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual.

#### **PSEUDONYMISED DATA**

Data relating to a specific individual where the identifiers have been replaced by artificial identifiers to prevent identification of the individual.

#### **PUBLIC DOMAIN**

Works that are publicly available and in which the intellectual property rights have expired or been waived

#### **PUBLIC SECTOR BODIES**

State, regional or local authorities, bodies governed by public law and associations formed by one or several such authorities or one or several such bodies governed by public law.

#### **PUBLIC SECTOR INFORMATION (PSI)**

The wide range of information that public sector bodies collect, produce, reproduce and disseminate in many areas of activity while accomplishing their Public Task.

## **R**

#### **RAW DATA**

In the context of PSI, raw data is data collected which has not been subjected to processing or any other manipulation beyond that necessary for its first use. Raw data, i.e. unprocessed data, is a relative term; data processing commonly occurs by stages, and the 'processed data' from one stage may be considered the 'raw data' of the next.

#### **RE-USE (NOUN/VERB)**

The use by persons or legal entities of documents held by public sector bodies, for commercial or non-commercial purposes other than the initial purpose within the public task for which the documents were produced. Exchange of documents between public sector bodies purely in pursuit of their public tasks does not constitute re-use.

#### **RESOURCE DESCRIPTION FRAMEWORK (RDF)**

RDF, a W3C standard, is the foundation of several technologies for modelling distributed knowledge and is meant to be used as the basis of the **Semantic Web**

## **S**

#### **SAMPLE OF ANONYMISED RECORDS (SARS)**

A set of unit records available for research where key information has been removed to ensure anonymity. (Specifically Census SARS)

#### **SEMANTIC WEB**

A web of data that can be processed directly and indirectly by machines, providing a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is based on the Resource Description Framework (RDF).

#### SHARE-ALIKE LICENCE

The **Creative Commons** Attribution Share-Alike license allows re-distribution and re-use of a licensed work on the conditions that the creator is appropriately credited and that any derivative work is made available under “the same, similar or a compatible license”.

#### STAR RATING

In UK Linked Data, a system of ranking data sources that indicates ease of machine readability. It is not a measure of the quality of the data content.

#### SYNTHETIC POPULATION

A particular application of simulated data that generates a complete micro-view of individuals in a population.

### T

#### TAXONOMY

The science or technique of classification.

#### THIRD PARTY RIGHTS

**Information**, the rights for which are not owned by the **Information** Provider or Licensor.

### U

#### UNIFORM RESOURCE IDENTIFIER (URI)

The generic term for all types of names and addresses that refer to objects on the World Wide Web. A URL is one kind of URI.

#### UNIFORM RESOURCE LOCATOR (URL)

A type of URI that identifies a resource via a representation of its network location

#### UNIT RECORDS

Individual items of information from surveys or observations that often contain confidential details.

### V

#### VALUE-ADDED INFORMATION (OR DATA) SEARCH FOR TERM

Data to which value has been added to enhance and facilitate its use and effectiveness by or for users. See **Derived data**.

## Appendix D: Acronym List

| Acronym | Definition  |
|---------|---|
| AP      | Authority and Purpose   |
| AR      | Accountability, Audit, and Risk Management                                |
| CJLEADS | Criminal Justice Law Enforcement Automated Data Services                  |
| COBIT   | Control Objectives for Information and Related Technology                 |
| COPPA   | Children's Online Privacy Protection Act of 1998                          |
| DHS     | Department of Human Services (UT)<br>Department of Homeland Security (US) |
| DI      | Data Quality and Integrity  |
| DM      | Data Minimization and Retention   |
| DOC     | Department of Corrections   |
| DOH     | Department of Health  |
| DTMB    | Department of Technology, Management and Budget                           |
| DTS     | Department of Technology Services   |
| DWS     | Department of Workforce Services  |
| EDW     | Enterprise Data Warehouses  |
| EIM     | Enterprise Information Management   |
| ePHI    | electronic protected health information                                   |
| FedRAMP | Federal Risk and Authorization Program                                    |
| FERPA   | Family Educational Rights and Privacy Act of 1974                         |
| FIPPS   | Fair Information Practice Principles                                      |
| FIPS    | Federal Information Processing Standards                                  |
| FISCAM  | Federal Information System Controls Audit Manual                          |
| FISMA   | Federal Information Security Act  |
| GAO     | Government Accountability Office  |
| GBICC   | Government Business Intelligence Competency Center                        |
| GCN     | Government Computer News  |
| GDAC    | Government Data Analytics Center  |
| HIPAA   | Health Insurance Portability and Accountability Act                       |
| HITECH  | Health Information Technology for Economic and Clinical Health            |
| HITRUST | Health Information Trust Alliance   |
| ICT     | Information and Communication Technology                                  |
| IEC     | International Electrotechnical Commission                                 |
| IP      | Individual Participation and Redress                                      |

| Acronym | Definition   |
|---------|--|
| IoT     | Internet of Things   |
| IOT     | Indiana Office of Technology   |
| ISO     | International Organization for Standardization                           |
| IT      | Information Technology   |
| MIMGB   | Michigan Information Management Governance Board                         |
| MPH     | Management and Performance Hub   |
| NCFACFS | North Carolina Financial Accountability and Compliance Technology System |
| NBDRA   | NIST Big Data Reference Architecture                                     |
| NIST    | National Institute for Standards and Technology                          |
| OMB     | Office of Management and Budget  |
| OSTP    | Office for Science and Technology Policy                                 |
| PCI     | Payment Card Industry  |
| PHI     | Public Health Information  |
| PIA     | Privacy Impact Assessment  |
| PII     | Personally Identifiable Information                                      |
| PL/SQL  | Procedural Language extension to SQL                                     |
| SE      | Security   |
| SORN    | System of Record Notice  |
| SOW     | Statement of Work  |
| SOX     | Sarbanes–Oxley Act of 2002   |
| SP      | Special Publication  |
| SQL     |  |
| TR      | Transparency   |
| UDOT    | Utah Department of Transportation  |
| UL      | Use Limitation   |
|         |  |
|         |  |
|         |  |
|         |  |
|         |  |
|         |  |
|         |  |
|         |  |
|         |  |

## Appendix E 17 Steps to Implement a Public Sector Big Data Project<sup>56</sup>

### Stage 1: Planning Your Big Data Project

Big data projects are complex undertakings at best. This holds especially true in the public sector, where such projects often require large infrastructure changes, program designs and agreements across agencies and departments. The first and biggest stage of any big data initiative is planning your project. Attention to detail can make or break your project before it even begins.

"The planning phase includes conceptualization of the project, which is vital for establishing a platform for success and ensuring that stakeholders are properly informed," Desouza says. "This is an opportunity to lay the foundation of a quality project."

#### Step 1: Do Your Homework Before Undertaking a Big Data Project

First, determine what big data can and cannot do for your organization. You need to learn how big data can benefit your organization and what the risks and challenges are. **Think through the complexities of governance and policies in place around data, processes and systems — especially if you have outdated policies that don't account for current technologies. Understanding the policies of other agencies to identify shared constituents and minimize duplication of effort can pay big dividends.**

"Dig into examples and look at what has worked and what has not and even contact individuals who have been featured in press stories," Desouza says. "If CIOs do not have time to do their homework on big data, they should probably not commission a big data project."

#### Step 2: Build a Coalition to Support Your Big Data Project

Articles and whitepapers rarely talk about big data project failures. Professional networks are essential to getting that information. Talk to peers at other agencies, academic institutions, think tanks and the private sector.

Build an advisory group within your organization to both extend your influence while also helping you place big data within the context of your working environment.

---

<sup>56</sup> <http://www.cio.com/article/2368491/big-data/144854-17-Steps-to-Implement-a-Public-Sector-Big-Data-Project.html>

"Coalitions can go a long way in furthering agendas and creating inroads to new partnerships or information," Desouza says. "CIOs will need to perfect their 'elevator pitch' for big data to engage people in a coalition. The elevator pitch should explain how an investment in data management will allow the agency to tackle an existing problem more effectively and efficiently or take advantage of a new opportunity."

### **Step 3: Define the Broader Opportunity Your Big Data Project Presents**

For your first big data project, focus on something that directly benefits citizens and stakeholders. This will draw attention and critical thought. Get down to specifics later; for now, it's about the broad opportunity.

"Keeping the opportunity broad at the start allows CIOs more flexibility to engage other stakeholders and give them an opportunity to shape the program," Desouza says. "A common strategy employed by CIOs is to outline the broad opportunity in the form of a working paper or position paper. This paper looks at the opportunities that exist within an agency for superior data management. The working paper then becomes the platform for having strategic discussions and deliberations."

### **Step 4: Start with the Lowest-Hanging Fruit**

The best place to begin is with the easiest opportunities. Begin a project by tackling public data rather than getting involved with private data. Modernize existing technologies and processes for efficiency before creating new processes.

"CIOs that have witnessed success with their big data efforts note that they began by addressing problems that were simple, yet were visible pain points for an agency," Desouza says. "Choosing the visible pain points and building a data-driven solution helps win support for the overall program."

This is also a time to build a map of data elements and their interconnections. These maps can help you uncover data dependencies, interactions among data elements and organizational and political elements.

### **Step 5: Ensure Strategic Alignment of the Big Data Project**

Build alignment between your big data project and other organizational efforts or risk having your project perceived as a distraction from core efforts that pulls away valuable resources. One way is to embed phases of a big data project into existing IT efforts. For instance, weave data governance issues into every IT project.

It is also essential to line up a sponsor from senior management.

"These projects need a sponsor — someone who is willing to champion the project during moments of controversy or discomfort," Desouza says. "It is important that someone with clout is willing to weather the proverbial storms that often accompany the initiation of big data efforts."

### **Step 6: Become a Privacy and Security Advocate**

**It is easy to overlook or even undermine privacy with a big data project. CIOs need to adopt the role of privacy advocates when undertaking these projects, especially since existing privacy laws may require updating as a result of new technologies.**

"CIOs should be acutely aware of privacy and security considerations as discussions on data are taking place," Desouza says. "This will be critical for project success. Ultimately, if CIOs are aware of these issues and advocate for care in their handling, this will be reflected positively in how the project proceeds and is perceived by stakeholders."

**Privacy and ethical considerations around data collection, integration, analysis and dissemination should be discussed openly and sincerely. Seeking clarity from legal counsel is essential."**

### **Step 7: Use Taskforces to Implement Your Big Data Project**

Build a taskforce with both technical and organizational expertise to oversee the project. Ideally the taskforce will include representatives from the IT team who understand the technology, representatives from the business side who perform the tasks that generate or use the data being managed, and representatives who understand the legal and governance restrictions on the data in question.

"Each of these perspectives is valuable and must be included so as to ensure that the big data project does not run into any major surprises," Desouza says. "One of the critical roles to assign to the taskforce is that of the spokesperson. Ideally, there should be one individual to give regular updates to stakeholders and keep the senior sponsor apprised of any issues."

### **Step 8: Outline Expected Resistance and Plan for It**

Expect resistance from parts of your organization. The best way to overcome it is to determine the likely sticky areas ahead of time.

"One CIO interviewed for this study notes that his city's open access program caused internal strife because it gave city employees access to other city employees' information, resulting in discomfort throughout the organization," Desouza says. "There will be political repercussions for analyzing data that was never looked at before. This is

especially true if the big data project has anything to do with increasing efficiency of operations. Outlining the various sources and types of resistance upfront can help CIOs build an educated campaign and pitch for the project."

### **Step 9: Develop Key Performance Indicators for Your Big Data Project**

You need to develop key performance indicators (KPIs) around your big data project that focus on both process and outcome measures. Process measures are about improving efficiency; they capture gains in quicker completion times, lower costs of operations and so on. Outcome measures are about customers' perception of the service; these measures include improved customer service, increased customer value and so on.

"Baseline data on organizational processes should be captured before the project begins. This will allow meaningful comparisons of outcomes, both before and after project commencement," Desouza says. "Performance indicators should make sense to the business units involved and offer information on what the unit actually needs (not useless esoteric measures)."

### **Step 10: Design a Risk Mitigation Plan**

**Public sector databases contain citizens' data, making them valuable targets. You must assess the potential impact of compromised data and develop a risk mitigation plan with processes for reducing the risks.**

"It is important to **consider who has access to data, how much sensitive information is returned when database queries are made and what the physical security surrounding server rooms is,**" Desouza says.

He notes that you should also **develop a communications plan alongside the risk mitigation plan to ensure that messages are accurate and advance the goals of your agency or program. The communications plan should include dealing with press, academia and other agencies.**

### **Stage 2: Executing Your Big Data Project**

With the planning stage complete, it's time to put the gears in motion. The effectiveness of your planning in the previous stage will play a big role in your success, but good project management at this stage is equally important.

"Executing a big data project requires ongoing attention from the project's advisory group and the staff managing the project," Desouza says. "Learning and establishing best practices for project management is important. Organizational proficiencies or inefficiencies can bring about the success or failure of the project."

**Step 11: Constantly Gauge the Pulse of Your Big Data Initiative**

There's no way around it: CIOs need to consistently monitor the project status to get in front of major problems and allow for the development of creative solutions. Desouza says many CIOs use formal or informal dashboards for their projects that leverage the KPIs developed in the planning stage.

"CIOs say they need to regularly check the pulse of the program both from a process and outcome perspective," Desouza says. "In addition, they need to constantly gauge the conditions in the environment, especially in terms of any sentiment toward the project. Appropriate and timely communications, along with other interventions, can help address the issues and nip potential problems in the bud."

**Step 12: Communicate, Communicate, Communicate**

Communication is vital at every step of your big data project. **Success in big data requires breaking down silos of data and information, and that makes sharing the information you have essential.**

"Communication about milestones, inefficiencies, successes and failures will help an agency and peers gain a better understanding of big data," Desouza says. "In instances where data is shared between agencies, constant coordination, communication and feedback is necessary to ensure mission success."

**Step 13: Manage Scope Creep in Your Big Data Project**

Executing a big data project is difficult enough without the inevitable scope creep that comes as stakeholders see progress and think of additional ways to use the data. You don't want to continuously adjust the project plan and deliverables.

"It is critical that CIOs keep a watchful eye for scope creep and be clear on the boundaries of the current effort and how future revisions and additions will be made," Desouza says. "One approach might be to take the model that Google follows and release products in beta." By doing so, you can capture new ideas and work them into the next release or update.

**Step 14: Stay Focused on the Data, Not the Technologies**

Big data technologies are evolving at an exceptional pace. But your project may not need all-new technology. Maybe you can repurpose existing technology assets. **Stay focused on managing your data. With a clear view of data management issues from an organizational and policy viewpoint, it should be relatively easy to choose the appropriate technology.**

"Multiple CIOs report that the minute their agency announces an effort on big data (or any other major data activity), they are bombarded with calls from sales consultants who inquire and try to sell products and services," Desouza says. "Having a clear focus on the goal of the project, which is to leverage data and manage it more effectively toward a business outcome, helps keep everyone focused."

### **Step 15: If Necessary, Pull the Plug on Your Big Data Project**

Sometimes big data projects fail. But it can be hard to pull the plug, given sunk costs. In these situations, failure to call it quits will not only make the situation worse, it could hurt the state of your entire IT department.

"One strategy suggested by a CIO is to **outline clearly at the project's beginning the conditions under which the project would be stopped**," Desouza says. "Thinking through these upfront not only helps in setting realistic expectations of the project, but also will sensitize the team to look for signals of trouble and discuss them openly during the team meetings."

### **Stage 3: Post-Implementation of Your Big Data Project**

Once your big data project is up and running, you're not done. It's time to review what the agency accomplished—including what went well, what failed and what could have been done better—and to plan for the next project.

### **Step 16: Conduct a Postmortem and Impact Analysis on Your Big Data Project**

It is a good idea to **document the entire project, including lessons learned from all stages, to retain the institutional knowledge you gained as a result**. This information can also be shared with peers.

"One important element of conducting a postmortem is that it should not be used for evaluation or to point fingers at individuals or events," Desouza says. "Unless people are protected to share their true experiences and learning episodes, the postmortem exercise will not be of any value." Also, conduct a thorough impact analysis to convey the value of the project, accounting for improvements in both process measures and organizational value measures. Once it's done, publicize it.

### **Step 17: Identify Your Next Big Data Project**

The results of your big data project and the lessons learned from it should help you **identify opportunities for new big data projects. You'll be able to build on the practices and processes you established with your first project.** That said, give your team a little time to recover before plunging into the next project.

"One additional benefit of waiting before launching the next effort is that it gives CIOs more time to collect evidence on the performance and benefit of the first project," Desouza says. "This information will help CIOs make a stronger case for the next project."

## Annotated Bibliography – Task 1.

- Anderson, Richard and Daniel Roberts. *Big Data: Strategic Risks and Opportunities – Looking Beyond the Technology Issues*. Crowe Horwath Global Risk Consulting, September 2012 (8 pages. Legal and policy gaps. While some jurisdictions limit the use of personal data, in other instances the collection and analysis of data are largely unrestricted. In addition, certain regulatory requirements, such as anti-money-laundering regulations, require businesses to develop an unprecedented level of insight into customer behavior and activities. Conflicting or inconsistent privacy frameworks make it more challenging to manage large volumes of data efficiently, in a way that allows access when needed but limits access when prohibited. Identifies challenges and suggests a plan of action: Identify roles and responsibilities; Define goals and priorities; Assess critical data issues; Identify key risk indicators, Identify opportunities to add value)
- Brooks, J. and M.Wills. *Data-Driven Approaches to Delivering Better Outcomes* (Washington, D.C.: National Governors Association Center for Best Practices, July 24, 2015 (11 pages. Focused on helping governors identify strategies to improve the efficiency and effectiveness of state government. Ideas and examples drawn from three *Delivering Results* experts roundtables held in the fall of 2014 and conversations with national and state experts. This issue brief highlights lessons learned during the yearlong *Delivering Results* initiative about ways in which governors are building a more data-driven state government. Key lessons learned are: Set a clear vision of statewide, cross-cutting goals; Establish a strategic budget process; and, Support a management approach that ensures agency policies and processes achieve the enunciated goals.)
- Campbell, Chris. *Top Five Differences between Data Lakes and Data Warehouses*, <http://www.blue-granite.com/blog/bid/402596/Top-Five-Differences-between-Data-Lakes-and-Data-Warehouses>, January 26, 2015
- CDT. *Health Big Data in the Government Context*, *Center for Democracy & Technology* (15 pages. To explore the privacy and security implications of health big data, and to develop concrete proposals for how to address those issues and at the same time reap the benefits of big data, CDT is undertaking a series of consultations with stakeholders and experts. We are examining three scenarios: (1) clinical and administrative data generated by health care providers and payers; (2) health data contributed by consumers using the Internet and other consumer-facing technologies; and (3) health data collected by federal, state, and local

governments. The enormous potential of health big data for governments is undeniable, and can be achieved in a privacy-protective way through responsible practices, such as those outlined in this paper. It is critical for governments seeking to use large amounts of citizen health information to institutionalize privacy-protective measures in advance of any collection and use of the data, guided by the FIPPs and by well-designed ethical rubrics. Government entities must carefully weigh the benefits and risks to using any kind of personal health information, provide contextual notice and transparency, and facilitate meaningful citizen engagement and government accountability for their data practices. )

CDWG. *Proactive Planning For Big Data – In government, Big Data presents both a challenge and an opportunity that will grow over time*, CDWG White Paper (8 pages. Addresses Barriers to Big Data Success; Tips for tackling Big Data; Big Data Toolbox; Data Insight Layers; Analysis and Visualization; Big Data Security)

Center for Digital Government. *BIG DATA AND ANALYTICS- Research Report from the Center for Digital Government*. Public CIO Special Report, Q3 2015 (30 pages. Includes articles: The Powerful Combination of Big Data and Analytics; What Are Big Data and Analytics?; Big Data and Analytics at Work; Solutions to a Complex World; Overcoming Data Challenges; 10 Steps to Success; and, Achieving Results)

DHS. *DHS Information Sharing and Safeguarding Strategy*, US Department of Homeland Security, January 2013 (29 pages. outline goals and objectives that guide the activities of participants in the Homeland Security Enterprise towards a common information sharing and safeguarding end within the context of our distributed homeland security architecture.)

Duncan, Jeffrey, Wu Wu, Scot P. Narrus, Stephen Clyde, Barry Nangle, Sid Thornton, Julio Facelli. A Focus Area Maturity Model for a Statewide Master Person Index. OJPHI (13 pages. The sharing of personally identifiable information across organizational boundaries to facilitate patient identification in Utah presents significant policy challenges. The focus area maturity model provides an orderly path that can guide the complex process of developing a functional statewide master person index among diverse, autonomous partners. While this paper focuses on our experience in Utah, we believe that the arguments for using a focus area maturity model to guide the development of state or regional MPIs is of general interest.)

Dixon, Chris. *Big Data in State & Local Government*, Deltek, June 3, 2014 (17 slides. Big data in will grow very gradually and organically in state and local government. It will grow in response to immediate business needs, such as

cybersecurity or traffic management, not grandiose theories of transparency, openness, or performance.)

Executive Office of the President. *Report to the President – Big Data and Privacy: A Technological Perspective*. Executive Office of the President – President’s Council of Advisors on Science and Technology, May 2014. (76 pages. Explores changing nature of privacy as computing technology has advanced and big data has come to the fore; new ways in which personal data are acquired, both from original sources, and through subsequent processing. Provides 5 recommendations: Policy attention should focus more on the actual uses of big data and less on its collection and analysis; Policies and regulation, at all levels of government, should not embed particular technological solutions, but rather should be stated in terms of intended outcomes; NITRD agencies should strengthen U.S. research in privacy-related technologies and in the relevant areas of social science that inform the successful application of those technologies; OSTP, should encourage increased education and training opportunities concerning privacy protection, including career paths for professionals; The United States should take the lead both in the international arena and at home by adopting policies that stimulate the use of practical privacy-protecting technologies that exist today.)

Executive office of the President. *Big Data: Seizing Opportunities, Preserving Values*, Executive Office of the President, May 2014. (85 pages. Review focuses on how the public and private sectors can maximize the benefits of big data while minimizing its risks. It also identifies opportunities for big data to grow our economy, improve health and education, and make our nation safer and more energy efficient. A significant finding of this report is that big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace. Americans’ relationship with data should expand, not diminish, their opportunities and potential.)

Experfy Editor. *Major Hurdles in Big Data: Risks and Threats*, Experfy Insights <http://www.experfy.com/blog/major-hurdles-big-data-risks-threats/>, July 1, 2014 (Privacy and Security; Deriving conclusions from erroneous data patterns; Too much reliance on data; Limitations of big data. Includes several videos of speakers addressing various topics)

GAO. Federal Information System Controls Audit Manual (FISCAM), Government Accountability Office, February 2009 (manual provides a methodology for performing information system (IS) control audits in accordance with “generally

accepted government auditing standards” (GAGAS), as presented in *Government Auditing Standards*.)

- GAO. *Data Analytics for Oversight and Law Enforcement*, US Government Accountability Office, July 2013. (32 pages. summarizes the key themes that emerged from the discussion in the forum. Specifically, the report discusses the challenges and opportunities in (1) accessing and using data and (2) sharing data. In addition, participants identified next steps to address these challenges and capitalize on opportunities.)
- GAO. *Human Services – Sustained and Coordinated Efforts Could Facilitate Data Sharing While Protecting Privacy*, US Government Accountability Office, February 2013 (62 pages. To address identified challenges, stakeholders suggested that federal agencies could clarify federal privacy requirements and consider harmonizing requirements. Nearly all stakeholders GAO surveyed said that coordinated, multiagency guidance that clarifies what data sharing is permissible would be extremely useful. They also suggested that developing model data sharing agreements and informed consent language that comply with federal privacy requirements, or providing existing examples, would be useful. Stakeholders also said it would be highly useful to reexamine requirements to ensure more consistent privacy rules for data sharing across human services programs and agencies.)
- HITRUST, *HITRUST Common Security Framework*, Version 6, 2014 (481 pages)
- Hughes, Jessica. *Utah Mapping Partnership Aims to Build Live Data-Sharing Framework*, February 19, 2015
- Idala, David A., Martha H. Somerville, Laura A. Spicer, Cynthia L Boddie-Willis, Jamie L. John, and Tricia Roddy. *Overcoming Interagency Data-Sharing Barriers: Lessons from the Maryland Kids First Act*, The Hilltop Institute, January 2011 (10 pages. The lack of a datasharing agreement between Maryland’s Medicaid and tax agencies initially hindered both the efficiency of the process and the ability to evaluate Maryland’s tax-based outreach. New legislation enacted by Maryland’s General Assembly, however, now authorizes data sharing between the state’s tax and Medicaid agencies, allowing the possibility of a full evaluation of the initiative.)
- InfoLawGroup LLP. *The Privacy Legal Implications of Big Data: A Primer*, <http://www.infolawgroup.com/2013/02/articles/big-data/the-privacy-legal-implications-of-big-data-a-primer/>, February 12, 2013 (The potential uses and benefits of Big Data are endless. Unfortunately, Big Data also poses some risk to both the companies seeking to unlock its potential, and the individuals whose

information is now continuously being collected, combined, mined, analyzed, disclosed and acted upon. This post explores the concept of Big Data and some of the privacy-related legal issues and risks associated with it.)

Informatica. *Safeguarding Sensitive Data in State and Local Government – Advancing Cybersecurity with the Informatica Solution for Data Privacy*, Informatica White Paper, March 2013 (20 pages. discusses the challenges to securing information in state and local government organizations, outlines common sources of vulnerability, and illustrates with a case study an example of an increasingly common data breach. It discusses the effectiveness and versatility of data masking—both traditional, persistent data masking and the newer, breakthrough technology of dynamic data masking—in addressing the data privacy requirements of the public sector. It also examines the pros and cons of complementary data protection techniques, such as encryption and database activity monitoring, and how they can be used alongside data masking software to provide optimal protection in specific scenarios. Finally, the paper outlines what to look for in a data privacy solution and advocates implementing Informatica® data masking products to achieve robust, transparent, and cost-effective data privacy.)

(ISC)2 Government Advisory Council Executive Writers Bureau. *Big data = big exposure. What can you do about it?*, GCN July 29, 2013 (3 pages. By applying the existing approaches under FISMA with mature change and configuration management processes, agencies can begin to securely leverage the power of big data. Security teams will need to become more integrated and involved in the lives of data scientists and business units to understand how they are operating and where they need support. While big data is new to many agencies, the principles in protecting information and bringing mature management to an operation often is not. Agencies should leverage their existing operational and managerial controls to protect new technologies while automated tools are developed to add further rigor, maturity and automation.)

Jones, Steve. *Big & Fast Data: The Democratization of Information – Moving from the Enterprise Data Warehouse to the Business Data Lake*, Capgemini, 2015 (20 pages. The key is for IT to put in place technologies and delivery methods that enable effective democratization of insights in a way that last-generation approaches such as the EDW have failed to do, meeting the increasingly pressing need for fast time to insight. In fact, more than half (54%) of respondents stated that they consider leveraging fast data to be more important than leveraging big

data. In this way IT can help the business to realize the available opportunities. Provides 7 guiding principles: Embark on the journey to insights within your business and technology context; Enable your data landscape for the flood coming from connected people and things; Master governance, security and privacy of your data assets; Develop an enterprise data science culture; Unleash data and Insights-as-a-Service; Make insights-driven value a crucial business KPI; Empower your people with insights at the point of action.)

Kash, Wyatt. *Agencies lay groundwork to make data more valuable*, <http://fedscoop.com/agencies-lay-groundwork-make-data-valuable/>, December 12, 2014 (Making DHS's vast amount of data independent but also accessible only to those with the appropriate viewing privileges remains a massive challenge for the DHS officials)

Loshin, David. *Big Data and Government: Business Drivers and Best Practices*, Terradata (8 pages. Covers Why Big Data; Introducing Big Data into the Environment; Use Cases for Big Data; Enabling Big Data as Part of a Unified Information Architecture, Getting Started with Big Data. With big data gaining momentum, we are on the cusp of a new age in information management. Big data can add value through the enhancement of analytics and predictive modeling practices that leverage massive amounts of data. However, it is worth investing effort in properly scoping a big data program in a way that is aligned with the existing environment. Big data will be an integral part of an overall analytics strategy, but it can not bypass the best practices associated with adhering to the system development lifecycle (SDLC))

Marr, Bernard. *The 5 Biggest Risks of Big Data*, <http://data-informed.com/the-5-biggest-risks-of-big-data/>, June 24, 2015 (Data security, Data Privacy, Costs, Bad Analytics, Bad Data)

Martinez Pacin, Adrian. *A Vision for Big Data*, Intel. (9 pages. Provides Key elements for a Big Data policy. End goals include: Use research funding to incentivize breakthrough innovation in Big Data; Leading role of the Public Sector in the Data Economy; Ensure citizens trust in Big Data solutions protecting privacy rights; Ensure the supply of data scientists and data analysts. Recommendations: Ensure research in societal challenges that can be addressed by Big Data solutions; Focus research on hardware and software that enables Big Data processing (e.g. HPC, data centres, analytics); Enable the Public Sector Information for re-use; Adopt Big Data solutions for evidence-based policy-making; Clarify the distinction between personal data and non-personal data; Build a strong but balanced data protection framework to enable citizen trust;

Raise network security to enable citizen trust; Research in anonymization technologies; Develop comprehensive and special regimes for scientific processing of data; Commit the Grand Coalition for Digital Skills to address data skills; Engage with academia and the ICT sector to develop data scientist curricula; Develop a network of Centers of Excellence for Big Data.)

Moss, Larissa T. and Sid Adelman. *The Role of the Chief Data Officer in the 21<sup>st</sup> Century*, <http://www.cutter.com/content-and-analysis/resource-centers/business-intelligence/sample-our-research/biar1302.html> , Vol 13, NO. 2 (19 pages. Includes Business case for a Chief Data Officer, Proposal for a Chief Data Officer, Qualifications, Activities, Responsibilities and Authority; Data Governance; Data Quality, etc. To achieve maximum benefit, all enterprise-class activities, including data warehousing, business intelligence, master data management, customer relationship management, data governance, data quality improvement initiatives, enterprise architecture, and so on, should be lead by a new chief officer whose primary responsibility is the standardization and management of data assets in the organization. This new position is the chief data officer.)

Nahra, Kirk J and Wiley Rein LLP. *The Evolving Worrlld of Privacy and Security*, Wiley Rein LLP, May 20, 2015 (24 slides. debate about “non-HIPAA” healthcare data and what it means for all health care data users and the future of health care privacy (and perhaps overall privacy))

NASCIO. *Enterprise Security Assessment: Information Security and Privacy – State of Utah*, NASCIO, 2008

NIST Big Data Public Working Group. Security and Privacy Subgroup. *NIST Special Publication 1500-4 – NIST Big Data Interoperability Framework: Volume 4, Security and Privacy*, Final Version, National Institute of Standards and Technology, US Department of Commerce, August 14, 2015 (70 pages. Exploration of security and privacy topics with respect to Big Data. This volume considers new aspects of security and privacy with respect to Big Data, reviews security and privacy use cases, proposes security and privacy taxonomies, presents details of the Security and Privacy Fabric of the NIST Big Data Reference Architecture (NBDRA), and begins mapping the security and privacy use cases to the NBDRA.)

NIST Big Data Public Working Group. *NIST Special Publication 1500-3 DRAFT NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements*, Draft Version 1, April 6, 2015  
<http://dx.doi.org/10.6028/NIST.SP.1500-3> (51 use cases gathered by the NBD-PWG Use Cases and Requirements Subgroup and the requirements generated

- from those use cases. The use cases are presented in their original and summarized form. Requirements, or challenges, were extracted from each use case, and then summarized over all of the use cases.)
- OECD. *EXPLORING DATA-DRIVEN INNOVATION AS A NEW SOURCE OF GROWTH – Mapping the Policy Issues Raised by “Big Data”*. Organisation for Economic Co-operation and Development, June 18, 2013 (35 pages. Explores the potential role of data and data analytics for the creation of significant competitive advantage and for the formation of knowledge-based capital (KBC), which can drive innovation and sustainable growth across the economy and society. Mapping the policy opportunities and challenges Pages 21-29)
- Olavsrud, Thor. *17 Steps to Implement a Public Sector Big Data Project*, CIO, March 18, 2014 <http://www.cio.com/article/2368491/big-data/144854-17-Steps-to-Implement-a-Public-Sector-Big-Data-Project.html> (Based on interviews with CIOs at every level of government, Kevin C. Desouza of Arizona State University lays out a three-stage, 17-step big data implementation plan)
- O’Reilly Radar Team. *Planning for Big Data – A CIO’s Handbook to the Changing Data Landscape*, February 2012. (88 pages. The aim of this book is to help you understand what big data is, why it matters, and where to get started. If you’re already working with big data, hand this book to your colleagues or executives to help them better appreciate the issues and possibilities)
- Patil, D.J. and Hilary Mason. *Data Driven – Creating a Data Culture*. O’Reilly, January 2015. (28 pages. Explores what it takes to be a data-driven organization and develop a data-driven culture.)
- Peppers & Rogers Group. *Achieving Excellence via Data-Driven Decision Making in Government - Serving the Future with Power of Analytics*. The Government Summit Thought Leadership Series in collaboration with Peppers & Rogers Group, February 2013. (Provides assessments of Data-Driven Decision making in Private Sector and Public Sector pages 6-17. Includes sections addressing: Smart Policing for Reducing Crime; Improving the Government Healthcare System using DDD; Ubiquity of DDD in Modern Public Service; From Service-Centric to Citizen-Centric Service Delivery)
- Reynolds, Paul. *Privacy Impact Assessment for the Common Entity Index Prototype*, US Department of Homeland Security, September 26, 2013 (18 pages. The CEI Prototype will enable DHS to correlate and consolidate a limited set of identity data from select component-level systems and organize key identifiers collected about individual members of the public. The purpose of this prototype is to determine the feasibility of establishing and effectively controlling access to a

centralized index of select biographic information, enabling DHS to provide correlated and consolidated identities.)

Reynolds, Paul. *Privacy Impact Assessment Update for the DHS Data Framework*, US Department of Homeland Security, August 29, 2014 (18 pages. The DHS Data Framework (“Framework”) is a scalable information technology program with built-in capabilities to support advanced data architecture and governance processes. The Framework is DHS’s “big data” solution to build in privacy protections while enabling more controlled, effective, and efficient use of existing homeland security-related information across the DHS enterprise and with other U.S. Government partners, as appropriate.)

Reynolds, Paul. *Privacy Impact Assessment Update for the Neptune*, US Department of Homeland Security, February 27, 2015 (14 pages. Neptune is the unclassified “data lake,” which DHS currently uses to receive, store, and tag the data from unclassified DHS information technology systems. Once tagged, unclassified DHS data sets from Neptune are transferred to Cerberus, which is the classified data lake that DHS currently uses to perform classified searches of unclassified DHS data sets. The Common Entity Index is an unclassified correlation engine that will allow DHS to connect disparate DHS data sets to view all available information about an identified individual.)

Robinson, Doug. *Managing Data as a Strategic Asset: Reality and Rewards*, NASCIO May 11, 2015 (33 slides. Covers Government Data Landscape: Data stored across multiple systems from multiple agencies; Lack of standards, consistency; Security concerns and privacy issues; Data quality issues: dirty and messy; Data sharing is difficult-format, language, access, culture, myths; Little insightful, usable data on “customers”. Managing Data: Need enterprise imperative and governance; Inventory data systems across the enterprise to identify the array; Understand security and privacy implications; Data Divide – The rise of Data Poverty; Power of Visualization and dashboards for transparency; Challenges with state skill sets, competencies, recruiting; Expect surprises and unintended consequences; Information Asset Portfolio. Data Management Capability Maturity Model Levels)

Russom, Miriam B, Robert H. Sloan, Richard Warner. *Legal Concepts Meet Technology: A 50 State Survey of Privacy Laws* (20 slides. Conclusions: Surveyed laws fail to address issue of how much control we should have over our information; Lack of consensus on what to protect under the PII rubric; Difficulties to delimit PII; Data security laws rely on reasonableness standards; In practice consent is taken as sufficient for authorization; Little legal constraint on data sharing; Practical and

technical problem with access and correction; Large aggregate costs vs low individual expected costs.)

SAGIROGLU, Seref and Duygu SINANC. *Big Data: A Review*, Gazi University Department of Computer Engineering, Faculty of Engineering, Ankara, Turkey, 2013 (6 pages. Reviews Privacy and Security concerns. Suggests security model for Big Data)

Stein, Brian and Alan Morrison. *Data Lakes and the Promise of Unsiloed Data*, <http://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/features/data-lakes.html> 2014, PWC (6 pages. Covers Data Lake Architecture, Why a Data Lake?; Motivating Factors Behind the Move to Data Lakes; Early Lessons and Pitfalls to Avoid; Data Flow in a Data Lake; How a data lake Matures)

Sweden, Eric. *Is Big Data a Big Deal for State Governments: The Big Data Revolution – Impacts for State Government – Timing is Everything*, NASCIO 2012 (15 pages. Big data carries many, big implications – for better and for worse. State government should be preparing now for the potential of big data and ensure current investment in technology allows for future leverage of big data capabilities. NASCIO will explore a number of topics in the future related to big data and will stay connected to developments at the federal and national level as the research agenda moves forward. Topics that can be anticipated in the future include enterprise architecture management, privacy, data lineage and quality, cost / benefit analysis, the necessary training agenda and applications in specific government lines of business.)

Tobin, Patrick. *Big Data Brings Four Big Risks*, Forbes, <http://www.forbes.com/sites/sungardas/2013/12/18/big-data-brings-four-big-risks/>, December 18, 2013 (Loss of agility; Loss of compliance; Loss of security; Loss of money. Companies today need to *manage their data to minimize their risk*. This involves having policies that are in compliance with regulatory standards, processes that cover all contingencies, retention schedules that are up to date, and a consistent self-evaluation to determine what data is necessary for the proper functioning of the company. The more efficiently companies store, manage, and host their data, the more agile, compliant, secure, and cost-effective they will be. And that will take the big risk out of big data.)

- Utah Code: Title 46. Notarization and Authentication of Documents and Electronic Signatures
- Utah Data Alliance. *Data Privacy and Security*, [www.UtahDataAlliance.org](http://www.UtahDataAlliance.org),
- Utah Data Release Policy for Utah's IBIS-PH Web-Based Query System, Utah Department of Health
- Utah Data Security Management Council, 2015 General Session (Bill creates a Data Security Management Council to develop recommendations for data security and risk assessment)
- Utah Department of Administrative Services Enterprise Information Security Policy
- Utah Department of Technology Services. Information Technology Plan FY 2015
- Utah Department of Technology Services. 2015-2018 Strategic Plan
- Utah Governmental Internet Information Privacy Act
- Utah Privacy Policy Statement
- Utah Rule R277-487. Public School Data Confidentiality and Disclosure
- Wood, Colin. *Data Governance: The Public Sector's Next Big Frontier – Managing and organizing data is the next phase of state and local government's hopeful metamorphosis*, <http://www.govtech.com/data/Data-Governance.html>, April 29, 2014 (Data has become the lifeblood of organizations, and to make the most of that information, organizations need a framework that addresses all of the issues of data quality, standards and management. Without good data governance, organizations are spending more to be less efficient and less effective, not to mention the degradation in transparency. Data governance is becoming critically important, and it's the CIO who's in the best position to be a champion for that cause.)
- Yiu, Chris. *The Big Data Opportunity – Making Government Faster, Smarter and More Personal*, <http://www.policyexchange.org.uk/images/publications/the%20big%20data%20oportunity.pdf> (Discusses the opportunity for data and analytics to transform public service delivery, to sound a note of caution about the challenges this agenda poses for the public sector, and to make recommendations for how government might begin to realise the former whilst addressing the latter. Capitalising on the full potential of big data in the public sector is a major challenge and will not be achieved overnight. Nevertheless we believe that the prize at stake – better services for real people, and a leaner, smarter public sector – merits a renewed effort to make better use of the public sector's data assets)

Zients, Jeffrey D. and Cass R. Sunstein. *Sharing Data While Protecting Privacy*, US Office of Management and Budget Memo M-11-02, November 3, 2010. (4 pages. When agencies share data, they must do so in a way that fully protects individual privacy. The public must be able to trust our ability to handle and protect personally identifiable information. In sharing data, agencies must comply with the Privacy Act of 1974<sup>2</sup> and all other applicable privacy laws, regulations, and policies. In addition to the legal framework that governs the use and disclosure of data, agencies are advised to consult established codes of Fair Information Practices. Memorandum is to direct agencies to find solutions that allow data sharing to move forward in a manner that complies with applicable privacy laws, regulations, and polices. These collaborative efforts should include seeking ways to facilitate responsible data sharing for the purpose of conducting rigorous studies that promote informed public policy decisions.)