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

Data Driven Decision Making Project

Volume 2: People – Organization and Skills
September 2015

Submitted to:
Dave Fletcher
Chief Technology Officer
State of Utah
dfletcher@utah.gov

Point of Contact:
Frank Crichton
Project Manager
SAIC
frank.h.crichton@saic.com

Executive Summary

Data-driven organizations in the business world have demonstrated significant improvements in services and efficiencies over their peers who are not data-driven. The same improvements in State services and quality of life improvements for the State's citizens can be realized by following the same processes and methods. Moving an organization to embrace and continually purse better evidence-based decisions requires leadership and skills on both the business and technical sides. On the business side, such an initiative must be funded by the legislature, led by executive leadership, guidance by key state decision-makers, have participation by agency leadership, and be leveraged by business analysts. On the technical side, new skillsets need to be developed or hired for managing cloud-based infrastructure, for creating and maintaining big data platforms, and for data science skills to convert operational data into actionable intelligence. Evidence-based management through data-driven decision-making has the potential to transform the efficiency of the State government.

1 Introduction

This document is one in a series of five documents, describing an assessment of the state of the art in technology and skills for a cost-effective deployment of a Big Data solution. This Task 3 Technology Roadmap document describes the technology choices for standing up a big data environment, along with some indication of the relative pricing for the technology choices. The companion documents in this study are Task 1 Policy and Governance, Task 3 Technology, Task 4 Business Case, Task 5 Process and Value.

2 Being Data-Driven

While the hype around Big Data and Data Science centers on the emerging technologies, the ability to use these technologies to affect change centers on organizational culture. Handling data and generating analytics has no impact on the organization unless the newly created insights can be actionable. The state must prioritize a continuous improvement mindset, one that considers the statewide needs, and not just the operational needs in individual agencies. In one report[1] data-driven businesses were estimated to have 5-6% greater output and productivity than their non-data-driven peers.

[1] Brynjolfsson, Erik, et. Al. "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance", http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486

Being a data-driven organization does not simply mean you have access to greater information, for example through reporting. Reports are essentially a rear-view mirror that provides “trailing indicators” on state services. To be data driven, the state needs the ability to answer the why questions for the trailing indicators. In other words to be data-driven is to identify the “leading indicators” that are causing the current results. Only by understanding why something happened can you plan a course of action to affect the leading indicators to cause performance improvement.

To support evidence-based decisions, you have to collect and integrate the right data, and assure that the cleansing and analysis of that data is valid. In addition to this you have to have people who can ask the right questions of the data. To be data-driven is to have a culture of continuous improvement with both the right processes and the agency-wide focus in place to become data-driven in critical business decisions. The state cannot of course become a data-driven organization without the technical underpinnings to bring the state’s data together to allow analysis. This pushes the state technically to adopt Big Data solutions.

> If your organization stores multiple petabytes of data, if the information most critical to your business resides in forms other than rows and columns of numbers, or if answering your biggest question would involve a “mashup” of several analytical efforts, you’ve got a big data opportunity.[2]

The technical roadmap for developing the Big Data solution that will allow the state to become a data-driven organization is presented companion report for Task 3: Technology. Having the data and the technology will not achieve results unless accompanied by organizational and individual changes.

In this report we cover the business roles that are important for organizational transformation. As a roadmap, Appendix A provides a simple guideline for measuring the organizational progress in adopting Big Data technologies.

3 Business Roles

The most critical element to being a data-driven organization is the clarity of leadership to push for the changes needed to get away from opinion-based decisions.

> *Without data you’re just another person with an opinion*
> *– William Edwards Deming*

[2] Thomas H. Davenport and D.J. Patil, “Data Scientist: The Sexiest Job of the 21st Century”, https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

The state must strive to identify ways to support decisions by factual analysis, and likewise use analysis to evaluate improvements in state services.

3.1 Legislative Leadership

Like the board of directors in a company, the Legislature needs to back the vision of increased information availability to support decisions. Leadership must have the backing of the legislature to obtain the funding to execute on this mission for evidence-based decision-making. Utah is pursuing this support in the request for funding, but must focus on continual communication of activities and results to demonstrate action.

3.2 Executive Leadership

Critical for a data-driven organization is the ability to reward those who share data, incentivizing the agencies and individuals to develop and nurture shareable data and analytics. There has to be a clear signal from the executive office that the data is not "owned" by the operational units, but that they belong to the state as a whole. A common hindrance to becoming a data-driven organization is that some agency heads or agency staff will view the directives to share data as an unfunded mandate that will impact their operational activities. It is important to evangelize the benefits of sharing data across the state, without compromising compliance or increasing risk.

There are a number of positions that are used to help distinguish from the oversight of technology systems (CTO), and the oversight of data and governance (CIO). Businesses use a number of titles to represent the role that focuses on the organization-wide analytics. The person whose focus is on deriving new insights from the data area are variously called:

- Chief Data Officer
- Chief Science Officer
- Chief Digital Officer
- Chief Analytics Officer

There is indeed value in having someone with a separate focus to move the organization forward to always looking to make evidence-based decisions. While the title doesn't matter, the culture change for an organization is a significant one and must be given a visible priority. For the state with CTO and CIO positions, those two must work closely together to ensure the new focus is prioritized across their spheres of responsibility.

Given that the emphasis must be on state-wide involvement and improvement, it is important that the leadership be at the executive level, and not in a particular agency. If not aligned with the executive, the activities will all be perceived by other organizations as a threat to their operations. The same question is asked for where the data scientists should be placed in the organization. While they will be working closely with business analysts in the agencies, they should remain as a central organization. Having the state-wide focus will enable them to better leverage the variety of data outside an agency, to address questions inside an agency.

The executive leadership's role is to energize the organization around the vision of becoming data-driven, get them aligned, and get everyone working towards this goal.

3.3 Advisory Board

Becoming a data-driven organization obviously requires an investment in technical resources and technical skills, but the success or failure of this initiative hinges more on culture change. It is vital that the organization understands and embraces the vision of thinking first of what data analytics can be applied to decisions. It is not enough to look at reports that provide key metrics that are measuring outcomes. It is critical that the State develop the mindset for analyzing the key drivers, or leading indicators that produce those outcomes. To percolate this decision-making culture change, an advisory board or governance board should be formed with participation of decision-makers across the state government. Led by executive leadership, this board would provide input into prioritizations, evaluate policy and regulatory concerns, provide vision in longitudinal studies that focus on individuals or communities from a holistic perspective. An advisory or governing board would guide the efficient utilization of this new capability for the state while also leading the culture change for data-driven decisions.

3.4 Business Leadership

> *"Do you have data to back that up?" should be a question that no one is afraid to ask (and everyone is prepared to answer).*[3]

It is difficult for someone with decades of experience to change to purse data-driven decisions. A somewhat humorous way to say this is that becoming data-driven means getting away from decisions by "HiPPO", or the Highest Paid Person's Opinion[4]. HiPPOs can be a significant problem, because the decisions are often based on –ill understood metrics or even pure guesswork. Having a "gut-feeling" for making a decision must give way to finding the critical factors from data analysis that will truly affect the outcomes. Having the best quality and timely data, analyzed by the best analysts will be for naught if the front-line decision maker has already made up their mind no matter what the data shows. It is critically important that agency leaders be on-board with the process if it is to succeed.

[3] Julie Arsenault, "How to Create a Data-driven Culture," PagerDuty, October 2, 2014

[4] HiPPO was a term coined by Avinash Kaushik and Ronny Kohavi in 2006, first blogged by Kaushik at http://www.kaushik.net/avinash/seven-steps-to-creating-a-data-driven-decision-making-culture/

3.5 Business Analysts

Data Science is a team sport. While the state will need to have dedicated data scientists that report to an executive level organization, they will be working closely with the business analysts who understand the complexities of their data and their mission. It is critically important that the business analysts can focus on what is important, and the data scientist can work with them to bring knowledge of the data outside that agency to achieve their goals.

4 Technical Roles

The technical roles can be broken down according to the architecture for Big Data systems, in other words dealing with the infrastructure, the platform, and the application. There is an overlap of understanding required to work at a given level. For example the person working with the platform needs to have an understanding of the underlying infrastructure to work with the cloud engineer, as well as an understanding of the data to be distributed, and the analytical approach to work with the data scientist or business analyst.

In all cases the roles are not meant to literally equate to an individual. How these skills are staffed depends on the skills of the folks currently at hand, and what skill gaps need to be filled. For convenience, the activities and needed skills will be discussed relative to the Architecture Layer.

4.1 Infrastructure Layer

While there are a number of common elements between managing a virtualized cluster and administering a physical computer, both the tools and the concepts have changed significantly in the emergence of cloud technologies. Traditionally there has been tight coupling in all aspects of administering servers, now there is a more modular approach that provides a separation of the infrastructure layers.

4.1.1 Cloud Engineer

When using the public cloud, the physical systems are maintained by the cloud provider, and the Cloud Engineer (CE) begins work at the virtual machine level. The engineer will need to create and manage the cluster through cloud management interfaces, to create the network structure to possibly including multiple networks. The virtual servers must be created, named, and assigned to run in specific regions. One element that can be overlooked is ensuring that the backup or disaster recovery solutions are in different regions to for additional fault tolerance to outages. The cloud engineer is responsible for documenting all the processes and operational deployment for the cluster. The Cloud Engineer works with the Platform Engineer to design the optimal choice for virtual server characteristics in terms of CPU and memory, and deploying server and storage components for optimal cost and performance. The CE works with security personnel to ensure that the systems are secure.

The activities for an on-premise solution follow the same criteria with the addition of the maintenance of the physical machines. For some compute-intensive applications, the cloud engineer would need to analyze the potential performance gains for running the platform on “bare metal”. The CE needs to know how to analyze the benchmarks in the different deployments to provide the optimal choice for the specific workload.

4.1.2 Cloud Engineer Skills

Cloud Engineer (CE) skills encompass system engineering, systems analysis, and systems administration skills. While much of the needed knowledge resides in the established areas of operating systems, virtualization and scripting, there are new specifics related to the orchestration of cloud services offered by cloud providers such as AWS. For cloud deployments, the CE needs to understand compute services including auto-scaling and load balancing; the range of storage services and their trade-offs in cost and retrieval latency; database services such as RedShift and DynamoDB; network tools such as those for creating a virtual private cloud; and security and identity services.

For on-premises work, the CE should know how to manage a cluster of servers using the equivalent open source cloud tools. Systems would be managed with software such as Chef and Puppet, with configuration management through tools such as Zookeeper, and workflow software such as Oozie.

In both cases, these skills are expected to be available in recent university graduates, and may also be obtained through vendor specific conferences and training, such as the AWS specific courses online and those available at conferences such as AWS-Invent. For open source tools, a number of technical conferences have tutorial sessions, and in some cases such as the Big Data Tech Con, the entire conference is focused on hands-on tutorials of the software tools in the Apache stack.

4.2 Big Data Platform Layer

The emerging Big Data technologies are changing the way organizations view data management and analytics. There has been an explosion over the course of the last ten years in software solutions at all layers of the IT architecture. The state will want to leverage their existing staff, and extend their skills or hire additional staff with skills in the new technologies. The advantage is that all current graduates will have skills in these techniques.

4.2.1 Big Data Engineer

Analytics skills are provided by Data Scientists. The platform their tools sit on top of is focused on the data management and data processing across a cluster. The new techniques for handling these skills focus primarily around software engineers. The number of tools available have skyrocketed in the last decade.[5] The specific choices for software are provided in Task 3: Technology. The Big Data Engineer (BDE) would need to be able to install a Hadoop distribution, such as Cloudera, Hortonworks, or MapR. They need to understand how to store and query data in new distributed file systems such as HDFS or storing data in S3 buckets. They need to understand how to distribute data across the nodes in a cluster to balance performance. They need to understand how to optimize the performance of the system as the amount of data in the platform grows. The BDEwould be expected to understand cloud management services, as well as open source tools for scheduling (such as zookeeper) and workflow (such as Oozie).

The engineer must be able to work with the cloud engineer to determine the optimal configuration of virtual machines in terms of CPU and memory for the most cost-effective configuration. Multiple storage options are available and must be evaluated for cost and performance. By analyzing data usage patterns, some data can be migrated to off-line storage for significant cost savings, such as moving data to Glacier on AWS.

4.2.2 Big Data Engineer Skills

The skills for a Big Data Engineer rely heavily on traditional programming, database management and database administration skills. To this should be added some knowledge of the cloud in order to work with the cloud engineer. Cloud skills can be obtained through online courses from cloud providers such as AWS. The Platform providers offer training classes tailored to their specific platform. To understand the new open source tools, Massively Open Online courses or hands-on tutorials at conferences are valuable.

4.3 Analytics Processing Layer

One of the new emerging rolls is for a Data Scientist. Data Science is a new discipline that has arisen in big data that is not a clear derivative of traditional skill lines, but is in fact a combination of multiple disparate disciplines. Data Science is not just analytics.

[5] Judy Qiu, Shantenu Jha, Geoffrey Fox, “HPC-ABDS: The Case for an Integrating Apached Big Data Stack with HPC”, http://www.slideshare.net/Foxsden/hpcabds-the-case-for-an-integrating-apache-big-data-stack-with-hpc

> *Data Science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing.*[6]

Data Science has been described as one of the most difficult skillsets for hiring. Data Science is a term that has arisen hand-in-hand with Big Data. To understand why, you can think of it as a progression from earlier analytical skills. Centuries ago the earliest analysts were statisticians who developed precise answers to specific questions using tightly controlled data. Decades ago, Data Miners began to be viewed as different since they were looking for approximate answers, to diverse business questions, using repurposed data that was often messy. To do their analysis they began to deploy a wider range of mathematical techniques often focusing on emerging Machine Learning methods. Data Science is now beginning to differentiate itself in that the magnitude of the data itself has become a significant problem. Analytics can, for example, no longer be separated from the methods to store the data across a cluster of computers. Data Scientists must now be cognizant of systems and software engineering issues prior to designing their analysis.

Data Science is the term for the management, analysis, and interpretation of data that is big, fast moving, or diverse. These data characteristics are described as the *volume*, *velocity*, and *variety* of data.

The goal of Data Science is to generate value for the state, which will derive primarily from the variety characteristic.

> *Business benefits are frequently higher when addressing the* ***variety*** *of data than when addressing* ***volume***"[7]

Priority areas of the state such as air-quality, inter-generational poverty, and fraud require the combination of data from a number of agencies and system to properly understand the causal contributing factors.

4.3.1 Data Scientist

A Data Scientist must embody a number of skills, and a number of characteristics. As described above it is no longer sufficient to understand analytical techniques. The successful Data Scientist must embody knowledge of the mission the state has prioritized, have an understanding of the data and the operational domains, must understand statistical and data mining techniques, and must understand systems and software engineering, as shown in Figure 2-1.

---

[6] NIST SP 1500-1, "NIST Big Data Interoperability Framework: Volume 1, Definitions, 2015

[7] Mark Beyer and Doug Laney, Gartner, 2012, "The Importance of Big Data: A Definition"

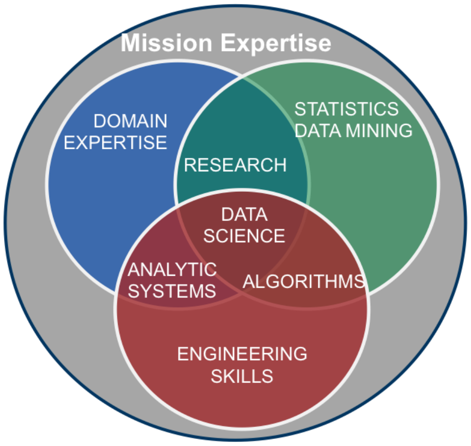

Figure 2.1 Data Science Skills

It is no surprise then that the Data Science "Rock Star" that can embody all these skills is a very rare person to find. What this means in actual practice is that you look to build Data Science Teams. Each member of the team must have some understanding of the other disciplines, but be primarily focused in only one or perhaps two of these disciplines. Since business analysts within the agencies will have the mission and domain expertise, and the big data engineering will be covered in the next section, the state should focus on finding data science candidates that have a background in mathematics or science, and have learned enough to communicate with team members in the other disciplines. Data Scientists are difficult to find. Ideally the state will locate a senior data scientist and then be able to hire some junior data scientists, or mathematically inclined university graduates. An ideal approach is to have a senior data scientist, then work with the state universities and colleges to offer internship positions for undergraduate and graduate students. As the senior data scientist mentors them, they can grow into the position. The state would thus be in a position to hire the most promising students when they graduate.

4.3.2 Data Science Skills

To find data scientists, or identify candidates that the state can grow into data scientists, the best source is people with math or science degrees. The reasoning is that these folks are used to getting a topic; figuring out how to ask a tractable questions; formulating a way to answer the question; figuring out a way to gather the data needed (either collecting existing data or instrumenting their own data collection); determining the right analytical approach; checking their work; and interpreting the results. They don't allow technical limitations to bog them down; if a tool doesn't exist to do what they need they'll program their own. This sounds much like the scientific method, and that's precisely what it is. The important characteristics are intense curiosity, a desire to roll up your sleeves and get to the heart of the problem, the ability to focus on the problem, and the ability to think outside the box when needed.

There are a number of MOOCs (Massive Open On-line Courses) that cover a number of data science skills. The most critical skills to pursue are in the areas of Machine Learning, and in the practical sense the most appropriate specific skill is to understand the Spark data analytics framework. Over the course of the last two years this framework has become the analytical approach of choice. There are a number of conferences such as the Big Data Tech Con and Strata that provide sessions with hands-on tutorials with a number of the new tools.

The combination of data science skills in someone who is experienced is very hard to find, but the state can pursue the path of mentoring recent graduates as they work with both business analysts and with big data engineers on specific problems.

4.4 Cloud Security Fabric

Because cloud security must address vulnerabilities at all system layers, it is often described as developing an overall security fabric. While on-premises solutions can rely on the traditional perimeter security, cloud systems introduce a number of additional vulnerabilities and requirements for a Cloud Security Engineer (CSE). The CSE would need to work with the cloud engineer, the platform engineer, and the data scientist to ensure that all vulnerabilities have been addressed. They would be expected to work closely with other State security officers to ensure the protection of the extended Utah perimeter that encompasses on-premises resources and the public cloud.

4.4.1 Cloud Security Engineer

The CSE needs to understand the full architectural stack and be able to address the security needs at each layer, in the network, and in the network connectivity to the system users. Given the extensive touch points in cloud security, we list a number of the specific areas where a CSE should have knowledge.

- Create, deploy and administer Virtual Machines, vCloud vApps, and vCenter, vApps. Managing virtual machine clones and templates.
- End-to-end understanding of virtual network security hardening best practices and techniques
- Create security configuration checklists (e.g., hardening or lock down guides) for virtual technology platform (e.g., operating systems, databases, firewalls, etc.).
- Must know FedRAMP process and documentation for cloud certification.
- Understand DISA STIGS and how to test and implement them in VMWare, Microsoft Windows 2008-2012 R2, Redhat RHEL 6.5, Applications and Database Servers etc.
- Clarify and implement National Institute of Standards and Technology (NIST) standards, including developing information security policies, standards, procedures, and checklists.
- Must have background knowledge in NIST 800-53 rev.4 security controls and Cloud guidance special publications.
- Must have a thorough understanding of the Cloud Risk Management Framework (CRMF) for certification and accreditation process in respect to virtual network environments.

- Research, communicate and apply Federal mandates pertaining to security configurations, policies and procedures within Federal Information Security Management Act (FISMA) and NIST.
- Apply a broad range of network security engineering skills to effectively perform complex assignments demanding familiarity with principles, theories, concepts and technologies surrounding firewalls, IDS/IPS, incident identification and analysis
- Apply experience with Microsoft Windows Active Directory, and file structure with security group permissions.
- Apply experience in virtual perimeter and application firewalls. Knowledge of virtual appliances and software Blade Servers etc.
- Apply strong understanding of performing and maintaining Continuous Monitoring in virtual environments, vulnerability scanning tools, software updates and security patches performed weekly of VMWare and virtual machine operating systems.
- Apply experience in VM to create and deploy virtual Machines with a security group profile established.

4.4.2 Cloud Security Skills

Cloud security skills reside at the overlap of traditional security protections, and the new technologies emerging around cloud and big data. The skills needed would first need to encompass traditional security skills for infrastructure and application protection, identity and key management, encryption at rest and in motion, and network security concepts. To this must be added sufficient understanding of each of the infrastructure, platform, and processing layers to work with the other members of the team whose main focus is in these areas. In addition to coursework related to cloud services and big data technologies in common with a cloud engineer, additional training can be found through organizations such as the Cloud Security Alliance.

5 Collaboration

A well-known saying is the reminder that “the smartest person doesn’t work for your organization”. As the State progresses into big data technologies to improve evidence-based decision making, there are a number of ways to enhance collaborations. One valuable way is to work with research universities. By creating datasets that mask privacy elements and providing data in research sandboxes, the state can collaborate with researchers to develop analytics and analytic methods for challenging problems. This would also provide a way for interns to come on board for specific projects, potentially providing a future employment pipeline for the state as well.

In Task 1 we discussed the policy issues with data sharing. Greater collaboration with other states could provide lessons learned in the ways policies and regulations have been dealt with. Targeted workshops at NASCIO meetings could be organized around data sharing, case studies, analytical methods, collaboration methods.

The state can benefit from attendance at technical meetings such as Strata which have different practical sessions such as, such as “Data-Driven Business Day” discussing real-world data science use cases; “Rday” and “Pyday” for practical tips on the use of these data science toolkits, “Law, Ethics, and Open Data” for examples of the interactions between these disciplines; “Security” for practical examples of protection big data infrastructures.

Appendix A: Organizational Readiness[8]

Technological readiness is useful for assessing the maturity of the technology components which make up Big Data implementations. However, successful utilization of Big Data technologies within an organization strongly benefits from an assessment of both the readiness of the organization and its level of adoption with respect to Big Data technologies. As with the domains and measures for the Technology Readiness scale, we choose definitions similar to those used for SOA.

A.1 Organizational Readiness Domains

- **Business and Strategy:** Capabilities that provide organizational constructs necessary for Big Data initiatives to succeed. These include a clear and compelling business motivation for adopting Big Data technologies, expected benefits, funding models etc.
- **Governance:** The readiness of governance policies and processes to be applied to the technologies adopted as part of a Big Data initiative. Additionally, readiness of governance policies and processes for application to the data managed and operated on as part of a Big Data initiative.
- **Projects, Portfolios, and Services:** Readiness with respect to the planning and implementation of Big Data efforts. Readiness extends to quality and integration of data, as well as readiness for planning and usage of Big Data technology solutions.
- **Organization:** Competence and skills development within an organization regarding the use and management of Big Data technologies. This includes, but is not limited to, readiness within IT departments (e.g., service delivery, security, and infrastructure) and analyst groups (e.g. methodologies, integration strategies, etc.).

A.2 Scale of Organizational Readiness

1. **No Big Data**
   - No awareness or efforts around Big Data exist in the organization
2. **Ad Hoc**
   - Awareness of Big Data exists
   - Some groups are building solutions
   - No Big Data plan is being followed
3. **Opportunistic**
   - An approach to building Big Data solutions is being determined
   - The approach is opportunistically applied, but is not widely accepted or adopted within the organization
4. **Systematic**
   - The organizational approach to Big Data has been reviewed and accepted by multiple affected parties.
   - The approach is repeatable throughout the organization and nearly-always followed.
5. **Managed**

[8] Thanks to David Boyd of the NIST Big Data Working Group for his original draft of this assessment.

- Metrics have been defined and are routinely collected for Big Data projects
- Defined metrics are routinely assessed and provide insight into the effectiveness of Big Data projects

6. **Optimized**
    - Metrics are always gathered and assessed to incrementally improve Big Data capabilities within the organization.
    - Guidelines and assets are maintained to ensure relevancy and correctness

A.3 Scale of Organizational Adoption

1. **No Adoption**
    - No current adoption of Big Data technologies within the organization
2. **Project**
    - Individual projects implement Big Data technologies as they are appropriate
3. **Program**
    - A small group of projects share an implementation of Big Data technologies
    - The group of projects share a single management structure and are smaller than a business unit
4. **Divisional**
    - Big Data technologies are implemented consistently across a business unit
5. **Cross-Divisional**
    - Big Data technologies are consistently implemented by multiple divisions with a common approach
    - Big Data technologies across divisions are at an organizational readiness level of Systematic or higher
6. **Enterprise**
    - Big Data technologies are implemented consistently across the enterprise
    - Organizational readiness is at level of Systematic or higher