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

Data Driven Decision Making Project

Volume 4: Business Case Version 1
September 30, 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

Table of Contents

EXECUTIVE SUMMARY **ES-1**

1 INTRODUCTION AND PURPOSE **1**

1.1 PURPOSE 1
1.2 BUSINESS DRIVERS 1

2 GAP ANALYSES **5**

2.1 "AS-IS" ENVIRONMENT 5
2.2 "TO-BE" ENVIRONMENT 8

3 GLOBAL ASSUMPTIONS AND METHODOLOGY **11**

3.1 ASSUMPTIONS 11

4 ALTERNATIVES **15**

4.1 ALTERNATIVE DESCRIPTIONS 15
4.2 PLATFORM OPTIONS 16
4.3 COMPARISON 16
4.4 ALTERNATIVES CHOSEN FOR FURTHER EVALUATION 19

5 ALTERNATIVE 1: ON-PREMISE SOLUTION **21**

5.1 DESCRIPTION 21
5.2 SPECIFIC ASSUMPTIONS 21
5.3 SCHEDULE 21
5.4 COST PROFILE 21

6 ALTERNATIVE 2: CLOUD SOLUTION **25**

6.1 DESCRIPTION 25
6.2 SPECIFIC ASSUMPTIONS 25
6.3 SCHEDULE 25
6.4 COST PROFILE 25

7 BENEFIT ANALYSIS **29**

7.1 TANGIBLE BENEFITS 29
7.2 BENEFIT ASSUMPTIONS 29
7.3 INTANGIBLE BENEFITS 35

8 CONCLUSIONS **37**

8.1 COMPARATIVE ANALYSIS 37
8.2 CONCLUSION 39
8.3 KEY MANAGEMENT IMPLICATIONS 41

APPENDIX A: ACRONYMS **A-1**

APPENDIX B: ALTERNATIVE 1—ON PREMISE COST NOTES **B-1**

APPENDIX C: ALTERNATIVE 2—CLOUD HOSTED COST NOTES **C-1**

APPENDIX D: KEY CONTRIBUTORS **D-1**

Executive Summary

This analysis assessed the benefits accruing to the State of Utah and its citizens from a big data investment. The following describes the purpose, constraints, alternatives to address the constraints, and key conclusions. Also assessed are the cost and benefit profiles and key management implications:

Purpose

The State has been at the forefront of using technology to better serve its citizens. From web technologies, to social media, to mobile, to open data, the State has leveraged technology for greater efficiencies, openness, and greater capabilities. Big Data is the newly emerging technology that fosters data-driven decision-making through advanced approaches to analytics and data-management. Underpinning these approaches are the need for a powerful, scalable, and cost-effective computing and storage infrastructure. This document assess the costs and benefits of providing two big data solutions – one based on-premise within the State and the second hosted by a public cloud service provider.

Growing Constraints Impede Addressing State Priorities

The following policy, technology, and personnel constraints are hindering progress in addressing state priorities.

Policy Constraints

Information sharing challenges include confusion or misperceptions around what information agencies are allowed to share. There is also a cautious, risk adverse tendency in interpreting federal privacy requirements.

Results-in-Brief

Constraints. Growing policy, technical, and personnel constraints are hindering the State's ability to achieve priorities.

- Data harboring results from security and privacy requirements inconsistently applied across agencies and organizational resistance to change.
- Throughput hindered from growing technical obsolescence and multiple independent data silos.

Solutions: Two alternatives address constraints:
Alternative 1: On-premise;
Alternative 2: Public cloud.

Benefits: Both alternatives generate $43 million in cost avoidance over 5-years.

Costs. Both alternatives have similar 5-year cost profiles.

- On-premise: $7.46 million
- Public cloud: $7.87 million

Financial Performance:

- ROI: 260%
- Payback Period: 2-years

Alternative selection: Cost is not a discriminator. Selection should be driven by initial Big Data project requirements and the State's strategic and technical priorities.

Policies addressing data sharing, personally identifiable information, security, and agency authorities may not be consistent or consistently interpreted across State agencies.

While data and information sharing is supported by the Office of the Governor, there is no legislation providing guidelines, enterprise level processes, or requiring participation. Agencies independently must interpret and determine data sharing strategies. Zero tolerance policies regarding inappropriate data access or use can result in employment termination. Collectively these issues promote data harboring.

Technology Constraints

Four technology challenges are impacting the State's capability to analyze data and apply results to address State priorities. The current environment is:

- **Inefficient as multiple discrete, "siloed" databases and warehouses are maintained.** The State maintains several hundred databases and a small number of data warehouses that are largely independent of one another.
- **Approaching technical obsolesce.** Some systems are older than 30 years, were programed in FORTRAN, and have reached or exceeded their economic and technological life. Current data center technologies are limited in the ability to process and analyze vast, diverse, and fast-streaming data.
- **Lacking sufficient technical capacity to manage an enterprise data environment.** At present the institutional capacity, governance processes, and data frameworks do not exist to fully realize enterprise level big data benefits that can positively impact executive and legislative initiatives.
- **Not fully interoperable.** Not all agencies have the IT interfaces to exchange data nor data structures that support data analyses. Across departments independent legacy systems with unique data structures make comparing and combining data difficult. One agency noted that data integrations between silos consume 80% of an analyst's time while only 20% is spent extracting value.

Personnel Constraints

Attempts to promote data sharing across departments is often hindered by institutional resistance to change. There are also concerns that an enterprise data management approach may expose poor performance. Several interviewees commented that a big data approach and automated reporting could eliminate positions and heighten employee job insecurity. This is exacerbated with the recognition that the new technologies are not well understood by staff whose formal training occurred long ago.

Alternatives to Address Constraints

To address these constraints, this assessment evaluated two alternatives:

- **Alternative 1: On-premise solution**. The on-premise alternative provides the State a pre-configured hardware system containing a cluster of racks, servers, storage, and networking devices. The vendor will drop ship these clusters to the State's datacenter where State and vendor personnel will install the clusters and perform testing. Post installation, customization, Hadoop configuration, system administration, and maintenance will be provided by the State.
- **Alternative 2: Public cloud solution**. Alternative 2 is hosted in a public cloud using a Hadoop solution that is acquired, configured, and provided as a service by the cloud vendor. Physical and virtual machine hosting is also provided by a cloud service provider. The State is responsible for fine-tuning the configuration to meet specific analytical needs when 'jobs' will be run. The cloud vender will provision compute (virtual machines), storage and networking services on its platform, install the Hadoop systems, and provide that capability as a service.

Conclusion–Select Alternative based on Non-financial Criteria

As shown below, the cost differential between each alternative is small – only $411 thousand – and each alternative delivers the same benefits. Therefore each alternative's strategic and technical strengths and weaknesses should be the decision criteria. These can only be determined by the State's operating environment the initial "quick win" projects chosen as big data candidates.

There are advantages to a local on-premises cloud alternative. Current regulations are written for an on-premise perimeter security so managing personally identifiable information can use the accepted current procedures; there is an efficiency in an always-on consistently high processing (which would incur higher charges on the cloud). While staff are developing in-depth knowledge of the new technologies, the local option provides a great transition environment.

There are clear advantages to utilizing a cloud service provider such as Amazon Public cloud. There is no capital outlay; lower cost may be achievable with long term pricing agreements and CSP price reductions. Other advantages include scalability; extensibility; ability to focus on the mission and not the machines; regular physical machine upgrades; and 24x7uptime for the physical infrastructure.

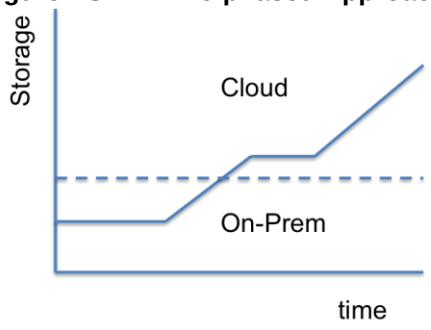
Collectively, the public cloud provides an easier long-term growth path and greater flexibility. It may have hindrances to deploying aspects of some data and solutions depending on regulatory concerns.

Time-phased strategy

Determining which alternative is superior is driven by the initial big data project requirements and any regulatory issues impacting data movement to the cloud. The quick-win scenario could provide the most immediate alternative selection decision criteria.

While a public cloud is undoubtedly the end-state or part of the end-state, policies and regulations could lead to early delays and effectively slow down the push to become operational and realize "quick wins". While the State will have to decide on the tolerance for this extra potential complexity, there are advantages to early development being with an on-premises cluster. As a simplistic representation consider ES-1.

Figure ES-1: Time-phased Approach



The initial learning system could be an on-premise approach, even one re-using existing resources, where the staff could learn and potential obtain quick results (using current perimeter security) on a small project. Then after an initial project or two, the team can expand to develop a hybrid cloud. While this adds some complexity, in actuality the cloud management tools (and Big Data platform tools) are fully capable of spanning a hybrid cloud using a single "pane of glass" control. Over time the State could begin to add more and more resources to the public cloud. Given the experience with the cost and actual performance, and the maturation of cloud usage policies and regulations, the State would be in a better position to determine their full requirements, and if indeed the full migration to the cloud would be the appropriate end-state. In either case this offers a mixed alternative for the early and late stages of the deployment, with no one-off work. This on-premise cloud could in the long-term provide a traditional perimeter security approach for PII data analysis making regulatory compliance easier.

This additional time-phased hybrid cloud option is presented to give DTS the greatest flexibility in achieving rapid value to the State.

Cost Summary – Alternative Costs are Similar

As shown in Figure ES-2, the On-premise alternative is $410,794 cheaper than the public cloud. As a result, both alternatives achieve nearly identical returns on investment, have similar net present values, and provide a 2-year payback period as shown in Figure ES-2. Figures ES-3 and ES-4 provide the cost profiles for both alternatives.

Figure ES-2: Financial Comparison

Element	Alternative 1: On-premise Solution	Alternative 2: Cloud Solution
Total Benefit	$43,303,590	$43,303,590
Total Cost	$7,461,160	$7,871,954
Net Benefit	$35,842,430	$35,431,635
Return on Investment	260%	260%
Net Present Value	$35,242,597	$34,839,266
Payback Period	2 years	2 years

Alternative 1 – On-premise requires $375,000 in capital expenditures to procure hardware and software clusters. Based on data growth assumptions, a $187,500 outlay is required in 2016 and again in 2017. Recurring costs reflect a 3-year maintenance contract procured with each cluster. Average annual recurring government and contract labor averages approximately $1.4 million per year. Average recurring total cost per year approaches $1.5 million.

Figure ES-3: Alternative 1 On-Premise Total Cost (Dollars)

Non-Recurring	2016	2017	2018	2019	2020	Total
Procure Hardware and Software	187,500	187,500				**375,000**
Provide Contractor Labor	40,000	40,000				**80,000**
Total Non-Recurring	**227,500**	**227,500**	**-**	**-**	**-**	**455,000**

Recurring	2016	2017	2018	2019	2020	Total
Procure Maintenance Contracts	50,290	50,290		50,290	50,290	**201,160**
Provide Government Labor	720,000	1,070,000	1,245,000	1,245,000	1,245,000	**5,525,000**
Provide Contractor Labor	500,000	315,000	265,000	150,000	50,000	**1,280,000**
Total Recurring	**1,270,290**	**1,435,290**	**1,510,000**	**1,445,290**	**1,345,290**	**7,006,160**
Total	**1,497,790**	**1,662,790**	**1,510,000**	**1,445,290**	**1,345,290**	**7,461,160**

Alternative 2 requires non-recurring contract labor to provide configuration support the first year. The recurring labor profile is the same as Alternative 1 per year. Total

average annual recurring costs, including cloud computing charges is approximately $1.6 million per year.

Figure ES-4: Alternative 2 Public Cloud Total Cost (Dollars)

WBS	Non-Recurring	2016	2017	2018	2019	2020	Total
1.1	Procure Hardware and Software						-
1.2	Provide Contractor Labor	48,000					**48,000**
	Total Non-Recurring	**48,000**	-	-	-	-	**48,000**

WBS	Recurring	2016	2017	2018	2019	2020	Total
2.1	Procure Cloud Computing	263,084	118,360	110,540	343,956	183,014	**1,018,954**
2.2	Provide Government Labor	720,000	1,070,000	1,245,000	1,245,000	1,245,000	**5,525,000**
2.3	Provide Contractor Labor	500,000	315,000	265,000	150,000	50,000	**1,280,000**
	Total Recurring	**1,483,084**	**1,503,360**	**1,620,540**	**1,738,956**	**1,478,014**	**7,823,954**
	Total	**1,531,084**	**1,503,360**	**1,620,540**	**1,738,956**	**1,478,014**	**7,871,954**

Benefits Exceed Costs

As shown in Figure ES-5, tangible, monetized benefits accrue in five categories. After 2019, approximately $16 million in benefits are anticipated. The greatest benefit is derived from decreasing tax fraud.

Figure ES-5: Benefit Summary

Category	2016	2017	2018	2019	2020	Total
Reduced Recidivism		455,700	911,400	1,822,800	1,822,800	**5,012,700**
Decreased Tax Fraud		1,986,519	3,973,037	7,946,075	7,946,075	**21,851,706**
Decreased Food Stamp Fraud		567,216	1,134,432	2,457,936	2,457,936	**6,617,520**
Decreased Child Support Fraud		46,321	92,642	200,725	200,725	**540,414**
Decreased Medicaid Fraud		843,750	1,687,500	3,375,000	3,375,000	**9,281,250**
Total	-	**3,899,506**	**7,799,012**	**15,802,536**	**15,802,536**	**43,303,590**

Intangible benefits are expected to help address key policy, technical, and personnel gaps. Most importantly, improved data-driven decision making provides:

- The ability to provide critical insights into complex issues
- A focus on evidence-based decision-making
- The identification of root causes that can be addressed

This in-turn helps the State address the following constraints:

Policy

- Improved policy decision making and resource allocation.
- Enterprise focus will increase throughput and service delivery to customers.
- Integrated policies for data sharing, security, privacy, governance, and management

Technology

- Infrastructure savings from eliminating silod data operations.
- Advanced data analytic and sharing infrastructure
- Interoperability across agencies will promote data sharing and analytics

Personnel

- Improved skill set alignment to support and manage a data driven enterprise

Management Implications

In implementing a solution, management should:

- **Assess and prioritize candidate Big Data programs.** This may influence the technical and acquisition strategy adopted by the State
- **Integrate lowest level of data.** To achieve full benefits, the lowest level of data must be integrated and made available.
- **Monitor Data growth**. If data growth is predicted to exceed 500 TB, then additional capacity will be needed.
- **Monthly Data Transmission Rate**. Should the value exceed 10 TB then the pricing assumptions should be reviewed.
- **Monitor usage patterns.** Usage statistics will provide insight into determining the optimal procurement strategy for acquiring additional capability for either alternative

1 Introduction and Purpose

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 4 Document: Big Data Business Cases assess the strategic, technical, and financial merits of two alternatives to provide a Big Data solution for the State. The companion documents in this study are Task 1 Policy and Governance, Task 2 People - Organization and Skills, Task 3 Technology Roadmap and Task 5 Process - Generating Value.

The following describes key business drivers and challenges and issues that a big data investment can help address.

1.1 Purpose

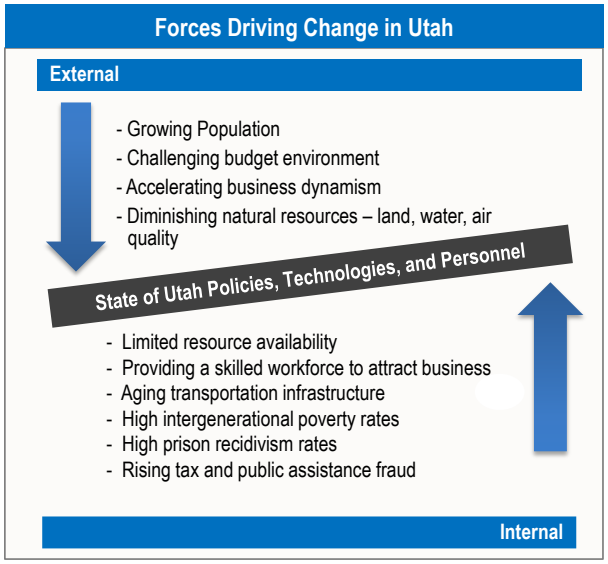
The State has been at the forefront of using technology to better serve the citizens of Utah. From web technologies, to social media, to mobile, to open data, the State has leveraged technology for greater efficiencies, openness, and greater capabilities. Big Data is the newly emerging technology that fosters data-driven decision-making through advanced approaches to analytics and data-management. Underpinning these approaches are the need for a powerful, scalable, and cost-effective computing and storage infrastructure. As opposed to the traditional method of buying ever more powerful monolithic commercial platforms, the big data paradigm moves compute and data systems to operate on a cluster of cheap commodity servers located either on-premise at the state data center or at a cloud provider. The question before the state of Utah is whether this cluster technology would provide cost-effective benefits to the State in order to enhance analytics to decision support. The purpose of this Business Analysis is to determine the cost to the State to develop Big Data capabilities, and the benefits that would accrue from being data-driven.

1.2 Business Drivers

As shown in Figure 1, external and internal constraints impact the State's policy, technology, and personnel decisions. Population growth is predicted to almost double to 5.8 million by 2050[1] and budgetary resource needs are growing. The State seeks to create an inviting climate to conduct business and live by leveraging its natural resources, improving air quality, and boost the economy by providing a skilled workforce.

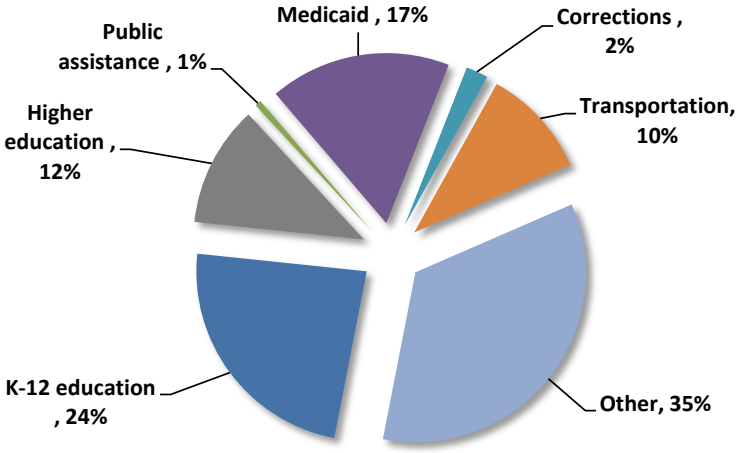
[1] Governor's Office of Planning and Budget 2012 Baseline Projections

Figure 1: External and Internal Constraints Driving Change



As shown in Figure 2, Medicaid, higher education, and transportation consume the most state resources. Note the "Other" category includes additional benefit programs.

Figure 2: Proportion of State Spending by Function[2,3]



[2] National Association of State Budget Officers (NASBO) State Expenditure Report FY 2012-2014 State Expenditure Report FY 2012-2014 http://www.nasbo.org/sites/default/files/State%20Expenditure%20Report%20%28Fiscal%202012-2014%29S.pdf

[3] "Other" expenditures include "Children's Health Insurance Program (CHIP), institutional and community care for the mentally ill and developmentally disabled, public health programs, employer contributions to pensions and health benefits, economic development, environmental projects, state police, parks and recreation, housing and general aid to local governments."

Efforts to more accurately identify and address underlying factors that contribute to rising health, education, public assistance, Medicaid, corrections, and transportation costs will benefit citizens and advance State priorities. Analytics that can identify relationships across programs can assist policy makers and administrator's better address root causes, more efficiently deliver services, and balance public and private sector interests. The proposed big data investment's purpose is to provide the technology infrastructure to achieve those ends.

The following sections assess the gap that must be addressed to implement a big data solution. Differing technical approaches are evaluated, and the costs and benefits and return on investment (ROI) for two alternatives are estimated.

2 Gap Analyses

The following summarizes the policy, technology, and personnel constraints impeding the State of Utah's ability to implement big data solutions to achieve State and departmental goals. The "to-be" big data environment is then presented.

2.1 "As-Is" Environment

The following baseline description reflects interviews with representatives from six Utah organizations and one user group[4].

- Automated Geographic Reference Center (AGRC)
- COGNOS User Group
- Commission on Criminal and Juvenile Justice
- Department of Natural Resources (DNS)
- Department of Technology Services (DTS)
- Department of Workforce Services (DWS)
- State office of Education

These interviews surfaced four common themes:

- **Sub-optimal policy decision making and resource allocation.** At the executive and legislative branch levels, priorities and resource allocation decisions are made with incomplete insight into relationships across agencies and programs. This impacts the ability to fully achieve program goals and prioritize budgets accordingly.
  Sub-optimal policy and resource allocation decisions also delay advances in reducing intergenerational poverty, recidivism, and fraud. This also negatively impacts the ability to balance natural resource and commercial interests, attract new business, and improve air quality.
- **Departmental, not enterprise focus, limits ability to achieve savings, and benefit customers.** Mission convergence is limited and some divisions wish to retain autonomy. Many departments are not looking to partner or leverage activities across the enterprise with other departments to improve throughput and service delivery.
- **Quantifying benefits is challenging but improvements to resident's lives and business community is expected.** All interviewed expect improved data analytics from varied sources will benefit their operations and customers

[4] Interviews were arranged through DTS and performed between September 2nd and September 9th, 2015.

directly and counties and townships indirectly. Most interviewees noted that until implementation, program relationships will not be understood but expected positive benefits. For example, correlations between human and health services and public safety will be revealed which will positively impact recidivism and intergenerational poverty rates.

- **Security and Privacy controls must be designed into solution.** Mechanisms to safeguard personally identifiable information (PII), credit card data, juvenile, and other sensitive data must be designed into the enterprise data solution. Uncertainty regarding privacy requirements limits data sharing.

The following describes the policy, technical, and personnel constraints in the current environment.

2.1.1 Policy Constraints

Multiple factors limit the value Utah could derive from the data it already has. DTS supports two data warehouses for the Division of Finance that includes statewide financial information which is used extensively across the enterprise. DTS also maintains data warehouse environments for: Department of Workforce Services (employment, labor, and economic data), Human Services, Health, Corrections, UDOT (transportation), and Tax (revenue).

Information sharing challenges include confusion or misperceptions around what information agencies are allowed to share. There is also a cautious, risk adverse tendency in interpreting federal privacy requirements. Policies addressing data sharing, personally identifiable information (PII), security, and agency authorities may not be consistent or consistently interpreted across State agencies.

Agencies report that they rely on legal review or analysis before agreeing to share data. This can be time consuming and costly. In a Government Accountability Office (GAO) report[5], GAO interviewed stakeholders from several states, including the Utah Department of Workforce Services. The report indicated an agency's legal counsel may advise against sharing data as a precaution rather than due to an explicit prohibition. Stakeholders also reported that potential data sharing inconsistencies in federal privacy requirements across programs are challenging. Some child welfare, education, and health information privacy protection policies require informed consent which creates additional compliance and data exchange challenges.

[5] GAO-13-106: Published: Feb 8, 2013. Publicly Released: Feb 21, 2013. http://www.gao.gov/products/GAO-13-106

While data and information sharing is supported by the Office of the Governor, there is no legislation providing guidelines, enterprise level processes, or requiring participation. Multiple data sharing agreements are in place between specific agencies to outline exactly what data will be shared, with whom, how, and for what purpose. There is currently no coordinated, multiagency guidance that clarifies what data sharing is permissible. Agencies or, in some cases individual agency employees, independently must interpret and determine appropriate information or data sharing. Zero tolerance policies regarding inappropriate use of or access to data resulting in termination of employment are strong disincentives to sharing information. Collectively these issues promote data harboring.

2.1.2 Technology Constraints

Four technology challenges are impacting the State's capability to use analyze data and apply results to address State priorities. The current environment is:

- **Inefficient as multiple discrete, siloed databases and warehouses are maintained.** The State maintains several hundred databases and a small number of data warehouses that are largely independent of one another. These include data warehouses in the Division of Finance, the Department of Workforce Services, the Department of Human Services, the Department of Health, and others.
- **Approaching technical obsolesce.** Some systems are older than 30 years, were programed in FORTRAN, and have reached or exceeded their economic and technological life. Some departments rely on relational database technologies that cannot process and analyze large and diverse data and high velocity data streams. Last generation business analytic tools such as IBM Cognos are still in use. Current data center technologies are limited in the ability to process and analyze vast, diverse, and fast-streaming data.
- **Lacking sufficient technical capacity to manage an enterprise data environment.** At present the institutional capacity, governance processes, and data frameworks do not exist to fully realize enterprise level big data benefits that can positively impact executive and legislative initiatives.
- **Not fully interoperable.** Not all agencies have the IT interfaces to exchange data nor data structures that support data analyses. Across departments independent legacy systems with unique data structures make comparing and combining data difficult. Further disparate data structures require manual data conversion. One agency noted that data integrations between silos consume 80% of an analyst's time while only 20% is spent extracting value.

2.1.3 Personnel Constraints

Attempts to promote data sharing across departments are often hindered by institutional resistance to change. There are also concerns that an enterprise data management approach may expose poor performance. Several interviewees commented that a big data approach and automated reporting could eliminate positions and heighten employee job insecurity.

2.2 "To-be" Environment

The target environment is to provide integrated policies, technologies, and personnel that support a data driven government.

2.2.1 Policy

The vision is to make data shared and accessible by default within its data lake. All State agencies will participate and share data to effectively and efficiently and develop analytics and business intelligence capabilities. Agency cross coordination will reduce unnecessary information silos and technology barriers. Collectively this will improve executive and legislative decision-making and resource allocations while improving agency service delivery throughput.

Legislation will provide the necessary funding and authorities, to establish the data lake and analytics within the Governor's Office of Management and Budget. A Chief Data Officer within DTS will be responsible for oversight and management of the data lake and analytics program. There will be a Data Strategy and Governance program outlining the roles and responsibilities for secure data sharing and management. The Chief Data Officer will establish a governance board composed of representatives from each agency to coordinate data requirements, ensure implementation is supported across state agencies and facilitate sharing through Partnership Agreements and data sharing agreements. Legal Counsel will coordinate with agencies to review policies and evaluate the feasibility of sharing sensitive data elements and data sets.

Data Security and management policies will establish the controls, processes, procedures, and associated tools to 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. Periodic audits will ensure compliance with privacy and security guidelines. The agency collecting the data will be the data owner responsible for coordination with citizens and managing privacy and consent as well as appropriate access, correction, and redress regarding use of PII. Data owners will 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.

2.2.2 Technology

Being data-driven means the State has the tools to bring a variety of data sources together to address specific problems from a state-wide perspective. Analytics requires the individual records, and not the aggregates in order to do a full analysis of a given problem, such as inter-generational poverty, or college drop-out rates. Bringing together the data from a range of state services provides a view into the problem that no single agency can provide. Handling this much data requires technology that can scale to handle the data, while remaining affordable in cost. The new Big Data technologies provide this.

A Big Data platform resides on a cloud-based infrastructure. It can be located on-premises, or can be hosted by a cloud provider such as AWS. It uses management software to control an infrastructure that creates virtual machines (VMs) on top of the physical hardware. On top of those VMs resides the software that hosts the data, creates database access, and provides workflow. On top of that resides the analytics capability, for the data scientist doing deep analysis of the blended datasets, to the easy access tools for business analysts.

2.2.3 Personnel

New technologies require new skills. For the infrastructure, an administrator needs to learn about virtualization technologies to become a Cloud Engineer. They need to learn about the installation of the Apache Big Data Stack for installing the data platform to become a Big Data Engineer. Security specialists need to learn about the relevant security policies for hosting data in a cloud-based environment. Analysts need to work with the Big Data Engineer to learn how the data will be stored, and work out how the data should be distributed across the nodes in the cluster, essentially learning about their “instrument” in order to be a Data Scientist. Analysts also need to learn the new software tools that allow full-dataset analytics. The Analyst and Big Data Engineer need to work with the Business Analysts to determine how best they can access the data they need, and provide easy access analytical tools. Leadership needs to promote the vision for becoming a data-driven organization, and work to establish the culture change needed to ensure that decision makers first look at ways to derive information from data before making their decisions.

3 Global Assumptions and Methodology

This section describes the global strategic, technical, and financial assumptions. These include a combination of assumptions, facts, and requirements used to guide the analysis. The methodology applied when preparing the business case concludes the section.

3.1 Assumptions

The following global assumptions are categorized as strategic, technical, financial, or schedule

3.1.1 Strategic

Figure 3 and Figure 4 identify "In-Scope" and "Out-of-Scope" functions and activities.

Figure 3: "In-Scope" Functions and Activities

In-Scope Functions and Activities
Hardware and Software procurement and installation

Figure 4: "Out-of-Scope" Functions and Activities

Out-of-Scope Functions and Activities
Data collection, formatting, and management
Security applications – these currently exist in the data warehouse

3.1.2 Technical

The technical assumptions listed below are based on the functional needs (performance and operational storage) rather than specific technical requirements. These characteristics were chosen to meet initial needs for analytical capacity and to provide a clear comparative analysis between on-premises and cloud-based clusters:

- Management nodes = 1
- Number of Name/Job Tracker nodes = 3
- Operational Data Growth: 20% annual compound growth
- Initial Operational data capacity = 150 TB total
- Target Headroom (unused space) per Node = 1 TB total
- Max Storage per Node = 24TB total
- High availability is required

- Use Open Source Operating System
- Existing security tools will be used for managing the Hadoop system security

3.1.3 Financial

- **Fiscal Year.** The Utah State fiscal year (FY) begins July 1 and ends June 30. All years shown are Utah State fiscal years, unless otherwise stated.
- **Investment Horizon.** The investment analysis spans 5 years from the initial installation (FY 2016 through FY 2020).
- **Sunk costs.** Sunk costs are any costs incurred prior to FY 2016. Sunk costs are excluded from financial metric calculations.
- **Schedule.** All alternatives will be implemented in FY 2016.
- **Technology refresh rate.** Hardware and software will be refreshed every five years. As the investment horizon is five years, no refresh cost are included.
- **Inflation.** Inflated dollars are not applied and should only be used for budgeting purposes.
- **Discount rate.** A 0.4% real discount rate is applied to calculate each alternative's present value[6].
- **Labor categories and annual salary.** Labor costs shown in Figure 5 are based on estimates contained in the 2015 Big Data Proposal.

Figure 5: Labor Cost Assumptions

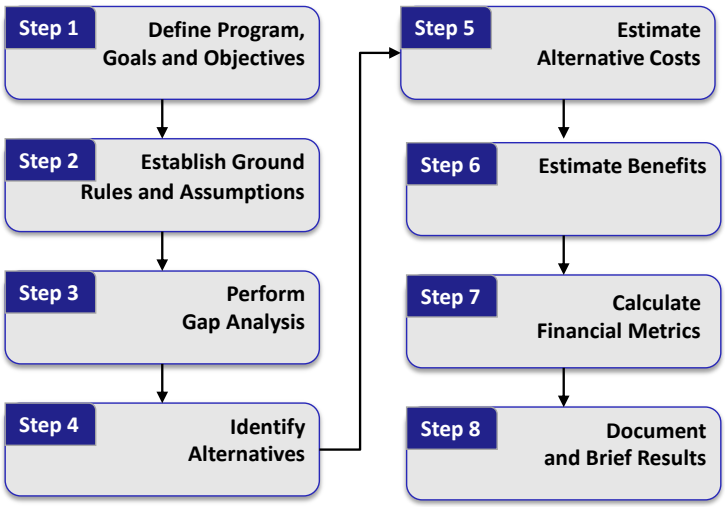
		Year 1	Year 2	Year 3	Year 4	Year 5
Category	Rate	FTEs	FTEs	FTEs	FTEs	FTEs
Hadoop Administrator	$150,000	1	1	1	1	1
Data Scientist	$175,000	1	3	4	4	4
Technical Support	$135,000	2	2	2	2	2
Data Security Administrator	$125,000	1	1	1	1	1

[6] https://www.whitehouse.gov/omb/circulars_a094/a94_appx-c. 5-year real rate.

3.2 Methodology

Figure 6 summarizes the 8-step methodology applied in preparing the business case.

Figure 6: Cost Benefit Methodology



- **Step 1: Define Program Goals and Objectives.** Big Data program goals and objectives were defined based on discussions with DTS representatives.
- **Step 2: Establish Ground Rules and Assumptions.** A project team was formed consisting of State and contractor representatives. The team developed and documented assumptions and other variables impacting the analysis. This step includes a table summarizing what is "in-scope" and "out-of-scope" and strategic, technical, and financial assumptions.
- **Step 3: Perform Gap Analysis.** Based on the program's mission and goals, the "As-Is" environment and capability gaps were documented to describe the objectives each alternative must achieve.
- **Step 4: Identify Alternatives.** Alternatives designed to close the gaps described in Step 3 are assessed. All reasonable alternatives were identified and described. Next, criteria were developed to "down-select" to a viable, mutually exclusive few for further analyses.
- **Step 5: Estimate Alternative Costs.** For both alternatives 5-year life-cycle cost estimates were prepared.
- **Step 6: Estimate Benefits.** Tangible and intangible program impacts were assessed. Where possible impacts were quantified then monetized.
- **Step 7: Calculate Financial Metrics.** The incremental time-phased costs and benefits were used to calculate each alternative's net present value, return on investment, and payback period.
- **Step 8: Document and Brief Results.** Based on this methodology's results, alternatives are ranked based on their strategic, technical, and financial

merits. Supporting appendices detailing each alternative's cost and benefit estimates and other supporting information are provided.

4 Alternatives

The following sections describe the technical alternatives evaluated and then down-selects to two alternatives for further cost and benefit analyses.

4.1 Alternative Descriptions

This section describes the on-premise and cloud hosting alternatives and the technology options within each. The section concludes by applying down-select criteria to identify two viable alternatives for further analyses.

4.1.1 On-premise Hosting

On-premise hosting options all involve making capital purchases in 2016 and 2017 (with annual maintenance). Capacity is fully available upon installation, is fixed based on those purchases, and the infrastructure are sunk costs regardless of utilization. Options include:

- **Custom Hadoop Cluster.** This involves buying servers, storage, and networking hardware/software; installing Hadoop software; and building a cluster within Utah's existing datacenter.
- **Hadoop Converged Infrastructure**. Involves buying pre-configured hardware system (rack of servers, storage, and networking); installing Hadoop systems; and "standing up" cluster within Utah's existing datacenter.
- **Big Database Appliance**: High-end machines are procured and tuned to handle very large databases using traditional technologies.

4.1.2 Cloud-based hosting

A secure network connection serves as the link between Utah and the cloud resources. Cloud services generally incur costs through usage on a pay-as-you-go basis. When compute jobs are requested, the number of compute servers in use can be scaled up temporarily so that the analysis is completed within a specified time. When the demand is over, the cost returns to the baseline maintained to ensure ready access. The storage and platform services will remain in place at all times.

Cloud hosting will be performed either in a:

- **Custom Hadoop Cluster**: This involves provisioning compute (virtual machines), storage and networking services on a cloud provider's platform, installing Hadoop systems, and building a Hadoop cluster.
- **Hadoop as a Service.** This involves provisioning and configuring a Hadoop service on a cloud provider's platform. This could include Amazon Elastic MapReduce (Amazon EMR) or Google Click to Deploy.

The primary difference in these options whether Utah staff acquires, installs, configure and maintain the Hadoop software (first option) or whether a turnkey service is provided by the cloud vendor (option two).

4.2 Platform Options

- **Commercial Big Data Appliance.** Involves either buying a bundling of hardware and software into an appliance, or licensing the software and implementing on recommended commodity hardware.
- **Open Source Big Data Platform.** Involves either buying a bundling of hardware and software into an appliance, or licensing the software and implementing on commodity hardware or on a public cloud infrastructure
- **Public Cloud Big Data Platform**. Involves leveraging cloud provider specific tools

4.3 Comparison

For on-premise and cloud-bases hosting options, Figure 7 and Figure 8 compare the advantages and disadvantages of each.

Figure 7: On-premise Options

Alternative and Option	Advantages	Disadvantages
Custom Hadoop Cluster	• Provides DTS with traditional role of selecting, purchasing, installing, configuring and maintaining this hardware infrastructure. ▪ Offers DTS full responsibility for implementing hardware-level security ▪ Dynamic access to data may be simpler than a cloud solution	▪ A fully-funded resource with fixed capacity ▪ May require additional staff for operations and maintenance ▪ Requires periodic technology refreshes ▪ Requires facility space and infrastructure ▪ Difficult to change configuration after purchase ▪ May require FedRamp accreditation
Hadoop Converged Infrastructure	▪ Provides DTS with traditional role of selecting, purchasing, installing, and maintaining this hardware infrastructure. ▪ Pre-configuration reduces time to use machines somewhat ▪ Offers DTS full responsibility for implementing hardware-level security	▪ Requires facility space and infrastructure ▪ May have special facility requirements ▪ Difficult to change configuration after purchase ▪ May require FedRamp accreditation

Alternative and Option	Advantages	Disadvantages
Big Database Appliance	▪ Provides DTS with traditional role of selecting, purchasing, installing, and maintaining this single-purpose hardware infrastructure. ▪ Ability to use familiar database technologies ▪ Possible to use existing database licenses	▪ A fully-funded resource with fixed capacity ▪ May require additional staff for operations and maintenance ▪ Typically very expensive ▪ May require FedRamp accreditation

Figure 8: Cloud-based Hosting Options

Alternative and Option	Advantages	Disadvantages
Custom Hadoop Cluster	▪ Offers flexibility to state in establishing an analytics function due to reduced sunk cost. ▪ Ability to scale up and down to accommodate surging demand ▪ Minimizes sunk costs; clusters are used only when needed ▪ Try/pilot multiple configurations to identify an optimal configuration ▪ Access "Hadoop as a Service" tools which are purchased and maintained by cloud vendor ▪ Reduces the need for new IT staff ▪ Focus on core business by outsourcing non-core IT functions ▪ Access to FedRAMP certified providers, such as AWS GovCloud	▪ Somewhat less control over internal configuration of servers ▪ Likely requires secure, dedicated network connection to cloud vendor ▪ Requires Utah to acquire, install, configure and maintain Hadoop software that will run on cloud vendor's servers. ▪ May require additional engineering to take advantage of the cloud's benefits; such as maintaining certain data in cloud storage to make processing efficient.
Hadoop as a Service	▪ Offers the least sunk costs and the most flexibility for the state to establish and test the value of an analytics program ▪ Relies on cloud vendor investment in acquisition, installation, configuration, and maintenance of the Hadoop	▪ Less control over internal configuration of servers ▪ Likely requires secure, dedicated network connection to cloud vendor ▪ May require additional engineering expense to maximize cloud benefits

Alternative and Option	Advantages	Disadvantages
	software. ▪ Fastest start up time since the Hadoop infrastructure is immediately available ▪ Reduces the need for state Hadoop configuration specialists ▪ Improves ability to scale clusters based on demand	maintaining certain data in cloud storage to make processing efficient. ▪ Less control over the core Hadoop software

4.3.1 Options Down-select

The integrated project team identified the following requirements against which options within each alternative were assessed on a five-point Reichert scale where:

1 = Unlikely to achieve requirements
2 = Could achieve the requirement with difficulty
3 = Will likely achieve the requirement
4 = Will achieve the requirement
5 = Will meet or exceed the requirement

The requirements that were scored for each alternative's option are:

- Minimize installation complexity
- Minimize on-going configuration ease
- Maximize local security management
- Other

The IPT scored each requirement and the results are depicted in Figure 9. As shown the Hadoop Converted Infrastructure best achieved the requirements.

Figure 9: Ability to Achieve Requirement (1 = least capable, 5 = most capable)

Alternative and Option	Minimize Installation Complexity	Minimize On-going Configuration Ease	Maximize Local Security Management	Total
Custom Hadoop Cluster	3	3	5	**11**
Hadoop Converged Infrastructure	5	3	5	**13**
Big Database Appliance	2	3	5	**10**
Supercomputer	1	1	5	**7**
Managed Services	1	4	4	**9**

Alternative and Option	Minimize Installation Complexity	Minimize On-going Configuration Ease	Maximize Local Security Management	Total
contract				

The same approach was applied to the Cloud Hosting alternative as shown in Figure 10 with a Custom Hadoop Cluster proving the most attractive.

Figure 10: Ability to Achieve Requirements (1 = least capable, 5 = most capable)

Alternative and Option	Minimize Installation Complexity	Minimize On-going Configuration Ease	Maximize Local Security Management	Total
Custom Hadoop Cluster	3	3	4	**10**
Hadoop as a Service	5	5	3	**13**

4.4 Alternatives Chosen for Further Evaluation

Based on the requirements the two alternatives chosen for further analyses are:

- Alternative 1: Hadoop Converged On-premise Infrastructure (referred to as On-premise for the rest of the analysis)
- Alternative 2: Cloud hosted Hadoop as a Service (referred to the Public cloud solution for the rest of the analyses)

5 Alternative 1: On-premise Solution

This section describes alternative 1, specific assumptions, summarizes each activity necessary to deploy and sustain the solution, and provides a cost profile.

5.1 Description

The Hadoop-converged on-premise alternative provides the State a pre-configured hardware system containing a cluster of racks, servers, storage, and networking devices. The state procurement system will seek quotes based on DTS specifications. After receiving the required number of bids, a selection will be made. The selected vendor will drop ship these clusters to the State's datacenter where State and vendor personnel will install the clusters and perform testing. Post installation, customization, Hadoop configuration, system administration, and maintenance will be provided by the State.

5.2 Specific Assumptions

- The following items are excluded from the analyses as it is assumed the host data center will provide:
  - Datacenter physical infrastructure elements including floor space, power, and cooling
  - Edge Firewall Security and provide all network infrastructure needs
  - Disaster recovery and continuity of operations
- FedRamp and other compliance costs are excluded.
- Sufficient funding to procure, implement, and maintain the alternative

5.3 Schedule

Implementation will occur in FY 2016. Once a purchase order is issued, approximately 30-45 days will elapse before system delivery. Installation through final acceptance testing is estimated to require four weeks. The system will be operational within FY 2016.

5.4 Cost Profile

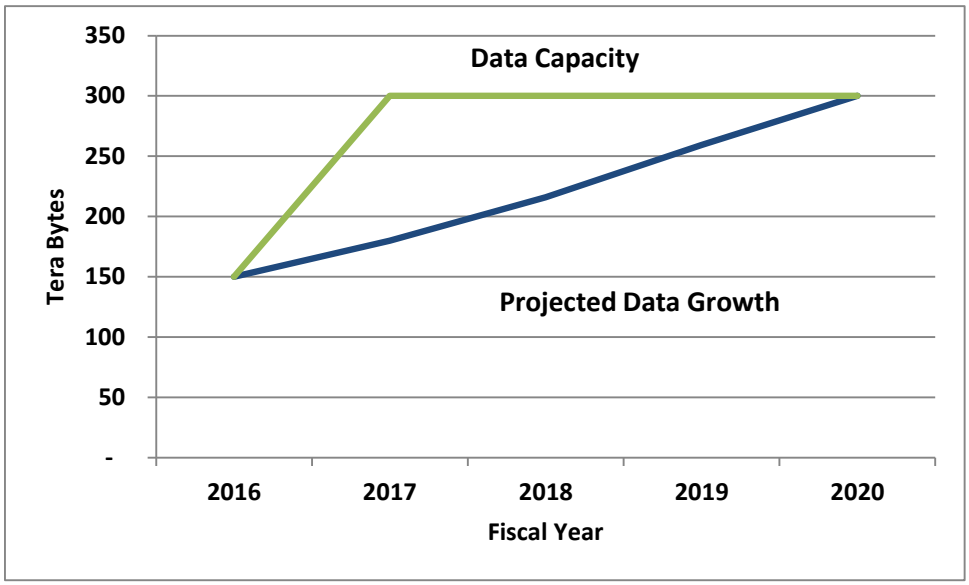
Activities and costs are categorized between non-recurring and recurring costs.

5.4.1 Non-recurring Costs

The following non-recurring costs will be incurred during FY 2016.

- **Procure Hardware.** This involves procuring converged infrastructure[7]. Each pre-configured cluster will contain server racks, storage, and networking devices. The Hadoop system will also be installed and the entire cluster tested prior to drop shipment.
  1 cluster will be procured in 2016 and another in 2017. Each cluster provides 150 TB capacity This provides sufficient capacity to accommodate an annual 20% compound data growth beginning with 150 TB in 2016 and growing to 311 TB in 2020 as shown in Figure 11.

Figure 11: Data Growth and Alternative 1 Data Capacity



- **Provide Contract Labor.** This represents labor to implement the system at the host datacenter. Two full time equivalent (FTE) are needed to install the hardware, install the operation system, and configure networking and security policies.

5.4.2 Recurring Costs

The following activities represent costs to maintain the alternative through 2020.

- **Procure maintenance contracts.** The maintenance contract provides onsite support by a vendor authorized service technician. The vendor will provide supported parts and materials necessary to maintain the covered hardware

[7] SuperMicro corporation's Bare Metal 42U cluster

equipment in operating condition. If a parts has been determined to be defective, the vendor representative will be onsite and replace the defective parts to enable proper operation of the hardware products. For this analysis we priced the 4-Hour Onsite Response option, which guarantees that a vendor authorized representative will arrive at the customer's site to begin hardware maintenance service within 4 hours after the service request has been received and spare parts have been received onsite.

The maintenance contracts for each cluster are paid in advance and cover three years. When each cluster is purchased in 2016 and 2017, a maintenance contract is also acquired. In 2019 and 2020 each maintenance contract will be renewed.

- **Provide government labor.** Based on earlier research[8] the following skill sets will be needed by the State. These represent an addition to the existing technical staff. The FTE number is shown in Figure 12:
  - Hadoop Administrator or Big Data Engineer
  - Data Scientists
  - Technical Support
  - Data Security Administrator

Figure 12: Annual Labor Categories and Number of FTE

Category	2016	2017	2018	2019	2020
Hadoop Administrator	1	1	1	1	1
Data Scientist	1	3	4	4	4
Technical Support	2	2	2	2	2
Data Security Administrator	1	1	1	1	1
Total Government Labor	**6**	**7**	**8**	**8**	**8**

- **Provide contract labor.** Contract labor will provide consulting support to efficiently move DTS to self-sufficiency. State personnel will initially to work with experts to architect the Big Data system to meet their needs going forward. This would entail working with Cloud and Security Engineers on the configuration of the infrastructure; with Big Data Engineers on setting up and maintaining the software platform including the Hadoop stack; and with Data Scientists on establishing a collaborative analytics program with business analysts across the state. This contract support would be expected to

[8] Big Data Whitepaper, 2015

diminish over time as the State gains experience and establishes an operational Data-Driven Decision Support program.

5.4.3 Cost Summary

Figure 13 shows the non-recurring, recurring, and total annual costs to procure and maintain the alternative through 2020. In 2016 and 2017 the clusters are purchased and installed the three year maintenance contracts are procured.

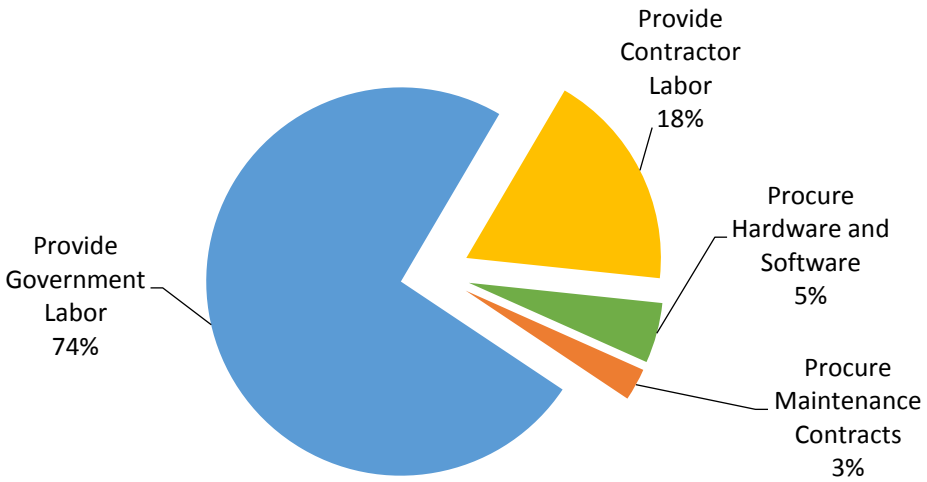
Figure 13: Alternative 1: On-Premise Total Cost (Dollars)

Non-Recurring	2016	2017	2018	2019	2020	Total
Procure Hardware and Software	187,500	187,500				**375,000**
Provide Contractor Labor	40,000	40,000				**80,000**
Total Non-Recurring	**227,500**	**227,500**	**-**	**-**	**-**	**455,000**

Recurring	2016	2017	2018	2019	2020	Total
Procure Maintenance Contracts	50,290	50,290		50,290	50,290	**201,160**
Provide Government Labor	720,000	1,070,000	1,245,000	1,245,000	1,245,000	**5,525,000**
Provide Contractor Labor	500,000	315,000	265,000	150,000	50,000	**1,280,000**
Total Recurring	**1,270,290**	**1,435,290**	**1,510,000**	**1,445,290**	**1,345,290**	**7,006,160**
Total	**1,497,790**	**1,662,790**	**1,510,000**	**1,445,290**	**1,345,290**	**7,461,160**

As shown in Figure 14, Government labor represents the majority of the investment's expense. This represents the Hadoop Administrator, Data Scientist, Technical Support, and Data Security Administrators.

Figure 14: Proportion of Cost by Category



6 Alternative 2: Public Cloud Solution

This section describes alternative 2, specific assumptions, the WBS, and schedule. The section concludes with a summary cost profile and analysis. Appendix D contains supporting cost notes.

6.1 Description

Alternative 2 is a public cloud based on using a Hadoop solution that is acquired, configured, and provided as a service by the cloud vendor. The State of Utah is responsible for fine-tuning the configuration to meet specific analytical needs when 'jobs' will be run. Hosting is provided by a cloud service provider. This involves provisioning compute (virtual machines), storage and networking services on a cloud provider's platform, installing the Hadoop systems, providing that capability as a service.

6.2 Specific Assumptions

- **Monthly Data Transfer Rate.** 10 TB of data are assumed to transfer monthly
- **Data Security.** One copy of core data in the Hadoop cluster, whereas another copy of the data is stored in persistent S3 storage

6.3 Schedule

Implementation will occur in FY 2016.

6.4 Cost Profile

Activities and costs are categorized between non-recurring and recurring costs.

6.4.1 Non-recurring Costs

Contract labor is required to assist the state in establishing the cloud environment. The contractors will perform Initial design, setup and configuration of the Hadoop cluster, including setting up Management and Name Tracker nodes; ingesting data from identified sources using Extract, Transform, Load (ETL) tools and techniques; formatting the data in AWS S3 data stores (buckets); and configuring Hadoop to use S3 data stores for processing analytics.

6.4.2 Recurring Costs

The following activities represent costs to maintain the alternative through 2020.

- **Procure Cloud Computing.** To provide the performance characteristics required the following initial set of virtual machines was priced on three year reserve basis paid for upfront and storage is priced on a per month basis at a fixed rate throughout the year (no ramp up):

- Management Node: 1x r3.4xlarge
- Name/Job Tracker Node: 3x r3.4xlarge
- Slave/Data Nodes: 7x d2.4xlarge
- S3 Storage: 150 TB
- AWS Data Transfer: 10TB

To account for the 20% compound growth year over year the following was added in each of the years as listed. Note that the increase is incremental, i.e., an additional Slave/Data node will be brought online when capacity of the units in use is exceed, even if only a small portion of the new node is utilized in the year that it is brought online.

- **Year 2:** Add two d2.4xlarge and storage increased to 180TB
- **Year 3:** Add one d2.4xlarge and storage increased to 216TB
- **Year 4:** Renew 3 year reserve contract for initial purchase and add two d2.4xlarge for a total of 9 d2.4xlarge paid for and increase storage to 259TB
- **Year 5:** Renew 3 year reserve contract for virtual machines added in year two and add three d2.4xlarge for a total of 5 d2.4xlarge paid for and increase storage to 311TB

- **Provide government labor.** The same labor profile in Alternative 1 is required for this alternative. These represent an addition to the existing technical staff. The FTE number is shown in Figure 15:

Figure 15: Annual Labor Categories and Number of FTE

Category	2016	2017	2018	2019	2020
Hadoop Administrator	1	1	1	1	1
Data Scientist	1	3	4	4	4
Technical Support	2	2	2	2	2
Data Security Administrator	1	1	1	1	1
Total Government Labor	**6**	**7**	**8**	**8**	**8**

6.4.3 Cost Summary

In Figure 16 total non-recurring and recurring costs are depicted. In FY 2016 contract labor assists the state in preparing for the cloud environment.

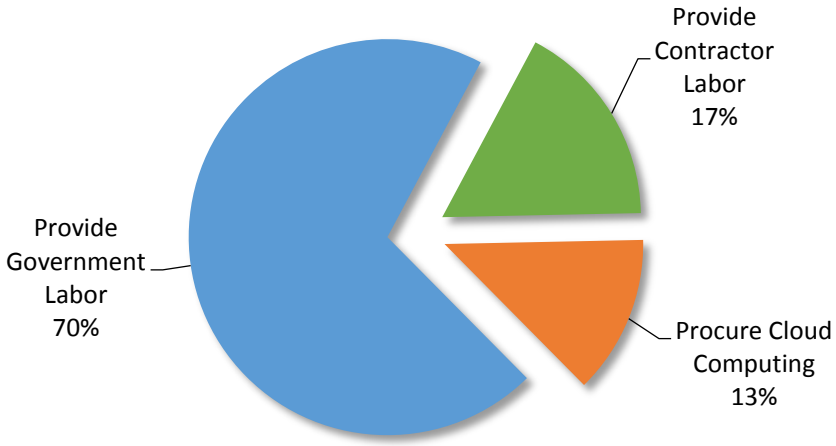
Figure 16: Alternative 2: Cloud Solution Total Cost (Dollars)

WBS	Non-Recurring	2016	2017	2018	2019	2020	Total
1.1	Procure Hardware and Software						-
1.2	Provide Contractor Labor	48,000					**48,000**
	Total Non-Recurring	**48,000**	-	-	-	-	**48,000**

WBS	Recurring	2016	2017	2018	2019	2020	Total
2.1	Procure Cloud Computing	263,084	118,360	110,540	343,956	183,014	**1,018,954**
2.2	Provide Government Labor	720,000	1,070,000	1,245,000	1,245,000	1,245,000	**5,525,000**
2.3	Provide Contractor Labor	500,000	315,000	265,000	150,000	50,000	**1,280,000**
	Total Recurring	**1,483,084**	**1,503,360**	**1,620,540**	**1,738,956**	**1,478,014**	**7,823,954**
	Total	**1,531,084**	**1,503,360**	**1,620,540**	**1,738,956**	**1,478,014**	**7,871,954**

As shown in Figure 17, Government labor represents the majority of the investment's expense. As with the first alternative, this represents the Hadoop Administrator, Data Scientist, Technical Support, and Data Security Administrators

Figure 17: Proportion of Cost by Category



7 Benefit Analysis

This section summarizes the tangible and intangible benefits resulting from the Big Data investment. When applied, these will advance the governor's priorities in education, energy, jobs, and self-determination. Example benefits are presented for Water, transportation, air quality, jobs, transportation, and self-determination.

7.1 Tangible Benefits

Given that the specific analytic targets and methods for Utah have not been developed, published cost savings for similar target analyses were selected and scaled to the characteristics of the State. Five benefit categories were assessed and annual impacts are summarized in Figure 18.

Figure 18: Annual Monetized Benefits

Category	2016	2017	2018	2019	2020	Total
Reduced Recidivism		455,700	911,400	1,822,800	1,822,800	**5,012,700**
Decreased Tax Fraud		1,986,519	3,973,037	7,946,075	7,946,075	**21,851,706**
Decreased Food Stamp Fraud		567,216	1,134,432	2,457,936	2,457,936	**6,617,520**
Decreased Child Support Fraud		46,321	92,642	200,725	200,725	**540,414**
Decreased Medicaid Fraud		843,750	1,687,500	3,375,000	3,375,000	**9,281,250**
Total	**-**	**3,899,506**	**7,799,012**	**15,802,536**	**15,802,536**	**43,303,590**

7.2 Benefit Assumptions

All benefits are assumed to ramp up over time. In year 1 no benefits are projected as the alternative is being deployed. Thereafter benefits begin accruing at a 25% rate growing from 25% in year 2, then 50% in year 3, and 100% starting in year 4.

7.2.1 Reduced recidivism rates

Prison recidivism is a significant problem across correctional systems in the United States. Many inmates return to prison within a short time period of release creating a cycle of incarceration. Aside from the social problem resulting from recidivism, there is a significant cost that is associated with this issue. According to the Bureau of Justice Statistics (BJS), the average per prisoner cost of incarceration in state prison in 2010 was approximately $28,000 per year.[9]

[9] Tracey Kyckelhahn, *State Corrections Expenditures, FY1982-2010*, U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, NCJ239672, Washington, DC, December 2012, p. 4, http://www.bjs.gov/content/pub/pdf/scefy8210.pdf

The Congressional Research Service published comprehensive review of research into prison recidivism in January of 2015.[10] The BJS initially studied recidivism in 1994 and recently published a new report that examined the recidivism rates for 404,638 prisoners released in 30 states for five years after their release from prison in 2005.[11] The prisoners included in the study represent approximately three-quarters of the inmates released in 2005. The 2005 BJS recidivism study used a larger sample and a longer follow-up period than the 1994 study. Data show that by the end of the five-year follow-up period, approximately three-quarters (76.6%) of prisoners released in 2005 were rearrested. Furthermore, the BJS found that most released prisoners were rearrested within one year of being released. By the end of the first year following release, 43.4% of inmates were rearrested. The longer released prisoners went without being rearrested, the less likely they were to be rearrested. Compared to the arrest rate of 43.4% in the first year, 28.5% released prisoners who were not arrested in the first year were arrested for the first time in the second year after release. The proportion of released prisoners who were arrested for the first time over the course of the last three years of the follow-up period continued to decrease.

Offender reentry is a complex issue that touches on a wide range of social and governmental networks and programs. Offender reentry policies can vary significantly from state to state, and from community to community within particular states. The policies affecting prisoners and the kinds of programs available to them both in and out of prison depend on a variety of factors, including the availability of funding for social programs within states and communities and the number of private nonprofit and religious organizations operating in a given community.[12] Studies have shown that there are 5 key areas impacting offender reentry[13]:

- Employment
- Substance Abuse Treatment
- Education
- Mental Health Treatment
- Housing

---

[10] Offender Reentry: Correctional Statistics, Reintegration into the Community, and Recidivism, Nathan James, Analyst in Crime Policy, January 12, 2015, Congressional Research Service

[11] Matthew R. Durose, Alexia D. Cooper, and Howard N. Snyder, *Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010*, NCJ 244205, April 2014.

[12] Supra note 10

[13] Richard P. Seiter and Karen R. Kadela, "Prisoner Reentry: What Works, What Does Not, and What Is Promising," *Crime & Delinquency*, vol. 49, no. 3, 2003,

According to BJS the state of Utah had 7075 prisoners under the jurisdiction of state or federal correctional authorities at the end of 2013.[14] The same study reports that 2988 prisoners were released in 2013. Given the recidivism statistics quoted above it can be estimated that roughly 1300 of the 2988 prisoners released will return to prison in the first year and total cost of $36M per year based on $28,000 per prisoner per year.

Using big data analysis to improve the utilization and application of existing programs in the 5 key areas mentioned above could improve these numbers. The State of Michigan used computerized assessments to help reduce the three year recidivism rate by 10 percent. Assuming a conservative 5% reduction in recidivism in Utah due to Big Data Analysis would mean nearly $2M in savings annually as shown in Figure 19, not accounting for the compounding effect of multiple cases of re-entry into the system.

Figure 19: Recidivism Reduction Benefits (Dollars)

Prison Recidivism	2016	2017	2018	2019	2020	Total
Prisoners Released	3,000	3,000	3,000	3,000	3,000	
Recidivism Rate	43.40%	43.40%	43.40%	43.40%	43.40%	
Number Re-arrested	1,302	1,302	1,302	1,302	1,302	
Annual Cost per prisoner	$ 28,000	$ 28,000	$ 28,000	$ 28,000	$ 28,000	
Annual Recidivism Cost ($000)	36,456	36,456	36,456	36,456	36,456	
Recidivism Reduction	0.00%	1.25%	2.50%	5.00%	5.00%	
Total Cost Avoided	**$ -**	**$ 455,700**	**$ 911,400**	**$1,822,800**	**$1,822,800**	**$ 5,012,700**

7.2.2 Decreased tax fraud

Big Data analysis can be used successfully to fight tax fraud. The state of Georgia implemented a pilot program using big data analysis to improve tax fraud detection in 2012. The pilot program, which detected that 2 percent of all returns filed were potentially fraudulent, is part of the state's comprehensive anti-fraud program, which stopped millions of dollars in fraudulent returns in 2011, state officials said.[15] During the pilot program, Georgia tax officials used identity-based filters, which screened tax refund requests against billions of identity records collected from public databases and commercial sources. To find fraud before it happens, government

[14] BJS Bulletin September 30, 2014 Prisoners in 2013, E Ann Carson, Ph.D. BJS Statistician, http://www.bjs.gov/content/pub/pdf/p13.pdf

[15] Georgia fights tax fraud with identity-based filters, Government Computer News, Rutrell Yasin, 5/15/2012, http://gcn.com/articles/2012/05/15/georgia-fights-tax-fraud-lexisnexis.aspx

agencies must be able to access records across city, county and state boundaries. Identity-based filters can uncover categories such as:

- People who have died
- Altered Social Security numbers, changed to elude to Do Not Pay lists
- People not associated with the given address they list
- People not associated with the Social Security number they list
- Identities not found in public-records searches
- People incarcerated in prisons

In 2015 1,245,435 state tax return were filled in Utah and the state collected about $2.9 Billion in state income tax. 1,077,317 filers received a tax refund averaging $368.79.[16] Applying these numbers and the Georgia pilot program results would indicate over 20,000 fraudulent refunds. As shown in Figure 20 the annual savings when fully operational would amount to nearly $8M per year. In addition to the benefits to the state there would be the secondary benefit of detecting fraudulent federal tax refunds. The average size of the federal tax refunds in Utah in 2011 was $2904.[17] Again assuming a 2% fraud detection rate based on the Georgia pilot results, this would result in $62M in fraudulent federal tax refunds detected and avoided.

Figure 20: State Tax Fraud Prevention Benefits (Dollars)

Tax Fraud Reduction	2016	2017	2018	2019	2020	Total
Number of Tax Refunds	1,077,317	1,077,317	1,077,317	1,077,317	1,077,317	
Average Refund	$ 369	$ 369	$ 369	$ 369	$ 369	
Total Annual Refunds Issued ($000)	397,304	397,304	397,304	397,304	397,304	
Fraud Detected	0.00%	0.50%	1.00%	2.00%	2.00%	
Total Cost Avoided	**$ -**	**$ 1,986,519**	**$ 3,973,037**	**$ 7,946,075**	**$ 7,946,075**	**$ 21,851,706**

7.2.3 Decreased benefits fraud

Big Data analysis can also be used to fight fraud in public assistance programs. The Utah Department of Workforce Services provides economic data and statistical data in 4 major program areas[18]:

- **Child Care:** Child care includes both employment support and FEP supported Child Care assistance.

[16] Utah State Tax Commission Annual Report, 2013-14, http://tax.utah.gov/commission/reports/fy14report.pdf
[17] http://www.governing.com/gov-data/average-irs-tax-refund.html
[18] http://jobs.utah.gov/wi/statewide/assistrecipients.html

- **Financial Assistance**: Financial assistance includes all financial programs including the family employment program, refugee cash assistance, general assistance, and transitional cash assistance programs.
- **Food Stamps:** Supplemental Nutrition Assistance Program (SNAP) and the food stamp program.
- **Medical:** Medical includes Family Medicaid, Aged Blind and Disabled programs, Pregnant Woman, Children's Health Insurance Program (CHIP), Utah's Premium Partnership for Health Insurance (UPP), Primary Care Network (PCN), Nursing Home, Medicare cost sharing programs, waiver programs, and other special medical programs.

For July of 2015 the number of participants in these programs is shown in Figure 21.

Figure 21: Utah Public Assistance Program Participants

Child Care	20,424
Financial Assiatance	10,528
Food Stamps	229,086
Medical	313,162

For the Food Stamp program the potential savings from fraud reduction due to big data analysis is estimated. The average monthly food stamp payment in Utah is $312 and there most recent data indicated 101,000 recipients.[19] The United States Department of Agriculture (USDA) indicates the average level of Food Stamp Fraud is 1.3%.[20] Assuming that half the fraudulent applications could be detected with the use of big data would eliminate 0.65% of all payments currently issued. The resulting savings potential is shown in Figure 22.

Figure 22: Food Stamp Fraud Reduction Benefits (dollars)

Food Stamp Fraud Reduction	2016	2017	2018	2019	2020	Total
Number of Recipients	101,000	101,000	101,000	101,000	101,000	
Average Monthly Payment	$ 312	$ 312	$ 312	$ 312	$ 312	
Total Payments Issued ($000)	378,144	378,144	378,144	378,144	378,144	
Fraud Detected	0.00%	0.15%	0.30%	0.65%	0.65%	
Total Cost Avoided	**$ -**	**$ 567,216**	**$ 1,134,432**	**$2,457,936**	**$2,457,936**	**$ 6,617,520**

[19] https://jobs.utah.gov/edo/annreport/annualreport2014.pdf
[20] http://www.fns.usda.gov/pressrelease/2013/fns-001213

The Child Care program had 12,867 participants receiving an average of $200 per month.[21] Using the same fraud reduction assumptions as for the Food Stamp program the Child Care program savings from big data analysis are shown in Figure 23.

Figure 23: Child Care Fraud Reduction Benefits (dollars)

Child Care Fraud Reduction	2016	2017	2018	2019	2020	Total
Number of Recipients	12,867	12,867	12,867	12,867	12,867	
Average Monthly Payment	$ 200	$ 200	$ 200	$ 200	$ 200	
Total Payments Issued ($000)	30,881	30,881	30,881	30,881	30,881	
Fraud Detected	0.00%	0.15%	0.30%	0.65%	0.65%	
Total Cost Avoided	**$ -**	**$ 46,321**	**$ 92,642**	**$ 200,725**	**$ 200,725**	**$ 540,414**

In 2014 Utah Medicaid payments totaled over $2.2 Billion. The state's share of funding is 30%, totaling $675 Million.[22] Due to the large payments made even a small reduction in fraudulent payments will have a significant financial impact. This analysis assumes a conservative 0.5% of all payments could be avoided based on fraud detection using big data analysis. This would result in a savings exceeding $3 Million per year once reached. The Medicare program savings from big data analysis are shown in Figure 24.

Figure 24: Medicaid Fraud Reduction Benefits (dollars)

Medicaid Fraud Reduction	2016	2017	2018	2019	2020	Total
Total Payments Issued ($000)	2,250,000	2,250,000	2,250,000	2,250,000	2,250,000	
Stae Share of Funding	30%	30%	30%	30%	30%	
Total State Medicare Payments ($000)	675,000	675,000	675,000	675,000	675,000	
Fraud Detected	0.00%	0.13%	0.25%	0.50%	0.50%	
Total Cost Avoided	**$ -**	**$ 843,750**	**$ 1,687,500**	**$3,375,000**	**$3,375,000**	**$ 9,281,250**

7.2.4 Improved air quality

Based on a 2013 Environmental Protection Agency report, Utah emitted approximately 67 million metric tons of carbon dioxide (CO_2) in 2013[23]. The White House Office of Management and Budget provides estimates measuring the impact of reducing carbon emissions expressed as the Social Cost of Carbon[24]. A July 2015 updated estimate assumes 1 metric ton of CO_2 causes $36 in economic damage.

[21] Supra note 19

[22] https://medicaid.utah.gov/Documents/pdfs/annual%20reports/medicaid%20annual%20reports/MedicaidAnnualReport_2014.pdf

[23] http://www.epa.gov/statelocalclimate/documents/pdf/CO2FFC_2013.pdf

[24] Office of Information and Regulatory Affairs, Office of Management and Budget; July 2015; https://www.whitehouse.gov/blog/2015/07/02/estimating-benefits-carbon-dioxide-emissions-reductions,

No basis for estimating a realistic percent reduction in CO_2 could be found. Therefore a monetized estimate was not feasible. Activities that could decrease CO_2 include:

- I Improved targeting incentives
- Improved targeting of pollution abatement strategies
  - Current pollution reduction
  - Future pollution avoidance (such as through growth management)
- Improved speed and precision in sending alerts
- Expanded capability for enforcement (e.g. 'red' no-burn days)

7.3 Intangible Benefits

Improved data-driven decision making provides:

- The ability to provide critical insights into complex issues
- A focus on evidence-based decision-making
- The identification of root causes that can be addressed

This in-turn helps the State address the following constraints:

7.3.1 Policy

- Improved policy decision making and resource allocation
- Enterprise focus will increase throughput and service delivery to customers
- Integrated policies for data sharing, security, privacy, governance, and management.

7.3.2 Technology

- Infrastructure savings from eliminating "silod" data operations.
- Advanced data analytic and sharing infrastructure
- Interoperability across agencies will promote data sharing and analytics

7.3.3 Personnel

- Improved skill set alignment to support and manage a data driven enterprise

8 Conclusions

The following provides a comparative analysis, conclusion, and key management implications to monitor post implementation.

8.1 Comparative Analysis

The project team developed criteria to independently assess the extent to which an alternative achieved a criterion's objective. Those shown in Figure 25 are for the State's consideration. The figure identifies which alternative or both can best meet the criterion.

Figure 25: Selection Criteria for Consideration

Criteria	Either	Alternative 1: On-Premise	Alternative 2: Public Cloud
Strategic			
Maximize flexibility in establishing and operating a Big Data program			●
Maximize value proposition.			●
Maximize regulatory compliance		●	
Technical			
Minimize response time to service disruption	●		
Maximize data availability		●	
Maximize Security	●		
Maximize ability to secure	●		
Minimize technical obsolescence			●
Maximize security perception		●	
Maximize ability to achieve funding - operating expenditures easier to achieve than capital expenditures			●
Ensure data remains in the United States if not using a "Gov" cloud solution	●		
Minimize maintenance involvement			●
Maximize scalability			●
Maximize Surge Capacity			●

Criteria	Either	Alternative 1: On-Premise	Alternative 2: Public Cloud
Minimize speed to expand			●
Maximize ability to experiment with CPU/Memory/Storage configurations		●	
Maximize steady, high compute efficiency		●	
Financial			
Return on Investment	●		

An analysis was performed for a subset of criteria. This can be refined to reflect the criteria most fitting for the State's strategic and technical priorities.

8.1.1 Strategic

The following strategic objectives were assessed:

- **Maximize flexibility in establishing and operating a Big Data program.** Alternative 2 offers the most agility and the lowest fixed cost for Utah. The cloud provider assumes virtually all of the risk of establishing, configuring, operating, and securing the Hadoop environment.
- **Maximize value proposition.** Utah leverages this investment by focusing its time and attention on the establishment of the program, collecting and curating the data, and achieving the value propositions of the individual analyses.
- **Maximize regulatory compliance.** As Alternative 1 resides within the State's data center the solution is theoretically more capable of ensuring compliance with the Sarbanes Oxley Act, Data Act, Health Insurance Portability and Accountability Act (HIPAA), and other Federal, State, and private sector regulations and guidance.

8.1.2 Technical

The following technical objectives were assessed:

- **Minimize response time to service disruption.** This item could be either. The On-premise alternative allows DTS to respond immediately to resolve operational and security issues and to rapidly safeguard data. However, monitoring and responsiveness may be slower outside of core business hours. Alternative 2 may require more coordination with the cloud vendor, changing service level agreements, and addressing other contractual issues which could require significant time and resources. However, cloud vendors are fully

staffed 24/7/365 and Utah benefits from the constant investment in upgrading security controls and tools.

- **Maximize data availability:** Overall the on-Premise Private Cloud provides easier access to data maintained within the state network. However, this may still be accomplished within the cloud. Additionally, the cloud may provide equal or even greater access to open data maintained by the Federal government, other governments, or NGOs.
- **Maximize Security**. This item could be either. Hosting systems locally provides the comfort of local control and accountability for security. Conversely, cloud vendors are highly motivated to avoid data loss and hacking through constant investment in the latest security technologies and the provision of constant monitoring.
- **Minimize technical obsolescence**. Alternative 1 requires DTS to make continued hardware and software upgrades. Alternative 2 separates DTS from recurring investments to maintain technical currency.

8.1.3 Financial

Both alternatives generate the same benefits, only the cost profiles differ and then only slightly. As shown in Figure 26, the On-premise alternative is $410,794 cheaper than the Cloud based alternative. As a result, both alternatives achieve nearly identical returns on investment, have similar net present values, and provide a 2-year payback period.

Figure 26: Financial Comparison

Element	Alternative 1: On-premise Solution	Alternative 2: Cloud Solution
Total Benefit	$43,303,590	$43,303,590
Total Cost	$7,461,160	$7,871,954
Net Benefit	$35,842,430	$35,431,635
Return on Investment	260%	260%
Net Present Value	$35,242,597	$34,839,266
Payback Period	2 years	2 years

8.2 Conclusion

As benefits and costs are not distinguishing selection criteria, the selection rational is based on each alternative's ability to meet the State's strategic and technical objectives.

Based on the cloud security environment, there is also no difference in the State's technical ability to secure either alternative with similar labor plus software costs.

There are technical strengths and weaknesses of each alternative. The relative importance of those strengths and weaknesses can only be determined by the State's operating environment, and the specific quick-win project that will be chosen.

8.2.1 On-premise Alternative

There are advantages to a local on-premises cloud alternative. Current regulations are written for an on-premise perimeter security so managing personally identifiable information can use the accepted current procedures; there is an efficiency in an always-on consistently high processing (which would incur higher charges on the cloud). While staff are developing in-depth knowledge of the new technologies the local option provides a great transition environment.

8.2.2 Public Cloud Alternative

There are clear advantages to utilizing a CSP such as AWS Public cloud. There is no capital outlay, lower cost may be achievable with long term pricing agreements and CSP price reductions. Other advantages include scalability; extensibility; ability to focus on the mission and not the machines; regular physical machine upgrades; and 24x7uptime for the physical infrastructure.

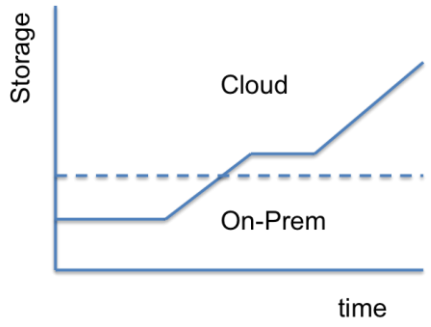
Collectively, the public cloud provides an easier long-term growth path and greater flexibility. It may have hindrances to deploying aspects of some data and solutions depending on regulatory concerns

8.2.3 Time-phased Consideration

Determining which alternative, the on-premise solution, or public cloud, is superior depends on what the specific requirements are for the State's first project, and any regulatory issues in moving data to the cloud. This quick-win scenario could provide the most immediate decision criteria between the two alternatives.

While a public cloud is undoubtedly the end-state or part of the end-state, policies and regulations could lead to early delays and effectively slow down the push to become operational and realize "quick wins". While the State will have to decide on the tolerance for this extra potential complexity, there are advantages to early development being with an on-premises cluster. As a simplistic representation consider Figure 27.

Figure 27: Time-phased Approach



The initial learning system could be an on-premise cloud, even one re-using existing resources, where the staff could learn and potential obtain quick results (using current perimeter security) on a small project. Then after an initial project or two, the team can expand to develop a hybrid cloud. While this adds some complexity, in actuality the cloud management tools (and Big Data platform tools) are fully capable of spanning a hybrid cloud using a single "pane of glass" control. Over time the State could begin to add more and more resources to the public cloud. Given the experience with the cost and actual performance, and the maturation of cloud usage policies and regulations, the State would be in a better position to determine their full requirements, and if indeed the full migration to the cloud would be the appropriate end-state. In either case this offers a mixed alternative for the early and late stages of the deployment, with no one-off work. This on-premise cloud could in the long-term provide a traditional perimeter security approach for PII data analysis making regulatory compliance easier.

This additional time-phased hybrid cloud option is presented to give DTS the greatest flexibility in achieving rapid value to the State

8.3 Key Management Implications

In implementing a solution, management should:

- **Assess and prioritize candidate Big Data programs.** This may influence the technical and acquisition strategy adopted by the State
- **Integrate lowest level of data.** To achieve full benefits, the lowest level of data must be integrated and made available.
- **Data growth.** If data growth is predicted to exceed 500 TB, then pricing assumptions should be assessed
- **Monthly Data Transmission Rate.** Should the value exceed 10 TB then the pricing assumptions should be reviewed.
- **Monitor usage patterns.** Usage statistics will provide insight into determining the optimal procurement strategy for acquiring additional capability for either alternative.

Appendix A: Acronyms

AWS	Amazon Web Services
AGRC	Automated Geographic Reference Center
BJS	Bureau of Justice Statistics
CHIP	Children Health Insurance Program
CPIC	Capital Planning and Investment Control
DNS	Department of Natural Resource Services
DTS	Department of Technology Services
DWS	Department of Workforce Services
EMR	Elastic MapReduce
ETL	Extract, transform, and load
FedRAMP	Federal Risk and Authorization Management Program
FISMA	Federal Information Security Management Act
FTE	Full Time Equivalent
GB	Gigabyte
IaaS	Infrastructure as a Service
IPT	Integrated Project Team
MTCS	Multi-Tier Cloud Security
PCN	Primary Care Network
NPV	Net Present Value
ROI	Return on Investment
SaaS	Software as a Service
SLA	Service Level Agreement
SNAP	Supplemental Nutrition Assistance Program
SQL	Structured Query Language
TB	Terabyte
UPP	Utah's Premium Partnership for Health Insurance
USDA	United States Department of Agriculture
VM	Virtual Machine

Appendix B: Alternative 1—On Premise Cost Notes

This section contains detailed cost notes that support the cost estimated for each On Premise Cost Element. Each note provides the following to the extend applicable:

- A description of the WBS cost element
- The methodology, assumptions, data, and basis for estimate
- Summary and detailed Cost Element constant year dollar estimates

1 Non-Recurring Costs

1.1 Procure Hardware and Software

In each of the first two years pre-configured SuperMicro Corporation's Bare Metal 42U cluster will be installed containing server racks, storage, networking devices, and the Hadoop system. The cost for each of the clusters is $187,500 as quoted by SuperMicro.

1.2 Provide Contractor Labor

The cost represents the labor needed to install the hardware, install the operation system, and configure networking and security policies. The assumption was made that it will take 2 FTEs 4 weeks each to complete this work. The labor rate was assumed to be $125 per hour. The resulting cost is $40,000

$$\$125 * 2 * 4weeks * 40hrs/week = \$40{,}000$$

2 Recurring Costs

2.1 Procure Maintenance Contracts

SuperMicro Corporation quoted $50,290 for a three year maintenance contract that provides onsite support by a vendor authorized service technician. This cost is incurred each of the first two years after the implementation of each cluster. Renewal costs are incurred in years 4 an5 and no change in cost is assumed at this time.

2.2 Provide Government Labor

The basis for the government labor estimate is shown in Figure 5 in Section 3 documenting the financial assumptions.

2.3 Provide Contractor Labor

These are the costs associated with experts needed to initially architect the Big Data system. This would include Cloud and Security Engineers providing the configuration of the infrastructure; Big Data Engineers setting up and maintaining the software

platform including the Hadoop stack; and Data Scientists establishing a collaborative analytics program with business analysts across the state. The assumed cost of this support, which is diminishing over time as work is transition to the government, is shown in Figure 28

Figure 28: Contractor Labor Costs ($'s)

Recurring ($s)	2016	2017	2018	2019	2020	Total
Provide Contractor Labor	500,000	315,000	265,000	150,000	50,000	**1,280,000**

Appendix C: Alternative 2—Cloud Hosted Cost Notes

This section contains detailed cost notes that support the cost estimated for each Cloud Hosted Cost Element. Each note provides the following to the extend applicable:

- A description of the WBS cost element
- The methodology, assumptions, data, and basis for estimate
- Summary and detailed Cost Element constant year dollar estimates

3 Non-Recurring Costs

3.1 Procure Hardware and Software

No hardware or software will be purchased for the Cloud Hosted Solution.

3.2 Provide Contractor Labor

The cost represents the labor needed to perform Initial design, setup and configuration. The assumption was made that it will take 2 FTEs one month to complete this work. The labor rate was assumed to be $150 per hour. The resulting cost is $48,000

$$\$150 * 2 * 4weeks * 40hrs/week = \$48{,}000$$

4 Recurring Costs

4.1 Procure Cloud Computing

The cloud computing resources are sized to match performance characteristics required. The following set of virtual machines was priced on three year reserve basis paid for upfront. Storage is priced on a per month basis at a fixed rate throughout the year.

- Cost for the r3.4xlarge instances is $12,906 for 3 years reserved.
- Cost for the d2.4xlarge instances is $20,780 for 3 years reserved.

The number of instances purchased for a 3 year reserved term are shown in Figure 29. The resulting instances in operation in each of the years of the analysis are shown in Figure 30.

Figure 29: Instances Purchased for 3 Year term

Instance Purchased	2016	2017	2018	2019	2020
r3.4xlarge	4			4	
d2.4xlarge	7	2	2	9	5

Figure 30: Instances running

Instance Running	2016	2017	2018	2019	2020
r3.4xlarge	4	4	4	4	4
d2.4xlarge	7	9	11	13	16

Cost for S3 Storage is $30/TB per month. Figure 31 shows the amount of storage purchased each on a monthly basis for the 5 years analyzed.

Figure 31: S3 Storage Assumptions

S3 Storage (TB)	2016	2017	2018	2019	2020
$30/TB per month	150	180	216	259	311

The Data Transfer requirement is assumed to be stable 10TB per month at $0.9 per GB, resulting in annual cost $10,800.

The resulting total cloud compute and storage costs are shown in Figure 32.

Figure 32: Cloud Compute and Storage Costs

Recurring	2016	2017	2018	2019	2020	Total
Procure Cloud Computing	263,084	118,360	110,540	343,956	183,014	**1,018,954**

4.2 Provide Government Labor

The basis for the government labor estimate is shown in Figure 5 in Section 3 documenting the financial assumptions.

4.3 Provide Contractor Labor

These are the costs associated with experts needed to initially architect the Big Data system. This would include Cloud and Security Engineers providing the configuration of the infrastructure; Big Data Engineers setting up and maintaining the software platform including the Hadoop stack; and Data Scientists establishing a collaborative analytics program with business analysts across the state. The assumed cost of this support, which is diminishing over time as work transitions to the government, is shown in Figure 28.

Figure 33: Contractor Labor Costs ($s)

Recurring ($s)	2016	2017	2018	2019	2020	Total
Provide Contractor Labor	500,000	315,000	265,000	150,000	50,000	**1,280,000**

Appendix D: Key Contributors

Name	Title	Area of Contribution
State of Utah		
Mark VanOrden	Chief Information Officer	High level vision for big data
Dave Fletcher	Chief Technology Officer	Project oversight and guidance
Robert (Bob) Woolley	Chief Technical Architect	Project oversight and guidance
Phil Bates	Chief Information Security Officer	Domain expertise
Bill Boudreaux	IT Director, Natural Resources, Agriculture, & Environmental Quality	Domain expertise
Bert Granberg	AGRC Data	Domain expertise
Norm Johnson	Cognos User Group	Domain expertise
Darrus McBride	DTS Hosting Manager	Domain expertise
Greg Mead	IT Director,	Domain expertise
Drew Mingl	Open Data Coordinator	Domain expertise
Dave Moffit	Technical Lead, Dept. of Workforce Services Big Data	Domain expertise
Mark Schultz	IT Director, Dept. of Workforce Services	Domain expertise
Boyd Webb	Information Security	Domain expertise
Wu Xu	Director of Health Informatics	Domain expertise
Dean Zumbrennan	DTS Database Manager	Domain expertise
SAIC		
Frank Crichton	Project Manager	Coordination and guidance
Nancy Grady	Principal Data Scientist / Technical Lead	Task 1,2,3,4,5
Colleen Apte	Policy and Research Analyst	Task 1 and 4
Jeremy Coones	Data Scientist	Task 3, 5
Shane Gellenthien	Director, CPIC Center of Excellence	Task 4
Holger Hinsch	Sr. inancial Analyst	Task 4, 5
Cary Landis	SME/Cloud	Task 2 and 3
Rick Martin	SME/Air Quality Use Case	Task 5
Merrick Watchorn	Cloud Security SME	Task 1, 2, and 3