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

Data Driven Decision Making Project

Volume 3: Technology Roadmap
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

To become a data-driven organization, Utah will need to bring together the data from across the state to one environment to enable interdisciplinary analytics. The technologies that will make this both possible and cost effective are cloud-based Big Data technologies. There are two main choices for hosting these tools, and on-premise cluster and a public cloud. Each option was carefully compared for the most cost-effective means to achieve the level of performance. From the price and performance analysis there is no technical or cost differentiation in the two options. The decision will need to be based on the prioritization Utah has for a number of options. In the next layer, the data platform has three leading options, and Cloudera is the best of those options. The analytics or processing layer is more dependent upon the nature of the data for the choice of database, and skills and comfort of the analyst. Data Scientists will want to focus on the Spark data framework, and Business Analysts will need the capability to use Business Intelligence or Excel themselves to perform their own analysis. Finally we note that the additional layers, and especially the addition of third part cloud providers introduce changes in the security fabric from a traditional perimeter security for an on-premise system. The different areas needing attention are detailed to provide an approach to securing the Big Data Environment.

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 2 People Skills and Collaborations, Task 4 Business Case, Task 5 Data Science and Value.

2 Business Challenges and Vision

In Utah, State and Local governments are steadily digitizing government data and providing online services. Due to the awareness of users, and availability of Internet access, use of E- government information has increased. As a result, the size and variety of data available to decision makers is increasing exponentially. *Volume*, *variety*, *velocity* and *complexity* of data resources within the State are increasing. In addition, external data sources such as sensors from the *Internet of Things* will continue to increase. The ubiquity of *smartphones* enables greater delivery of services to citizens, as well as a mechanism for data input and feedback from citizens. Enabling state-wide analytics through better management and access of State data holds the potential both for more efficient and effective government services, as well as for enhanced citizen engagement in the decision making process. The vision is to ensure that Utah has the resources and methodologies in place to become a *data-driven organization*.

2.1 Business Challenges

Like most organizations, the State has developed around operational needs. Each agency within the state developed their own silos for data storage and analytics, to meet the pressing needs of their operational mission. From an enterprise-wide point of view, this results in an environment where each agency only has their own data at their disposal for making decisions. For cost-effectiveness, the State's data systems have been centralized under DTS control in the data center, but the data remain under the isolated stewardship of their agencies.

Current data warehouse technologies are no longer cost-effective, when trying to handle large volumes of data, or handle high velocity data. Furthermore, the large number of operational datasets makes the development of an *Enterprise Data Warehouse* a costly and time-consuming proposition. Before embarking on a project, it is vital to have a vision for the value that these technologies can provide above and beyond the current State capabilities. The challenge is to determine the most cost-effective approach to provide the State value through enterprise-wide analytics, and to ensure that there is an expected return on any investment to enhance the technologies. The Task 4 Business Case will explore the expected return on investment that could be expected from a Big Data deployment, and Task 5 Data Science and Value in this series will describe the economic analysis of using Big Data technologies, and the scenarios for creating additional citizen value from the use of these technologies.

2.2 Business Vision

Data-Driven government is the goal for the State of Utah. Analytics performed across the breadth of data in the State provides an opportunity for the creation of value that is not possible through analytics on individual data silos within operational units. While few datasets will reach the criteria for “Volume” in the new Big Data paradigm, Variety is the Big Data characteristic that will bring the most value.

> *Business benefits are frequently higher when addressing the* ***variety*** *of data than when addressing* ***volume***”[1]

The vision is to bring together datasets from across the State to enable enterprise-wide analytics and evidence-based management. Using Big Data for decision-making can result in the transformation of government by increasing opportunities for efficiency, effectiveness, and access to previously siloed data resources.

3 Technical Assumptions and Vision

The State of Utah has been at the forefront of technology innovations for government. The State has incorporated social media, mobile technologies, and open data to provide greater services to its citizens. The State is unable to reach its full decision-making potential in our current digital age, however, using existing relational database technologies. Traditional data center technologies are limited in the ability to process and analyze vast, diverse and fast-streaming data.

3.1 Utah Environment

While Utah has centralized control of the systems across the state, the data remains in separate databases. Being in their own silos, it is very difficult to integrate the data to provide any form of longitudinal analysis. Each agency only has access to their own data, and their ability to use data analysis for decision-making is correspondingly limited. While data is requested and shared between organizations, it can be a lengthy process. If some needed data was left out of the original request, the process has to start all over again.

3.2 Technical Context

Beginning around the year 2005, new approaches began to emerge to parallelize data management and analysis. This paradigm shift is described as *Big Data*. Unfortunately this term is used to represent a number of concepts from the size of the datasets, the Hadoop ecosystem of tools, the prominence of unstructured data,

[1] Mark Beyer and Doug Laney, Gartner, 2012, “The Importance of Big Data: A Definition”

data in the cloud, the desire for value, the loss of security and privacy, the immediacy of results, the ability for individual personalization, and many others. Fundamentally the paradigm shift is that data-intensive applications have gone parallel, analogously to the parallelization shift that occurred 20 years ago for compute-intensive scientific simulations. It's not just that data is "bigger" than before, since data has been getting "bigger" every year for several decades and new approaches have been developed to handle it. Just as the relational database was a paradigm shift in the way data could be handled for analysis, Big Data is a paradigm shift to parallelism for scalability.

> ***Big Data*** *consists of extensive datasets—primarily in the characteristics of volume, variety, velocity, and/or variability—that require a scalable architecture for efficient storage, manipulation, and analysis.*[2]

The important point is to note that it is the increased efficiency in scalable affordable cost or speed of analysis that determines when data is considered big. This term is somewhat recursive in that data is "big" if you need to use parallel techniques to handle it, and of course you use the Big Data Engineering tools because the data is "big".

We will be using the term Big Data more in the sense of Big Data Engineering, investigating the tools and techniques that can enable evidence-based analytical results.

3.3 Technical Assumptions

In this study, several assumptions were made for the features of an overall Big Data environment.

- Cost-effectiveness
- Avoiding vendor lock-in
- Using industry-standard skillsets
- Initial expense: OpEx easier than CapEx
- Incremental development
- Focus on Security and Privacy
- Current on-premises environment changes/expansion
- Availability of the system
- Extensiveness of storage
- Improved processing capability

3.3 Technical Vision

[2] NIST SP1500-1, "NIST Big Data Interoperability Framework: Volume 1, Defintions

One of the major goals in developing a Big Data capability is to create what is called a ***Data Lake*** for enabling enterprise-wide analytics, versus an ***Enterprise Data Warehouse (EDW).*** In a Data Lake, the data from operational systems is brought together to enable analytics across the datasets, without the traditional step of full integration between datasets. An EDW is the best solution when (1) the data to be used is well known, (2) there are only a few datasets being brought together, (3) the desired analytics is well known, (4) the analytics will be run on a regular basis, and (5) the high performance needs of the analytics justifies the significant time and expense for a "big bang" EDW project. A big bang project is one in which there is essentially no value generated until the project is completely finished.

The technical goal for the state of Utah is to create a data lake where the operational datasets can be gathered into one environment to enable state-wide analytics for decision support. This would represent the technical component of becoming a data-driven government.

4 Roadmap

We recognize that Big Data technologies are changing rapidly. This section provides a roadmap for the hardware and software technology to obtain Big Data capabilities in the most cost effective manner at the current time.

There are multiple ways to evaluate a Big Data deployment. The NIST reference architecture for Big Data systems is shown in Figure 4-1.

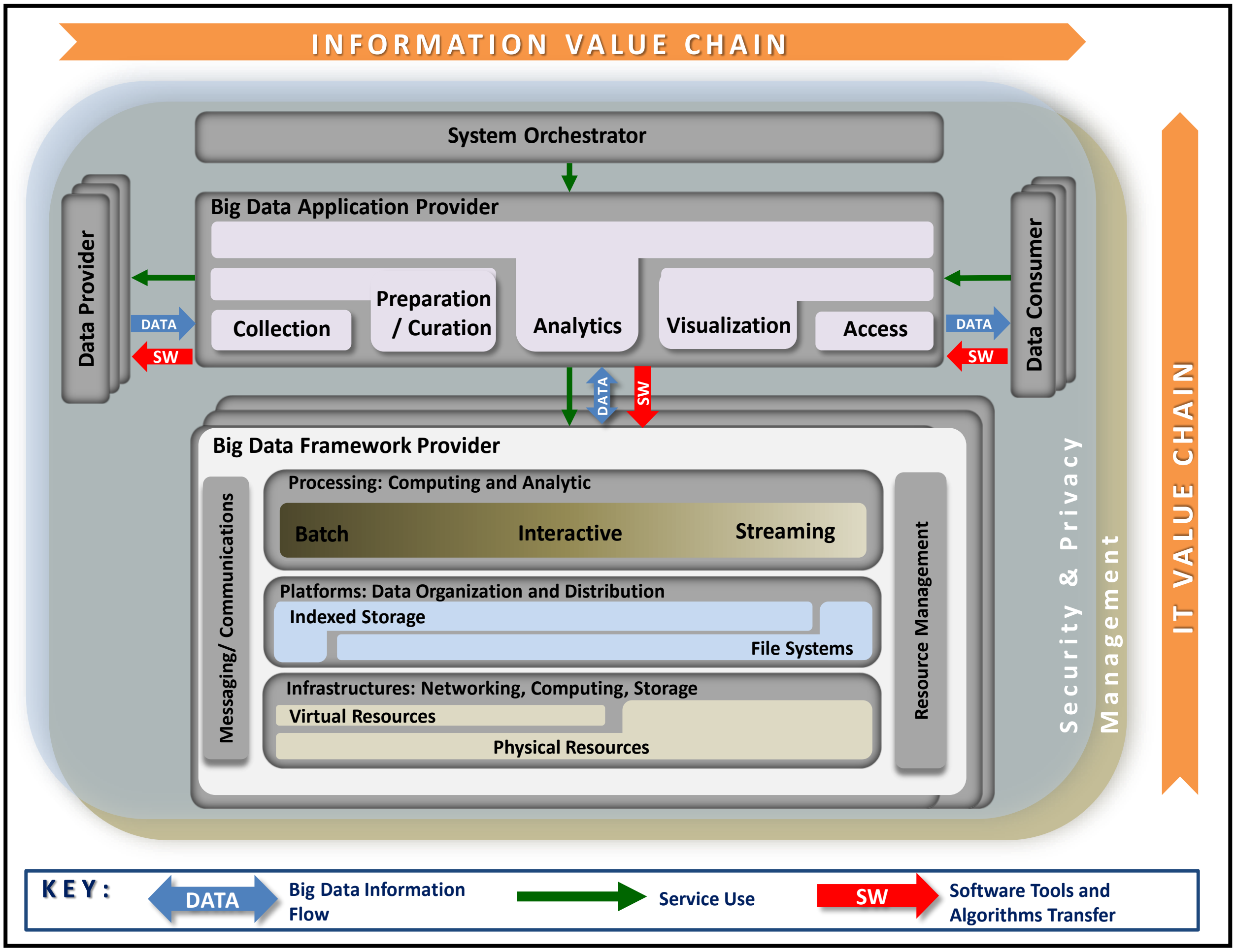


Figure 4-1: NIST Big Data Reference Architecture[3]

In this architecture, the *Data Providers* are the internal operational systems, or any external sources of data. The *System Orchestrator* represents the business ownership, governance and technical requirements for the system. The ownership and governance aspects of the system will be described in Task 1: Policy and Governance. The *Data Consumer* represents the State's business user that will use the analytics to develop their data-driven decisions. The *Application Provider* represents the enterprise-wide data and analytics capability, which is built using the tools within the *Framework Provider*. Big Data applications that can generate value for the State are described in Task 5: Data Science and Value. Both *Security and Privacy*, and *Management* are considered as a fabric that touches every component of this entire ecosystem. Security and Privacy will be discussed in within each

[3] "NIST Big Data Interoperability Framework: Volume 6, Reference Architecture", NIST Special Publication 1500-6, eds David Boyd and Wo Chang, 2015.

component. It is assumed that the development and management processes for this environment will follow current State practices.

In the following we will be describing the most cost-effective frameworks for developing the State's Big Data capability, describing the Infrastructure options, the Platform options, and the Analytics options.

4.1 Infrastructure Layer

The architecture of a Hadoop cluster is largely driven by the customer's processing and storage (including data redundancy) requirements. There are many possible ways to size and architect any given Hadoop cluster. For example, a 100TB Hadoop cluster might be implemented using four 25TB nodes, twenty-five 4TB nodes, or a hundred 1TB nodes. Another major consideration is that Hadoop is designed to provide high reliability by storing extra copies of data across nodes; effectively multiplying the storage costs depending on how many copies of the data are stored. The architecture ultimately drives the cost of the cluster, which is based on the size, number and type of slave servers (nodes); the amount of storage available on each of those nodes; and related support costs.

Hadoop can be deployed in a traditional onsite datacenter as well as in the cloud. The cloud allows organizations to deploy Hadoop without hardware to acquire or specific setup expertise. The cloud also makes it possible to temporarily "spin up" many (potentially thousands) of virtual machine (VM) nodes to run a Hadoop job; and then terminate the VMs when the job is complete. Hadoop is therefore well-suited for implementations that leverage the cloud.
A common misconception about Hadoop is that it is very inexpensive because it can use "commodity hardware". However, what is typically true is that Hadoop is significantly less costly than alternatives (such as a supercomputer), but may still require a noteworthy investment to support a large enterprise.

Our preliminary estimate for on-premises hosting is approximately $2,906,160 for 5-years; versus $4,367,174 for 5 years based a cloud computing model; whereas the on-premises model requires datacenter investments to be provided by the State of Utah. The following describes our fundamental assumptions, and high-level analysis of on-premises versus cloud-based hosting approach Utah's Big Data platform.

<u>**High-level Assumptions**</u>

We established a few assumptions in order to draw as close of a comparison as possible. Our assumptions are based on the functional needs (performance and operational storage) rather than being based on technical requirements:

1) One Management node
2) Number of Name/Job Tracker nodes = 3
3) Target Operational data = 150 TB total
4) Target Headroom (unused space) per Node = 1 TB total
5) Max Storage per Node = 24TB total

6) 20% growth annually in operational data
High availability is required

4.1.1 On-Premises Options

SAIC evaluated three main options for a Big Data on-premises infrastructure:

A. **Custom Hadoop Cluster:** Involves buying servers, storage, and networking hardware/software; installing Hadoop systems; and building a cluster within Utah's existing datacenter.

Examples: Hybrid cluster of HP and Dell servers, and EMC storage

Pros	Cons
• allows for full control of hardware • allows for ability to implement custom hardware-level security	• requires upfront capital purchases • requires facility space and costs • requires investment into professional services to analyze, design and build the cluster • requires periodic technology refreshes • difficult to change after the investment • may require FedRamp accreditation • Typically costly

B. **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.

Examples: Supermicro Bare Metal 42U; HP HDP; and Cisco Big Data Hadoop

Pros	Cons
• Save time/money setting up/configuring hardware & software • allows for ability to implement custom hardware-level security	• requires upfront capital purchases • requires facility space and costs • may have special facility requirements • difficult to change after the investment • may require FedRAMP accreditation

C. **Big Database Appliance:** High-end machines that are tuned to handle very large databases using traditional technologies.

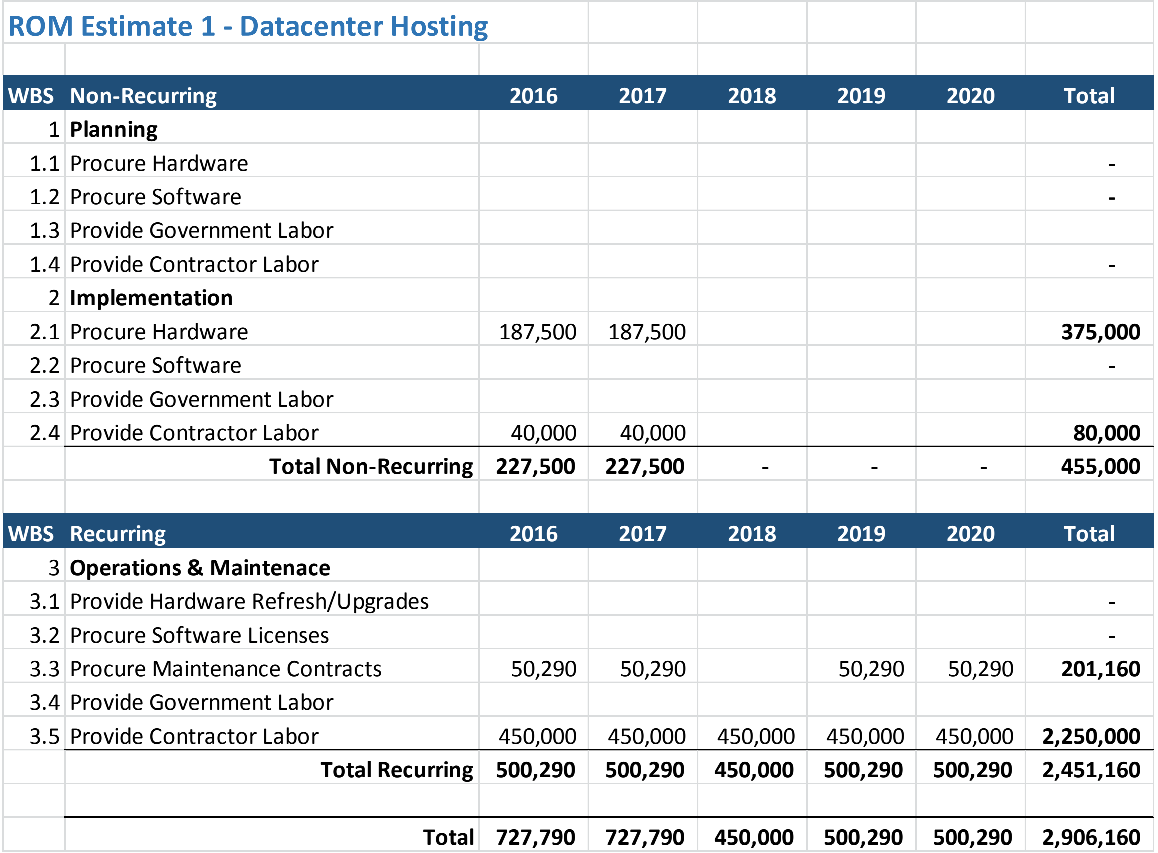
Example: Oracle Exadata

Pros	Cons

• ability to use familiar database technologies • possibility to take advantage of existing database licenses	• requires upfront capital purchases • typically very expensive • may require FedRamp accreditation

4.1.2 On-premises Pricing

SAIC explored several viable providers of converged infrastructure. We focused our analysis on converged infrastructure because – based on the prior experiences of our team members, converged infrastructure generally offers the highest value at the lowest cost. Specifically, we researched available options from Supermicro, HP and Cisco. We also researched Nutanix, however the company was not responsive to our inquiries. The SuperMicro converged infrastructure offered the best pricing for machines. To support our abovementioned assumptions (See Section 3.3), SAIC estimates an initial rough-order-of-magnitude cost of about $187,500 for the machines. To accommodate 20% growth per year, Utah would need to acquire an additional machine in Year 2, which would support growth over the next few years. The maintenance are approximately $50,290 per year; and labor is approximately $40,000 for years 1-2 setup; plus $450,000 per year ongoing; arriving at a ROM 5-year total of $2,906,160 for on-premises hosting. The following figure summarized the costs for Years 1-5 using SuperMicro machines.



ROM Estimate 1 - Datacenter Hosting

WBS	Non-Recurring	2016	2017	2018	2019	2020	Total
1	**Planning**						
1.1	Procure Hardware						-
1.2	Procure Software						-
1.3	Provide Government Labor						
1.4	Provide Contractor Labor						-
2	**Implementation**						
2.1	Procure Hardware	187,500	187,500				**375,000**
2.2	Procure Software						-
2.3	Provide Government Labor						
2.4	Provide Contractor Labor	40,000	40,000				**80,000**
	Total Non-Recurring	**227,500**	**227,500**	**-**	**-**	**-**	**455,000**
WBS	**Recurring**	**2016**	**2017**	**2018**	**2019**	**2020**	**Total**
3	**Operations & Maintenace**						
3.1	Provide Hardware Refresh/Upgrades						-
3.2	Procure Software Licenses						-
3.3	Procure Maintenance Contracts	50,290	50,290		50,290	50,290	**201,160**
3.4	Provide Government Labor						
3.5	Provide Contractor Labor	450,000	450,000	450,000	450,000	450,000	**2,250,000**
	Total Recurring	**500,290**	**500,290**	**450,000**	**500,290**	**500,290**	**2,451,160**
	Total	**727,790**	**727,790**	**450,000**	**500,290**	**500,290**	**2,906,160**

Figure 1: ROM pricing for Years 1 through 5

Additional Assumptions:
The on premises solution assumes that Utah will support all datacenter facility costs, including network infrastructure, edge firewall security, security tools, floor space, power, cooling and Disaster Recovery / Continuity of Operations (DR/COOP) expenses; as well as any investments required to support FedRAMP and/or additional compliance requirements.

4.1.3 Public Cloud Options

The advent of cloud computing makes possible several potential benefits:

- Ability to structure costs as Operational Expenses (OPEX) based on consumption, as an alternative to upfront capital expenses
- Scale up and down with increased/decreased demand and surges
- Ability to temporarily stop/run large clusters based on need
- Ability to try/pilot multiple configurations to identify an optimal configuration
- Access to special "Hadoop as a Service" tools that simplify (by automation) the setup and management of Hadoop clusters
- Enable Utah to focus on its core business by outsourcing non-core IT functions

There are dozens of viable cloud computing vendors. According to Gartner, the leading vendors include Amazon Web Services (AWS), Microsoft (Azure), VMware (vCloud® Government Service), IBM (Softlayer) and Google (Google Compute Engine (GCE). The cloud providers are largely similar in terms of the capabilities of their compute and storage offerings, as well as performance and pricing. These factors (capabilities, performance and pricing) are representative of commoditized services. The primary differentiators between vendors are more related to the market share of their offerings, their accreditations (such as FedRAMP), and their specialized "Hadoop as a Service" offerings. Therefore, SAIC focused our analysis on these major differentiators (market share, FedRAMP, and availability of Hadoop services), as well as alignment with the State of Utah's environment.

The vendor's market share especially important when choosing a technology platform because it largely reflects the results of comprehensive AoAs that were performed by other customers. Market share typically reflects the stability of the vendor. As the cloud computing marketspace continues to unfold, SAIC suggests selecting a vendor with major market share to ensure the availability of the vendor and services for the next several years.

As a disclaimer, our analysis does not represent a comprehensive Analysis of Alternatives (AoA), which would inherently involve significantly more effort to compare detailed features and granular pricing scenarios.

Based on our analysis, SAIC suggests that AWS offers Utah the best value based on the above-mentioned factors. Below is a summary of our analysis:

- **Market share:** Amazon Web Services (AWS) leads Gartner Research's 2015 Cloud Computing Infrastructure as a Service (IaaS) market in terms of the company's

"Ability to Execute" and "Completeness of Vision". The AWS leadership position in the market is significant compared to nearest competitors. According to a 2014 Synergy Research Group report, Amazon is dominating the worldwide cloud infrastructure market with a 27% share; followed by Microsoft, IBM and Google. Microsoft is leading the growth rate with 96%; followed by Amazon at 51% growth.

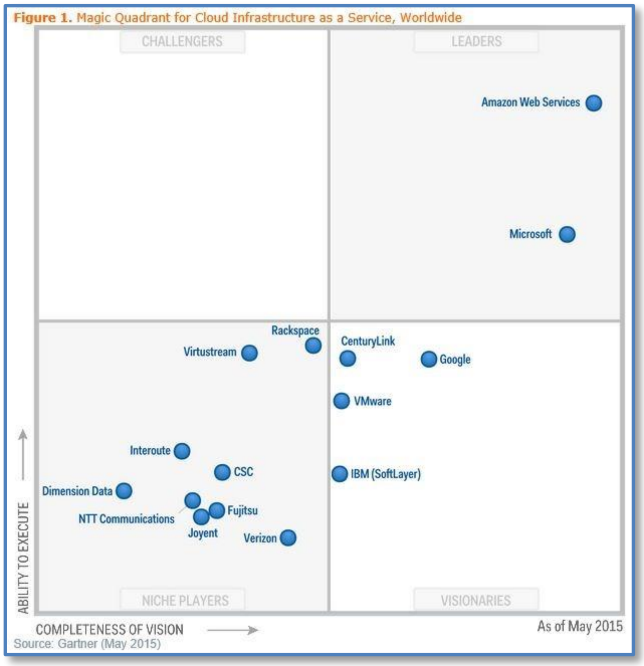


Figure 4-2: Gartner 2015 Magic Quadrant

- **FedRAMP Accreditation:** Of the vendors mentioned, only AWS (East/West and GovCloud), Microsoft Azure (Azure Cloud Infrastructure and Public Infrastructure) and VMware (vCloud® Government Service) are FedRAMP certified; whereas Google is pursuing FedRAMP accreditation. IBM does not yet appear to be pursuing its FedRAMP accreditation for its Softlayer offering, though IBM certified its predecessor SmartCloud offering.
- **Hadoop as a Service:** Amazon and Google offer automated services to simplify the setup and management of Hadoop for their respective clouds:
  - **Amazon Elastic MapReduce:** Amazon Elastic MapReduce (Amazon EMR) is a "Hadoop as a Service" offering that makes it easier to setup and manage Hadoop clusters that make use of Amazon's EC2 (compute). Amazon EMR allows for running other Hadoop frameworks such as Spark and Presto; and interacting with data in other AWS data stores such as Amazon S3 and Amazon DynamoDB. The EMR service costs extra with the potential benefit of reducing the manual labor required to set up and manage the Hadoop cluster.
  - **Google (Hadoop) Click to Deploy:** Google simplifies the task of setting up Hadoop on GCE. It supports Hadoop 2, automating the setup of distributed file

system (HDFS), resource manager (YARN), and modifying the configuration (worker node count, virtual machine type, and other parameters) prior to deployment. It also automated deployment of Apache Spark on the cluster nodes.

- **Alignment with Utah's Environment:** SAIC also considered that Utah is currently using some Amazon cloud services.

In short, AWS has a substantially greater market share than other vendors; has obtained its FedRAMP accreditation; offers a "Hadoop as a Service" (namely Elastic Map Reduce) for possible future consideration; and aligns with Utah's current environment. Additionally, AWS pricing and features are competitive, with availability of specialized virtual machines (D2 instances) that are tuned for big data clusters. We recommend AWS based on its high value, low risk and competitive pricing.

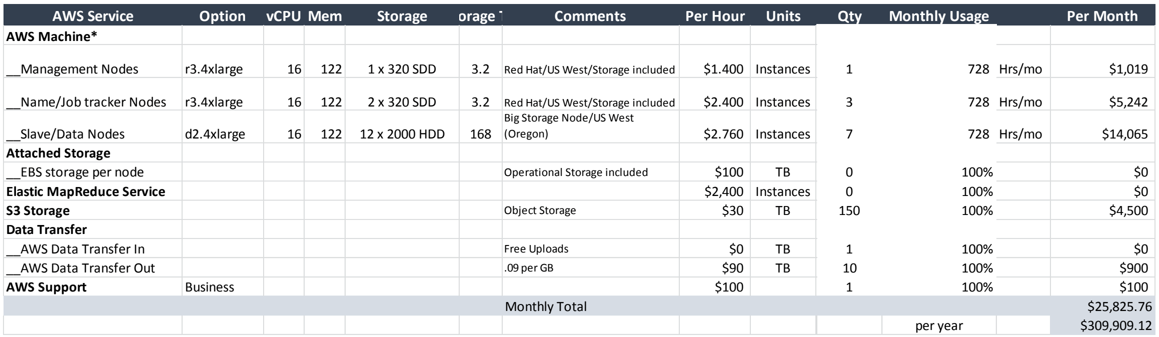
4.1.4 Public Cloud Pricing

AWS virtual machines are referred to as EC2 (Elastic Compute Cloud). There exist several types of virtual machines (instances). The best selection depends on the needs of the application at hand. The full list of instances is available at the following link: https://aws.amazon.com/ec2/instance-types/

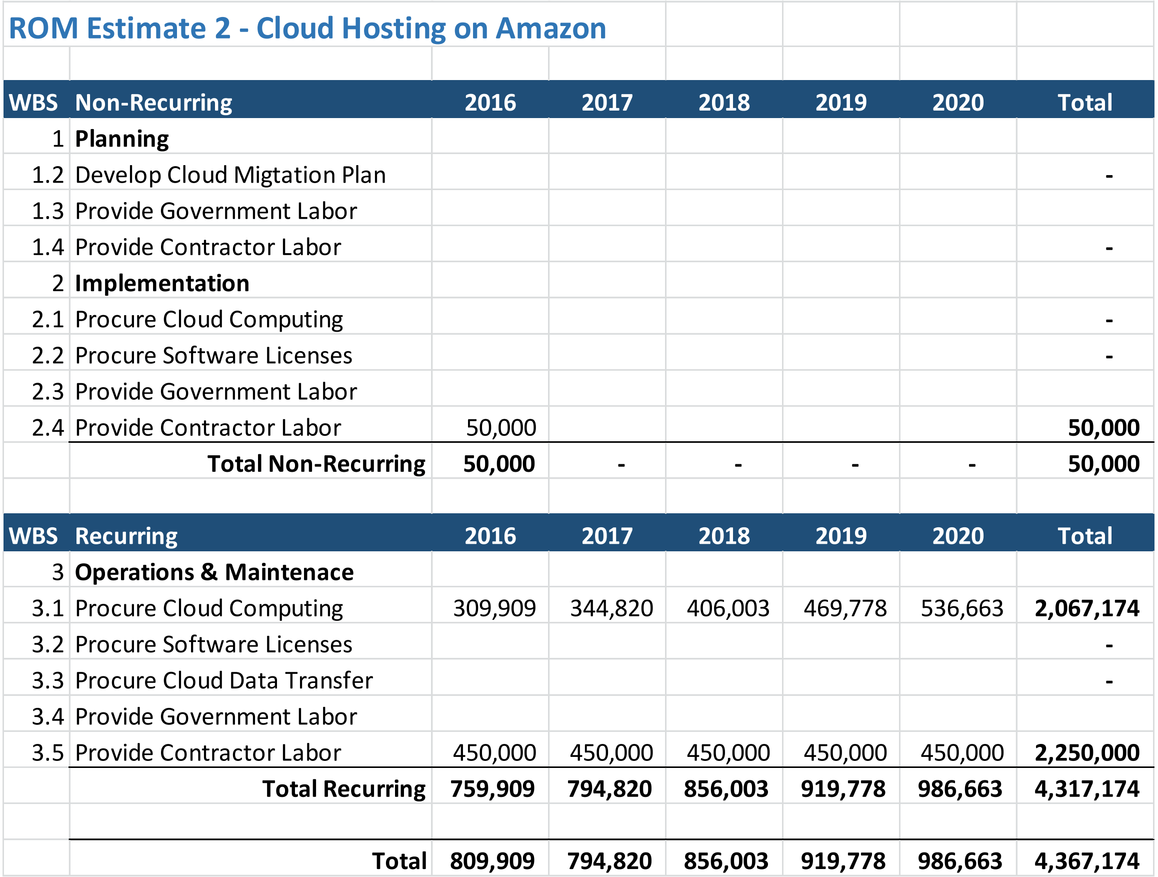
Analysis of Instance Types

SAIC suggests a cluster that uses AWS R3 instances for master (Management and Name/Job Tracker) nodes; and D2 instances for slave nodes. The R3 instances are well-suited for Hadoop master nodes, based on high memory availability. The D2 instances are well-suited for Hadoop slave nodes because they include bundled storage that "lives close to" the processor, thus delivering low latency. AWS offers two main types of storage: EBS (Elastic Block Storage) and S3 (Simple Storage Service). EC2 is comparable to attached virtual hard-drive storage; whereas S3 is more affordable object storage that is capable of holding large volumes of data. The D2 instances include sufficient bundled storage, and therefore EBS storage is unnecessary. The S3 storage is relatively low-cost and highly reliable, and therefore can serve as an alternative to relying on costly Hadoop replication for high-availability. For that reason, SAIC is assuming one copy of core data in the Hadoop cluster, whereas a copy of the data is stored in persistent S3 storage.

To support our above mentioned assumptions (See Section 3.3), SAIC estimates an initial rough-order-of-magnitude cost of about $309,909 for AWS-based hosting (including compute, storage and networking). To accommodate 20% growth per year, Utah would gradually increase its hosting requirements to approximately $536,663 by Year 5. The maintenance is included in the AWS pricing, with minimal support costs also included. The labor is approximately $50,000 for Year 1 (Setup); plus $450,000 per year ongoing; arriving at a ROM 5-year total of $4,367,174 for cloud-based hosting. The following figure summarized the cloud hosting costs.



AWS Service	Option	vCPU	Mem	Storage	orage ⁊	Comments	Per Hour	Units	Qty	Monthly Usage		Per Month
AWS Machine*												
__Management Nodes	r3.4xlarge	16	122	1 x 320 SDD	3.2	Red Hat/US West/Storage included	$1.400	Instances	1	728	Hrs/mo	$1,019
__Name/Job tracker Nodes	r3.4xlarge	16	122	2 x 320 SDD	3.2	Red Hat/US West/Storage included	$2.400	Instances	3	728	Hrs/mo	$5,242
__Slave/Data Nodes	d2.4xlarge	16	122	12 x 2000 HDD	168	Big Storage Node/US West (Oregon)	$2.760	Instances	7	728	Hrs/mo	$14,065
Attached Storage												
__EBS storage per node						Operational Storage included	$100	TB	0	100%		$0
Elastic MapReduce Service							$2,400	Instances	0	100%		$0
S3 Storage						Object Storage	$30	TB	150	100%		$4,500
Data Transfer												
__AWS Data Transfer In						Free Uploads	$0	TB	1	100%		$0
__AWS Data Transfer Out						.09 per GB	$90	TB	10	100%		$900
AWS Support	Business						$100		1	100%		$100
						Monthly Total						$25,825.76
										per year		$309,909.12



ROM Estimate 2 - Cloud Hosting on Amazon

WBS	Non-Recurring	2016	2017	2018	2019	2020	Total
1	**Planning**						
1.2	Develop Cloud Migtation Plan						-
1.3	Provide Government Labor						
1.4	Provide Contractor Labor						-
2	**Implementation**						
2.1	Procure Cloud Computing						-
2.2	Procure Software Licenses						-
2.3	Provide Government Labor						
2.4	Provide Contractor Labor	50,000					**50,000**
	Total Non-Recurring	**50,000**	**-**	**-**	**-**	**-**	**50,000**
WBS	**Recurring**	**2016**	**2017**	**2018**	**2019**	**2020**	**Total**
3	**Operations & Maintenace**						
3.1	Procure Cloud Computing	309,909	344,820	406,003	469,778	536,663	**2,067,174**
3.2	Procure Software Licenses						-
3.3	Procure Cloud Data Transfer						-
3.4	Provide Government Labor						
3.5	Provide Contractor Labor	450,000	450,000	450,000	450,000	450,000	**2,250,000**
	Total Recurring	**759,909**	**794,820**	**856,003**	**919,778**	**986,663**	**4,317,174**
	Total	**809,909**	**794,820**	**856,003**	**919,778**	**986,663**	**4,367,174**

Additional Assumptions:
The cloud-based hosting is based on Amazon's Public Cloud pricing. The Amazon GovCloud (closed community for government entities) pricing will vary. . Our pre-year estimates do not take into account details such as ramp up time for utilization of cloud services; and do not take into account the likely steady decline in cloud-based hosting costs over time. The prices will vary incrementally based on hosting with other public cloud providers: E.g. Google, Azure versus Amazon. All datacenter operations costs are included; as well as FedRAMP compliance investments.

Alternative Cloud Strategies for Reducing Costs

This estimate is based on a Hadoop cluster that runs 24/7. It is possible to take advantage of the cloud's unique capabilities to reap added performance and cost benefits. For example, alternative strategies include:

- Shutting down Hadoop clusters during non-core business hours. For example, a cluster that runs 10 hours every day might reduce costs by 30% or more
- Detecting and spinning up clusters when jobs are submitted – only paying for the compute time that the Hadoop cluster is running, plus low-cost storage
- Spinning up faster (for example 50 node) clusters during peak hours; and smaller-node (5 node) clusters during non-core hours
- taking advantage of "Hadoop as a Service" offerings to simplify/lower costs of setup, licensing and maintenance.

4.1.5 Recommendation

Having gone through the process to develop an apples-to-apples comparison for an on-premise cloud and a public cloud implementation, Task 4: Business Case demonstrates that there is not a significant difference in the five year cost between the two options. Based on our knowledge of cloud security, there is also no difference in your technical ability to secure either environment with similar labor plus software costs.

There are clear advantages to utilizing the AWS Public cloud - in terms of OpEx; lower cost with long term pricing agreements (as well as AWS price reductions); scalability; extensibility; ability to focus on the mission and not the machines; regular physical machine upgrades; and 24x7 uptime for the physical infrastructure.

There are advantages to a local on-premises cloud. Current regulations are written for an on-premise perimeter security so handling PII can use the accepted current procedures; there is an efficiency in always-on consistently high processing (which would incur higher charges on the cloud); and the local option provides a great transition sandbox.

Determining which solution, the public cloud or the on-premise cluster, is best truly depends on what the specific requirements are for your first project, and any regulatory issues in moving data to the cloud. The quick-win scenario could be the most immediate factor in your choice. Some sample criteria to consider are shown in Figure 4-3.

Factor	Either	Public Cloud	On-Premise Cluster
Cost	√		
Ability to Secure	√		
Perception of Security			√
OpEx easier than CapEx		√	
Regulatory Issues			√
Data Remains in US	√		

Maintenance		√	
Scalability		√	
Speed to expand		√	
Experiment w/ CPU/Memory/Storage		√	
Surge Capacity		√	
Cost effective for steady, high compute			√

Figure 4-3: Determining Factors

On the whole, the public cloud will give you easier growth and greater flexibility, but may have hindrances to deploying aspects some data and solutions depending on regulatory concerns.

While a public cloud is undoubtedly the end-state or part of the end-state, policies and regulations could require early attention and effectively slow down the push to get operational and get a “quick win”. While the State will have to decide on the tolerance for this extra potential complexity, there are advantages to early development being on an on-premises cluster. As a simplistic representation of a scenario, we suggest the graph in Figure 4-3. 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 actually the cloud management tools (and platform tools discussed in the next section) are fully capable of spanning a hybrid cloud using the single “pane of glass”. 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 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.

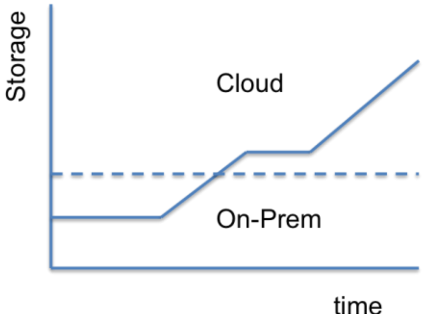


Figure 4-3: Time-phased approach

4.2 Platform Layer

In the same way that cloud technologies have seen an explosion in growth and adoption due to the cost-effectiveness in managing infrastructure. Beginning roughly in 2005, a paradigm shift occurred in Big Data engineering. The shift was to parallelize the data handling for data-intensive applications; much as the computational science community shifted to parallel applications in the 90s for their compute-intensive applications. The open source community has seen an explosion in the software tools available for building applications. Figure 4.2-1 shows Professor

Geoffrey Fox's representation of the Apache Big Data Stack. For all the tools mentioned, there are only roughly a dozen of these that pre-date 2005.

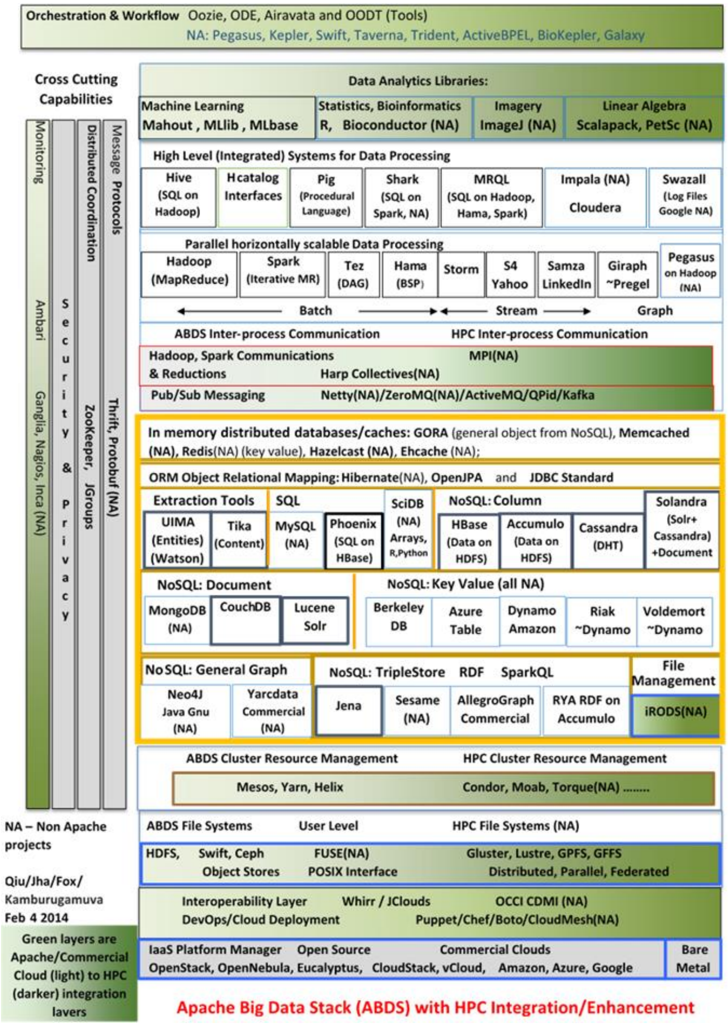

Orchestration & Workflow Oozie, ODE, Airavata and OODT (Tools)
NA: Pegasus, Kepler, Swift, Taverna, Trident, ActiveBPEL, BioKepler, Galaxy
Cross Cutting Capabilities
Monitoring
Ambari
Ganglia, Nagios, Inca (NA)
Security & Privacy
Distributed Coordination
ZooKeeper, JGroups
Message Protocols
Thrift, Protobuf (NA)
Data Analytics Libraries:
Machine Learning
Mahout , MLlib , MLbase
Statistics, Bioinformatics
R, Bioconductor (NA)
Imagery
ImageJ (NA)
Linear Algebra
Scalapack, PetSc (NA)
High Level (Integrated) Systems for Data Processing
Hive (SQL on Hadoop)
Hcatalog Interfaces
Pig (Procedural Language)
Shark (SQL on Spark, NA)
MRQL (SQL on Hadoop, Hama, Spark)
Impala (NA) Cloudera
Swazall (Log Files Google NA)
Parallel horizontally scalable Data Processing
Hadoop (MapReduce)
Spark (Iterative MR)
Tez (DAG)
Hama (BSP)
Storm
S4 Yahoo
Samza LinkedIn
Giraph ~Pregel
Pegasus on Hadoop (NA)
Batch
Stream
Graph
ABDS Inter-process Communication
HPC Inter-process Communication
Hadoop, Spark Communications & Reductions
MPI(NA)
Harp Collectives(NA)
Pub/Sub Messaging
Netty(NA)/ZeroMQ(NA)/ActiveMQ/QPid/Kafka
In memory distributed databases/caches: GORA (general object from NoSQL), Memcached (NA), Redis(NA) (key value), Hazelcast (NA), Ehcache (NA);
ORM Object Relational Mapping: Hibernate(NA), OpenJPA and JDBC Standard
Extraction Tools
UIMA (Entities) (Watson)
Tika (Content)
SQL
MySQL (NA)
Phoenix (SQL on HBase)
SciDB (NA) Arrays, R,Python
NoSQL: Column
HBase (Data on HDFS)
Accumulo (Data on HDFS)
Cassandra (DHT)
Solandra (Solr+ Cassandra) +Document
NoSQL: Document
MongoDB (NA)
CouchDB
Lucene Solr
NoSQL: Key Value (all NA)
Berkeley DB
Azure Table
Dynamo Amazon
Riak ~Dynamo
Voldemort ~Dynamo
NoSQL: General Graph
Neo4J Java Gnu (NA)
Yarcdata Commercial (NA)
NoSQL: TripleStore RDF SparkQL
Jena
Sesame (NA)
AllegroGraph Commercial
RYA RDF on Accumulo
File Management
iRODS(NA)
ABDS Cluster Resource Management
HPC Cluster Resource Management
Mesos, Yarn, Helix
Condor, Moab, Torque(NA) ……..
ABDS File Systems
User Level
HPC File Systems (NA)
HDFS, Swift, Ceph Object Stores
FUSE(NA) POSIX Interface
Gluster, Lustre, GPFS, GFFS Distributed, Parallel, Federated
Interoperability Layer
Whirr / JClouds
OCCI CDMI (NA)
DevOps/Cloud Deployment
Puppet/Chef/Boto/CloudMesh(NA)
IaaS Platform Manager Open Source
OpenStack, OpenNebula, Eucalyptus, CloudStack, vCloud,
Commercial Clouds
Amazon, Azure, Google
Bare Metal
NA – Non Apache projects
Qiu/Jha/Fox/ Kamburugamuva Feb 4 2014
Green layers are Apache/Commercial Cloud (light) to HPC (darker) integration layers
Apache Big Data Stack (ABDS) with HPC Integration/Enhancement

Figure 4.2-1: Spectrum of Open Source tools for both compute and data-intensive applications[4]

This chart is shown just to indicate the extent of the recent technology advancements in data-handling methods and tools. The shift to using these open source tools is industry-wide, as all vendors have adapted their tools to run against HDFS or Hadoop (=HDFS + MapReduce). Given the large number of tools, no one installs and runs them individually, but uses either COTS tools that leverage Hadoop, or open source tools that bundle a set of tools into a full platform. The data platform choices revolve around three main options.

A. **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.

Examples: Teradata, Oracle Exabyte

[4] Geoffrey Fox, private communication, February 2014

Pros	Cons
• Optimized performance • Single control for infrastructure and platform	• Upfront capital purchases • Facility space and costs • Investment into professional services to analyze, design and build the cluster • License maintenance cost • Vendor lock-in • Typically costly

B. **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.

Examples: Cloudera, Hortonworks, MapR

Pros	Cons
• Optimized performance • Single control for infrastructure and platform • Common staff skillsets for hiring and training • Community versions	• Version control can be an issue • Security not mature • Requires manual installation and updates

C. **Public Cloud Big Data Platform**: Involves leveraging cloud provider specific tools

Examples: AWS EC2, Redshift,

Pros	Cons
• Optimized performance • Single control for infrastructure and platform	• More specialized skillsets • Additional service usage charges • Vendor lock-in

Given the expense of COTS hardware/software appliances, and their vendor lock-in, Option A was not considered to be a viable option. While AWS (the market-leading cloud provider choice) has a number of specific data services, their use would restrict migration between public and private cloud, and incur additional service charges above the compute and storage instances. For these reasons, the recommendation is to stay with the open source platforms, which can be run on-premises or on the public cloud.

4.2.1 Major Platforms

While there are a number of vendors seeking to integrate a full platform out of the Apache stack, there are three main vendors, Cloudera, Hortonworks, and MapR. In each case the vendor offers a fully-functioning free edition. The companies make their money in the addition off extra tools for management convenience, and support. As the prices quote for the enterprise editions would be highly dependent on the configuration, government discount, and desired support options. We note that both Cloudera and Hortonworks are on GSA Schedule 70. In Appendix B we provide Gartner's 2015 assessment of the three companies.

Hadoop has been out long enough to have moved into what is being called their 2.0 version that uses Yarn to overcome the inherent latencies in MapReduce. The major cloud providers have images for the different platforms, and provide some of the same services themselves, such as AWS Elastic Map Reduce (EMR). This does however mean that there is an additional charge for using the cloud provider's version.

4.2.1.1 Cloudera

Cloudera is arguably the market-leading platform. Originally funded by the intelligence community angel investment program In-Q-Tel, Cloudera is the platform of choice in the intelligence and defense communities, and has a very large market penetration. They are estimated to have roughly a 51% share of the market in big data platforms. They have a large commercial presence as well, and have developed a number of innovations in open source components. While they utilize the vendor-neutral components of the "open core", such as HDFS, they have developed some proprietary tools such as Impala to obtain enhanced performance. Cloudera follows a mixed model, they were initially a services and support company that is migrating to do more selling of their Data Hub Editions.

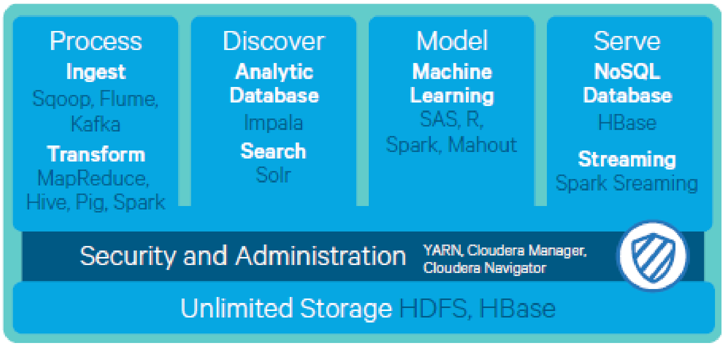
Versions[5]**:**

Express, Basic, Flex or Data Hub Editions

Pros:

Cloudera Management Suite[6]– Config, Manage, Monitor, etc.

Default Tools:



[5] http://www.cloudera.com/content/cloudera/en/products-and-services/product-comparison.html

[6] http://www.cloudera.com/content/cloudera/en/products-and-services/cloudera-enterprise/cloudera-manager/cloudera-manager-features.html

Deployment
Flexibility
On-Premises
Appliances
Engineered Systems
Public Cloud
Private Cloud
Hybrid Cloud

Figure 4.2-2: Cloudera toolset

4.2.1.2 Hortonworks

Hortonworks is fully open source, and has been working on partnerships to tune it for specific connections. Hortonworks has a large following, and has received a large boost in credibility given their investment from Microsoft, and the partnering relationship with Pivotal. They are fully open-source compliant, so they are dependent on the pace of open source development. Being open source, their primary revenue model is services and support.

Versions: Free, Enterprise, Enterprise Plus

Appears to be the same platform/tools across the versions. Enterprise and Enterprise Plus are a paid subscription support service.

Pros:

Microsoft connectors

Native Windows version available

Default Tools:

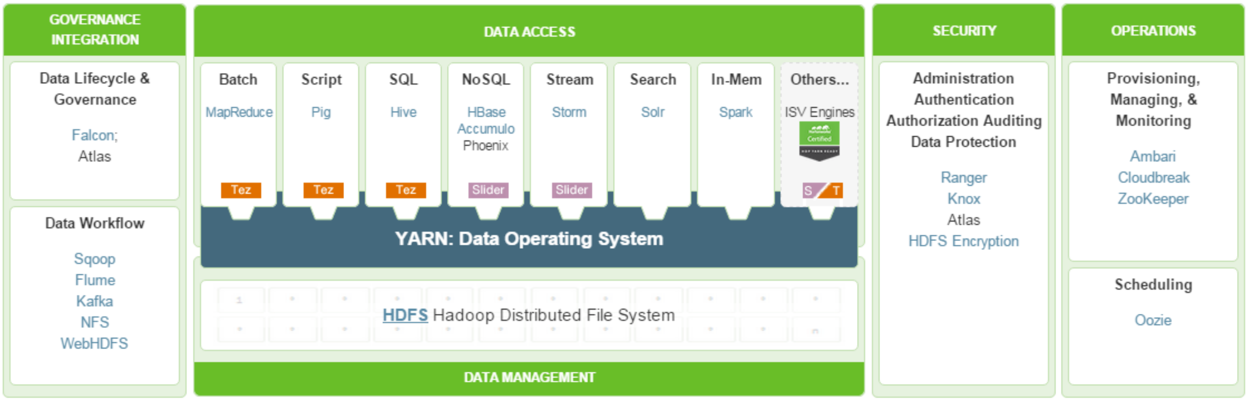


Figure 4.2-3: Hortonworks Toolset

4.2.1.3 MapR

MapR has chosen the route to developing proprietary elements in the "core" of the Apache open source stack. MapR has a reduced marketshare, but has a number of differentiated features in terms of security. In terms of the Gartner Report presented in Appendix A, it is viewed as equally innovative and capable to Cloudera.

Versions[7]: Community (M3), Enterprise (M5), or Enterprise Database (M7) Editions

M5 - For critical deployments requiring business continuity (HA/DR).

[7] https://www.mapr.com/products/mapr-distribution-editions

M7 - For critical deployments requiring (HA/DR) and NoSQL.

Pros:

Proprietary data format – Claimed to be more efficient than HDFS

Cons:

Proprietary data format – NOT HDFS

Default Tools[8]:

Not broken down by edition, M3, 5, or 7.

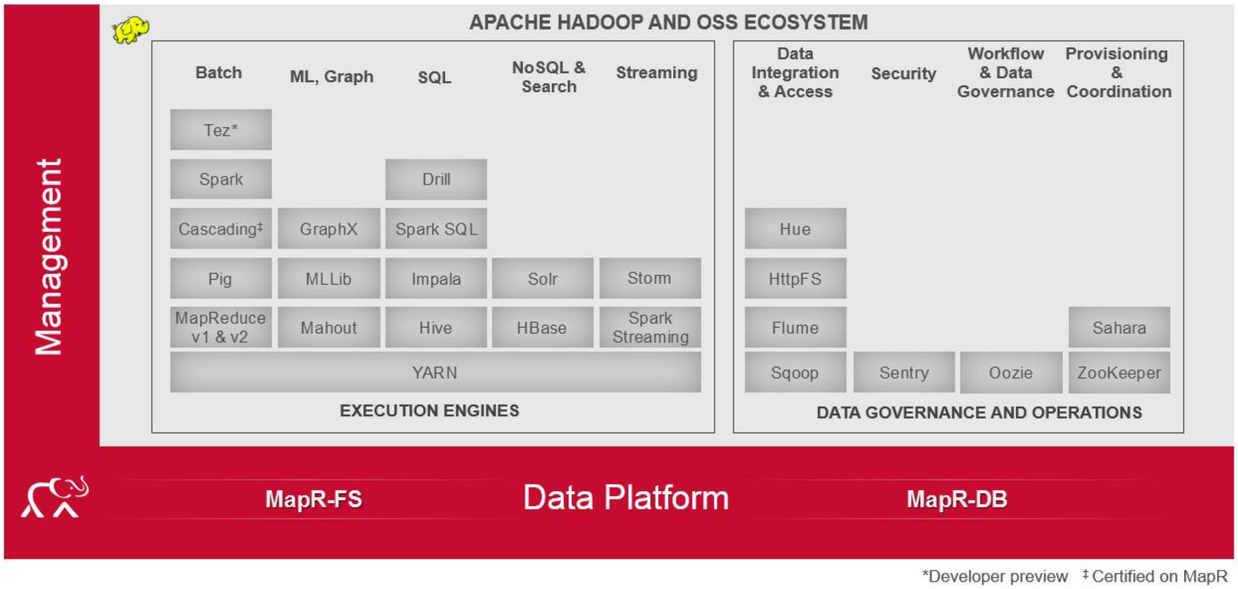


Figure 4.3-4: MapR Toolset

4.2.2 Recommendation

The three leading platforms are Cloudera, Hortonworks, and MapR based on partnerships, revenue, and marketshare. Each has strengths and weaknesses for different benchmarks algorithms and different cluster sizes.[9] Since the State is not a strong Microsoft shop, then Hortonworks is less of a choice with their integration with Azure and Microsoft tools. Should the State choose to go with Azure, Hortonworks could be more of a candidate. Given the marketshare, the use of the open source core (for the expansive community development advances), and the ability to use some enhanced performance tools we recommend Cloudera as a solid choice. The Cloudera manager is a capable management tool, and it can be used in on-premise and public cloud infrastructures, as well as span hybrid clouds. The only caveat, is that the State should review security requirements and compare Cloudera with MapR specifically for those capabilities before making the final choice.

4.3 Processing Layer

[8] https://www.mapr.com/products/whats-included

[9] See for example http://www.altoros.com/hadoop_benchmark.html

The data processing, or analytics application tools are the most difficult to specify since they are the most dependent upon the actual data and the desired analytics. The choices for this layer are easily changed and therefore not as critical in the initial startup of a Big Data capability. None-the-less some important aspects can be identified.

There are four components of the Big Data Application layer that are important architecture elements. While actually a part of the platform, the first choice centers around the database software to be used. The second choice is for the methods for blending the data. The third choice is in the analytics tools for the team's data scientists, and the fourth is the choice of tools for the State's business analysts.

4.3.1 Database

The imagery for a Data Lake is that data flows into this reservoir from the feeder systems, and is initially stored in HDFS, the Hadoop Distributed File System. We note that if AWS is the State's choice for its infrastructure there will be additional options such as S3 buckets for this initial storage. Within any of the platforms chosen above there is the choice of database. Open Source options consist of Hive and HBase; proprietary options consist of Cloudera's Impala and Amazon's Redshift. It is beyond the scope of this project to determine the best NoSQL platforms since this is highly dependent on the data types and characteristics of the datasets. We do note that there can be a significant difference in performance among these non-relational databases including those that operate in memory[10].

4.3.2 Data Blending

The typical data workflow will consist of moving the data from operational systems into the Data Lake. The platforms have a number of tools to accomplish the workflow. While not creating an EDW, it is still important to note that the metadata about the datasets is critically important. Data transferred as a snapshot must be refreshed so it does not go stale. Metadata can be used to record the workflow and understand the provenance of any data in the Data Lake. Care must be taken that while the data is being stored according to the organization of the source, it is imperative that the keys that will allow integration are identified. For security and privacy reasons identity fields can be masked, encrypted, or transformed. Whatever method is chosen it is important that this transformation is applied to all data sources so the datasets in the Data Lake can indeed be joined.

The most common scenario is to pull the data into the data lake, and then depending on the analytics task the data scientist is pursing with the business analyst, the Big Data Engineer or Data Scientist will likely pull the data together into a data mart, so the business analyst does not have to be aware of the storage structure of the individual datasets. It is this integration that allows "agile analytics" to rapidly generate and test candidate analytics, pursing those that initially look promising.

[10] https://amplab.cs.berkeley.edu/benchmark/

Another approach for data blending is to look at data virtualization solutions for creating a view in between the analytics and the data. This ensures the analytics are insulated against changes made in the underlying storage.

There are times when it is not possible or practical to pull all the data from an operational resource. In this case, the data would be pulled dynamically from the source when needed. This is Gartner's concept of a Logical Data Warehouse[11]. SAIC has developed a proof of concept for a LDW through the controlling usage of metadata, there are no mature solutions that implement this concept.

4.3.3 Data Science Analytics

The Hadoop ecosystem has now matured into Hadoop 2.0. The basic shift is the realization that the batch processing of MapReduce is very slow, too slow for many applications. Consequently both YARN and Mesos have been developed to address this issue, with current attempts to integrate the two approaches. What has become clear in the last year is that the Spark framework for analytics has been shown to be the clear "winner" in the analytics development area. Spark is an open source data analytics framework that provides an abstraction layer among a number of data sources, and provides the ability to implement SQL-like queries, high velocity streaming analytics, machine learning, or graphical analysis pipelines, as shown in Figure 4-4.

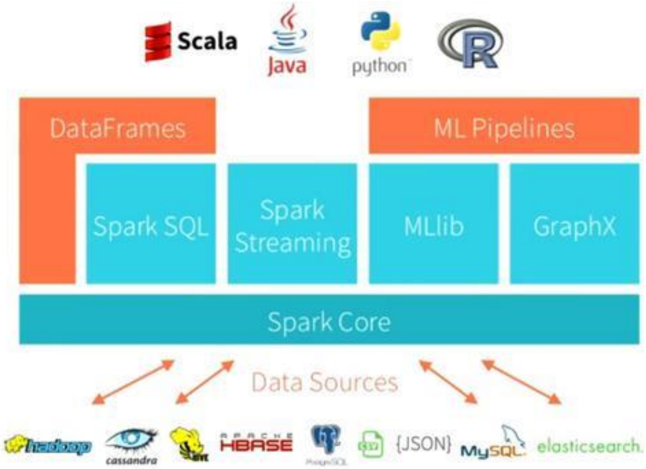


Figure 4.3-1: Spark Framework[12]

You can also leverage the data science programming languages of choice, Python and R. (Note that if a data scientist has a statistics background they likely use R, and if they come from a programming background they use Python. Both have large libraries of modules to make tasks easier.) Like Redhat, Spark has a company built

[11] Beyer, Mark A., Edjlali, Roxane, "Understanding the Logical Data Warehouse: The Emerging Practice", G00234996, 6/21/2012.

[12] http://www.informationweek.com/big-data/big-data-analytics/apache-spark-3-promising-use-cases/a/d-id/1319660

around it named Databricks. They are the largest committers to the Spark open source project and are continuing to develop the software and provide services to those wanting to use it.

4.3.4 Business Analytics

The business analyst would be expected to work with the data scientist as a team. The business analyst would provide the domain expertise for that area, as well as an understanding of the analytical needs. The data scientist would provide assistance with the data lake and the creation of any data marts needed for analysis. The important element is that the business analyst should not be dependent on the data scientist to test out their hypotheses. In the old tradition, business analysts would need to ask an IT report-writer to create the report they needed, and wait until the report was available. To avoid this legacy pitfall, business analysts should be provided the tools to access the data in the Data Lake or data mart, and investigate the data for themselves.

There are a number of business intelligence (BI) tools that can work here. Tableau is the leading BI tool of choice. It is a commercial tool that has a desktop version for creating analytical documents, a server for sharing these documents, and a reader for managers to get familiar with the data presentation. Tableau is not strong in the Extract-Transform-Load (ETL) portion of connecting to the data, but the interface is intuitive, powerful, and fairly easy to pick up. Pentaho is an open source version that is strong on the ETL side, but weaker in the BI options and ease of use.

In spit of the strength of the BI tools, there is no question that Excel is still the dominant tool for business analysts. If this is the preference for Utah's analysts, then additional work needs to be done to prepare data marts with sampled or summarized data to be able to fit in the limits of excel.

In each of these cases, the choices are highly dependent on the current analytical tools that are already in place, and on the skills and background of the analysts.

5 Security and Privacy Fabric

Given the critical importance of security and privacy as Utah creates a Data Lake, the critical aspects are discussed separately from the individual layers discussed in Section 4.

5.1 Security Conservation Principle

Federal, state and local government agencies need a structured, yet flexible approach for managing the portion of risk resulting from the incorporation of cloud-based information systems into the mission and business processes of the organization. The State of Utah Agencies as Consumers can require cloud Providers to implement all steps in the Risk Management Framework (RMF) process. The only exception is the security authorization step, as that remains an inherent Federal responsibility that is directly linked to the management of risk related to the use of cloud services. The core concept of the Security Conservation Principle, first introduced in the SP 500-299: NIST Cloud Computing Security Reference Architecture is that, for a particular service migrated to the cloud, the full set of necessary Security Components and controls that should be implemented to secure the cloud computing environment is always the same; however, the division of responsibility for those Components and controls changes based upon the characteristics of the cloud, particularly the service deployment. Figure 5-1 depicts the *Security Conservation Principle* for the cloud computing environment.

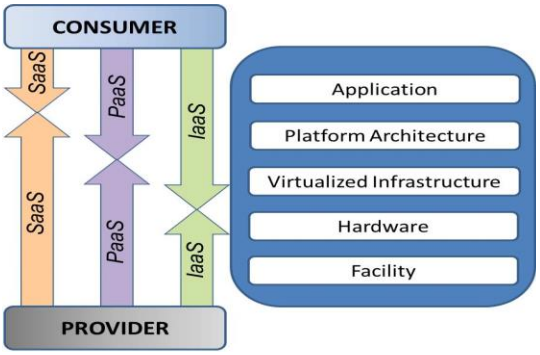
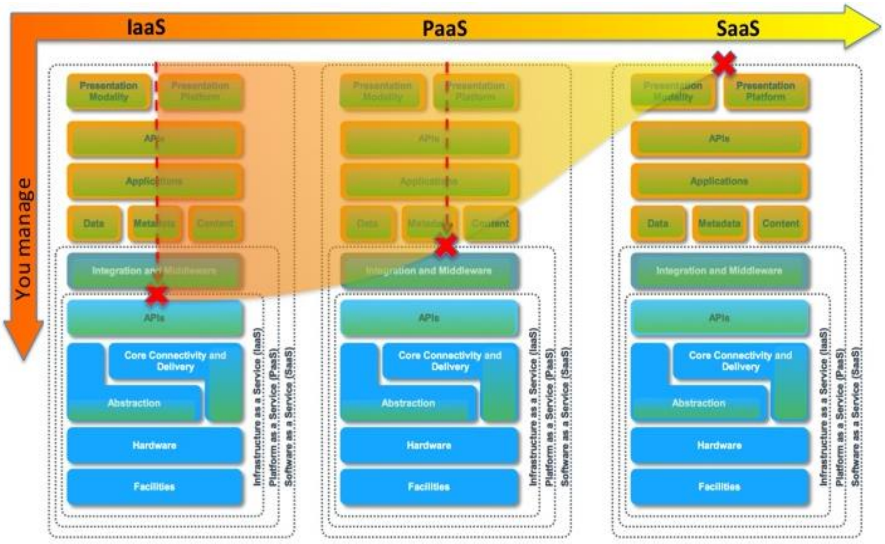


Figure 5-1: Security Conservation Principle

Figure 5-2 provides a high level overview of the responsibility for implementation may not rest solely with the Consumer, it is ultimately the Consumer's responsibility to exercise due diligence and specify what controls must be put in place and to provide oversight and accountability for the security of the cloud Ecosystem. In all of the Cloud deployment models previously described, a Consumer must be able to define his security requirements and visibility necessary into each environment to ensure the controls are working as advertised and that they meet their Regulatory requirements. For Hybrid Cloud environments a Consumer must be able to ensure that the same security controls (network separation, inter-VM firewalls, location-based High-Availability and Fault-Tolerance, Data Loss Prevention, Data at Rest Encryption, Regulatory, etc.) to protect their data will be available across Private and Public Clouds based on the available consumer/provider security control points. This assurance can take place via manual controls (Consumer configuration


IaaS
PaaS
SaaS
You manage
Presentation Modality
Presentation Platform
APIs
Applications
Data
Metadata
Content
Integration and Middleware
APIs
Core Connectivity and Delivery
Abstraction
Hardware
Facilities
Infrastructure as a Service (IaaS)
Platform as a Service (PaaS)
Software as a Service (SaaS)

The image of the cloud layered stack,, Cloud Security Alliance specification, 2009

Figure 5-2: Implementation Responsibility

or verification) or through a cloud Broker service that can match Consumer Security requirements to Cloud Providers that can meet those requirements.

The *Cloud-adapted Risk Management Framework* (CRMF) provides a risk-based approach to security control selection and specification considers effectiveness, efficiency, and constraints due to applicable laws, directives, Executive Orders, policies, standards, or regulations. While the framework is flexible and easily adaptable in most cases, it assumes a traditional IT environment, and requires some customization to address the unique characteristics of cloud-based services. The risk management is a cyclically executed process comprising of a set of coordinated activities for overseeing and controlling risks. This set of activities is composed of:

- Risk assessment,
- Risk treatment, and
- Risk control tasks that collectively target the enhancement of strategic and tactical security.

How confident cloud costumers feel about whether the amount of risk related to using cloud services is acceptable depends on how much trust they place on those involved in the surrounding cloud ecosystem's orchestration. The risk management process ensures that issues are identified and mitigated early on in the investment cycle and followed by periodic reviews. Since cloud customers, and other cloud actors involved in securely orchestrating a cloud ecosystem have differing degrees of control over cloud-based IT resources, they need to share the responsibility of implementing the security requirements. Regardless of the deployment model or service type, cloud consumers need to identify the threats, perform a risk assessment, and evaluate the security requirements of their individual cloud architectural context. The requirements also need to be mapped to the proper security controls and practices in the technical, operational, and management classes.

The type of cloud delivery model and the service type that are chosen for adoption, in association with security controls selected for the ecosystem need to be selected in such a way that the system in preserving its security posture. Therefore, a properly performed risk management cycle should ensure that the residual risk remaining after securing the ecosystem is minimal and that the system achieves a security posture equivalent to the security posture of an on premise technology architecture or solution. Conversely, the type of deployment model that is selected does have an impact on the distribution of security responsibilities among the cloud actors, which relates to the security conservation principle as discussed in the NIST Special Publication 500-299. While the risk management framework is adaptable to most scenarios, it defaults to the traditional IT environment and requires customization to successfully address the unique characteristics of cloud-based services and solutions.

5.2 Cloud Information Systems Boundaries

The cloud ecosystem is complex and dynamic. Security management scope varies based upon specific objectives and assurances needed. It may be from a high-level business perspective, by role classification or lower, more technical boundaries. Additionally, it may also have complex partner options, including cascading partners. Many dependencies may not be readily apparent. When an organization assumes the role of cloud consumer to access cloud-based IT resources, it needs to extend its *trust* beyond the physical boundary of the organization to include parts of the cloud environment.

- An *organizational boundary* represents the physical perimeter that surrounds a set of IT resources that are owned and governed by an organization. The organizational boundary does not represent the boundary of an actual organization, only an organizational set of IT assets and IT resources.
- A *trust boundary* is a logical perimeter that typically spans beyond physical boundaries to represent the extent to which IT resources are trusted. When analyzing cloud ecosystems, the *trust boundary* is most frequently associated with the trust issued by the organization acting as the cloud consumer. Another type of boundary relevant to cloud environments is the logical network perimeter. This type of boundary is classified as a cloud computing mechanism. The business perspectives of the *trust boundary* are significantly different from more technical perspectives, where underlying layers become more visible. Security risks can occur around technical and non-technical considerations, at the business boundary, but also at the sub-layers in the cloud ecosystem.

5.3 Cloud Security & Operations Management (CS&OM)

The SAIC team recommends that CS&OM is broken down into seven logical domains, which follows the NIST, CSA, CNSS, JARM, DHS, and GSA industry best practices for CS&OM, which are: (1) Business Operational Support Services (BOSS), (2) Information Technology Support Services (ITOS), (3) Security and Risk Management (SRM), (4) Presentation Services, (5) Application Services, (6) Information Services and (7) Infrastructure Services see detailed description from the high side description provided above. SAIC has a *Cloud Migration Edge* (CME) methodology that aids in providing tools, processes, and best practices for the entire migration life cycle. In addition to applying Information Technology Infrastructure Library (ITIL) best practices, CME supports the systematic production of System Development Life Cycle (SDLC) documentation. It supports the step-by-step implementation of a net-centric, web-enabled cloud computing/virtualization environment by explicitly breaking down the cloud/virtualization migration process into standardized components. Figure 5-3 illustrates the CME methodology.

The selection of Cloud security products are mapped against know threat vectors and provides for faster analytical queries and aggregation than traditional tools, particularly on larger data sets. This toolset, can deliver sophisticated log analytics, real-time monitoring, search and analytics, and is coupled with a dashboard for stored queries, reports and alerts. The State of Utah can derive meaningful insights from terabytes of log data and correlate events across multiple tiers of the Low environment in a single location, cutting down troubleshooting times, improving operational efficiency and reducing IT costs.

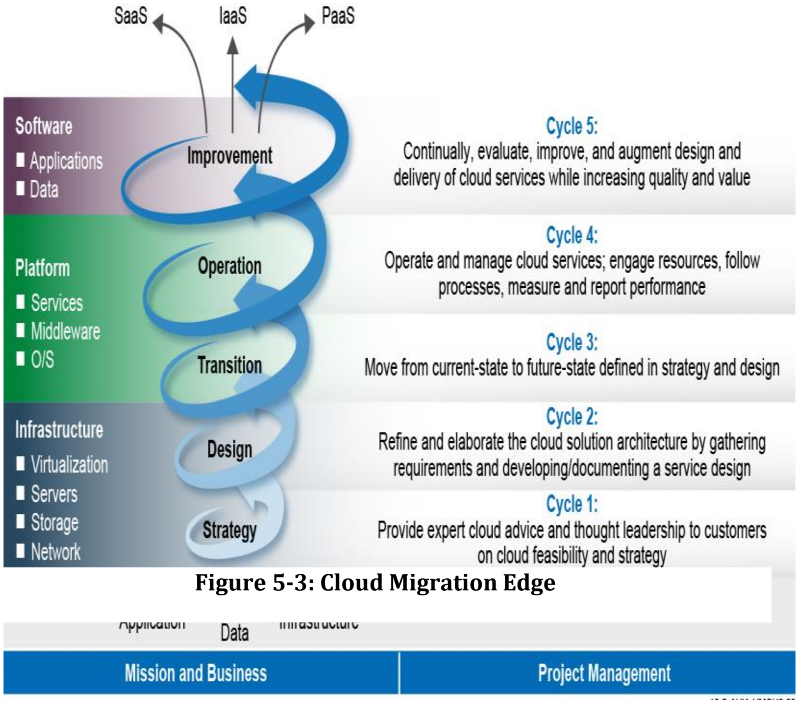

SaaS
IaaS
PaaS
Software
Applications
Data
Platform
Services
Middleware
O/S
Infrastructure
Virtualization
Servers
Storage
Network
Improvement
Operation
Transition
Design
Strategy
Cycle 5:
Continually, evaluate, improve, and augment design and delivery of cloud services while increasing quality and value
Cycle 4:
Operate and manage cloud services; engage resources, follow processes, measure and report performance
Cycle 3:
Move from current-state to future-state defined in strategy and design
Cycle 2:
Refine and elaborate the cloud solution architecture by gathering requirements and developing/documenting a service design
Cycle 1:
Provide expert cloud advice and thought leadership to customers on cloud feasibility and strategy
Figure 5-3: Cloud Migration Edge
Data
Mission and Business
Project Management

Figure 5-3: Cloud Migration Edge

Security Monitoring Data Management capability includes:
1) Session Events,
2) Authorization Events,
3) Authentication Events,
4) Application Events,
5) Network Events,
6) Computer Events,
7) Network Intrusion Prevention Services Events,
8) Privilege Usage Events,
9) eDiscovery Events,
10) Data Leakage Prevention Events,
11) Host-based Intrusion Prevention System Events,
12) Compliance Monitoring,
13) Certificate Revocation List,
14) Access Control Lists,
15) Database Events and
16) Transformation Services.

5.3 Roadmap for Implementation:

1. The identification all known internal and external compliance and governance standards mapped against the NIST SP 800-53, dated April 14, 2014.

2. Establish the adoption of the Cloud Security Alliance (CSA) & NIST Cloud Security Risk Architecture for initial security controls baseline to include FedRAMP low and moderate baselines.

3. Conduct risk and gaps analysis against known security compliance and regulatory oversight and begin the security controls tailoring as depicted in the current NIST guidance SP 800-53.

4. Upon completion of step 3, the development of all known cloud services mapped to security controls to determine the Confidentiality, Integrity and Assurance required for Utah.

5. Develop a phase security approach to include Provisional Authority To Operate (PATO) and Authority to Operate (ATO) as outlined in security compliance standards.

6. Develop the security tools list by each Cloud service model such as IaaS, PaaS and SaaS. Figure 5-4 provides the generic Service Level Agreement, with common Governance, Regulatory and Compliance (GRC) mapping with a common security model taxonomy.

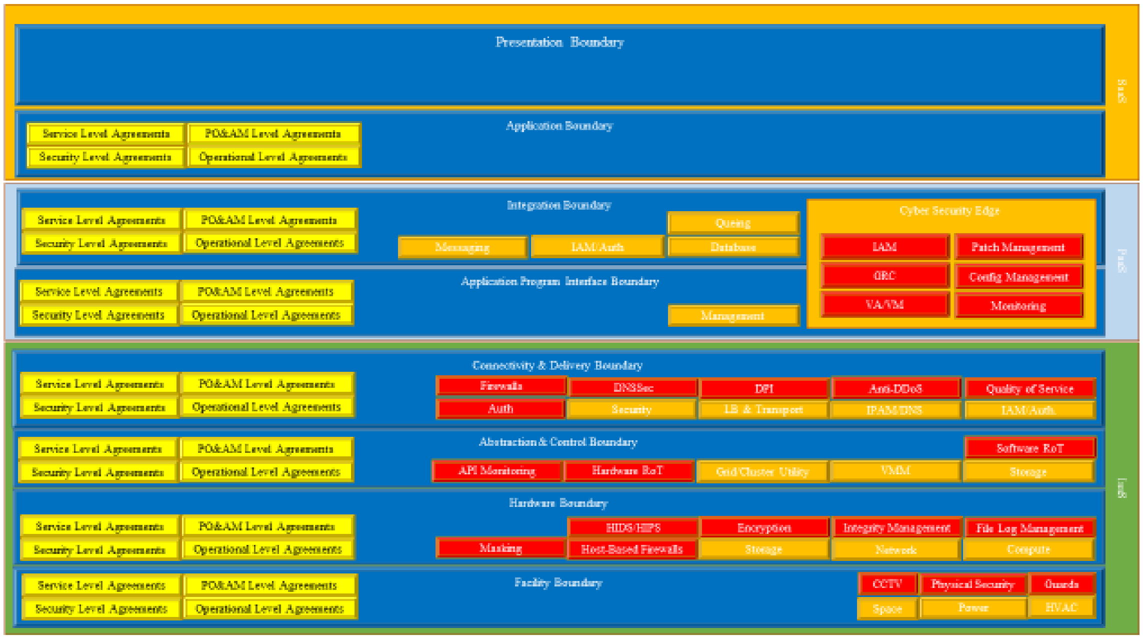


Figure 5-4: Mapping to a Security Model Taxonomy

Appendix A: Gartner Platform Comments

Gartner lists the three main vendors in their 2015 Magic Quadrant.[13] Note that this chart applies to both traditional RDBMS solutions as well as the new NoSQL solutions.

Cloudera and MapR are judged fairly similar. Pivotal is migrating to using Hortonworks, adapting their Greenplum database to function as an analytics data mart. Hortonworks has also established a partnership with Microsoft, which should increase its prominence. For the purposes of this chart, however, Gartner did not judge Hortonworks to qualify for their chart.

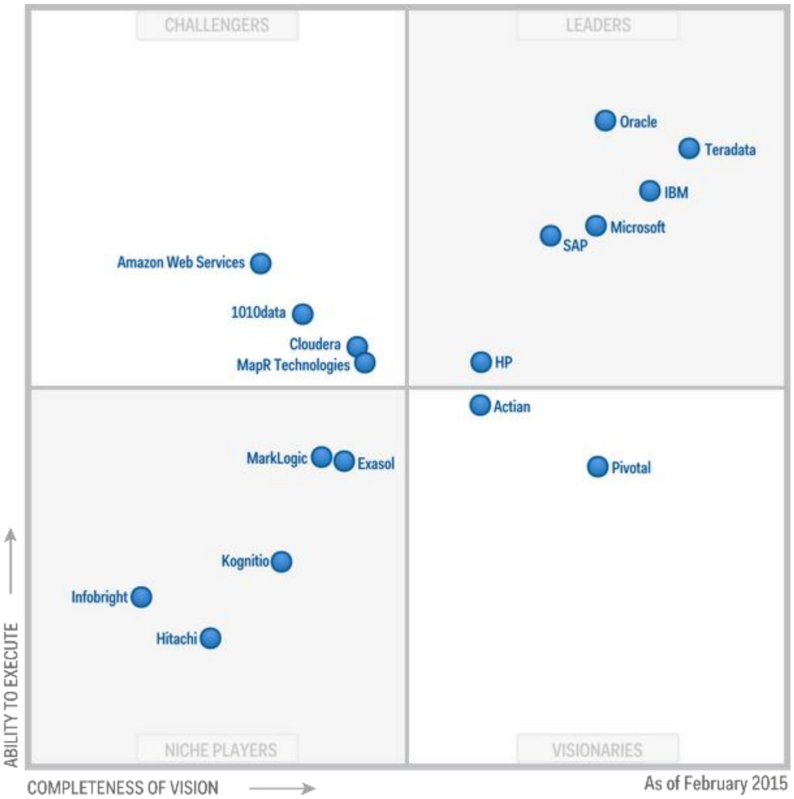


Figure 4-1: Gartner 2015 Magic Quadrant

A.1 Cloudera

[13] Gartner “Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics”, G00263133, Feb 2, 2015

Cloudera (www.cloudera.com) provides a data storage and processing platform based upon an Apache Hadoop Project, as well as proprietary system and data management tools for design, deployment, operation and production management.

Strengths

- Cloudera is tightly focused on embedding customer experiences in emerging product functionality such as with Impala, Cloudera Director or Cloudera Navigator — and a focus on intellectual property in solutions and software instead of consulting and professional services. This is an important shift from its early days and adds value to its offerings.
- Unlike many emerging solutions, Cloudera benefits from a large set of BI and data integration partners. This eases adoption for organizations with existing investments in these tools. It has also been able to partner with large traditional players in this market such as Teradata or Microsoft.
- Customer references reinforce Cloudera's execution strategy, reporting excellent support and services, broad interoperability with industry standard tools, and a high-speed performance. This is specific to analytics use-case customers.

Cautions

- The combination of a lack of metadata management, available skills in the market and difficulties in finding standard approaches to loading the data make Cloudera adoption slow and sometimes painful — according to its references.
- The uneven mix of its user base, sometimes for repetitive batch run processing and at other times heavily loaded with data analyst and data scientist users, makes it difficult for Cloudera's solution to find a "home base" in the market for DMSAs. However, the user base is maturing along with the tools and this is expected to be a short-lived issue.
- With only the beginnings of growth in EMEA and Asia/Pacific, Cloudera remains a vendor with largely North American experience. We anticipate further growth in both these other large regions during the next two years.

A.2 HortonWorks

Located in Palo Alto, California, U.S., and founded in 2011, Hortonworks markets the Hortonworks Data Platform (HDP), derived entirely from the open-source Apache Hadoop stack. The company was a leading contributor to Hive for SQL interfacing. HDP includes services to support security, data governance and operations. Hortonworks participates in the "Stinger" Initiative to advance Apache Hive for interactive query capabilities and claims HDP enables interactive query operations at the petabyte scale.

A.3 MapR Technologies

Founded in 2009, MapR Technologies (www.mapr.com) offers a Hadoop distribution with performance and storage optimizations, high availability improvements and administrative and management tools. It offers training and education services.

Strengths

- Based on information provided by MapR, it is the Hadoop distribution vendor with the largest number of paying customers — which is an important indication of adoption for an emerging category of technology.
- MapR's strategy is to deliver a data platform that combines Hadoop and operational database technologies to support a wide range of workloads in a single deployment. To enable this strategy, the company has compensated for Hadoop deficiencies by creating alternatives to Apache components while including a number of open-source projects from other distributions. For example, it substitutes Hadoop Distributed File System (HDFS) with its Posix-compliant, standard Network File System (NFS) file system, and it also supports Impala. This inclusive strategy offers customers the greatest number of options.
- MapR is praised by references for its reliability, performance and scalability, making it a solution suitable for enterprise use.

Cautions

- MapR has a smaller partner ecosystem than the other Hadoop distribution vendors. For example, the number of DBMS, BI or data integration partners is modest. However, MapR is actively addressing this by recently adding partnerships with Teradata, SAS and HP Vertica (for example).
- Reference customers struggle to find enough skilled resources in the market. The growing interest in Hadoop will help to relieve some of this pressure, but this is a multiyear cycle rather than one measured in quarters or months.
- Reference customers indicate that it can take time for MapR to support the latest Hadoop capabilities, although it can support multiple versions of the same Hadoop project. To address this concern, MapR accelerated its ecosystem update process during March 2013, and now has monthly Hadoop ecosystem releases.

Appendix B: Technological Readiness[14]

The technological readiness for Big Data serves as metric useful in assessing both the overall maturity of a technology across all implementers as well as the readiness of a technology for broad use within an organization. Technology readiness evaluates readiness types in a manner similar to that of technology readiness in Service-Oriented Architectures (SOA). However, the scale of readiness is adapted to better mimic the growth of open source technologies, notably those which follow models similar to the Apache Software Foundation (ASF). Figure 1 provides a superimposition of the readiness scale on a widely recognized "hype curve." This ensures that organizations which have successfully evaluated and adopted aspects of SOA can apply similar processes to assessing and deploying Big Data technologies.

B.1 Types of Readiness

- **Architecture**: Capabilities concerning the overall architecture of the technology and some parts of the underlying infrastructure
- **Deployment**: Capabilities concerning the architecture realization infrastructure deployment, and tools
- **Information**: Capabilities concerning information management: data models, message formats, master data management, etc.
- **Operations, Administration and Management**: Capabilities concerning post-deployment management and administration of the technology

B.2 Scale of Technological Readiness

1. **Emerging**
   - Technology is largely still in research and development
   - Access is limited to the developers of the technology
   - Research is largely being conducted within academic or commercial laboratories
   - Scalability of the technology is not assessed
2. **Incubating**
   - Technology is functional outside laboratory environments
   - Builds may be unstable
   - Release cycles are rapid
   - Documentation is sparse or rapidly evolving
   - Scalability of the technology is demonstrated but not widely applied
3. **Reference Implementation**
   - One or more reference implementations are available
   - Reference implementations are usable at scale

[14] Thanks to Dan McCreary of the NIST Big Data Working Group for his original draft of this assessment.

- The technology may have limited adoption outside of its core development community
- Documentation is available and mainly accurate

4. **Emerging Adoption**
   - Wider adoption beyond the core community of developers
   - Proven in a range of applications and environments
   - Significant training and documentation is available
5. **Evolving**
   - Enhancement-specific implementations may be available
   - Tool suites are available to ease interaction with the technology
   - The technology competes with others for market share
6. **Standardized**
   - Draft standards are in place
   - Mature processes exist for implementation
   - Best practices are defined