**NIST Special Publication 1500-9**

**DRAFT NIST Big Data Interoperability Framework: Volume 9, Adoption and Modernization**

NIST Big Data Public Working Group

Technology Roadmap Subgroup

Draft Version 2,

March 20, 2017

[http://dx.doi.org/10.6028/-------](http://dx.doi.org/10.6028/NIST.SP.1500-1)



NIST Special Publication 1500-9

Information Technology Laboratory

**DRAFT NIST Big Data Interoperability Framework:**

**Volume 9, Adoption and Modernization**

**Draft Version 2**

NIST Big Data Public Working Group (NBD-PWG)

Technology Roadmap Subgroup

National Institute of Standards and Technology

Gaithersburg, MD 20899

[http://dx.doi.org/10.6028/------](http://dx.doi.org/10.6028/NIST.SP.1500-1)

March 2017



U. S. Department of Commerce

*Wilbur L. Ross, Jr., Secretary*

National Institute of Standards and Technology

*Dr. Kent Rochford, Acting Under Secretary of Commerce for Standards and Technology*

*and Acting NIST Director*

**National Institute of Standards and Technology (NIST) Special Publication 1500-9**

pages (March 20, 2017)

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

There may be references in this publication to other publications currently under development by NIST in accordance with its assigned statutory responsibilities. The information in this publication, including concepts and methodologies, may be used by Federal agencies even before the completion of such companion publications. Thus, until each publication is completed, current requirements, guidelines, and procedures, where they exist, remain operative. For planning and transition purposes, Federal agencies may wish to closely follow the development of these new publications by NIST.

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. All NIST publications are available at <http://www.nist.gov/publication-portal.cfm>.

**Comments on this publication may be submitted to Wo Chang**

National Institute of Standards and Technology

Attn: Wo Chang, Information Technology Laboratory

100 Bureau Drive (Mail Stop 8900) Gaithersburg, MD 20899-8930

Email: [SP1500comments@nist.gov](mailto:SP1500comments@nist.gov)

Request for Contributions

The NIST Big Data Public Working Group (NBD-PWG) requests contributions to this draft version 2 of the *NIST Big Data Interoperability Framework Volume 9, Adoption and Modernization.* All contributions are welcome, especially comments or additional content for the current draft.

The NBD-PWG is actively working to complete version 2 of the set of NBDIFdocuments. The goals of version 2 are to enhance the version 1 content, define general interfaces between the NIST Big Data Reference Architecture (NBDRA) components by aggregating low-level interactions into high-level general interfaces, and demonstrate how the NBDRA can be used.

To contribute to this document, please follow the steps below as soon as possible but no later than May 1, 2017.

1. Register as a user of the NIST Big Data Portal (https://bigdatawg.nist.gov/newuser.php)
2. Record comments and/or additional content in one of the following methods:
   1. **TRACK CHANGES**: make edits to and comments on the text directly into this Word document using track changes
   2. **COMMENT TEMPLATE**: capture specific edits using the Comment Template (<http://bigdatawg.nist.gov/_uploadfiles/SP1500-1-to-7_comment_template.docx>), which includes space for Section number, page number, comment, and text edits
3. Submit the edited file from either method above to SP1500comments@nist.gov with the volume number in the subject line (e.g., Edits for Volume 1).
4. Attend the weekly virtual meetings on Tuesdays for possible presentation and discussion of your submission. Virtual meeting logistics can be found at https://bigdatawg.nist.gov/program.php

Please be as specific as possible in any comments or edits to the text. Specific edits include, but are not limited to, changes in the current text, additional text further explaining a topic or explaining a new topic, additional references, or comments about the text, topics, or document organization.

The comments and additional content will be reviewed by the subgroup co-chair responsible for the volume in question. Comments and additional content may be presented and discussed by the NBD-PWG during the weekly virtual meetings on Tuesday.

Three versions are planned for the NBDIF set of documents, with Versions 2 and 3 building on the first. Further explanation of the three planned versions and the information contained therein is included in Section 1.5 of each NBDIF document.

Please contact Wo Chang ([wchang@nist.gov](mailto:wchang@nist.gov))with any questions about the feedback submission process.

Big Data professionals are always welcome to join the NBD-PWG to help craft the work contained in the volumes of the NBDIF. Additional information about the NBD-PWG can be found at <http://bigdatawg.nist.gov>. Information about the weekly virtual meetings on Tuesday can be found at https://bigdatawg.nist.gov/program.php.

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at NIST promotes the U.S. economy and public welfare by providing technical leadership for the Nation’s measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology (IT). ITL’s responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in Federal information systems. This document reports on ITL’s research, guidance, and outreach efforts in IT and its collaborative activities with industry, government, and academic organizations.

Abstract

An abstract will be written when Volume 9 is close to completion.

Keywords

The keywords will be compiled when Volume 9 is close to completion.

Acknowledgements

This document reflects the contributions and discussions by the membership of the NBD-PWG, co-chaired by Wo Chang (NIST ITL), Nancy Grady (SAIC), Geoffrey Fox (University of Indiana), Arnab Roy (Microsoft), Mark Underwood (Krypton Brothers), David Boyd (InCadence Corp), Russell Reinsch (Loci), and Gregor von Laszewski (University of Indiana).

The document contains input from members of the NBD-PWG Technology Roadmap Subgroup, led by Russell Reinsch (CFGIO); Definitions and Taxonomies Subgroup, led by Nancy Grady (SAIC); Use Cases and Requirements Subgroup, led by Geoffrey Fox (University of Indiana); Security and Privacy Subgroup, led by Arnab Roy (Fujitsu) and Mark Underwood (Krypton Brothers); and Reference Architecture Subgroup, led by David Boyd (InCadence Strategic Solutions), .

NIST SP1500-9, Version 2 has been collaboratively authored by the NBD-PWG. As of the date of this publication, there are \_ NBD-PWG participants from industry, academia, and government. Federal agency participants include the National Archives and Records Administration (NARA), National Aeronautics and Space Administration (NASA), National Science Foundation (NSF), and the U.S. Departments of Agriculture, Commerce, Defense, Energy, Health and Human Services, Homeland Security, Transportation, Treasury, and Veterans Affairs.

NIST acknowledges the specific contributions[[1]](#footnote-1) to this volume by the following NBD-PWG members:

A list of contributors to version 2 of this volume will be added here.

The editors for this document were Russell Reinsch and Wo Chang

Table of Contents

[Executive Summary viii](#_Toc478543728)

[1 Introduction 9](#_Toc478543729)

[1.1 Background 9](#_Toc478543730)

[1.2 NIST Big Data Public Working Group 10](#_Toc478543731)

[1.3 Scope and Objectives of the Technology Roadmap Subgroup 11](#_Toc478543732)

[1.4 Report Production 11](#_Toc478543733)

[1.5 Future Work on this Volume 11](#_Toc478543734)

[2 Landscape Perspective 12](#_Toc478543735)

[2.1 Big Data Trends and Forecasts 12](#_Toc478543736)

[3 Adoption and Barriers 13](#_Toc478543737)

[3.1 Exploring Big Data Adoption 13](#_Toc478543738)

[3.1.1 Adoption by Industry 13](#_Toc478543739)

[3.1.2 Functional Perspective of Adoption 13](#_Toc478543740)

[3.2 Technical and non-technical Barriers to adoption 14](#_Toc478543741)

[3.2.1 Non-Technical Barriers 14](#_Toc478543742)

[3.2.2 Technical barriers to adoption 16](#_Toc478543743)

[4 Maturity 18](#_Toc478543744)

[4.1 functional and Technologic Maturity 18](#_Toc478543745)

[4.2 Organizational Maturity 19](#_Toc478543746)

[4.3 Big Data Trends and Forecasts 19](#_Toc478543747)

[5 Considerations for Implementation and Modernization 21](#_Toc478543748)

[5.1 Implementation 21](#_Toc478543749)

[5.2 System Modernization 21](#_Toc478543750)

[5.3 Recommendations for Organizational Modernization 23](#_Toc478543751)

[Appendix A: Acronyms A-1](#_Toc478543752)

[Appendix B: References B-1](#_Toc478543753)

List of Figures

[Figure 1: New System Implementation 22](#_Toc478543754)

[Figure 2: Requirement Considerations 23](#_Toc478543755)

List of Tables

[Table 1: Spending by Industry 12](#_Toc478543756)

[Table 2: Non-Technical and Technical Barriers to Adoption 14](#_Toc478543757)

[Table 3: Non-Technical Barriers to Adoption 15](#_Toc478543758)

[Table 4: Technical Barriers to Adoption 16](#_Toc478543759)

[Table 5: Maturity Projections 19](#_Toc478543760)

[Table 6: Possible Directions for Modernization 22](#_Toc478543761)

Executive Summary

This section will be written when the document content is near finalization. The NIST Big Data Public Working Group created this volume during stage 2 activities.

# Introduction

## Background

There is broad agreement among commercial, academic, and government leaders about the remarkable potential of Big Data to spark innovation, fuel commerce, and drive progress. Big Data is the common term used to describe the deluge of data in today’s networked, digitized, sensor-laden, and information-driven world. The availability of vast data resources carries the potential to answer questions previously out of reach, including the following:

* How can a potential pandemic reliably be detected early enough to intervene?
* Can new materials with advanced properties be predicted before these materials have ever been synthesized?
* How can the current advantage of the attacker over the defender in guarding against cyber-security threats be reversed?

There is also broad agreement on the ability of Big Data to overwhelm traditional approaches. The growth rates for data volumes, speeds, and complexity are outpacing scientific and technological advances in data analytics, management, transport, and data user spheres.

Despite widespread agreement on the inherent opportunities and current limitations of Big Data, a lack of consensus on some important fundamental questions continues to confuse potential users and stymie progress. These questions include the following:

* How is Big Data defined?
* What attributes define Big Data solutions?
* What is the significance of possessing Big Data?
* How is Big Data different from traditional data environments and related applications?
* What are the essential characteristics of Big Data environments?
* How do these environments integrate with currently deployed architectures?
* What are the central scientific, technological, and standardization challenges that need to be addressed to accelerate the deployment of robust Big Data solutions?

Within this context, on March 29, 2012, the White House announced the Big Data Research and Development Initiative.[[2]](#endnote-1) The initiative’s goals include helping to accelerate the pace of discovery in science and engineering, strengthening national security, and transforming teaching and learning by improving the ability to extract knowledge and insights from large and complex collections of digital data.

Six federal departments and their agencies announced more than $200 million in commitments spread across more than 80 projects, which aim to significantly improve the tools and techniques needed to access, organize, and draw conclusions from huge volumes of digital data. The initiative also challenged industry, research universities, and nonprofits to join with the federal government to make the most of the opportunities created by Big Data.

Motivated by the White House initiative and public suggestions, the National Institute of Standards and Technology (NIST) has accepted the challenge to stimulate collaboration among industry professionals to further the secure and effective adoption of Big Data. As one result of NIST’s Cloud and Big Data Forum held on January 15–17, 2013, there was strong encouragement for NIST to create a public working group for the development of a Big Data Standards Roadmap. Forum participants noted that this roadmap should define and prioritize Big Data requirements, including interoperability, portability, reusability, extensibility, data usage, analytics, and technology infrastructure. In doing so, the roadmap would accelerate the adoption of the most secure and effective Big Data techniques and technology.

On June 19, 2013, the NIST Big Data Public Working Group (NBD-PWG) was launched with extensive participation by industry, academia, and government from across the nation. The scope of the NBD-PWG involves forming a community of interests from all sectors—including industry, academia, and government—with the goal of developing consensus on definitions, taxonomies, secure reference architectures, security and privacy, and—from these—a standards roadmap. Such a consensus would create a vendor-neutral, technology- and infrastructure-independent framework that would enable Big Data stakeholders to identify and use the best analytics tools for their processing and visualization requirements on the most suitable computing platform and cluster, while also allowing value-added from Big Data service providers.

The *NIST Big Data Interoperability Framework* will be released in three versions, which correspond to the three stages of the NBD-PWG work. The three stages aim to achieve the following with respect to the NIST Big Data Reference Architecture (NBDRA.)

1. Identify the high-level Big Data reference architecture key components, which are technology, infrastructure, and vendor agnostic.
2. Define general interfaces between the NBDRA components.
3. Validate the NBDRA by building Big Data general applications through the general interfaces.

On September 16, 2015, seven volumes *NIST Big Data Interoperability Framework* V1.0 documents were published (http://bigdatawg.nist.gov/V1\_output\_docs.php), each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The seven volumes are as follows:

* Volume 1, Definitions
* Volume 2, Taxonomies
* Volume 3, Use Cases and General Requirements
* Volume 4, Security and Privacy
* Volume 5, Architectures White Paper Survey
* Volume 6, Reference Architecture
* Volume 7, Standards Roadmap

Currently the NBD-PWG is working on Stage 2 with the goals to enhance the version 1 content, define general interfaces between the NBDRA components by aggregating low-level interactions into high-level general interfaces, and demonstrate how the NBDRA can be used. As a result, the following two additional volumes have been identified.

* Volume 8, Reference Architecture Interfaces
* Volume 9, Adoption and Modernization

Potential areas of future work for each volume during Stage 3 are highlighted in Section 1.5 of each volume. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

## NIST Big Data Public Working Group

This section will be written at a later date.

## Scope and Objectives of the Technology Roadmap Subgroup

This section will be written at a later date.

## Report Production

This section will be written at a later date.

To achieve technical and high quality document content, this document will go through public comments period along with NIST internal review.

## Future Work on this Volume

This section will be written at a later date.

# Landscape Perspective

Section Scope: To understand particular aspects of industries, technologies and economic impacts of big data by viewing them in the contexts of the broader landscape.

Organizations face many challenges in the course of validating their existing integrations and observing the potential operational implications of the rapidly changing environment. Effectiveness is dependent on a clear understanding of new technologies. One way of looking at it is to examine where the money has been spent.

The following spend figures have not been researched in depth enough to be verified and adjusted for accuracy. Other categories that could be included in Table 1 include life sciences, insurance, and professional services.

Table 1: Spending by Industry

|  |  |  |  |
| --- | --- | --- | --- |
| Industry | Spend | Certainty of Spend assumption | Adoption Rate [Economist] |
| Telecommunications and media | $1.2b | Medium | Highest, 62% |
| Telecommunications and IT | $2b |  |  |
| Banking Financial services | $6.4b | Medium | 38% |
| Government and defense | $3b | High | 45% |
| IT, software, internet | $3b | Medium [for software]6 | 57% |
| Natural resources, Energy, and utilities | $1b | Medium | 45% |
| Healthcare | $1b | Low | Slowest, 21% |
| Retail | $.8b | Low | Highest, 68% |
| Transportation, logistics | $.7b | Low |  |
| Biotechnology |  |  | Slowest, 21% |
| Pharmaceuticals |  |  | Slowest, 21% |
| Construction and real estate |  |  | 52% |
| Education |  | Low | 53% |
| Manufacturing and automotive |  | Low | 40% |

## Big Data Trends and Forecasts

Subsection Scope: This subsection will discuss the direction of industries with respect to Big Data adoption, forecasts for Big Data systems, and potential future developments in Big Data.

# 

# Adoption and Barriers

## Exploring Big Data Adoption

### Adoption by Industry

Subsection Scope: Which industries have adopted Big Data? How have Big Data systems been received in these industries? What percentage of the industry has adopted Big Data? Are there any patterns noticed across industries? What are the factors that lead to successful adoptions of Big Data systems?

In 2012, Gartner reported that the potential opportunities available from the adoption of big data were higher for the Government sector than for any other sector or industry.

The telecommunications industry has been one of the most prominent adopters of big data projects. Reports from US Bureau of Economic Analysis (citation needed) and McKinsey Global Institute (citation needed) suggest that the utility sector in general has an advantage when it comes to capturing big data, indicating that the ability to access or own the data results has a direct reflection on adoption of big data systems.

While other sectors and industries such as banking and finance, insurance, healthcare providers, and manufacturing also enjoy ability to capture and access big data; many that are expected to glean the highest value or competitive advantage from adoption of big data projects, such as government, education services, and administration, support and waste management, do not have ease of capture. Difficulty in capturing, accessing and owning the data itself must therefore be considered a barrier to adoption.

Larger and mid-size organizations are considered to be the earlier adopters. In 2015, IDG reported that top priorities were integration into existing infrastructure [48%], security [38%], ease of use [35 to 51% depending on firm size], and support and services [37%]. In 2012, Gartner reported the potential opportunities were highest for Government; followed closely by Communications, Media and Services; Manufacturing and Natural Resources; and Banking and Securities, respectively. Healthcare Providers; Retail; and Insurance were to expect moderate potential, while Transportation; Utilities; Education; and Wholesale Trade industries had lower potential. More often than not, Hadoop is cited as overkill; and big data projects in general as too costly in comparison to expected ROI.

This Big Data Adoption in 2016 draft prepared for NIST does not adhere to or necessarily agree with the categorization schemes or findings of any reports cited.

### Functional Perspective of Adoption

Subsection Scope: What Big Data technologies have been adopted? What types of companies/industries are adopting these technologies?

Despite the obvious need for improved search technologies, very few organizations have implemented ‘real’ search systems within their stack. AIIM polled 353 members of its global community and found that over 70% considered search to be essential or vital to operations, and equivalent in importance to both big data projects and TAR, yet the majority do not have a mature search function and only 18% have federated capability [3.2 ref]. Very little adoption of open source technologies [~15% on average [small, medium, large company]; and reduced spend forecast for DIY built OS search apps.

## Technical and non-technical Barriers to adoption

Section Scope: Discussion of the types of barriers to Big Data adoption. What are the challenges that organizations face when implementing a Big Data system?

Table 2: Non-Technical and Technical Barriers to Adoption

| Non-Technical Barriers | Technical Barriers |
| --- | --- |
| Lack of stakeholder definition and product agreement | Lack of practitioners with the ability to handle for complexity of software |
| Budget / expensive licenses | Integration with existing infrastructure |
| Lack of established processes to go from proof-of-concept to production systems | Security of systems |
| Compliance with privacy and regulations | Cloud: concerns over liabilities, security, and performance |
| Inconsistent metadata standards | Cloud: connectivity bandwidth is a most significant constraint |
| Some silos of data and access restriction | Cloud: Mesh, cell, and internet network components |
| WRT cloud, shift from centralized stewardship toward decentralized and granular model |  |
| Legacy access methods present tremendous integration and compliance challenges |  |
| Proprietary, patented access methods have been a barrier to construction of connectors |  |
| Organizational maturity |  |

### Non-Technical Barriers

Subsection Scope: What are the non-technical challenges that organizations face when adopting Big Data systems?

Frequently cited non-technical barriers include lack of stakeholder definition and product agreement; budget, expensive licenses, small ROI in comparison to Big Data project costs, and unclear ROI. (citations needed) Establishing processes to go from proof-of-concept to production systems; and compliance with privacy and regulations are also major concerns.

Adoption of Access technologies: For legal and security reasons, some silos of data and access restriction is necessary. Metadata standards within individual organizations are often inconsistent.

Table 3: Non-Technical Barriers to Adoption

| Non-Technical Barriers | Aggregate Surveys | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Category   * Sub-category | CDW | Accenture | Knowledgent | Hitachi | TDWI | Information Week |
| Developing an overall management program |  |  |  |  |  |  |
| Budget; expensive licenses | 32% | 47% | 47% |  |  | 34% |
| Stakeholder definition and product agreement |  |  | 45% |  |  | 40% |
| Establishing processes to go from POC to production |  |  | 43% |  |  |  |
| Compliance, privacy and regulatory concerns |  |  | 42% |  | 29% |  |
| * S&P challenge in regulation understanding or compliance |  |  |  |  |  |  |
| * Governance: monitoring; doc operating model |  |  |  |  |  |  |
| * Governance: ownership |  |  |  |  |  |  |
| * Governance: adapting rules for quickly changing end users |  |  |  |  |  |  |
| Difficulty operationalizing insights |  |  | 33% | 31% |  |  |
| Lack of access to sources |  |  |  |  |  |  |
| Silos |  |  |  |  |  |  |
| * Lack of willingness to share; departmental communication. |  |  |  | 36% |  |  |
| Healthcare Info Tech (HIT) |  |  |  |  |  |  |
| * Defining the data that needs to be collected | 35% |  |  |  |  |  |
| * Resistance to change | 30% |  |  |  |  |  |
| * Lack of industry standards | 21% |  |  |  |  |  |
| Lack of buy-in from management |  |  |  | 18% | 29% |  |
| Lack of compelling use case |  |  |  |  | 31% |  |
| No clear ROI |  |  |  |  |  | 36% |

Additional Table: Incorporate figure 4.4 from analytics + rdm.pdf.

### Technical barriers to adoption

Subsection Scope: What are the technical challenges that organizations face when adopting Big Data systems? Discussion of barriers posed by technologies used for Big Data, use of the technologies, interconnectivity of the technologies, etc.

The lack of practitioners with the ability to handle complexity of software, and integration with existing infrastructure are frequently cited as the most significant difficulties. Security of Hadoop systems?

Connectivity routes are especially important for interface interoperability of patient health information. Existing standards such as CCD and CCR for clinical document exchange provide a simple query and retrieve model for integration where care professionals can selectively transmit data. … not result in a horizontally interoperable system for holistic viewing platforms that can connect the query activities of independent professionals over time and over disparate systems regardless of the underlying infrastructure or operating system for maintaining the data [FHIR subscription web services approach]]. This area will benefit from additional standards work.

Barriers to the use of cloud for analytics: concerns over liabilities, security, and performance. Physically, connectivity bandwidth is a most significant constraint. Mesh, cell, and internet network components.

Barriers for search include difficulties managing cloud sharing, mobile tech, and notetaking technologies [evernote]. The cloud creates additional challenges for governance. Shift from centralized stewardship, toward decentralized and granular model where user roles have structure for individual access rules. This runs parallel to or congruent to market demand for self-service analytics application capabilities.

Table 4: Technical Barriers to Adoption

| Technical Barriers | Aggregate Surveys | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Category   * Subcategory | CDW | Accenture | Knowledgent | Hitachi | TDWI | Information Week |
| Lack of practitioners for complexity of software | 27% | 40% | 40% | 40% | 42% | 46% |
| Performance during concurrent usage |  |  |  |  |  |  |
| Integration with existing infrastructure |  | 35% | 35% |  |  |  |
| * Moving data from source to analytics environment NRT |  |  |  |  |  |  |
| * Blending internal & external data; merging sources | 45% |  |  |  |  |  |
| * Organization-wide view of data movement between apps |  |  |  |  |  |  |
| * Moving data between on-premise systems and clouds |  |  |  |  |  |  |
| * Hadoop data |  |  |  |  |  |  |
| Hadoop specific |  |  |  |  |  |  |
| * Backup and recovery |  |  |  |  |  |  |
| * Availability |  |  |  |  |  |  |
| * Performance at scale |  |  |  |  |  |  |
| * Lack of user friendly tools |  |  |  |  | 27% |  |
| * Security |  | 50% |  |  | 29% |  |
| Compliance, privacy and regulatory concerns |  |  | 42% |  |  |  |
| * S&P securing deployments from hack |  |  |  |  |  |  |
| * S&P inability to mask, de-identify sensitive data |  |  |  |  |  |  |
| * S&P lack of fine control to support hetero user population |  |  |  |  |  |  |
| * Governance: auditing access; logging / tracking data lineage |  |  |  |  |  |  |
| Analytics layer Technical ‘mis-specifications’ |  |  |  |  |  |  |
| Lack of suitable software |  |  |  | 42% |  |  |
| Lack of metadata mgmt. |  |  | 25% |  | 28% |  |
| Providing end users with self service capability |  |  | 33% |  |  |  |
| Providing business level context for understanding |  |  | 33% |  |  |  |

# 

# Maturity

Section Scope: This section will look at maturity within a framework of three (3) stages: 1) R&D, 2) demo and deployment, and 3) commercialization. This subsection will discuss the maturity of technology, both hardware and software, related to Big Data systems. The maturity of an organization affects the likelihood of a successful Big Data implementation. Organization maturity with respect to Big Data implementations will be discussed.

## functional and Technologic Maturity

IaaS is driving opportunities for technology infrastructure. Maturity: As costs associated with both open source and commercial computing technologies fall drastically, it becomes easier for organizations to implement big data projects, increasing the overall knowledge levels and adding to a tide effect where all boats in the marina are raised toward maturity. Applications receive a great deal of attention in articles written for business audiences, however on the whole, the challenges in applications are proving less difficult to solve than challenges in infrastructure.

Maturity of in-memory technologies: It is not always simple to distinguish between in-memory DBMS, in-memory analytics, and in-memory data grids. It is safe to say however that all in-memory technologies will provide a high benefit to organizations that have valid business use cases for adopting these technologies. In terms of maturity, in-memory technologies have essentially reached mainstream adoption.

Maturity of Access technologies and information retrieval techniques. While access methods for traditional computing are in many cases brought forward into big data use cases, legacy access methods present tremendous integration and compliance challenges for organizations tackling big data. Solutions to the various challenges remain a work in progress. In some cases, proprietary, patented access methods have been a barrier to construction of connectors required for federated search and \_.

Maturity of internal search: “Only 12% have an agreed search strategy, and only half of those have a specific budget” [AIIM]. The top two challenges seem to be a lack of available staff with the skills to support the function, and the organizations ability to dedicate personnel to maintain the \_ servers. Departments are reluctant to take ownership of the function due to the problematic levels of the issues. The consensus amongst AIIM’s survey was that the Compliance, IG or Records management department should actually be the responsible for the search function.

Metadata standards are ranked a significant technical issue, as well as an inability for the organization to even agree on a local taxonomy snapshot, much less manage the difficulties of taxonomy dynamics [an organizational issue]. Interfaces round out the list. Less than 10% of organizations are following CMIS.

Another situation in the field is the multiplicity of search tools within an organization, each used by small groups. Some larger organizations having five or more search products in use simultaneously.

Maturity of stream processing: continued adoption of streaming data will benefit from technologies that provide the capability to cross reference [aka unify] streaming data with data at rest. Challenges in query planning. The age of big data applied a downward pressure on the use of standard indexes, reducing their use for data at rest. This trend is carried into adoption of unified architectures [Lambda], as unified architectures update indexes in batch intervals. An opportunity exists for open source technologies [Lucene and SOLR] which apply incremental indexing, to reduce updating costs and increase loading speeds for unified architectures.

Maturity of open source technologies: not as much maturity as media/ surface sources would indicate.

Maturity of open data: Some transformations are underway in the biology and cosmology domains, with new activity in climate science and materials science [NIH, USGEO]. Various agencies are considering mandating the management of curation and metadata activities in funded research projects.

## Organizational Maturity

Subsection Scope: What are the characteristics of a mature organization that will increase the likelihood of a successful Big Data implementation? Why is it important for an organization to be mature to have a successful Big Data implementation?

Maturity. “An organization’s ability to drive transformation with big data is directly correlated with its organizational maturity.” [IDC]

## Big Data Trends and Forecasts

Subsection Scope: This subsection will discuss the direction of industries with respect to Big Data adoption, forecasts for Big Data systems, and potential future developments in Big Data.

Transition [sentence] from maturity section. In-memory, private cloud infrastructure, complex event processing, have reached the mainstream. Modern data science and machine learning are on a very fast pace to maturity.

Table 5: Maturity Projections

|  |  |
| --- | --- |
| 2017 – 2020 | 2020 - 2025 |
| High performance message infrastructure | Internet of things |
| Search based analysis | Semantic web |
|  | Text and entity analysis |
| PMML |  |

Governance: information management roles; and stewardship applications.

Stewardship: Within any single organization, data stewardship may take on one of a handful of particular models.

In a data stewardship model that is function oriented or organization oriented, the components of the stewardship are often framed in terms of the lines of business or departments that use the data. These departments might be Customer Service, Finance, Marketing, or Sales, for example. All of these organization functions may be thought of as components of a larger enterprise process applications layer, supported by an organization wide standards layer.

Integration: increased usage of lightweight iPaaS platforms; and APIs for enabling microservices, mashup data from multiple sources. Scarcity of general use interfaces that are capable of supporting diverse data management requirements / capability to work with container frameworks, data APIs, metadata standards, etc. Demand for interfaces with flexibility to handle heterogeneous user types, each having unique conceptual needs.

Logically: an increase in application of semantic technologies for data enrichment. Several applications of text analysis technology including fast moving consumer goods, fraud detection, and healthcare. An area where cloud deployment is working.

Search over emails is a very large necessity for organizations. Considering the lack of products with this capability in the market, this area is ripe for standards to support product development.

# Considerations for Implementation and Modernization

Section Scope: This section will discuss very high level pathways for Big Data system implementation and system modernization.

## Implementation

Subsection Scope: What are the typical stages that an organizations goes through to implement a Big Data system? The stages will begin with considerations and decisions that need to be addressed when preparing to move forward with the Big Data system implementation. The stages will end with the considerations post implementation of the technology (e.g,. personnel training, maintenance, governance).

Implementing a Big Data system, here are the first couple of things to do or consider.

Where to start. A big data system implementation starts with a pilot project. In this initial stage, \_.

In the second stage, the prototype system is used primarily [only] by those in the IT department. Often limited to storage and data transformation, possibly some exploratory activity.

In the third stage, the project is expanding beyond storage and integration functions and providing a function for one or two lines of business, perhaps performing unstructured data or predictive analysis.

In the fourth stage, the project has achieved integration with organizations’ governance protocols, metadata standards and data quality management.

In the fifth or final stage, a big data project is providing a full range of services including business user abstractions, and collaboration and data sharing capabilities.

Data integration and preparation are often reported as the greatest challenges to successful big data projects. Nontechnical issues such as change management, solution approach or problem definition and framing are more likely to be greater challenges.

Consider modernizing part of a legacy system, as opposed to modernizing the entire system.

## System Modernization

Subsection Scope: System modernization involves the updating of a current data system to one that is prepared to handle Big Data requirements. This section will discuss the pathways that an organization can take to modernize existing systems.

An organization preparing to develop a big data system will typically find themselves considering one of two possible directions for modernization. For the time being we will refer to these two directions as

Augmentation and Replacement. Each of the modernization directions have unique advantages and disadvantages.

* Updating / augmenting the supporting architecture. Updating the supporting architecture has noticeable advantages of including more mature technologies amidst the stack; and also a flexibility in the timeline for implementation. Augmentation allows for a phased implementation that can be stretched out over more than one fiscal budget...
* Updating / replacing an existing system by implementing an entirely new system. Modernizing an existing system by replacing the whole architecture has notable disadvantages. In comparison to the augmentation approach, the level of change management required when replacing entire systems is significantly higher.

Table 6: Possible Directions for Modernization

|  |  |
| --- | --- |
| Advantages and Disadvantages of the Two Directions for Modernization | |
| Augmentation | |
| Advantages   * Phased approach * Not entirely immature stack of technology | **Disadvantages**   * X |
| Replacement | |
| Advantages   * X | **Disadvantages**   * Longer development cycle * Increased change management * Less mature technologies |

Additional Text: Discuss the implementation period middle road (e.g., two years) before system replacement where retraining, testing, etc. needs to be performed.



Figure 1: New System Implementation

Situation commonly referred to as ‘build or buy’ [or outsource]. In the ‘build’ or do-it-yourself [DIY] scenario the organization may modify their existing system or build an entirely new system separate of the existing system. One of the largest barriers organizations face when building their own systems is the scarcity of engineers with the skillset covering newer technologies such as streaming / near real time analysis.

If the DIY implementation is erected concurrent to the existing system, the organization is required to operate two systems for the length of time it will take to migrate / combine…

The alternative to the DIY scenario is for the organization to buy or rent a new big data system. Renting = cloud. Advantages include ease of scale; and not having to operate a congruent systems [or possibly not having to modify an existing system].

Hybrid parallel system [not 100% integrated with existing]. Organizations can use the cloud for storage but develop their own applications. Costly to move data to the cloud. Standards for hybrid implementations should accelerate the adoption and interoperability of… 2

Whether the organization goes a DIY route and builds their own system, or buys / rents a system, or a mix of internal development, open source tools and third party hosting, still face certain challenges. Data cleansing, plumbing, etc.

Figure 2: Requirement Considerations

## Recommendations for Organizational Modernization

Subsection Scope: This section will discuss the pathways that an organization can take to modernize the organization itself, to facilitate the successful transition from existing systems to a more modern system.

In search related projects, set a balance for governance and retrieval; determine ownership aka departmental responsibility for the function; aim for unified or single point search capability; consider outsourcing if necessary.

1. Acronyms

The acronym list will be revised once Volume 9 is closer to completion.

CCD Continuity of Care Document

CCR Continuity of Care Record

DIY do-it yourself

FHIR Fast Healthcare Interoperability Resources

HIT Healthcare Info Tech

ITL Information Technology Laboratory (in NIST)

NBDIF NIST Big Data Interoperability Framework Volume

NBD-PWG NIST Big Data Public Working Group

NBDRA NIST Big Data Reference Architecture

NIST National Institute of Standards and Technology

ROI return on investment

1. References

Economist / Hitachi report: <https://www.hds.com/assets/pdf/the-hype-and-the-hope-summary.pdf>

IDG survey: <http://www.idgenterprise.com/resource/research/2015-big-data-and-analytics-survey/>

Informationweek: Survey of 665 business technology professionals. TDWI\_LV13\_trip%20report\_F.pdf

3.2. Search and Discovery; exploiting knowledge, minimizing risk. AIIM. Search and Discovery Industry Watch. 2014.

4.2 IDC InfoBrief. Retrieved from Knowledge Hub Media, Using big data and analytics to drive business transformation. <http://knowledgehubmedia.com/using-big-data-analytics-to-drive-business-transformation-2/> [/ Raconteur]

Knowledgent: 2015 Survey: Current implementation challenges. <https://knowledgent.com/whitepaper/2015-big-data-survey-current-implementation-challenges/>

TDWI: Hadoop for the Enterprise: Making data management massively scalable, agile, feature rich and cost effective. <http://www.idgconnect.com/view_abstract/34448/hadoop-enterprise-making-data-management-massively-scalable-agile-feature-rich-cost-effective?source=connect>

6. DataRPM in Big Data Analytics News: <http://bigdataanalyticsnews.com/big-data-trends-2015-infographic/>

World Economic Forum Insight Report: The Global IT Report 2014, Rewards and Risks of Big Data. Econ\_IT\_Report.pdf

1. “Contributors” are members of the NIST Big Data Public Working Group who dedicated great effort to prepare and substantial time on a regular basis to research and development in support of version 2 of this document. [↑](#footnote-ref-1)
2. The White House Office of Science and Technology Policy, “Big Data is a Big Deal,” *OSTP Blog*, accessed February 21, 2014, <http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal>. [↑](#endnote-ref-1)