**NIST Special Publication 1500-10**

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

NIST Big Data Public Working Group

Standards Roadmap Subgroup

Version 2

January 29, 2018

<http://dx.doi.org/10.6028/NIST.SP.1500-10>



NIST Special Publication 1500-10

Information Technology Laboratory

**NIST Big Data Interoperability Framework:**

**Volume 9, Adoption and Modernization**

**Version 2**

NIST Big Data Public Working Group (NBD-PWG)

Standards Roadmap Subgroup

National Institute of Standards and Technology

Gaithersburg, MD 20899

This draft publication is available free of charge from:

<http://dx.doi.org/10.6028/NIST.SP.1500-10>

January 2018



U. S. Department of Commerce

*Wilbur L. Ross, Jr., Secretary*

National Institute of Standards and Technology

*Dr. Walter Copan, Under Secretary of Commerce for Standards and Technology*

*and NIST Director*

**National Institute of Standards and Technology (NIST) Special Publication 1500-10**

54 pages (January 29, 2018)

Certain commercial entities, equipment, or materials may be identified in this document 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 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)

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

The potential for organizations to capture value from Big Data improves every day as the pace of the Big Data revolution continues to increase, but the level of value captured by companies deploying Big Data initiatives has not been equivalent across all industries. Most companies are struggling to capture a small fraction of the available potential in Big Data initiatives. The healthcare and manufacturing industries, for example, have so far been less successful at taking advantage of data and analytics than other industries such as logistics and retail. Effective capture of value will likely require organizational investment in change management strategies that support transformation of the culture, and redesign of legacy processes.

In some cases, the less-than-satisfying impacts of Big Data projects are not for lack of significant financial investments in new technology. It is common to find reports pointing to a shortage of technical talent as one of the largest barriers to undertaking projects, and this issue is expected to persist into the future.

This volume explores the adoption of Big Data systems and barriers to adoption; factors in maturity of Big Data projects, organizations implementing those projects, and the Big Data technology market; and considerations for implementation and modernization of Big Data systems.

Keywords

Big Data; adoption; barriers; market maturity; project maturity; organizational maturity; implementation; system modernization.

Acknowledgements

This document reflects the contributions and discussions by the membership of the NBD-PWG, co-chaired by Wo Chang (NIST ITL), Bob Marcus (ET-Strategies), and Chaitan Baru (San Diego Supercomputer Center; National Science Foundation). For all versions, the Subgroups were led by the following people: Nancy Grady (SAIC), Natasha Balac (SDSC), and Eugene Luster (R2AD) for the Definitions and Taxonomies Subgroup; Geoffrey Fox (Indiana University) and Tsegereda Beyene (Cisco Systems) for the Use Cases and Requirements Subgroup; Arnab Roy (Fujitsu), Mark Underwood (Krypton Brothers; Synchrony Financial), and Akhil Manchanda (GE) for the Security and Privacy Subgroup; David Boyd (InCadence Strategic Solutions), Orit Levin (Microsoft), Don Krapohl (Augmented Intelligence), and James Ketner (AT&T) for the Reference Architecture Subgroup; and Russell Reinsch (Center for Government Interoperability), David Boyd (InCadence Strategic Solutions), Carl Buffington (Vistronix), and Dan McClary (Oracle), for the Standards Roadmap Subgroup.

The editors for this document were the following:

* ***Version 1***: This volume resulted from Stage 2 work and was not part of the Version 1 scope.
* ***Version 2***: Russell Reinsch (Center for Government Interoperability) and Wo Chang (NIST)

Laurie Aldape (Energetics Incorporated) and Elizabeth Lennon (NIST) provided editorial assistance across all NBDIF volumes.

NIST SP1500-10, Version 2 has been collaboratively authored by the NBD-PWG. As of the date of this publication, there are over six hundred 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 would like to acknowledge the specific contributions[[1]](#footnote-2) to this volume, during Version 1 and/or Version 2 activities, by the following NBD-PWG members:

|  |  |  |
| --- | --- | --- |
| **David Boyd**  *InCadence Strategic Solutions*  **Frank Farance**  *Consultant*  **Geoffrey Fox**  *Indiana University*  **Nancy Grady**  *SAIC* | **Zane Harvey**  *QuantumS3*  **Haiping Luo**  *Department of the Treasury*  **Russell Reinsch**  *Center for Government Interoperability*  **Arnab Roy**  *Fujitsu* | **Mark Underwood**  *Krypton Brothers Synchrony Financial*  **Gregor von Lasewski**  *Indiana University*  **Timothy Zimmerlin**  *Consultant* |

Regenerate the table of contents

Table of Contents

[Executive Summary vii](#_Toc516844628)

[1 Introduction 1](#_Toc516844629)

[1.1 Background 1](#_Toc516844630)

[1.2 Scope and Objectives of the Standards Roadmap Subgroup 2](#_Toc516844631)

[1.3 Report Production 3](#_Toc516844632)

[1.4 Report Structure 3](#_Toc516844633)

[1.5 this section will be deleted before the doc is released for public comment 4](#_Toc516844634)

[2 Adoption and Barriers 5](#_Toc516844635)

[2.1 Exploring Big Data Adoption 5](#_Toc516844636)

[2.1.1 Adoption by Use Case 5](#_Toc516844637)

[2.1.2 Adoption by Industry 6](#_Toc516844638)

[2.1.3 Levels of Spending 6](#_Toc516844639)

[2.2 Barriers to adoption: Nontechnical and Technical 7](#_Toc516844640)

[2.2.1 Nontechnical Barriers 8](#_Toc516844641)

[2.2.2 Technical Barriers to Adoption 10](#_Toc516844642)

[3 Maturity 13](#_Toc516844643)

[3.1 project maturity 14](#_Toc516844644)

[3.1.1 Level 1: Ad hoc 14](#_Toc516844645)

[3.1.2 Level 2: Department Adoption 14](#_Toc516844646)

[3.1.3 Level 3 Enterprise Adoption 14](#_Toc516844647)

[3.1.4 Level 4: Culture of Governance 15](#_Toc516844648)

[3.2 Organizational Maturity 16](#_Toc516844649)

[3.2.1 Evolution of organizational maturity 16](#_Toc516844650)

[3.3 market maturity of technologies 18](#_Toc516844651)

[3.4 Big Data Trends and Forecasts 19](#_Toc516844652)

[4 Implementation and Modernization 20](#_Toc516844653)

[4.1 System Modernization 20](#_Toc516844654)

[4.2 Implementation 22](#_Toc516844655)

[5 Specific solution Techniques, Dependent on the Problem Space 24](#_Toc516844656)

[6 Big Data readiness 29](#_Toc516844657)

[6.1 Tactical actions 29](#_Toc516844658)

[6.2 Fundamental Changes needed to enable big data 30](#_Toc516844659)

[6.3 Big Data readiness 31](#_Toc516844660)

[6.4 Disrupting the status quo 34](#_Toc516844661)

[6.5 Harnessing Big Data 35](#_Toc516844662)

[6.6 Checklists 37](#_Toc516844663)

[Appendix A: Big Data readiness checklists A-1](#_Toc516844664)

[Appendix B: Acronyms B-1](#_Toc516844665)

[Appendix C: Bibliography C-1](#_Toc516844666)

[Appendix D: Related Reading D-1](#_Toc516844667)

List of Figures

Regenerate this list

[**Figure 1. Evolution of Big Data systems as a function of project and organizational data governance maturity.** 17](#_Toc516844668)

[**Figure 2. New system implementation** 21](#_Toc516844669)

[**Figure 3. Requirement decision tree** 24](#_Toc516844670)

[**Figure 4. Machine learning algorithm application workflow** 25](file:///C:\Users\Claire\Dropbox%20(AQ)\Big%20data%20analytics\SP1500-june15%20(4.45%20pm).1150-Vol9-hp_CCA_RR.docx#_Toc516844671)

[**Figure 5. Supervised Machine Learning Algorithms** 26](file:///C:\Users\Claire\Dropbox%20(AQ)\Big%20data%20analytics\SP1500-june15%20(4.45%20pm).1150-Vol9-hp_CCA_RR.docx#_Toc516844672)

[**Figure 7. Dataset fragmentation.** 32](#_Toc516844674)

[**Figure 8. Data fragmentation** 33](#_Toc516844675)

[**Figure 9. Time spent making data analysis ready** 34](#_Toc516844676)

[**Figure 10. Improved data governance.** 36](#_Toc516844677)

[**Figure 11. Linear data flows for authoritative data** 37](#_Toc516844678)

List of Tables

Regenerate this list

[**Table 1. Approximate Adoption by Use Case and Industry** 5](#_Toc516844679)

[**Table 4. Nontechnical Barriers to Adoption** 9](#_Toc516844680)

[**Table 5. Technical Barriers to Adoption** 10](#_Toc516844681)

[**Table 6. Summary of Barriers to Big Data** 12](#_Toc516844682)

[**Table 7. Maturity projections** 19](#_Toc516844683)

[**Table 8. Advantages and Disadvantages of System Modernization via the Augmentation Pathway** 21](#_Toc516844684)

[**Table 9. Advantages and Disadvantages of System Modernization via the Replacement Pathway** 22](#_Toc516844685)

[**Table 10. Supervised Learning Regression Algorithms** 27](#_Toc516844686)

[**Table 11. Supervised Learning Classification Algorithms** 27](#_Toc516844687)

[**Table 12. Unsupervised Clustering Algorithms** 28](#_Toc516844688)

[**Table 13. Dimensionality Reduction Techniques** 29](#_Toc516844689)

Executive Summary

The NIST Big Data Public Working Group (NBD-PWG) Standards Roadmap Subgroup prepared this *NIST Big Data Interoperability Framework (NBDIF): Volume 9, Adoption and Modernization* to address nontechnical and technical barriers to Big Data adoption; explore project, organization, and technology maturity; consider future technology trends; and examine implementation and modernization strategies.

The NBDIF consists of nine volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The nine NBDIF volumes, which can be downloaded from <https://bigdatawg.nist.gov/V2_output_docs.php>, are as follows:

* Volume 1, Definitions [1]
* Volume 2, Taxonomies [2]
* Volume 3, Use Cases and General Requirements [3]
* Volume 4, Security and Privacy [4]
* Volume 5, Architectures White Paper Survey [5]
* Volume 6, Reference Architecture [6]
* Volume 7, Standards Roadmap [7]
* Volume 8, Reference Architecture Interfaces [8]
* Volume 9, Adoption and Modernization (this document)

The *NBDIF* is being released in three versions, which correspond to the three development 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; and
3. Validate the NBDRA by building Big Data general applications through the general interfaces.

Potential areas of future work for the Standards Roadmap Subgroup during Stage 3 are highlighted in Section 1.5 of this volume. The current effort (Stage 2) documented in this Volume 9 reflects concepts developed within the rapidly evolving field of Big Data.

# 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 cybersecurity threats be reversed?

Big Data by definition overwhelms traditional approaches to storage, computing, and retrieval of data. 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 new in Big Data?
* What is the difference between Big Data and *bigger data* that has been collected for years?
* 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, secure Big Data solutions?

Within this context, on March 29, 2012, the White House announced the Big Data Research and Development Initiative. [9] 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 analysts’ 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) 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 added value from Big Data service providers.

The *NIST Big Data Interoperability Framework* (NBDIF) is being 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; and
3. Validate the NBDRA by building Big Data general applications through the general interfaces.

On September 16, 2015, seven NBDIFVersion 1 volumes 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 [1]
* Volume 2, Taxonomies [2]
* Volume 3, Use Cases and General Requirements [3]
* Volume 4, Security and Privacy [4]
* Volume 5, Architectures White Paper Survey [5]
* Volume 6, Reference Architecture [6]
* Volume 7, Standards Roadmap [7]

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 of the Stage 2 work, the following two additional NBDIF volumes have been developed.

* Volume 8, Reference Architecture Interfaces [8]
* Volume 9, Adoption and Modernization [this document]

Version 2 of the NBDIF volumes, resulting from Stage 2 work, can be downloaded from the NBD-PWG website (<https://bigdatawg.nist.gov/V2_output_docs.php>). 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.

## Scope and Objectives of the Standards Roadmap Subgroup

The NBD-PWG Standards Roadmap Subgroup focused on forming a community of interest from industry, academia, and government, with the goal of developing a standards roadmap. The Subgroup’s approach included the following:

* Collaborate with the other four NBD-PWG subgroups;
* Review products of the other four subgroups including taxonomies, use cases, general requirements, and reference architecture;
* Gain an understanding of what standards are available or under development that may apply to Big Data;
* Perform a standards gap analysis and document the findings;
* Document vision and recommendations for future standards activities;
* Identify possible barriers that may delay or prevent adoption of Big Data; and
* Identify a few areas in which new standards could have a significant impact.

The goals of the Subgroup will be realized throughout the three planned phases of the NBD-PWG work, as outlined in Section 1.1.

Within the multitude of standards applicable to data and information technology (IT), the Subgroup focused on standards that: (1) apply to situations encountered in Big Data; (2) facilitate interfaces between NBDRA components (difference between Implementer [encoder] or User [decoder] may be nonexistent); (3) facilitate handling Big Data *characteristics*; and 4) represent a fundamental function.

## Report Production

The *NBDIF: Volume 9, Adoption and Modernization* is one of nine volumes, whose overall aims are to define and prioritize Big Data requirements, including interoperability, portability, reusability, extensibility, data usage, analytic techniques, and technology infrastructure to support secure and effective adoption of Big Data. The *NBDIF: Volume 9, Adoption and Modernization* arose from discussions during the weekly NBD-PWG conference calls. Topics included in this volume began to take form in Phase 2 of the NBD-PWG work, and this volume represents the groundwork for additional content planned for Phase 3.

During the discussions, the NBD-PWG identified the need to examine the landscape of Big Data implementations, challenges to implementing Big Data systems, technological and organizational maturity, and considerations surrounding implementations and system modernization. Consistent with the vendor-agnostic approach of the NBDIF, these topics were discussed without specifications for a particular technology or product to provide information applicable to a broad reader base. The Standards Roadmap Subgroup will continue to develop these and possibly other topics during Phase 3. The current version reflects the breadth of knowledge of the Subgroup members. The public’s participation in Phase 3 of the NBD-PWG work is encouraged.

To achieve high-quality technical content, this document has been reviewed and improved through a public comment period along with NIST internal review.

## Report Structure

Following the introductory material presented in Section 1, the remainder of this document is organized as follows:

* Section 2 examines the Big Data landscape at a high level.
* Section 3 explores the panorama of Big Data adoption thus far and the technical and nontechnical challenges faced by adopters of Big Data.
* Section 4 considers the influence of maturity (market, project, and organizational) to adoption of Big Data.
* Section 5 summarizes considerations when implementing Big Data systems or when modernizing existing systems to deal with Big Data.
* Appendices provide acronyms and bibliography for this document.

## this section will be deleted before the doc is released for public comment

A number of topics have not been discussed and clarified sufficiently to be included in Version 2. Topics that remain to be addressed in Version 3 of this document include the following:

* Technical challenges with data integration and preparation, specifically dealing with variables of different magnitudes; and
* Pathways for organizations to modernize to facilitate the successful transition from existing systems to more modern systems.

# Adoption and Barriers

## Exploring Big Data Adoption

This section views the adoption landscape from the perspectives of users and use cases, various industries, and levels of spending.

### Adoption by Use Case

Adoption of Big Data analysis technologies has been recently pegged at 53 percent [11]. Simple ways of looking at the big data environment are from the perspectives of use cases, both by organizational department, aka ‘function,’ and by industry; although each function and each industry adopting Big Data today have different levels of priorities. Overall, data warehouse optimization is reported as the top use case for big data projects, especially so for the healthcare industry, however the education and IT industries have placed higher priority on customer / social network analysis use cases (***Table 1***).

**Table 1. Approximate Adoption by Use Case and Industry**

|  |  |  |
| --- | --- | --- |
| Industry | Top Use Case | Random adoption metric. Priority? |
| Financial services | DW adoption | 83 |
| Healthcare | DW adoption | 80 |
| IT | Customer / social network analysis | 75 |
| Telecommunications | DW adoption | 74 |
| Education | Customer / social network analysis | 70 |

Departmentally, IT departments, business intelligence departments, and R&D are adopting big data for data warehouse optimization at the highest rate, but sales and marketing departments, finance departments, and executive management place higher priority on customer / social network analysis use cases. Different departments, and different sizes of organizations also have varying levels of interest in particular types of technologies. For example, executive management, and smaller organizations, have been found to show higher interest in service based products. The Dresner 2017 Big Data Study [11] cites financial services and telecommunications industries as the earliest adopters, with education lagging. In a 2016 report by Aman Naimat, [12] the numbers of personnel working on Big Data projects were used to determine Big Data adoption rates.

In this report, the IT, software and Internet, and banking and financial services industries appear to have been early Big Data adopters, while the oil and energy, and healthcare and pharmaceutical industries adopted Big Data at a slower rate. [12]

### Adoption by Industry

Adoption of Big Data systems has not been uniform across all industries or sectors. A 2014 report [Datameer] ranked financial services as the top industry in terms of big data usage, at 22%. Technology, telecommunications, and retail rounded out the top four. Government, fifth, and healthcare usage sixth, were each listed at 7%.

One condition effecting adoption is the fact that different industries inherently have different potential to capture value from the data. In this situation the higher difficulty of capturing value from the data equates to a barrier to adoption, and the reverse holds true as barriers, some of which are higher than others, impact the potential for the various industries to capture value from big data, for different reasons.

“[T]he public sector, including education, faces higher hurdles because of a lack of data-driven mind-set and available data. Capturing value in health care faces challenges given the relatively low investment performed so far.” [MGI]

While clear differences exist, there are however some common challenges that show up across all sectors that can delay the adoption of Big Data. A report by the U.S. Bureau of Economic Analysis and McKinsey Global Institute (MGI) suggests that the most obvious barrier to leveraging Big Data is access to the data itself. [15] The MGI report indicates a definite relationship between the ability to access data, and the potential to capture economic value, across all sectors / industries.

For example, the education industry is in the lowest percentile for availability of data, and consequently is also in the lowest 20% for producing economic value. The government sector, which is considered well positioned to benefit from Big Data, suffers from low access to data and may not fully realize the positive impacts of these technologies [15]. Table 2 lists industries that have the best access to data and rate highest on MGI’s value index.

**Table 2. Data Availability and Value Index from MGI Big Data Report**

|  |  |
| --- | --- |
| Data Availability | Value Index |
| Manufacturing, top 20 percentile | Manufacturing, top 20 percentile |
| Utilities, top 20% | Utilities, top 20% |
| Information, top 20% | Information, top 40% |
| Healthcare and social assistance, top 40% | Healthcare and social assistance, top 20% |
| Natural resources, top 40% | Natural resources, top 20% |

### Levels of Spending

Another way of looking at this environment is to view the landscape from the perspective of where money has been spent. Table 3 shows a sample breakdown of Big Data spending by industry across the Asia-Pacific region in 2016 [13], which as a region places big data slightly higher as a priority than Europe, Middle East and Africa; and North America.

**Table 3. Sample Spending by Industry**

|  |  |  |  |
| --- | --- | --- | --- |
| Industry | Sample Expenditure  (b = billion) | Certainty of Spend Assumption | Adoption Rate |
| Telecommunications and Media | US$1.2b | Medium | Highest, 62% |
| Telecommunications and IT | US$2b |  |  |
| Banking Financial Services | US$6.4b | Medium | 38% |
| Government and Defense | US$3b | High | 45% |
| IT, Software, Internet | US$3b | Medium (for software) [14] | 57% |
| Natural Resources, Energy, and Utilities | US$1b | Medium | 45% |
| Healthcare | US$1b | Low | Lowest, 21% |
| Retail | US$0.8b | Low | Highest, 68% |
| Transportation, Logistics | US$0.7b | Low |  |
| Biotechnology |  |  | Lowest, 21% |
| Pharmaceuticals |  |  | Lowest, 21% |
| Construction and Real Estate |  |  | 52% |
| Education |  | Low | 53% |
| Manufacturing and Automotive |  | Low | 40% |

## Barriers to adoption: Nontechnical and Technical

As organizations attempt to implement Big Data systems, they can be faced with a multitude of challenges. Generally, these challenges are of two types: nontechnical and technical. Nontechnical challenges involve issues surrounding the technical components of a Big Data system, but not considered hardware or software related. The nontechnical barriers could include issues related to workforce preparedness and availability, high cost, too many or too few regulations, or organizational culture. Technical challenges encompass issues resulting from the hardware or software and the interoperability between them. Technical barriers arise from factors which often include functional components of a Big Data system, integration with those functional components, or the security of those components.

Some barriers span both technical and non-technical. The adoption of Access technologies for example can involve nontechnical organizational departments, for legal and security reasons. Some silos of data and data access restriction policies are necessary, however poorly defined policies could result in inconsistent metadata standards within individual organizations, which can hinder interoperability.

Much like the market demand that is seen for self-service analytics application capabilities, is a shift from centralized stewardship toward a decentralized and granular model where user roles contain structures for individual access rules. This shift presents barriers for a search function, including difficulties managing cloud sharing, mobile tech, and notetaking technologies. Despite the obvious need for improved search technologies, very few organizations have implemented *full function* 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 technology-assisted review, yet the majority do not have a mature search function and only 18% have federated capability [16].

Placeholder to [elaborate on when federation matters].

As for Open Source search technologies, there has been very little adoption of these on average (approximately 15%) across small, medium, and large companies. Furthermore, forecasts indicate reduced spending on do-it-yourself (DIY)-built OS search apps.

### Nontechnical Barriers

Frequently cited nontechnical barriers are listed in ***Table 4*** and include lack of stakeholder definition and product agreement, budget, expensive licenses, small return on investment (ROI) in comparison to Big Data project costs, and unclear ROI. Workforce issues also affect the adoption of Big Data. The lack of practitioners with the ability to handle the complexities of software, and integration issues with existing infrastructure are frequently cited as the most significant difficulties. Other major concerns are establishing processes to progress from proof-of-concept to production systems and compliance with privacy and other regulations.

As previously noted, particular industries or organizations will likely face barriers that are specific to their situation. Barriers listed in Table 4 were considered serious enough to adversely impact a large number of potential Big Data adoptions. The number of survey respondents that cited a particular barrier are expressed as a percentage. Lower numbers are hidden; only higher numbers are shown in order to make them easier to locate.

New source: <https://sloanreview.mit.edu/article/implement-first-ask-questions-later-or-not-at-all/?social_token=fe68d2fc77840cfb52730af3e33dbafe&utm_source=linkedin&utm_medium=social&utm_campaign=sm-direct>

**Table 4. Nontechnical Barriers to Adoption**

| Nontechnical Barriers | Aggregate Surveys (% of respondents that identified the Big Data barrier) | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Category   * Sub-category | CDW | Accenture | Knowledgent | Hitachi | TDWI | Information Week |
| Difficulty developing an overall management program |  |  |  |  |  |  |
| Limited budgets; expensive licenses | 32% | 47% | 47% |  |  | 34% |
| Lack of stakeholder definition and product agreement |  |  | 45% |  |  | 40% |
| Difficulty 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% |
| Lack of practitioners for complexity of software | 27% | 40% | 40% | 40% | 42% | 46% |

### Technical Barriers to Adoption

Technical barriers include a broad range of issues involving the hardware and software for the Big Data systems. Technical barriers identified in ***Table 5*** are described along a functional orientation, intended to relate to the parts of Big Data systems as represented by the components and fabrics of the NBDRA. The *NBDIF: Volume 6, Reference Architecture* provides detailed discussion of the NBDRA and its functional components [here [link]].

**Table 5. Technical Barriers to Adoption**

| Technical Barriers | Aggregate Surveys (% of respondents that identified the Big Data barrier) | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Category   * Subcategory | CDW | Accenture | Knowledgent | Hitachi | TDWI | Information Week |
| Reduced performance during concurrent usage |  |  |  |  |  |  |
| Integration problems 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 misspecifications |  |  |  |  |  |  |
| Lack of suitable software |  |  |  | 42% |  |  |
| Lack of metadata management |  |  | 25% |  | 28% |  |
| Difficulty providing end users with self-service analytic capability |  |  | 33% |  |  |  |
| Complexity in providing business level context for understanding |  |  | 33% |  |  |  |

***Table 6*** Reorganizes some of the more significant nontechnical and technical barriers to adoption that were identified in sections 3.2.1, 3.2.2, and elsewhere.

**Table 6. Summary of Barriers to Big Data**

|  |  |  |
| --- | --- | --- |
| **AREA** | **NON-TECHNICAL BARRIERS** | **TECHNICAL BARRIERS** |
| **CULTURE** | • Data viewed simply as a means to an end. |  |
| • Lack of willingness to share. |  |
| • Resistance to change. |  |
| **DATA GOVERNANCE** | • Non-existant or inconsistent data governance. | • Merging data sources. |
| • Fragmented datasets | • Transferring data from source to analytics environment. |
| • Multiple "copies" of the same dataset that don't match. | • Blending internal and external data. |
| • Disparate data from different sources. | • Inconsistent metadata management. |
| • Data "silos." | • Inconsistent metadata standards. |
| • Lack of FAIRb, analysis ready data. |  |
| • Legacy access methods that present tremendous integration and compliance challenges. |  |
| • Proprietary, patented access methods a barrier to the construction of connectors. |  |
| • Inconsistent metadata standards. |  |
| **DATA ACCESS** | • Privacy regulations and confidentiality requirements. | • Concerns about liabilities and systems security. |
| • Data access restrictions. |  |
| **SKILL AND EXPERTISE** | • Lack of people with the ability to handle the complexity of software and analysis. |  |
| • Lack of people with ‘deep analytical’ trainingc. |  |
| • Lack of data savvy managersd. |  |
| • Lack of supporting technology personnel who develop, implement, and maintain the hardware and software tools such as databases and analytic programs needed to make use of Big Data. |  |
| **MANAGEMENT** | • Lack of buy-in from management. | • Integration with existing infrastructure. |
| • Lack of buy-in from data providers. |  |
| • Lack of organizational maturity. |  |
| • Shifting from centralized data stewardship toward decentralized and granular model. |  |
| • Difficulty operationalizing insights. |  |
| • Lack of process to go from proof-of-concept to production systems. |  |
| • Lack of definitions and product agreement. |  |
| • Lack of proof-of-concept examples. |  |
| **SOFTWARE AND COMPUTING SYSTEMS** | • Slow to switch from proprietary to open source software. | • Concerns about performance in the Cloud. |
|  | • Connectivity bandwidth in the Cloud is a most significant constraint. |
|  | • Cloud mesh, cell, and Internet network components. |
|  | • Legacy software and code. |
|  | • Lack of suitable software. |
|  | • Lack of suitable computing power. |
| **BUDGET** | • Lack of resources. |  |
| aAdapted from Manyika (2011). | |  |
| bData that are Findable, Accessible, Interoperable, and Reusable. | | |
| cPeople with advanced training in statistics and/or machine learning and who conduct data analysis. | | |
| dPeople with enough conceptual knowledge and quantitative skills to be able to frame and interpret analyses in an effective way (i.e., capable of posing the right questions for analysis, interpreting and challenging the results, and making appropriate decisions). | | |

To assist in viewing some of the other large barriers to adoption, it is helpful to organize them by their domains. Two important domains are healthcare and cloud computing.

Within the healthcare domain, connectivity routes are especially important for interface interoperability of patient health information. Existing standards, such as Continuity of Care Record (CCR) and Continuity of Care Document (CCD) for clinical document exchange, provide a simple query and retrieve model for integration where care professionals can selectively transmit data. These models do 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 (Fast Healthcare Interoperability Resources [FHIR] subscription web services approach). Additional standards work in this area could help alleviate the barrier.

In cloud implementations, cloud technologies have facilitated some aspects of Big Data adoption; however, challenges have arisen as the prevalence of cloud grows. Big Data challenges stemming from cloud usage include concerns over liabilities, security, and performance; the significant constraint of physical connectivity bandwidth; and interoperability of mesh, cell, and Internet network components.

The cloud increases the challenges for governance. As a project matures the challenges for managing governance concerns increase. (See Section 4.1, Project Maturity). Governance may become an even larger challenge than other regulatory and compliance concerns such as security and privacy. For example, privacy programs are frequently concerned with protection of private information, but often not with data in ERP applications; and security programs are frequently focused on protecting critical data and infrastructure, but not with data in analytics applications. While governance, security, and privacy programs have overlapping areas of concern, governance stakeholders frequently need to be concerned with a wider range of systems and related data.

# Maturity

Like most things, maturity can be viewed from multiple perspectives. For purposes in this document, the following three perspectives are used for shaping discourse on the concept: project maturity, organizational maturity, and market maturity. For purposes of this discussion, project maturity will describe the pathway that begins at the point where a team or small department is addressing a small need with a focused solution to implementation of a large, organization-wide Big Data system servicing a multitude of users and business needs. Characteristics of a particular maturity level may not be exclusive to a single level, and there may be some overlapping of characteristics, as the boundaries between stages of maturity are actually fuzzy.

Organizational maturity will describe some general changes across the organization, such as workflows, culture within the organization, worker training, executive support, and other factors that lead to a successful implementation of a Big Data system. Market maturity will describe the progression of technologies from immature to mid-maturity to mature. This section provides a high-level overview of the three perspectives of maturity. Other resources provide a more in-depth examination of maturity models. [add references]

## project maturity

Big Data systems adoption often progresses along a path that can be partitioned into a series of distinctly different stages. In the first stage, an application is pilot-tested in an ad hoc project, where a small set of users run some simple models. This prototype system will likely be used primarily (or only) by those in the IT department and is often limited to storage and data transformation tasks, and possibly some exploratory activity.

In the second stage, the project grows to department-wide levels of adoption, where a wider range of user types work with the system. The project may expand beyond storage and integration functions and begin providing a function for one or two lines of business, perhaps performing unstructured data or predictive analysis. The project then faces its largest hurdle of the maturity process, when it attempts to scale from departmental adoption to an enterprise-level project.

Governance is one of the key obstacles to a project during this transition because an enterprise-grade application will be required to have better-defined user roles, better-developed metadata policies and procedures, better control over information silo problems, as well as improvement in other related areas. In the enterprise setting, the project must align more closely with organizational strategies that require higher orders of data quality, data protection, and partnership between IT and business departments.

### Level 1: Ad hoc

In this level, the organization is capturing information in an ad hoc manner. The organization’s departments may be collecting data separately from each other. The data is stored and analyzed using a variety of systems, which may or may not be compatible with one another.

Characteristics of this level include:

* Data not consistently captured and/or stored;
* Spreadsheets frequently used, which could lead to inaccurate information and analytical errors;
* Procedures throughout data life cycle could be nonexistent or could vary across departments;
* Information is siloed; and
* Analytics could be inconsistent across departments.

### Level 2: Department Adoption

In this level, the individual business groups or departments select technologies that satisfy the project need or take advantage of existing worker expertise. ETL (Extract, Transform, Load) / ELT (Extract, Load, Transform) is performed on an as-needed basis and is tailored to specific requests. The system usually cannot readily incorporate new data sources or perform advanced analytics.

Characteristics of this level include:

* Information could be siloed;
* Small systems are developed for individual needs, and interoperability within the systems usually is not a priority;
* Procedures throughout data life cycle could be nonexistent or could vary across departments; and
* A general awareness of data governance is beginning, **perhaps in a single, local application**

### Level 3 Enterprise Adoption

In this level, the enterprise adopts a more systematic approach to Big Data across the organization. Big Data systems begin to address the needs across the organization. An organizational-wide governance program is tackling a larger problem-set, **such as a data warehouse or data lake use case.**

Characteristics of this level include:

* Many systems are integrated to provide cross-company information;
* Data management procedures begin to be developed and implemented; and
* Involves a wider range of personnel expertise.

### Level 4: Culture of Governance

In this level, the organization has fully adopted the Big Data system and utilizes the data and resulting analytics to optimize business processes. A fully developed governance program is tightly integrated across the organization.

Characteristics of this level include:

* Advanced analytics;
* Data or analytical results available to users, level may be based on user groups;
* External users able to access data and/or analytics;
* Greater use of external data;
* Involves a wide range of personnel expertise, from people to develop and maintain the system to data analysts to data visualization experts; and
* Systematic data governance **effort** across the organization.

Data governance refers to administering, or formalizing, discipline (e.g., behavior patterns) around the management of data. While some Big Data projects do not require the observation of governance practices, many, especially in regulated industries such as finance, have serious mandates to observe data governance policy that will need to persist across the entire data life cycle.

In the software development lifecycle [SDLC], there is an old saying known as the Triple Constraint, which states that a project can be completed fast, good, or cheap, but not more than two of the three. As various use cases in Big Data projects have differing requirements along the fast / cheap / good dimensions, we can also see variance in the types of governance program requirements, and roles of the personnel involved, along those same three dimensions.

In terms of types of governance programs, governance for a local business-application use case will not have to cover the same requirements as would a data warehouse use case, or a data lake use case. A data scientist, working in a data lake, may require fast access to raw data that has not been expensive to get into the lake, and would not be considered “good” data in terms of quality; whereas a data warehouse worker does not expect fast access to the data, but does require good data in terms of quality. Each of these facets presents a unique challenge for the creation of appropriate governance measures.

Information management roles and stewardship applications are two of the primary data management challenges organizations face with respect to governance. 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. For example, these departments might be Customer Service, Finance, Marketing, or Sales. 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.

In the early part of Level 4 (Figure 1), the project has achieved integration with organizations’ governance protocols, metadata standards, and data quality management. Finally, a Big Data initiative evolves to a point where it can provide a full range of services including business user abstractions, and collaboration and data-sharing capabilities.

## Organizational Maturity

While technical difficulties such as data integration and preparation are often reported as the greatest challenges to successful Big Data projects, the importance of nontechnical issues such as change management, solution approach, or problem definition and framing should not be underestimated and require significant attention and forethought. As stated in a report from IDC, “An organization’s ability to drive transformation with Big Data is directly correlated with its organizational maturity.” [17] In fact, organizational maturity is often the number-one barrier to success of Big Data projects

### Evolution of organizational maturity

Organizations mature at different rates, depending on a variety of factors, and can take months or years. Organizational maturity is considered below in relation to the four project maturity levels presented in Section 4.1. As a project develops from ad-hoc testing to a fully realized culture of governance, certain organizational changes should be considered for successful system implementations.

These organizational changes are presented below at a very high level. Specific activities to affect organizational change will be dependent on project specifics, an organization’s culture, executive leadership, industry characteristics, and other relevant factors.

Within each level, four broad areas of organizational change can be identified. These broad areas target different aspects of organizational change that should be considered. Each of these general areas involves different actions depending on the level of organizational maturity. For example, in Level 2, training workers might involve a few users on the entire small system, while in Level 4, groups of users might be defined, each of which receives specialized training on a portion of the system. The four broad areas of organizational change are as follows:

* Training of workers, including addressing overall system operations, focused process operations, and cultural changes;
* Management of the technology implementation and change, including a vision of the systems needed, strategic business vision for adopting Big Data systems;
* Workflow development, implementation, and adherence—this could include the development of standards and processes; and
* Technology evaluation, adoption, and implementation.

**Figure 1** maps organization maturity to project maturity and lists some organizational changes that are needed to reach the corresponding level. The lists of considerations are not all-inclusive and can vary depending on the industry, organizational needs, and organizational culture.

Additional references should be consulted for more in-depth examination of the organizational change activities specific to a particular industry, project type, organization type, or other defining project characteristic.

The levels are presented as a continuum with increasingly comprehensive activities to implement Big Data systems. Some of the items might begin in one level with a few activities and jump to a higher level creating gaps in data governance at lower levels that will need to be addressed later. In real life organizations, there is fuzzy boundary between levels and the development of data governance may not occur in a linear and orderly fashion.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **DIFFICULT TRANSITION**   |  | | --- | |  | | **MATURITY** | **PROJECT CHARACTERISTICS** | **ORGANIZATIONAL CHARACTERISTICS** |
|  | **LEVEL 4  Culture of Data Governance** | • External users are able to access the data, computer code, and analytics results.  • Internal user groups are able to access the data, computer code, and analytical results. • Involves a wide range of expertise, including people to develop and maintain the system, data analysts, and data visualization experts. • Uses advanced predictive analytics. • Greater use of external data. • Systematic data governance effort across the organization. | • Consistently uses standardized processes and models across the organization, with slight modifications for nonstandard project or regional needs. • Trains workers in overall system functioning, focused processes, workflows, and safety procedures. • Implements a fully developed and organizational-wide governance policy. • Anticipates organizational needs and responds with appropriate methods or technologies. • Uses external data, including open data, as appropriate. |
|  | **LEVEL 3  Cross- Organizational Adoption** | • Many systems are integrated to provide cross-organizational information. • A wider range of personnel expertise is used.  • Written data management procedures begin to be developed and implemented. | • Federates metadata. • Trains workers in implemented technologies workflows, and safety procedures. • Implements technology Standards. • Develops and implements an organization-wide governance program. • Initiates a master data management (MDM) program. • Appoints a system leader from upper management. |
|  |  | | |
|  | **LEVEL 2  Division Adoption** | • Information may still be siloed. • Interoperability between systems is not a priority. • A general awareness of data governance is beginning, perhaps in some isolated local projects. • Small systems are developed for individual or project needs. • Written procedures applicable to the entire data life cycle may still be nonexistent or incomplete. | • Begins a governance program. • Applies Big Data solutions to well-defined business processes. • There is an unorganized approach. • Appoints a leader for system implementation. |
|  | **LEVEL 1  Ad Hoc functionning** | • Information is siloed. • There is no or inconsistent use of predictive analytics. • Spreadsheets are frequently used, which may lead to inconsistencies, inaccuracies, analytical errors, and lack of interoperability. • Data are not consistently captured and/or stored.  • Written procedures applicable to the entire data life cycle are nonexistent or incomplete. | • The technology used depends on what is available at the time, or on the skill set of the workers. • Little to no training is provided to the workers. • Data collection and/or analysis is designed in response to a particular mandate or need in the moment. • Written procedures governing the data life cycle are nonexistent or vary across projects, groups, divisions, or departments. |

**Figure 1. Evolution of Big Data systems as a function of project and organizational data governance maturity.**

## market maturity of technologies

Technologies progress through a series of stages as they mature, which in broad terms are research and development (R&D), demonstration and deployment, and commercialization, in order of maturation development.

As costs associated with both open source and commercial computing technologies fall drastically, it becomes easier for organizations to implement Big Data projects, increasing overall knowledge levels and adding to a tide effect where all boats in the marina are raised toward maturity. The following technologies represent some of the more recent advances into demonstration and deployment:

* Open source: Maturity of open source technologies is not as prevalent as many media reports would indicate. Open source distributed file systems are essentially still immature stacks, especially in smaller enterprises, although streaming and real-time technology adoption is growing at a fast rate. [12]
* Unified architectures: Challenges persist 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 [18], as unified architectures update indexes in batch intervals. An opportunity exists for open source technologies which are able to apply incremental indexing, to reduce updating costs and increase loading speeds for unified architectures.
* Open data: Some transformations are under way in the biology and cosmology domains, with new activity in climate science and materials science. [15] Various agencies are considering mandating the management of curation and metadata activities in funded research projects. Metadata standards are frequently ranked as a significant technical issue. While agreeing on a local taxonomy snapshot is a large challenge for an organization, managing the difficulties of taxonomy dynamics (which are organizational issues) presents an even more challenging barrier.

Add discussion of Open Data in the context of Open Science

The following technologies represent some of the more recent advances into commercialization.

* Infrastructure as a Service (IaaS): Applications receive a great deal of attention in articles written for business audiences. However, overall, the challenges in applications are proving less difficult to solve than challenges in infrastructure. IaaS is driving many opportunities for commercialization of technology.
* In-memory technologies: It is not always simple to distinguish between in-memory DBMS (Database Management System), in-memory analytics, and in-memory data grids. However, 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 and commercialization.
* 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 connectivity.
* Internal search: In one survey of organizations considering Big Data adoption, “Only 12% have an agreed-upon search strategy, and only half of those have a specific budget.” [16] The top two challenges to internal search seem to be a lack of available staff with the skills to support the function, and the organization’s ability to dedicate personnel to maintain the related servers. Departments are reluctant to take ownership of the search function due to the problematic levels of the issues. The consensus amongst AIIM’s survey respondents was that the Compliance, Inspector General, or Records Management department should be the responsible owner for the search function. An underlying problem persists in some larger organizations, however, where five or more competing search products can be found, due to small groups each using their own tools.
* Stream processing: Continued adoption of streaming data will benefit from technologies that provide the capability to cross-reference (i.e., unify) streaming data with data at rest.

## Big Data Trends and Forecasts

In the early years of Big Data, organizations approached projects with the goal to exploit internal data, leaving the challenges of dealing with external data for later.

The usage of a *hub and spoke* architecture for data management emerged as a pattern in production environment implementations [19], which still relied heavily on ETL processes. The hub and spoke architecture provides multiple options for working with data in the hub, or for moving data out to the spokes for more specific task requirements, enabling for data persistence capabilities on one hand and data exposure (i.e., for analytics) capabilities on the other.

In 2017, in-memory, private cloud infrastructure, and complex event processing have reached the mainstream. Modern data science and machine learning are slightly behind but moving at a very fast pace to maturity.

***Table 7*** lists select technologies that are projected to mature in the near future and have a significant impact on the advancement of Big Data.

**Table 7. Maturity projections**

|  |  |
| --- | --- |
| 2017 – 2020 | 2020 - 2025 |
| * High-performance message infrastructure * Search-based analysis * Predictive Model Markup Language | * Internet of things * Semantic web * Text and entity analysis * Integration |

An increase is expected in the application of semantic technologies for data enrichment. Semantic data enrichment is an area that has experienced successes in cloud deployments. Several applications of text analysis technology are driving the demand for standards development including fast-moving consumer goods, fraud detection, and healthcare.

Integration is also an area of projected maturity growth. Increased usage is expected of lightweight iPaaS (integration Platform as a Service) platforms. Use of application programming interfaces (APIs) for enabling microservices and mashup data from multiple sources are also anticipated to grow. Currently, there is a scarcity of general use interfaces that are capable of supporting diverse data management requirements, container frameworks, data APIs, and metadata standards. Demand is increasing for interfaces with flexibility to handle heterogeneous user types, each having unique conceptual needs.

# Implementation and Modernization

## System Modernization

Organizations face many challenges in the course of validating their existing integrations and observing the potential operational implications of the rapidly changing Big Data environment. Start with a plan for going through the process of modernization. Stages like a, b, c, and d.

Many modernization projects follow some method for portfolio roadmapping. One such method is referred to as a technology brick roadmap. Brick structures classify products along the lifecycle, such as emerging, mainstream, and retirement, and also map out the implementation timelines for each technology on a chart. The following link shows a brick roadmap in figure 3.2.8: <http://www.statcan.gc.ca/pub/11-634-x/2016001/section3/chap2-eng.htm> which also has a description.

ROI [placeholder]. Maybe mention an impact filter here, or find place for impact table in N.

Stage A. or B, at some point Effectiveness of the plan is dependent on a clear understanding of new technologies. This section attempts to look at the industries and technologies related to Big Data and economic impacts by viewing them in context of the broader landscape.

Emerging technologies are at the beginning stage of maturity. Maturity is conditional on adoption. Looping back to Section 4.3, Pull concepts from EmergingT+V3 doc.

An organization preparing to develop a Big Data system will typically consider one of two possible directions for modernization. For simplification, these two directions will be referred to as Augmentation and Replacement. Each of these two modernization directions has unique advantages and disadvantages. The following bullets summarize the two directions:

* Augmentation: This direction involves updating to a Big Data system by augmenting the supporting architecture. Advantages of updating the supporting architecture include incorporation of more mature technologies amidst the stack and flexibility in the implementation timeline. Augmentation allows for a phased implementation that can be stretched out over more than one fiscal budget year.
* Replacement: This direction involves updating to a Big Data system by replacing the existing system with 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. One advantage of complete system replacement is reduced compatibility problems with legacy systems. Partial modernizations, by replacing a portion of the existing system, are also possible. However, the same advantages and disadvantages of complete system replacement may not apply.

Once system augmentation or replacement has been elected, a method of implementation can be chosen. **Figure 2** diagrams a decision situation, commonly referred to as *build or buy* (or outsource) that organizations face when modernizing to a Big Data system. In the build, or 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 skill set covering the newer technologies such as streaming or near real-time analysis.



**Figure 2. New system implementation**

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 get the new system running and migrate data or combine components.

The alternative to the DIY scenario is for the organization to buy or rent a new Big Data system. Renting usually refers to cloud solutions. Advantages to buying or renting include the ease of scale and not having to operate two systems simultaneously (or not having to modify an existing system).

Hybrid parallel systems are those that are not 100% integrated with the existing system. For example, organizations can use the cloud for storage but develop their own applications. One disadvantage is the high cost of moving data to the cloud. Developing standards for hybrid implementations should accelerate the adoption and interoperability of analytics applications.

Challenges exist with any of the implementation routes (DIY, buy or rent new system, or hybrid parallel systems). For example, data cleansing and systems plumbing are persistent hurdles no matter which type of project is undertaken. [20] [21]

When considering the augmentation pathway, the advantages and disadvantages should be examined. While the full list of advantages and disadvantages will be project-specific, ***Table 8*** provides a high-level list.

**Table 8. Advantages and Disadvantages of System Modernization via the Augmentation Pathway**

|  |  |
| --- | --- |
| Advantages | Disadvantages |
| Build | |
| * Phased approach | * Technically demanding * Fewer support options |
| Buy |  |
| * Phased approach * Not entirely immature stack of technology | * Potential vendor lock in issues |
| Hybrid |  |
| * Phased approach | * Potential compatibility problems with legacy systems |

In a similar fashion, ***Table 9*** provides a high-level list of advantages and disadvantages of the replacement pathway.

**Table 9. Advantages and Disadvantages of System Modernization via the Replacement Pathway**

|  |  |
| --- | --- |
| Advantages | Disadvantages |
| Build | |
| * Reduced compatibility problems with legacy systems | * Longer development cycle * Increased change management * Less mature technologies |
| Buy | |
| * Reduced compatibility problems with legacy systems | * Longer development cycle * Increased change management * Less mature technologies |
| Hybrid | |
| * Reduced compatibility problems with legacy systems | * Longer development cycle * Increased change management * Less mature technologies |

In every case, lower-level or lower-layer components of the system must be considered as equally (if not more) important as analysis or analytics functions. Future work on this volume may include improved coverage of an entire system modernization.

In addition to the modernization of complete systems, the modernization of analytics applications will be considered—specifically with respect to machine learning. Some motivations for modernizing analytics include the following:

* Improved monitoring and reporting: Basic descriptive business intelligence may be improved though use of Big Data systems;
* Improved diagnostics, forecasting, and predictive analysis: The term *predictive analysis* is often used to refer to analysis which is not exactly predictive in the common sense of the word;
* Enriched decision making: This function comprises 70% of the demand for analytics in 2017. [22] While operational decisions can be rule-based, not involving analytics, strategic decisions are optimization tasks.

The next section covers some of the questions related to system capability that an organization may need to consider when planning their own system.

## Implementation

Characteristics of a Big Data project implementation depend on the needs and capabilities of the particular organization undertaking the effort. This section attempts to provide some high-level issues for deliberation during the Big Data project planning stage. This is not intended to be a prescription covering the entire range or depth of considerations that an organization may face, but rather an initial list to supplement with project-specific concerns. During the planning phase, Big Data project considerations could include the following:

* Data quality: Consider the level of quality that will be required from the data model. As data quality increases, cost increases. A minimum viable quality of data, which will provide desired results, should be determined.
* Data access: Many factors can affect data access including organizational cultural challenges and security and privacy compliance. Cultural challenges are unique to each project but many are alleviated with sufficient support from upper management (e.g., corporate officers, influential advocates). Security and privacy affects multiple areas in a Big Data project including data access. Additional information on security and privacy considerations are provided in the *NBDIF: Volume 4, Security and Privacy* document.
* Component interoperability: For a complicated system, a comprehensive appraisal of system component interoperability can be critical. Advantages of commercial products are frequently lauded while the limitations, dependencies, and deficiencies are often not obvious. Exploration of component interoperability during the planning phase could prevent significant issues during later phases of Big Data projects.
* Potential bottlenecks: Projects requiring high performance often expose storage and network bottlenecks.
* For search-oriented projects: Organizations should strive to set a balance between governance and retrieval, determine ownership (i.e., departmental responsibility) for the function, aim for unified or single-point search capability; and unless the organization is a strong IT company, identify needed outsourced expertise.

[Placeholder for] New philosophy to “implement first” / quickly, without going thru the traditional steps. [mit] <https://sloanreview.mit.edu/article/implement-first-ask-questions-later-or-not-at-all/?social_token=fe68d2fc77840cfb52730af3e33dbafe&utm_source=linkedin&utm_medium=social&utm_campaign=sm-direct>

# Specific solution Techniques, Dependent on the Problem Space

Add definitions and discussion of data mining, machine learning, and artificial intelligence.

**Figure 3** very much oversimplifies some of the questions related to system capability that an organization may need to consider when planning their own system; its purpose here is to demonstrate how project requirements can drive decision making. The list of choices presented is not intended to be comprehensively complete. Inclusion is not an endorsement for usage, and no solutions have been intentionally excluded.



**Figure 3. Requirement decision tree**

After the scalability and latency requirements are considered as shown in **Figure 3**, the systems planning process will require continued consideration on whether machine learning is necessary. ***Figure 4***, ***Figure 5***, and ***Figure 6*** map the workflow of the machine learning decision trees and show the decision points in the application of machine learning algorithms. ***Table 10***, ***Table 11***, ***Table 12***, and ***Table 13*** list specific algorithms for each algorithm subgroup.

[placeholder] area for no correct answer to this selection.

[

**Figure 4. Machine learning algorithm application workflow**

***Figure 4*** shows the decision steps for application of a machine learning algorithm including the input preparation phase (e.g., feature engineering, data cleaning, transformations, scaling). ***Figure 5*** and ***Figure 6*** expand on algorithm choices for each problem subclass. ***Table 10*** and ***Table 11*** continue from ***Figure 5*** to provide additional information for the regression or classification algorithms. ***Table 12*** and ***Table 13*** provide additional information on the unsupervised algorithms and techniques shown in ***Figure 6***.



**Figure 5. Supervised Machine Learning Algorithms**

Figure 6. Unsupervised or Reinforcement Machine Learning Algorithms

Super vised learning problems involve datasets that have the feature which is trying to be predicted / measured for all observations or a subset of all observations (semi-supervised learning). The measurements for the feature which is trying to be predicted by the machine learning model are called labels. In supervised learning problems, the labeled data is used to train the model to produce accurate predictions.

Supervised learning problems can be classified into two subgroups of algorithms: regression or classification. Regression algorithms predict a continuous variable (a number), and classification algorithms predict a category from a finite list of possible categories. ***Table 10*** and ***Table 11*** compare supervised learning regression algorithms using four categories and supervised learning classification algorithms using the same four categories.

**Table 10. Supervised Learning Regression Algorithms**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Training Speed | Interpretability | Pre-Processing | Other Notes |
| Linear Regression | Fast | High | Centering and Scaling, Remove Highly Correlated Predictors | Speed at the expense of accuracy |
| Decision Tree | Fast | Medium |  | Speed at the expense of accuracy |
| Random Forest | Fast | Medium |  | Fast and accurate |
| Neural Network | Slow | Low | Centering and Scaling, Remove Highly Correlated Predictors | Accurate |
| K Nearest Neighbors | Fast | Low |  | Scales over medium size datasets |
| Ridge Regression | Fast | High | Centering and Scaling |  |
| Partial Least Squares | Fast | High | Centering and Scaling |  |
| Cubist | Slow | Low |  |  |
| Multivariate Adaptive Regression Splines (MARS) | Fast | Medium |  |  |
| Bagged / Boosted Trees | Fast | Low |  | Accurate, large memory requirements |

**Table 11. Supervised Learning Classification Algorithms**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Training Speed | Interpretability | Pre-Processing | Other Notes |
| Support Vector Machine | Slow | Low | Centering and Scaling | Speed at the expense of accuracy |
| Logistic Regression | Fast | High | Centering and Scaling, Remove Highly Correlated Predictors | Speed at the expense of accuracy |
| Decision Tree | Fast | Medium |  | Speed at the expense of accuracy |
| Random Forest | Slow | Medium |  | Accurate |
| Naïve Bayes | Fast | Low |  | Scales over vary large datasets. Speed at the expense of accuracy |
| Neural Network | Slow | Low | Centering and Scaling, Remove Highly Correlated Predictors |  |
| K Nearest Neighbors | Fast | Low |  | Scales over medium size datasets |
| Ridge Regression | Fast | High | Centering and Scaling |  |
| Nearest Shrunken Centroids | Fast | Medium |  |  |
| MARS | Fast | High |  |  |
| Bagged / Boosted Trees | Slow | Low |  | Accurate |

Unsupervised learning problems do not have labeled data and can be classified into two subgroups: clustering algorithms and dimensionality reduction techniques. Clustering algorithms attempt to find underlying structure in the data by determining groups of similar data. Dimensionality reduction algorithms are typically used for preprocessing of datasets prior to the application of other algorithms. ***Table 12*** lists common clustering algorithms, and ***Table 13*** lists common dimensionality reduction techniques.

**Table 12. Unsupervised Clustering Algorithms**

| Name | Pre-Processing | Interpretability | Notes |
| --- | --- | --- | --- |
| K -means | Missing value sensitivity, Centering and Scaling | Medium | Scales over large datasets for clustering tasks, must specify number of clusters (k) |
| Fuzzy c-means |  |  | Must specify number of clusters (k) |
| Gaussian | Specify k for probability tasks |  | Must specify number of clusters (k) |
| Hierarchical |  |  | Must specify number of clusters (k) |
| DBSCAN |  |  | Do not have to specify number of clusters (k) |

While technically dimension reduction may be a preprocessing technique, which transforms predictors, usually driven for computational reasons, some consider dimensionality reduction (or data reduction) techniques a class of unsupervised algorithms because they are also a solution for unlabeled data.

In that these methods attempt to *reduce* the data by capturing as much information as possible with a smaller set of predictors, they are very important for Big Data. Many machine learning models are sensitive to highly correlated predictors, and dimensionality reduction techniques are necessary for their implementation. Dimensionality reduction methods can increase interpretability and model accuracy, and reduce computational time, noise, and complexity.

**Table 13. Dimensionality Reduction Techniques**

|  |  |  |
| --- | --- | --- |
| Name | Interpretability | Notes |
| Principal Component Analysis (PCA) | Low | Scales to medium or large datasets |
| Correlation Filters |  |  |
| Linear Discriminant Analysis (LDA) |  |  |
| Generalized Discriminant Analysis (GDA) |  |  |
| Backward Feature Elimination |  |  |
| Singular Value Decomposition (SVD) |  |  |

While a wide array of algorithms has been classified in the preceding tables, a technique called ensemble modeling is widely used to combine the results of different types of algorithms to produce a more accurate result. Ensemble methods are learning algorithms that take a weighted vote of their different model’s predictions to produce a final solution. In practice, many applications will use an ensemble model to maximize predictive power.

# Big Data readiness

Big Data has the potential to answer questions, provide new insights previously out of reach, and strengthen evidence-informed decision making. However, the harnessing of Big Data can also very easily overwhelm existing resources and approaches, keeping those answers and insights out of reach.

Lack of data governance at the corporate level and data management gaps at the working levels, reflect the state of affairs in the private and public sectors in developed countries that are dealing with decades old legacy systems and ways of doing things. *Big Data* is a rapidly evolving area as evidenced by current efforts to develop new international Standards to provide guidance as we collectively move forward. These efforts will inevitably and necessarily profoundly impact overall data management practices at all levels and for all types of data.

It is important that an organization identify the technical and non-technical barriers to Big Data (**Table 6**). Contextualization of a path to Big Data readiness within a framework that describes Big Data reference architecture and Big Data governance and metadata management is also important. However, an effective first step will emphasize what can be done now in the present taking into account current realities (**Figure 1**) to position the organization to meet opportunities provided by the Big Data revolution.

Governance will be the foundation for readiness. As the success of Big Data system adoption relies heavily on organizational maturity (**Figure 1**), this section offers suggestions for a PATH TO BIG DATA READINESS (**Vol 6, Fig 2, top left corner**).

## Tactical actions

Big Data transformation does not need to happen all at once; nor does an organization need to wait for the development of an entire Big Data Framework (**Vol 6, Fig 2, 4, 6-8**) or Governance model before taking action to help accelerate the implementation of *Big Data*.

Putting the initial focus on structured digital scientific data and the identification of a pathway from Small Data to Big Data will provide a rational stepwise approach to harnessing Big Data. Implement actions that are generic and independent of systems currently in place. This means that they can be implemented “*now*.”

With small steps toward Big Data, tactical actions directed at the working level can help enable *Big Data* without overwhelming workers, managers, or stakeholders, and increase the chances of success of a Big Data project.

1. Create awareness of “*Big Data readiness*” from the bottom up in operations and research contexts via communications such as newsletters, bulletins, and a dedicated website or wiki.
2. Provide online training modules to increase digital literacy across the organization.
3. Deploy “It’s good enough” checklists for data Findability, Accessibility, Interoperability, and Re-usability (FAIR data) to help data providers produce data that are ready for Big Data workflows.
4. Implement “*user-centric,*” approaches to data preparation to replace project- and client-centric approaches. [[2]](#footnote-3)
5. Create “*linear data pathways*” to authoritative data sources to eliminate data fragmentation, duplication, and to preserve data lineage.
6. Develop and pilot test models of data-intensive scientific workflows for the preparation of FAIR, tidy, and analysis ready data and “*reproducible science*” in line with national and international best practices.
7. Implement semi-automated data verification and feedback loops to ensure that data are ready for integration into Big Data workflows.
8. Maximize chances of success of Actions 1-7 by including data providers in the development of solutions.

## **Fundamental Changes needed to enable big data**

**Breaking out of “*Lock-in*”**

Not to be confused with vendor lock-in, organizations can be locked into old ways of thinking and old ways of doing things that impede Big Data. Best practices in data management have not kept up with changes in technology that resulted in a rapid increase in the speed of generation, quantity, variety, complexity, variability and new uses for the data collected. There is in addition, uncertainty regarding data accuracy, and inconsistency in vocabulary and confusion over the meaning of Big Data. Meanwhile, organizations are still struggling to emerge from a paper-based world governed in silos to a digitally interconnected world. This is a difficult transition. It requires the transformation of longstanding, well-adapted thinking processes that no longer work well, to new thinking processes adapted to a new world.

**Change in thinking**

*Big Data* is being propelled from an emerging area to the fore of Open Data and Open Science. However, data that may be “locked in” traditional approaches are largely inaccessible to Big Data end users. This limits an organization’s ability to use Big Data approaches for knowledge acquisition and innovation. Changes in thinking across organizations are needed to achieve a coordinated and harmonized system that is simple, effective and geared to meet organizational needs.

**Culture change**

Operational and research programs have developed data management processes that work for them internally. They tend to be project- or client-centric to meet their specific mandate and needs, but not necessarily user-centric in the context of Open Science and Big Data.

A paradigm shift in thinking and culture is needed across organizations to achieve agile delivery of “*analysis-ready*” data that can be incorporated seamlessly into a *Big Data* workflow. The underlying principle for success is a “*Big Data readiness*” approach from the bottom up at the working level, operations, research, and business lines. Targeted generic actions will help create the necessary conditions on the ground – culture change will follow.

## Big Data readiness

**FAIR data**

Data providers in the field, laboratory, and other organizational levels need to recognize at the outset that the data users are unknown. They also need to accept that how the data will be used and for what purpose will remain unknown to them. Data transmitted from one person or group to the next must be FAIR and tidy.

FAIR data include all related metadata and documentation so that an unknown end-user can completely understand the data and the data quality without having to contact the data provider. FAIR data have been verified by the data provider to be “fit for use” by any future unknown user who is then in a position to assess whether or not the data are “fit for purpose” in some specific context. FAIR, tidy, analysis ready data can be easily integrated into a Big Data workflow.

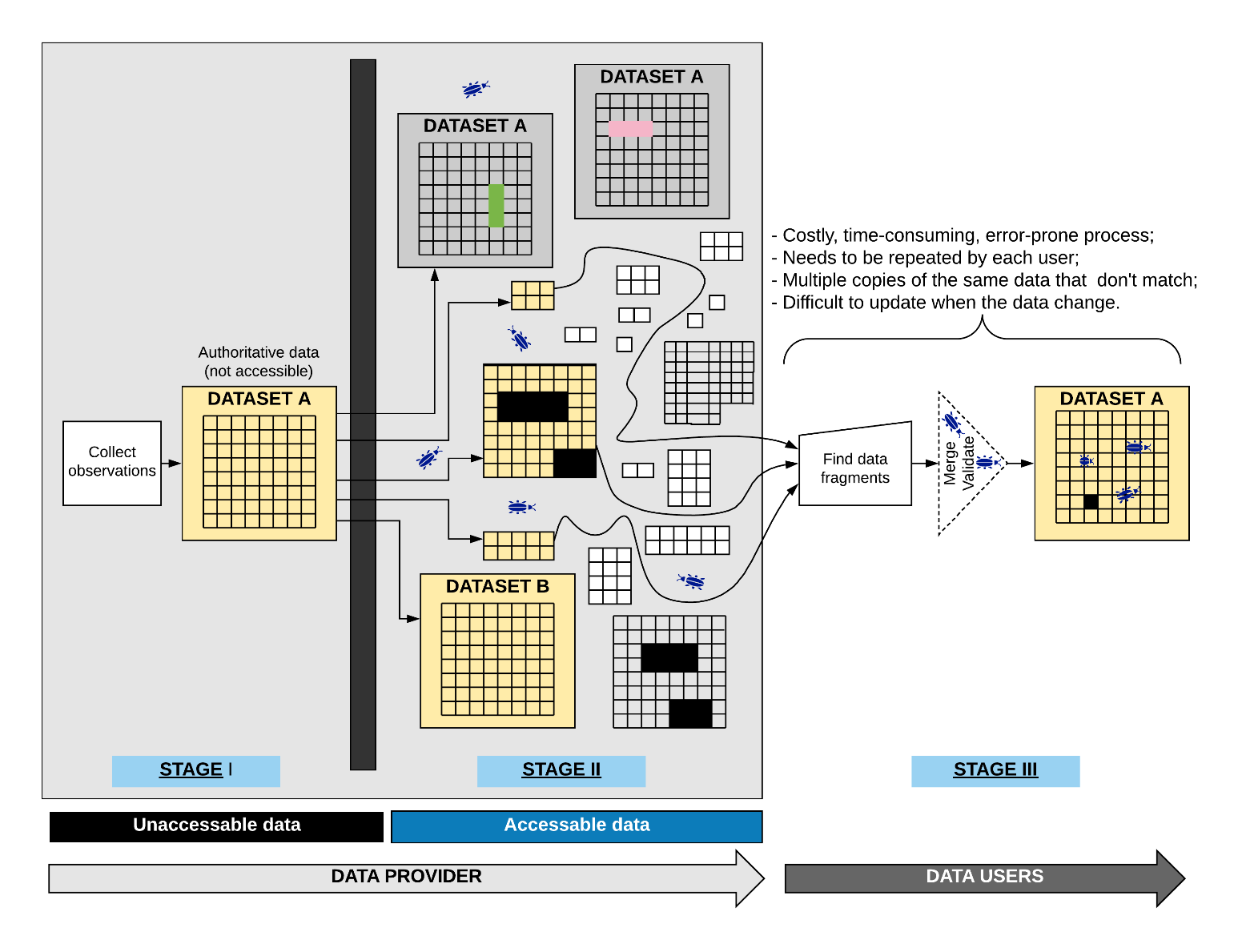
A *Big Data* *readiness* approach at the working level will concomitantly help solve existing data flow and data quality issues irrespective of whether or not the data will eventually enter a *Big Data* workflow. A *Big Data* readiness approach will improve an organization’s overall data stewardship and governance, help make Open Data and Open Science a reality, and improve the chances of success of future corporate solutions related to Big Data and analytics.

**The problem space**

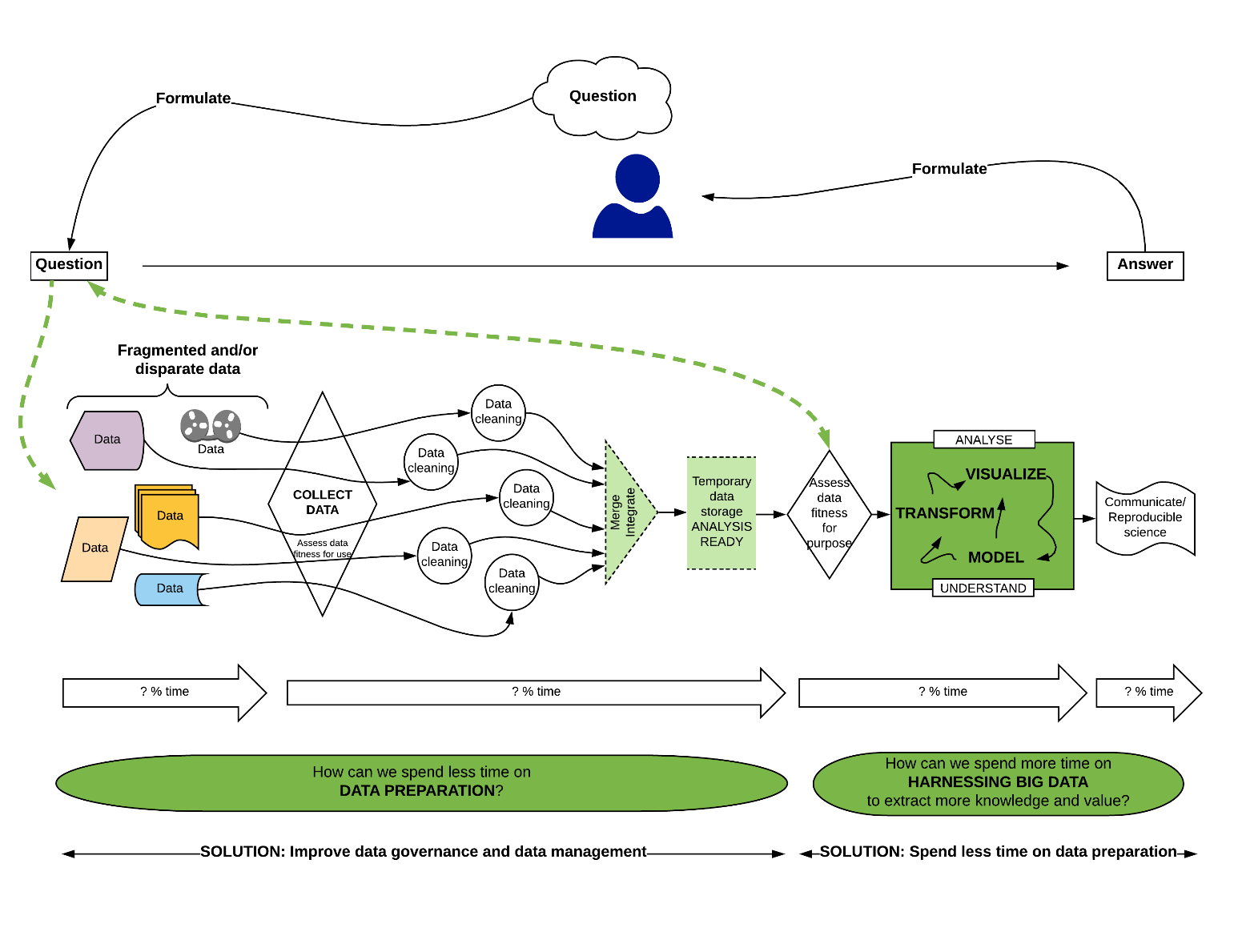
There is a need for common data Standards for the preparation and updating of FAIR data. Previous approaches to Data Governance may have led to data fragmentation (**Figure 6 and Figure 7**), variation in data quality, and incomplete information concerning the data. Where this may be satisfactory within specific mandates, it is problematic for *Big Data*. **Figure 6** shows how a very high-quality dataset produced under strict quality assurance/quality control (QA/QC) protocols ends up fragmented in the absence of data governance that encompasses the complete data life cycle management. From the viewpoint of the data providers, they have produced high quality data. From the viewpoint of the data users, they see poor quality data. In order to use such data, each user inherits the task of reassembling the data before being able to use them yet lacks all the information needed to perform the task reliably. This is an error-prone, costly, time consuming, and inefficient use of resources. Furthermore, it is unlikely that data reassembled by different end-users will result in matching datasets. The problem compounds exponentially when trying to integrate these data into *Big Data.* Targeted actions address gaps in Data Governance and data management to improve the ability to integrate data from multiple sources and to extract reliably new knowledge and insights from large and complex collections of digital data. Adopting a Big Data readiness approach in an organization can help enable Big Data analytics, machine learning and Artificial Intelligence (AI).

A common workflow is illustrated in **Figure 8** where multiple datasets from different sources somehow have to be merged. In addition to the problem of dataset fragmentation, there is confusion about which version is the approved copy, lack of version control, and absent or incomplete metadata.

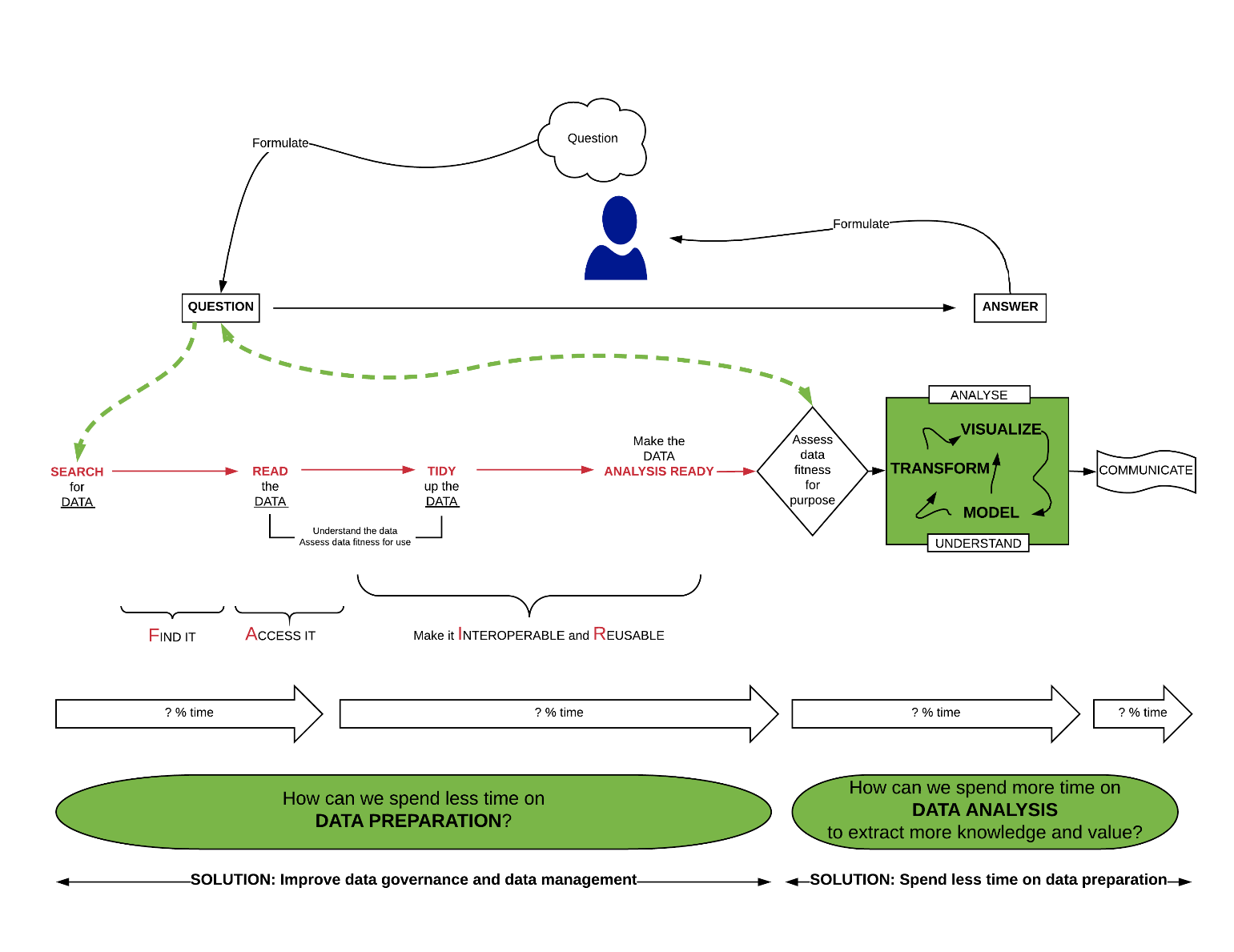
Data governance is the solution to these dataset, data flow, and metadata problems and the hidden costs that result from them.



**Figure 7. Dataset fragmentation.**



**Figure 8. Data fragmentation**



**Figure 9. Time spent making data analysis ready**

## Disrupting the status quo

**Implementation of a *Big Data readiness***

Implementation of a *Big Data readiness* approach at the working level may be easier to implement than imagined. The person best equipped to prepare “*analysis-ready*” data is the data provider – the person at the data source who knows the data best. Success requires inclusion of data providers – especially those who are experiencing the greatest challenges – in developing solutions. Inclusion means going beyond providing support. It means saying not only, “*What can we do for you?*” but also, “*This is what we need from you*.” It means disrupting the status quo. *Big Data readiness* requires a paradigm shift in thinking at the working levels that is revolutionary, not evolutionary.

**It’s “good enough”**

People are easily overwhelmed by disruption of the status quo. This can be mitigated by developing well thought out, “*It’s good enough*” modular checklists that will result in what is needed now to move forward on the pathway to *Big Data*. It is unrealistic to expect that people at the working level, in the field and in the laboratories, have or can acquire the necessary skills and tools to design and maintain databases or to output their data in unfamiliar formats. However, it is realistic and necessary to expect that they can output their data in a form that can be easily understood and used by other people and systems. If this is achieved, it will be good enough.

## Harnessing Big Data

Harnessing Big Data means extracting more knowledge from existing data. A major hurdle is data preparation which can take up to 70% or more of the total time (**Figure 6** and **Figure 7**), essentially performing tasks left undone when data providers release data that are not FAIR (**Figure 8**).

**The solution space**

*Data governance*

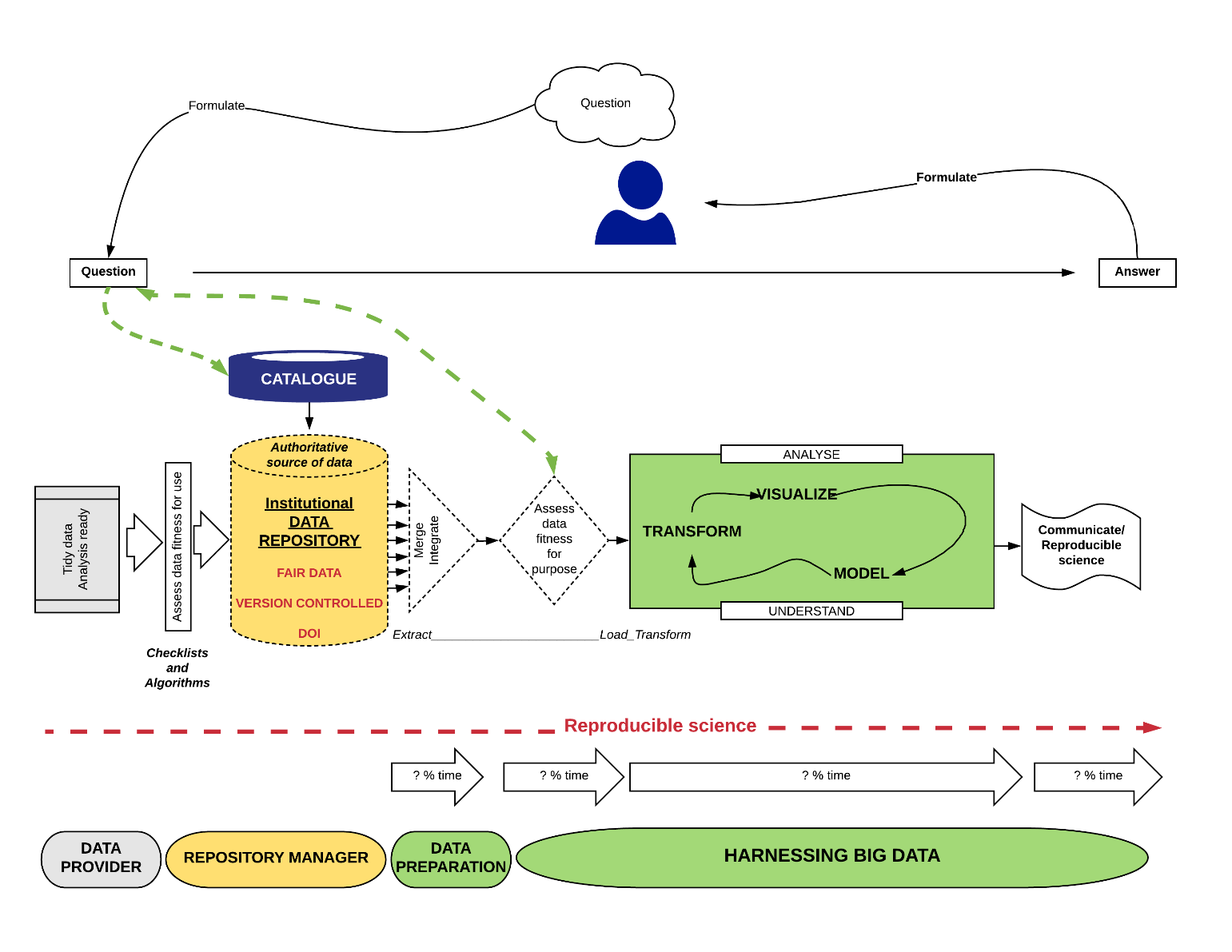
Data governance is the solution to extensive data preparation time. Data governance and improved data management frees up time for analysts to do analysis instead of data cleaning and preparation. The onus needs to be put on the data provider to provide FAIR data that are ready for analysis. Time is thus freed-up can then be used for the harnessing of Big Data in the continuum of reproducible science. **Figure 9** is a solution diagram for a single organization. The yellow container is the input/output control point and source of authoritative data.

*Cost savings*

Big Data and Data Governance are also tools to reduce costs. Big Data reduces costs by using existing data instead of collecting more data unnecessarily. Big Data may also reduce costs getting better answers quicker. However, Big Data will not improve data quality, solve data management problems, or obviate the need for good quality, well managed data. Good data governance and FAIR data will result in the reduction or elimination of inefficiencies and costly errors. Improved data quality, usability and discoverability will increase the value of data products thereby providing a bigger return on investment.

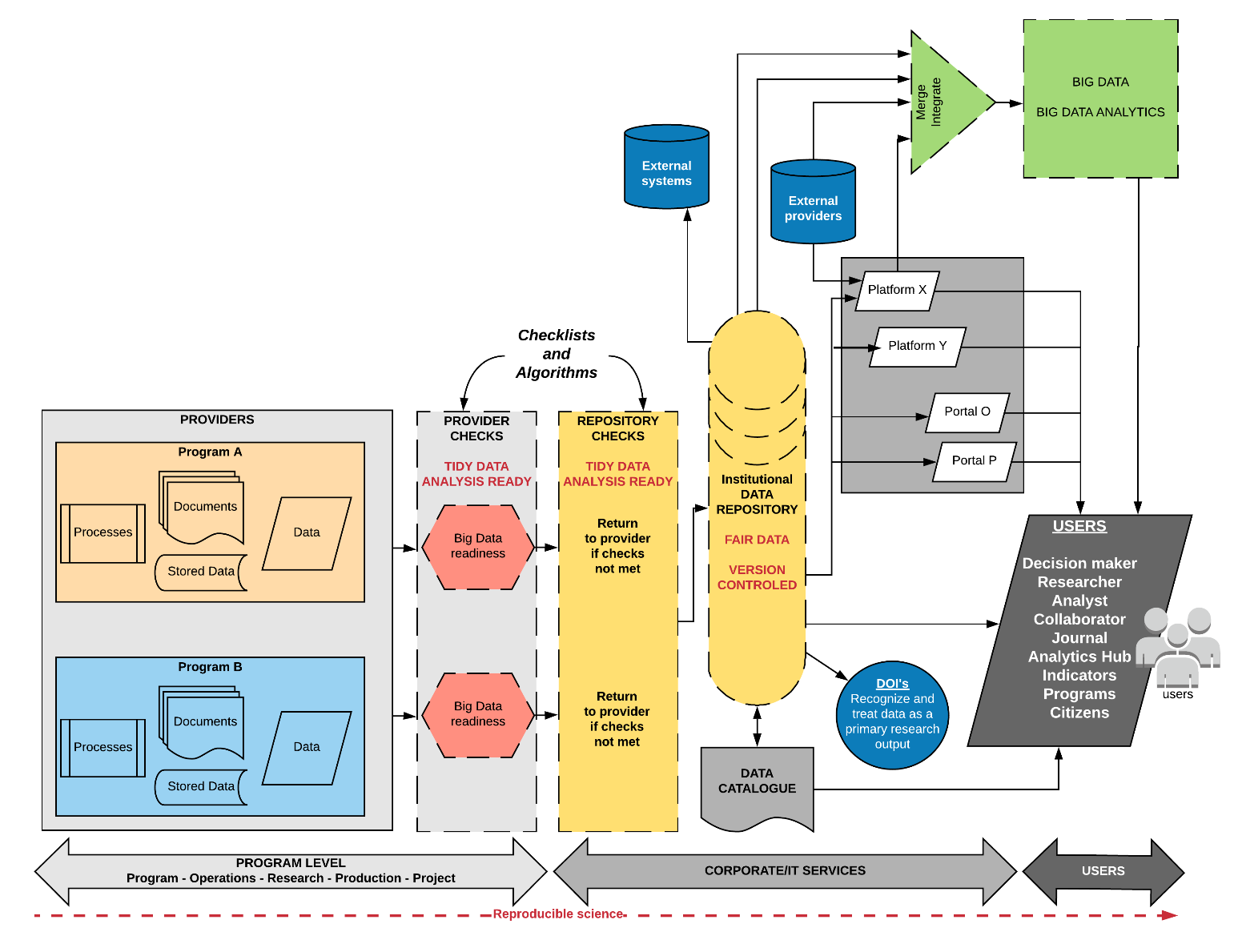
*Linear data pathways*

As the organization matures it will be able to implement linear pathways to authoritative data (**Figure 10**). Checklists implemented on the input side of the control point in **Figure 9** and **Figure 10** are a critical component to ensure that the organization has a single source of authoritative data.



**Figure 10. Improved data governance.**

Checklists implemented on the input of the data repository (yellow container) are critical component.



**Figure 11. Linear data flows for authoritative data**

## Checklists

Checklists are a useful data management tool to help generate FAIR data and to maintain data quality, consistency, and transparency. [23-26] In addition, a well-designed checklist can serve multiple purposes, for example:

* The data provider can use the checklist as an auto-evaluation tool.
* Checklist results can be submitted to management in lieu of the actual data for the purpose of data approval.
* The institutional digital repository can use the checklist to identify datasets for acceptance into the repository, and to return to the provider for correction datasets that fail to meet all the criteria.
* Management can easily merge checklist results received from across an organization to get a snapshot of the overall state of data management.
* Management can quickly scan the results to identify areas in need of improvement.

Examples of checklists are provided in Appendix B. The questions are formulated such that the “correct” answer is ‘yes’ which makes it easy to scan. Controlled answers are: ‘*yes*’, ‘*no*’, ‘*I don’t know*’, or ‘*not applicable*’.

1. Big Data readiness checklists

|  |  |  |  |
| --- | --- | --- | --- |
| line | **MODULES Scientific computing category** | **CHECKLISTS Big Data readiness** | **Does your dataset comply with the items in the checklist?** |
| 1 | Metadata management | Do the metadata include a description of the dataset? | yes |
| 2 | Metadata management | Does the dataset have a persistent identifier? | yes |
| 3 | Metadata management | Do the metadata include a dataset creation date? | yes |
| 4 | Metadata management | Do the metadata include a dataset update date? | yes |
| 5 | Metadata management | Do the metadata include a description of the temporal coverage? | yes |
| 6 | Metadata management | Do the metadata include a description of the geospatial coverage? | yes |
| 7 | Metadata management | Do the metadata identify the creator of the dataset? | yes |
| 8 | Metadata management | Do the metadata identify the contributors to the dataset? | yes |
| 9 | Metadata management | Do the metadata include a link to related publications? | yes |
| 10 | Metadata management | Do the metadata include a link to related data products? | yes |
| 11 | Metadata management | Do the metadata include keywords to improve dataset discoverability? | yes |
| 12 | Metadata management | Are all metadata provided in a machine-readable format? | yes |
| 13 | Metadata management | Are the terms used in the metadata compliant with relevant metadata standards or ontologies? | yes |
| 14 | Metadata management | Do the metadata include a citation that is compliant with JDDCP? | yes |
| 15 | Metadata management | Do the metadata include a description of the methods used for data collection? | yes |
| 16 | Metadata management | If the dataset comes from model output, do the metadata include a description of the model that was used? | yes |
| 17 | Metadata management | Do the metadata include a description of the experimental set-up? | yes |
| 18 | Metadata management | Is this dataset part of a data collection? | yes |
| 19 | Metadata management | Do the metadata include a description of the data collection, if applicable? | yes |
| 20 | Metadata management | Is there a data dictionary that describes the contents, format, and structure of the tables in the data collection, and the relationship between the tables? | yes |
| 21 | Data collection | Was a quality control technique such as" Statistical Process Control" used to ensure that collected data are accurate? | yes |
| 22 | Data collection | If the dataset includes data from a testing or calibration laboratory, was the laboratory method accredited? e.g., ISO/IEC 17025:2017 standard (originally known as ISO/IEC Guide 25). | yes |
| 23 | Data preparation | Were check-digits used on known unique identifiers to ensure valid values? | yes |
| 24 | Data preparation | Were drop-down menus, look-up tables or reference lists used for variables that should have a fixed code set? | yes |
| 25 | Data preparation | Are dates formatted according to the ISO 8601 Standard (e.g., YYYY-MM-DD)? | yes |
| 26 | Data preparation | Are times formatted according to the ISO 8601 Standard (e.g., HH:MM)? | yes |
| 27 | Data preparation | Where the dataset contains measured observations, are the units provided in a separate column? | yes |
| 28 | Data preparation | If the dataset contains latitude/longitude, is the date d a t u m provided? | yes |
| 29 | Data preparation | Are the data files tabular? i.e. There is one rectangular table per file, systematically arranged in rows and columns with the headers (column names) in the 1st row. Every record (row) has the same set of column names. Every column contains the same type of data, and only one type of data. | yes |
| 30 | Data management | Are the raw data available online? | yes |
| 31 | Data management | Are the raw data backed up in more than one location? | yes |
| 32 | Data management | Are all the steps used to process the data recorded and available online? | yes |
| 33 | Data management | Does each record (row) have a unique identifier? | yes |
| 34 | Data management | Have you anticipated the need to use multiple tables? | yes |
| 35 | Data management | Can the tables in a data collection be linked via common fields (columns)? | yes |
| 36 | Data management | Have the data been submitted to a reputable DOI repository? | yes |
| 37 | Data management | Do the files have names that are meaningful to humans? | yes |
| 38 | Data management | Do the variables (column) have names that are meaningful to humans? | yes |
| 39 | Data management | Have the data been deduplicated? | yes |
| 40 | Data management | Are the data FAIR (Findable, Accessible, Interoperable, Re-usable)? | yes |
| 41 | Data management | Was a logical, documented naming convention used for variables (column names)? | yes |
| 42 | Data management | Was a logical, documented naming convention used for file names? | yes |
| 43 | Data management | Were the data documented, "*as-you-go*" rather than at end the end of the process? | yes |
| 44 | Data management | Is a description of the quality control and quality assurance (QA/QC) procedures available online? | yes |
| 45 | Data management | Were measures taken to protect security of data in all holdings and all transmissions through encryption or other techniques? | yes |
| 46 | Data management | Were measures taken to protect against disclosure or theft of confidential information? | yes |
| 47 | Data management | Is a description of the measures taken to protect against disclosure or theft of confidential information available online? | yes |
| 48 | Data management | Were measures taken to ensure a "*single source of truth*" to minimize duplication of information and effort? | yes |
| 49 | Data management | Were standard formats used for names? | yes |
| 50 | Data management | Were standard formats used for civic addresses? | yes |
| 51 | Data management | Are the datasets prepared at the lowest possible level of granularity? (i.e. the data are not summary statistics or aggregated data) | yes |
| 52 | Data management | Are new datasets output at regular, predictable intervals (e.g., the last day of every month, the last day of the year)? | yes |
| 53 | Data management | Is the dataset located in a repository meeting CoreTrustSeal standards? | yes |
| 54 | Data management | Is there a description of the steps performed during data preparation? | yes |
| 55 | Data fitness for use | Are the data tidy? i.e. the data can be read by statistical or database software (other than Excel, Word, or Acrobat) without the need to write extensive computer code to extract information to put it in a machine useable form. | yes |
| 56 | Data fitness for use | Are the data analysis ready? | yes |
| 57 | Data fitness for use | Are the data machine readable? | yes |
| 58 | Data fitness for use | Can the data be ingested directly into statistical or database software (other than Excel, Word, or Acrobat) without the need to write extensive computer code? | yes |
| 59 | Data fitness for use | Are the data in CSV (i.e. comma separated, or character separated) format? | yes |
| 60 | Data fitness for use | Was a "user-centric" (i.e. the end-user is unknown), rather than a project- or client-centric approach used for data preparation? | yes |
| 61 | Data fitness for use | Can the data be incorporated seamlessly into a Big Data workflow? | yes |
| 62 | Data fitness for use | Are the data files in a non-proprietary format? | yes |
| 63 | Data fitness for use | Are new data appended to existing data files? | yes |
| 64 | Data fitness for use | Did you follow specified data quality assurance practices in the production of these data? | yes |
| 65 | Data fitness for use | Do the metadata include all concepts, definitions and descriptions of all of the variables? | yes |
| 66 | Data fitness for use | Do the metadata include descriptions of methods, procedures and quality assurance practices followed during production of the data? | yes |
| 67 | Data fitness for use | Are the metadata accurate, complete, up to date, and free of contradictions? | yes |
| 68 | Data fitness for use | Are accuracy indicators provided for all of the measured variables? | yes |
| 69 | Data fitness for use | Are there matching variables such as age, sex, address, industry, occupation? | yes |
| 70 | Data fitness for use | Is a description available online of any exceptions or limitations in these data? | yes |
| 71 | Data fitness for use | Do the data meet domain specific standards or requirements? | yes |
| 72 | Data fitness for use | Are the data fit-for-use? | yes |
| 73 | Computer code | Is there a brief explanatory comment at the start of the code? | yes |
| 74 | Computer code | Has the code been decomposed into functions? | yes |
| 75 | Computer code | Has duplication been eliminated? | yes |
| 76 | Computer code | Does the code include well researched libraries or packages to perform needed tasks? | yes |
| 77 | Computer code | Have you tested the libraries or packages before relying on them? | yes |
| 78 | Computer code | Do the functions and variables have meaningful names? | yes |
| 79 | Computer code | Have dependencies and requirements been made explicit? | yes |
| 80 | Computer code | Have you avoided using comment/uncomment for sections of code to control the program's behavior? | yes |
| 81 | Computer code | Have you provided a simple example or test dataset? | yes |
| 82 | Computer code | Has the code been submitted to a reputable DOI-issuing repository? | yes |
| 83 | Computer code | Is an overview of the project available online? | yes |
| 84 | Computer code | Is a shared "to-do" list for the project available online? | yes |
| 85 | Computer code | Is a description of the communication strategy available online? | yes |
| 86 | Computer code | Is there an explicit license? | yes |
| 87 | Computer code | Is the project citable? | yes |
| 88 | Project organization | Is each project in its own directory which is named after the project? | yes |
| 89 | Project organization | Are text documents associated with the project in a documents directory? | yes |
| 90 | Project organization | Are the raw data and metadata in a data directory? | yes |
| 91 | Project organization | Are the files generated during cleanup and analysis in a results directory? | yes |
| 92 | Project organization | Is the project source code in a ‘source’ directory? | yes |
| 93 | Project organization | Are external scripts or compiled programs in a bin directory? | yes |
| 94 | Project organization | Do all filenames reflect their content or function? | yes |
| 95 | Keeping track of changes | Is (almost) everything created by a human being backed up as soon as it is created? | yes |
| 96 | Keeping track of changes | Are changes kept small? | yes |
| 97 | Keeping track of changes | Are changes shared frequently? | yes |
| 98 | Keeping track of changes | Is a checklist created, maintained, and used for saving and sharing changes to the project? | yes |
| 99 | Keeping track of changes | Is each project stored in a folder that is mirrored off the researcher's working machine? | yes |
| 100 | Keeping track of changes | Is there a file called CHANGELOG.txt in the project's docs subfolder? | yes |
| 101 | Keeping track of changes | Is the entire project copied whenever a significant change has been made? | yes |
| 102 | Keeping track of changes | Is a version control system used? | yes |
| 103 | Keeping track of changes | Are changes conveyed to all users in a timely fashion? | yes |
| 104 | Reproducibility | Are the data the result of a reproducible workflow? | yes |
| 105 | Reproducibility | Are all methods documented in detail such that a 3rd party could reproduce the workflow and obtain the same results without needing to consult with the data provider? | yes |
| 106 | Reproducibility | Given the data and information provided, are the data and the limitations of the data completely understandable by a 3rd party without needing to consult with the data provider? | yes |
| 107 | Manuscripts | Are manuscripts written using reference management software? | yes |
| 108 | Manuscripts | Are manuscripts written in a plain text format? | yes |
| 109 | Manuscripts | Are manuscripts deposited in a pre-print repository? | yes |
| 110 | Manuscripts | Are manuscripts submitted to an open source, peer reviewed journal? | yes |
| 111 | Manuscripts | Do manuscripts identify individual authors and co-authors? | yes |
| 112 | Manuscripts | Are manuscripts version controlled? | yes |

1. Acronyms

API application programming interface

CCD Continuity of Care Document

CCR Continuity of Care Record

DBMS Database Management System

DIY Do-It-Yourself

ELT Extract, Load, Transform

ERP Enterprise Resource Planning

ETL Extract, Transform, Load

FAIR Find, Access, make it Interoperable, and Reusable

FHIR Fast Healthcare Interoperability Resources

HIT Healthcare Info Tech

IaaS Infrastructure as a Service

iPaaS integration Platform as a Service

IT information technology

ITL Information Technology Laboratory at NIST

MARS Multivariate Adaptive Regression Splines

MGI McKinsey Global Institute

NBDIF NIST Big Data Interoperability Framework

NBD-PWG NIST Big Data Public Working Group

NBDRA NIST Big Data Reference Architecture

NIST National Institute of Standards and Technology

OS operating system

R&D research and development

ROI return on investment

1. Bibliography

[1] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 1, Definitions,” 2015.

[2] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies,” 2015.

[3] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements,” 2015.

[4] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 4, Security and Privacy,” 2015.

[5] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 5, Architectures White Paper Survey,” *Spec. Publ. (NIST SP) - 1500-5*, vol. 5, 2015.

[6] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 6, Reference Architecture,” 2015.

[7] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap,” 2015.

[8] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 8, Reference Architecture Interface,” *Spec. Publ. (NIST SP) - 1500-9*, vol. 8, 2017.

[9] T. White House Office of Science and Technology Policy, “Big Data is a Big Deal,” *OSTP Blog*, 2012. [Online]. Available: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal. [Accessed: 21-Feb-2014].

[10] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 9, Adoption and Modernization,” *Spec. Publ. (NIST SP) - 1500-10*, vol. 9, 2017.

[11] Dresner Advisory Services, “2017 Big Data Analytics Market Study,” 2017.

[12] A. Naimat, *The Big Data Market: A Data-Driven Analysis of Companies Using Hadoop, Spark, and Data Science*. O’Reilly, 2016.

[13] C. Ross, “The hype and the hope: The road to big data adoption in Asia-Pacific,” *The Economist Intelligence Unit Perspectives*, 2013. [Online]. Available: https://www.eiuperspectives.economist.com/technology-innovation/hype-and-hope-road-big-data-adoption-asia-pacific.

[14] DataRPM, “Big Data Trends for 2015 Infographic,” *Big Data Analytics News*, 2015. [Online]. Available: http://bigdataanalyticsnews.com/big-data-trends-2015-infographic/.

[15] McKinsey & Company, “Big data: The next frontier for innovation, competition, and productivity,” *McKinsey Glob. Inst.*, no. June, p. 156, 2011.

[16] AIIM, “Search and Discovery – Exploiting Knowledge, Minimizing Risk,” 2014.

[17] IDC, “Using Big Data + Analytics to Drive Business Transformation,” 2015.

[18] MemSQL, “The Lambda Architecture Simplified,” Apr. 2016.

[19] B. Hopkins, L. Owens, and J. Keenan, “The Patterns Of Big Data: A Data Management Playbook Toolkit,” 2013.

[20] D. Neef, *Digital Exhaust: What Everyone Should Know About Big Data, Digitization and Digitally Driven Innovation*. O’Reilly, Safari, 2014.

[21] D. Mysore, S. Khupat, and S. Jain, “How to know if a big data solution is right for your organization,” *IBM developerWorks, Big data architecture and patterns, Part 2*, 2013. [Online]. Available: https://www.ibm.com/developerworks/library/bd-archpatterns2/index.html.

[22] J. Taylor, “Analytics Capability Landscape,” 2015.

[23] Broman KW, Woo KH (2017). Data organization in spreadsheets. The American Statistician, 72(1): Special Issue on Data Science <https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989>

[24] Kitzes J (2016). Reproducible workflows. <http://datasci.kitzes.com/lessons/python/reproducible_workflow.html>

[25] Wickham H (2014). Tidy data. Journal of Statistical Software, 59(40), 1-23. <https://www.jstatsoft.org/article/view/v059i10>

[26] Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Tea TK (2017). Good enough practices in scientific computing. PLOS Computational Biology | <https://doi.org/10.1371/journal.pcbi.1005510http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001745>

[n] datameer. https://vizworld.com/2014/12/big-data-a-competitive-weapon-for-the-enterprise-infographic/

1. Related Reading

Albert, C, Rosemergy, S. A framework for evaluating common operating environments: piloting, lessons learned, and opportunities. 2011. Pdf.

Why discovery tools leave gaps. <https://www.betterbuys.com/bi/data-discovery-tools-enterprises/#comments>

Anatomy of a Hadoop failure. <https://www.datanami.com/2017/03/17/anatomy-hadoop-project-failure/>

1. “Contributors” are members of the NIST Big Data Public Working Group who dedicated great effort to prepare, and/or gave substantial time on a regular basis to research and development in support of this document. [↑](#footnote-ref-2)
2. A pivotal turning point is the release of all data in human readable and machine-readable format. For example, CSV files in tabular form can be understood by humans and can be read by statistical or database software (other than Excel, Word, or Acrobat) without the need to write extensive computer code to extract information and put it in a machine useable form. [↑](#footnote-ref-3)