NIST Special Publication 1500-10r1

NIST Big Data Interoperability Framework:

Volume 9, Adoption and Modernization

Version 3

NIST Big Data Public Working Group
Definitions and Taxonomies Subgroup

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-10r1

National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Special Publication 1500-10r1

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

Version 3

NIST Big Data Public Working Group
Definitions and Taxonomies Subgroup
Information Technology Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-10r1

October 2019

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

National Institute of Standards and Technology (NIST) Special Publication 1500-10r1
76 pages (October 2019)

NIST Special Publication series 1500 is intended to capture external perspectives related to NIST standards, measurement, and testing-related efforts. These external perspectives can come from industry, academia, government, and others. These reports are intended to document external perspectives and do not represent official NIST positions.

Copyrights and Permissions

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

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; considerations for implementation and modernization of Big Data systems; and, Big Data readiness.

Keywords

Technology adoption; barriers to adoption; market maturity; project maturity; organizational maturity; implementation; system modernization, digital transformation, Big Data readiness.

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 began during Stage 2 work and was not part of the Version 1 scope.
- ***Version 2***: Russell Reinsch (Center for Government Interoperability) and Wo Chang (NIST)
- ***Version 3***: Russell Reinsch (Center for Government Interoperability), Claire C. Austin (Department of the Environment, Canada), and Wo Chang (NIST)

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

NIST SP1500-10, Version 3 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] to this volume, during Version 2 and/or Version 3 activities, by the following NBD-PWG members:

Claire C. Austin
Department of the Environment
Government of Canada

David Boyd
InCadence Strategic Solutions

Frank Farance
Consultant

Geoffrey Fox
Indiana University

Nancy Grady
SAIC

Zane Harvey
QuantumS3

Haiping Luo
Department of the Treasury
Government of the United States

Gary Mazzaferro
alloyCloud, Inc.

Russell Reinsch
Center for Government Interoperability

Arnab Roy
Fujitsu

Mark Underwood
Krypton Brothers
Synchrony Financial

Gregor von Laszewski
Indiana University

Timothy Zimmerlin
Consultant

[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.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

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 *NIST Big Data Interoperability Framework* (NBDIF) was released in three versions, which correspond to the three stages of the NBD-PWG work. Version 3 (current version) of the NBDIF volumes resulted from Stage 3 work with major emphasis on the validation of the NBDRA Interfaces and content enhancement. Stage 3 work built upon the foundation created during Stage 2 and Stage 1. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data. The three stages (in reverse order) aim to achieve the following with respect to the NIST Big Data Reference Architecture (NBDRA).

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

The *NBDIF* consists of nine volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The nine 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]
- Volume 8, Reference Architecture Interfaces [8]
- Volume 9, Adoption and Modernization (this volume)

During Stage 1, Volumes 1 through 7 were conceptualized, organized, and written. The finalized Version 1 documents can be downloaded from the V1.0 Final Version page of the NBD-PWG website (https://bigdatawg.nist.gov/V1_output_docs.php).

During Stage 2, the NBD-PWG developed Version 2 of the NBDIF Version 1 volumes, with the exception of Volume 5, which contained the completed architecture survey work that was used to inform Stage 1 work of the NBD-PWG. The goals of Stage 2 were 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 need for NBDIF Volume 8 and NBDIF Volume 9 was identified and the two new volumes were created. Version 2 of the NBDIF volumes, resulting from Stage 2 work, can be downloaded from the V2.0 Final Version page of the NBD-PWG website (https://bigdatawg.nist.gov/V2_output_docs.php).

1 INTRODUCTION

1.1 BACKGROUND

There is broad agreement among commercial, academic, and government leaders about the 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) was released in three versions, which correspond to the three stages of the NBD-PWG work. Version 3 (current version) of the NBDIF volumes resulted from Stage 3 work with major emphasis on the validation of the NBDRA Interfaces and content enhancement. Stage 3 work built upon the foundation created during Stage 2 and Stage 1. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data. The three stages (in reverse order) aim to achieve the following with respect to the NIST Big Data Reference Architecture (NBDRA).

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

The *NBDIF* consists of nine volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The nine 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]
- Volume 8, Reference Architecture Interfaces [8]
- Volume 9, Adoption and Modernization (this volume)

During Stage 1, Volumes 1 through 7 were conceptualized, organized, and written. The finalized Version 1 documents can be downloaded from the V1.0 Final Version page of the NBD-PWG website (https://bigdatawg.nist.gov/V1_output_docs.php).

During Stage 2, the NBD-PWG developed Version 2 of the NBDIF Version 1 volumes, with the exception of Volume 5, which contained the completed architecture survey work that was used to inform Stage 1 work of the NBD-PWG. The goals of Stage 2 were 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 need for NBDIF Volume 8 and NBDIF Volume 9 was identified and the two new volumes were created. Version 2 of the NBDIF volumes, resulting from Stage 2 work, can be downloaded from the V2.0 Final Version page of the NBD-PWG website (https://bigdatawg.nist.gov/V2_output_docs.php).

The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

1.2 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.

1.3 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.

1.4 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 (technology, product, 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.

While each NBDIF volume was created with a specific focus within Big Data, all volumes are interconnected. During the creation of the volumes, information from some volumes was used as input for other volumes. Broad topics (e.g., definition, architecture) may be discussed in several volumes with each discussion circumscribed by the volume's particular focus. Arrows shown in Figure 1 indicate the main flow of information input and/or output from the volumes. Volumes 2, 3, and 5 (blue circles) are essentially standalone documents that provide output to other volumes (e.g., to Volume 6). These volumes contain the initial situational awareness research. During the creation of Volumes 4, 7, 8, and 9 (green circles), input from other volumes was used. The development of these volumes took into account work on the other volumes. Volumes 1 and 6 (red circles) were developed using the initial situational awareness research and continued to be modified based on work in other volumes. The information from these volumes was also used as input to the volumes in the green circles.

Figure 1: NBDIF Documents Navigation Diagram Provides Content Flow Between Volumes

2 ADOPTION AND BARRIERS

2.1 EXPLORING BIG DATA ADOPTION

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

2.1.1 ADOPTION BY USE CASE

Adoption of Big Data analysis technologies has been recently estimated to be 53 percent [10]. Simple ways of looking at the Big Data environment are from the perspectives of use cases, both by organizational department (i.e., 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	Adoption metric = Priority?
Financial services	Data warehouse adoption	83
Healthcare	Data warehouse adoption	80
IT	Customer / social network analysis	75
Telecommunications	Data warehouse adoption	74
Education	Customer / social network analysis	70

Departmentally, IT departments, business intelligence (BI) 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 [10] cites financial services and telecommunications industries as the earliest adopters, with education lagging. In a 2016 report by Aman Naimat [11], 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 [11].

2.1.2 ADOPTION BY INDUSTRY

Adoption of Big Data systems has not been uniform across all industries or sectors. A 2014 report [12] 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 affecting 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.

“The 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 [13].”

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 [13]. 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 [13]. 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%

2.1.3 LEVELS OF SPENDING

One indicator of maturity is financial investment into research and development, so in some cases, viewing the landscape from the perspective of where money has been spent, can shed some light into level of adoption. Table 3 shows a sample breakdown of Big Data spending by industry across the Asia-Pacific region in 2016 [14] 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) [15]	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%

2.2 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 technologyassisted review, yet the majority do not have a mature search function and only 18% have federated search capability [16].

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.

2.2.1 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 barriers listed in Table 4 were compiled from multiple surveys, as indicated in the column headers. Each survey contained both similar and distinct questions as compared to other surveys in the group. 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. The blank cells do not correspond to zero percent.

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 budget; 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 Information Technology (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%

2.2.2 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.

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						
• Data from distributed file based computing systems						
Specific to distributed file based computing systems						
• 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 2.2.1, 2.2.2, and elsewhere. distributed file based computing systems

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-existent or inconsistent data governance • Lack of vision • Fragmented datasets • Multiple "copies" of the same dataset that don't match • Disparate data from different sources • Data "silos" • Lack of Findable, Accessible, Interoperable, and Reusable (FAIR), 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	• Merging data sources • Transferring data from source to analytics environment • Blending internal and external data • Inconsistent metadata management • Inconsistent metadata standards • Inconsistent data standards
Data Access	• Privacy regulations and confidentiality requirements • Sensitive data • Data access restrictions	• Concerns about liabilities and systems security
Skill and Expertise	• Lack of people with the ability to handle the complexity of software and analysis • Lack of people with 'deep analytical' training [b] • Lack of data-savvy managers [c] • 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 • 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 and pilot testing	• Integration with existing infrastructure • Integration with existing workflows
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 human and technical resources	

Notes:
[a] Adapted from Big data: The next frontier for innovation, competition, and productivity [13].
[b] People with advanced training in statistics and/or machine learning and who conduct data analysis.

[c] People 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 3.1, 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 enterprise resource planning (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.

3 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.

3.1 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.

3.1.1 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 the following:

- 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 in silos; and
- Analytics could be inconsistent across departments.

3.1.2 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. Extract, transform, load (ETL) / extract, load, transform (ELT) 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 the following:

- Information may be in silos;
- 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.

3.1.3 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 the following:

- 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.

3.1.4 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 the following:

- 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. These departments might be Customer Service, Finance, Marketing, Sales, or Research. 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 2), 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.

3.2 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.

3.2.1 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 2 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.

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. • Reproducible science and preservation of data lineage.	• 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. • In a science-based organization, develops and implements an Open Science policy. • Initiates a master data management (MDM) program. • Appoints a system leader from upper management.
	DIFFICULT TRANSITION	
LEVEL 2 **Division Adoption**	• Information may still be in silos. • 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 Functioning**	• Information is in silos. • 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 2: Evolution of Big Data Systems as a Function of Project and Organizational Data Governance Maturity

Klievink et al. [18] evaluated the ability of public sector organizations to use Big Data on the basis of organizational maturity, organizational capabilities, and organizational alignment. Increased organizational maturity was observed where there was more structural collaboration between organizations. Organizational capabilities for Big Data use were described in terms of: internal attitude, external attitude, legal compliance, IT resources, data science expertise, IT governance, and data governance. The last three (i.e., data science expertise, IT governance, and data governance) were found to have the greatest impact on improvements in organizational capability. Organizational alignment (i.e., whether or not Big Data applications are suited for the organization in question) was found to be vital for the success of Big Data. In addition, when evaluating organizational alignment, it was found that the intensity of data use was a determinant of the readiness for Big Data. Paradoxically, intensity of data collection was not necessarily associated with data quality or with readiness for Big Data. This is an important observation to keep in mind in the cases where the intensity of data collection is high, but the intensity of data use is low because the primary data users are found elsewhere within the organization or externally to the organization. In some cases, it may well be that the greatest barrier to Big Data is not organizational maturity or capability, but alignment with the data provider’s priorities.

3.3 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. 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 [11].
- 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 [19], 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 [13]. Various agencies are considering mandating the management of curation and metadata activities in funded research projects. However, metadata standards are frequently ranked as a significant technical issue. While agreeing on a local taxonomy snapshot is a major challenge for an organization, managing the difficulties of taxonomy dynamics (which are organizational issues) presents an even more challenging barrier.

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 database management system (DBMS), 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.

3.4 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 [20], 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 2018, in-memory, private cloud infrastructure, and complex event processing reached the mainstream. Modern data science and machine learning are slightly behind but moving at a very fast pace to maturity.

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 integration Platform as a Service (iPaaS) platforms. Use of application programming interfaces (API) for enabling micro services 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.

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.

561 ***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

562

4 MODERNIZATION AND IMPLEMENTATION

4.1 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. Beginning with transition plans, modernization projects often follow some method for portfolio road mapping. One such method is a technology brick approach comprising a strategy and a roadmap [21]. Brick structures classify products applying one or more ways to describe lifecycles, such as emerging, mainstream, and retirement. Within the methodology, it is common to map out the implementation timelines for each technology on a chart. Additional organizations and groups are exploring methodologies, processes, and frameworks that facilitate Big Data projects [22].

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

- Augmentation: 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: 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.

Hybrid parallel systems: Hybrid systems is a modular approach towards modernization where new Big Data capabilities may Replace and Augment existing systems. For example, organizations can use the cloud for storage but develop their own applications. One disadvantage of this route is the high cost of moving data to the cloud. Developing standards for hybrid implementations should accelerate the adoption and interoperability of analytics applications.

When considering pathways, the potential advantages and disadvantages should be examined. While the full list of advantages and disadvantages will be project-specific, Tables 6 and 7 provide a high-level comparison.

Table 8 provides a high-level list of advantages and disadvantages of the augmentation pathway, while Table 9 provides a high-level list of advantages and disadvantages of the replacement pathway.

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

Augmentation Type	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

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

Replacement Type	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

4.2 IMPLEMENTATION

In the earliest stages of planning, effectiveness of the plan is dependent on a clear understanding of new technologies. When evaluating technologies, it is prudent to make sure that the solution is being evaluated against the organization's actual use case, not something else that the vendor is promoting. Evaluating solutions for a single application is easy compared to evaluating solutions to fill broader use cases. The solutions for broad use cases are usually platforms, which are difficult to evaluate without implementing a proof of concept or pilot.

When ready to look at technology, the proper starting point is to ensure understand what data is involved and evaluate options from the standpoint of what capabilities are needed to work with that particular data. Some organizations may not actually have a Big Data use case; most use cases in 2018 are still BI, consisting of mainly transaction processing, and index-oriented queries on structured and trusted data. Big Data use cases are notoriously unstructured, with data that are not vetted, and not with adequate quality levels or compatible with standards which simplify integration.

Once a system augmentation or replacement path has been selected, a method of implementation can be chosen. When planning Big Data system modernization projects, organizations often find themselves at a second fork in the road decision point. Figure 3 diagrams this decision point, commonly referred to as the *build or buy* question.

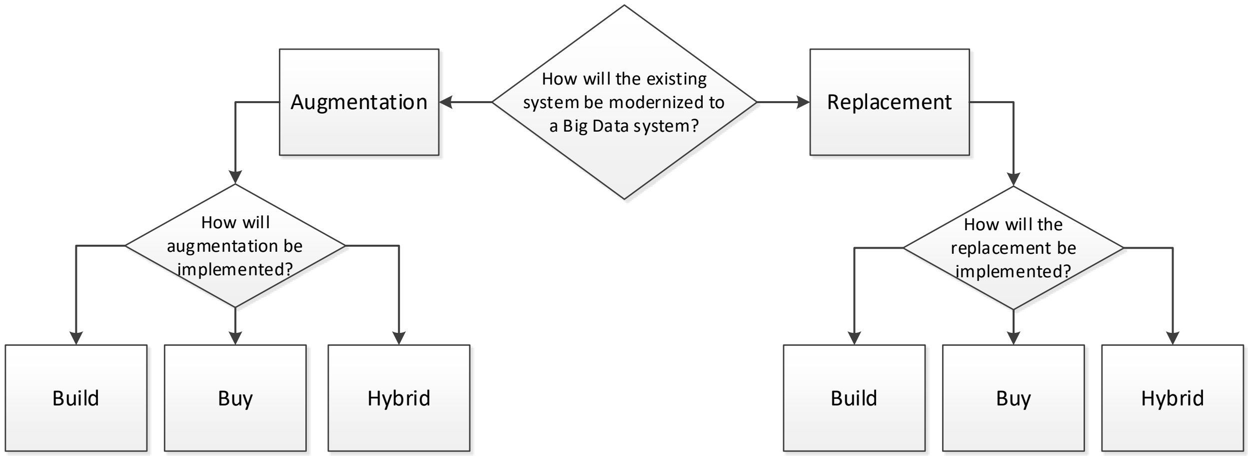

Figure 3: New System Implementation

In the build versus buy discussion, proponents from each side may disagree on the best approach.

4.2.1 BUY

On the one side, are the "buy" proponents who prefer purchasing commercial off the shelf (COTS) products and will articulate the benefits organizations realize when they focus on their core business and reduce IT project distractions; and also that custom or open source systems can result in a form of lock-in leverage for the developers, as the system is ultimately only understood by the key team member(s) who built it. Proponents of COTS typically also argue that COTS systems have a much higher success rate. The alternative to the pure buy scenario is for the organization to 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).

4.2.2 BUILD

On the other side, the "build" proponents prefer the benefits of developing a system in house and will articulate advantages of custom coded systems. Advantages of this option are realized for organizations that have unique requirements, as opposed to COTS systems which have been found to often be one size fits all, which can simultaneously fall short in some areas and be overkill in others. The build route can be a fit for organizations having a skilled IT department. Custom, "good enough" capabilities can have lower total cost of ownership (TCO). The downside of the build option is that these systems are tough to build, ergo risky. 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 Structured Query Language (SQL) layers for distributed storage, or construction of interfaces for 'real-time' analysis.

In the build, or do-it-yourself (DIY) scenario, the organization may modify their existing system, or build an entirely new system separate of the existing system. If the DIY implementation is implemented 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.

Developing an open source solution: part of the build philosophy, is the option of developing an open source solution. Proponents point out that full stack open source systems are more flexible than COTS; and secondly, are less expensive than COTS. This situation can hold true, however it can also be entirely false. While the open source technology itself may initially be very low cost or free, the cost of human resources required to build these systems are much higher, potentially causing the final TCO to be higher. Note also that all open source licenses are not the same. If the organization does have experience developing systems but not with open source technologies, they have the option to build using open source by partnering.

4.2.3 PARTNERING WITH THIRD PARTY SYSTEM INTEGRATORS

A third, perhaps less talked about option is partnership with a third party, where the third party provides outsourced development or integration services. Partnering may be the preferred implementation option. A 2017 Digital Banking Report on artificial intelligence (AI) implementations indicates that less than 10% of organizations plan to build a solution for any of the seven use cases surveyed for the report. Two to three times more organizations plan to purchase a commercial solution. But far and away the highest percentage of organizations plan to partner with an industry provider to implement an AI solution, in some cases over 50% of the respondents making this declaration [23].

4.2.4 PROJECT ISSUES

Certain challenges will persist with any of the implementation routes whether it be build / DIY; buying or renting new systems; or going with hybrid parallel systems. For example, data cleaning and systems plumbing are persistent hurdles no matter which type of project is undertaken [24], [25]. 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. Lower layer components of the system must be considered as equally important as (if not more important than) the analysis or analytics functions.

4.3 NEXT STEPS

Whether an organization decides to build custom, build open source, or buy COTS, many experts agree that internal culture may emerge as the largest obstacle to new program success [26], [27]. One best practice is to assess the organization's current state of readiness for such a project, before considering which technology to evaluate.

One task that organizations traditionally perform early in the decision process is to estimate the ROI for the project. If there is a clear but low ROI for the project, the organization may be a good candidate for evaluating leading and more established COTS solutions where benefits include reduced risk due to the maturity of available products.

If there is a clear, high ROI for the project, then an organization has two good options, depending on whether they already have a system development department. If they do not have a system development team, they are a good candidate to evaluate newer and more innovative COTS solutions where benefits often include responsive support departments. If there is no clear ROI, the organization must consider whether going forward with the project is an acceptable risk.

As previously noted, reducing costs is often the number one factor in enterprise wide modernization plans. The same holds true for departments and smaller business units. The cost factor is also driving adoption of newer implementation philosophies. Traditionally, organizations clarify requirements before implementing new software or technology projects. Unfortunately, the collection of requirements process often results in an output that is either not accurate or not valuable. A newer philosophy which prescribes an ‘implement first, ask questions later’ approach, eschews the traditional order of gathering requirements first; and prescribes a launch first mentality which recommends end users experiment with technology and pilot solutions first, and adopt those solutions if they work, with the belief that even if the pilot projects fail, the cost of failure is still lower than what would have been the total cost of a traditional project’s processes. The idea is that partly due to the availability of cloud-based technologies which can be implemented inexpensively, the ROI of a project is less important because the investment factor of the ROI is lower. The costs of experimenting with a new cloud-based solution can be very low in comparison to the time consuming and financially expensive processes of gathering requirements, and the implementation-first philosophy has had some success, although critics sound the shadow IT alarm bell. Shadow IT may be a problem for governance, but practices of experimenting with small, rapid tests of new technologies has gained so much traction that it is now commonplace [28].

5 SPECIFIC SOLUTION TECHNIQUES, DEPENDENT ON THE PROBLEM SPACE

Section 5 examines the industries and technologies related to Big Data and economic impacts by viewing them in context of the broader landscape.

Figure 4 is a simplified representation of some of the questions related to system capability that an organization may need to consider when planning their own system. Its purpose 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.

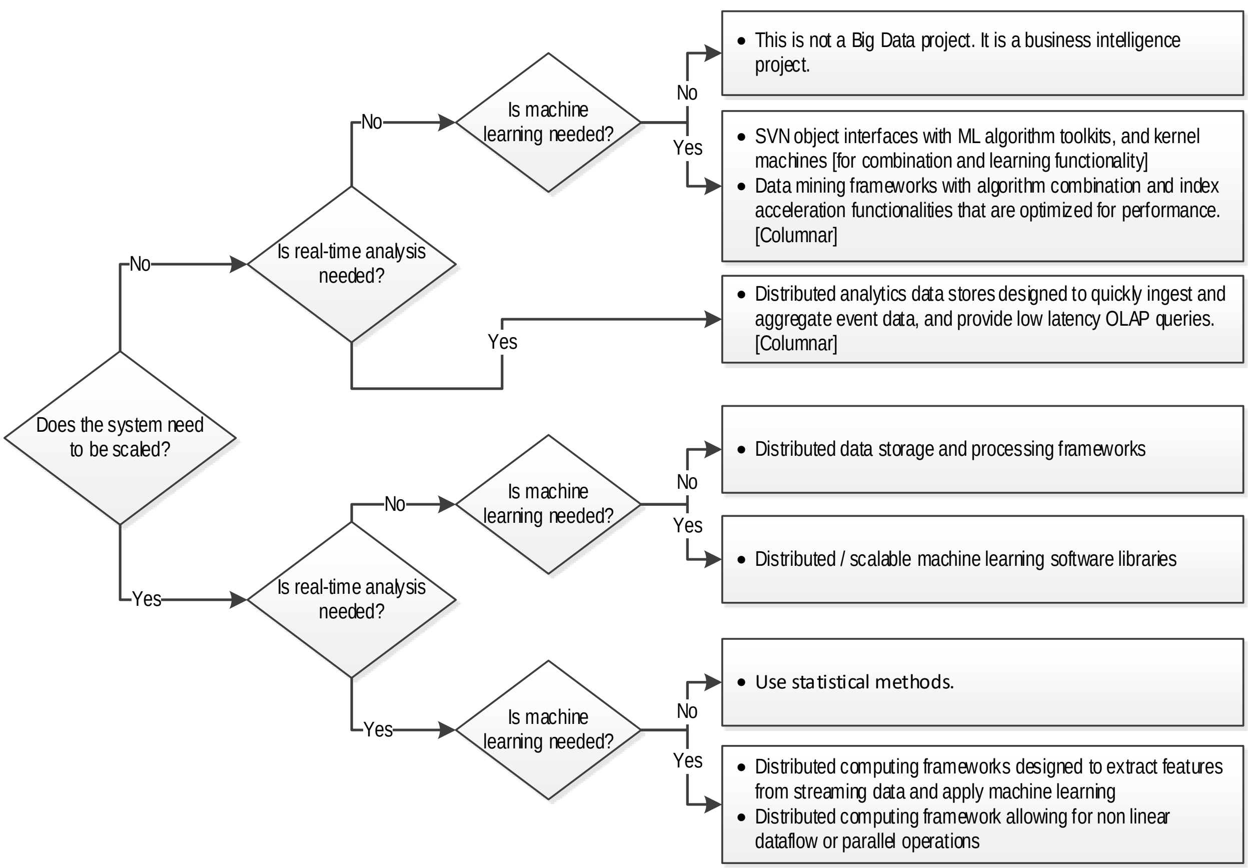

Figure 4: Requirement Decision Tree

After the scalability and latency requirements are identified as shown in Figure 4, the systems planning process will require continued consideration on whether machine learning is necessary. Figure 5, Figure 6, and Figure 7 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. There is no “correct” answer to the question of which algorithms to select. In fact, several tests should be run with different algorithms in order to validate various model results.

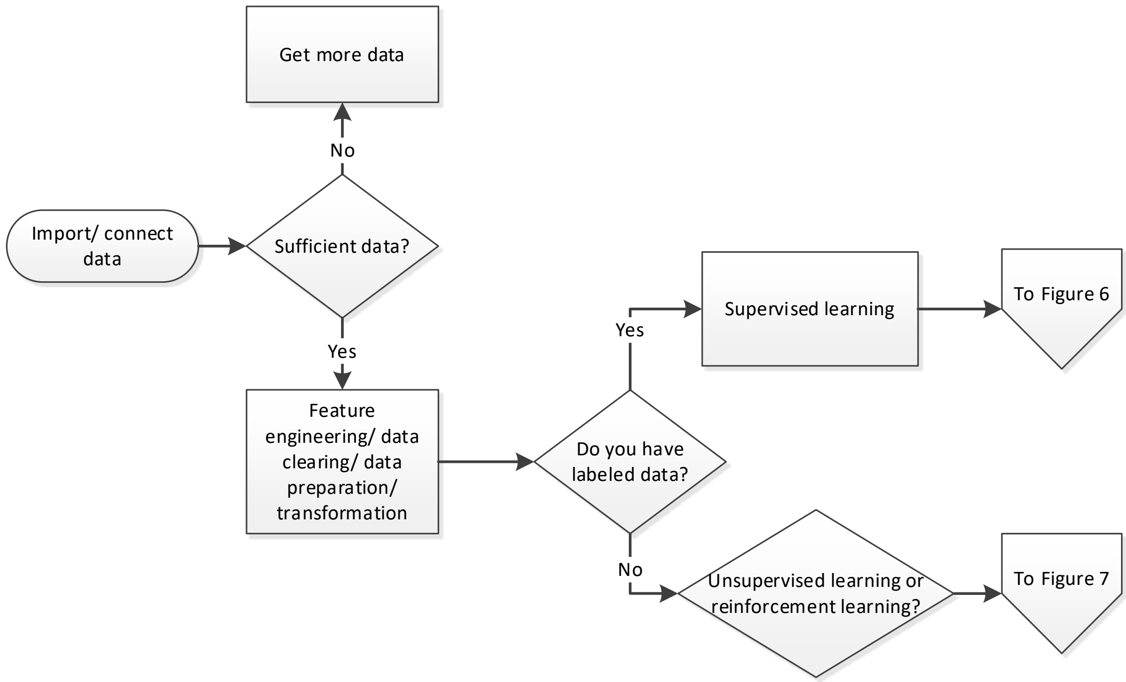

Figure 5: Machine Learning Algorithm Application Workflow

Figure 5 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 6 and Figure 7 expand on algorithm choices for each problem subclass. Table 10 and Table 11 continue from Figure 6 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 7.

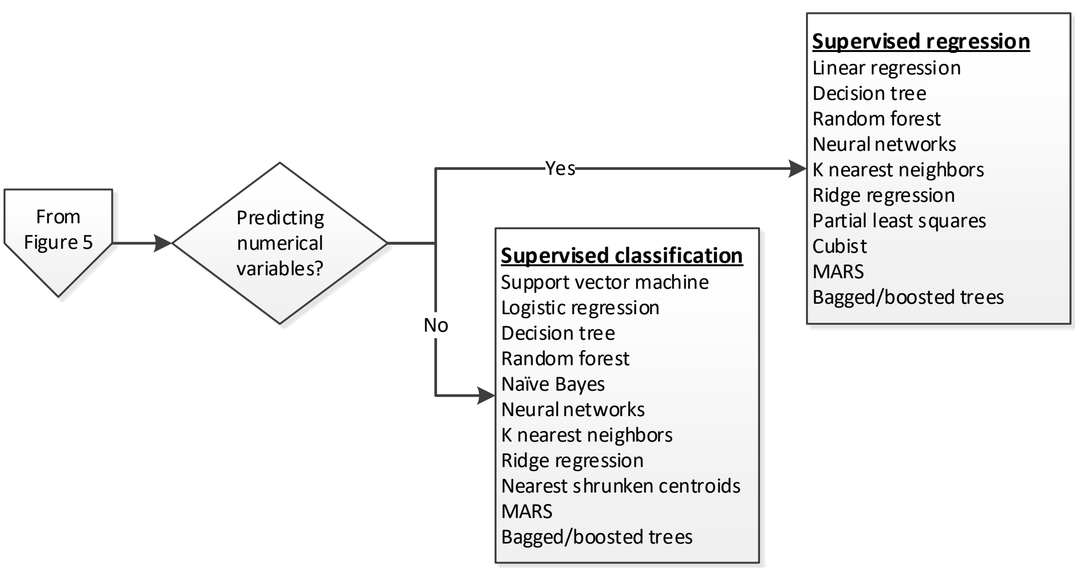

Figure 6: Supervised Machine Learning Algorithms

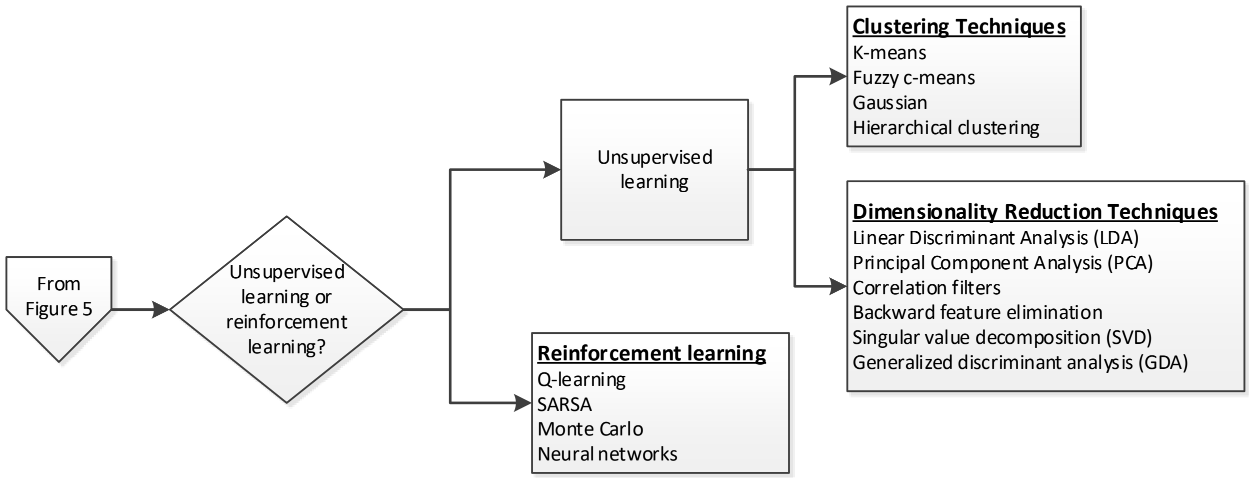

Figure 7: Unsupervised or Reinforcement Machine Learning Algorithms

Supervised 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, another 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.

6 BIG DATA READINESS

6.1 INTRODUCTION

Big Data[2] has the potential to answer questions, provide new insights previously inaccessible, and strengthen evidence-informed decision making. However, the harnessing of data into the Big Data net can also very easily overwhelm existing resources and approaches, keeping those answers and insights out of reach.

Big Data readiness begins at the source where data are first created and extends along a path through an organization to the outside world. Section 6 focuses on practical solutions to common problems experienced when integrating diverse datasets from disparate sources.

Business data, administrative data, health data, research data, etc. can potentially end up in the Big Data net. According to the Research Data Domain of the CASRAI dictionary [29], research data is defined as:

> *"Data that are used as primary sources to support technical or scientific enquiry, research, scholarship, or artistic activity, and that are used as evidence in the research process and/or are commonly accepted in the research community as necessary to validate research findings and results. All other digital and non-digital content have the potential of becoming research data. Research data may be experimental data, observational data, operational data, third party data, public sector data, monitoring data, processed data, or repurposed data".*

Many organizations hold important data assets for a variety of uses, including confidential and sensitive data, and may be faced with inconsistent data quality and multiple, sometimes uncontrolled, data flow pathways. This heterogeneity presents people at the working level and upper management alike with enormous challenges in developing and implementing solutions that will enable Big Data and Big Data Analytics.

The purpose of Section 6 is to contribute to the development of innovative thinking transferable to a wide range of organizations and domains with the goal of effecting changes needed to achieve Big Data. To support corporate governance and data management planning and strategies that may not yet be fully developed, Section 6 offers suggestions for a path to *Big Data readiness* based on Open Science, FAIR [30], [31], [32], [33], [34] data and an "*It's good enough*" approach [35]. FAIR data, endorsed by the G20 in 2016 means that the data are Findable, Accessible, Interoperable, and Reusable. "*It's good enough*" means doing what can be done now to make things work with the tools and the people currently in place. A *"Big Data readiness"* approach will support long-term planning and enable short-term solutions for data management in general. It will also support and enable the NBDRA, thereby enabling the data provider to feed data into the architecture at the blue arrow in the top left corner of Figure 8, which is discussed in detail in *NBDIF: Volume 6, Reference Architecture.*

---

[2] Big Data consists of extensive datasets—primarily in the characteristics of volume, variety, velocity, and/or variability—that require a scalable architecture for efficient storage, manipulation, and analysis [1], [51].

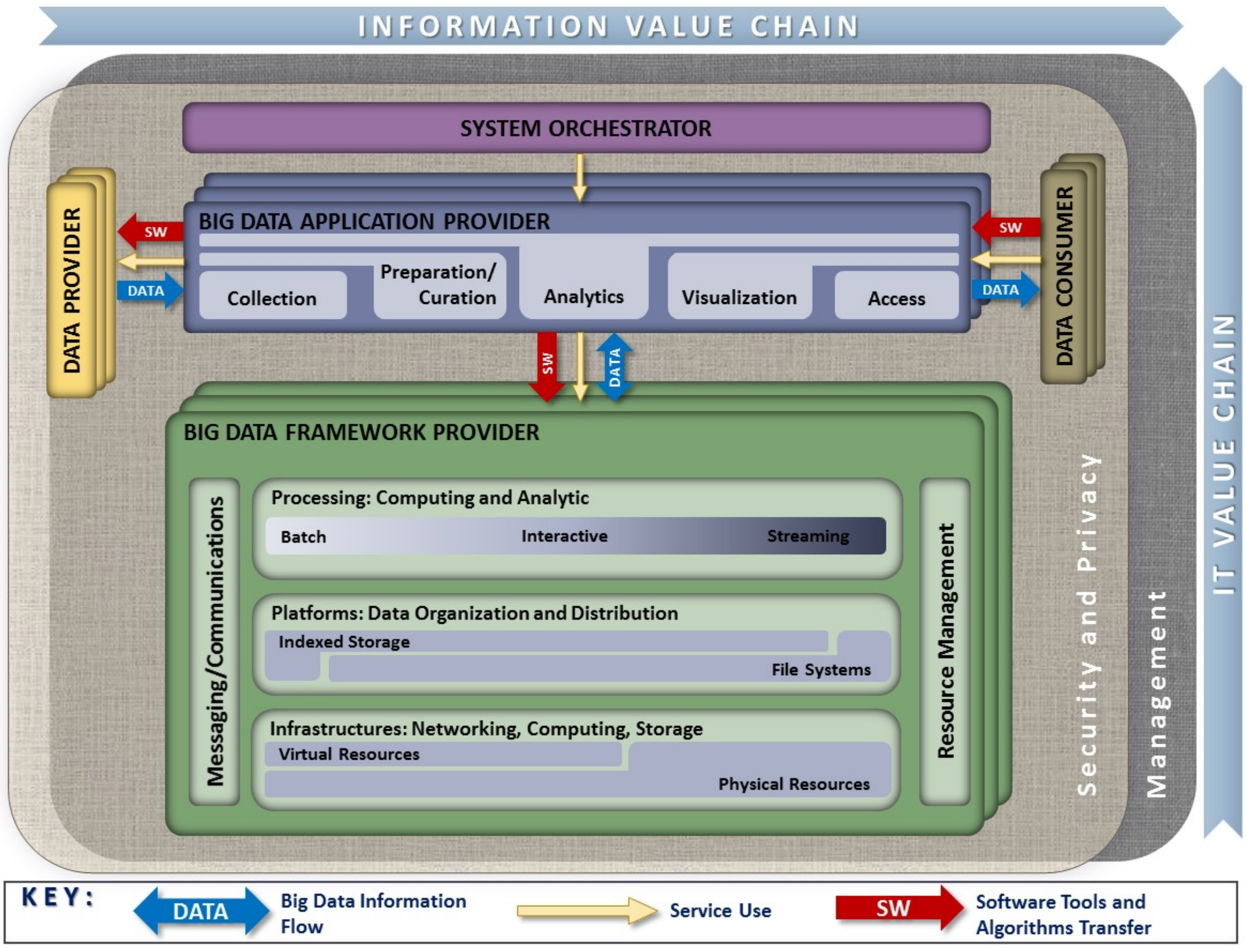

Figure 8: NIST Big DataRreference Architecture (NBDRA)

Section 6 proposes a generic strategy and tactical actions directed primarily at the working level that can be anticipated to have significantly positive short-term impacts without overwhelming workers, managers, or stakeholders, and to increase the chances of success of a Big Data project and implementation of a future data strategy. It will take some time to realize the business value of data strategies that may be under development in an organization and for some scenarios, an organization cannot afford to wait until implementation.

It is important that an organization identify the technical and nontechnical barriers to Big Data (Table 6). Contextualization of a path to Big Data readiness within a framework that describes the NBDRA. Big Data governance and metadata management is also important. However, Big Data transformation does not need to happen all at once; nor does the organization or its base need to wait for the development of a Big Data Framework, governance model, data policy, data strategy, master data management, or Open Science plan before taking action to help accelerate the implementation of Big Data. The actions proposed in Section 6 can be an effective first step for what can be done now in the present (taking into account current organizational maturity, capabilities, and data flow realities; Figure 2) to position an organization to meet opportunities provided by the Big Data revolution.

6.2 A BIG DATA PROBLEM SPACE

6.2.1 BARRIERS TO BIG DATA

6.2.1.1 Legacy Systems

Data management gaps at the working level and lack of data governance at the corporate level have been identified in organizations in private and public sectors dealing with decades old systems and procedures. Legacy systems do not only refer to the dark data buried in printed output, on compact disks (CD), in notebooks, on external hard drives, and on personal computers, etc. Legacy systems also refer to hardware and software that are still in use in the organization but are no longer supported by either the original vendor or by the organization's IT department, and to in-house computer code that may be poorly documented or developed without a well-structured approach. Additional challenges include more recent hardware and software that fail to meet the demands of Big Data and modern analytics, and people who experience challenges in adapting to new ways of doing things. Developing countries and new organizations may have a competitive advantage in that they have the opportunity to build state of the art systems from scratch relatively inexpensively, unencumbered by legacy systems or by other technical and non-technical barriers that are a function of an organization's overall readiness for Big Data measured by organizational maturity, organizational capability, and organizational alignment. See Section 3.

6.2.1.2 "Lock-in"

Not to be confused with vendor lock-in, which can also be a problem, organizations can be locked in to 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 sources and uses for the data collected. In addition, there is uncertainty regarding data accuracy, inconsistency in vocabulary, and confusion over the meaning of Big Data, data mining, and artificial intelligence. Meanwhile, many organizations are still struggling to emerge from a paper-based world governed in siloed organizations to a digitally literate and interconnected world. This is a very 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.

6.2.1.3 Culture Change

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" from traditional approaches are largely inaccessible to Big Data end users. This limits an organization's ability to use Big Data approaches for knowledge acquisition, innovation, and decision-making. Changes in thinking across organizations are needed to achieve a coordinated and harmonized system that is simple, effective and geared to meet organizational needs.

Organizations and various groups within them 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 where the user is unknown. A paradigm shift in thinking and culture is needed in many 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, in operations, research, and business lines. Targeted generic actions will help create the necessary conditions on the ground. Culture change will follow.

This bottom up change in thinking and culture must work hand-in-hand with top down culture change that also needs to happen if data are to become a strategic asset. Resources assigned to data life-cycle management must become a priority for program areas, supported appropriately by senior managers. Ultimately, sustainable culture change needs to work in both directions.

6.2.1.4 Degradation of Data Quality

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 uncontrolled data flows, data fragmentation, variation in data quality, and incomplete information concerning the data (Figure 9). Where this may be satisfactory within specific mandates, it is problematic for Open Science, reproducible research and Big Data.

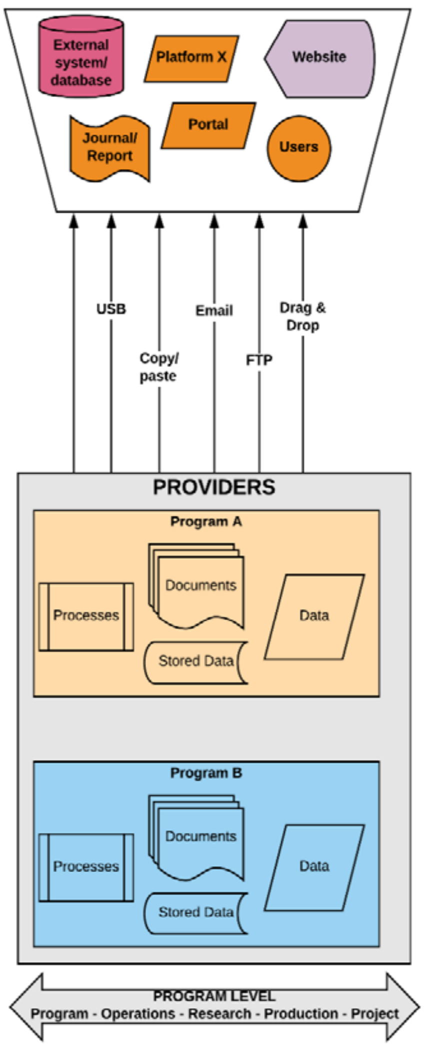

Figure 9: Uncontrolled Data Flow Pathways

Gartner estimates that poor data quality costs an average organization $13.5 million per year and that data governance problems are worsening [36]. There are seven levels of data quality:

1. Quality of the observations or measurements;
2. Quality of the recording of the observations and measurements;
3. Quality of the descriptors associated with the observations and measurements;
4. Quality of the information needed for an end user to completely understand the data and their limitations;
5. Organization of the observations/measurements/descriptors in a dataset or collection;
6. Compliance with recognized consensus Standards; and,
7. Quality of the management of the data and information, including sharing.

While there is a need for shared responsibilities across all six levels, the first two levels are primarily the realm of domain expertise, the fourth requires domain and information management expertise, and the last two are primarily data management expertise.

A very high-quality dataset produced under strict quality assurance/quality control (QA/QC) protocols can become fragmented in the absence of data governance encompassing the complete data life cycle (Figure 10). From the viewpoint of the data providers, they have produced extremely high-quality data. From the viewpoint of the data users, they see poor quality data that are difficult or impossible to use. 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.

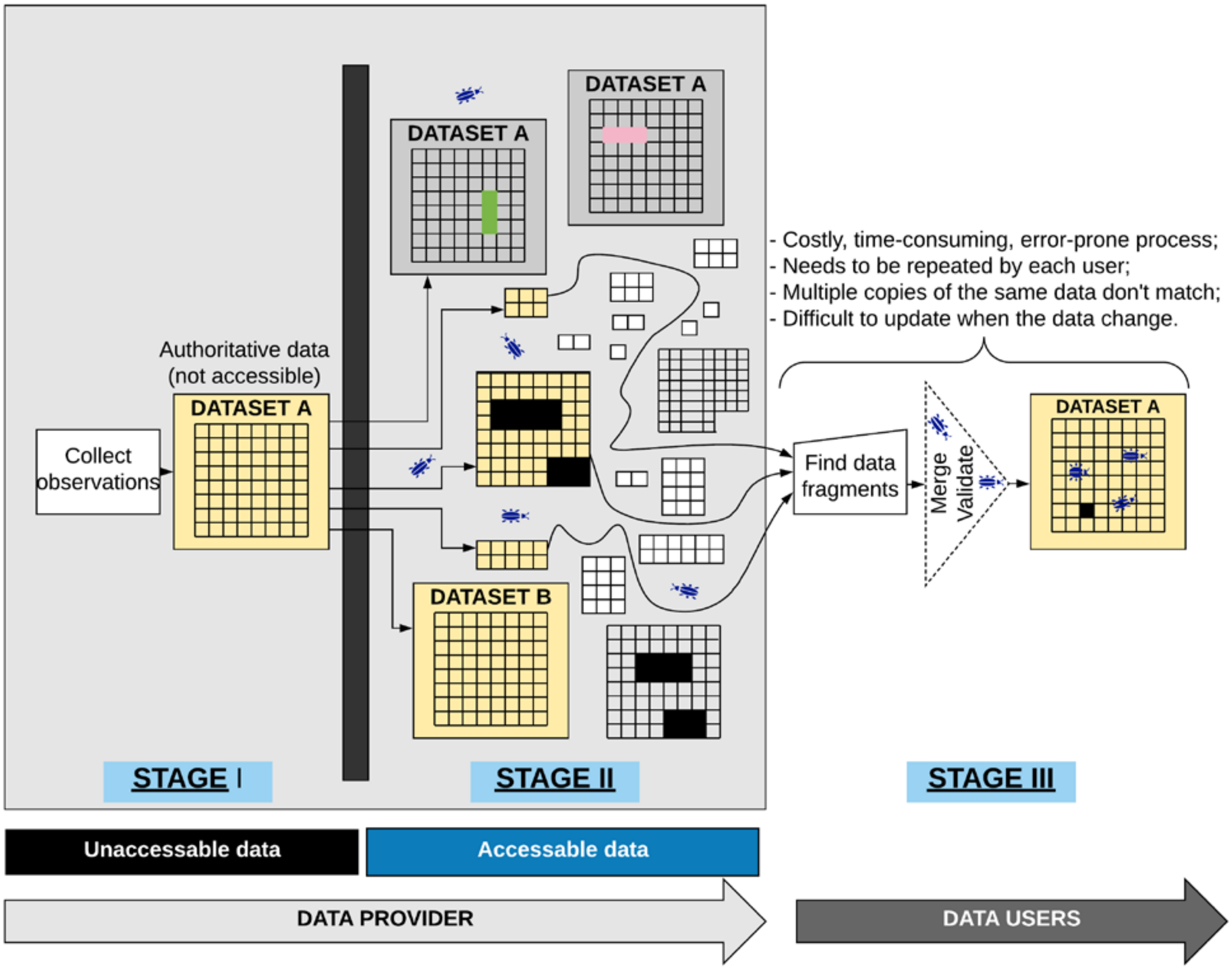

Figure 10: Dataset Fragmentation

The different stages of dataset fragmentation are as follows. Stage I: the data provider produces high quality observations and measurements that have undergone intensive QA/QC. Stage II: the data are published to various platforms and portals, during which data fragmentation and duplication may occur and data lineage lost. Stage III: the data user must find all of the data fragments and reassemble them into something resembling the original dataset in Stage I.

6.2.1.5 Merging Datasets from Diverse Sources

A commonly seen workflow is illustrated in Figure 11 where multiple datasets from different sources somehow have to be merged. In addition to the problem of dataset fragmentation and simply finding the data, there is confusion about other issues such as which one is the approved copy, lack of version control, absent or incomplete metadata, lack of common fields, variety in nomenclature and measurement units, and inconsistent data structures.

Before the analyst can use the data, there may be unavoidable manual work involved in collecting and cleaning each of the data streams before they can be used (Stage III of Figure 10), and in integrating these disparate data from diverse sources (Figure 11). All of these data would be lost to Big Data where reliance on manual processes is no longer possible, or an inordinate amount of time would need to be spent on data preparation.

6.2.1.6 Data Preparation

A major hurdle for the researcher or data scientist is data cleaning which can take up to 70% or more of the total time spent for the analysis [37], essentially performing tasks left undone when data providers release data that are not FAIR (Figure 10 and Figure 11). It takes enormous time, effort, and money to output small datasets to meet a variety of requests in Stage II of Figure 10, and an even greater amount of time, effort and money for an analyst to reassemble the data before they can be used (Figure 10: Stage III). Elimination of Stages II and III would eliminate the associated costs and wasted time, and result in more reliable analyses and stronger insights. Long-term data governance is the solution to the dataset, data flow, and metadata problems and to eliminating the hidden costs that result from them.

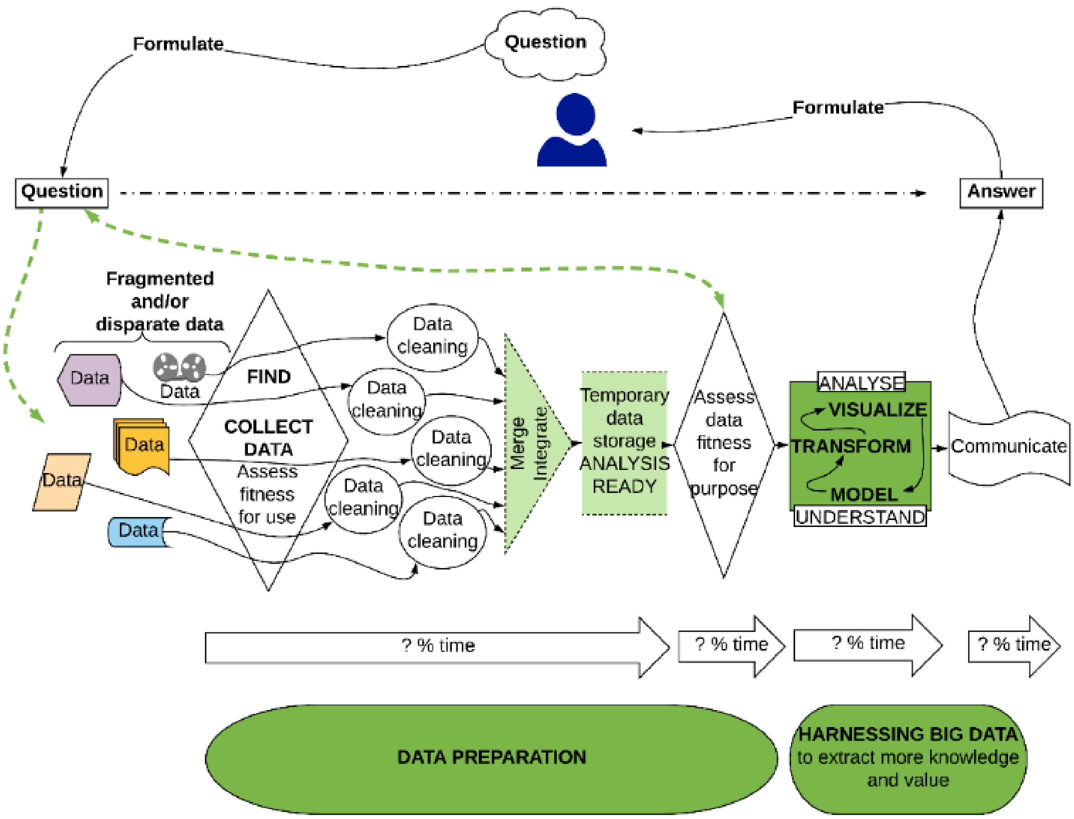

Figure 11: Integration of Data From Diverse Sources

If data providers published FAIR data that are analysis ready, data users would not need to spend 70-80% of their time on data preparation.

Short term targeted actions that address gaps in Data Governance and data management will improve the ability to integrate data from multiple sources and to reliably extract new knowledge and insights from large and complex collections of digital data. Adopting a "Big Data readiness" approach within an organization will help enable Big Data analytics, machine learning, and AI.

6.3 A BIG DATA SOLUTION SPACE

6.3.1 THE "BIG DATA READINESS" APPROACH

The respective roles of data providers and data users require clarification. Data providers in the field, laboratory, and other organizational levels need to recognize at the outset that there will be unknown data users and that it is an integral part of their job to prepare their data to a standard that meets the requirements of these unknown users. Data providers also need to accept that how the data will be used and for what purpose will remain unknown to them. It is not the role of the data provider to assess if their data are fit for the purpose envisaged by some unknown user. That is the responsibility of the data user. However, to implement Open Data and Big Data it must be part of the data provider's role to make sure that data transmitted from one person or group to the next throughout the data life cycle are FAIR and tidy (organized for ease of use).

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.

Best practices, standards, and training are key to data providers being able to prepare data appropriately. The organization must take on the responsibility of defining those practices and standards so that data can be integrated easily. A "*Big Data readiness*" approach should be included in organizational data strategies for short-term success in Big Data projects. For example, defining data quality and data standards strategies to support a Data Management Operational Plan could also include components of a "*Big Data readiness*" approach.

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 such as a Big Data interoperability framework and Reference Architecture that support Big Data and analytics.

6.3.2 DISRUPTING THE STATUS QUO

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 in implementation of "*Big Data readiness*" 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.

6.3.3 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.

6.3.4 DATA GOVERNANCE

Big Data will not improve data quality, solve data management problems, reduce the need for good quality, well-managed data, or obviate requirements for competent statistical analysis. Grappling with poor quality data (Figure 10 and Figure 11) is not the essence of what it means to "harness" Big Data. Harnessing Big Data refers to analysts and systems extracting more knowledge from existing data. Data governance that also includes *Big Data readiness* is a fundamental and essential piece of the solution to extensive data preparation time and eliminating hidden costs.

Figure 12 is a solution diagram for an organization. Data governance is the solution to extensive data preparation time and eliminating hidden costs.[3] 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 thus freed-up can then be used for the harnessing of Big Data in the continuum of reproducible science.

Good data governance and FAIR data will result in 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. Big Data can then reduce costs by reusing existing data instead of collecting more data unnecessarily. Big Data can also reduce costs by getting better answers more quickly.

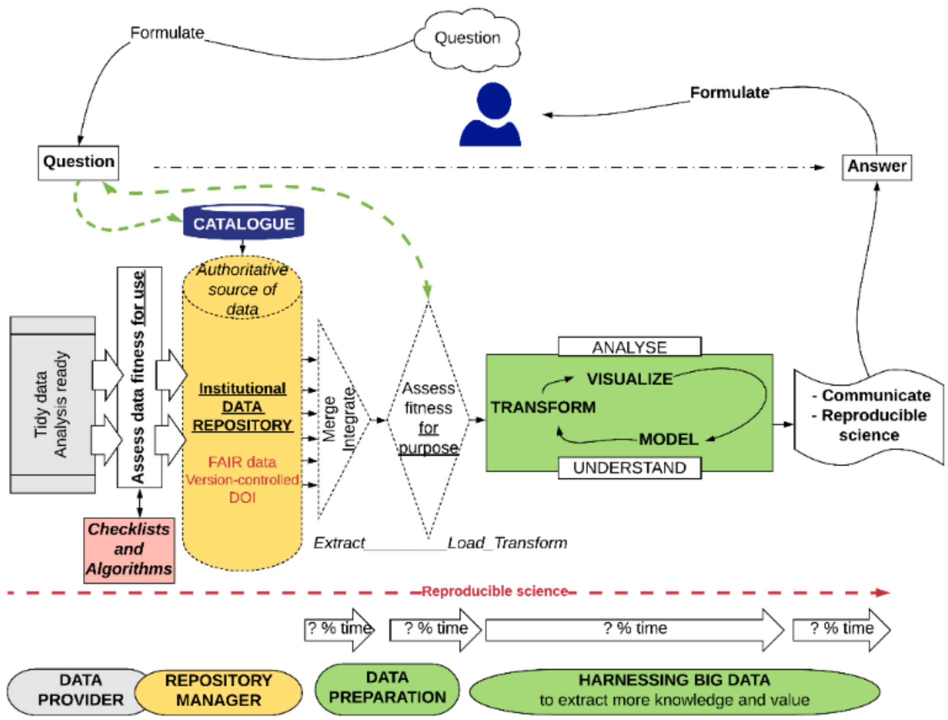

Figure 12: Improved Data Governance

[3] See also: *NBDIF: Volume 1, Definitions* (Section 7.2); *NBDIF: Volume 2, Big Data Taxonomies* (Section 2.2-B); and *NBDIF: Volume 4, Security and Privacy* (Sections 4.2.3, 4.3.2 and 8.5).

The data repository (yellow container) is the input/output control point and source of authoritative data. FAIR and analysis ready data on the left side of the diagram are released by the data providers. Semi-automated checklists implemented on the input side of the data repository are a critical component to ensure that the data are, in fact, FAIR.

As the organization matures, uncontrolled data flows (Figure 9) can be shut down and replaced by a data architecture that is able to provide an authoritative source of data for external systems, platforms, portals, data consumers, and can feed data into the NIST Big Data Reference Architecture (Figure 13). For organizations that have not yet achieved this, the use of data checklists can be an effective tactical action to accelerate the process (See Section 6.3.5).

While a stepwise move toward "*Big Data readiness*" and reproducible science means changing the way things are done with the tools currently in place, it also means adopting new tools and new competencies. Lowndes et al. have published a refreshingly candid account of their path to adoption and implementation of open data science tools and reproducible science in a complex environmental sciences framework [38]. Consult the EU funded Education for Data Intensive Science to Open New science frontiers (EDISON) for a comprehensive curriculum to train competent data scientists [39].

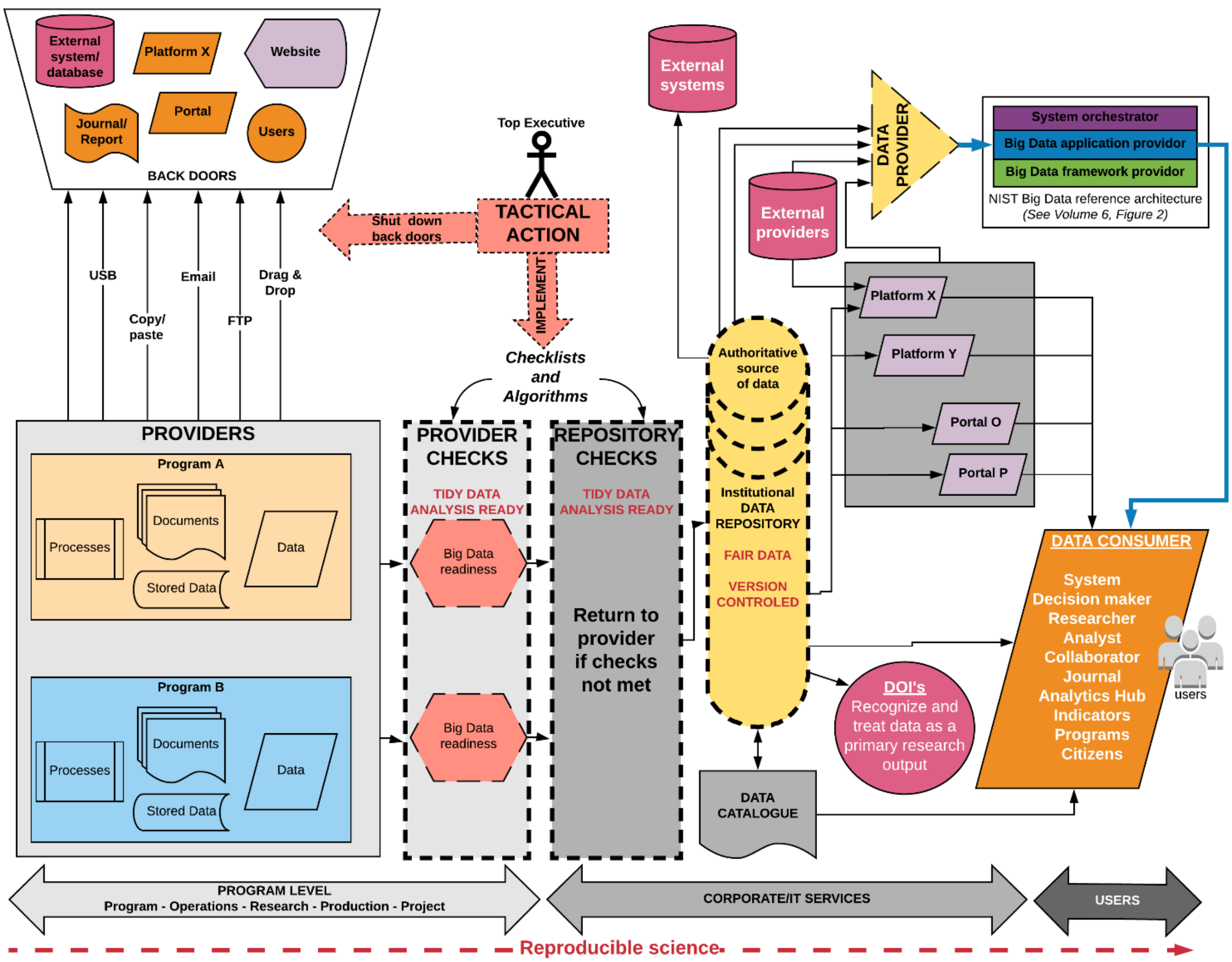

Figure 13: Linear Data Flows for Authoritative Data

6.3.5 PROPOSED STRATEGY

This proposal focuses on structured digital scientific data and the identification of a pathway from Small Data to Big Data, providing a rational stepwise approach to harnessing Big Data. Implementing actions that are generic and independent of systems currently in place means that they can be implemented "*now*":

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 FAIR data to help data providers produce data that are ready for Big Data workflows.
4. Implement a "*user-centric,*" approach to data preparation and release to replace project- and client-centric approaches.[4]
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. Encourage the use of open data science practices and tools.
8. Implement semi-automated data verification and feedback loops to ensure that data are ready for integration into Big Data workflows.
9. Maximize chances of success of Actions 1-8 by including data providers in the development of solutions.

6.3.6 MULTI-FUNCTIONAL DATA CHECKLISTS

Data checklists can be a useful data management tool for data providers and data repositories, as well as for data stewards and managers who need to approve data without having been involved in their production. The use of checklists will help promote consistency, awareness, understanding, and efficiency in data governance. Implemented on the input side of the data repository in Figure 12 and Figure 13, they are a critical component to help ensure FAIR data and to maintain data quality, consistency, and transparency.

[4] A pivotal turning point is the release of data in human readable and machine-readable format. For example, comma-separated values (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. In the case where the data reside in a relational database, users should be able to query the database remotely and/or downloaded it in a freely available format that supports the SQL.

6.3.6.1 Multiple Uses for Data Checklists

Well-designed checklists can serve multiple functions, for example:

1. The data provider can use the checklists as a data auto-evaluation tool.
2. The checklists can be used as a learning tool.
3. Checklist results can be submitted to data stewards and/or management along with or in lieu of the actual data for the purpose of data approval.
4. The institutional digital repository can use the checklists to identify datasets for acceptance into the repository, and to return to the provider for correction datasets that fail to meet all the criteria.
5. Management can easily merge checklist results received from across the organization to get a snapshot of the overall state of data quality and data management.
6. Management can quickly scan the results to identify areas that may require closer attention within a project, identify gaps and areas in general need of improvement across the organization, or identify special cases that legitimately depart from general guidelines.

6.3.6.2 Data Checklist Design

Model data checklists were developed based on insights from the literature [40], [41], [42], [43], [44], [45], [46], [47] and lessons learned from downloading and using a wide range of research, monitoring, and crowd-sourced data. The checklists developed for this paper comprise eight thematic modules with a total of 23 modules or sub-modules (Table 14).

Table 14: Model Data Checklist Modules

Module	Sub-modules
1. Metadata	a) Metadata management b) Provenance c) Multilingualism d) Accessibility
2. Data	a) Raw data b) Data format/structure c) Data collection d) Data preparation e) Geospatial data – additional considerations f) Data management g) Data fitness for use
3. Source	a) Data repository b) Website
4. Visualization	a) Graphics b) Cartography
5. Software	a) Computer code b) Project organization c) File organization d) Computer code changes
6. Reproducibility	
7. Manuscripts	
8. Standards	
9. Confidentiality	

In order to keep implementation of the modules manageable, each module or submodule comprises no more than 10-20 questions each. The model checklists are not meant to be prescriptive, nor are they exhaustive. There is no "one-size-fits-all" or "off-the-shelf" solution. Organizations should adapt the modules and questions to their particular needs—modifying, removing, or adding new ones where necessary—and, implementation should be incremental.

Questions are formulated such that the *preferred* answer is "yes." The items are a mix of general questions (e.g., Are the data FAIR? Are the data accurate?) and detailed sentinel or *canary-in-the-coalmine* questions (Are dates consistently formatted, e.g., YYYY-MM-DD?). When results are compiled across an organization this structure makes it easy to scan and zero in on areas that may require closer attention within a project, identify gaps and areas in general need of improvement, identify training needs, or identify special cases where a "no" response is in fact acceptable. Controlled responses are: "yes", "no", "I don't know", or "not applicable".

The complete set of 23 modules and submodules and the detailed questions included in each of them are provided in Appendix A.

6.3.6.3 Implementation of Checklists

If the checklists are to achieve their intended goal, how they are used is as important as their content. The aim is to improve the organization's data quality and enable Big Data. This should be done in a context of a process of modernization. It requires a phased in approach and a supportive environment, including training both at the working level and for managers. The checklists and the way they are used must have strong support at the highest level of upper management.

An implementation plan should be developed to roll the checklists out in a manner that will ensure effective uptake. The model checklists, offered as a starting point, may not all apply in all situations. They should be pilot tested within the organization prior to implementation, and implementation should be iterative.

1. *Iterative implementation of the checklists.* Create a working group to adapt the checklists to the needs and realities of the organization. Each item should be assigned a level of priority, with approximately one third of the items tagged as either essential, valuable, or desirable for the first round of implementation.
2. *Pilot test the checklist module and sub-module subject headings and adjust as necessary.*
3. *Pilot test the module and sub-module checklist questions and adjust as necessary.*
4. *Round 1 should be implemented uniformly across the organization in conjunction with a data inventory in order to give management a good sense of the overall state of the data.* This should provide the organization with a good understanding of the state of the data in the organization as a whole and in each work unit. This exercise will yield valuable information for long term planning and for identification of where priorities need to be placed in the short term.
5. *In Round 2, the levels of importance should be adjusted to establish "Round 2" goals.* Round 2 goals could target low hanging fruit and what can be done in the short term without increasing resources, as well as a few of the most pressing needs to maximize short term impact and valuable outcomes. Since data management and data quality vary across the organization, Round 2 checklists should be adapted to the needs and realities of each work unit.
6. *Round 3 and successive iterations in each work unit should modify the importance levels of the various items, adding items as necessary, until the final round of implementation when all items would achieve the level of importance, "essential," and all data will be compliant.* At this point, the checklists will have evolved from "It's good enough" to "Best practices," and

will have achieved uniformity across the organization. Thereafter, checklist modules should be revised on a regular basis to keep them relevant to evolving realities.

6.4 NEXT STEPS

Although each organization will need to develop its own path to "*Big Data readiness*", these paths will have a number of similarities: effecting culture change, treating even small data as an organizational and inter-organizational asset, and adopting common standards so that data are useable beyond their original purpose by unimagined systems. This will require commitment both on the part of the individual data creator and on the part of the organization.

Work will need to be done to automate (or semi-automate) the data checklists in order to reduce the amount of manual labor needed to implement them. This work should be done to complement, not compete with open initiatives such as GO FAIR [48], CoreTrustSeal [49], Data Documentation Initiative (DDI) alliance [50], and RDA FAIR Data Maturity Model [33].

Appendix A: Data Checklists

Section 6.3.6 provides guidelines on how to use the data checklists. The columns contain the following information:

- The “Category” column contains the names of the 23 modules or sub-modules, which correspond to those listed in Table 14.
- Entries in the “Current Priority” column are based on organizational maturity (e.g., It's *good enough* for now).
- The “Data Checklist Questions” are potential general and sentinel questions within each category. Questions are formulated such that the preferred answer is “yes.”
- The “Answers” column provides a space for answers to the data checklist questions. The following are allowed answers: “yes”; “no”; “I don't know”; or “not applicable” (with “yes” being the preferred answer).

Table A-1: Model Data Checklist Questions

ID	Category	Current Priority	Data Checklist Questions	Answers
1a-1	Metadata management	1. essential	Do the metadata include a description of the dataset?	
1a-2	Metadata management	3. desirable	Do the metadata include a dataset creation date?	
1a-3	Metadata management	1. essential	Do the metadata include a dataset update date?	
1a-4	Metadata management	3. desirable	Do the metadata include a link to related publications?	
1a-5	Metadata management	3. desirable	Do the metadata include a link to related data products?	
1a-6	Metadata management	2. valuable	Are all metadata provided in a machine-readable format?	
1a-7	Metadata management	2. valuable	Are all metadata provided in a human-readable format?	
1a-8	Metadata management	3. desirable	Are the terms used in the metadata compliant with relevant metadata standards or ontologies?	
1a-9	Metadata management	3. desirable	Do the metadata include a citation that is compliant with JDDCP (Joint Declaration of Data Citation Principles)?	
1a-10	Metadata management	2. valuable	Do the metadata include a description of the methods used for data collection?	

ID	Category	Current Priority	Data Checklist Questions	Answers
1a-11	Metadata management	3. desirable	Do the metadata include a description of the experimental set-up, if applicable?	
1a-12	Metadata management	2. valuable	Is this dataset part of a data collection and, if so, is this described in the metadata?	
1a-13	Metadata management	2. valuable	Is there a data dictionary that describes the contents, format, structure, data collection, and relationship between tables, if applicable?	
1a-14	Metadata management	2. valuable	Do the metadata include all concepts, definitions, and descriptions of all of the variables?	
1a-15	Metadata management	2. valuable	Do the metadata include descriptions of methods, procedures and QA/QC practices followed during production of the data?	
1a-16	Metadata management	1. essential	Are the metadata accurate, complete, up to date, and free of contradictions?	
1a-17	Metadata management	1. essential	Does the documentation match the data files received?	
1a-18	Metadata management	3. desirable	Do the metadata contain keywords selected from a controlled vocabulary, and is the controlled vocabulary properly cited?	
1a-19	Metadata management	2. valuable	Do the metadata distinguish between types of research data, such as primary (original), derived, dynamic, raw, or aggregated data?	
1a-20	Metadata management	2. valuable	Are the metadata registered or indexed in a searchable resource?	
1a-21	Metadata management	2. valuable	Are the metadata assigned a globally unique and eternally persistent identifier?	
1b-1	Provenance	2. valuable	Is the name of the principal investigator included in the metadata record?	
1b-2	Provenance	1. essential	Is the provenance of the data fully and accurately documented in the metadata?	
1b-3	Provenance	1. essential	If applicable, is the data integration process fully and accurately documented in the metadata?	
1b-4	Provenance	1. essential	Is your name and contact information included in the metadata record?	
1b-5	Provenance	1. essential	If the dataset comes from model output, do the metadata include a description of the model that was used?	
1c-1	Multilingualism	2. valuable	Are all elements available in English (i.e., filename, metadata, associated resources, exposed elements in Web services)?	
1c-2	Multilingualism	3. desirable	Are all elements available in an official language other than English (i.e., filename, metadata, associated resources, exposed elements in Web services)?	

ID	Category	Current Priority	Data Checklist Questions	Answers
1c-3	Multilingualism	2. valuable	In the case where multilingual column names are a requirement, are separate rows used for column names in the different languages (e.g., French in row 1, Spanish in row 2, Cree in row 3, English in row 4)?	
1c-4	Multilingualism	2. valuable	In the case of multilingualism, do the metadata include translations of all the column names into the relevant languages (e.g., French, Spanish, Cree, English)?	
1c-5	Multilingualism	3. desirable	In the case of multilingualism and datasets containing text fields in English, do the metadata include translation(s) of all the possible text entries for each variable?	
1d-1	Accessibility	2. valuable	Are the data downloadable?	
1d-2	Accessibility	2. valuable	Are the data available for bulk download?	
1d-3	Accessibility	2. valuable	Does the dataset have a persistent identifier?	
1d-4	Accessibility	3. desirable	Are the metadata files in a non-proprietary format?	
1d-5	Accessibility	2. valuable	Are the data available via an open user licence (i.e., anyone can freely access, use, modify, and share for any purpose—subject, at most, to requirements that preserve provenance and openness [e.g., Creative Commons licence CC0, BY, or BY-SA, or equivalent])?	
1d-6	Accessibility	1. essential	Are the metadata available online?	
1d-7	Accessibility	2. valuable	Are the QA/QC results available online?	
1d-8	Accessibility	2. valuable	Are the raw data available online?	
1d-9	Accessibility	1. essential	Is the encryption used documented in the metadata, if applicable?	
1d-10	Accessibility	1. essential	Is the compression used documented in the metadata, if applicable?	
1d-11	Accessibility	3. desirable	Do users receive notifications of changes?	
1d-12	Accessibility	2. valuable	Are the data updated in a timely fashion?	
1d-13	Accessibility	1. essential	If a standard format was used for the data, is the relevant standard and its version number documented in the metadata?	

ID	Category	Current Priority	Data Checklist Questions	Answers
1d-14	Accessibility	2. valuable	Has long term maintenance of the data been planned, and have sufficient resources been allocated and secured?	
1d-15	Accessibility	2. valuable	Are the metadata retrievable by their identifier using a standardized communications protocol?	
1d-16	Accessibility	2. valuable	Are the metadata accessible, even when the data are not accessible or no longer available?	
2a-1	Raw data	2. valuable	In the case of line-oriented data, is the format CSV (preferred), TSV, or fixed-width (fixed width can be problematic)?	
2a-2	Raw data	2. valuable	In the case of textual works, is the character encoding UTF-8 (preferred), UTF-16 (with Byte Order Mark [BOM]), US-ASCII, or ISO 8859?	
2a-3	Raw data	2. valuable	In the case of textual works, is the format PDF, rich text format, or plain text?	
2a-4	Raw data	2. valuable	In the case of raster images, is the format the same format as the master copy (TIFF, JPEG2000, PNG, JPEG/JFIF, DNG, BMP, or GIF)?	
2a-5	Raw data	2. valuable	In the case of vector images, is the format SVG, DXF, EPS, or shapefile?	
2a-6	Raw data	2. valuable	In the case of audio, is the format PCM WAVE, Broadcast WAVE, CD audio, DSD, or LP?	
2b-1	Data format/structure	2. valuable	In the case of self-describing digital datasets, is the format either JSON (preferred) or XML-based using a well-known schema (or accompanied by the schema employed)?	
2b-2	Data format/structure	2. valuable	In the case where the data reside in a relational database, is the database in third normal form?	
2b-3	Data format/structure	1. essential	In the case where the data do not reside in a relational database, are the data files tabular? In other words, there is one rectangular table per file, systematically arranged in rows and columns with the headers (column names) in the first row. Every record (row) has the same column name. Every column contains the same type of data, and only one type of data.	
2b-4	Data format/structure	1. essential	Are the field types (column types) used appropriate (e.g., date field for dates, alphanumeric field for text, numerical field for numbers)?	
2b-5	Data format/structure	2. valuable	Was a logical, documented naming convention used for variables (column names)?	
2b-6	Data format/structure	1. essential	Are the column names in the first row of the data file?	
2b-7	Data format/structure	1. essential	If these data have undergone analysis and/or visualization, do these results appear in a separate file from the data file?	

ID	Category	Current Priority	Data Checklist Questions	Answers
2b-8	Data format/structure	1. essential	Are the data organized so that both humans and machines can easily read it?	
2b-9	Data format/structure	1. essential	Has the data file been examined for the presence of hidden information which, if found, has been either: made visible, moved somewhere else, or removed?	
2b-10	Data format/structure	1. essential	Do all the columns have a column name (i.e., variable name)?	
2b-11	Data format/structure	1. essential	Are the column names consistent with the documentation?	
2b-12	Data format/structure	2. valuable	Where possible, is human understandable information preferred over coded information (e.g., "cat", "dog" instead of "1", "2" to represent cat and dog, respectively).	
2b-13	Data format/structure	1. essential	Does each record (row) have a unique identifier?	
2b-14	Data format/structure	1. essential	Can the tables in a data collection be linked via common fields (columns)?	
2b-15	Data format/structure	1. essential	Can the data tables be linked to the metadata via common fields (columns)?	
2b-16	Data format/structure	2. valuable	Are the filenames consistent, descriptive, and informative (clearly indicates content) to humans?	
2b-17	Data format/structure	3. desirable	Do the filenames follow the convention: less than 70 characters; most unique content at start of filename; no acronyms; no jargon; no organization name?	
2b-18	Data format/structure	2. valuable	Was a logical, documented naming convention used for file names?	
2b-19	Data format/structure	3. desirable	Are standard/controlled vocabularies used within the data?	
2c-1	Data collection	2. valuable	Is there a written data management plan?	
2c-2	Data collection	2. valuable	Were drop-down menus, look-up tables or reference lists used for variables that should have a fixed code set?	
2c-3	Data collection	2. valuable	Was a quality control technique such as "Statistical Process Control" used to ensure that collected data are accurate?	
2c-4	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])?	

ID	Category	Current Priority	Data Checklist Questions	Answers
2c-5	Data collection	1. essential	Where the dataset contains measured observations, are the units appropriately indicated? (e.g., a separate column for units (preferably), or units as part of the variable name (column name), or units indicated in the metadata for each measurement variable - whichever works best for data usability.	
2c-6	Data collection	1. essential	If there are comments included with the data, is there a separate column for comments?	
2c-7	Data collection	2. valuable	Are consistent phrases used in comment fields?	
2c-8	Data collection	1. essential	Do all empty cells contain a consistent common code for missing data?	
2c-9	Data collection	1. essential	In the case if measurement methods using instruments or analyzers (e.g., in the field or laboratory), are "below detection limit" values included in the data?	
2c-10	Data collection	1. essential	Are the replicate data used to calculate the intraday and interday method detection limits provided?	
2c-11	Data collection	3. desirable	If applicable, is a description of the temporal coverage provided in the metadata?	
2c-12	Data collection	1. essential	Does the information entered in each column correspond to the designated field type (e.g., no non-numeric characters in numeric columns)?	
2c-13	Data collection	1. essential	Where coded information is present in the dataset, is a description of the codes provided in the metadata?	
2c-14	Data collection	1. essential	Do the variables (column) have names that are meaningful to humans (i.e., consistent, descriptive, informative, clearly indicating content)?	
2c-15	Data collection	1. essential	Are dates consistently formatted as YYYY-MM-DD?	
2d-1	Data preparation	1. essential	Are consistent identifiers used for categorical variables?	
2d-2	Data preparation	1. essential	Is a consistent data structure used across all files containing the same type of data?	
2d-3	Data preparation	1. essential	Have stray spaces been removed from the data file?	
2d-4	Data preparation	1. essential	Have apparently empty rows and columns been purged of all unintentional hidden codes?	
2d-5	Data preparation	2. valuable	Are the laboratory-calculated detection limits provided in the metadata?	
2d-6	Data preparation	3. desirable	Do the variables follow a standard?	

ID	Category	Current Priority	Data Checklist Questions	Answers
2d-7	Data preparation	3. desirable	Do the units follow a standard?	
2d-8	Data preparation	3. desirable	Were standard formats used for names of people?	
2d-9	Data preparation	2. valuable	Were standard formats used for civic addresses?	
2d-10	Data preparation	1. essential	Have reference data been used where applicable (e.g., a set of permissible values to be used in specific fields [columns] as defined by third party standard authorities?	
2d-11	Data preparation	1. essential	Is the dataset updated with changes in the reference data as they occur (e.g., standard country codes and time zones change frequently)?	
2d-12	Data preparation	3. desirable	If applicable, are calibrations provided?	
2d-13	Data preparation	2. valuable	Have values been checked to ensure that they fall within a valid range?	
2d-14	Data preparation	1. essential	Have the data been visualized (e.g., plot, map, or both)?	
2d-15	Data preparation	1. essential	Has the dataset been deduplicated?	
2d-16	Data preparation	1. essential	Is the dataset complete?	
2d-17	Data preparation	1. essential	Has the dataset been assessed for accuracy?	
2d-18	Data preparation	1. essential	If timestamps are included in the data is the synchronization methodology documented in the metadata?	
2e-1	Geospatial data—additional considerations	1. essential	If the dataset contains latitude/longitude, is the datum provided?	
2e-2	Geospatial data—additional considerations	3. desirable	Do the metadata include a description of the geospatial coverage?	
2e-3	Geospatial data—additional considerations	1. essential	Do the metadata include a description of the map projection?	
2e-4	Geospatial data—additional considerations	1. essential	Do the latitude/longitude match the data description (e.g., land/water, mountain/valley, northern/southern hemisphere)?	
2e-5	Geospatial data—additional considerations	1. essential	In the case of geospatial data, is the most complete data (e.g., all layers, appendices) provided, even if proprietary?	
2e-6	Geospatial data—additional considerations	1. essential	In the case of geospatial data, is the format compatible with widely adopted geographic information systems (GIS; e.g., ArcGIS)?	

ID	Category	Current Priority	Data Checklist Questions	Answers
2e-7	Geospatial data—additional considerations	1. essential	In the case of geospatial data, is the format developed or endorsed by the Open Geospatial Consortium (OGC; e.g., Geography Markup Language [GML])?	
2f-1	Data management	3. desirable	Was the file integrity checked (e.g., checksum, file size, number of files)	
2f-2	Data management	2. valuable	Are the raw data available online?	
2f-3	Data management	3. desirable	Are the raw data backed up in more than one location?	
2f-4	Data management	2. valuable	Are all the steps used to process the data recorded and available online?	
2f-5	Data management	1. essential	Has the need to join multiple tables been anticipated?	
2f-6	Data management	2. valuable	Is the file organization in a data collection consistent and appropriate?	
2f-7	Data management	2. valuable	Has a unique persistent identifier been associated with each data file (e.g., digital object identifier [DOI])?	
2f-8	Data management	2. valuable	Were the data documented, "as-you-go" rather than at the end of the process?	
2f-9	Data management	2. valuable	Were measures taken to protect security of data in all holdings and all transmissions through encryption or other techniques?	
2f-10	Data management	1. essential	Were measures taken to ensure a "single source of truth" to minimize duplication of information and effort?	
2f-11	Data management	1. essential	Are the datasets prepared at the lowest possible level of granularity (i.e., the data are not summary statistics or aggregated data)?	
2f-12	Data management	3. desirable	Are new datasets output at regular, predictable intervals (e.g., the last day of every month, the last day of the year)?	
2f-13	Data management	2. valuable	Have the data been registered and assigned a DOI?	
2f-14	Data management	2. valuable	Are the data FAIR?	
2f-15	Data management	3. desirable	Was this dataset produced under an organizational data stewardship plan?	
2g-1	Data fitness for use	1. essential	In the case where the data reside in a relational database, can the full database be downloaded in a freely available database format that supports the SQL?	

ID	Category	Current Priority	Data Checklist Questions	Answers
2g-2	Data fitness for use	1. essential	Are the data machine readable?	
2g-3	Data fitness for use	2. valuable	Are the data human readable?	
2g-4	Data fitness for use	1. essential	Can the data be ingested directly into statistical or database software (other than Excel, Word, or Acrobat) without the need to write more than three lines of computer code?	
2g-5	Data fitness for use	1. essential	Are the data in CSV (i.e., comma separated, or character separated) format?	
2g-6	Data fitness for use	3. desirable	In the case of CSV files, is delimiter collision avoided by using a character that is not found elsewhere in the file as the delimiter (e.g., | or ~)?	
2g-7	Data fitness for use	2. valuable	Was a user-centric (i.e., the end-user is unknown), rather than a project- or client-centric, approach used for data preparation?	
2g-8	Data fitness for use	2. valuable	Can the data be incorporated seamlessly into a Big Data workflow?	
2g-9	Data fitness for use	2. valuable	Are the data files in a non-proprietary format?	
2g-10	Data fitness for use	1. essential	Has the file been checked that it can be opened?	
2g-11	Data fitness for use	1. essential	Were new data appended to existing data files?	
2g-12	Data fitness for use	1. essential	If data were appended to existing files, was the documentation updated to reflect changes in the record counts or data layout?	
2g-13	Data fitness for use	1. essential	Were specified data quality assurance practices followed in the production of these data?	
2g-14	Data fitness for use	2. valuable	Are accuracy indicators provided for all of the measured variables?	
2g-15	Data fitness for use	1. essential	Is there absence of matching variables that could be used singly or combined to re-identify anonymized data (e.g., name, address, age, sex, address, industry, occupation) in order to circumvent privacy protection?	
2g-16	Data fitness for use	2. valuable	Is a description available online of any exceptions or limitations in these data?	
2g-17	Data fitness for use	2. valuable	Do the data meet domain specific standards or requirements?	
2g-18	Data fitness for use	1. essential	Are the data fit-for-use by an unknown third party user?	

ID	Category	Current Priority	Data Checklist Questions	Answers
2g-19	Data fitness for use	3. desirable	Is the data file directory structure documented in the metadata?	
3a-1	Data repository	1. essential	Does the repository perform basic curation (e.g., checking, addition of basic metadata or documentation)?	
3a-2	Data repository	1. essential	Does the repository have an explicit mission to provide access to and preserve data?	
3a-3	Data repository	3. desirable	Does the repository maintain all applicable licenses covering data access and use and monitor compliance?	
3a-4	Data repository	3. desirable	Does the repository have a written continuity plan to ensure ongoing access to and preservation of its holdings?	
3a-5	Data repository	3. desirable	Does the repository ensure that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms?	
3a-6	Data repository	1. essential	Does the repository have adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission?	
3a-7	Data repository	1. essential	Does the repository have clear written mechanisms in place to secure ongoing expert guidance and feedback, including scientific guidance?	
3a-8	Data repository	2. valuable	Does the repository guarantee the integrity and authenticity of the data?	
3a-9	Data repository	1. essential	Does the repository accept only data and metadata that meet defined criteria to ensure relevance and understandability for data users?	
3a-10	Data repository	2. valuable	Does the repository apply documented processes and procedures in managing archival storage of the data?	
3a-11	Data repository	2. valuable	Does the repository assume responsibility for long-term preservation and manage this function in a planned and documented way?	
3a-12	Data repository	1. essential	Does the repository have appropriate expertise to address technical data and metadata quality and ensure that sufficient information is available for end users to make quality-related evaluations?	
3a-13	Data repository	2. valuable	Does repository archiving take place according to defined workflows from ingest to dissemination?	
3a-14	Data repository	2. valuable	Does the repository enable users to discover the data and refer to them in a persistent way through proper citation?	
3a-15	Data repository	3. desirable	Does the repository enable reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data?	

ID	Category	Current Priority	Data Checklist Questions	Answers
3a-16	Data repository	2. valuable	Does the repository function on well-supported operating systems and other core infrastructural software and is it using hardware and software technologies appropriate to the services it provides to its Designated Community?	
3a-17	Data repository	1. essential	Does the technical infrastructure of the repository provide for protection of the facility and its data, products, services, and users?	
3a-18	Data repository	3. desirable	Does the repository meet all "Core Trustworthy Data Repositories" requirements?	
3b-1	Website	3. desirable	Does the web page use schema.org dataset markup?	
3b-2	Website	3. desirable	In the case where the dataset does not use shema.org for dataset markup, does it use an equivalent, such as W3C's "Data Catalog Vocabulary (DCAT) format"?	
3b-3	Website	1. essential	Does the web page contain metatags in the <head> section of the html page to provide search information about the content (e.g., title, description)?	
3b-4	Website	2. valuable	In the case of tabular data, WC3 best practices guidelines adhered to for "Tabular Data and Metadata on the Web?"	
3b-5	Website	3. desirable	Is the content optimized for dataset discoverability by Google dataset search?	
4a-1	Graphics	1. essential	In the case of time series data, do the time series display as expected?	
4a-2	Graphics	3. desirable	Are the symbols effective and appropriate to content; do they display well and contribute to ease of understanding?	
4a-3	Graphics	3. desirable	Are standard or standardized symbols used (e.g., thematically standardized symbols for hazards, resources)?	
4a-4	Graphics	3. desirable	Do the symbols convey attribute information (i.e., information about the thing represented by the symbol)?	
4a-5	Graphics	3. desirable	Is a clearly legible legend present?	
4a-6	Graphics	3. desirable	Is the legend meaningful (i.e., informative and clearly indicating the content)?	
4a-7	Graphics	3. desirable	Does the legend include measurement units where applicable?	
4a-8	Graphics	2. valuable	Does the visualization load in a reasonable time period?	
4a-9	Graphics	2. valuable	Is the colour palette effective?	

ID	Category	Current Priority	Data Checklist Questions	Answers
4a-10	Graphics	2. valuable	Is the colour palette perceivable by most forms of colour blindness?	
4a-11	Graphics	2. valuable	Is the visualization clearly rendered (i.e., the quality of the visualization is high, quickly and easily understood at appropriate scale)?	
4b-1	Cartography	2. valuable	In the case of digital maps, is the format GeoTIFF, GeoPDF, GeoJPEG2000, or shapefile?	
4b-2	Cartography	1. essential	Is the map title unique and specific?	
4b-3	Cartography	1. essential	Does the map display what the title says?	
4b-4	Cartography	3. desirable	Are Web mapping services available?	
4b-5	Cartography	3. desirable	Are the contents of the Web Map Service visible at all scales?	
4b-6	Cartography	3. desirable	Is the Web Map Service visible at appropriate scales for the level of detail of the datasets(s)?	
4b-7	Cartography	3. desirable	Are the contents of the Web Map Service consistent between scales?	
4b-8	Cartography	3. desirable	Are the symbols effective and appropriate to content; does it display well and contribute to ease of understanding?	
4b-9	Cartography	3. desirable	Are standard or standardized symbols used (e.g., thematically standardized symbols for hazards, resources)?	
4b-10	Cartography	3. desirable	Do the symbols convey attribute information (i.e., information about the thing represented by the symbol)?	
4b-11	Cartography	1. essential	Is a clearly legible legend present?	
4b-12	Cartography	1. essential	Is the legend meaningful (i.e., informative and clearly indicating the content)?	
4b-13	Cartography	3. desirable	Does the legend include measurement units where applicable?	
4b-14	Cartography	1. essential	Is the map scale shown?	
4b-15	Cartography	1. essential	Is the orientation (north/south) shown?	
4b-16	Cartography	1. essential	Is the map projection shown?	

ID	Category	Current Priority	Data Checklist Questions	Answers
4b-17	Cartography	1. essential	Are the map credits shown (e.g., date of the map data, source of the map data, name of the map creator)?	
5a-1	Computer code—documentation	1. essential	Is there a brief explanatory comment at the start of the code?	
5a-2	Computer code—documentation	1. essential	Is the code liberally commented so that a third party can easily understand what was done at each step?	
5a-3	Computer code—documentation	2. valuable	Has the use of comment/uncomment for sections of code to control the program's behavior been avoided?	
5a-4	Computer code—documentation	2. valuable	Is an overview of the project available online?	
5a-5	Computer code—documentation	2. valuable	Is a shared "to-do" list for the project available online?	
5a-6	Computer code—documentation	3. desirable	Is a description of the communication strategy available online?	
5a-7	Computer code—documentation	1. essential	Are interfaces (inputs and outputs) to code modules well documented?	
5a-8	Computer code—documentation	1. essential	Are all prior assumptions and results of the code described?	
5a-9	Computer code—documentation	3. desirable	Is a checklist created, maintained, and used for saving and sharing changes to the project?	
5a-10	Computer code—documentation	2. valuable	Is there a file called CHANGELOG.txt in the project's docs subfolder?	
5a-11	Computer code—documentation	2. valuable	Is a README file included with the code?	
5b-1	Computer code	2. valuable	Has the code been decomposed into functions?	
5b-2	Computer code	1. essential	Has duplication been eliminated?	
5b-3	Computer code	2. valuable	Does the code include well researched libraries or packages to perform needed tasks?	
5b-4	Computer code	1. essential	Have the libraries and packages used been tested before relying on them?	
5b-5	Computer code	1. essential	Do the functions and variables have meaningful names?	

ID	Category	Current Priority	Data Checklist Questions	Answers
5b-6	Computer code	1. essential	Have dependencies and requirements been made explicit?	
5b-7	Computer code	1. essential	Is a simple example using test dataset provided?	
5b-8	Computer code	2. valuable	Has the code been submitted to a reputable DOI-issuing repository?	
5b-9	Computer code	2. valuable	Is there an explicit license?	
5b-10	Computer code	2. valuable	Are unit tests included with the code?	
5b-11	Computer code	2. valuable	Is the code readable and understandable?	
5b-12	Computer code	2. valuable	Is the code written according to relevant software standards and guidelines?	
5b-13	Computer code	2. valuable	Were static code analysis tools used?	
5b-14	Computer code	2. valuable	Are all libraries and dependencies openly available, in their current versions, and supported?	
5b-15	Computer code	2. valuable	Does the code follow defined architectures and design patterns?	
5b-16	Computer code	2. valuable	Is the code configurable and extensible wherever possible?	
5b-17	Computer code	2. valuable	Are exceptions handled gracefully?	
5b-18	Computer code	1. essential	Are all of the resources used, cleaned up or closed before completion?	
5c-1	Computer code—file organization	3. desirable	Is each project in its own directory, which is named after the project?	
5c-2	Computer code—file organization	3. desirable	Are text documents associated with the project in a ‘documents’ directory?	
5c-3	Computer code—file organization	3. desirable	Are the raw data and metadata in a ‘data’ directory?	
5c-4	Computer code—file organization	3. desirable	Are the files generated during cleanup and analysis in a ‘results’ directory?	
5c-5	Computer code—file organization	3. desirable	Is the project source code in a ‘source’ directory?	

ID	Category	Current Priority	Data Checklist Questions	Answers
5c-6	Computer code—file organization	3. desirable	Are external scripts or compiled programs in a ‘bin’ directory?	
5c-7	Computer code—file organization	1. essential	Do all filenames reflect their content or function?	
5d-1	Computer code changes	3. desirable	Is (almost) everything created by a human backed up as soon as it is created?	
5d-2	Computer code changes	3. desirable	Are changes kept small?	
5d-3	Computer code changes	3. desirable	Are changes shared frequently?	
5d-4	Computer code changes	2. valuable	Is each project stored in a folder that is mirrored off the researcher's working machine?	
5d-5	Computer code changes	2. valuable	Is the entire project copied whenever a significant change has been made?	
5d-6	Computer code changes	1. essential	Is computer code version controlled?	
5d-7	Computer code changes	3. desirable	Are changes conveyed to all users in a timely fashion?	
6a-1	Reproducibility	2. valuable	Are the data the result of a ‘reproducible’ workflow?	
6a-2	Reproducibility	2. valuable	Are known issues/limitations clearly described?	
6a-3	Reproducibility	2. valuable	Are all methods documented in detail such that a third party could reproduce the workflow and obtain the same results without needing to consult with the data provider?	
6a-4	Reproducibility	1. essential	Given the data and information provided, are the data and the limitations of the data completely understandable by a third party without needing to consult with the data provider?	
7a-1	Manuscripts	2. valuable	Is there a peer reviewed data article describing the data? - excluding articles that use the data.	
7a-2	Manuscripts	2. valuable	Are manuscripts written using reference management software?	
7a-3	Manuscripts	3. desirable	Are manuscripts written in a plain text format?	
7a-4	Manuscripts	3. desirable	Are manuscripts deposited in a pre-print repository?	

ID	Category	Current Priority	Data Checklist Questions	Answers
7a-5	Manuscripts	2. valuable	Are manuscripts submitted to an open source, peer reviewed journal?	
7a-6	Manuscripts	1. essential	Do manuscripts identify individual authors and co-authors?	
7a-7	Manuscripts	1. essential	Are manuscripts version controlled?	
7a-8	Manuscripts	1. essential	Are statements properly referenced?	
8a-1	Standards	1. essential	Are date and time compliant with ISO 8601?	
8a-2	Standards	1. essential	IANA (Internet Assigned Numbers Authority) time zone database	
8a-3	Standards	2. valuable	In the case of geospatial data, are the metadata compliant with ISO 19115-NAP?	
8a-4	Standards	2. valuable	Are measurement units compliant with the unified code for units of measure?	
8a-5	Standards	2. valuable	Are Web mapping services compliant with ISO 19128?	
8a-6	Standards	2. valuable	In the case of geospatial data, is the supporting documentation compliant with ISO 19131 (Data product specification)?	
8a-7	Standards	3. desirable	ISO 3166 (Parts 1-3) - Codes for the representation of names of countries and their subdivisions	
8a-8	Standards	3. desirable	ISO 14721 - Open Archival Information System (OAIS) Reference Model	
8a-9	Standards	3. desirable	ISO 15489 - Information and documentation -- Records management	
8a-10	Standards	3. desirable	ISO 27000 - Information security standards	
8a-11	Standards	3. desirable	ISO 639-1 - Codes for the representation of names of languages (Parts 1-5)	
8a-12	Standards	3. desirable	ISO 15836/ANSI Z39.85 (NISOZ3985) - Dublin Core Metadata Element Set	
8a-14	Standards	3. desirable	ISO/IEC 11179 - Information technology -- Metadata registries (MDR)	
9a-1	Confidential or sensitive information	1. essential	Are the data free of confidential information?	

ID	Category	Current Priority	Data Checklist Questions	Answers
9a-2	Confidential or sensitive information	1. essential	Are the data free of sensitive information?	
9a-3	Confidential or sensitive information	1. essential	Were measures taken to protect against disclosure or theft of the confidential information?	
9a-4	Confidential or sensitive information	1. essential	Were measures taken to protect against disclosure or theft of the sensitive information?	
9a-5	Confidential or sensitive information	3. desirable	Is a description of the measures taken to protect against disclosure or theft of confidential information available online?	
9a-6	Confidential or sensitive information	3. desirable	Is a description of the measures taken to protect against disclosure or theft of sensitive information available online?	
9a-7	Confidential or sensitive information	1. essential	Have the data been de-identified by the "safe harbor" method?	
9a-8	Confidential or sensitive information	2. valuable	Have the data been de-identified by a statistical method?	
9a-9	Confidential or sensitive information	3. desirable	Have direct personal identifiers been removed and replaced by codes?	
9a-10	Confidential or sensitive information	2. valuable	Are the data free of restrictions on their use?	
9a-11	Confidential or sensitive information	2. valuable	Are the data managed without an embargo?	
9a-12	Confidential or sensitive information	2. valuable	If applicable, do the metadata contain information on subject consent?	
9a-13	Confidential or sensitive information	2. valuable	If applicable, do the metadata contain information on ethics reviews?	
9a-14	Confidential or sensitive information	2. valuable	If there are restrictions on the use of the data, are the reasons for these restrictions explained in the metadata?	
9a-15	Confidential or sensitive information	2. valuable	If there are restrictions on the use of the data, do the metadata provide information on how to gain controlled access to the data?	

Appendix B: Acronyms

AI	artificial intelligence
API	application programming interface
CCD	continuity of care document
CCR	continuity of care record
COTS	commercial off the shelf
DBMS	database management system
DIY	do-it-yourself
ELT	extract, load, transform
ERP	enterprise resource planning
ETL	extract, transform, load
FAIR	findable, accessible, interoperable, and reusable
FHIR	Fast Healthcare Interoperability Resources
HIT	Healthcare Information Technology
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

Appendix C: Bibliography

[1] W. L. Chang (Co-Chair), N. Grady (Subgroup Co-chair), and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 1, Definitions (NIST SP 1500-1 VERSION 3),” Gaithersburg MD, Sep. 2019 [Online]. Available: https://doi.org/10.6028/NIST.SP.1500-1r2

[2] W. L. Chang (Co-Chair), N. Grady (Subgroup Co-chair), and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies (NIST SP 1500-2 VERSION 3),” Gaithersburg, MD, Sep. 2019 [Online]. Available: https://doi.org/10.6028/NIST.SP.1500-2r2

[3] W. L. Chang (Co-Chair), G. Fox (Subgroup Co-chair), and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 3, Big Data Use Cases and General Requirements (NIST SP 1500-3 VERSION 3),” Gaithersburg, MD, Sep. 2019 [Online]. Available: https://doi.org/10.6028/NIST.SP.1500-3r2

[4] W. L. Chang (Co-Chair), A. Roy (Subgroup Co-chair), M. Underwood (Subgroup Co-chair), and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 4, Security and Privacy (NIST SP 1500-4 VERSION 3),” Gaithersburg, MD, Sep. 2019 [Online]. Available: https://doi.org/10.6028/NIST.SP.1500-4r2

[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 [Online]. Available: https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-5-architectures-white-paper-survey

[6] W. L. Chang (Co-Chair), D. Boyd (Subgroup Co-chair), O. Levin (Version 1 Subgroup Co-Chair), and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 6, Reference Architecture (NIST SP 1500-6 VERSION 3),” Gaithersburg, MD, Sep. 2019 [Online]. Available: https://doi.org/10.6028/NIST.SP.1500-6r2

[7] W. L. Chang (Co-Chair), R. Reinsch (Subgroup Co-chair), D. Boyd (Version 1 Subgroup Co-chair), C. Buffington (Version 1 Subgroup Co-chair), and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap (NIST SP 1500-7 VERSION 3),” Gaithersburg, MD, Sep. 2019 [Online]. Available: https://doi.org/10.6028/NIST.SP.1500-7r2

[8] W. L. Chang (Co-Chair), G. von Laszewski (Editor), and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 8, Reference Architecture Interfaces (NIST SP 1500-9 VERSION 2),” Gaithersburg, MD, Sep. 2019 [Online]. Available: https://doi.org/10.6028/NIST.SP.1500-9r1

[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] Dresner Advisory Services, "2017 Big Data Analytics Market Study," 2017 [Online]. Available: http://dresneradvisory.com/products/2015-big-data-analytics-market-study

[11] A. Naimat, *The Big Data Market: A Data-Driven Analysis of Companies Using Hadoop, Spark, and Data Science*. O'Reilly, 2016 [Online]. Available: https://www.oreilly.com/ideas/the-big-data-market

[12] Datameer, "Big Data - A competitive weapon for the enterprise." p. 1, 2015 [Online]. Available: https://www.datameer.com/wp-content/uploads/2015/09/State_of_the_Industry.pdf

[13] James Manyika *et al.*, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Glob. Inst.*, no. June, p. 156, 2011.

[14] 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

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

[16] AIIM, "Search and Discovery – Exploiting Knowledge, Minimizing Risk," 2014 [Online]. Available: http://www.aiim.org/Resources/Research/Industry-Watches/2014/2014_Sept_Search-and-Discovery

[17] D. Vesset *et al.*, "IDC MaturityScape: Big Data and Analytics 2.0," June, 2015 [Online]. Available: https://www.cacp.ca/index.html?asst_id=1637

[18] B. Klievink, B. J. Romijn, S. Cunningham, and H. de Bruijn, "Big data in the public sector: Uncertainties and readiness," *Inf. Syst. Front.*, vol. 19, no. 2, pp. 267–283, 2017.

[19] MemSQL, "The Lambda Architecture Simplified," Apr. 2016 [Online]. Available: https://insidebigdata.com/white-paper/memsql-lambda-architecture/

[20] B. Hopkins, L. Owens, and J. Keenan, "The Patterns Of Big Data: A Data Management Playbook Toolkit," Jun. 2013 [Online]. Available: https://www.forrester.com/report/The+Patterns+Of+Big+Data/-/E-RES96101

[21] S. Canada, "Compendium of Management Practices for Statistical Organizations from Statistics Canada's International Statistical Fellowship Program," Jul. 2016 [Online]. Available: https://www150.statcan.gc.ca/n1/en/pub/11-634-x/11-634-x2016001-eng.pdf?st=9uKSPAOT

[22] IEEE Big Data Workshop, "Methodologies to Improve Big Data Projects, Homepage," 2019. [Online]. Available: http://www.midp-info.org/

[23] J. Marous, "Innovation in Retail Banking 2017," Oct. 2017 [Online]. Available: https://www.digitalbankingreport.com/trends/innovation-in-retail-banking-2017/

[24] D. Neef, *Digital Exhaust: What Everyone Should Know About Big Data, Digitization and Digitally Driven Innovation*. O'Reilly, Safari, 2014 [Online]. Available: https://www.safaribooksonline.com/library/view/digital-exhaust-what/9780133838190/

[25] 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

[26] J. Reed, "The biggest AI obstacle is culture, not data science skills - a services view from Globant," *diginomica*, 23-Jan-2018. [Online]. Available: https://diginomica.com/the-biggest-ai-obstacle-is-culture-not-data-science-skills-a-services-view-from-globant

[27] E. E. Wood, "Company culture the leading obstacle to digital transformation, survey finds," *itbusiness.ca*, 25-Jul-2017. [Online]. Available: https://www.itbusiness.ca/news/company-culture-the-leading-obstacle-to-digital-transformation-survey-finds-infographic/92984

[28] S. Andriole, "7 Ways for CIOs to Deliver Business Value," *Forbes*, 23-Jan-2019. [Online]. Available: https://www.forbes.com/sites/steveandriole/2019/01/23/7-ways-for-cios-to-morph-into-business-value/#21ce1a543967

[29] CASRAI, "CASRAI dictionary: Research Data Domain," 09-Oct-2018. [Online]. Available: https://dictionary.casrai.org/Category:Research_Data_Domain

[30] E. Commission, "G20 Leaders' Communique Hangzhou Summit," *Press Release Database*, 05-Sep-2016. [Online]. Available: https://europa.eu/rapid/press-release_STATEMENT-16-2967_en.htm

[31] M. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, p. 160018, Mar. 2016 [Online]. Available: http://dx.doi.org/10.1038/sdata.2016.18

[32] The European Open Science Cloud for Research Pilot Project, "Stakeholders Open Consultation," 2018. [Online]. Available: https://eoscpilot.eu/open-consultation

[33] Research Data Alliance, "FAIR Data Maturity Model WG." [Online]. Available: https://www.rd-alliance.org/groups/fair-data-maturity-model-wg

[34] Research Data Alliance, "FAIRSharing Registry: connecting data policies, standards & databases WG." [Online]. Available: https://rd-alliance.org/group/fairsharing-registry-connecting-data-policies-standards-databases.html

[35] C. C. Austin, "A Path to Big Data Readiness," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018 [Online]. Available: https://ieeexplore.ieee.org/document/8622229

[36] Gartner, "Gartner Says CIOs and CDOs Must 'Digitally Remaster' Their Organizations," *Press Release*, 02-Feb-2015. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2015-02-02-gartner-says-cios-and-cdos-must-digitally-remaster-their-organizations

[37] R. Delgado, "Why Your Data Scientist Isn't Being More Inventive," *Dataconomy*, 15-Mar-2016.

[Online]. Available: https://dataconomy.com/2016/03/why-your-datascientist-isnt-being-more-inventive/

[38] J. S. S. Lowndes *et al.*, “Our path to better science in less time using open data science tools,” *Nat. Ecol. Evol.*, vol. 1, 2017 [Online]. Available: https://www.nature.com/articles/s41559-017-0160

[39] EDISON, “EDISON Data Science Framework,” *EDISONcommunity/EDSF*. [Online]. Available: https://github.com/EDISONcommunity/EDSF

[40] K. W. Broman and K. H. Woo, “Data Organization in Spreadsheets,” *Am. Stat.*, vol. 72, no. 1, pp. 2–10, 2018 [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989

[41] J. Kitzes, “Reproducible workflows,” *Data Science Lessons*, 2016. [Online]. Available: http://datasci.kitzes.com/lessons/python/reproducible_workflow.html. [Accessed: 03-Nov-2018]

[42] Government of Canada, “Data quality toolkit,” *Statistics Canada*, Jul-2017. [Online]. Available: https://www.statcan.gc.ca/eng/data-quality-toolkit

[43] H. Wickham, “Tidy Data,” *J. Stat. Softw.*, vol. 59, no. 10, pp. 1–23, 2014 [Online]. Available: https://www.jstatsoft.org/v059/i10

[44] G. Wilson, J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal, “Good enough practices in scientific computing,” *PLoS Comput. Biol.*, vol. 13, no. 6, 2017 [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005510

[45] Google, “Google Search: Dataset,” *Google Developers*, 29-Jul-2019. [Online]. Available: https://developers.google.com/search/docs/data-types/dataset

[46] *ISO 19115-1:2014, Geographic information – Metadata – Part 1: Fundamentals*. International Organization for Standardization, 2014 [Online]. Available: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798

[47] CoreTrustSeal, “Data Repositories Requirements.” [Online]. Available: https://www.coretrustseal.org/why-certification/requirements/

[48] GO FAIR, “GO FAIR, Homepage,” 2019. [Online]. Available: https://www.go-fair.org/

[49] CoreTrustSeal, “CoreTrustSeal, Homepage,” 2019. [Online]. Available: https://www.coretrustseal.org/

[50] Data Documentation Initiative, “Data Documentation Initiative (DDI), Homepage,” 2019. [Online]. Available: https://www.ddialliance.org/

[51] *ISO/IEC 20546 Information technology -- Big data -- Overview and vocabulary*. International Organization for Standardization / International Electrotechnical Commission, 2019.