NIST Special Publication 1500-10

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

NIST Big Data Public Working Group
Standards Roadmap Subgroup

June 2018

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

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

NIST Special Publication 1500-10

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

NIST Big Data Public Working Group (NBD-PWG)
Standards Roadmap 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.1500-10

June 2018

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 1500-10
Natl. Inst. Stand. Technol. Spec. Publ. 1500-10, 38 pages (June 2018) CODEN: NSPUE2

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

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

Keywords

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

Acknowledgements

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

The editors for this document were the following:

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

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

NIST SP1500-10 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[a] to this volume, during Version 1 and/or Version 2 activities, by the following NBD-PWG members:

David Boyd
InCadence Strategic Solutions

Frank Farance
Consultant

Geoffrey Fox
Indiana University

Nancy Grady
SAIC

Zane Harvey
QuantumS3

Haiping Luo
Department of the Treasury

Russell Reinsch
Center for Government Interoperability

Arnab Roy
Fujitsu

Mark Underwood
Krypton Brothers; Synchrony Financial

Gregor von Lasewski
Indiana University

Timothy Zimmerlin
Consultant

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

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

The *NBDIF* is being released in three versions, which correspond to the three development stages of the NBD-PWG work. The three stages aim to achieve the following with respect to the NIST Big Data Reference Architecture (NBDRA).

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

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

1 INTRODUCTION

1.1 BACKGROUND

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

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

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

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

- How is Big Data defined?
- What attributes define Big Data solutions?
- What is new in Big Data?
- What is the difference between Big Data and *bigger data* that has been collected for years?
- How is Big Data different from traditional data environments and related applications?
- What are the essential characteristics of Big Data environments?
- How do these environments integrate with currently deployed architectures?
- What are the central scientific, technological, and standardization challenges that need to be addressed to accelerate the deployment of robust, secure Big Data solutions?

Within this context, on March 29, 2012, the White House announced the Big Data Research and Development Initiative. [9] The initiative's goals include helping to accelerate the pace of discovery in science and engineering, strengthening national security, and transforming teaching and learning by improving analysts' ability to extract knowledge and insights from large and complex collections of digital data.

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

Motivated by the White House initiative and public suggestions, the National Institute of Standards and Technology (NIST) accepted the challenge to stimulate collaboration among industry professionals to further the secure and effective adoption of Big Data. As one result of NIST's Cloud and Big Data Forum held on January 15–17, 2013, there was strong encouragement for NIST to create a public working group for the development of a Big Data Standards Roadmap. Forum participants noted that this roadmap should define and prioritize Big Data

requirements, including interoperability, portability, reusability, extensibility, data usage, analytics, and technology infrastructure. In doing so, the roadmap would accelerate the adoption of the most secure and effective Big Data techniques and technology.

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

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

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

On September 16, 2015, seven NBDIF Version 1 volumes were published (http://bigdatawg.nist.gov/V1_output_docs.php), each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The seven volumes are as follows:

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

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

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

Version 2 of the NBDIF volumes, resulting from Stage 2 work, can be downloaded from the NBD-PWG website (https://bigdatawg.nist.gov/V2_output_docs.php). Potential areas of future work for each volume during Stage 3 are highlighted in Section 1.5 of each volume. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

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 (market, project, and organizational) to adoption of Big Data.
- Section 5 summarizes considerations when implementing Big Data systems or when modernizing existing systems to deal with Big Data.
- Appendices provide acronyms and bibliography for this document.

1.5 FUTURE WORK ON THIS VOLUME

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

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

2 LANDSCAPE PERSPECTIVE

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

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

Table 1: Approximate Adoption by Industry

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

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

Another way of looking at this environment is to view the landscape from the perspective of where money has been spent. Table 2 shows a sample breakdown of Big Data spending by industry across the Asia-Pacific region in 2016 [13], which as a region places big data slightly higher as a priority than Europe, Middle East and Africa; and North America.

Table 2: Sample Spending by Industry

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

3 ADOPTION AND BARRIERS

3.1 EXPLORING BIG DATA ADOPTION

3.1.1 ADOPTION BY INDUSTRY

Adoption of Big Data systems has not been uniform across all industries or sectors. While different industries have different potential to capture value, there are some common challenges that show up across all sectors that could delay adoption of Big Data. A report by the U.S. Bureau of Economic Analysis and McKinsey Global Institute (MGI) suggests that the most obvious barrier to leveraging Big Data is access to the data itself. [15] The MGI report indicates a definite relationship between the ability to access data, and the potential to capture economic value, across all sectors / industries.

For example, the education industry is in the lowest percentile for availability of data, and consequently is also in the lowest 20% for producing economic value. The government sector, which is considered well positioned to benefit from Big Data, suffers from low access to data and may not fully realize the positive impacts of these technologies. [15] Table 3 lists industries that have the best access to data and rate highest on MGI's value index.

Table 3: 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%

3.1.2 FUNCTIONAL PERSPECTIVE OF ADOPTION

Despite the obvious need for improved search technologies, very few organizations have implemented *real* search systems within their stack. AIIM polled 353 members of its global community and found that over 70% considered search to be essential or vital to operations, and equivalent in importance to both Big Data projects and technology-assisted review, yet the majority do not have a mature search function and only 18% have federated capability. [16] There has been very little adoption of open source technologies (~15% on average) across small, medium, and large companies. Forecasts indicate reduced spending on do-it-yourself (DIY)-built OS search apps.

3.2 NONTECHNICAL AND TECHNICAL BARRIERS TO ADOPTION

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 issues. The nontechnical barriers could include issues related to workforce preparedness and availability, high cost, too many or lack of regulations, and organizational culture. Technical challenges encompass issues resulting from the hardware or software, and the interoperability between them, of a Big Data system. Technical barriers arise from

various factors, which include functional components of a Big Data system, integration with those functional components, and the security of those components.

Table 4 lists some of the more significant nontechnical and technical barriers to adoption that were identified in the surveys. Particular industries or organizations could 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. Some barriers not listed in Table 4 may be specific to an industry or a particular organization.

Table 4: Nontechnical and Technical Barriers to Adoption

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

3.2.1 NONTECHNICAL BARRIERS

Frequently cited nontechnical barriers are listed in Table 5 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. Other major concerns are establishing processes to progress from proof-of-concept to production systems and compliance with privacy and other regulations.

In addition to technical considerations, there are also nontechnical barriers to adoption of Big Data. For example, the adoption of access technologies involves nontechnical organizational departments, for legal and security reasons; some silos of data and data access restriction policies are necessary. Poorly defined policies could result in inconsistent metadata standards within individual organizations, which can hinder interoperability.

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.

Table 5 lists several nontechnical barriers to Big Data adoption and the number of respondents that identified the Big Data barrier.

Table 5: Nontechnical Barriers to Adoption

Nontechnical Barriers	**Aggregate Surveys (% of respondents that identified the Big Data barrier)**					
Category • **Sub-category**	**CDW**	**Accenture**	**Knowledgent**	**Hitachi**	**TDWI**	**Information Week**
Difficulty developing an overall management program						
Limited budgets; expensive licenses	32%	47%	47%			34%
Lack of stakeholder definition and product agreement			45%			40%
Difficulty establishing processes to go from POC to production			43%			
Compliance, privacy and regulatory concerns			42%		29%	
• S&P challenge in regulation understanding or compliance						
• Governance: monitoring; doc operating model						
• Governance: ownership						
• Governance: adapting rules for quickly changing end users						
Difficulty operationalizing insights			33%	31%		
Lack of access to sources						
Silos: Lack of willingness to share; departmental communication				36%		

Nontechnical Barriers Category • Sub-category	Aggregate Surveys (% of respondents that identified the Big Data barrier)					
	CDW	Accenture	Knowledgent	Hitachi	TDWI	Information Week
Healthcare Info Tech (HIT)						
• Defining the data that needs to be collected	35%					
• Resistance to change	30%					
• Lack of industry standards	21%					
Lack of buy-in from management				18%	29%	
Lack of compelling use case					31%	
No clear ROI						36%
Lack of practitioners for complexity of software	27%	40%	40%	40%	42%	46%

3.2.2 TECHNICAL BARRIERS TO ADOPTION

Technical barriers include a broad range of issues involving the hardware and software for the Big Data systems. These issues affect every part of the Big Data system, 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. Technical barriers have been identified in the literature, some which are summarized in Table 6. The amount of survey respondents that cited a particular barrier are expressed as a percentage in the table.

Table 6: 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						

Technical Barriers	Aggregate Surveys (% of respondents that identified the Big Data barrier)					
Category • **Subcategory**	**CDW**	**Accenture**	**Knowledgent**	**Hitachi**	**TDWI**	**Information Week**
• Moving data between on-premise systems and clouds						
• Hadoop data						
Hadoop specific						
• Backup and recovery						
• Availability						
• Performance at scale						
• Lack of user friendly tools					27%	
• Security		50%			29%	
Compliance, privacy, and regulatory concerns			42%			
• S&P securing deployments from hack						
• S&P inability to mask, de-identify sensitive data						
• S&P lack of fine control to support hetero user population						
• Governance: auditing access; logging / tracking data lineage						
Analytics layer technical misspecifications						
Lack of suitable software				42%		
Lack of metadata management			25%		28%	
Difficulty providing end users with self-service analytic capability			33%			
Complexity in providing business level context for understanding			33%			

Parallel to market demand for self-service analytics application capabilities is a shift from centralized stewardship, toward a decentralized and granular model where user roles have structure for individual access rules. This shift presents barriers for search, including difficulties managing cloud sharing, mobile tech, and notetaking technologies. In addition, the cloud increases the challenges for governance.

Amongst privacy, security, and regulatory compliance concerns, governance appears to produce significant challenges. Often, privacy stakeholders may not need to be concerned with data in enterprise resource planning (ERP) systems, and security stakeholders may not need to be concerned with business intelligence and analytics systems; but governance stakeholders almost always need to be concerned with those systems, as well as with partner and financial data, and infrastructure components (e.g., database management system [DBMS] and networks). (Reference: Bowles)

The data in Table 6 is organized in a functional orientation. 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.

4 MATURITY

Maturity can be considered from the following three perspectives: 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.

4.1 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. Figure 1 depicts these stages and the significant hurdle from departmental to enterprise adoption.

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.

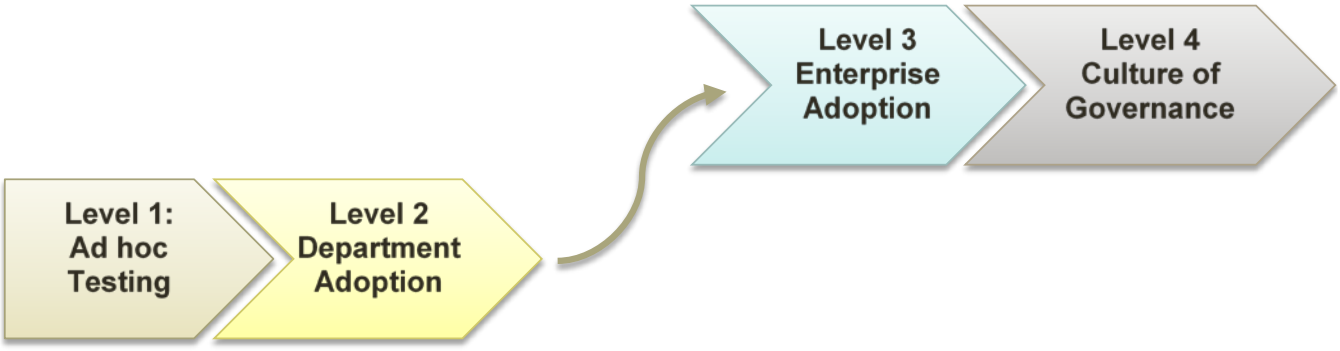

Figure 1: Governance Gap Diagram

4.1.1 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 is siloed; and
- Analytics could be inconsistent across departments.

4.1.2 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. ETL (Extract, Transform, Load)/ELT (Extract, Load, Transform) is performed on an as-needed basis and is tailored to specific requests. The system usually cannot readily incorporate new data sources or perform advanced analytics.

Characteristics of this level include the following:

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

4.1.3 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 developed during this level.

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.

4.1.4 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 application 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.

Information management roles and stewardship applications are two of the primary data management challenges organizations face with respect to governance. Within any single organization, data stewardship may take on one of a handful of particular models. In a data stewardship model that is function-oriented or organization-oriented, the components of the stewardship are often framed in terms of the lines of business or departments that use the data. For example, these departments might be Customer Service, Finance, Marketing, or Sales. All of these organization functions may be thought of as components of a larger enterprise process applications layer, supported by an organization-wide standards layer.

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

4.2 ORGANIZATIONAL MATURITY

Success of Big Data system adoption relies heavily on organizational maturity. Organizations mature at different rates, depending on a variety of factors, and can take months or years. Technical difficulties such as data integration and preparation are often reported as the greatest challenges to successful Big Data projects. However, 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.

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 achieved for successful system implementations. These organizational changes are considered 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 could be considered. 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 lists some organizational changes that should be considered 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.

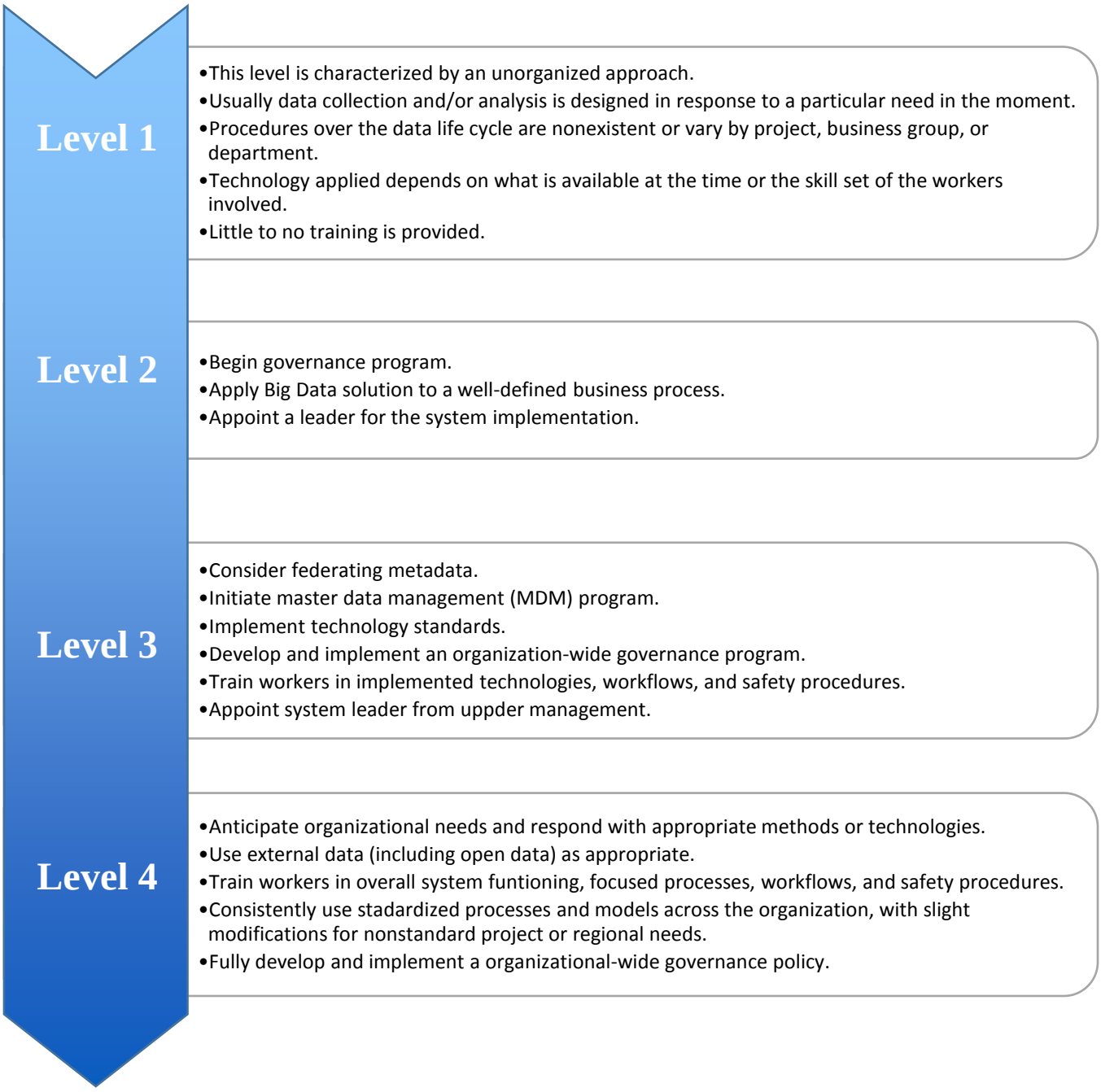

Figure 2: Select Organizational Changes Suggested for Levels of Maturity

As depicted by the gradient on the left of Figure 2, the levels are considered 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 continue through higher levels, including more comprehensive activities, with a fuzzy boundary between levels.

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

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

- Infrastructure as a Service (IaaS): Applications receive a great deal of attention in articles written for business audiences. However, overall, the challenges in applications are proving less difficult to solve than challenges in infrastructure. IaaS is driving many opportunities for commercialization of technology.
- In-memory technologies: It is not always simple to distinguish between in-memory DBMS (Database Management System), in-memory analytics, and in-memory data grids. However, all in-memory technologies will provide a high benefit to organizations that have valid business use cases for adopting these technologies. In terms of maturity, in-memory technologies have essentially reached mainstream adoption and commercialization.
- Access technologies and information retrieval techniques: While access methods for traditional computing are in many cases brought forward into Big Data use cases, legacy access methods present tremendous integration and compliance challenges for organizations tackling Big Data. Solutions to the various challenges remain a work in progress. In some cases, proprietary, patented access methods have been a barrier to construction of connectors required for federated search and connectivity.
- Internal search: In one survey of organizations considering Big Data adoption, “Only 12% have an agreed-upon search strategy, and only half of those have a specific budget.” [16] The top two challenges to internal search seem to be a lack of available staff with the skills to support the function, and the organization’s ability to dedicate personnel to maintain the related servers. Departments are reluctant to take ownership of the search function due to the problematic levels of the issues. The consensus amongst AIIM’s survey respondents was that the Compliance, Inspector General, or Records Management department should be the responsible owner for the search function. An underlying problem persists in some larger organizations, however, where five or more competing search products can be found, due to small groups each using their own tools.
- Stream processing: Continued adoption of streaming data will benefit from technologies that provide the capability to cross-reference (i.e., unify) streaming data with data at rest.

4.4 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 [19], which still relied heavily on ETL processes. The hub and spoke architecture provides multiple options for working with data in the hub, or for moving data out to the spokes for more specific task requirements, enabling for data persistence capabilities on one hand and data exposure (i.e., for analytics) capabilities on the other.

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

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

Table 7: Maturity Projections

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

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

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

5 CONSIDERATIONS FOR IMPLEMENTATION AND MODERNIZATION

5.1 SYSTEM MODERNIZATION

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

- Augmentation: This direction involves updating to a Big Data system by augmenting the supporting architecture. Advantages of updating the supporting architecture include incorporation of more mature technologies amidst the stack and flexibility in the implementation timeline. Augmentation allows for a phased implementation that can be stretched out over more than one fiscal budget year.
- Replacement: This direction involves updating to a Big Data system by replacing the existing system with an entirely new system. Modernizing an existing system by replacing the whole architecture has notable disadvantages. In comparison to the augmentation approach, the level of change management required when replacing entire systems is significantly higher. One advantage of complete system replacement is reduced compatibility problems with legacy systems. Partial modernizations, by replacing a portion of the existing system, are also possible. However, the same advantages and disadvantages of complete system replacement may not apply.

Once system augmentation or replacement has been elected, a method of implementation can be chosen. Figure 3 diagrams a decision situation, commonly referred to as *build or buy* (or outsource) that organizations face when modernizing to a Big Data system. In the build, or DIY scenario, the organization may modify their existing system or build an entirely new system separate of the existing system. One of the largest barriers organizations face when building their own systems is the scarcity of engineers with the skill set covering the newer technologies such as streaming or near real-time analysis.

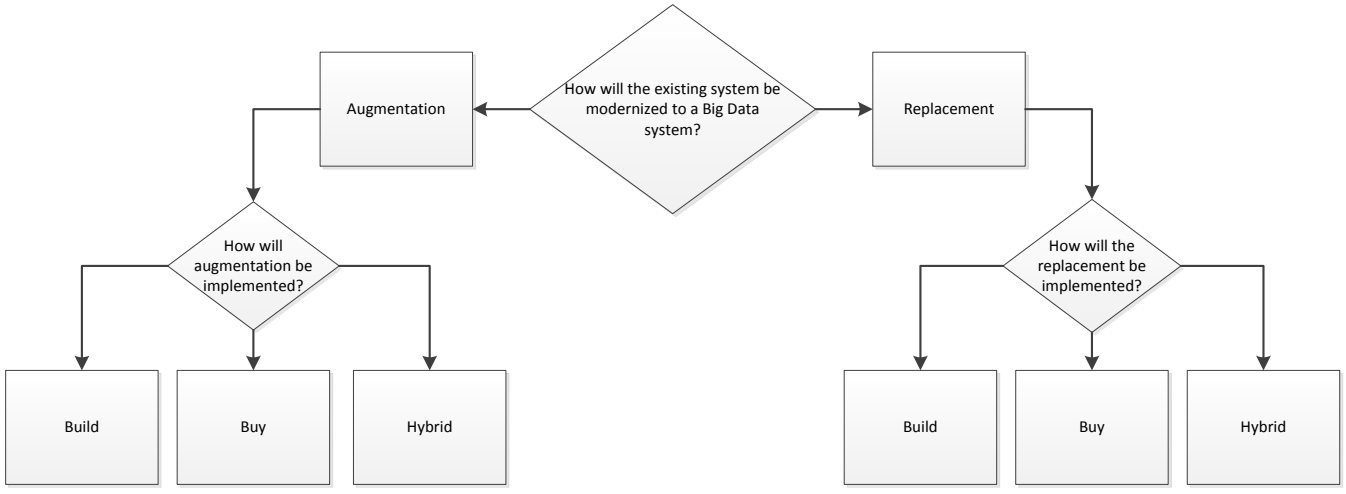

Figure 3: New System Implementation

If the DIY implementation is erected concurrent to the existing system, the organization is required to operate two systems for the length of time it will take to get the new system running and migrate data or combine components.

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

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

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

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

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

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

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

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

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

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

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

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

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

5.2 IMPLEMENTATION

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

- Data quality: Consider the level of quality that will be required from the data model. As data quality increases, cost increases. A minimum viable quality of data, which will provide desired results, should be determined.
- Data access: Many factors can affect data access including organizational cultural challenges and security and privacy compliance. Cultural challenges are unique to each project but many are alleviated with sufficient support from upper management (e.g., corporate officers, influential advocates). Security and privacy affects multiple areas in a Big Data project including data access. Additional information on

security and privacy considerations are provided in the *NBDIF: Volume 4, Security and Privacy* document.

- Component interoperability: For a complicated system, a comprehensive appraisal of system component interoperability can be critical. Advantages of commercial products are frequently lauded while the limitations, dependencies, and deficiencies are often not obvious. Exploration of component interoperability during the planning phase could prevent significant issues during later phases of Big Data projects.
- Potential bottlenecks: Projects requiring high performance often expose storage and network bottlenecks.
- For search-oriented projects: Organizations should strive to set a balance between governance and retrieval, determine ownership (i.e., departmental responsibility) for the function, aim for unified or single-point search capability; and unless the organization is a strong IT company, identify needed outsourced expertise.

6 SPECIFIC TECHNIQUES DEPENDENT ON THE PROBLEM SPACE

Figure 4 very much oversimplifies some of the questions related to system capability that an organization may need to consider when planning their own system; its purpose here is to demonstrate how project requirements can drive decision making. The list of choices presented is not intended to be comprehensively complete. Inclusion is not an endorsement for usage, and no solutions have been intentionally excluded.

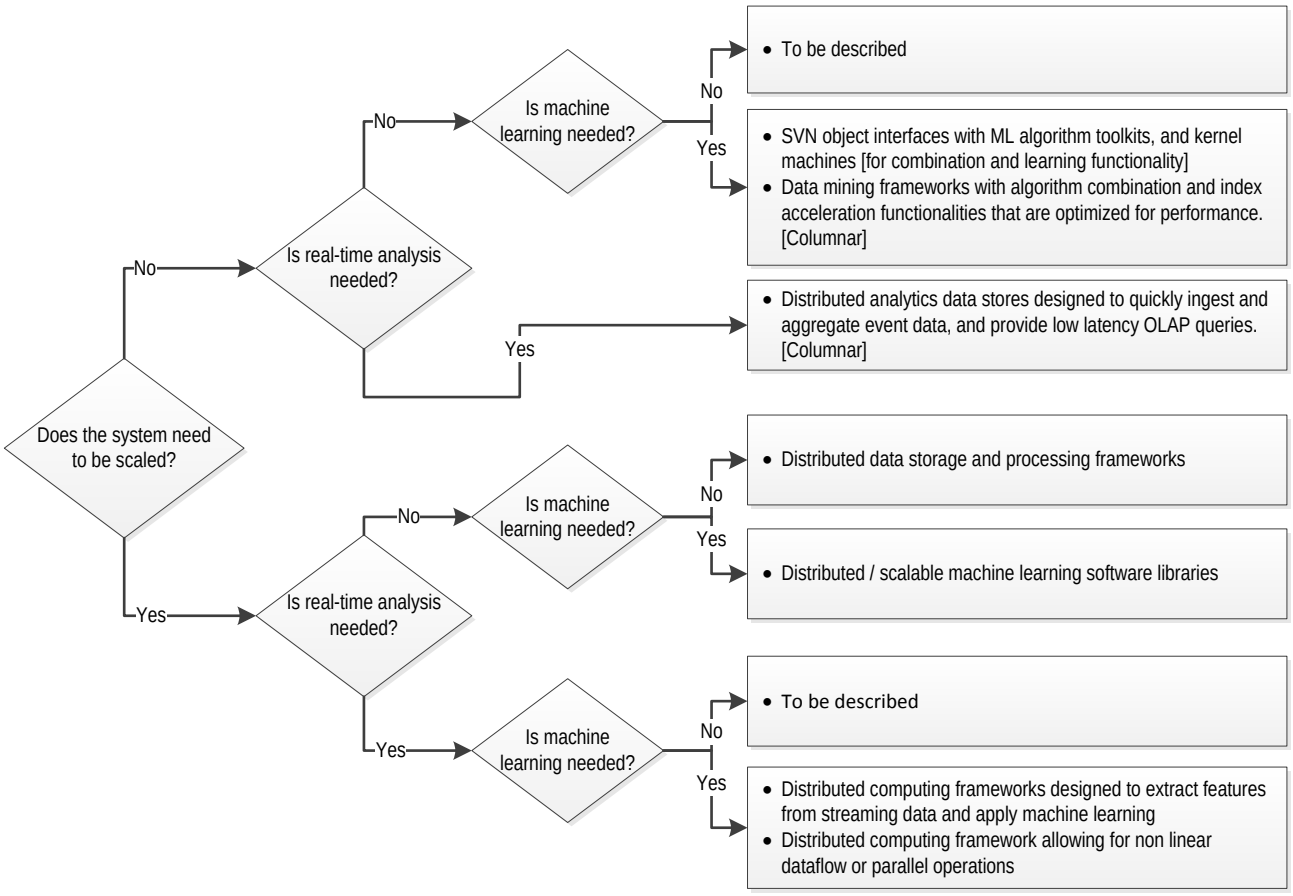

Figure 4: Requirement Decision Tree

After the scalability and latency requirements are considered as shown in Figure 4, the systems planning process will require continued consideration on whether machine learning is necessary. Figures 5, 6, and 7 map the workflow of the machine learning decision trees and show the decision points in the application of machine learning algorithms. Tables 10, 11, 12, and 13 list specific algorithms for each algorithm subgroup.

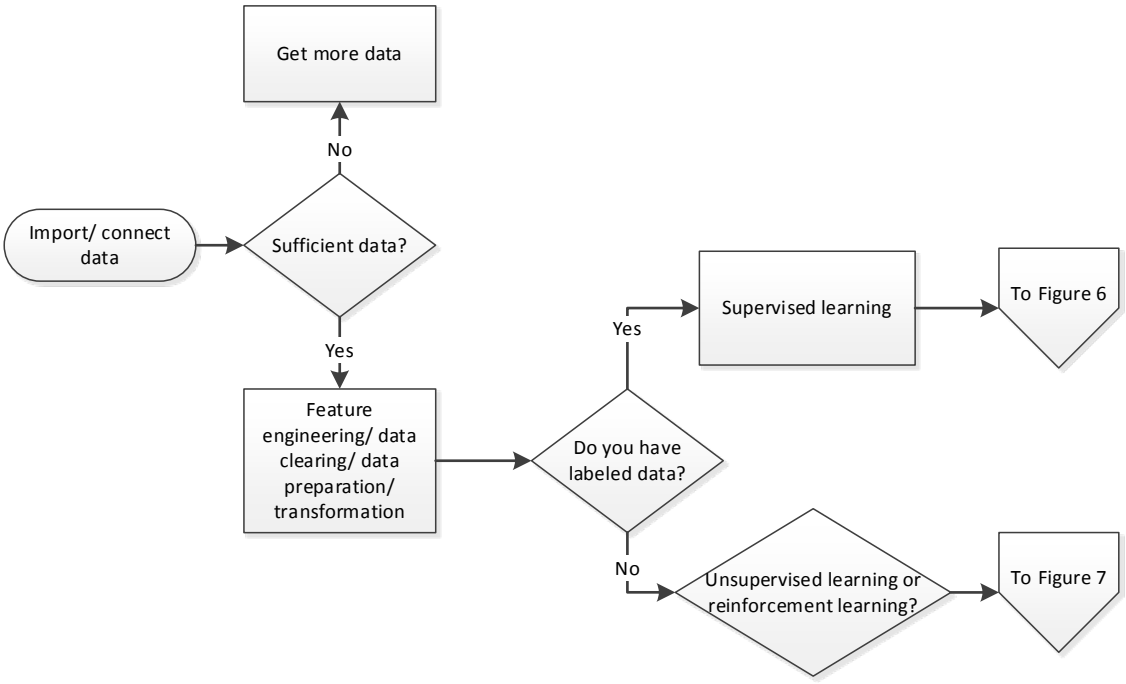

Figure 5: Machine Learning Algorithm Application Workflow

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). Figures 6 and 7 expand on algorithm choices for each problem subclass. Tables 10 and 11 continue from Figure 6 to provide additional information for the regression or classification algorithms. Tables 12 and 13 provide additional information on the unsupervised algorithms and techniques shown in Figure 7.

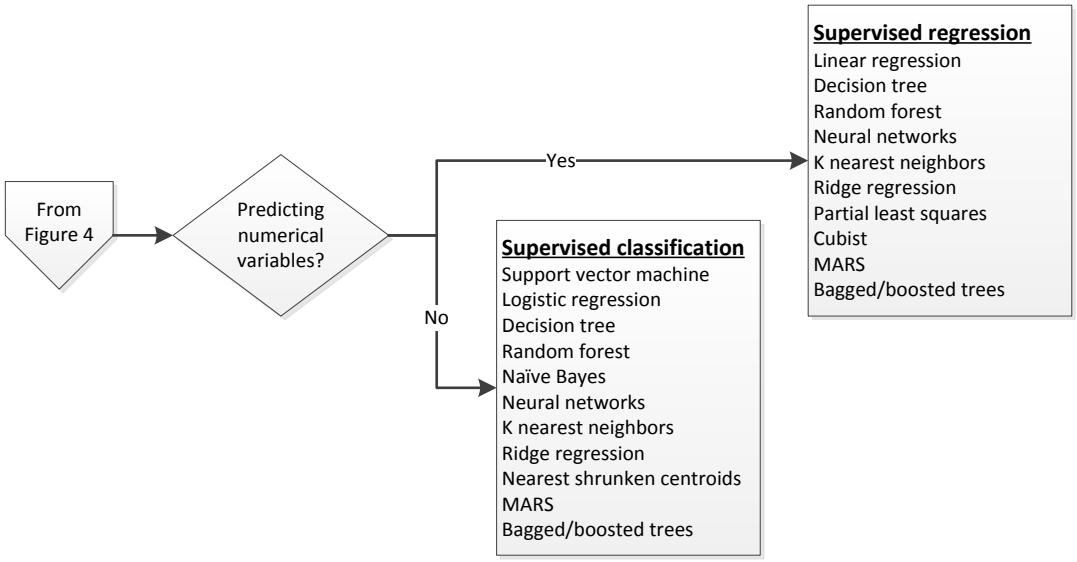

Figure 6: Supervised Machine Learning Algorithms

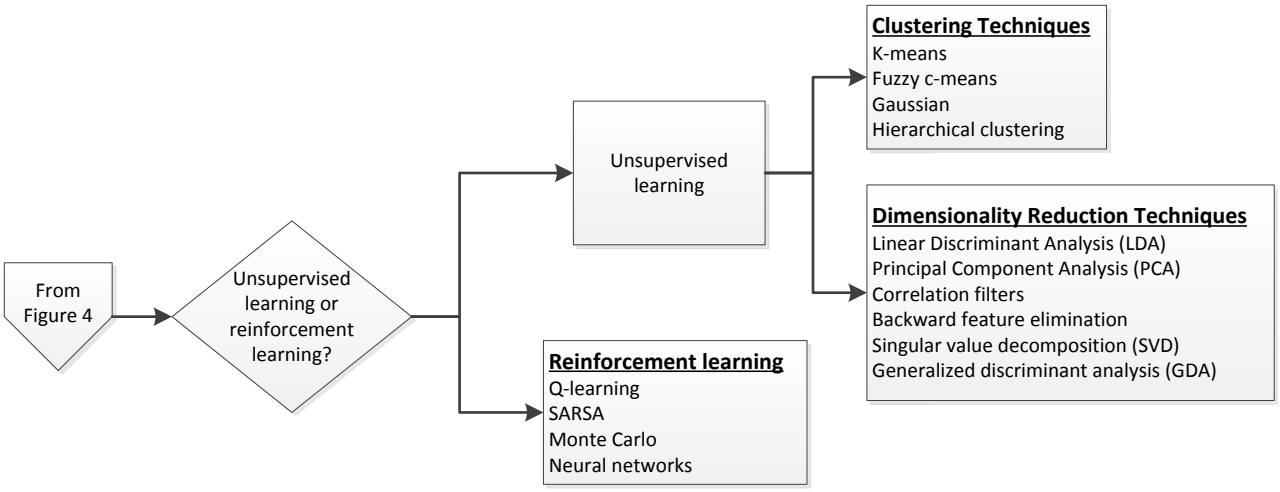

Figure 7: Unsupervised or Reinforcement Machine Learning Algorithms

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 algorithm two subgroups: 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 11 compare supervised learning regression algorithms using four categories and supervised learning classification algorithms using the same four categories.

Table 10: Supervised Learning Regression Algorithms

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

Table 11: Supervised Learning Classification Algorithms

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

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

Table 12: Unsupervised Clustering Algorithms

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

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

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

Table 13: Dimensionality Reduction Techniques

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

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

Appendix A: Acronyms

API	application programming interface
CCD	Continuity of Care Document
CCR	Continuity of Care Record
DBMS	Database Management System
DIY	Do-It-Yourself
ELT	Extract, Load, Transform
ERP	Enterprise Resource Planning
ETL	Extract, Transform, Load
FHIR	Fast Healthcare Interoperability Resources
HIT	Healthcare Info Tech
IaaS	Infrastructure as a Service
iPaaS	integration Platform as a Service
IT	information technology
ITL	Information Technology Laboratory at NIST
MARS	Multivariate Adaptive Regression Splines
MGI	McKinsey Global Institute
NBDIF	NIST Big Data Interoperability Framework
NBD-PWG	NIST Big Data Public Working Group
NBDRA	NIST Big Data Reference Architecture
NIST	National Institute of Standards and Technology
OS	operating system
R&D	research and development
ROI	return on investment

Appendix B: Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[22] J. Taylor, "Analytics Capability Landscape," 2015.