NIST Special Publication 1500-7r2

NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap

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.1500-7r2

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

NIST Special Publication 1500-7r2

NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap

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.1500-7r2

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 1500-7r2
89 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.

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

While opportunities exist with Big Data, the data can overwhelm traditional technical approaches. To advance progress in Big Data, the NIST Big Data Public Working Group (NBD-PWG) is working to develop consensus on important, fundamental concepts related to Big Data. The results are reported in the *NIST Big Data Interoperability Framework* (BDIF) series of volumes. This volume, Volume 7, contains summaries of the work presented in the other six volumes, an investigation of standards related to Big Data, and an inspection of gaps in those standards.

Keywords

Big Data; Big Data Application Provider; Big Data characteristics; Big Data Framework Provider; Big Data standards; Big Data taxonomy; Data Consumer; Data Provider; Management Fabric; reference architecture; Security and Privacy Fabric; System Orchestrator; use cases.

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 (San Diego Supercomputer Center), 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***: David Boyd (InCadence Strategic Solutions), Carl Buffington (Vistronix), and Wo Chang (NIST)
- ***Version 2***: Russell Reinsch (Center for Government Interoperability) and Wo Chang (NIST)
- ***Version 3:*** 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 SP1500-7, 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, Census, 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, Version 2, and/or Version 3 activities, by the following NBD-PWG members:

Claire C. Austin
Department of the Environment Government of Canada

Chaitan Baru
University of California, San Diego, Supercomputer Center

David Boyd
InCadence Strategic Services

Carl Buffington
Vistronix

Zane Harvey
QuantumS3

Bruno Kelpsas
Microsoft Consultant

Pavithra Kenjige
PK Technologies

Brenda Kirkpatrick
Hewlett-Packard

Donald Krapohl
Augmented Intelligence

Sara Mazer
Marklogic

Shawn Miller
U.S. Department of Veterans Affairs

William Miller
MaCT USA

Sanjay Mishra
Verizon

Quyen Nguyen
NARA

[a] "Contributors" are members of the NIST Big Data Public Working Group who dedicated great effort to prepare and gave substantial time on a regular basis to research and development in support of this document. All opinions are the authors' own.

Wo Chang
NIST

Yuri Demchenko
University of Amsterdam

Kate Dolan
CTFC

Frank Farance
Farance, Inc.

Nancy Grady
SAIC

Keith Hare
JCC Consulting, Inc.

Luca Lepori
Data Hold

Orit Levin
Microsoft

Jan Levine
kloudtrack

Serge Mankovski
CA Technologies

Robert Marcus
ET-Strategies

Gary Mazzaferro
AlloyCloud, Inc.

Russell Reinsch
Center for Government Interoperability

John Rogers
Hewlett-Packard

Doug Scrimager
Slalom Consulting

Cherry Tom
IEEE-SA

Mark Underwood
Krypton Brothers; Synchrony Financial

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

To provide a common Big Data framework, the NIST Big Data Public Working Group (NBD-PWG) is creating vendor-neutral, technology- and infrastructure-agnostic deliverables, which include the development of consensus-based definitions, taxonomies, a reference architecture, and a roadmap. This document, *NIST Big Data Interoperability Framework (NBDIF): Volume 7, Standards Roadmap*, summarizes the work of the other NBD-PWG subgroups (presented in detail in the other volumes of this series) and presents the work of the NBD-PWG Standards Roadmap Subgroup. The NBD-PWG Standards Roadmap Subgroup investigated existing standards that relate to Big Data, initiated a mapping effort to connect existing standards with both Big Data requirements and use cases (developed by the Use Cases and Requirements Subgroup), and explored gaps in the Big Data standards.

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 (this volume)
- Volume 8, Reference Architecture Interfaces [7]
- Volume 9, Adoption and Modernization [8]

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?

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

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

- How is Big Data defined?
- What attributes define Big Data solutions?
- What is 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 (this volume)
- Volume 8, Reference Architecture Interfaces [7]
- Volume 9, Adoption and Modernization [8]

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).

1.2 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 standards gap analysis and document the findings;
- Identify possible barriers that may delay or prevent adoption of Big Data; and
- Identify a few areas where 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, 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 *characteristics*; and (4) represent a fundamental function. The aim is to enable data scientists to perform analytics processing for their given data sources without worrying about the underlying computing environment.

1.3 REPORT PRODUCTION

The *NBDIF: Volume 7, Standards Roadmap* 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 7, Standards Roadmap* is dedicated to developing a consensus vision with recommendations on how Big Data should move forward specifically in the area of standardization. In the first phase, the Subgroup focused on the identification of existing standards relating to Big Data and inspection of gaps in those standards. During the second phase, the Subgroup mapped standards to requirements identified by the NBD-PWG, mapped standards to use cases gathered by the NBD-PWG, and discussed possible pathways to address gaps in the standards. To achieve technical and high-quality document content, this document will go through a public comments 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 summarizes the work developed by the other four subgroups and presents the mapping of standards to requirements and standards to use cases.

- Section 3 reviews existing standards that may apply to Big Data, provides two different viewpoints for understanding the standards landscape, and considers the maturation of standards.
- Section 4 presents current gaps in Big Data standards, and examines areas where the development of standards could have significant impact.

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 NBDIF ECOSYSTEM

The exponential growth of data is already resulting in the development of new theories addressing topics from synchronization of data across large distributed computing environments, to addressing consistency in high-volume and high-velocity environments. The NBDIF is intended to represent the overall topic of Big Data, grouping the various aspects of the topic into high-level facets of the ecosystem. At the forefront of the construct, the NBD-PWG laid the groundwork for construction of a reference architecture. Development of a Big Data reference architecture involves a thorough understanding of current techniques, issues, concerns, and other topics.

To this end, the NBD-PWG collected use cases to gain an understanding of current applications of Big Data, conducted a survey of reference architectures to understand commonalities within Big Data architectures in use, developed a taxonomy to understand and organize the information collected, and reviewed existing Big Data-relevant technologies and trends. From the collected use cases and architecture survey information[b], the NBD-PWG created the NBDRA, which is a high-level conceptual model designed to serve as a tool to facilitate open discussion of the requirements, structures, and operations inherent in Big Data. These NBD-PWG activities and functional components were used as input during the development of the entire NIST Big Data Interoperability Framework. The remainder of Section 2 summarizes the NBD-PWG work contained in other NBDIF Volumes.

2.1 DEFINITIONS

There are two fundamental concepts in the emerging discipline of Big Data that have been used to represent multiple concepts. These two concepts, Big Data and Data Science, are broken down into individual terms and concepts in the following subsections. As a basis for discussions of the NBDRA and related standards, associated terminology is defined in subsequent subsections. The *NBDIF: Volume 1, Definitions* explores additional concepts and terminology surrounding Big Data.

2.1.1 DATA SCIENCE DEFINITIONS

In its purest form, data science is the fourth paradigm of science, following theory, experiment, and computational science. The fourth paradigm is a term coined by Dr. Jim Gray in 2007 to refer to the conduct of data analysis as an empirical science, learning directly from data itself. Data science as a paradigm would refer to the formulation of a hypothesis, the collection of the data—new or preexisting—to address the hypothesis, and the analytical confirmation or denial of the hypothesis (or the determination that additional information or study is needed.) As in any experimental science, the result could in fact be that the original hypothesis itself needs to be reformulated. The key concept is that data science is an empirical science, performing the scientific process directly on the data. Note that the hypothesis may be driven by a business need, or can be the restatement of a business need in terms of a technical hypothesis.

> ***Data science** is the extraction of useful knowledge directly from data through a process of discovery, or of hypothesis formulation and hypothesis testing.*

While the above definition of the data science paradigm refers to learning directly from data, in the Big Data paradigm, this learning must now implicitly involve all steps in the data life cycle, with analytics

[b] See NBDIF: Volumes 3, 5, and 6, version 1 for additional information on the use cases, reference architecture information collection, and development of the NBDRA.

being only a subset. Data science can be understood as the activities happening in the data layer of the system architecture to extract knowledge from the raw data.

*The **data life cycle** is the set of processes that transform raw data into actionable knowledge, which includes data collection, preparation, analytics, visualization, and access.*

Traditionally, the term analytics has been used as one of the steps in the data life cycle of collection, preparation, analysis, and action.

***Analytics** is the synthesis of knowledge from information.*

2.1.2 BIG DATA DEFINITIONS

Big Data refers to the inability of traditional data architectures to efficiently handle the new datasets. Characteristics of Big Data that force new architectures are ***volume*** (i.e., the size of the dataset) and ***variety*** (i.e., data from multiple repositories, domains, or types), and the data in motion characteristics of ***velocity*** (i.e., rate of flow) and ***variability*** (i.e., the change in other characteristics). These characteristics—volume, variety, velocity, and variability—are known colloquially as the Vs of Big Data and are further discussed in the *NBDIF: Volume 1, Definitions*.

Each of these characteristics influences the overall design of a Big Data system, resulting in different data system architectures or different data life cycle process orderings to achieve needed efficiencies. A number of other terms are also used, several of which refer to the analytics process instead of new Big Data characteristics. The following Big Data definitions have been used throughout the seven volumes of the NBDIF and are fully described in the *NBDIF: Volume 1, Definitions*.

***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.*

*The **Big Data paradigm** consists of the distribution of data systems across horizontally coupled, independent resources to achieve the scalability needed for the efficient processing of extensive datasets.*

***Veracity** refers to accuracy of the data.*

***Value** refers to the inherent wealth, economic and social, embedded in any dataset.*

***Volatility** refers to the tendency for data structures to change over time.*

***Validity** refers to appropriateness of the data for its intended use*

Like many terms that have come into common usage in the current information age, Big Data has many possible meanings depending on the context from which it is viewed. Big Data discussions are complicated by the lack of accepted definitions, taxonomies, and common reference views. The products of the NBD-PWG are designed to specifically address the lack of consistency. The NBD-PWG is aware that both technical and nontechnical audiences need to keep abreast of the rapid changes in the Big Data landscape as those changes can affect their ability to manage information in effective ways.

For each of these two unique audiences, the consumption of written, audio, or video information on Big Data is reliant on certain accepted definitions for terms. For nontechnical audiences, a method of expressing the Big Data aspects in terms of volume, variety and velocity, known as the *Vs*, became popular for its ability to frame the somewhat complex concepts of Big Data in simpler, more digestible ways.

Similar to the who, what, and where interrogatives used in journalism, the Vs represent checkboxes for listing the main elements required for narrative storytelling about Big Data. While not precise from a terminology standpoint, they do serve to motivate discussions that can be analyzed more closely in other settings such as those involving technical audiences requiring language which more closely corresponds to the complete corpus of terminology used in the field of study.

Tested against the corpus of use, a definition of Big Data can be constructed by considering the essential technical characteristics in the field of study. These characteristics tend to cluster into the following five distinct segments:

1. Irregular or heterogeneous data structures, their navigation, query, and data-typing (aka, variety);
2. The need for computation and storage parallelism and its management during processing of large datasets (aka, volume);
3. Descriptive data and self-inquiry about objects for real-time decision making (aka, validity/veracity);
4. The rate of arrival of the data (aka, velocity); and
5. Presentation and aggregation of such datasets (i.e., visualization) [10]

With respect to computation parallelism, issues concern the unit of processing (e.g., thread, statement, block, process, and node), contention methods for shared access, and begin-suspend-resume-completion-termination processing.

Descriptive data is also known as metadata. Self-inquiry is often referred to as reflection or introspection in some programming paradigms.

With respect to visualization, visual limitations concern how much information a human can usefully process on a single display screen or sheet of paper. For example, the presentation of a connection graph of 500 nodes might require more than 20 rows and columns, along with the connections or relationships among each of the pairs. Typically, this is too much for a human to comprehend in a useful way. Big Data presentation concerns itself with reformulating the information in a way that makes the data easier for humans to consume.

It is also important to note that Big Data is not necessarily about a large amount of data because many of these concerns can arise when dealing with smaller, less than gigabyte datasets. Big Data concerns typically arise in processing large amounts of data because some or all of the four main characteristics (irregularity, parallelism, real-time metadata, presentation / visualization) are unavoidable in such large datasets.

2.1.3 ADDITIONAL DEFINITIONS

As a result of analysis performed during work on this volume, the need arose for a modern definition of integration, as it would apply to Big Data in 2018. The term integration has often been used to refer to a broad range of activities or functions related to data processing. Those activities or functions can include application integration middleware (for business line communications processes), message queues, data integration, Application Programming Interfaces (APIs), or even systems integration or continuous integration (i.e., code versioning). While the NBD-PWG respects the importance of all of these activities, not all activities are within the scope of this Version 3 of the *NBDIF: Volume 7, Standards Roadmap*.

Within the scope of this document, a modern definition for integration can be thought of in terms of a structure for database coupling in the storage layer; extract, load, and transform (ELT) and extract, transform, load (ETL) in the compute layer; app integration and event updating in the app layer; and query processing in the presentation layer.

As of the publication date of this document, data integration is widely recognized as one of the primary elements required for leveraging Big Data environments [11], [12], [13], [14], [15].

2.1.3.1 Connectivity in Integration

Connectivity is normally the first step in data processing, and support for all types of connections and all types of data are the dreams of Big Data users everywhere. Most off-the-shelf data warehouse data acquisition products offer a stable of connectors as part of the package. However, the 'usability' of a connector is just as important as the availability of the connector. The diversity of data types and data sources frequently means that custom middleware code must be written in order for a connector to work.

An area ripe for development is compatibility with different ETL techniques. This is not to imply that ETL is always required. It is important to note the current lack of standards for connectors to content management systems, collaboration apps, web portals, social media apps, customer relationship management systems, file systems, databases, and APIs.

Truly modern data acquisition workflows require easier-to-use graphic interfaces that abstract the complexities of programming a connector, away from the casual user. As the range of sources for data capture widens, the probability is greater that a more capable Master Data Management (MDM) or governance solution would be appropriate.

Aside from the types of data being captured, the modes of interaction or 'speed' of the data may dictate the type of integration required. The data warehouse is the traditional use case for data integration. In this scenario, large batches of transactions are extracted from a location point where they are at-rest, then processed in a single run that can take hours to complete. In some Big Data processing scenarios, users want immediate access to data that is streaming in-motion, so the system delivers results in real time, by capturing and processing small chunks of data within seconds. Real-time systems are more difficult to build and implement.

2.1.3.2 Translation in Integration

Big Data use cases brought about changes to traditional data integration scenarios. Traditional data integration focused on the mechanics of moving structured data to or from different types of data structures via extraction from the source, transformation of that data into a format recognized by the target application, and then loading transformed data into the target application. The most notable change to data integration approaches comes in the form of a process where data is loaded immediately into a target location without any transformation; thus the transformation takes place inside the target system.

Legacy ETL techniques historically configured separate tools for change data capture (CDC), replication, migration, etc. As the demand for additional capabilities required technologies with wider scopes, basic product lines in the ETL industry took on additional capabilities. Some technologies specialized in functions such as federation and data virtualization, synchronization, or data preparation.

ETL is still important to data integration; however, with modern Big Data use cases, organizations are challenged to deal with unstructured data and fast moving data in motion, either of which results in a Big Data program requiring more attention to additional related systems such as MDM, synchronization, and data quality [16]. As such, there is a serious need for improved standardization in metadata and business rule management.

Modern translation workflows require metadata interfaces that provide nontechnical users with functionality for working with metadata. One concern often left unchecked, however, is for a consistent version of the data. Federation and data virtualization allow for stability of the data while integration work is performed. For example, an end user need not necessarily coordinate access to annual sales data in the access layer of the data warehouse, daily sales data in the staging layer of the data warehouse, and new data in the source layer database. Users can have an operational view combined with historic view. These services work by metadata mapping, where the federation layer takes the metadata from the ETL component.

2.2 TAXONOMY

The NBD-PWG Definitions and Taxonomy Subgroup developed a hierarchy of Reference Architecture components. Additional taxonomy details are presented in the *NBDIF: Volume 2, Taxonomy*. The NIST Big Data Reference Architecture Taxonomy outlines potential actors for the seven roles developed by the NBD-PWG Definition and Taxonomy Subgroup.

2.3 USE CASES

A consensus list of Big Data requirements across stakeholders was developed by the NBD-PWG Use Cases and Requirements Subgroup. The development of requirements included gathering and understanding various use cases from the nine diversified areas, or application domains, listed below.

- Government Operation;
- Commercial;
- Defense;
- Healthcare and Life Sciences;
- Deep Learning and Social Media;
- The Ecosystem for Research;
- Astronomy and Physics;
- Earth, Environmental, and Polar Science; and
- Energy.

Participants in the NBD-PWG Use Cases and Requirements Subgroup and other interested parties supplied publicly available information for various Big Data architecture examples from the nine application domains, which developed organically from the 51 use cases collected by the Subgroup.

After collection, processing, and review of the use cases, requirements within seven Big Data characteristic categories were extracted from the individual use cases. Requirements are the challenges limiting further use of Big Data. The complete list of requirements extracted from the use cases is presented in the document *NBDIF: Volume 3, Use Cases and General Requirements*.

The use case specific requirements were then aggregated to produce high-level general requirements, within seven characteristic categories. The seven categories are as follows:

- ***Data source requirements*** (relating to data size, format, rate of growth, at rest, etc.);
- ***Data transformation provider*** (i.e., data fusion, analytics);
- ***Capabilities provider*** (i.e., software tools, platform tools, hardware resources such as storage and networking);
- ***Data consumer*** (i.e., processed results in text, table, visual, and other formats);
- ***Security and privacy***;
- ***Life cycle management*** (i.e., curation, conversion, quality check, pre-analytic processing); and
- ***Other requirements***.

The general requirements, created to be vendor-neutral and technology-agnostic, are organized into seven categories in Table 1 below.

Table 1: Seven Requirements Categories and General Requirements

DATA SOURCE REQUIREMENTS (DSR)	
DSR-1	Needs to support reliable real-time, asynchronous, streaming, and batch processing to collect data from centralized, distributed, and cloud data sources, sensors, or instruments.
DSR-2	Needs to support slow, bursty, and high-throughput data transmission between data sources and computing clusters.
DSR-3	Needs to support diversified data content ranging from structured and unstructured text, document, graph, web, geospatial, compressed, timed, spatial, multimedia, simulation, and instrumental data.
TRANSFORMATION PROVIDER REQUIREMENTS (TPR)	
TPR-1	Needs to support diversified compute-intensive, analytic processing, and machine learning techniques.
TPR-2	Needs to support batch and real-time analytic processing.
TPR-3	Needs to support processing large diversified data content and modeling.
TPR-4	Needs to support processing data in motion (e.g., streaming, fetching new content, tracking).
CAPABILITY PROVIDER REQUIREMENTS (CPR)	
CPR-1	Needs to support legacy and advanced software packages (software).
CPR-2	Needs to support legacy and advanced computing platforms (platform).
CPR-3	Needs to support legacy and advanced distributed computing clusters, co-processors, input output processing (infrastructure).
CPR-4	Needs to support elastic data transmission (networking).
CPR-5	Needs to support legacy, large, and advanced distributed data storage (storage).
CPR-6	Needs to support legacy and advanced executable programming: applications, tools, utilities, and libraries (software).
DATA CONSUMER REQUIREMENTS (DCR)	
DCR-1	Needs to support fast searches (~0.1 seconds) from processed data with high relevancy, accuracy, and recall.
DCR-2	Needs to support diversified output file formats for visualization, rendering, and reporting.
DCR-3	Needs to support visual layout for results presentation.
DCR-4	Needs to support rich user interface for access using browser, visualization tools.
DCR-5	Needs to support high-resolution, multidimensional layer of data visualization.
DCR-6	Needs to support streaming results to clients.
SECURITY AND PRIVACY REQUIREMENTS (SPR)	
SPR-1	Needs to protect and preserve security and privacy of sensitive data.
SPR-2	Needs to support sandbox, access control, and multilevel, policy-driven authentication on protected data.
LIFE CYCLE MANAGEMENT REQUIREMENTS (LMR)	
LMR-1	Needs to support data quality curation including preprocessing, data clustering, classification, reduction, and format transformation.
LMR-2	Needs to support dynamic updates on data, user profiles, and links.
LMR-3	Needs to support data life cycle and long-term preservation policy, including data provenance.
LMR-4	Needs to support data validation.
LMR-5	Needs to support human annotation for data validation.
LMR-6	Needs to support prevention of data loss or corruption.
LMR-7	Needs to support multisite archives.
LMR-8	Needs to support persistent identifier and data traceability.
LMR-9	Needs to support standardizing, aggregating, and normalizing data from disparate sources.

OTHER REQUIREMENTS (OR)	
OR-1	Needs to support rich user interface from mobile platforms to access processed results.
OR-2	Needs to support performance monitoring on analytic processing from mobile platforms.
OR-3	Needs to support rich visual content search and rendering from mobile platforms.
OR-4	Needs to support mobile device data acquisition.
OR-5	Needs to support security across mobile devices.

The preceding requirements were also mapped to 51 use cases in the *NBDIF: Volume 3, Use Cases and General Requirements* document, as shown below in Figure 2.

Table D-4: Data Consumer	
General Requirements	
1. Needs to support fast searches (~0.1 seconds) from processed data with high relevancy, accuracy, and high recall.	Applies to 4 use cases: M0148, M0160, M0165, M0176
2. Needs to support diversified output file formats for visualization, rendering, and reporting.	Applies to 16 use cases: M0078, M0089, M0090, M0157, M0c161, M0164, M0164, M0165, M0166, M0166, M0167, M0167, M0174, M0177, M0213, M0214
3. Needs to support visual layouts for results presentation.	Applies to 2 use cases: M0165, M0167
4. Needs to support rich user interfaces for access using browsers, visualization tools.	Applies to 1 use cases: M0089, M0127, M0157, M0160, M0162, M0167, M0167, M0183, M0184, M0188, M0190
5. Needs to support a high-resolution multi-dimension layer of data visualization.	Applies to 21 use cases: M0129, M0155, M0155, M0158, M0161, M0162, M0171, M0172, M0173, M0177, M0179, M0182, M0185, M018c6, M0188, M0191, M0213, M0214, M02c15, M0219, M0222
6. Needs to support streaming results to clients.	Applies to 1 use case: M0164

Figure 2: Example of the Data Consumer Requirements Mapped to 51 Use Cases.

The requirements and use cases provide a foundation for development of the NBDRA, and the standards mapping and tracking exercises described in Section 3. Additional information about the Use Cases and Requirements Subgroup, use case collection, analysis of the use cases, and generation of the use case requirements are presented in the *NBDIF: Volume 3, Use Cases and General Requirements* document.

2.4 SECURITY AND PRIVACY

Security and privacy measures for Big Data involve a different approach than traditional systems. Big Data is increasingly stored on public cloud infrastructure built by various hardware, operating systems, and analytical software. Traditional security approaches usually addressed small-scale systems holding static data on firewalled and semi-isolated networks. The surge in streaming cloud technology necessitates extremely rapid responses to security issues and threats [17]. Security and privacy considerations are a fundamental aspect of Big Data and affect all components of the NBDRA. This comprehensive influence is depicted in Figure 2 by the grey rectangle marked "Security and Privacy" surrounding all the Reference Architecture components.

At a minimum, a Big Data Reference Architecture will provide verifiable compliance with both governance, risk management, and compliance (GRC) and confidentiality, integrity, and availability (CIA) policies, standards, and best practices. Additional information on the processes and outcomes of the NBD PWG Security and Privacy Subgroup are presented in *NBDIF: Volume 4, Security and Privacy*.

The NBD-PWG Security and Privacy Subgroup began this effort by identifying ways that security and privacy in Big Data projects can be different from traditional implementations. While not all concepts apply all the time, the following observations were considered representative of a larger set of differences:

1. Big Data projects often encompass heterogeneous components in which a single security scheme has not been designed from the outset.
2. Most security and privacy methods have been designed for batch or online transaction processing systems. Big Data projects increasingly involve one or more streamed data sources that are used in conjunction with data at rest, creating unique security and privacy scenarios.
3. The use of multiple Big Data sources not originally intended to be used together can compromise privacy, security, or both. Approaches to de-identify personally identifiable information (PII) that were satisfactory prior to Big Data may no longer be adequate, while alternative approaches to protecting privacy are made feasible. Although de-identification techniques can apply to data from single sources as well, the prospect of unanticipated consequences from the fusion of multiple datasets exacerbates the risk of compromising privacy.
4. A huge increase in the number of sensor streams for the Internet of Things (e.g., smart medical devices, smart cities, smart homes) creates vulnerabilities in the Internet connectivity of the devices, in the transport, and in the eventual aggregation.
5. Certain types of data thought to be too big for analysis, such as geospatial and video imaging, will become commodity Big Data sources. These uses were not anticipated and/or may not have implemented security and privacy measures.
6. Issues of veracity, context, provenance, and jurisdiction are greatly magnified in Big Data. Multiple organizations, stakeholders, legal entities, governments, and an increasing amount of citizens will find data about themselves included in Big Data analytics.
7. Volatility is significant because Big Data scenarios envision that data is permanent by default. Security is a fast-moving field with multiple attack vectors and countermeasures. Data may be preserved beyond the lifetime of the security measures designed to protect it.
8. Data and code can more readily be shared across organizations, but many standards presume management practices that are managed inside a single organizational framework. A related observation is that smaller firms, subject to fewer regulations or lacking mature governance practices, can create valuable Big Data systems.

The NBD-PWG security and privacy fabric sets forth three levels of voluntary conformance. The levels offer incremental increases in security and privacy Big Data risk mitigation. The approach taken unifies both models of information security—such as presented in the NIST Cybersecurity Framework—with domain-specific models.

The three-level technique reveals important differences between domains as disparate as astronomy and health care; some aspects must be addressed in ways particular to the specialization and by specialists. Recognizing that security can be viewed as a reduction in risk or harm caused, not necessarily a 100% assurance, the NBDPWG security fabric is framed as a safety- and harm-reduction framework. It recognizes the importance of scalability to Big Data by emphasizing the increased importance of modeling and simulation. The fabric adapts key concepts from safety engineering, such as the Material Data Safety Sheet (29 CFR 1910 1200(g)), for tracing risk associated with "toxic" privacy data.

The framework offers a smooth transition to broader adoption of time-dependent, attribute-based access controls (NIST SP 800-162, SP 1800-3) and processes in support of the NIST Risk Management Framework (NIST 800-37 Rev 2).

The security fabric outlined here envisions an infrastructure of monitoring, simulation, analytics and governance that leverages Big Data to such an extent where data volumes could well exceed those of the systems they were designed to make safe.

2.5 REFERENCE ARCHITECTURE SURVEY

The NBD-PWG Reference Architecture Subgroup conducted the reference architecture survey to advance understanding of the operational intricacies in Big Data and to serve as a tool for developing system-specific architectures using a common reference framework. The Subgroup surveyed currently published Big Data platforms by leading companies or individuals supporting the Big Data framework and analyzed the collected material. This effort revealed a remarkable consistency between Big Data architectures. Survey details, methodology, and conclusions are reported in *NBDIF: Volume 5, Architectures White Paper Survey*.

2.6 REFERENCE ARCHITECTURE

2.6.1 OVERVIEW

The goal of the NBD-PWG Reference Architecture Subgroup is to develop a Big Data open Reference Architecture that facilitates the understanding of the operational intricacies in Big Data. It does not represent the system architecture of a specific Big Data system, but rather is a tool for describing, discussing, and developing system-specific architectures using a common framework of reference. The Reference Architecture achieves this by providing a generic high-level conceptual model that is an effective tool for discussing the requirements, structures, and operations inherent to Big Data. The model is not tied to any specific vendor products, services, or reference implementation, nor does it define prescriptive solutions that inhibit innovation.

The design of the NBDRA does not address the following:

- Detailed specifications for any organization's operational systems;
- Detailed specifications of information exchanges or services; and
- Recommendations or standards for integration of infrastructure products.

Building on the work from other subgroups, the NBD-PWG Reference Architecture Subgroup evaluated the general requirements formed from the use cases, evaluated the Big Data Taxonomy, performed a reference architecture survey, and developed the NBDRA conceptual model. The *NBDIF: Volume 3, Use Cases and General Requirements* document contains details of the Subgroup's work. The use case characterization categories (from *NBDIF: Volume 3, Use Cases and General Requirements*) are listed below on the left and were used as input in the development of the NBDRA. Some use case characterization categories were renamed for use in the NBDRA. Table 2 maps the earlier use case terms directly to NBDRA components and fabrics.

Table 2: Mapping Use Case Characterization Categories to Reference Architecture Components and Fabrics

Use Case Characterization Categories		Reference Architecture Components and Fabrics
Data sources	→	Data Provider
Data transformation	→	Big Data Application Provider
Capabilities	→	Big Data Framework Provider
Data consumer	→	Data Consumer
Security and privacy	→	Security and Privacy Fabric
Life cycle management	→	System Orchestrator; Management Fabric
Other requirements	→	To all components and fabrics

2.6.2 NBDRA Conceptual Model

As discussed in Section 2, the NBD-PWG Reference Architecture Subgroup used a variety of inputs from other NBD-PWG subgroups in developing a vendor-neutral, technology- and infrastructure-agnostic conceptual model of Big Data architecture. This conceptual model, the NBDRA, is shown in Figure 2 and represents a Big Data system composed of five logical functional components connected by interoperability interfaces (i.e., services). Two fabrics envelop the components, representing the interwoven nature of management and security and privacy with all five of the components. The NBDRA is intended to enable system engineers, data scientists, software developers, data architects, and senior decision makers to develop solutions to issues that require diverse approaches due to convergence of Big Data characteristics within an interoperable Big Data ecosystem. It provides a framework to support a variety of business environments, including tightly integrated enterprise systems and loosely coupled vertical industries, by enhancing understanding of how Big Data complements and differs from existing analytics, business intelligence, databases, and systems.

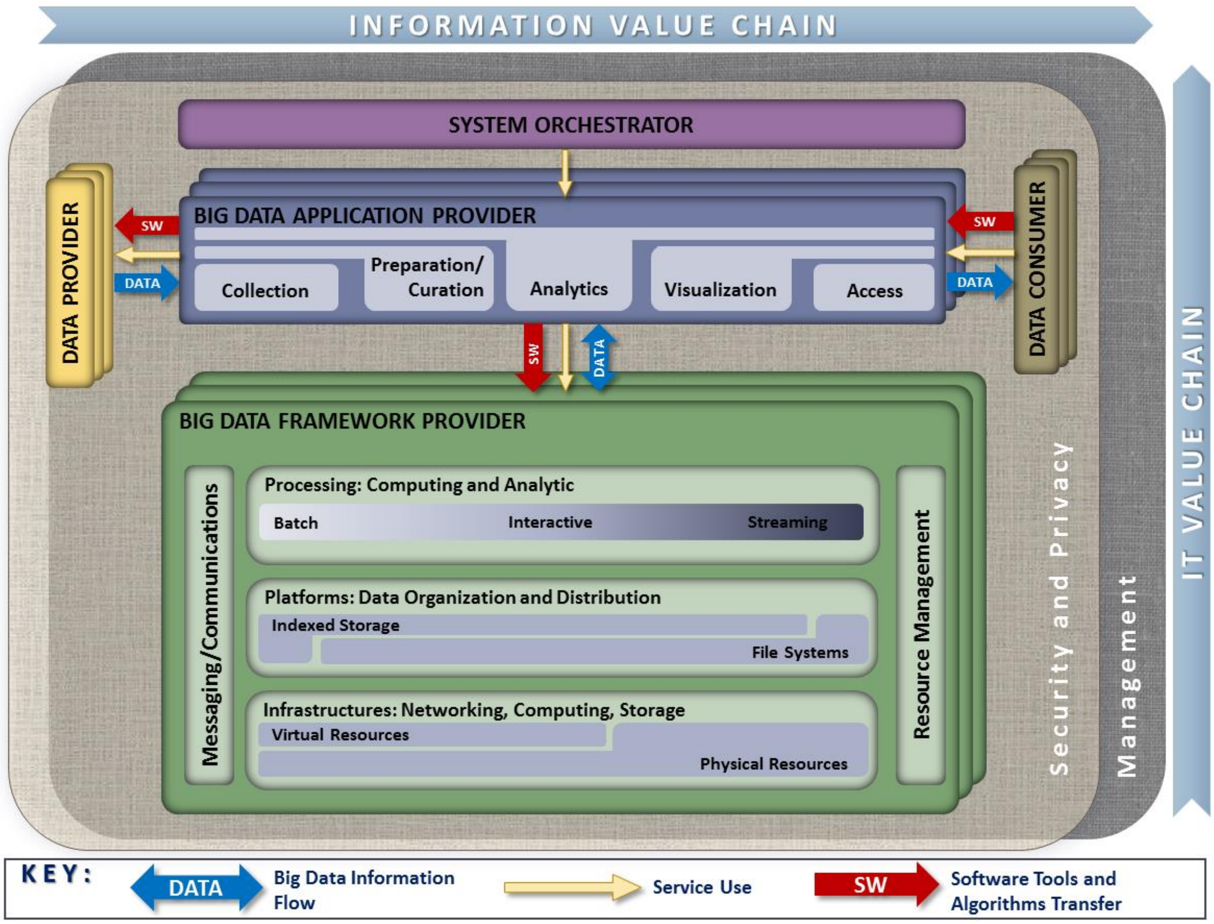

Figure 3: NIST Big Data Reference Architecture (NBDRA) Conceptual Model

Note: None of the terminology or diagrams in these documents is intended to be normative or to imply any business or deployment model. The terms *provider* and *consumer* as used are descriptive of general roles and are meant to be informative in nature.

The NBDRA is organized around five major roles and multiple sub-roles aligned along two axes representing the two Big Data value chains: Information Value (horizontal axis) and Information Technology (IT; vertical axis). Along the information axis, the value is created by data collection, integration, analysis, and applying the results following the value chain. Along the IT axis, the value is created by providing networking, infrastructure, platforms, application tools, and other IT services for hosting of and operating the Big Data in support of required data applications. At the intersection of both axes is the Big Data Application Provider role, indicating that data analytics and its implementation provide the value to Big Data stakeholders in both value chains.

The five main NBDRA roles, shown in Figure 2, represent different technical roles that exist in every Big Data system. These roles are the following:

- System Orchestrator,
- Data Provider,
- Big Data Application Provider,
- Big Data Framework Provider, and
- Data Consumer.

Traditional siloed behavior of these players contributes to locking in old ways of thinking. Changing behavior from inward looking (i.e., meeting own needs) to outward looking (i.e., meeting others' needs) may help solve this phenomenon of siloed behavior. For example:

1. Data providers should provide findable, accessible, interoperable, and reusable (FAIR), analysis ready, open data that are useable for unknown third parties.
2. Data platform providers should develop platforms that meet the needs of both data providers (contributing open data) and data consumers.
3. Data application providers should develop web applications that meet the needs of both data providers (contributing open data) and data consumers.
4. Data consumers can join in participatory design by creating use cases for the developing data applications.
5. Data orchestrators should create use cases that provide insight to data providers and data consumers about the data life cycle.

The two fabric roles shown in Figure 2 encompassing the five main roles are:

- Management, and
- Security and Privacy.

These two fabrics provide services and functionality to the five main roles in the areas specific to Big Data and are crucial to any Big Data solution. The **DATA** arrows in Figure 2 show the flow of data between the system's main roles. Data flows between the roles either physically (i.e., by value) or by providing its location and the means to access it (i.e., by reference). The **SW** arrows show transfer of software tools for processing of Big Data *in situ*. The **Service Use** arrows represent software programmable interfaces. While the main focus of the NBDRA is to represent the run-time environment, all three types of communications or transactions can happen in the configuration phase as well. Manual agreements (e.g., service-level agreements) and human interactions that may exist throughout the system are not shown in the NBDRA. The roles in the Big Data ecosystem perform activities and are implemented via functional components.

In system development, actors and roles have the same relationship as in the movies, but system development actors can represent individuals, organizations, software, or hardware. According to the Big Data taxonomy, a single actor can play multiple roles, and multiple actors can play the same role. The NBDRA does not specify the business boundaries between the participating actors or stakeholders, so the roles can either reside within the same business entity or can be implemented by different business entities. Therefore, the NBDRA is applicable to a variety of business environments, from tightly integrated enterprise systems to loosely coupled vertical industries that rely on the cooperation of independent stakeholders. As a result, the notion of internal versus external functional components or roles does not apply to the NBDRA. However, for a specific use case, once the roles are associated with specific business stakeholders, the functional components would be considered as internal or external, subject to the use case's point of view.

The NBDRA does support the representation of stacking or chaining of Big Data systems. For example, a Data Consumer of one system could serve as a Data Provider to the next system down the stack or chain. The NBDRA is discussed in detail in the *NBDIF: Volume 6, Reference Architecture.* The Security and Privacy Fabric, and surrounding issues, are discussed in the NBDIF: *Volume 4, Security and Privacy*. From the data provider's viewpoint, getting ready for Big Data is discussed in *NBDIF: Volume 9, Adoption and Modernization*. Once established, the definitions and Reference Architecture formed the basis for evaluation of existing standards to meet the unique needs of Big Data and evaluation of existing implementations and practices as candidates for new Big Data-related standards. In the first case, existing efforts may address standards gaps by either expanding or adding to the existing standard to accommodate Big Data characteristics or developing Big Data unique profiles within the framework of the existing standards.

3 ANALYZING BIG DATA STANDARDS

Big Data has generated interest in a wide variety of multi-stakeholder, collaborative organizations. Some of the most involved to date have been organizations participating in the de jure standards process, industry consortia, and open source organizations. These organizations may operate differently and focus on different aspects, but they all have a stake in Big Data.

Integrating additional Big Data initiatives with ongoing collaborative efforts is a key to success. Identifying which collaborative initiative efforts address architectural requirements and which requirements are not currently being addressed is a starting point for building future multi-stakeholder collaborative efforts. Collaborative initiatives include, but are not limited to the following:

- Subcommittees and working groups of American National Standards Institute (ANSI);
- Accredited standards development organizations (SDOs; the de jure standards process);
- Industry consortia;
- Reference implementations; and
- Open source implementations.

Some of the leading SDOs and industry consortia working on Big Data-related standards include the following:

- IEC—International Electrotechnical Commission, http://www.iec.ch/;
- IEEE—Institute of Electrical and Electronics Engineers, https://www.ieee.org/index.html, de jure standards process;
- IETF—Internet Engineering Task Force, https://www.ietf.org/;
- INCITS—International Committee for Information Technology Standards, http://www.incits.org/, de jure standards process;
- ISO—International Organization for Standardization, http://www.iso.org/iso/home.html, de jure standards process;
- OASIS—Organization for the Advancement of Structured Information Standards, https://www.oasis-open.org/, Industry consortium;
- OGC®—Open Geospatial Consortium, http://www.opengeospatial.org/, Industry consortium;
- OGF—Open Grid Forum, https://www.ogf.org/ogf/doku.php, Industry consortium; and
- W3C—World Wide Web Consortium, http://www.w3.org/, Industry consortium.

In addition, the Research Data Alliance (RDA) https://www.rd-alliance.org/ develops relevant guidelines. RDA is a community-driven organization of experts, launched in 2013 by the European Commission, the United States National Science Foundation, National Institute of Standards and Technology, and the Australian Government's Department of Innovation with the goal of building the social and technical infrastructure to enable open sharing of data.

The organizations and initiatives referenced in this document do not form an exhaustive list. More standards efforts addressing additional segments of the Big Data mosaic may exist.

There are many government organizations that publish standards relative to their specific problem areas. The U.S. Department of Defense alone maintains hundreds of standards. Many of these are based on other standards (e.g., ISO, IEEE, ANSI) and could be applicable to the Big Data problem space.

However, a fair, comprehensive review of these standards would exceed the available document preparation time and may not be of interest to most of the audience for this report. Readers interested in domains covered by government organizations and standards are encouraged to review available standards for applicability to their specific needs.

Open source implementations are providing useful new technologies used either directly or as the basis for commercially supported products. These open source implementations are not just individual products. As actual implementations of technologies are proven, reference implementations will evolve based on some of these community accepted efforts. Organizations will likely need to integrate an ecosystem of multiple products to accomplish their goals. Because of the ecosystem complexity and the difficulty of fairly and exhaustively reviewing open source implementations, many such implementations are not included in this section. However, it should be noted that those implementations often evolve to become the de facto reference implementations for many technologies.

Standards can be of different types, and serve different functions along the lifecycle of technology diffusion. Semantic standards that enable the reduction of information or transaction costs, are applicable to the basic research stages of technology development. Measurement and testing standards, are applicable to the transition point where basic research advances to the applied research stage. Interface standards that enable interoperability between *components*, are applicable to the stage when applied research advances into experimental development. Compatibility and quality standards which enable economies of scale, interoperability between *products*, and reduced risk, are applicable to the ultimate diffusion of technology [13].

Several pathways exist for the development of standards. The trajectory of this pathway is influenced by the SDO through which the standard is created and the domain to which the standard applies. For example, *ANSI/ Standards Engineering Society (SES) 1:2012, Recommended Practice for the Designation and Organization of Standards*, and *SES 2:2011, Model Procedure for the Development of Standards*, set forth documentation on how a standard itself must be defined.

Standards often evolve from requirements for certain capabilities. By definition, established de jure standards endorsed by official organizations, such as NIST, are ratified through structured procedures prior to the standard receiving a formal stamp of approval from the organization. The pathway from de jure standard to ratified standard often starts with a written deliverable that is given a *Draft Recommendation* status. If approved, the proposed standard then receives a higher *Recommendation* status, and continues up the ladder to a final status of *Standard* or perhaps *International Standard*.

Standards may also evolve from implementation of best practices and approaches which are proven against real-world applications, or from theory that is tuned to reflect additional variables and conditions uncovered during implementation. In contrast to formal standards that go through an approval process to meet the definition of ANSI/SES 1:2012, there are a range of technologies and procedures that have achieved a level of adoption in industry to become the conventional design in practice or method for practice, though they have not received formal endorsement from an official standards body. These dominant in-practice methods are often referred to as market-driven or de facto standards.

De facto standards may be developed and maintained in a variety of different ways. In *proprietary* environments, a single company will develop and maintain ownership of a de facto standard, in many cases allowing for others to make use of it. In some cases, this type of standard is later released from proprietary control into the *Open Source* environment.

The open source environment also develops and maintains technologies of its own creation, while providing platforms for decentralized peer production and oversight on the quality of, and access to, the open source products.

The phase of development prior to the de facto standard is referred to as specifications. “When a tentative solution appears to have merit, a detailed written spec must be documented so that it can be implemented and codified [18]”. Specifications must ultimately go through testing and pilot projects before reaching the next phases of adoption.

At the most immature end of the standards spectrum are the emerging technologies that are the result of R&D. Here the technologies are the direct result of attempts to identify solutions to particular problems.

Since specifications and de facto standards can be very important to the development of Big Data systems, this volume attempts to include the most important standards and classify them appropriately.

Big Data efforts require a certain level of data quality. For example, metadata quality can be met using ISO 2709 (Implemented as MARC21) and thesaurus or ontology quality can be met by using ISO 25964. In the case of Big Data, ANSI/NISO (National Information Standards Organization) has a number of relevant standards; many of these standards are also ISO Standards under ISO Technical Committee (TC) 46, which are Information and Documentation Standards. NISO and ISO TC 46 are working on addressing the requirements for Big Data standards through several committees and work groups.

U.S. federal departments and agencies are directed to use voluntary consensus standards developed by voluntary consensus standards bodies:

> *"'Voluntary consensus standards body' is a type of association, organization, or technical society that plans, develops, establishes, or coordinates voluntary consensus standards using a voluntary consensus standards development process that includes the following attributes or elements:*
>
> *i. Openness: The procedures or processes used are open to interested parties. Such parties are provided meaningful opportunities to participate in standards development on a nondiscriminatory basis. The procedures or processes for participating in standards development and for developing the standard are transparent.*
> *ii. Balance: The standards development process should be balanced. Specifically, there should be meaningful involvement from a broad range of parties, with no single interest dominating the decision making.*
> *iii. Due process: Due process shall include documented and publicly available policies and procedures, adequate notice of meetings and standards development, sufficient time to review drafts and prepare views and objections, access to views and objections of other participants, and a fair and impartial process for resolving conflicting views.*
> *iv. Appeals process: An appeals process shall be available for the impartial handling of procedural appeals.*
> v. *Consensus: Consensus is defined as general agreement, but not necessarily unanimity. During the development of consensus, comments and objections are considered using fair, impartial, open, and transparent processes* [19]".

3.1 EXISTING STANDARDS / THE CURRENT STATE

The NBD-PWG embarked on an effort to compile a list of standards that are applicable to Big Data with a goal to assemble Big Data-related standards that may apply to a large number of Big Data implementations across several domains. The enormity of the task hinders the inclusion of every standard that could apply to every Big Data implementation. Appendix B presents a partial list of existing standards, with descriptions, from the above listed organizations that are relevant to Big Data and the NBDRA. Appendix C and Appendix D describe different aspects of the same list of standards presented in Appendix B. Determining the relevance of standards to the Big Data domain is challenging since almost all standards in some way deal with data. Whether a standard is relevant to Big Data is generally determined by the impact of Big Data characteristics (i.e., volume, velocity, variety, and variability) on the standard or, more generally, by the scalability of the standard to accommodate those characteristics. A standard may also be applicable to Big Data depending on the extent to which that standard helps to address one or more of the Big Data characteristics. Finally, a number of standards are also very domain- or problem-specific and, while they deal with or address Big Data, they support a very specific functional domain. Developing even a marginally comprehensive list of such standards would require a massive undertaking involving subject matter experts in each potential problem domain, which is currently beyond the scope of the NBD-PWG. In selecting standards to include in Appendix B, C, and D, the NBD-PWG focused on standards that met the following criteria:

- Facilitate interfaces between NBDRA components;
- Facilitate the handling of data with one or more Big Data characteristics; and
- Represent a fundamental function needing to be implemented by one or more NBDRA components.

Appendix B, C, and D represent a table of potentially applicable standards from a portion of contributing organizations working in the Big Data domain. As most standards represent some form of interface between components, the standards table in Appendix C indicates whether the NBDRA component would be an Implementer or User of the standard. For the purposes of this table, the following definitions were used for Implementer and User.

> ***Implementer****: A component is an implementer of a standard if it provides services based on the standard (e.g., a service that accepts Structured Query Language (SQL) commands would be an implementer of that standard) or encodes or presents data based on that standard.*

> ***User:*** *A component is a user of a standard if it interfaces to a service via the standard or if it accepts/consumes/decodes data represented by the standard.*

While the above definitions provide a reasonable basis for some standards, the difference between Implementer and User may be negligible or nonexistent. Appendix B contains the entire Big Data standards catalog collected by the NBD-PWG to date.

3.1.1 MAPPING EXISTING STANDARDS TO SPECIFIC REQUIREMENTS

During Stage 2 work the NBD-PWG began mapping the general requirements, which are summarized in Table 1, to applicable standards, with the goal of simply aggregating potentially applicable standards to the general requirement statements from Volume 3. The requirements-to-standards matrix in Table 3 illustrates the mapping of the DCR category of general requirements to existing standards. The approach links a requirement with related standards by setting the requirement code and description in the same row as related standards descriptions and standards codes.

Table 3: Data Consumer Requirements-to-Standards Matrix

Requirement	Requirement Description	Standards Description	Standard / Specification
DCR-1	Fast search, with high precision and recall.		
DCR-2	Support diversified output file formats for visualization, rendering and reporting.	KML: data vector format. Image format: RPF raster product format based specification, derived from ADRG and other sources.	(1) KML. (2) Military Spec CADRG. (2) NITF; GeoTiff.
DCR-3	Support visual layout of results for presentation.	Suggested charts and tables for various purposes.	International Business Communication Standards (IBCS) notation; related: ACRL

Requirement	Requirement Description	Standards Description	Standard / Specification
DCR-4	Support for rich user interfaces for access using browsers, and visualization tools.	1. Programming interface represents documents as objects.	(1) Document object model (DOM). (2) CSS selector, JSON, Canvas, SVG. (3) WebRTC
DCR-5	Support high resolution Multidimensional visualization Layer	ISO 13606 compliant interface generator visualizes multidimensional (medical) concepts.	BMC Visualization [20]
DCR-6	Streaming results to clients	(1) Defines file format and real time transport protocol (RTP) payload format for video and audio.	(1) IEEE 1857.2, 1857.3. (2) DASH. (3) Daala.

One example of a simple, rich user interface which may satisfy basic requirements of DCR-4 can be seen on the Smart Electric Power Alliance website. The interactive online catalog of standards for the smart grid employs modern navigation features to represent standards in an interactive webpage accessible to browsers. The Catalog of Standards Navigation Tool provides hover overlays and effective dialog boxes (i.e., divs) for exploring "the domains, subdomains, components and standards of the Smart Grid." The website can be accessed at www.gridstandardsmap.com .

The work undertaken in Table 3 is representative of work that should be continued with the other six General Requirements categories (i.e., TPR, CPR, DCR, SPR, LMR, and OR) listed in Table 1 and explained fully in the *NBDIF: Volume 3, Use Cases and General Requirements*.

Incomplete population of the DCR requirements in Table 3 reflect only the unfinished nature of this work, as of the date of this publication, due to limited available resources of the NBD-PWG, and should not be interpreted as standards gaps in the technology landscape. As more fields of the resulting matrix are completed, denser areas in the matrix will provide a visual summary of where an abundance of standards exist, and most importantly, sparsely populated areas will highlight gaps in the standards catalog as of the date of publication.

Potentially, the fields in Table 3 would become heavily populated with standards that are not specifically mapped to particular requirements, exposing the need for a more detailed activity that links specific requirements to standards. One way to accomplish this is to have standards mapped to the sub-component sections of use cases, as described in the next section, 3.1.2.

3.1.2 MAPPING EXISTING STANDARDS TO SPECIFIC USE CASE SUBCOMPONENTS

Similar to the standards to requirements mapping in Section 3.1.1, use cases were also mapped to standards (Table 4). Three use cases were initially selected for mapping and further analysis in Versions 2 and 3 of this document. These use cases were selected from the 51 Version 1 use cases collected by the NBD-PWG and documented in the *NBDIF: Volume 3, Use Cases and Requirements*.

The mapping illustrates the intersection of a domain-specific use case with standards related to Big Data. In addition, the mapping provides a visual summary of the areas where standards exist and most importantly, highlights gaps in the standards catalog as of the date of publication of this document. The aim of the use case to standards mapping is to link a use case number and description with codes and descriptions for standards related to the use case, providing a more detailed mapping than that in Table 3.

Table 4: General Mapping of Select Use Cases to Standards

Use Case Number and Type	Use Case Description	Standards Description	Standard / Specification
8: Commercial	Web search	For XML, XIRQL works independent of schema, to identify attributes; integrates with ranking computations; selects specific elements for retrieval.	W3C99 (XPpath), W3C03 (XQuery), full-text, elixir, XIRQL, XXL. INEX.
13: Defense	Geospatial Analysis and Visualization	netCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. Compressed ARC Digitized Raster Graphics is a general purpose product comprising computer readable digital map and chart images.	CF-netCDF3, Opensearch_EO, MapML, KML, CADRG
15: Defense	Intelligence data processing	Collection of formats, specifies Geo and Time extensions, supports sharing of search results	OGC OpenSearch, WCPS

In addition to mapping standards that relate to the overall subject of a use case, specific portions of the original use cases (i.e., the categories of Current Solutions, Data Science, and Gaps) were mapped to standards.

The detailed mapping provides additional granularity in the view of domain-specific standards. The data from the Current Solutions, Data Science, and Gaps categories, along with the subcategory data, was extracted from the raw use cases in the *NBDIF: Volume 3, Use Cases and Requirements* document. This data was tabulated with a column for standards related to each subcategory. The process of use case subcategory mapping was initiated with two use cases, Use Case 8 and Use Case 15, as evidenced below.

USE CASE 8: WEB SEARCH

Table 5 demonstrates mapping of related standards to the selected sub-components of the web search use case.

Table 5: Excerpt from Use Case Document M0165—Detailed Mapping to Standards

Information from Use Case 8			Related Standards / Specification
Category	**Subcategory**	**Use Case Data**	
Current Solutions	Compute system	Large cloud	
	Storage	Inverted index	
	Networking	External most important	SRU, SRW, CQL, Z39.50; OAI PMH; Sparql (de facto), representational state transfer (REST), Href;
	Software		Spark (de facto)
Data Science (collection, curation, analysis, action)	Veracity	Main hubs, authorities	
	Visualization	Page layout is critical. Technical elements inside a website affect content delivery.	IBCS Notation
	Data Quality		SRank
	Data Types	Plain text ASCII format; binary image formats; sound files; video. HTML.	Txt; gif, jpeg and png; wav; mpeg. UTF-8.
	Data Analytics	Crawl, preprocess, index, rank, cluster, recommend. Crawling / collection: connection elements including mentions from other sites.	Sitemap.xml, responsive design (spec), browser automation and APIs
Gaps		Links to user profiles, social data. Access to deep web.	Schema.org

USE CASE 13: LARGE SCALE GEOSPATIAL ANALYSIS AND VISUALIZATION

Table 6 demonstrates mapping of related standards to the selected sub-components of the geospatial analysis and visualization use case.

Table 6: Excerpt from Use Case Document M0213---Detailed Mapping to Standards

Information from Use Case 13			Related Standards / Specification
Category	**Subcategory**	**Use Case Data**	
Current Solutions	Compute	System should support visualization components on handhelds and laptops	
	Storage	Visualization components use local disk and flash ram	
	Network	Displays are operating at the end of low bandwidth wireless networks	CF-netCDF3 Data Model Extension standard. Maps to ISO 19123 coverage schema. http://www.opengeospatial.org/docs/is

Information from Use Case 13			Related Standards / Specification
Category	**Subcategory**	**Use Case Data**	
	Software		Opensearch-EO specification: Browser usable descriptions of search filter parameters for response support and query formulation. Also defines a “default response encoding based on Atom 1.0 XML (RD.22). (http://www.opengeospatial.org/standards/requests/172) OGC WCPS standard: spatio-temporal data cube analytics language for server-side evaluation
Data Science	Veracity		
	Visualization	Spatial data is not natively accessible by browsers.	MapML Testbed 14 (T14): http://www.opengeospatial.org/blog/2772 MapML conveys map semantics similar to hypertext. Four threads: EOC, Next Gen, MoPoQ, and CITE: http://www.opengeospatial.org/blog/2773
	Data Quality	The typical problem is visualization implying quality / accuracy not available in the original data. All data should include metadata for accuracy or circular error probability.	
	Data Types	Imagery: (various formats NITF, GeoTiff, CADRG). Vector: (various formats shape files, KML, text streams: Object types include points, lines, areas, polylines, circles, ellipses.	KML is one of several 3D modeling standards dealing with cartographic, geometric and semantic viewpoints in an earth-browser, for indoor navigation. KML provides a single language for first responders to navigate indoor facilities. Others include CityGML and IFC. KML leverages OpenGIS.
Gaps	Geospatial data requires unique approaches to indexing and distributed analysis.		Note: There has been some work with in DoD related to this problem set. Specifically, the DCGS-A standard cloud (DSC) stores, indexes, and analyzes some Big Data sources. Many issues still remain with visualization however.

USE CASE 15: DEFENSE INTELLIGENCE DATA PROCESSING AND ANALYSIS

Table 7 demonstrates mapping of related standards to the selected sub-components of the defense intelligence data processing use case.

Table 7: Excerpt from Use Case Document M0215—Detailed Mapping to Standards

Information from Use Case 15			Related Standards / Specification
Category	**Subcategory**	**Use Case Data**	
Current Solutions	Compute system	Fixed and deployed computing clusters ranging from 1000s of nodes to 10s of nodes.	
	Storage	Up to 100s of PBs for edge and fixed site clusters. Dismounted soldiers have at most 100s of GBs.	
	Networking	Connectivity to forward edge is limited and often high latency and with packet loss. Remote communications may be Satellite or limited to radio frequency / Line of sight radio.	
	Software	Currently baseline leverages: 1. Distributed storage 2. Search 3. Natural Language Processing (NLP) 4. Deployment and security 5. Storm (spec) 6. Custom applications and visualization tools	1: Distributed File Systems (HDFS; de facto) 2. Opensearch - EO 3: GrAF (spec), 4: Puppet (spec),
Data Science (collection, curation, analysis, action)	Veracity (Robustness Issues, semantics)	1. Data provenance (e.g., tracking of all transfers and transformations) must be tracked over the life of the data. 2. Determining the veracity of "soft" data sources (generally human generated) is a critical requirement.	1: ISO/IEC 19763, W3C Provenance
	Visualization	Primary visualizations will be Geospatial overlays and network diagrams. Volume amounts might be millions of points on the map and thousands of nodes in the network diagram.	
	Data Quality (syntax)	Data Quality for sensor-generated data (image quality, sig/noise) is generally known and good. Unstructured or "captured" data quality varies significantly and frequently cannot be controlled.	
	Data Types	Imagery, Video, Text, Digital documents of all types, Audio, Digital signal data.	
	Data Analytics	1. Near real time Alerts based on patterns and baseline changes. 2. Link Analysis 3. Geospatial Analysis 4. Text Analytics (sentiment, entity extraction, etc.)	3: GeoSPARQL, 4: SAML 2.0,
Gaps		1. Big (or even moderate size data) over tactical networks 2. Data currently exists in disparate silos which must be accessible through a semantically integrated data space. 3. Most critical data is either unstructured or imagery/video which requires significant processing to extract entities and information.	1. 2: SAML 2.0, W3C OWL 2, 3:

3.2 GAPS IN STANDARDS

Section 3.1 provides a structure for identification of relevant existing Big Data standards, and the current state of the landscape. A number of technology developments are considered to be of significant importance and are expected to have sizeable impacts heading into the next decade. Any list of *important* items will obviously not satisfy every community member; however, the list of gaps in Big Data standardization provided in this section describe broad areas that may span across the range of interest to SDOs, consortia, and readers of this document.

The list below, which was produced through earlier work by an ISO/IEC Joint Technical Committee 1 (JTC1) Study Group on Big Data, served as a potential guide to ISO in their establishment of Big Data standards activities [21]. The 16 potential Big Data standardization gaps identified by the study group, described broad areas that were of interest to this community. These gaps in standardization activities related to Big Data in the following areas:

1. Big Data use cases, definitions, vocabulary, and reference architectures (e.g., system, data, platforms, online/offline);
2. Specifications and standardization of metadata including data provenance;
3. Application models (e.g., batch, streaming);
4. Query languages including non-relational queries to support diverse data types (e.g., XML, Resource Description Framework [RDF], JSON, multimedia) and Big Data operations (i.e., matrix operations);
5. Domain-specific languages;
6. Semantics of eventual consistency;
7. Advanced network protocols for efficient data transfer;
8. General and domain-specific ontologies and taxonomies for describing data semantics including interoperation between ontologies;
9. Big Data security and privacy access controls;
10. Remote, distributed, and federated analytics (taking the analytics to the data) including data and processing resource discovery and data mining;
11. Data sharing and exchange;
12. Data storage (i.e., memory storage system, distributed file system, data warehouse);
13. Human consumption of the results of Big Data analysis (i.e., visualization);
14. Energy measurement for Big Data;
15. Interface between relational (i.e., SQL) and non-relational (i.e., not only [or no] Structured Query Language [NoSQL]) data stores; and
16. Big Data quality and veracity description and management (includes master data management).

The NBD-PWG Standards Roadmap Subgroup began a more in-depth examination of the topics listed above, to identify potential opportunities to close the gaps in standards. Version 2 of this volume explored four of the 16 gaps identified above in further detail.

- Gap 2: Specifications of metadata
- Gap 4: Non-relational database query, search and information retrieval (IR)
- Gap 10: Analytics
- Gap 11: Data sharing and exchange

Version 3 of this volume explored four more of the 16 gaps in further detail.

- Gap 12: Data storage.
- Gap 13: Human consumption of the results of Big Data analysis (i.e., visualization).
- Gap 15: Interface between relational and non-relational data stores.
- Gap 16: Big data quality and veracity description and management.

All of the issues related to the gaps in standards are important. Due to constraints in available resources, some of the gaps have not been addressed by the completion of version 3. Additional resources will be required to continue this work. The following table shows the current disposition of the 16 original gaps. Security and privacy issues are addressed in the *NBDIF: Volume 4, Security and Privacy* document.

3.3 UPDATES TO THE LIST OF GAPS

3.3.1 OUT OF SCOPE GAPS

In the process of investigating the original 16 gaps, the Subgroup found it appropriate to classify Gap #3 and Gap #14 as outside the scope of this document, which focuses on interoperability. Gaps #3 and #14 describe Big Data issues but are not really interoperability scenarios. For example, Gap #3 real time processing improves wait-times for access to data, and improves exception handling or error handling, but these are not interoperability issues.

3.3.2 ADDITION OF NEW GAPS

In the process of investigating the original 16 gaps, the subgroup found it appropriate to add new gaps to the list. Four such gaps have been added. Additionally, recent progress in other NBDIF volumes may have alignment with the gaps in this volume. In the process of updating the list of gaps from version 2 and considering new gaps, the NBD-PWG has attempted to keep the focus on gap closures that can be expected to provide a large impact in terms of enabling greater economic, financial, or work productivity improvements; and also to keep the focus as closely as possible on core areas of Big Data interoperability. Internet bandwidth, for example, can affect NLP, data mining, distributed storage, cloud computing, and query performance, but whether the network connectivity is a core Big Data interoperability issue is debatable. Impact can be expected to change over time. What is described as having little impact today may be expected to have moderate or higher impact any number of years into the future. According to a BCG+MIT report, the financial services industry is one which has a high potential to take advantage of improvements in analytics technologies, in the near future.

In an effort to keep this document relevant to the current state of the market, no more than five years into the future is considered, concentrating on the time period prior to 2023. The following list of four gaps have been added to the original list of 16.

NEW GAPS FOR VERSION 3:

Gap 17 Blending data, faster integration of external data sources (n5); transformation, integration running on distributed storage and computing systems. Issues surround data formats (e.g., log formats, JSON)

Gap 18 Real time synchronization for data quality. Integration. Introduced in Section 2.

Gap 19 Joining traditional and big architectures. Interoperability. Legacy systems are inflexible.

Gap 20 Single version of the truth; drivers of Trust. Introduced in Section 4.2.2.

3.3.3 SCHEME FOR ORDERING GAPS

Earlier versions of the Standards Roadmap presented the 16 gaps in an unordered list. For purposes of better readability, the subgroup set out to order the earlier list. Below is a proposed grouping of the gaps, shaped by functional groups discussed in the early work of the NBD-PWG, detailed in document M0054. Additional work on the hierarchy could be completed, namely, to articulate that integration can be viewed as a higher level parent of interoperability. The proposed scheme for ordering the gaps is as follows:

(CENTRAL TO) INTEROPERABILITY

Gap 2 Specifications and standardization of metadata including data provenance

Gap 13 Human consumption of the results of Big Data analysis (e.g., visualization)

Gap 8 General and domain-specific ontologies and taxonomies for describing data semantics including interoperation between ontologies

Gap 5 Domain-specific languages

Gap 4 Query languages including non-relational queries to support diverse data types (e.g., XML, Resource Description Framework (RDF), JSON, multimedia) and Big Data operations (i.e., matrix operations)

Gap 15 Interface between relational (i.e., SQL) and non-relational (i.e., NoSQL) data stores

Gap 19 Joining traditional and big architectures

QUALITY AND DATA INTEGRITY

Gap 6 Semantics of eventual consistency

Gap 12 Data storage (e.g., memory storage system, distributed file system, data warehouse)

Gap 20 Trust

MANAGEMENT, ADMINISTRATION, RESOURCE PLANNING AND COSTS

Gap 1 Big Data use cases, definitions, vocabulary, and reference architectures (e.g., system, data, platforms, online/offline);

Gap 3 Application models (e.g., batch, streaming);

Gap 16 Big Data quality and veracity description and management (includes master data management [MDM]).

Gap 14 Energy measurement for Big Data;

DEPLOYMENT, OPTIMIZATION

Gap 10 Remote, distributed, and federated analytics (taking the analytics to the data) including data and processing resource discovery and data mining

Gap 11 Data sharing and exchange

Gap 7 Advanced network protocols for efficient data transfer

SECURITY

Gap 9 Big Data security and privacy access controls (See *NBDIF: Volume 4, Security and Privacy)*

4 GAP DISCUSSION POINTS

4.1 GAPS CENTRAL TO INTEROPERABILITY

Interoperability can be decomposed down to two main types of capabilities: connectivity and translation.

4.1.1 STANDARDS GAP 2: SPECIFICATION OF METADATA

Metadata is one of the most significant of the Big Data problems. Metadata is the only way of finding items, yet 80% of data lakes are not applying metadata effectively [14]. Metadata layers are ways for lesser technical users to interact with data mining systems. Metadata layers also provide a means for bridging data stored in different locations, such as on premise and in the cloud. A definition and concept description of metadata is provided in the *NBDIF: Volume 1, Definitions* document.

Metadata issues have been addressed in ISO 2709-ANSI/NISO Z39.2 (implemented as MARC21) and cover not only metadata format but, using the related Anglo-American Cataloging Rules, content and input guidance for using the standard.

The metadata management field appears to now be converging with master data management (MDM) and somewhat also with analytics. Metadata management facilitates access control and governance, change management, and reduces complexity and the scope of change management, with the top use case likely to be data governance [14]. Demand for innovation in the areas of automating search capabilities such as semantic enrichment during load and inclusion of expert / community enrichment / crowd governance, and machine learning, remains strong and promises to continue.

Organizations that have existing metadata management systems will need to match any new metadata systems to the existing system, paying special attention to federation and integration issues. Organizations initiating new use cases or projects have much more latitude to investigate a range of potential solutions. Note that there is not always a need for a separate system; metadata could be inline markup of ICD-10 codes for example, in a physician's report.

Perhaps a more attainable goal for standards development will be to strive for standards for supporting interoperability beyond the defining of ontologies, or XML, where investment of labor concentrates on the semantic mappings instead of syntactic mapping in smaller blocks that can be put together to form a larger picture, for example, to define conveying the semantics of who, what, where, and when of an event and translation of an individual user's terms (in order to create a module that can then be mapped to another standard).

Metadata is a pervasive requirement for integration programs and new standards for managing relationships between data sources; and automated discovery of metadata will be key to future Big Data projects. Recently, new technologies have emerged that analyze music, images, or video and generate metadata automatically. In the linked data community, efforts continue toward developing metadata techniques that automate construction of knowledge graphs and enable the inclusion of crowdsourced information.

There are currently approximately 30 Metadata standards listed on the Digital Curation Centre (DCC) website (http://www.dcc.ac.uk/). Some of the lesser-known standards of a more horizontal data integration type are listed below:

- Data Package, version 1.0.0-beta.17 (a specification) released March of 2016;
- Observ-OM, integrated search. LGPLv3 Open Source licensed;
- PREMIS, independent serialization, preservation actor information;
- PROV, provenance information;
- QuDEx, agnostic formatting;
- Statistical Data and Metadata Exchange (SDMX), specification 2.1 last amended May 2012; and
- Text Encoding and Interchange (TEI), varieties and modules for text encoding.

Metadata is really the central control mechanism for all integration activity. Metadata can track changes and rules application, across enrichment, movement, parsing, cleaning, auditing, profiling, lineage services, transformation, matching, and scheduling services. For successful systems, it must be pervasive throughout. If data integration is important, Metadata needs to be integrated too, so that users can bring in new metadata from other datasets, or share metadata with other systems.

A primary use case is the data lake. Data lakes are also not environments where events over time are easily correlated with historical analysis. One solution attempts to resolve both of these problems by combining metadata services with clearly defined business taxonomy. The metadata services are a centralized, common storage framework consisting of three types of metadata: business metadata such as business definitions; operational metadata such as when operations occurred, which includes logs and audit trails; and technical metadata such as column names, data types, and table names.

The taxonomy framework consists of a mechanism for organizing metadata vocabulary into folder and sub-folder type data classification hierarchies; and a mechanism for definition and assignment of business vocabulary tags to columns in physical data stores. The hierarchies serve to reduce duplications and inconsistencies and increase visibility into workflows that are otherwise missing in data lake systems. For privacy and security compliance functions, the tags enlist a notification trigger which alert administrators or users whenever tagged data has been accessed or used.

For lineage functions, log events are combined with logical workflow models at runtime, allowing for more than simple forensic validation and confidence of compliance requirements. Metatag rules can prevent unification violations incurred by the joining of separate, otherwise compliant datasets.

The host of Satellite data lake components required to make data lake ecosystems useful each operate out of unique interfaces. The combination metadata and taxonomy solution sits atop the data lake, in a single interface that oversees the whole system, enabling improved governance, and integration and exchange (import / export) of metadata. Data steward tasks such as tagging can be separated from policy protection tasks, allowing for dual role operation or specialization of human resources. A prominent open source query tool is a key component. The connector for the query tool includes a capability to track structured query activity. REST based APIs provide data classification navigation paths that are pre-defined.

4.1.2 STANDARDS GAP 4: NON-RELATIONAL DATABASE QUERY, SEARCH AND INFORMATION RETRIEVAL (IR)

Search serves as a function for interfacing with data in both retrieval and analysis use cases. As a non-relational database query function, search introduces a promise of *self-service* extraction capability over multiple sources of unstructured (and structured) Big Data in multiple internal and external locations. Search has capability to integrate with technologies for accepting natural language, and also for finding and analyzing patterns, statistics, and providing conceptual summary and consumable visual formats.

This is an area where the ISO 23950 [22] / ANSI/NISO Z39.50 [23] approach could help. The NISO standard "defines a client/server based service and protocol for Information Retrieval. It specifies procedures and formats for a client to search a database provided by a server, retrieve database records, and perform related information retrieval functions. The protocol addresses communication between information retrieval applications at the client and server; it does not address interaction between the client and the end-user. [23]"

In that we live in an age where one web search engine maintains the mindshare of the American public, it is important to clearly differentiate between the use of search as a data analysis method and the use of search for IR. Significantly different challenges are faced by business users undertaking search for information retrieval activities as opposed to using a search function for analysis of data that resides within an organization's storage repositories. In web search, casual users are familiar with the experience of the technology, namely, instant query expansion, ranking of results, rich snippets and knowledge graph containers. Casual users are also familiar with standard file folder functionality for organizing documents and information in personal computers. For large enterprises and organizations needing search functionality over their own documents, deeper challenges persist and are driving significant demand for enterprise-grade solutions. In that these enterprise requirements may be unfamiliar to small business users; some clarification on the differences are described below.

Web Search

Current web search engines provide a substantial service to citizens but have been identified as applying bias over how and what search results are delivered back to the user. The surrender of control that citizens willingly trade in exchange for the use of free web search services is widely accepted as a worthwhile tradeoff for the user; however, future technologies promise even more value for the citizens who will search across the rapidly expanding scale of the world wide web. The notable case in point is commonly referred to as the semantic web.

Current semantic approaches to searching almost all require content indexing as a measure for controlling the enormous corpus of documents that reside online. In attempting to tackle this problem of enormity of scale via automation of content indexing, solutions for the semantic web have proven to be difficult to program, meaning that the persistent challenges for development of a semantic web continue to delay its development.

Two promising approaches for developing the semantic web are ontologies and linked data technologies; however, neither approach has proven to be a complete solution. Standard Ontological alternatives, OWL and RDF, which would benefit from the addition of linked data, suffer from an inability to effectively use linked data technology. Reciprocally, linked data technologies suffer from the inability to effectively use ontologies. Not apparent to developers is how standards in these areas would be an asset to the concept of an all-encompassing semantic web, or how they can be integrated to improve retrieval over that scale of data.

Using Search for Enterprise Data Analysis

A steady increase in the belief that logical search systems are the superior method for information retrieval on data at rest can be seen in the market. Generally speaking, analytic search indexes can be constructed more quickly than natural language processing (NLP) search systems, meanwhile NLP technologies requiring semi-supervision can have unacceptable error rates. Currently, Contextual Query Language (CQL) [24], declarative logic programming languages, and RDF [25] query languages are aligned with the native storage formats of the Big Data platforms. Often only one language is supported, however multi-model platforms may support more than one language. Some query languages are managed by standards organizations, while other query languages are defacto standards "in-the-wild".

With the exception of multi-model databases, any product's underlying technology will likely be document, metadata, or numerically focused, but not all three. Architecturally speaking, indexing is the centerpiece, while metadata provides context, and machine learning can provide enrichment. Markup metadata can also provide document enrichment, with tags such as ICD-10 codes for example.

The age of Big Data has applied a downward pressure on the use of standard indexes, which are good for small queries but have three issues: (1) they cause slow loading; (2) ad hoc queries, for the most part, require advance column indexing; and (3) the constant updating that is required to maintain indexes quickly becomes prohibitively expensive. One open source search technology provides an incremental indexing technique that solves some part of this problem. Another technology provides capability to perform indexing upon either ingest or changing of the data, through the use of a built-in universal index. After indexing, query planning functionalities are of primary importance.

Generally speaking, access and IR functions will remain areas of continual work in progress. In some cases, silo architectures for data are a necessary condition for running an organization, with legal and security reasons being the most obvious. There are several Big Data technologies that support RBAC with cell / element / field level security which can alleviate the need to have different silos for legal and security reasons.

Other technologies are emerging in the area of ‘federated search.’ The main barrier to effective federated search functionality is the difficulty in merging results into relevancy ranking algorithms. Proprietary, patented access methods are also a barrier to building connectors required for true federated search.

Ultimately, system speed is always constrained by the slowest component. The future goal for many communities and enterprises in this area is the development of unified information access solutions (i.e., UIMA). Unified indexing and multi-model databases present an alternative to challenges in federated search.

Incredibly valuable external data is underused in most search implementations because of the lack of an appropriate architecture. Frameworks that separate content acquisition from content processing by putting a data buffer (a big copy of the data) between them have the capability to provide potential solutions to this problem. With this approach, data can be gathered without the requirement to immediately make content processing decisions; content processing decisions can be settled later. Documents would have to be *pre-joined* when they are processed for indexing, and large, mathematically challenging algorithms for relevancy and complex search security requirements (such as encryption) could be run separately at index time. With such a framework, search could potentially become superior to traditional structured query languages for online analytical processing (OLAP) and data warehousing. Search systems can be faster than query languages, more powerful, scalable, and schema free. Records can be output in XML and JSON and then loaded into a search engine. Fields can be mapped as needed.

Tensions remain between any given search system’s functional power and its ease of use. The concept of Discovery, as the term is understood in the IR domain, was initially relegated to the limited functionality of filtering (facets) in a sidebar. The facets have historically been loaded when a search system returned a result set. Emerging technologies are focusing on supplementing the faceted search user’s experience. Content Representation standards were initially relied upon in the Wide Area Information Servers (WAIS) system but newer systems must contend with the fact that there are now hundreds of data formats. In response, open source technologies promise power and flexibility to customize, but this promise comes with a high price tag of being either technically demanding and requiring skilled staff to setup and operate, or requiring a third party to maintain.

Standards for content processing are still needed to enable compatibility with normalizing techniques, records merging formats, external taxonomies or semantic resources, regular expression, and/or use of metadata for supporting interface navigation functionality.

Standards for describing relationships between different data sources, and standards for maintaining metadata context relationships will have substantial impact. Semantic platforms to enhance information discovery and data integration applications may provide solutions in this area; RDF and ontology mapping seem to be the front runners in the race to provide semantic uniformity. RDF graphs are leading the way for visualization, and ontologies have become accepted methods for descriptions of elements. While the cross-walking of taxonomies, and ontologies is still a long way off, technologic advances in this area should be helpful for the success of data analytics across silos, and the semantic web.

QUERY LANGUAGES TO SUPPORT BIG DATA CUBE OPERATIONS

Two main data model extensions beyond the relational model are graph and array databases. They offer declarative graph and array queries which are optimizable on the server side, paralleling the traditional advantages the relational model offers on sets (of records or tables). Matrix operations are a special case of general multi-dimensional tensor operations, which in turn fall under the category of Linear Algebra.

Array databases [26] offer a multi-dimensional “data cube” view [27], which is suitable for spatio-temporal sensor, time series image simulation, and statistics data. Data models like the OGC/ISO Coverage Implementation Schema adds support for regular and irregular space/time grids. Declarative array query languages like domain-independent ISO SQL/MDA [28], [29] and geo-specific OGC WCPS [30] have been demonstrated to be highly optimizable, parallelizable, and amenable to distributed processing, up to location-transparent data center federations [27].

In a service setup the question arises how such data extraction and processing functionality can be offered. Offering programming access (e.g., in python) to a server is: (1) inconvenient and restricting access to coding experts; and (2) insecure as there is no way to check whether malicious code is coming in for execution in the server. As a result, database research has rejuvenated the concept of query languages where a query describes what the result should look like and not how it gets computed.

Such “declarative” (as opposed to “procedural”) languages allow for much more compact expressions without programming ballast. Further, the database server needs to find a strategy to evaluate such a query there is ample room for optimization, parallelization, and other techniques which make query processing faster than a naïve algorithm. Finally, such languages are “safe in evaluation” meaning that every query is guaranteed, by construction of the language, to finalize after a finite number of steps. Therefore, query languages represent the preferred way of giving flexible data retrieval, filtering, and processing access to data via the Internet.

Volume 2 includes a brief discussion on four types of data structures: sets, hierarchies, graphs, and arrays. Each of those data categories have appropriate available languages for query. For Sets, classical SQL maintains a long standing dominance. In Hierarchies, data can be queried through XPath / XQuery; Graph queries can apply languages such as Cypher. For Arrays, SQL/MDA (Multi-Dimensional Arrays) provides for domain-independent array queries and Open Geospatial Consortium (OGC) Web Coverage Processing Service (WCPS) serves as a geo-oriented spatio-temporal data cube query language.

4.1.3 STANDARDS GAP 11: DATA SHARING AND EXCHANGE

The overarching goal of data sharing and exchange is to maximize access to data across heterogeneous repositories while still adhering to protect confidentiality and personal privacy. The objective is to improve the ability to locate and access digital assets such digital data, software, and publications while enabling proper long-term stewardship of these assets by optimizing archival functionality, and (where appropriate) leveraging existing institutional repositories, public and academic archives, as well as community and discipline-based repositories of scientific and technical data, software, and publications.

From the new global Internet, to Big Data economy opportunities in Internet of Things, smart cities, and other emerging technical and market trends, it is critical to have a standard data infrastructure for Big Data that is scalable and can apply the FAIR (Findability, Accessibility, Interoperability, and Reusability) data principle between heterogeneous datasets from various domains without worrying about data source and structure.

A very important component as part of the standard data infrastructure is the definition of new Persistent Identifier (PID) types. PIDs such as Digital Object Identifiers (DOIs) are already widely used on the Internet as durable, long-lasting references to digital objects such as publications or datasets.

An obvious application of PIDs in this context is to use them to store a digital object's location and state information and other complex core metadata. In this way, the new PID types can serve to hold a combination of administration, specialized, and/or extension metadata. Other functional information, such as the properties and state of a repository or the types of access protocols it supports, can also be stored in these higher layers of PIDs. Assigning PIDs to static datasets is straightforward. However, datasets that are updated with corrections or new data, or that are subsets of a larger dataset present a challenge.

Mechanisms for making evolving data citable have been proposed by the Research Data Alliance data citation working group and others [31], [32], [33], [34], [35], [36]. Because the PIDs are themselves digital objects, they can be stored in specialized repositories, similar to metadata registries that can also expose services to digital object users and search portals. In this role, the PID types and the registries that manage them can be viewed as an abstraction layer in the system architecture, and could be implemented as middleware designed to optimize federated search, assist with access control, and speed the generation of cross-repository inventories. This setting can enable data integration/mashup among heterogeneous datasets from diversified domain repositories and make data discoverable, accessible, and usable through a machine-readable and actionable standard data infrastructure.

Organizations wishing to publish open data will find that there are certain legal constraints and licensing standards to be conscious of; data may not necessarily be 100% *Open* in every sense of the word. There are, in fact, varying degrees to the openness of data; various licensing standards present a spectrum of licensing options, where each type allows for slightly differing levels of accommodations. Some licensing standards, including the Open Government License, provide truly open standards for data sharing. Use of Creative Commons licenses is increasingly common (https://creativecommons.org/licenses/).

Organizations wishing to publish open data must also be aware that there are some situations where the risks of having the data open, outweigh the benefits; and where certain licensing options are not appropriate, including situations when interoperability with other datasets is negatively affected. See the next sub section Inter-Organization Data Sharing for additional discussion.

DATA MIGRATION

Migration and consolidation are fundamental activities in legacy data processing. An opportunity is presented in data migration scenarios to ensure data quality and, additionally, to clean and enrich the data to improve it during the migration process. A common-sense approach here is to apply business rules during the migration project that leverage metadata to synchronize new data and update it as it is offloaded to a new system. Data providers are the best actors to ensure metadata is good prior to migration, and that data is still accurate after it is consolidated in its new location.

Previously, loading was a high cost step because data had to be structured first. The beauty of the non-relational architecture was the ease of loading, and the schema on write or schema on query capability of the ELT model which offered a less complicated data transformation workflow, thereby reducing the high cost of loading and migrating data. Multi-model databases technologies also promise a reduction in the level of migration that is required for data processing.

INTER-ORGANIZATIONAL DATA SHARING

The financial services, banking, and insurance (FSBI) sector has been an industry at the forefront of Big Data adoption. As such, FSBI can provide information about the challenges related to integration of external data sources. Due to the heterogeneous nature of external data, many resources are required for integrating external data with an organization's internal systems. In FSBI, the number of sources can also be high, creating a second dimension of difficulty.

By some reports, the lack of integration with internal systems is the largest organizational challenge when attempting to leverage external data sources [37]. Many web portals and interfaces for external data sources do not provide APIs or capabilities that support automated integration, causing a situation where the majority of organizations currently relinquish expensive resources on manual coding methods to solve this problem. Of special interest in this area are designs offering conversion of SOAP protocol to REST (REpresentational State Transfer) protocol.

Aside from the expense, another problem with the hard coding methods is the resulting system inflexibility. Regardless of those challenges, the penalty for not integrating with external sources is even higher in the FSBI industry, where the issues of error and data quality are significant. The benefits of data validation and data integrity ultimately outweigh the costs.

4.1.4 STANDARDS GAP 13: VISUALIZATION, FOR HUMAN CONSUMPTION OF THE RESULTS OF DATA ANALYSIS

A key to a successful Big Data or data science analysis is providing the results in a human interpretable format either through statistical results (e.g., p-values, Mean Squared Error) or through visualization. Visualization of data is a very effective technique for human understanding. Data and results are typically displayed in condensed statistical graphics such as scatter plots, bar charts, histograms, box plots, and other graphics.

The increase in the amounts of real-time data that are typically generated in Big Data analysis will require increasingly complex visualizations for human interpretation. Sensor data, for example, coming from Internet of Things (IoT) applications is driving use cases for real-time processing and visualization of data and results, which require Big Data tools.

Another use case which deals with the human consumption of the results of Big Data analysis is cyber analytics. The key to cyber analytics is to flag certain data for additional inspection by a competent cybersecurity professional; but the amount of network traffic which needs filtering and algorithms applied in real time is staggering for even small networks.

Usage of data visualization in 2D or 3D renderings is also increasing. Capable of depicting both temporal and spatial changes in data, these advanced renderings are used for the visualization of transport containers, air traffic, ships, cars, people or other movements across the globe in a real-time fashion and may require Big Data tools.

Projections on the total global amount of data available for analysis and visualization involve exponential growth over the foreseeable future. Effective visualization and human consumption of this explosion of data will need associated standards.

4.1.5 STANDARDS GAP 15: INTERFACE BETWEEN RELATIONAL AND NON-RELATIONAL DATA STORES

Every interface consists of four essential facets, which each interface must deal with (i.e., tempo, quantity, content, and packaging) or in other words the inputs, the outputs, how long the processing takes, and how much material (in this case data) is delivered to the end user.

In many situations, unstructured data constitutes the majority of data available for analysis. In reality, most so called unstructured data does have some type of structure, because all data has some pattern that can be used to parse and process the data. However, there is an increase in the need for tools to help parse the data or to enforce a traditional relational database management system (RDBMS) structure to the data. While non-relational style databases are easier to scale than schema based relational databases, the lack of ACID (Atomicity, Consistency, Isolation, and Durability) can affect accuracy and confidence in Big Data analyses.

Algorithms which can parse “unstructured data” into a RDBMS format are useful in creating ACID compliant data sources and a more mature ecosystem for analysis. Although non-relational databases offer advantages in scalability and are often better suited for the extreme volumes of data associated with Big Data analyses, many applications require the traditional RDBMS format to use legacy tools and analysis approaches. Therefore, the need for “hybrid” approaches between non-relational and relational style data storage is greatly increasing and associated standards for these approaches are necessary. Two main data model extensions beyond the relational model are graph and array databases.

4.2 GAPS IN QUALITY AND DATA INTEGRITY

4.2.1 STANDARDS GAP 12: DATA STORAGE

Some of the most key concerns in Big Data storage in general include the consistency of the data, scalability of the systems, and dealing with the heterogeneity of data and sources. Capabilities for dealing with challenges of data heterogeneity are less mature.

4.2.1.1 Big Data Storage Problems and Solutions in Data Clustering

Many solutions for Big Data storage problems optimize the storage resource in some kind of way, to facilitate either the pre-processing or processing of the data. One such approach attempts to use data clustering techniques in order to optimize computing resources. Solutions using data clustering (Table 8) to resolve storage and compute problems are not necessarily concerned with the integrity of data**.**

In dealing with problems of optimizing storage for high dimensional data, Hierarchical Agglomerative Clustering (HAC) mechanisms have capabilities for supporting efficient storage of data, by reducing the demand requirements for space. HAC methods have capabilities which implement data clustering methods for dataset decomposition, merging columns to compress data, and efficient access to the data. HAC techniques include approaches for finding optimal decomposition by locating a partition strategy that minimizes storage space requirements, prior to the pre-processing stage. HAC methods can address availability, scalability, resource optimization, and data velocity aspects of data storage problems [38].

K-means algorithms have the capability to work along with MapReduce processing and assist by partitioning and merging of data subsets which results in a form of compression similar to HAC methods, thus reducing main memory requirements.

General purpose K-means algorithms allow for the handling of larger datasets by reducing data cluster centroid distances; and the scalability aspect of applicable storage problems, but HAC methods for resolving heterogeneity, availability, or velocity aspects of Big Data are not fully mature or standardized [39].

The class of Artificial Bee Colony (ABC) algorithms have demonstrated functionality for resolving accessibility aspects of later stage processing execution problems through the use of storage partitioning, but features for dealing with heterogeneity, or velocity of data with respect to latency during processing tasks are also immature [40].

Table 8: Clustering Solutions

Challenge	Solution Research	Solution Description
Storage of high dimensional data	Hierarchical Agglomerative Clustering (HAC)	A variant of the class of HAC mechanisms, SOHAC, is described for optimizing storage resources. SOHAC research covers a method which addresses many aspects for storing high dimensional data, but not those of heterogeneity.
Prediction difficulty	K means algorithm	K-Means has been used to address scalability and resource optimization problems but not velocity, heterogeneity, or availability issues.
Processing latency	Artificial Bee Colony algorithm	ABCs may resolve availability and resource optimization problems, but not velocity, heterogeneity, or scalability issues.

4.2.1.2 Data Storage Problems and Solutions in Data Indexing

Query optimization is a difficult function in Big Data use cases. Technology implementers can expect to make tradeoffs between lookup capabilities and throughput capabilities.

In the quest for solutions to challenges in data indexing, Composite Tree and Fuzzy Logic methods were each found to resolve many aspects of slow retrieval and other problems; however, few solutions were responsive to data velocity aspects of storage problems. Note that data heterogeneity does not necessarily affect the process of an indexing mechanism, therefore indexing systems do not necessarily need to design for these features. The details of the methods reviewed in this indexing section are so overtly technical, as to make consumable summary of the performance descriptions especially difficult. Given the limits of resources for Version 3, the overview of these capabilities (in Table 9 and in the text) is obtusely generalized.

Regarding latency in data retrieval, the capabilities which deploy Composite Tree methods described here have shown promise in fast retrieval of data for all aspects of the problem, except for challenges in velocity of data. Variations of K-nearest-neighbor methods promise resolution of many aspects of Big Data, but mature Composite Tree methods for fast moving data unresolved are especially immature [41].

When applied to problems in indexing, the class of support vector machines (SVMs) promise the capability to perform cost effective entity extraction from video at rest. SVMs are able to reduce search filter ‘ball’ sizes, which is the area within a radius of points surrounding the center of the group of documents relevant to the query. SVM variants for resolution of heterogeneity, velocity, resource optimization, or scalability aspects of Big Data indexing problems are areas in search of solutions [42].

Table 9: Indexing Solutions

Challenge	Solution Research	Solution Description
Latency in data retrieval	Composite Tree / composite quantization for nearest neighbor	Speeds query response on data at rest and streams. Resolves all but the issue of index loading times on data with velocity.
Result accuracy	Support vector machines (SVM)	SVMs can reduce data dimensionality (+) and allow for data to fit in-memory. SVMs resolve availability and integrity issues.
Index updating	Fuzzy Logic	Updates quickly and remains lean by deleting old index images.

4.2.1.3 Big Data Storage Problems and Solutions in Data Replication

In data replication functions, integrity of the data is critical. Replication of data is also an important function for supporting access. Traditional data replication technologies are mature; several commercial products have offered replication solutions for regular data, for years.

Fuzzy Logic, ABC, and dynamic data replication (D2RS) techniques (Table 10) have been described as solutions to availability, integrity, resource optimization, and scalability aspects of Big Data management problems. However, descriptions of techniques addressing heterogeneity and velocity are much less common.

Fuzzy Logic techniques work on the premise that there are degrees of membership for entities or objects within categories, and to what extent an entity or object belongs to or deviates from a category or 'set", is extremely useful for classification tasks and if-then rules. Fuzzy Logic techniques have the capability to improve data consistency problems in data replication functions [43].

Table 10: Replication Solutions

Challenge	Solution Research	Solution Description
Data inconsistency	Fuzzy Logic	Data replication technique which uses fuzzy logic to select a peer.
Coordinating storage with computing environments	Artificial bee colony (ABC) algorithm	ABC addresses job scheduling issues in grid environments.
Site access speed limits	Dynamic data replication (D2RS) [44]	

4.2.2 STANDARDS GAP 16: BIG DATA QUALITY AND VERACITY DESCRIPTION AND MANAGEMENT

Amidst most of the use cases for data integration is an absolute need to maximize data quality, which helps to ensure accuracy. Data must be cleaned to provide quality and accurate analytic outputs. This is especially true in cases where automated integration systems are in play. Applying data quality processes too late is more costly than adherence to quality processes early on because poor quality gets amplified downstream. In many ways, quality is the top concern [16].

A need exists for semantic auditing and metrics to determine authority of data. Traditionally, 'trusted data' is a data state validated across multiple authoritative sources. However, trusted data assumes no semantic variation, an important aspect in distributed systems. Trusted data also lacks hard metrics which denote trust. For example, multiple authoritative sources may be inconsistent leading to degradation of trust in its value(s). Another example is having a sufficient quorum of sources to establish trust. Another use case is rate of change at authoritative sources.

For values which assume a common semantic, automated methods may be applied to derive trust levels. However, there is no such technology available to measure semantics progression. One example is programmers hijacking a field in a data structure to represent some other data not available in the message structure (e.g., Over the Horizon Targeting and Over the Horizon Gold message specifications). If a data field is subjugated for a unique application, documentation or communication of the resulting semantic alterations are often left to channels of tribal knowledge, and not formally or appropriately recorded. The only way to discover this type of shift is through manual audits.

Similar to the need for code vulnerability audit tools, a semantic audit tool is required. Unfortunately, semantic audit tools still cannot combat users entering semantically shifted data into a form which are ultimately wrongly certified by authoritative processes. One of the causes of these problems is standard data structures which do not keep pace with application semantics. Another cause is applications which do not keep pace with user's needs. Yet another cause is application developers who do not fully understand the entire specification, for example the Common Message Format.

In an ontology structure for formalizing semantics, denotative and connotative solutions work together, and ultimately support a saliency map for associating data sources with applications. The saliency map communicates and transfers information from one domain to another; automated intention detection is possible; and decoding of context is possible. In this structure connotative spaces fit into denotative spaces and provide meaning, and meanings lead to trust.

Trusted data is a quality benchmark signifying the degree of confidence a consumer has for acquired data products. Acquired data products may be incorporated into newly created data products and actionable intelligence which includes insights, decision making, and knowledge building. Means and metrics to gauge trust are often arbitrary and vary between industries, applications, and technological domains.

Established trust has broad impacts on Big Data analytic processes as well as results created by the analytics. Trusted data benefits consumers by shrinking production costs and accelerating the delivery of analytic results. Valuable analytic processing directed at validating data's veracity can be reduced or even eliminated.

Traditionally, trust levels are established through personal relationships, mechanisms of apportionment, and transitive means exemplified by authoritative mandate and Friend of a Friend (FOAF) relationships, as well as homegrown, ad-hoc methods. Trust signifiers themselves are commonly informal and often acquired by transitive means.

Concepts of trust within Big Data domains are often sourced from a cyber-security world application; validating the identities of remotely communicating participants. Identity establishment commonly includes one or more methods of exchanging information, and the use of third-party, authoritative entities. Big Data applications with increasing emphasis on analytic correctness and liability concerns are expanding the definition of trust past concepts found in cyber-security identity applications. Ideas surrounding data trust are shifting expectations toward data quality.

To ensure Big Data product trust, formal and standardized practices are required to consistently improve results and reduce potential civil and possibly criminal liabilities. Formalization should include applying best practices source for other areas within the computing industry as well as other mature industries. Trust practices could require application profiles identifying significant measures and quality levels, hard and soft metrics, and measure supporting processes and technologies to enable a proof driven infrastructure guaranteeing and certifying product quality.

DATA CLEANING

Cleaning is the keystone for data quality. The tasks of data cleaning and preparation to make the data useable have been cited as consuming the majority of time and expense in data analysis. A 2016 CrowdFlower survey of data scientists found that 19% of their time was spent on finding data, and 60% of their time was spent on cleaning and organizing the data [45]. In the 2017 CrowdFlower survey, "access to quality data" was cited as the number one roadblock to success for artificial intelligence (AI) initiatives. Fifty-one percent of respondents listed issues related to quality data ("getting good training data" or "improving the quality of your training dataset") as the biggest bottleneck to successfully completing projects [46]. Gartner estimated that poor data quality costs an average organization $13.5 million per year [47]. Other surveys have found similar results. Failure to properly clean 'dirty' data can lead to inaccurate analytics, incorrect conclusions, and wrong decisions.

Cleaning of dirty data may involve correcting hundreds of types of errors and inconsistencies, such as the following: removing duplicates, standardizing descriptors (e.g., addresses), adding metadata, removing commas, correcting data type errors, poorly structured data, incorrect units, spelling errors, various inconsistencies, and typos.

While quality is not mandatory for integration, it is commonly the most important element. Unstructured data is especially difficult to transform. Graphical interfaces, sometimes referred to as self-service interfaces, provide data preparation features which offer a promise of assisting business / casual users to explore data, transform and blend datasets, and perform analytics on top of a well-integrated infrastructure.

One set of capabilities which present a potential solution to data cleaning issues creates callable business rules, where, for example, the name and address attributes of a data record are checked upon data entry into an application, such as a customer relationship management system, which then uses custom exits to initiate a low-latency data quality process. Implementation of these capabilities requires hand-coded extensions for added flexibility over the base ETL tool, which need to be carefully constructed to not violate the vendor's support of the base ETL tool.

Intra-Organization Data Consistency, and Cross-System Data Synchronization

Data consistency has a close association with data quality, and data synchronization, the latter of which has substantial overlap with change data capture (CDC). Changes (updates) are an inevitable part of data processing, in both batch and real time workloads. Batch CDC predates Big Data and is therefore, not an area that warrants explication here; although it may be interesting to note that some modern metadata technologies can also perform some CDC functionality.

Real-time CDC, however, is new to Big Data use cases and reflects a need for a change in broker or message queue technologies, both of which are ripe areas for standardization. As noted elsewhere, data quality is also an area of concern here, as anyone can appreciate the unfortunate results if inaccurate data is propagated from one application within a department, across an entire enterprise. When the time comes to move data, best of breed synchronization services provide CDC, message Queue capability, and triggers for initiating a transfer process. Some MDM solutions also provide synchronization capabilities as part of their programs.

4.3 GAPS IN MANAGEMENT AND ADMINISTRATION

Supporting Master Data Management, MDM

The modernization of MDM product capabilities is underway in the industry; and the boundaries between integration solutions and MDM solutions are increasingly blurred every year, with several functional sub-components including organization and data consistency between apps, and data warehousing, having significant overlap.

Multi-model databases that maintain a copy of the original content in a staging database, master a subset of key information, and use RDF to support data merging have been suggested as a modern alternative to traditional MDM platforms. Multi-model databases reduce the need for up-front ETL allowing for simpler data integration. Flexible schemas and flexible metadata support allow for different lenses to be placed upon the data supporting a wider user base. RDF and OWL can be used to augment facts and business rules used to merge records in MDM.

Single truth

The concept of single truth can be based on metadata management as a part of larger reference data management (RDM) functions. Some modern MDM architectures that perform integration and mastering distinguish between a 'trust-based' model instead of a 'truth-based' model that chases elusive perfection in a Big Data environment. In contrast to the truth-based model that masters a small subset of entity attributes (those that can be virtually assured to be correct or true), a trust-based model leverages a larger amount of data; entities retain the data from the original sources along with the metadata to provide historical context, data lineage or provenance, and timestamps on each data element. This approach allows users, application developers, or business stakeholders to see all the data and decide what is closest to the true copy—and what will be most useful for the business. Some modern MDM tools use visual interfaces that accommodate all types of users, to see lineage and provenance of the data processing, and to reach a higher level of trust with the data. Using the same interface for system requirements gathering and translation to developers also reduces confusion in projects and increases the chances for successful implementations. Metadata management techniques are critical to MDM programs, as metadata itself is a central control mechanism for all integration activity.

Supporting Governance

By some perspectives governance plays an integration role in the life cycle of Big Data, serving as the glue that binds the primary stages of the life cycle together. From this perspective, acquisition, awareness, and analytics of the data compose the full life cycle. The acquisition and awareness portions of this life cycle deal directly with data heterogeneity problems. Awareness, in this case, would generally be that the system, which acquires heterogeneous data from external sources, must have a contextual semantic framework (i.e., model) for integration of that data to make it usable.

The key areas where standards can promote the usability of data in this context are with global resource identifiers, models for storing data relationship classifications (such as RDF) and the creation of resource relationships [48]. Hence information architecture plays an increasingly important role. The awareness part of the cycle is also where the framework for identifying patterns in the data is constructed, and where metadata processing is managed. It is quite possible that this phase of the larger life cycle is the area most prepared for innovation, although the analytics phase may be the part of the cycle currently undergoing the greatest transformation.

As the wrapper or glue that holds the parts of the Big Data life cycle together, a viable governance program will likely require a short list of properties for assuring the novelty, quality, utility, and validity of its data. As an otherwise equal partner in the Big Data life cycle, governance is not a technical function as the others, but rather more like a policy function that should reach into the cycle at all phases. In some sense, governance issues present more serious challenges to organizations than other integration topics listed at the beginning of this section. Better data acquisition, consistency, sharing, and interfaces are highly desired. However, the mere mention of the term *governance* often induces thoughts of pain and frustration for an organization's management staff. Some techniques in the field have been found to have higher rates of end user acceptance and thus satisfaction of the organizational needs contained within the governance programs.

One of the more popular methods for improving governance-related standardization on datasets and reports is through a requirement that datasets and reports go through a review process that ensures that the data conforms to a handful of standards covering data ownership and aspects of IT. See, also, Volume 9, Section 6.5.3 Upon passage of review, the data can be given a 'watermark' which serves as an organization-wide seal of approval that the dataset or the report has been vetted and certified to be appropriate for sharing and decision making.

This process is popular partly because it is rather quick and easy to implement, minimizing push back from employees who must adopt a new process. The assessment for a watermark might include checks for appropriate or accurate calculations or metrics applied to the data, a properly structured dataset for additional processing, and application of proper permissions controls for supporting end-user access. A data container, such as a data mart, can also serve as a form of data verification [49].

4.4 GAPS IN DEPLOYMENT AND OPTIMIZATION

4.4.1 STANDARDS GAP 10: ANALYTICS

Strictly speaking, analytics can be completed on small datasets without Big Data processing. The advent of more accessible tools, technologically and financially, for distributed computing and parallel processing of large datasets has had a profound impact on the discipline of analytics. Both the ubiquity of cloud computing and the availability of open source distributed computing tools have changed the way statisticians and data scientists perform analytics. Since the dawn of computing, scientists at national laboratories or large companies had access to the resources required to solve many computationally expensive and memory-intensive problems. Prior to Big Data, most statisticians did not have access to supercomputers and near-infinitely large databases. These technology limitations forced statisticians to consider trade-offs when conducting analyses and many times dictated which statistical learning model was applied. With the cloud computing revolution and the publication of open source tools to help setup and execute distributed computing environments, both the scope of analytics and the analytical methods available to statisticians changed, resulting in a new analytical landscape. This new analytical landscape left a gap in associated standards. Continual changes in the analytical landscape due to advances in Big Data technology are only worsening this standards gap.

Some examples of the changes to analytics due to Big Data are the following:

- Allowing larger and larger sample sizes to be processed and thus changing the power and sampling error of statistical results;
- Scaling out instead of scaling up, due to Big Data technology, has driven down the cost of storing large datasets;
- Increasing the speed of computationally expensive machine learning algorithms so that they are practical for analysis needs;
- Allowing in-memory analytics to achieve faster results;
- Allowing streaming or real-time analytics to apply statistical learning models in real time;
- Allowing enhanced visualization techniques for improved understanding;
- Cloud-based analytics made acquiring massive amounts of computing power for short periods of time financially accessible to businesses of all sizes and even individuals;
- Driving the creation of tools to make unstructured data appear structured for analysis;
- Shifting from an operational focus to an analytical focus with databases specifically designed for analytics;
- Allowing the analysis of more unstructured (non-relational) data;
- Shifting the focus on scientific analysis from causation to correlation;
- Allowing the creation of data lakes, where the data model is not predefined prior to creation or analysis;
- Enhanced machine learning algorithms—training and test set sizes have been increased due to Big Data tools, leading to more accurate predictive models;
- Driving the analysis of behavioral data—Big Data tools have provided the computational capacity to analyze behavioral datasets such as web traffic or location data; and
- Enabling deep learning techniques.

With this new analytical landscape comes the need for additional knowledge beyond just statistical methods. Statisticians are required to have knowledge of which algorithms scale well and which algorithms deal with particular dataset sizes more efficiently.

For example, without Big Data tools, a random forest may be the best classification algorithm for a particular application provided project time constraints. However, with the computational resources afforded by Big Data, a deep learning algorithm may become the most accurate choice that satisfies the same project time constraints. Another prominent example is the selection of algorithms which handle streaming data well.

Standardizing analytical techniques and methodologies that apply to Big Data will have an impact on the accuracy, communicability, and overall effectiveness of analyses completed in accordance with this NBDIF.

With respect to the shifting of focus on scientific analysis from causation to correlation, traditional scientific analysis has focused on the development of causal models, from which predictions can be made. Causal models focus on understanding the relationships that drive change in the physical world. However, the advent of Big Data analysis has brought about a shift in what is practical in terms of model development. Big Data has allowed a shift of the focus from causal driven to correlation driven. Ever more frequently, knowing that variables are correlated is enough to make progress and better decisions. Big data analytics has allowed this shift from focusing on understanding why (causal) to the what (correlation). Some technologists have even purported that Big Data analysis focusing on correlation may make the scientific method obsolete [50]. From a pragmatic standpoint, deriving correlations instead of causal models will continue to be increasingly important as Big Data technologies mature.

DATA VIRTUALIZATION

Another area for consideration in Big Data systems implementation is that of data virtualization, sometimes referred to as 'federation.' As one of the basic building blocks of a moderately mature integration program, data virtualization is all about moving analysis to the data, in contrast to pulling data from a storage location into a data warehouse for analysis. Data virtualization programs are also applicable in small dataset data science scenarios.

However, data virtualization and data federation systems struggle with many things. For example, federated systems go down when any federate goes down, or require complex code to support partial queries in a degraded state. Often, live source systems do not have capacity for even minimal real-time queries, much less critical batch processes, so the federated virtual database may bring down or impact critical up-stream systems.

Another shortcoming is that every query to the overall system must be converted into many different queries or requests, one for every federated silo. This creates additional development work and tightly couples the federated system to silos.

There is also the least common denominator query issue: if any source system or silo does not support a query—because that query searches by a particular field, orders by a particular field, uses geospatial coordinate search, uses text search, or involves custom relevance scores—then the overall system cannot support it. This also means that any new systems added later may actually decrease the overall capabilities of the federation, rather than increase it. Emerging data-lake and multi-model database technologies introduce functionalities for remedy of these challenges. However, Big Data systems built on a data lake face a difficult task when attempting to support governance. Data manipulation functions in data lake architectures remain black boxes, overly restrictive in their ability to meet governance requirements. The result is frequently a situation of inconsistency, a governance condition referred to as the data swamp.

5 PATHWAYS TO ADDRESS STANDARDS GAPS

Note that the impact of gap closures is not expected to be even for all industries. For example, the development of interoperability standards for predictive analytics applications which are believed to generally provide value to a number of industries and use cases, notably in healthcare [51], is not expected to have a higher than average impact on the automotive industry. In contrast, predictive maintenance capabilities are expected to have a high impact in the automotive industry, but not so in the healthcare industry.

5.1 MIDDLEWARE

A key solution for many Big Data interoperability problems will be Middleware. We can almost come to this hypothesis through the process of elimination. Due to the lack of consensus on lower level technologies such as network protocols, operating systems, programming languages, etc., middleware is the remaining piece of the architecture puzzle which is in a position to successfully mask heterogeneity and also connect to other levels of the architecture. Middleware can be platform independent, acting as an abstraction of system behavior, and structure. Middleware can also map to platform specific models, and be reused for multiple applications, through reasonable levels of effort. A standard will be required for these mappings, to ensure that the different implementations that will be based on them, follow certain consistent engineering practice.

5.2 PERIPHERALS

Best practices suggest that practitioners maintain sight of peripherals to interoperability, including governance.

Appendix A: Acronyms

ACRL	Association of College and Research Libraries
AMQP	Advanced Message Queuing Protocol
ANSI	American National Standards Institute
API	Application Programming Interface
AVC	Advanced Video Coding
AVDL	Application Vulnerability Description Language
BDAP	Big Data Application Provider
BDFP	Big Data Framework Provider
BIAS	Biometric Identity Assurance Services
CCD	Continuity of Care Document
CCMS	Common Core Metadata Schema
CCR	Continuity of Care Record
CDC	Change Data Capture
CGM	Computer Graphics Metafile
CIA	Confidentiality, Integrity, and Availability
CIS	Coverage Implementation Schema
CMIS	Content Management Interoperability Services
CPR	Capability Provider Requirements
CQL	Contextual Query Language
CTAS	Conformance Target Attribute Specification
DC	Data Consumer
DCAT	Data Catalog Vocabulary
DCC	Digital Curation Centre
DCIP	Data Catalog Interoperability Protocol
DCR	Data Consumer Requirements
DOI	Digital Object Identifier
DOM	Document Object Model
DP	Data Provider
DSML	Directory Services Markup Language
DSR	Data Source Requirements
DSS	Digital Signature Service
EPP	Extensible Provisioning Protocol
ETL	Extract, Transform, Load
EXI	Efficient XML Interchange
FAIR	Findable, Accessible, Interoperable, and Reusable
FSBI	Financial Services, Banking, and Insurance
GeoXACML	Geospatial eXtensible Access Control Markup Language
GML	Geography Markup Language
GRC	Governance, Risk management, and Compliance
HDFS	Hadoop Distributed File System
HEVC	High Efficiency Video Coding
HITSP	Healthcare Information Technology Standards Panel
HLVA	High-Level Version Architecture
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol

IBCS	International Business Communication Standards
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
INCITS	International Committee for Information Technology Standards
iPaaS	integration platform as a service
IR	Information Retrieval
ISO	International Organization for Standardization
IT	Information Technology
ITL	Information Technology Laboratory
ITS	Internationalization Tag Set
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
JSR	Java Specification Request
JTC1	Joint Technical Committee 1
LMR	Life Cycle Management Requirements
M	Management Fabric
MDM	Master Data Management
MDX	Multidimensional expressions
MFI	Metamodel Framework for Interoperability
MOWS	Management of Web Services
MPD	Model Package Description
MPEG	Moving Picture Experts Group
MQTT	Message Queuing Telemetry Transport
MUWS	Management Using Web Services
MWaaS	Middleware as a Service
NARA	National Archives and Records Administration
NASA	National Aeronautics and Space Administration
NBD-PWG	NIST Big Data Public Working Group
NBDIF	NIST Big Data Interoperability Framework
NBDRA	NIST Big Data Reference Architecture
NCAP	Network Capable Application Processor
NCPDP	National Council for Prescription Drug Programs
NDR	Naming and Design Rules
netCDF	network Common Data Form
NIEM	National Information Exchange Model
NISO	National Information Standards Organization
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NoSQL	Not Only or No Structured Query Language
NSF	National Science Foundation
OASIS	Organization for the Advancement of Structured Information Standards
OData	Open Data
ODMS	On Demand Model Selection
OGC	Open Geospatial Consortium
OGF	Open Grid Forum
OLAP	Online Analytical Processing
OpenMI	Open Modelling Interface Standard
OR	Other Requirements
OWS Context	Web Services Context Document
P3P	Platform for Privacy Preferences Project

PICS	Platform for Internet Content Selection
PID	Persistent Identifier
PII	Personally Identifiable Information
PMML	Predictive Modeling Markup Language
POWDER	Protocol for Web Description Resources
RDF	Resource Description Framework
REST	Representational State Transfer
RFID	Radio Frequency Identification
RIF	Rule Interchange Format
RPM	RedHat Package Manager
S&P	Security and Privacy Fabric
SAF	Symptoms Automation Framework
SAML	Security Assertion Markup Language
SDMX	Statistical Data and Metadata Exchange
SDOs	Standards Development Organizations
SES	Standards Engineering Society
SFA	Simple Features Access
SKOS	Simple Knowledge Organization System Reference
SLAs	Service-Level Agreements
SML	Service Modeling Language
SNMP	Simple Network Management Protocol
SO	System Orchestrator Component
SOAP	Simple Object Access Protocol
SPR	Security and Privacy Requirements
SQL	Structured Query Language
SWE	Sensor Web Enablement
SWS	Search Web Services
TC	Technical Committee
TCP/IP	Transmission Control Protocol / Internet Protocol
TEDS	Transducer Electronic Data Sheet
TEI	Text Encoding and Interchange
TJS	Table Joining Service
TPR	Transformation Provider Requirements
TR	Technical Report
UBL	Universal Business Language
UDDI	Universal Description, Discovery and Integration
UDP	User Datagram Protocol
UIMA	Unstructured Information Management Architecture
UML	Unified Modeling Language
UOML	Unstructured Operation Markup Language
VoID	Vocabulary of Interlinked Datasets
WAIS	Wide Area Information Servers
W3C	World Wide Web Consortium
WCPS	Web Coverage Processing Service Interface
WCS	Web Coverage Service
WebRTC	Web Real-Time Communication
WFS	Web Feature Service
WMS	Web Map Service
WPS	Web Processing Service
WS-BPEL	Web Services Business Process Execution Language
WS-Discovery	Web Services Dynamic Discovery

WSDL	Web Services Description Language
WSDM	Web Services Distributed Management
WS-Federation	Web Services Federation Language
WSN	Web Services Notification
XACML	eXtensible Access Control Markup Language
XDM	XPath Data Model
X-KISS	XML Key Information Service Specification
XKMS	XML Key Management Specification
X-KRSS	XML Key Registration Service Specification
XMI	XML Metadata Interchange
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations

Appendix B: Collection of Big Data Related Standards

The following table contains a collection of standards that pertain to a portion of the Big Data ecosystem. This collection is current, as of the date of publication of Volume 7. It is not an exhaustive list of standards that could relate to Big Data but rather a representative list of the standards that significantly impact some area of the Big Data ecosystem.

In selecting standards to include in Appendix B, the working group focused on standards that fit the following criteria:

- Facilitate interfaces between NBDRA components;
- Facilitate the handling of data with one or more Big Data characteristics; and
- Represent a fundamental function needing to be implemented by one or more NBDRA components.

Appendix B represents a portion of potentially applicable standards from a portion of contributing organizations working in Big Data domain. Appendix C and Appendix D describe different aspects of the same list of standards presented in Appendix B.

Table B-1: Big Data-Related Standards

Standard Name/Number	Description
ISO/IEC 9075-*	ISO/IEC 9075 defines SQL. The scope of SQL is the definition of data structure and the operations on data stored in that structure. ISO/IEC 9075-1, ISO/IEC 9075-2 and ISO/IEC 9075-11 encompass the minimum requirements of the language. Other parts define extensions. Specifically, 9075-15:2018 defines model and queries on multi-dimensional arrays (data cubes).
ISO/IEC Technical Report (TR) 9789	Guidelines for the Organization and Representation of Data Elements for Data Interchange
ISO/IEC 11179-*	The 11179 standard is a multipart standard for the definition and implementation of Metadata Registries. The series includes the following parts: • Part 1: Framework • Part 2: Classification • Part 3: Registry metamodel and basic attributes • Part 4: Formulation of data definitions • Part 5: Naming and identification principles • Part 6: Registration

ISO/IEC 10728-*	Information Resource Dictionary System Services Interface
ISO/IEC 13249-*	Database Languages – SQL Multimedia and Application Packages
ISO/IEC TR 19075-*	This is a series of TRs on SQL related technologies. • Part 1: Xquery • Part 2: SQL Support for Time-Related Information • Part 3: Programs Using the Java Programming Language • Part 4: Routines and Types Using the Java Programming Language
ISO/IEC 19503	Extensible Markup Language (XML) Metadata Interchange (XMI)
ISO/IEC 19773	Metadata Registries Modules
ISO/IEC TR 20943	Metadata Registry Content Consistency
ISO/IEC 19763-*	Information Technology—Metamodel Framework for Interoperability (MFI) ISO/IEC 19763, Information Technology –MFI. The 19763 standard is a multipart standard that includes the following parts: • Part 1: Reference model • Part 3: Metamodel for ontology registration • Part 5: Metamodel for process model registration • Part 6: Registry Summary • Part 7: Metamodel for service registration • Part 8: Metamodel for role and goal registration • Part 9: On Demand Model Selection (ODMS) TR • Part 10: Core model and basic mapping • Part 12: Metamodel for information model registration • Part 13: Metamodel for forms registration • Part 14: Metamodel for dataset registration • Part 15: Metamodel for data provenance registration
ISO/IEC 9281:1990	Information Technology—Picture Coding Methods
ISO/IEC 10918:1994	Information Technology—Digital Compression and Coding of Continuous-Tone Still Images
ISO/IEC 11172:1993	Information Technology—Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1,5 Mbit/s
ISO/IEC 13818:2013	Information Technology—Generic Coding of Moving Pictures and Associated Audio Information
ISO/IEC 14496:2010	Information Technology—Coding of Audio-Visual Objects
ISO/IEC 15444:2011	Information Technology—JPEG (Joint Photographic Experts Group) 2000 Image Coding System
ISO/IEC 21000:2003	Information Technology—Multimedia Framework (MPEG (Moving Picture Experts Group)-21)
ISO 6709:2008	Standard Representation of Geographic Point Location by Coordinates
ISO 19115-*	Geographic Metadata. ISO 19115-2:2009 contains extensions for imagery and gridded data; and ISO/TS 19115-3:2016 provides an XML schema implementation for the fundamental concepts compatible with ISO/TS

	19138:2007 (Geographic Metadata XML, or GMD).
ISO 19110	Geographic Information Feature Cataloging
ISO 19139	Geographic Metadata XML Schema Implementation
ISO 19119	Geographic Information Services
ISO 19157	Geographic Information Data Quality
ISO 19114	Geographic Information—Quality Evaluation Procedures
IEEE 21451 -*	Information Technology—Smart transducer interface for sensors and actuators • Part 1: Network Capable Application Processor (NCAP) information model • Part 2: Transducer to microprocessor communication protocols and Transducer Electronic Data Sheet (TEDS) formats • Part 4: Mixed-mode communication protocols and TEDS formats • Part 7: Transducer to radio frequency identification (RFID) systems communication protocols and TEDS formats
IEEE 2200-2012	Standard Protocol for Stream Management in Media Client Devices
ISO/IEC 15408-2009	Information Technology—Security Techniques—Evaluation Criteria for IT Security
ISO/IEC 27010:2012	Information Technology—Security Techniques—Information Security Management for Inter-Sector and Inter-Organizational Communications
ISO/IEC 27033-1:2009	Information Technology—Security Techniques—Network Security
ISO/IEC TR 14516:2002	Information Technology—Security Techniques—Guidelines for the Use and Management of Trusted Third-Party Services
ISO/IEC 29100:2011	Information Technology—Security Techniques—Privacy Framework
ISO/IEC 9798:2010	Information Technology—Security Techniques—Entity Authentication
ISO/IEC 11770:2010	Information Technology—Security Techniques—Key Management
ISO/IEC 27035:2011	Information Technology—Security Techniques—Information Security Incident Management
ISO/IEC 27037:2012	Information Technology—Security Techniques—Guidelines for Identification, Collection, Acquisition and Preservation of Digital Evidence
JSR (Java Specification Request) 221 (developed by the Java Community Process)	JDBC™ 4.0 Application Programming Interface (API) Specification
W3C XML	XML 1.0 (Fifth Edition) W3C Recommendation 26 November 2008
W3C Resource Description Framework (RDF)	The RDF is a framework for representing information in the Web. RDF graphs are sets of subject-predicate-object triples, where the elements are used to express descriptions of resources.
W3C JavaScript Object Notation (JSON)-LD 1.0	JSON-LD 1.0 A JSON-based Serialization for Linked Data W3C Recommendation 16 January 2014

W3C Document Object Model (DOM) Level 1 Specification	This series of specifications define the DOM, a platform- and language-neutral interface that allows programs and scripts to dynamically access and update the content, structure and style of HyperText Markup Language (HTML) and XML documents.
W3C XQuery 3.0	The XQuery specifications describe a query language called XQuery, which is designed to be broadly applicable across many types of XML data sources.
W3C XProc	This specification describes the syntax and semantics of *XProc: An XML Pipeline Language*, a language for describing operations to be performed on XML documents.
W3C XML Encryption Syntax and Processing Version 1.1	This specification covers a process for encrypting data and representing the result in XML.
W3C XML Signature Syntax and Processing Version 1.1	This specification covers XML digital signature processing rules and syntax. XML Signatures provide integrity, message authentication, and/or signer authentication services for data of any type, whether located within the XML that includes the signature or elsewhere.
W3C XPath 3.0	XPath 3.0 is an expression language that allows the processing of values conforming to the data model defined in (XQuery and XPath Data Model (XDM) 3.0). The data model provides a tree representation of XML documents as well as atomic values and sequences that may contain both references to nodes in an XML document and atomic values.
W3C XSL Transformations (XSLT) Version 2.0	This specification defines the syntax and semantics of XSLT 2.0, a language for transforming XML documents into other XML documents.
W3C Efficient XML Interchange (EXI) Format 1.0 (Second Edition)	This specification covers the EXI format. EXI is a very compact representation for the XML Information Set that is intended to simultaneously optimize performance and the utilization of computational resources.
W3C RDF Data Cube Vocabulary	The Data Cube vocabulary provides a means to publish multidimensional data, such as statistics on the Web using the W3C RDF standard.
W3C Data Catalog Vocabulary (DCAT)	DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. This document defines the schema and provides examples for its use.
W3C HTML5 A vocabulary and associated APIs for HTML and XHTML	This specification defines the 5th major revision of the core language of the World Wide Web—HTML.
W3C Internationalization Tag Set (ITS) 2.0	The ITS 2.0 specification enhances the foundation to integrate automated processing of human language into core Web technologies and concepts that are designed to foster the automated creation and processing of multilingual Web content.
W3C OWL 2 Web Ontology Language	The OWL 2 Web Ontology Language, informally OWL 2, is an ontology language for the Semantic Web with formally defined meaning.
W3C Platform for Privacy Preferences (P3P) 1.0	The P3P enables Web sites to express their privacy practices in a standard format that can be retrieved automatically and interpreted easily by user agents.
W3C Protocol for Web Description Resources (POWDER)	POWDER—the Protocol for Web Description Resources—provides a mechanism to describe and discover Web resources and helps the users to decide whether a given resource is of interest.

W3C Provenance	Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. The Provenance Family of Documents (PROV) defines a model, corresponding serializations and other supporting definitions to enable the inter-operable interchange of provenance information in heterogeneous environments such as the Web.
W3C Rule Interchange Format (RIF)	RIF is a series of standards for exchanging rules among rule systems, in particular among Web rule engines.
W3C Service Modeling Language (SML) 1.1	This specification defines the SML, Version 1.1 used to model complex services and systems, including their structure, constraints, policies, and best practices.
W3C Simple Knowledge Organization System Reference (SKOS)	This document defines the SKOS, a common data model for sharing and linking knowledge organization systems via the Web.
W3C Simple Object Access Protocol (SOAP) 1.2	SOAP is a protocol specification for exchanging structured information in the implementation of web services in computer networks.
W3C SPARQL 1.1	SPARQL is a language specification for the query and manipulation of linked data in a RDF format.
W3C Web Service Description Language (WSDL) 2.0	This specification describes the WSDL Version 2.0, an XML language for describing Web services.
W3C XML Key Management Specification (XKMS) 2.0	This standard specifies protocols for distributing and registering public keys, suitable for use in conjunction with the W3C Recommendations for XML Signature (XML-SIG) and XML Encryption (XML-Enc). The XKMS comprises two parts: • The XML Key Information Service Specification (X-KISS) • The XML Key Registration Service Specification (X-KRSS).
OGC® OpenGIS® Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile	This series of standard covers Catalogue Services based on ISO19115/ISO19119 are organized and implemented for the discovery, retrieval and management of data metadata, services metadata and application metadata.
OGC® OpenGIS® GeoAPI	The GeoAPI Standard defines, through the GeoAPI library, a Java language API including a set of types and methods which can be used for the manipulation of geographic information structured following the specifications adopted by the Technical Committee 211 of the ISO and by the OGC®.
OGC® OpenGIS® GeoSPARQL	The OGC® GeoSPARQL standard supports representing and querying geospatial data on the Semantic Web. GeoSPARQL defines a vocabulary for representing geospatial data in RDF, and it defines an extension to the SPARQL query language for processing geospatial data.
OGC® OpenGIS® Geography Markup Language (GML) Encoding Standard	The GML is an XML grammar for expressing geographical features. GML serves as a modeling language for geographic systems as well as an open interchange format for geographic transactions on the Internet.

OGC® Geospatial eXtensible Access Control Markup Language (GeoXACML) Version 1	The Policy Language introduced in this document defines a geo-specific extension to the XACML Policy Language, as defined by the OASIS standard eXtensible Access Control Markup Language (XACML), Version 2.0”
OGC® network Common Data Form (netCDF)	netCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.
OGC® Open Modelling Interface Standard (OpenMI)	The purpose of the OpenMI is to enable the runtime exchange of data between process simulation models and also between models and other modelling tools such as databases and analytical and visualization applications.
OGC® OpenSearch Geo and Time Extensions	This OGC standard specifies the Geo and Time extensions to the OpenSearch query protocol. OpenSearch is a collection of simple formats for the sharing of search results.
OGC® Web Services Context Document (OWS Context)	The OGC® OWS Context was created to allow a set of configured information resources (service set) to be passed between applications primarily as a collection of services.
OGC® Sensor Web Enablement (SWE)	This series of standards support interoperability interfaces and metadata encodings that enable real time integration of heterogeneous sensor webs. These standards include a modeling language (SensorML), common data model, and sensor observation, planning, and alerting service interfaces.
OGC® OpenGIS® Simple Features Access (SFA)	Describes the common architecture for simple feature geometry and is also referenced as ISO 19125. It also implements a profile of the spatial schema described in ISO 19107:2003.
OGC® OpenGIS® Georeferenced Table Joining Service (TJS) Implementation Standard	This standard is the specification for a TJS that defines a simple way to describe and exchange tabular data that contains information about geographic objects.
OGC® OpenGIS® Web Coverage Processing Service Interface (WCPS) Standard	Defines a protocol-independent language for the extraction, processing, and analysis of multidimensional gridded coverages representing sensor, timeseries image, simulation, or statistics data.
OGC® OpenGIS® Web Coverage Service (WCS)	Defines a modular, flexible suite of functionality for offering multidimensional, spatio-temporal coverage data for access over the Internet. WCS Core, mandatory for a WCS implementation to be compliant, establishes subsetting and format encoding; WCS extensions add optional functionality facets, from simple band extract up to complex analytics with WCPS.
OGC® Web Feature Service (WFS) 2.0 Interface Standard	The WFS standard provides for fine-grained access to geographic information at the feature and feature property level. This International Standard specifies discovery operations, query operations, locking operations, transaction operations and operations to manage stored, parameterized query expressions.
OGC® OpenGIS® Web Map Service (WMS) Interface Standard	The OpenGIS® WMS Interface Standard provides a simple HTTP (Hypertext Transfer Protocol) interface for requesting geo-registered map images from one or more distributed geospatial databases.

OGC® OpenGIS® Web Processing Service (WPS) Interface Standard	The OpenGIS® WPS Interface Standard provides rules for standardizing how inputs and outputs (requests and responses) for geospatial processing services, such as polygon overlay. The standard also defines how a client can request the execution of a process, and how the output from the process is handled. It defines an interface that facilitates the publishing of geospatial processes and clients' discovery of and binding to those processes.
OASIS AS4 Profile of ebMS 3.0 v1.0	Standard for business to business exchange of messages via a web service platform.
OASIS Advanced Message Queuing Protocol (AMQP) Version 1.0	The AMQP is an open internet protocol for business messaging. It defines a binary wire-level protocol that allows for the reliable exchange of business messages between two parties.
OASIS Application Vulnerability Description Language (AVDL) v1.0	This specification describes a standard XML format that allows entities (such as applications, organizations, or institutes) to communicate information regarding web application vulnerabilities.
OASIS Biometric Identity Assurance Services (BIAS) Simple Object Access Protocol (SOAP) Profile v1.0	This OASIS BIAS profile specifies how to use XML (XML10) defined in ANSI INCITS 442-2010—BIAS to invoke SOAP -based services that implement BIAS operations.
OASIS Content Management Interoperability Services (CMIS)	The CMIS standard defines a domain model and set of bindings that include Web Services and RESTful AtomPub that can be used by applications to work with one or more Content Management repositories/systems.
OASIS Digital Signature Service (DSS)	This specification describes two XML-based request/response protocols - a signing protocol and a verifying protocol. Through these protocols a client can send documents (or document hashes) to a server and receive back a signature on the documents; or send documents (or document hashes) and a signature to a server, and receive back an answer on whether the signature verifies the documents.
OASIS Directory Services Markup Language (DSML) v2.0	The DSML provides a means for representing directory structural information as an XML document methods for expressing directory queries and updates (and the results of these operations) as XML documents
OASIS ebXML Messaging Services	These specifications define a communications-protocol neutral method for exchanging electronic business messages as XML.
OASIS ebXML RegRep	ebXML RegRep is a standard defining the service interfaces, protocols and information model for an integrated registry and repository. The repository stores digital content while the registry stores metadata that describes the content in the repository.
OASIS ebXML Registry Information Model	The Registry Information Model provides a blueprint or high-level schema for the ebXML Registry. It provides implementers with information on the type of metadata that is stored in the Registry as well as the relationships among metadata Classes.
OASIS ebXML Registry Services Specification	An ebXML Registry is an information system that securely manages any content type and the standardized metadata that describes it. The ebXML Registry provides a set of services that enable sharing of content and metadata between organizational entities in a federated environment.
OASIS eXtensible Access Control Markup Language (XACML)	The standard defines a declarative access control policy language implemented in XML and a processing model describing how to evaluate access requests according to the rules defined in policies.

OASIS Message Queuing Telemetry Transport (MQTT)	MQTT is a Client Server publish/subscribe messaging transport protocol for constrained environments such as for communication in Machine to Machine and Internet of Things contexts where a small code footprint is required and/or network bandwidth is at a premium.
OASIS Open Data (OData) Protocol	The OData Protocol is an application-level protocol for interacting with data via RESTful interfaces. The protocol supports the description of data models and the editing and querying of data according to those models.
OASIS Search Web Services (SWS)	The OASIS SWS initiative defines a generic protocol for the interaction required between a client and server for performing searches. SWS define an Abstract Protocol Definition to describe this interaction.
OASIS Security Assertion Markup Language (SAML) v2.0	The SAML defines the syntax and processing semantics of assertions made about a subject by a system entity. This specification defines both the structure of SAML assertions, and an associated set of protocols, in addition to the processing rules involved in managing a SAML system.
OASIS SOAP-over-UDP (User Datagram Protocol) v1.1	This specification defines a binding of SOAP to user datagrams, including message patterns, addressing requirements, and security considerations.
OASIS Solution Deployment Descriptor Specification v1.0	This specification defines schema for two XML document types: Package Descriptors and Deployment Descriptors. Package Descriptors define characteristics of a package used to deploy a solution. Deployment Descriptors define characteristics of the content of a solution package, including the requirements that are relevant for creation, configuration and maintenance of the solution content.
OASIS Symptoms Automation Framework (SAF) Version 1.0	This standard defines reference architecture for the Symptoms Automation Framework, a tool in the automatic detection, optimization, and remediation of operational aspects of complex systems,
OASIS Topology and Orchestration Specification for Cloud Applications Version 1.0	The concept of a "service template" is used to specify the "topology" (or structure) and "orchestration" (or invocation of management behavior) of IT services. This specification introduces the formal description of Service Templates, including their structure, properties, and behavior.
OASIS Universal Business Language (UBL) v2.1	The OASIS UBL defines a generic XML interchange format for business documents that can be restricted or extended to meet the requirements of particular industries.
OASIS Universal Description, Discovery and Integration (UDDI) v3.0.2	The focus of UDDI is the definition of a set of services supporting the description and discovery of (1) businesses, organizations, and other Web services providers, (2) the Web services they make available, and (3) the technical interfaces which may be used to access those services.
OASIS Unstructured Information Management Architecture (UIMA) v1.0	The UIMA specification defines platform-independent data representations and interfaces for text and multi-modal analytics.
OASIS Unstructured Operation Markup Language (UOML) v1.0	UOML is interface standard to process unstructured document; it plays the similar role as SQL to structured data. UOML is expressed with standard XML.
OASIS/W3C WebCGM v2.1	Computer Graphics Metafile (CGM) is an ISO standard, defined by ISO/IEC 8632:1999, for the interchange of 2D vector and mixed vector/raster graphics. WebCGM is a profile of CGM, which adds Web linking and is optimized for Web applications in technical illustration, electronic documentation, geophysical data visualization, and similar fields.

OASIS Web Services Business Process Execution Language (WS-BPEL) v2.0	This standard defines a language for specifying business process behavior based on Web Services. WS-BPEL provides a language for the specification of Executable and Abstract business processes.
OASIS/W3C - Web Services Distributed Management (WSDM): Management Using Web Services (MUWS) v1.1	MUWS defines how an IT resource connected to a network provides manageability interfaces such that the IT resource can be managed locally and from remote locations using Web services technologies.
OASIS WSDM: Management of Web Services (MOWS) v1.1	This part of the WSDM specification addresses management of the Web services endpoints using Web services protocols.
OASIS Web Services Dynamic Discovery (WS-Discovery) v1.1	This specification defines a discovery protocol to locate services. The primary scenario for discovery is a client searching for one or more target services.
OASIS Web Services Federation Language (WS-Federation) v1.2	This specification defines mechanisms to allow different security realms to federate, such that authorized access to resources managed in one realm can be provided to security principals whose identities and attributes are managed in other realms.
OASIS Web Services Notification (WSN) v1.3	WSN is a family of related specifications that define a standard Web services approach to notification using a topic-based publish/subscribe pattern.
IETF Simple Network Management Protocol (SNMP) v3	SNMP is a series of IETF sponsored standards for remote management of system/network resources and transmission of status regarding network resources. The standards include definitions of standard management objects along with security controls.
IETF Extensible Provisioning Protocol (EPP)	This IETF series of standards describes an application-layer client-server protocol for the provisioning and management of objects stored in a shared central repository. Specified in XML, the protocol defines generic object management operations and an extensible framework that maps protocol operations to objects.
National Council for Prescription Drug Programs (NCPDPD) Script standard	Electronic data exchange standard used in medication reconciliation process. Medication history, prescription info (3), census update.
ASTM Continuity of Care Record (CCR)	Electronic data exchange standard used in medication reconciliation process. CCR represents a summary format for the core facts of a patient's dataset.
Healthcare Information Technology Standards Panel (HITSP) C32 HL7 Continuity of Care Document (CCD)	Electronic data exchange standard used in medication reconciliation process. Summary format for CCR document structure.
PMML Predictive Model Markup Language	XML based data handling. Mature standard defines and enables data modeling, and reliability and scalability for custom deployments. Pre / post processing, expression of predictive models.
Dash7	Dynamic adaptive streaming over HTTP. Media presentation description format. Wireless sensor and actuator protocol; home automation, based on ISO IEC 18000-7
H.265	High efficiency video coding (HEVC) MPEG-H part 2. Potential compression successor to Advanced Video Coding (AVC) H.264. Streaming video.

VP9	Royalty free codec alternative to HEVC. Successor to VP8, competitor to H.265. Streaming video.
Daala	Video coding format. Streaming video.
WebRTC	Browser to browser communication
X.509	Public key encryption for securing email and web communication.
MDX	Multidimensional expressions (MDX) became the standard for OLAP query.
NIEM-HLVA	National Information Exchange Model (NIEM) High-Level Version Architecture (HLVA): Specifies the NIEM version architecture.
NIEM-MPD	NIEM Model Package Description (MPD) Specification: Specifies rules for organizing and packaging MPDs in general and IEPDs specifically.
NIEM-Code List Specifications	NIEM Code Lists Specification: Establishes methods for using code list artifacts with NIEM information exchange specifications.
NIEM Conformance Specification	Defines general conformance to NIEM.
NIEM-CTAS	NIEM Conformance Target Attribute Specification (CTAS): Specifies XML attributes to establish a claim that the document conforms to a set of conformance targets.
NIEM-NDR	NIEM Naming and Design Rules (NDR): Specifies principles and enforceable rules for NIEM-conformant schema documents, instance XML documents and data components.
Non-Normative Guidance in Using NIEM with JSON	Non-Normative Guidance in Using NIEM with JSON: Guidance for using NIEM with JSON-LD specified by RFC4627. Note: A normative NIEM-JSON specification is under development and scheduled for release in Dec 2017.
DCC Data Package, version 1.0.0-beta.17 (a specification) released March of 2016	
DCC Observ-OM \	Observation representation (features, protocols, targets and values). It is intended to lower the barrier for future data sharing and facilitate integrated search across panels and species. All models, formats, documentation, and software are available under LGPLv3.
DCC PREMIS	Independent serialization, preservation of actor information
DCC PROV	Provenance information
DCC QuDEx	Agnostic formatting

DCC SDMX, specification 2.1 last amended May of 2012	Efficient exchange and sharing of statistical data and metadata.
DCC TEI	Varieties and modules for text encoding
BMC Visualization	A dual layer XML based approach to the definition of archetypes and their visual layout that will allow automatic generating of efficient medical data interfaces, allows different views for one MDV model. The same software can provide different interfaces for different devices and users. Meets the following requirements: 1. Complies with the requirements and constraints of an ISO 13606 reference model. The dual model approach of ISO 13606 allows separating the medical knowledge from the software implementation and permits healthcare professionals to define medical concepts without the need to understand how the concepts will be implemented within the EHR. 2. Provides multiple device support. 3. Supports different views on the same data. The same information can be displayed in different ways according to its needed context. This feature is useful for healthcare professionals who may need different views according to their specialization. Patients will need less data but the data have to be presented in more convenient form to ensure that in will be understood without medical background. 4. Is stored separately from the visualized data. The dual model approach that is used as the basis for archetypes has proven to be efficient and flexible. 5. Platform independent.
IEEE 1857.3	Real time transmission of audiovisual content, including internet media streaming, IPTV, and video on demand.
Open Group C172, O-BDL	Describes a set of architectural patterns, and key concepts for setting up data centric strategies.
ISO 10646	Defines character encoding relevant to UTF, and backward compatibility with ASCII.
ISA-Tab	The Investigation/Study/Assay (ISA) tab-delimited (TAB) format is a general purpose framework for complex metadata.
Dublin Core	
ISO/IEC 19123	Coverages, i.e., spatio-temporal regular and irregular grids, point clouds, and general meshes. In particular, this establishes ISO's geospatial data cube model. 19123-1 (in preparation): Abstract Coverage Model 19123-2 (adopted): Coverage Implementation Schema (identical to OGC CIS 1.0)
OGC® Coverage Implementation Schema (CIS)	Defines a format-independent data model for spatio-temporal coverages, i.e.: regular and irregular grids, point clouds, and meshes. In particular, this establishes OGC's data cube model. Various extensions define mappings to data formats such as XML, JSON, RDF, GeoTIFF; NetCDF, GRIB2, etc.

W3C DCIP	A specification designed to "facilitate interoperability between data catalogs published on the Web" (spec.datacatalogs.org) and is complementary to DCAT. It provides an "agreed" protocol (REST API) to access the data defined in DCAT.
VoID	An "RDF Schema vocabulary for describing metadata about RDF data sets" (VOID). Its primary purpose is to bridge the gap between data publishers and data consumers using an exclusive vocabulary to describe different data set attributes.

Appendix C: Standards and the NBDRA

As most standards represent some form of interface between components, the standards table in Appendix C indicates whether the NBDRA component would be an Implementer or User of the standard. For the purposes of this table, the following definitions were used for Implementer and User.

> ***Implementer****: A component is an implementer of a standard if it provides services based on the standard (e.g., a service that accepts Structured Query Language (SQL) commands would be an implementer of that standard) or* ***encodes*** *or presents data based on that standard.*

> ***User:*** *A component is a user of a standard if it interfaces to a service via the standard or if it accepts/consumes/****decodes*** *data represented by the standard.*

While the above definitions provide a reasonable basis for some standards, the difference between implementation and use may be negligible or nonexistent. The NBDRA components and fabrics are abbreviated in the table header as follows:

- SO = System Orchestrator
- DP = Data Provider
- DC = Data Consumer
- BDAP = Big Data Application Provider
- BDFP = Big Data Framework Provider
- S&P = Security and Privacy Fabric
- M = Management Fabric

Table C-1: Standards and the NBDRA

Standard Name/Number	NBDRA Components						
	SO	DP	DC	BDAP	BDFP	S&P	M
ISO/IEC 9075-*		I	I/U	U	I/U	U	U
ISO/IEC Technical Report (TR) 9789		I/U	I/U	I/U	I/U		
ISO/IEC 11179-*		I	I/U	I/U		U	
ISO/IEC 10728-*							
ISO/IEC 13249-*		I	I/U	U	I/U		
ISO/IEC TR 19075-*		I	I/U	U	I/U		

ISO/IEC 19503		I	I/U	U	I/U	U	
ISO/IEC 19773		I	I/U	U	I/U	I/U	
ISO/IEC TR 20943		I	I/U	U	I/U	U	U
ISO/IEC 19763-*		I	I/U	U	U		
ISO/IEC 9281:1990		I	U	I/U	I/U		
ISO/IEC 10918:1994		I	U	I/U	I/U		
ISO/IEC 11172:1993		I	U	I/U	I/U		
ISO/IEC 13818:2013		I	U	I/U	I/U		
ISO/IEC 14496:2010		I	U	I/U	I/U		
ISO/IEC 15444:2011		I	U	I/U	I/U		
ISO/IEC 21000:2003		I	U	I/U	I/U		
ISO 6709:2008		I	U	I/U	I/U		
ISO 19115-*		I	U	I/U	U		
ISO 19110		I	U	I/U			
ISO 19139		I	U	I/U			
ISO 19119		I	U	I/U			
ISO 19157		I	U	I/U	U		
ISO 19114				I			
IEEE 21451 -*		I	U				
IEEE 2200-2012		I	U	I/U			
ISO/IEC 15408-2009	U					I	
ISO/IEC 27010:2012		I	U	I/U			
ISO/IEC 27033-1:2009		I/U	I/U	I/U	I		
ISO/IEC TR 14516:2002	U					U	
ISO/IEC 29100:2011						I	
ISO/IEC 9798:2010		I/U	U	U	U	I/U	
ISO/IEC 11770:2010		I/U	U	U	U	I/U	
ISO/IEC 27035:2011	U					I	
ISO/IEC 27037:2012	U					I	
JSR (Java Specification Request) 221 (developed by the Java Community Process)		I/U	I/U	I/U	I/U		
W3C XML	I/U	I/U	I/U	I/U	I/U	I/U	I/U
W3C Resource Description Framework (RDF)		I	U	I/U	I/U		

W3C JavaScript Object Notation (JSON)-LD 1.0		I	U	I/U	I/U		
W3C Document Object Model (DOM) Level 1 Specification		I	U	I/U	I/U		
W3C XQuery 3.0		I	U	I/U	I/U		
W3C XProc	I	I	U	I/U	I/U		
W3C XML Encryption Syntax and Processing Version 1.1		I	U	I/U			
W3C XML Signature Syntax and Processing Version 1.1		I	U	I/U			
W3C XPath 3.0		I	U	I/U	I/U		
W3C XSL Transformations (XSLT) Version 2.0		I	U	I/U	I/U		
W3C Efficient XML Interchange (EXI) Format 1.0 (Second Edition)		I	U	I/U			
W3C RDF Data Cube Vocabulary		I	U	I/U	I/U		
W3C Data Catalog Vocabulary (DCAT)		I	U	I/U			
W3C HTML5 A vocabulary and associated APIs for HTML and XHTML		I	U	I/U			
W3C Internationalization Tag Set (ITS) 2.0		I	U	I/U	I/U		
W3C OWL 2 Web Ontology Language		I	U	I/U	I/U		
W3C Platform for Privacy Preferences (P3P) 1.0		I	U	I/U		I/U	
W3C Protocol for Web Description Resources (POWDER)		I	U	I/U			
W3C Provenance		I	U	I/U	I/U	U	
W3C Rule Interchange Format (RIF)		I	U	I/U	I/U		
W3C Service Modeling Language (SML) 1.1	I/U	I	U	I/U			
W3C Simple Knowledge Organization System Reference (SKOS)		I	U	I/U			
W3C Simple Object Access Protocol (SOAP) 1.2		I	U	I/U			
W3C SPARQL 1.1		I	U	I/U	I/U		
W3C Web Service Description Language (WSDL) 2.0	U	I	U	I/U			
W3C XML Key Management Specification (XKMS) 2.0	U	I	U	I/U			
OGC® OpenGIS® Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile		I	U	I/U			
OGC® OpenGIS® GeoAPI		I	U	I/U	I/U		
OGC® OpenGIS® GeoSPARQL		I	U	I/U	I/U		
OGC® OpenGIS® Geography Markup Language (GML) Encoding Standard		I	U	I/U	I/U		
OGC® Geospatial eXtensible Access Control Markup Language (GeoXACML) Version 1		I	U	I/U	I/U	I/U	
OGC® network Common Data Form (netCDF)		I	U	I/U			
OGC® Open Modelling Interface Standard (OpenMI)		I	U	I/U	I/U		

OGC® OpenSearch Geo and Time Extensions		I	U	I/U	I		
OGC® Web Services Context Document (OWS Context)		I	U	I/U	I		
OGC® Sensor Web Enablement (SWE)		I	U	I/U			
OGC® OpenGIS® Simple Features Access (SFA)		I	U	I/U	I/U		
OGC® OpenGIS® Georeferenced Table Joining Service (TJS) Implementation Standard		I	U	I/U	I/U		
OGC® OpenGIS® Web Coverage Processing Service Interface (WCPS) Standard		I	U	I/U	I		
OGC® OpenGIS® Web Coverage Service (WCS)		I	U	I/U	I		
OGC® Web Feature Service (WFS) 2.0 Interface Standard		I	U	I/U	I		
OGC® OpenGIS® Web Map Service (WMS) Interface Standard		I	U	I/U	I		
OGC® OpenGIS® Web Processing Service (WPS) Interface Standard		I	U	I/U	I		
OASIS AS4 Profile of ebMS 3.0 v1.0		I	U	I/U			
OASIS Advanced Message Queuing Protocol (AMQP) Version 1.0		I	U	U	I		
OASIS Application Vulnerability Description Language (AVDL) v1.0		I	U	I		U	
OASIS Biometric Identity Assurance Services (BIAS) Simple Object Access Protocol (SOAP) Profile v1.0		I	U	I/U		U	
OASIS Content Management Interoperability Services (CMIS)		I	U	I/U	I		
OASIS Digital Signature Service (DSS)		I	U	I/U			
OASIS Directory Services Markup Language (DSML) v2.0		I	U	I/U	I		
OASIS ebXML Messaging Services		I	U	I/U			
OASIS ebXML RegRep		I	U	I/U	I		
OASIS ebXML Registry Information Model		I	U	I/U			
OASIS ebXML Registry Services Specification		I	U	I/U			
OASIS eXtensible Access Control Markup Language (XACML)		I	U	I/U	I/U	I/U	
OASIS Message Queuing Telemetry Transport (MQTT)		I	U	I/U			
OASIS Open Data (OData) Protocol		I	U	I/U	I/U		
OASIS Search Web Services (SWS)		I	U	I/U			
OASIS Security Assertion Markup Language (SAML) v2.0		I	U	I/U	I/U	I/U	
OASIS SOAP-over-UDP (User Datagram Protocol) v1.1		I	U	I/U			
OASIS Solution Deployment Descriptor Specification v1.0	U						I/U
OASIS Symptoms Automation Framework (SAF) Version 1.0							I/U
OASIS Topology and Orchestration Specification for Cloud Applications Version 1.0	I/U			U	I		I/U
OASIS Universal Business Language (UBL) v2.1		I	U	I/U	U		

OASIS Universal Description, Discovery and Integration (UDDI) v3.0.2		I	U	I/U			U
OASIS Unstructured Information Management Architecture (UIMA) v1.0				U	I		
OASIS Unstructured Operation Markup Language (UOML) v1.0		I	U	I/U	I		
OASIS/W3C WebCGM v2.1		I	U	I/U	I		
OASIS Web Services Business Process Execution Language (WS-BPEL) v2.0	U			I			
OASIS/W3C - Web Services Distributed Management (WSDM): Management Using Web Services (MUWS) v1.1	U			I	I	U	U
OASIS WSDM: Management of Web Services (MOWS) v1.1	U			I	I	U	U
OASIS Web Services Dynamic Discovery (WS-Discovery) v1.1	U	I	U	I/U			U
OASIS Web Services Federation Language (WS-Federation) v1.2		I	U	I/U		U	
OASIS Web Services Notification (WSN) v1.3		I	U	I/U			
IETF Simple Network Management Protocol (SNMP) v3				I	I	I/U	U
IETF Extensible Provisioning Protocol (EPP)	U						I/U
NCPDPD Script standard	.	.	.	.	.	.	.
ASTM Continuity of Care Record (CCR) message	.	.	.	.	.	.	.
Healthcare Information Technology Standards Panel (HITSP) C32 HL7 Continuity of Care Document (CCD)	.	.	.	.	.	.	.
PMML Predictive Model Markup Language	.	.	.	.	.	.	.
Dash7							
H.265							
VP9							
Daala							
WebRTC							
X.509							
MDX							
NIEM-HLVA		I/U	I/U	I/U			
NIEM-MPD		I/U	I/U	I/U			
NIEM-Code List Specifications		I/U	I/U	I/U			
NIEM Conformance Specification		I/U	I/U	I/U			
NIEM-CTAS		I/U	I/U	I/U			
NIEM-NDR		I/U	I/U	I/U			
Non-Normative Guidance in Using NIEM with JSON		I/U	I/U	I/U			
DCC Data Package, version 1.0.0-beta.17 (a specification) released March of 2016							

DCC Observ-OM \							
DCC PREMIS							
DCC PROV							
DCC QuDEx							
DCC SDMX, specification 2.1 last amended May of 2012							
DCC TEI							
ISO/IEC 19123							
OGC® Coverage Implementation Schema (CIS)							
Open Group C172, O-BDL							
ISO 10646							
ISA-Tab							
Dublin Core							
BMC Visualization							
IEEE 1857.3							
W3C DCIP							
VoID							

Appendix D: Categorized Standards

Large catalogs of standards, such as the collection in Appendix B and C, describe the characteristics and relevance of existing standards. In the catalog format presented in Appendix D, the NBD-PWG strives to provide a structure for an ongoing process that supports continuous improvement of the catalog to ensure the usefulness of it in the years to come, even as technologies and requirements evolve over time.

The approach is to identify standards with one or more category terms, allowing readers to cross-reference the list of standards either by application domains or classes of activities defined in the NBDRA. The categorized standards could help to reduce the long list of standards to a shorter list that is relevant to the reader's area of concern.

Additional contributions from the public are invited. Please see the *Request for Contribution* in the front matter of this document for methods to submit contributions. First, contributors can identify standards that relate to application domains and NBDRA activities category terms and fill in the columns in Table E-1. Second, additional categorization columns could be suggested, which should contain classification terms and should be broad enough to apply to a majority of readers.

The application domains and NBDRA activities defined to date are listed below. Additional information on the selection of application domains is contained in the *NBDIF: Volume 3, Use Cases and Requirements*. The *NBDIF: Volume 6, Reference Architecture* expounds on the NBDRA activities.

Application domains defined to date:

- Government Operations
- Commercial
- Defense
- Healthcare and Life Sciences
- Deep Learning and Social Media
- The Ecosystem for Research
- Astronomy and Physics
- Earth, Environmental and Polar Science
- Energy
- IoT
- Multimedia

NBDRA classes of activities defined to date:

- **System Orchestrator (SO)**
  - Business Ownership Requirements and Monitoring
  - Governance Requirements and Monitoring
  - System Architecture Requirements Definition
  - Data Science Requirements and Monitoring
  - Security/Privacy Requirements Definition and Monitoring
- **Big Data Framework Provider (BDFP)**
  - Messaging
  - Resource Management
  - Processing: Batch Processing
  - Processing: Interactive Processing
  - Processing: Stream Processing
  - Platforms: Create
  - Platforms: Read
  - Platforms: Update
  - Platforms: Delete
  - Platforms: Index
  - Infrastructures: Transmit
  - Infrastructures: Receive
  - Infrastructures: Store
  - Infrastructures: Manipulate
  - Infrastructures: Retrieve
- **Security and Privacy (SP)**
  - Authentication
  - Authorization
  - Auditing
- **Management (M)**
  - Provisioning
  - Configuration
  - Package Management
  - Resource Management
  - Monitoring
- **Big Data Application Provider (BDAP)**
  - Collection
  - Preparation
  - Analytics
  - Visualization
  - Access

Whereas the task of categorization is immense and resources are limited, completion of this table relies on new and renewed contributions from the public. The NBD-PWG invites all interested parties to assist in the categorization effort.

Table D-1: Categorized Standards

Standard Name/Number	Application Domain	NBDRA Activities
ISO/IEC 9075-*		
ISO/IEC Technical Report (TR) 9789		
ISO/IEC 11179-*		

ISO/IEC 10728-*		
ISO/IEC 13249-*		
ISO/IEC TR 19075-*		
ISO/IEC 19503		
ISO/IEC 19773		
ISO/IEC TR 20943		
ISO/IEC 19763-*		
ISO/IEC 9281:1990		
ISO/IEC 10918:1994		
ISO/IEC 11172:1993		
ISO/IEC 13818:2013		
ISO/IEC 14496:2010	Multimedia coding (from IoT doc)	
ISO/IEC 15444:2011		
ISO/IEC 21000:2003		
ISO 6709:2008		
ISO 19115-*		
ISO 19110		
ISO 19139		
ISO 19119		
ISO 19157		
ISO 19114		
IEEE 21451 -*	IoT (from IoT doc)	
IEEE 2200-2012	IoT (from IoT doc)	
ISO/IEC 15408-2009		
ISO/IEC 27010:2012		
ISO/IEC 27033-1:2009		
ISO/IEC TR 14516:2002		
ISO/IEC 29100:2011		
ISO/IEC 9798:2010		SP: Authentication
ISO/IEC 11770:2010		
ISO/IEC 27035:2011		
ISO/IEC 27037:2012		

JSR (Java Specification Request) 221 (developed by the Java Community Process)		
W3C XML		
W3C Resource Description Framework (RDF)		
W3C JavaScript Object Notation (JSON)-LD 1.0		
W3C Document Object Model (DOM) Level 1 Specification		
W3C XQuery 3.0		
W3C XProc		
W3C XML Encryption Syntax and Processing Version 1.1		
W3C XML Signature Syntax and Processing Version 1.1		SP: Authentication
W3C XPath 3.0		
W3C XSL Transformations (XSLT) Version 2.0		
W3C Efficient XML Interchange (EXI) Format 1.0 (Second Edition)		
W3C RDF Data Cube Vocabulary		
W3C Data Catalog Vocabulary (DCAT)		
W3C HTML5 A vocabulary and associated APIs for HTML and XHTML		
W3C Internationalization Tag Set (ITS) 2.0		
W3C OWL 2 Web Ontology Language		
W3C Platform for Privacy Preferences (P3P) 1.0		
W3C Protocol for Web Description Resources (POWDER)		
W3C Provenance	Defense,	
W3C Rule Interchange Format (RIF)		
W3C Service Modeling Language (SML) 1.1		
W3C Simple Knowledge Organization System Reference (SKOS)		
W3C Simple Object Access Protocol (SOAP) 1.2		
W3C SPARQL 1.1		
W3C Web Service Description Language (WSDL) 2.0		
W3C XML Key Management Specification (XKMS) 2.0		
OGC® OpenGIS® Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile		

OGC® OpenGIS® GeoAPI		
OGC® OpenGIS® GeoSPARQL		
OGC® OpenGIS® Geography Markup Language (GML) Encoding Standard		
OGC® Geospatial eXtensible Access Control Markup Language (GeoXACML) Version 1		
OGC® network Common Data Form (netCDF)		
OGC® Open Modelling Interface Standard (OpenMI)		
OGC® OpenSearch Geo and Time Extensions		
OGC® Web Services Context Document (OWS Context)		
OGC® Sensor Web Enablement (SWE)		
OGC® OpenGIS® Simple Features Access (SFA)		
OGC® OpenGIS® Georeferenced Table Joining Service (TJS) Implementation Standard		
OGC® OpenGIS® Web Coverage Processing Service Interface (WCPS) Standard	BDFP processing, infrastructures, access, visualization, analytics	
OGC® OpenGIS® Web Coverage Service (WCS)	BDFP infrastructures, access	
OGC® Web Feature Service (WFS) 2.0 Interface Standard		
OGC® OpenGIS® Web Map Service (WMS) Interface Standard		
OGC® OpenGIS® Web Processing Service (WPS) Interface Standard		
OASIS AS4 Profile of ebMS 3.0 v1.0		
OASIS Advanced Message Queuing Protocol (AMQP) Version 1.0		
OASIS Application Vulnerability Description Language (AVDL) v1.0		
OASIS Biometric Identity Assurance Services (BIAS) Simple Object Access Protocol (SOAP) Profile v1.0		
OASIS Content Management Interoperability Services (CMIS)		
OASIS Digital Signature Service (DSS)		
OASIS Directory Services Markup Language (DSML) v2.0		
OASIS ebXML Messaging Services		
OASIS ebXML RegRep		

OASIS ebXML Registry Information Model		
OASIS ebXML Registry Services Specification		
OASIS eXtensible Access Control Markup Language (XACML)		
OASIS Message Queuing Telemetry Transport (MQTT)		
OASIS Open Data (OData) Protocol		
OASIS Search Web Services (SWS)		
OASIS Security Assertion Markup Language (SAML) v2.0		
OASIS SOAP-over-UDP (User Datagram Protocol) v1.1		
OASIS Solution Deployment Descriptor Specification v1.0		
OASIS Symptoms Automation Framework (SAF) Version 1.0		
OASIS Topology and Orchestration Specification for Cloud Applications Version 1.0		
OASIS Universal Business Language (UBL) v2.1		
OASIS Universal Description, Discovery and Integration (UDDI) v3.0.2		
OASIS Unstructured Information Management Architecture (UIMA) v1.0		BDAP: Analytics
OASIS Unstructured Operation Markup Language (UOML) v1.0		
OASIS/W3C WebCGM v2.1		BDAP: Visualization
OASIS Web Services Business Process Execution Language (WS-BPEL) v2.0		
OASIS/W3C - Web Services Distributed Management (WSDM): Management Using Web Services (MUWS) v1.1		
OASIS WSDM: Management of Web Services (MOWS) v1.1		
OASIS Web Services Dynamic Discovery (WS-Discovery) v1.1		
OASIS Web Services Federation Language (WS-Federation) v1.2		
OASIS Web Services Notification (WSN) v1.3		
IETF Simple Network Management Protocol (SNMP) v3		

IETF Extensible Provisioning Protocol (EPP)		
NCPDPD Script standard		
ASTM Continuity of Care Record (CCR) message		
Healthcare Information Technology Standards Panel (HITSP) C32 HL7 Continuity of Care Document (CCD)		
PMML Predictive Model Markup Language		
Dash7		
H.265		BDFP: Processing: Stream Processing;
VP9		BDFP: Processing: Stream Processing;
Daala		BDFP: Processing: Stream Processing;
WebRTC		
X.509		
MDX		
NIEM-HLVA	Government Operations, Defense, Commercial	BDAP: collection; BDFP: messaging
NIEM-MPD	Government Operations, Defense, Commercial	BDAP: collection; BDFP: messaging
NIEM-Code List Specifications	Government Operations, Defense, Commercial	BDAP: collection; BDFP: messaging
NIEM Conformance Specification	Government Operations, Defense, Commercial	BDAP: collection; BDFP: messaging
NIEM-CTAS	Government Operations, Defense, Commercial	BDAP: collection; BDFP: messaging
NIEM-NDR	Government Operations, Defense, Commercial	BDAP: collection; BDFP: messaging
Non-Normative Guidance in Using NIEM with JSON	Government Operations, Defense, Commercial	BDAP: collection; BDFP: messaging
DCC Data Package, version 1.0.0-beta.17 (a specification) released March of 2016		
DCC Observ-OM \		
DCC PREMIS		
DCC PROV		
DCC QuDEx		
DCC SDMX, specification 2.1 last amended May of 2012		
DCC TEI		
ISO/IEC 19123		
OGC® Coverage Implementation Schema (CIS)		
Open Group C172, O-BDL		
ISO 10646		
ISA-Tab		

Dublin Core		
BMC Visualization		
IEEE 1857.3		
W3C DCIP		
VoID		

Appendix E: 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, Big Data 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, Big Data 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 (SP1500-5)," 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, Big Data 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), G. von Laszewski (Editor), and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 8, Big Data Reference Architecture Interfaces (NIST SP 1500-9 VERSION 2)," Gaithersburg, MD, Sep. 2019 [Online]. Available: https://doi.org/10.6028/NIST.SP.1500-9r1

[8] W. L. Chang (Co-Chair), R. Reinsch (Subgroup Co-chair), C. Austin (Editor), and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 9, Adoption and Modernization (NIST SP 1500-10 VERSION 2)," Gaithersburg, MD, Sep. 2019 [Online]. Available: https://doi.org/10.6028/NIST.SP.1500-10r1

[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] F. Farance, “Adapted from the Refactoring Metadata Status Report,” 2016.

[11] P. Russom, “Data Lakes: Purposes, practices, patterns, and platforms,” 2017 [Online]. Available: https://info.talend.com/rs/talend/images/WP_EN_BD_TDWI_DataLakes.pdf

[12] NEXLA, “The Definitive Data Operations Report,” 2018 [Online]. Available: https://www.nexla.com/definitive-data-operations-report/

[13] K. Blind and S. Gauch, “Research and standardisation in nanotechnology: evidence from Germany,” *J. Technol. Transf.*, vol. 34, no. 3, pp. 320–342, 2009 [Online]. Available: http://link.springer.com/10.1007/s10961-008-9089-8

[14] C. Idoine, P. Krensky, E. Brethenoux, J. Hare, S. Sicular, and S. Vashisth, “Magic Quadrant for data science and machine-learning platforms,” Feb. 2018 [Online]. Available: https://www.gartner.com/en/documents/3860063/magic-quadrant-for-data-science-and-machine-learning-pla0

[15] G. De Simoni and R. Edjlali, “Magic Quadrant for Metadata Management Solutions,” *Gart. Repr.*, pp. 1–26, 2017 [Online]. Available: https://www.gartner.com/doc/3778891/magic-quadrant-metadata-management-solutions

[16] SAS, “The new data integration landscape: Moving beyond ad hoc ETL to an enterprise data integration strategy.” [Online]. Available: https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/new-data-integration-landscape-106221.pdf

[17] Cloud Security Alliance, “Expanded Top Ten Big Data Security and Privacy Challenges,” *Cloud Security Alliance*, 2013. [Online]. Available: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf

[18] A. DiStefano, K. E. Rudestam, and R. Silverman, *Encyclopedia of Distributed Learning*, Annotated. SAGE Publications, 2003 [Online]. Available: http://sk.sagepub.com/reference/distributedlearning

[19] *Revision of OMB Circular No. A-119, ‘Federal Participation in the Development and Use of Voluntary Consensus Standards and in Conformity Assessment Activities*. U.S. Office of Management and Budget, 2016, pp. 4673–4674 [Online]. Available: https://www.federalregister.gov/d/2016-01606

[20] G. Kopanitsa, H. Veseli, and V. Yampolsky, “Development, implementation and evaluation of an information model for archetype based user responsive medical data visualization,” *J. Biomed. Inform.*, vol. 55, no. June, pp. 196–205, 2015 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046415000763?via%3Dihub

[21] ISO/IEC JTC 1, “Big Data, Preliminary Report 2014,” 2015 [Online]. Available: https://www.iso.org/files/live/sites/isoorg/files/developing_standards/docs/en/big_data_report-

jtc1.pdf

[22] International Organization for Standardization (ISO), *ISO 23950:1998 Information and documentation — Information retrieval (Z39.50) — Application service definition and protocol specification*. International Organization for Standardization, 2014 [Online]. Available: https://www.iso.org/obp/ui/#iso:std:27446:en

[23] ANSI/NISO, *ANSI/NISO Z39.50-2003 (S2014), Information Retrieval (Z39.50): Application Service Definition and Protocol Specification*,. ANSI/NISO, 2015 [Online]. Available: https://groups.niso.org/apps/group_public/download.php/14978/z39-50-2003_s2014.pdf

[24] The Library of Congress, "CQL: Contextual Query Language," *Search/Retrival via URL*, 2013. [Online]. Available: http://www.loc.gov/standards/sru/cql/contextSets/. [Accessed: 02-Jul-2017]

[25] W3C, "Resource Description Framework (RDF)," *Semantic Web*, 2014. [Online]. Available: https://www.w3.org/RDF/. [Accessed: 02-Jul-2017]

[26] P. Baumann, D. Misev, V. Merticariu, B. P. Huu, B. Bell, and K.-S. Kuo, "Array Databases: Concepts, Standards, Implementations," 2019 [Online]. Available: https://www.rd-alliance.org/system/files/Array-Databases_final-report.pdf

[27] P. Baumann, D. Misev, V. Merticariu, and B. P. Huu, "Datacubes: Towards Space/Time Analysis-Ready Data," in *Service-Oriented Mapping: Changing Paradigm in Map Production and Geoinformation Management*, J. Döllner, M. Jobst, and P. Schmitz, Eds. Cham: Springer International Publishing, 2019, pp. 269–299 [Online]. Available: https://doi.org/10.1007/978-3-319-72434-8_14

[28] D. Misev and P. Baumann, "Extending the SQL Array Concept to Support Scientific Analytics," in *Proc. Intl. Conf. on Scientific and Statistical Database Management (SSDBM'2014)*, 2014, p. 11 [Online]. Available: https://dl.acm.org/citation.cfm?id=2618255

[29] International Organization for Standardization / International Electrotechnical Commission, *ISO/IEC 9075-15:2019, Information technology database languages— SQL—Part15:Multi-dimensional arrays (SQL/MDA)*. International Organization for Standardization / International Electrotechnical Commission, 2019 [Online]. Available: https://www.iso.org/standard/67382.html

[30] P. Baumann, "The OGC web coverage processing service (WCPS) standard," *Geoinformatica*, vol. 14, no. 4, pp. 447–479, Oct. 2010 [Online]. Available: https://doi.org/10.1007/s10707-009-0087-2

[31] A. Rauber and M. Parsons, "Scalable Dynamic Data Citation - RDA-WG-DC Position Paper," Mar. 2015 [Online]. Available: https://www.rd-alliance.org/groups/data-citation-wg/wiki/scalable-dynamic-data-citation-rda-wg-dc-position-paper.html

[32] Australian National Data Service (ANDS), "Citing dynamic data," 2017. [Online]. Available: https://www.ands.org.au/working-with-data/citation-and-identifiers/data-citation/citing-dynamic-data. [Accessed: 02-Mar-2019]

[33] A. Ball and M. Duke, "How to Cite Datasets and Link to Publications," *DCC How-to Guides.*

Edinburgh: Digital Curation Centre, 2015. [Online]. Available: http://www.dcc.ac.uk/resources/how-guides/cite-datasets

[34] E. D. P. and S. Committee, "Data Citation Guidelines for Earth Science Data. Ver. 2," Jun. 2019 [Online]. Available: https://esip.figshare.com/articles/Data_Citation_Guidelines_for_Earth_Science_Data_Version_2/8441816

[35] J. Klump, H. Robert, and L. A. I. Wyborn, "Scalable Approaches for Identifiers of Dynamic Data and Linked Data in an Evolving World," in *eResearch Australasia 2015*, 2015.

[36] L. Wyborn, N. Car, B. Evans, and J. Klump, "How do you assign persistent identifiers to extracts from large, complex, dynamic data sets that underpin scholarly publications?," in *EGU General Assembly, 2016*, 2016 [Online]. Available: http://meetingorganizer.copernicus.org/EGU2016/EGU2016-11639-1.pdf

[37] Kofax, "Integrating Data Sources is an Expensive Challenge for the Financial Services Sector (White Paper)," 2015 [Online]. Available: https://www.kofax.com/~/media/Files/Kofax/whitepaper/wp-integrating-data-sources-is-an-expensive-challenge-for-the-financial-services-sector-en.pdf

[38] G. I. Nagy and K. Buza, "SOHAC: Efficient Storage of Tick Data That Supports Search and Analysis," in *Advances in Data Mining. Applications and Theoretical Aspects*, 2012, pp. 38–51.

[39] W. Zhao, H. Ma, and Q. He, "Parallel K-Means Clustering Based on MapReduce," in *Cloud Computing*, 2009, pp. 674–679.

[40] A. Siddiqa, I. A. T. Hashem, I. Yaqoob, M. Marjani, and S. Shamshirband, "A survey of big data management: taxonomy and state of the art," *Elsevier J. Netw. Comput. Appl.*, vol. 71, pp. 151–166, Aug. 2016 [Online]. Available: https://doi.org/10.1016/j.jnca.2016.04.008

[41] T. Zhang, C. Du, and J. Wang, "Composite Quantization for Approximate Nearest Neighbor Search," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14)*, 2014, pp. II–838--II–846 [Online]. Available: http://proceedings.mlr.press/v32/levine14.pdf

[42] S. Paul, C. Boutsidis, M. Magdon-Ismail, and P. Drineas, "Random Projections for Linear Support Vector Machines," *CoRR*, vol. abs/1211.6, 2012 [Online]. Available: http://arxiv.org/abs/1211.6085

[43] E. Spaho, L. Barolli, F. Xhafa, A. Biberaj, and O. Shurdi, "P2P data replication and trustworthiness for a JXTA-Overlay P2P system using fuzzy logic," *Appl. Soft Comput.*, vol. 13, no. 1, pp. 321–328, Jan. 2013 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494612004012. [Accessed: 22-Aug-2019]

[44] D.-W. Sun, G.-R. Chang, S. Gao, L.-Z. Jin, and X.-W. Wang, "Modeling a Dynamic Data Replication Strategy to Increase System Availability in Cloud Computing Environments," *J. Comput. Sci. Technol.*, vol. 27, no. 2, pp. 256–272, Mar. 2012 [Online]. Available:

http://link.springer.com/10.1007/s11390-012-1221-4. [Accessed: 30-Nov-2018]

[45] CrowdFlower, “Data Science Report 2016,” 2016 [Online]. Available: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

[46] CrowdFlower, “Data Scientist Report 2017,” 2017 [Online]. Available: https://visit.figure-eight.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport.pdf

[47] Gartner, “Gartner Says CIOs and CDOs Must ‘Digitally Remaster’ Their Organizations,” Egham, UK, 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

[48] K. Cagle, “Understanding the Big Data Life-Cycle,” 2015. [Online]. Available: https://www.linkedin.com/pulse/four-keys-big-data-life-cycle-kurt-cagle. [Accessed: 10-Jun-2017]

[49] W. W. Eckerson, “How to Create a Culture of Governance,” *The New BI Leader*, 2017. [Online]. Available: https://www.eckerson.com/articles/how-to-create-a-culture-of-governance. [Accessed: 10-Jun-2017]

[50] C. Anderson, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete,” *Wired*, Jun-2008 [Online]. Available: https://www.wired.com/2008/06/pb-theory/

[51] McKinsey Global Institute, “The Age of Analytics: Competing in a Data-Driven World,” Dec. 2016 [Online]. Available: https://www.mckinsey.com/~/media/McKinsey/Business Functions/McKinsey Analytics/Our Insights/The age of analytics Competing in a data driven world/MGI-The-Age-of-Analytics-Full-report.ashx