NIST Special Publication 1500-7

NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap

Final Version 1

NIST Big Data Public Working Group
Technology Roadmap Subgroup

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.SP.1500-7

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

NIST Special Publication 1500-7

NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap

Final Version 1

NIST Big Data Public Working Group (NBD-PWG)
Technology Roadmap Subgroup
Information Technology Laboratory

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.SP.1500-7

September 2015

U. S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Under Secretary of Commerce for Standards and Technology and Director

National Institute of Standards and Technology (NIST) Special Publication 1500-7
46 pages (September 16, 2015)

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.

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

Big Data is a term used to describe the large amount of data in the networked, digitized, sensor-laden, information-driven world. While opportunities exist with Big Data, the data can overwhelm traditional technical approaches, and the growth of data is outpacing scientific and technological advances in data analytics. 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* series of volumes. This volume, Volume 7, contains summaries of the work presented in the other six volumes and an investigation of standards related to Big Data.

Keywords

Big Data; Big Data Application Provider; Big Data characteristics; Big Data taxonomy; Big Data standards; 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 of the NIST ITL, Robert Marcus of ET-Strategies, and Chaitanya Baru, University of California San Diego Supercomputer Center.

The document contains input from members of the NBD-PWG: Technology Roadmap Subgroup, led by Carl Buffington (Vistronix), David Boyd (InCadence Strategic Solutions), and Dan McClary (Oracle); Definitions and Taxonomies Subgroup, led by Nancy Grady (SAIC), Natasha Balac (SDSC), and Eugene Luster (R2AD); Use Cases and Requirements Subgroup, led by Geoffrey Fox (University of Indiana) and Tsegereda Beyene (Cisco); Security and Privacy Subgroup, led by Arnab Roy (Fujitsu) and Akhil Manchanda (GE); and Reference Architecture Subgroup, led by Orit Levin (Microsoft), Don Krapohl (Augmented Intelligence), and James Ketner (AT&T).

NIST SP1500-7, Version 1, has been collaboratively authored by the NBD-PWG. As of the date of this publication, there are over six hundred NBD-PWG participants from industry, academia, and government. Federal agency participants include the National Archives and Records Administration (NARA), National Aeronautics and Space Administration (NASA), National Science Foundation (NSF), and the U.S. Departments of Agriculture, Commerce, Defense, Energy, Health and Human Services, Homeland Security, Transportation, Treasury, and Veterans Affairs.

NIST acknowledges the specific contributions[a] to this volume by the following NBD-PWG members:

Chaitan Baru
University of California, San Diego, Supercomputer Center

David Boyd
InCadence Strategic Services

Carl Buffington
Vistronix

Pw Carey
Compliance Partners, LLC

Wo Chang
NIST

Yuri Demchenko
University of Amsterdam

Nancy Grady
SAIC

Keith Hare
JCC Consulting, Inc.

Bruno Kelpsas
Microsoft Consultant

Pavithra Kenjige
PK Technologies

Brenda Kirkpatrick
Hewlett-Packard

Donald Krapohl
Augmented Intelligence

Luca Lepori
Data Hold

Orit Levin
Microsoft

Jan Levine
kloudtrack

Serge Mankovski
CA Technologies

Robert Marcus
ET-Strategies

Gary Mazzaferro
AlloyCloud, Inc.

Shawn Miller
U.S. Department of Veterans Affairs

William Miller
MaCT USA

Sanjay Mishra
Verizon

Quyen Nguyen
NARA

John Rogers
Hwelett-Packard

Doug Scrimager
Slalom Consulting

Cherry Tom
IEEE-SA

Timothy Zimmerlin
Automation Technologies Inc.

The editors for this document were David Boyd, Carl Buffington, and Wo Chang

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

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, reference architecture, and roadmap. This document, *NIST Interoperability Framework: Volume 7, Standards Roadmap*, summarizes the deliverables of the other NBD-PWG subgroups (presented in detail in the other volumes of this series) and presents the work of the NBD-PWG Technology Roadmap Subgroup. In the first phase of development, the NBD-PWG Technology Roadmap Subgroup investigated existing standards that relate to Big Data and recognized general categories of gaps in those standards.

The *NIST Big Data Interoperability Framework* consists of seven volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The seven volumes are:

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

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

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

Potential areas of future work for the Subgroup during stage 2 are highlighted in Section 1.5 of this volume. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

1 INTRODUCTION

1.1 BACKGROUND

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

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

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

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

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

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

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

Motivated by the White House initiative and public suggestions, the National Institute of Standards and Technology (NIST) has accepted the challenge to stimulate collaboration among industry professionals to further the secure and effective adoption of Big Data. As one result of NIST's Cloud and Big Data Forum held on January 15–17, 2013, there was strong encouragement for NIST to create a public working group for the development of a Big Data Interoperability Framework. 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 requirements, and—from these—a standards roadmap. Such a consensus would create a vendor-neutral, technology- and infrastructure-independent framework that would enable Big Data stakeholders to identify and use the best analytics tools for their processing and visualization requirements on the most suitable computing platform and cluster, while also allowing value-added from Big Data service providers.

The *NIST Big Data Interoperability Framework* consists of seven volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The seven volumes are:

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

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

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

The NBDRA, created in Stage 1 and further developed in Stages 2 and 3, 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. It is discussed in detail in *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*. Potential areas of future work for the Subgroup during stage 2 are highlighted in Section 1.5 of this volume. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

1.2 NIST BIG DATA PUBLIC WORKING GROUP

The focus of the NBD-PWG is to form a community of interest from industry, academia, and government, with the goal of developing consensus-based Big Data definitions, taxonomies, reference architectures, and standards roadmap. The aim is to create vendor-neutral, technology- and infrastructure-agnostic deliverables to enable Big Data stakeholders to select the best analytics tools for their processing and visualization requirements on the most suitable computing platforms and clusters while allowing value-added from Big Data service providers and flow of data between the stakeholders in a cohesive and secure manner.

To achieve this goal, five subgroups were formed to address specific issues and develop the deliverables. These subgroups are as follows:

- NIST Big Data Definitions and Taxonomies Subgroup
- NIST Big Data Use Case and Requirements Subgroup
- NIST Big Data Security and Privacy Subgroup

- NIST Big Data Reference Architecture Subgroup
- NIST Big Data Technology Roadmap Subgroup

This volume and its companions were developed based on the following guiding principles:

- Deliverables are technologically agnostic;
- The audience is multi-sector, comprised of industry, government, and academia;
- Findings from all subgroups are aligned; and
- Deliverables represent the culmination of concepts from all subgroups.

1.3 SCOPE AND OBJECTIVES OF THE TECHNOLOGY ROADMAP SUBGROUP

The NBD-PWG Technology Roadmap Subgroup focused on forming a community of interest from industry, academia, and government, with the goal of developing a consensus vision with recommendations on how Big Data should move forward. The Subgroup's approach was to perform a gap analysis through the materials gathered from all other subgroups. This included setting standardization and adoption priorities through an understanding of what standards are available or under development as part of the recommendations. The goals of the Subgroup will be realized throughout the three planed phases of the NBD-PWG work, as outlined in Section 1.1. The primary tasks of the NBD-PWG Technology Roadmap Subgroup include the following:

- Gather input from NBD-PWG subgroups and study the taxonomies for the actors' roles and responsibility, use cases and general requirements, and secure reference architecture;
- Gain understanding of what standards are available or under development for Big Data;
- Perform a gap analysis and document the findings;
- Identify what possible barriers may delay or prevent adoption of Big Data; and
- Document vision and recommendations.

1.4 REPORT PRODUCTION

The *NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap* is one of seven volumes in the document, whose overall aims are to define and prioritize Big Data requirements, including interoperability, portability, reusability, extensibility, data usage, analytic techniques, and technology infrastructure in order to support secure and effective adoption of Big Data. The *NIST Big Data Interoperability Framework: 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.

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

- Section 2 summarizes the Big Data definitions presented in the *NIST Interoperability Framework: Volume 1, Definitions* document;
- Section 3 summarizes the assessment of the Big Data ecosystem, which was used to develop the NBDRA and this roadmap;
- Section 4 presents an overview of the NBDRA;
- Section 5 presents an overview of the security and privacy fabric of the NBDRA; and
- Section 6 investigates the standards related to Big Data and the gaps in those standards.

1.5 FUTURE WORK ON THIS VOLUME

The NIST Big Data Interoperability Framework will be released in three versions, which correspond to the three stages of the NBD-PWG work, as outlined in Section 1.1.

Version 2 activities will focus on the following:

- Continue to build and refine the gap analysis and document the findings;
- Identify where standards may accelerate the adoption and interoperability of Big Data technologies;
- Document recommendations for future standards activities; and
- Further map standards to NBDRA components and the interfaces between them.

2 BIG DATA DEFINITION

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 and measurement technology, associated terminology is defined in subsequent subsections. *NIST Big Data Infrastructure Framework: Volume 1, Definitions* contains additional details and terminology.

2.1 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 Section 3. 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 *NIST Big Data Interoperability Framework* and are fully described in *Volume 1*.

> ***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.*

2.2 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 end 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 empirical synthesis of actionable knowledge from raw data through the complete data life cycle process.*
>
> *The* ***data science paradigm*** *is extraction of actionable knowledge directly from data through a process of discovery, hypothesis, 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 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.*

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.*

3 INVESTIGATING THE BIG DATA ECOSYSTEM

The 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 information, 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 were used as input during the development of the entire NIST Big Data Interoperability Framework.

3.1 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
- Energy

Participants in the NBD-PWG Use Cases and Requirements Subgroup and other interested parties supplied publically 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 *NIST Big Data Interoperability Framework: 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 were as follows:

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

The general requirements, created to be vendor-neutral and technology-agnostic, are listed below.

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, statistical and graph 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, and 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 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, multi-dimension 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 pre-processing, 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.

Additional information about the Subgroup, use case collection, analysis of the use cases, and generation of the use case requirements are presented in the *NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements* document.

3.2 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 *NIST Big Data Interoperability Framework: Volume 5, Architectures White Paper Survey*.

3.3 TAXONOMY

The NBD-PWG Definitions and Taxonomy Subgroup developed a hierarchy of reference architecture components. Additional taxonomy details are presented in the *NIST Big Data Interoperability Framework: Volume 2, Taxonomy* document.

Figure 1 outlines potential actors for the seven roles developed by the NBD-PWG Definition and Taxonomy Subgroup. The dark blue boxes contain the name of the role at the top with potential actors listed directly below.

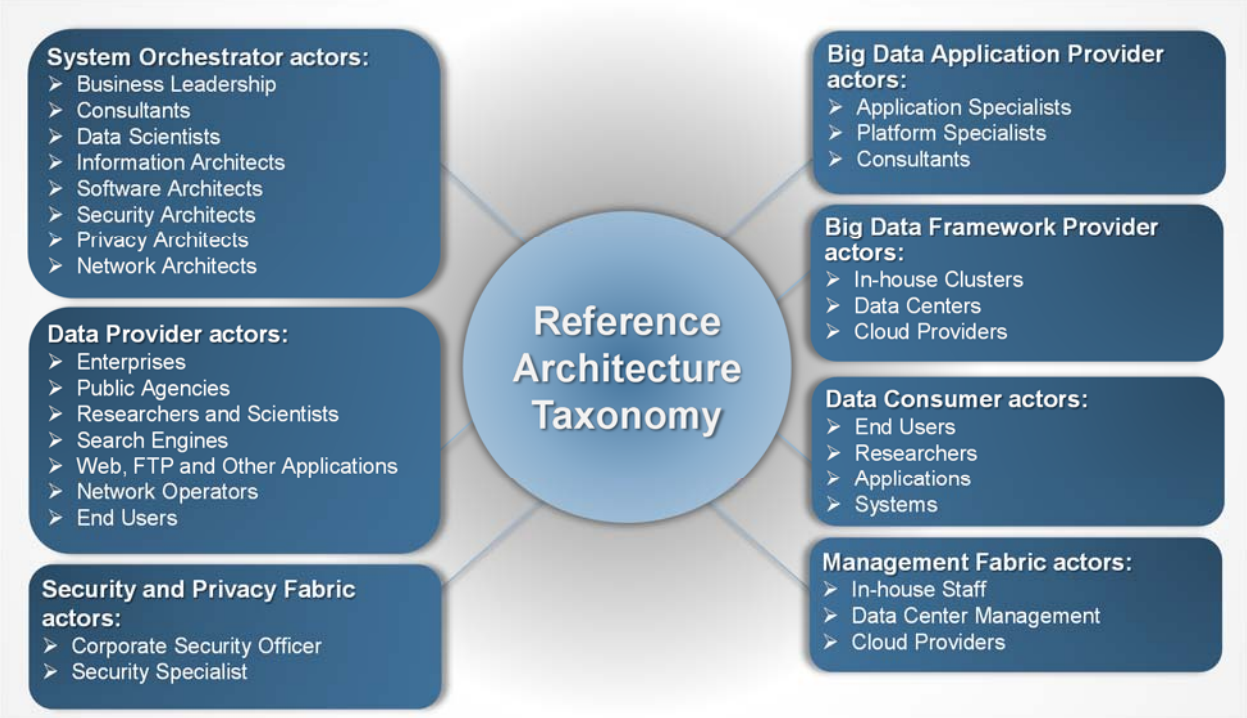

Figure 1: NIST Big Data Reference Architecture Taxonomy

4 BIG DATA REFERENCE ARCHITECTURE

4.1 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 *NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements* document contains details of the Subgroup's work.

The NBD-PWG Use Case Subgroup developed requirements in seven categories, which correspond to the reference architecture components as shown in Table 1. The requirements from each category were used as input for the development of the corresponding NBDRA component.

Table 1: Mapping of Use Case Categories to the NBDRA Components

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 fabric

4.2 NBDRA CONCEPTUAL MODEL

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 comprised 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.

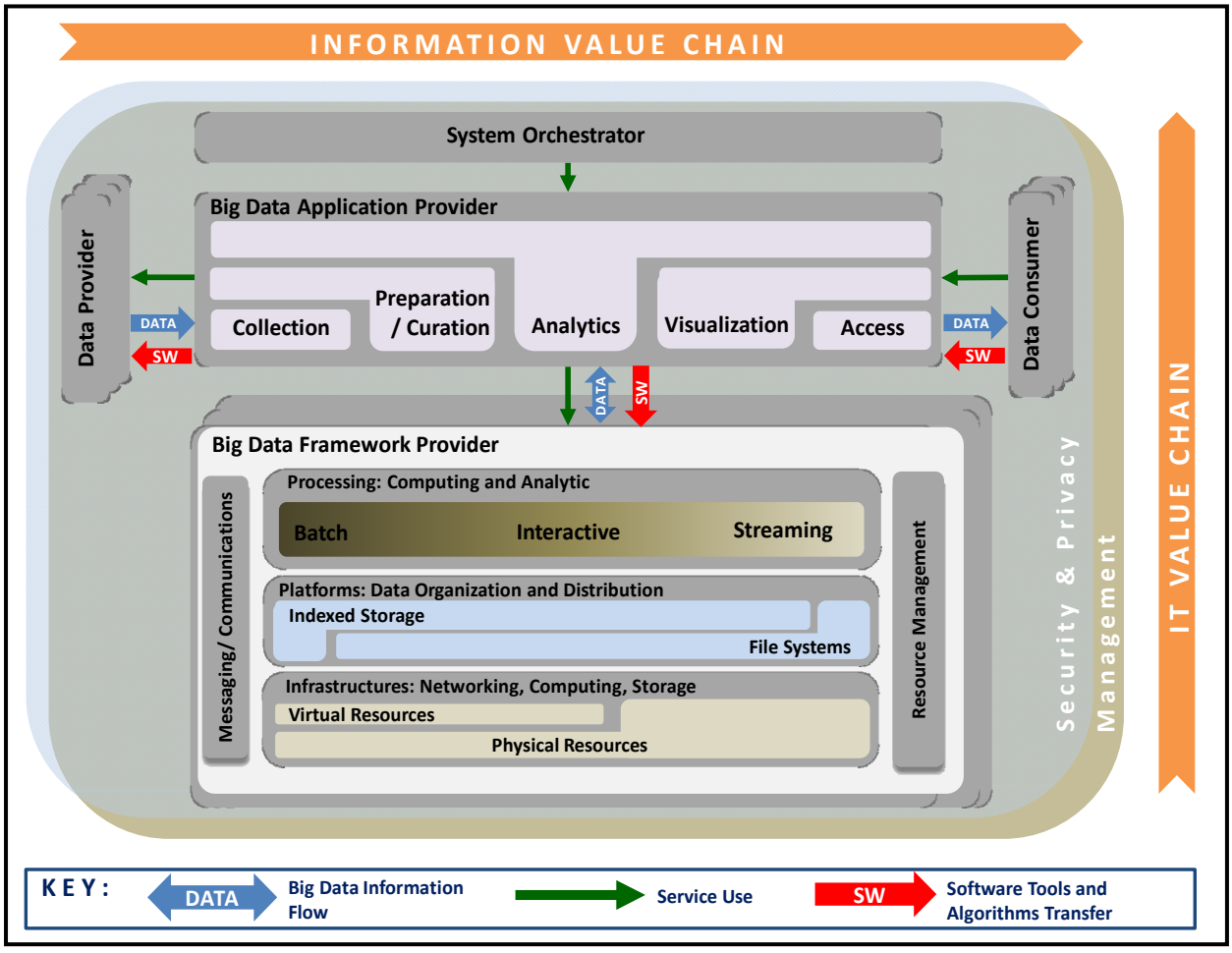

Figure 2: NBDRA Conceptual Model

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 two axes representing the two Big Data value chains: the information (horizontal axis) and the 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 component, indicating that data analytics and its implementation provide the value to Big Data stakeholders in both value chains. The names of the Big Data Application Provider and Big Data Framework Provider components contain "providers" to indicate that these components provide or implement a specific technical function within the system.

The five main NBDRA components, shown in Figure 2 and discussed in detail in Section 4, represent different technical roles that exist in every Big Data system. These functional components are:

- System Orchestrator
- Data Provider
- Big Data Application Provider
- Big Data Framework Provider
- Data Consumer

The two fabrics shown in Figure 2 encompassing the five functional components are:

- Management
- Security and Privacy

These two fabrics provide services and functionality to the five functional components 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 components. Data flows between the components 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 [SLAs]) and human interactions that may exist throughout the system are not shown in the NBDRA.

The components represent functional roles in the Big Data ecosystem. 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 five main components and the two fabrics of the NBDRA are discussed in the *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture* and *Volume 4, Security and Privacy*.

5 BIG DATA 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.[2]

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 of 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 *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy.*

6 BIG DATA STANDARDS

Big Data has generated interest in a wide variety of multi-stakeholder collaborative organizations, including those involved 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
- Open source implementations

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

- International Committee for Information Technology Standards (INCITS) and International Organization for Standardization (ISO)—de jure standards process
- Institute of Electrical and Electronics Engineers (IEEE)—de jure standards process
- International Electrotechnical Commission (IEC)
- Internet Engineering Task Force (IETF)
- World Wide Web Consortium (W3C)—Industry consortium
- Open Geospatial Consortium (OGC®)—Industry consortium
- Organization for the Advancement of Structured Information Standards (OASIS)—Industry consortium
- Open Grid Forum (OGF)—Industry consortium

The organizations and initiatives referenced in this document do not form an exhaustive list. It is anticipated that as this document is more widely distributed, more standards efforts addressing additional segments of the Big Data mosaic will be identified.

There are a number of 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 the majority of the audience for this report. Readers interested in domains covered by the government organizations and standards are encouraged to review the standards for applicability to their specific needs.

Open source implementations are providing useful new technology that is being used either directly or as the basis for commercially supported products. These open source implementations are not just individual products. One needs to integrate an ecosystem of products to accomplish one's goals. Because of the ecosystem complexity, and because of the difficulty of fairly and exhaustively reviewing open source implementations, 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.

6.1 EXISTING STANDARDS

This section presents a list of existing standards from the above listed organizations that are relevant to Big Data and the NBDRA. 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 veracity) 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 beyond the scope of the NBD-PWG.

In selecting standards to include in Table 2, the working group focused on standards that would do the following:

- 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.

Table 2 represents a portion of potentially applicable standards from a portion of contributing organizations working in Big Data domain.

As most standards represent some form of interface between components, Table 2 is annotated with 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 are abbreviated in the table header as follows:

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

Table 2: Existing Big Data Standards

Standard Name/Number	Description	NBDRA Components						
		SO	**DP**	**DC**	**BDAP**	**BDFP**	**S&P**	**M**
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.		I	I/U	U	I/U	U	U
ISO/IEC Technical Report (TR) 9789	Guidelines for the Organization and Representation of Data Elements for Data Interchange		I/U	I/U	I/U	I/U		
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		I	I/U	I/U		U	
ISO/IEC 10728-*	Information Resource Dictionary System Services Interface							
ISO/IEC 13249-*	Database Languages – SQL Multimedia and Application Packages		I	I/U	U	I/U		
ISO/IE 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		I	I/U	U	I/U		
ISO/IEC 19503	Extensible Markup Language (XML) Metadata Interchange (XMI)		I	I/U	U	I/U	U	

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
ISO/IEC 19773	Metadata Registries Modules		I	I/U	U	I/U	I/U	
ISO/IEC TR 20943	Metadata Registry Content Consistency		I	I/U	U	I/U	U	U
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		I	I/U	U	U		
ISO/IEC 9281:1990	Information Technology—Picture Coding Methods		I	U	I/U	I/U		
ISO/IEC 10918:1994	Information Technology—Digital Compression and Coding of Continuous-Tone Still Images		I	U	I/U	I/U		
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		I	U	I/U	I/U		
ISO/IEC 13818:2013	Information Technology—Generic Coding of Moving Pictures and Associated Audio Information		I	U	I/U	I/U		

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
ISO/IEC 14496:2010	Information Technology—Coding of Audio-Visual Objects		I	U	I/U	I/U		
ISO/IEC 15444:2011	Information Technology—JPEG (Joint Photographic Experts Group) 2000 Image Coding System		I	U	I/U	I/U		
ISO/IEC 21000:2003	Information Technology—Multimedia Framework (MPEG [Moving Picture Experts Group]-21)		I	U	I/U	I/U		
ISO 6709:2008	Standard Representation of Geographic Point Location by Coordinates		I	U	I/U	I/U		
ISO 19115-*	Geographic Metadata		I	U	I/U	U		
ISO 19110	Geographic Information Feature Cataloging		I	U	I/U			
ISO 19139	Geographic Metadata XML Schema Implementation		I	U	I/U			
ISO 19119	Geographic Information Services		I	U	I/U			
ISO 19157	Geographic Information Data Quality		I	U	I/U	U		
ISO 19114	Geographic Information—Quality Evaluation Procedures				I			
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		I	U				
IEEE 2200-2012	Standard Protocol for Stream Management in Media Client Devices		I	U	I/U			

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
ISO/IEC 15408-2009	Information Technology—Security Techniques—Evaluation Criteria for IT Security	U					I	
ISO/IEC 27033-1:2009	Information Technology—Security Techniques—Network Security		I/U	I/U	I/U	I		
ISO/IEC TR 14516:2002	Information Technology—Security Techniques—Guidelines for the Use and Management of Trusted Third Party Services	U					U	
ISO/IEC 29100:2011	Information Technology—Security Techniques—Privacy Framework						I	
ISO/IEC 9798:2010	Information Technology—Security Techniques—Entity Authentication		I/U	U	U	U	I/U	
ISO/IEC 11770:2010	Information Technology—Security Techniques—Key Management		I/U	U	U	U	I/U	
ISO/IEC 80000-13:2008	ISO/IEC 80000-13:2008 - Quantities and units -- Part 13: Information science and technology standard cancels and replaces subclauses 3.8 and 3.9 of IEC 60027-2:2005 (those related to Information theory and Prefixes for binary multiples).	U	U	U	U	I/U	U	U
ISO/IEC 27000	ISO/IEC 27000, Information security management systems—Overview and vocabulary	I	U	U	U	U	I/U	U
ISO/IEC 27001	ISO/IEC 27001, Information security management systems—Requirements	I	U	U	U	U	I/U	U
ISO/IEC 27002	ISO/IEC 27002, Code of practice for information security controls		I/U	I/U	U	U	I/U	U
ISO/IEC 27003	ISO/IEC 27003, Information security management system implementation guidance	I	U	U	U	U	U	U
ISO/IEC 27004	ISO/IEC 27004, Information security management—Measurement		I/U	I/U	U	U	I/U	U
ISO/IEC 27005	ISO/IEC 27005, Information security risk management	I	U	U	U	U	U	U

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
ISO/IEC 27006	ISO/IEC 27006, Requirements for bodies providing audit and certification of information security management systems		U	U	U	U	I/U	U
ISO/IEC 27007	ISO/IEC 27007, Guidelines for information security management systems auditing		U	U	U	U	I/U	U
ISO/IEC/TR 27008	ISO/IEC/TR 27008, Guidelines for auditors on information security controls	I						
ISO/IEC/DIS 27009	ISO/IEC/DIS 27009, Sector-specific application of ISO/IEC 27001–Requirements				I			
ISO/IEC 27010	ISO/IEC 27010, Information security management for inter-sector and inter-organizational communications		I/U	I/U	I/U		I/U	
ISO/IEC 27011	ISO/IEC 27011, Information security management guidelines for telecommunications organizations based on ISO/IEC 27002	I	U	U	U	U	U	U
ISO/IEC 27013	ISO/IEC 27013, Guidance on the integrated implementation of ISO/IEC 27001 and ISO/IEC 20000-1	I	U	U	U	U	U	U
ISO/IEC 27014	ISO/IEC 27014, Governance of information security	I	U	U	U	U	U	U
ISO/IEC/TR 27015	ISO/IEC/TR 27015, Information security management guidelines for financial services	I	U	U	U	U	U	U
ISO/IEC/TR 27016	ISO/IEC/TR 27016, Information security management — Organizational economics	I	U	U	U	U	U	U
ISO/IEC 27017	ISO/IEC 27017, Code of practice for information security controls based on ISO/IEC 27002 for cloud services	I	U	U	U	U	I/U	U
ISO/IEC 27018	ISO/IEC 27018, Code of practice for PII protection in public clouds acting as PII processors	I	U	U	U	U	I/U	U
ISO/IEC 27019	ISO/IEC 27019, Information security management guidelines based on ISO/IEC 27002 for process control systems specific to the energy utility industry	I	U	U	U	U	U	U

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
ISO/IEC 27035:2011	Information Technology—Security Techniques—Information Security Incident Management	U					I	
ISO/IEC 27037:2012	Information Technology—Security Techniques—Guidelines for Identification, Collection, Acquisition and Preservation of Digital Evidence	U					I	
JSR (Java Specification Request) 221 (developed by the Java Community Process)	JDBC™ 4.0 Application Programming Interface (API) Specification		I/U	I/U	I/U	I/U		
W3C XML	XML 1.0 (Fifth Edition) W3C Recommendation 26 November 2008	I/U	I/U	I/U	I/U	I/U	I/U	I/U
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.		I	U	I/U	I/U		
W3C JavaScript Object Notation (JSON)-LD 1.0	JSON-LD 1.0 A JSON-based Serialization for Linked Data W3C Recommendation 16 January 2014		I	U	I/U	I/U		
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.		I	U	I/U	I/U		
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.		I	U	I/U	I/U		
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.	I	I	U	I/U	I/U		

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
W3C XML Encryption Syntax and Processing Version 1.1	This specification covers a process for encrypting data and representing the result in XML.		I	U	I/U			
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.		I	U	I/U			
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.		I	U	I/U	I/U		
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.		I	U	I/U	I/U		
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.		I	U	I/U			
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.		I	U	I/U	I/U		
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.		I	U	I/U			

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
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.		I	U	I/U			
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.		I	U	I/U	I/U		
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.		I	U	I/U	I/U		
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.		I	U	I/U		I/U	
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 make a decision whether a given resource is of interest.		I	U	I/U			
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 interoperable interchange of provenance information in heterogeneous environments such as the Web.		I	U	I/U	I/U	U	

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
W3C Rule Interchange Format (RIF)	RIF is a series of standards for exchanging rules among rule systems, in particular among Web rule engines.		I	U	I/U	I/U		
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.	I/U	I	U	I/U			
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.		I	U	I/U			
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.		I	U	I/U			
W3C SPARQL 1.1	SPARQL is a language specification for the query and manipulation of linked data in a RDF format.		I	U	I/U	I/U		
W3C Web Service Description Language (WSDL) 2.0	This specification describes the WSDL Version 2.0, an XML language for describing Web services.	U	I	U	I/U			
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) and the XML Key Registration Service Specification (X-KRSS).	U	I	U	I/U			
OGC® OpenGIS® Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile	This series of standard covers Catalogue Services based on ISO19115/ISO19119, which are organized and implemented for the discovery, retrieval and management of data metadata, services metadata, and application metadata.		I	U	I/U			

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
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®.		I	U	I/U	I/U		
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.		I	U	I/U	I/U		
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.		I	U	I/U	I/U		
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.		I	U	I/U	I/U	I/U	
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.		I	U	I/U			

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
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.		I	U	I/U	I/U		
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.		I	U	I/U	I		
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.		I	U	I/U	I		
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.		I	U	I/U			
OGC® OpenGIS® Simple Features Access	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.		I	U	I/U	I/U		
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.		I	U	I/U	I/U		

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
OGC® OpenGIS® Web Coverage Processing Service Interface (WCPS) Standard	Defines a protocol-independent language for the extraction, processing, and analysis of multi-dimensional gridded coverages representing sensor, image, or statistics data.		I	U	I/U	I		
OGC® OpenGIS® Web Coverage Service (WCS)	This document specifies how a WCS offers multidimensional coverage data for access over the Internet. This document specifies a core set of requirements that a WCS implementation must fulfill.		I	U	I/U	I		
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.		I	U	I/U	I		
OGC® OpenGIS® Web Map Service (WMS) Interface Standard	The OpenGIS® WMS Interface Standard provides a simple HTTP interface for requesting geo-registered map images from one or more distributed geospatial databases.		I	U	I/U	I		
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.		I	U	I/U	I		
OASIS AS4 Profile of ebMS 3.0 v1.0	Standard for business to business exchange of messages via a web service platform.		I	U	I/U			

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
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.		I	U	U	I		
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.		I	U	I		U	
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.		I	U	I/U		U	
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.		I	U	I/U	I		
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.		I	U	I/U			

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
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		I	U	I/U	I		
OASIS ebXML Messaging Services	These specifications define a communications-protocol neutral method for exchanging electronic business messages as XML.		I	U	I/U			
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.		I	U	I/U	I		
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.		I	U	I/U			
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.		I	U	I/U			
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.		I	U	I/U	I/U	I/U	

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
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.		I	U	I/U			
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.		I	U	I/U	I/U		
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.		I	U	I/U			
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.		I	U	I/U	I/U	I/U	
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.		I	U	I/U			

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
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.	U						I/U
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.							I/U
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.	I/U			U	I		I/U
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.		I	U	I/U	U		
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.		I	U	I/U			U

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
OASIS Unstructured Information Management Architecture (UIMA) v1.0	The UIMA specification defines platform-independent data representations and interfaces for text and multimodal analytics.				U	I		
OASIS Unstructured Operation Markup Language (UOML) v1.0	UOML is an interface standard to process unstructured documents; it plays the similar role as SQL to structured data. UOML is expressed with standard XML.		I	U	I/U	I		
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.		I	U	I/U	I		
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.	U			I			
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.	U			I	I	U	U
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.	U			I	I	U	U
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.	U	I	U	I/U			U

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
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.		I	U	I/U		U	
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.		I	U	I/U			
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.				I	I	I/U	U
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.	U						I/U

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

6.2 GAP IN STANDARDS

The potential gaps in Big Data standardization are provided in this section to describe broad areas that may be of interest to SDOs, consortia, and readers of this document. The list provided below was produced by an ISO/IEC Joint Technical Committee 1 (JTC1) Study Group on Big Data to serve as a potential guide to ISO in their establishment of Big Data standards activities.[3] The potential Big Data standardization gaps, identified by the study group, described broad areas that may be of interest to this community. These gaps in standardization activities related to Big Data are 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, RDF, JSON, multimedia) and Big Data operations (e.g., 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 (e.g., memory storage system, distributed file system, data warehouse)
13. Human consumption of the results of big data analysis (e.g., 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
16. Big Data quality and veracity description and management

6.3 PATHWAY TO ADDRESS STANDARDS GAPS

Standards often 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 the case of Big Data, most standards are evolving from existing standards modified to address the unique characteristics of Big Data. 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. Recognizing this lack of a common framework on which to build standards, ISO/IEC JTC1 has specifically charted a working group, which will first focus on developing common definitions and a reference architecture. Once established, the definitions and reference architecture will form the basis for evolution 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 standards efforts may address these 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. 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 velocity environments. As actual implementations of technologies are proven, reference implementations will evolve based on community-accepted open source efforts.

Acronyms A: Acronyms

AMQP	Advanced Message Queuing Protocol
ANSI	American National Standards Institute
API	application programming interface
AVDL	Application Vulnerability Description Language
BDAP	Big Data Application Provider component
BDFP	Big Data Framework Provider component
BIAS	Biometric Identity Assurance Services
CGM	Computer Graphics Metafile
CIA	confidentiality, integrity, and availability
CMIS	Content Management Interoperability Services
CPR	Capability Provider Requirements
DC	Data Consumer component
DCAT	Data Catalog Vocabulary
DCR	Data Consumer Requirements
DOM	Document Object Model
DP	Data Provider component
DSML	Directory Services Markup Language
DSR	Data Source Requirements
DSS	Digital Signature Service
EPP	Extensible Provisioning Protocol
EXI	Efficient XML Interchange
GeoXACML	Geospatial eXtensible Access Control Markup Language
GML	Geography Markup Language
GRC	governance, risk management, and compliance
HTML	HyperText Markup Language
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
INCITS	International Committee for Information Technology Standards
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
MFI	Metamodel Framework for Interoperability
MOWS	Management of Web Services
MPEG	Moving Picture Experts Group
MQTT	Message Queuing Telemetry Transport
MUWS	Management Using Web Services
NARA	National Archives and Records Administration
NASA	National Aeronautics and Space Administration
NBD-PWG	NIST Big Data Public Working Group
NCAP	Network Capable Application Processor

netCDF	Network Common Data Form
NIST	National Institute of Standards and Technology
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
OpenMI	Open Modelling Interface Standard
OR	Other Requirements
OWS Context	OGC Web Services Context Document
P3P	Platform for Privacy Preferences Project
PICS	Platform for Internet Content Selection
POWDER	Protocol for Web Description Resources
RDF	Resource Description Framework
RFID	radio frequency identification
RIF	Rule Interchange Format
S&P	Security and Privacy Fabric
SAF	Symptoms Automation Framework
SAML	Security Assertion Markup Language
SDO	standards development organization
SKOS	Simple Knowledge Organization System Reference
SLA	service-level agreement
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
TEDS	Transducer Electronic Data Sheet
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
UOML	Unstructured Operation Markup Language
W3C	World Wide Web Consortium
WCPS	Web Coverage Processing Service Interface
WCS	Web Coverage Service
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	XSL Transformations

Appendix B: References

GENERAL RESOURCES

Institute of Electrical and Electronics Engineers (IEEE). https://www.ieee.org/index.html

International Committee for Information Technology Standards (INCITS). http://www.incits.org/

International Electrotechnical Commission (IEC). http://www.iec.ch/

International Organization for Standardization (ISO). http://www.iso.org/iso/home.html

Open Geospatial Consortium (OGC). http://www.opengeospatial.org/

Open Grid Forum (OGF). https://www.ogf.org/ogf/doku.php

Organization for the Advancement of Structured Information Standards (OASIS). https://www.oasis-open.org/

World Wide Web Consortium (W3C). http://www.w3.org/

DOCUMENT REFERENCES

[1] The White House Office of Science and Technology Policy, "Big Data is a Big Deal," *OSTP Blog*, accessed February 21, 2014, http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal.

[2] Cloud Security Alliance, *Expanded Top Ten Big Data Security and Privacy Challenges,* April 2013. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf.

[3] "Big Data, Preliminary Report 2014," ISO/IEC JTC1: Information Technology. http://www.iso.org/iso/big_data_report-jtc1.pdf. (Accessed March 2, 2015). Pages 21-23.