DRAFT NIST Big Data Interoperability Framework: Volume 6, Reference Architecture

Version 1

NIST Big Data Public Working Group (NBD-PWG)
Reference Architecture Subgroup
National Institute of Standards and Technology
Gaithersburg, MD 20899

Month 2014

U. S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Under Secretary of Commerce for Standards and Technology and Director
Authority

This publication has been developed by National Institute of Standards and Technology (NIST) to further its statutory responsibilities …

Nothing in this publication should be taken to contradict the standards and guidelines made mandatory and binding on Federal agencies …. 

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

There may be references in this publication to other publications currently under development by NIST in accordance with its assigned statutory responsibilities. The information in this publication, including concepts and methodologies, may be used by Federal agencies even before the completion of such companion publications. Thus, until each publication is completed, current requirements, guidelines, and procedures, where they exist, remain operative. For planning and transition purposes, Federal agencies may wish to closely follow the development of these new publications by NIST.

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. All NIST Information Technology Laboratory publications, other than the ones noted above, are available at http://www.nist.gov/publication-portal.cfm.

Comments on this publication may be submitted to:
National Institute of Standards and Technology
Attn: Information Technology Laboratory
100 Bureau Drive (Mail Stop 890os0) Gaithersburg, MD 20899-8930
Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (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. 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 Information Technology and its collaborative activities with industry, government, and academic organizations.

National Institute of Standards and Technology Special Publication XXX-series
xxx pages (April 23, 2014)

DISCLAIMER

This document has been prepared by the National Institute of Standards and Technology (NIST) and describes issues in Big Data computing.

Certain commercial entities, equipment, or material may be identified in this document in order to describe a concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that these entities, materials, or equipment are necessarily the best available for the purpose.
Acknowledgements

This document reflects the contributions and discussions by the membership of the NIST Big Data Public Working Group (NBD-PWG), co-chaired by Wo Chang of the NIST Information Technology Laboratory, 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: Reference Architecture Subgroup, led by Orit Levin (Microsoft), Don Krapohl (Augmented Intelligence), and James Ketner (AT&T); Technology Roadmap Subgroup, led by Carl Buffington (Vistronix), David Boyd (Data Tactic), 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).

NIST SP xxx-series, 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 would like to acknowledge the specific contributions to this volume by the following NBD-PWG members:

Chaitan Baru, University of California, San Diego, Supercomputer Center
Janis Beach, Information Management Services, Inc.
Scott Brim, Internet2
Gregg Brown, Microsoft
Carl Buffington, Vistronix
Jill Gemmill, Clemson University
Nancy Grady, SAIC
Ronald Hale, ISACA
Keith Hare, JCC Consulting, Inc.
Richard Jones, The Joseki Group LLC
Pavithra Kenjige, PK Technologies
James Kobielus, IBM
Donald Krapohl, Augmented Intelligence
Orit Levin, Microsoft
Eugene Luster, DISA/R2AD
Serge Manning, Huawei USA
Robert Marcus, ET-Strategies
Gary Mazzaferro, AlloyCloud, Inc.
Shawn Miller, U.S. Department of Veterans Affairs
Sanjay Mishra, Verizon
Vivek Navale, National Archives and Records Administration
Quyen Nguyen, National Archives and Records Administration
Felix Njeh, U.S. Department of the Army
Gururaj Pandurangi, Avyan Consulting Corp.
Linda Pelekoudas, Strategy and Design Solutions
Dave Raddatz, Silicon Graphics International Corp.
John Rogers, HP
Arnab Roy, Fujitsu
Michael Seablom, National Aeronautics and Space Administration
Rupinder Singh, McAfee, Inc.
Anil Srivastava, Open Health Systems Laboratory
Glenn Wasson, SAIC
Timothy Zimmerlin, Automation Technologies Inc.
Alicia Zuniga-Alvarado, Consultant

The editors for this document were Orit Levin and Wo Chang.
Table of Contents

Executive Summary .................................................................................................................................... 1
1 Introduction ......................................................................................................................................... 2  
  1.1 Background ................................................................................................................................. 2  
  1.2 Scope and Objectives of the Reference Architectures Subgroup .............................................. 3  
  1.3 Report Production ....................................................................................................................... 3  
  1.4 Report Structure .......................................................................................................................... 4  
2 High Level Reference Architecture Requirements .......................................................................... 5  
3 Conceptual Model .............................................................................................................................. 7  
4 Functional Components of the NBDRA ............................................................................................. 1  
  4.1 Data Provider ............................................................................................................................... 1  
  4.2 Big Data Application Provider .................................................................................................... 2  
  4.3 Big Data Framework Provider ..................................................................................................... 4  
    4.3.1 Physical and Virtual Resources ............................................................................................ 4  
    4.3.2 Infrastructures ...................................................................................................................... 4  
    4.3.3 Platforms ............................................................................................................................ 5  
    4.3.4 Processing Frameworks ....................................................................................................... 5  
  4.4 Data Consumer ............................................................................................................................. 6  
  4.5 System Orchestrator ...................................................................................................................... 6  
5 Management Component of the NBDRA ............................................................................................. 1  
  5.1 System Management ................................................................................................................... 1  
  5.2 Lifecycle Management ............................................................................................................... 1  
6 Security and Privacy Component of the NBDRA ............................................................................. 1  
7 Reference Architecture Taxonomy .................................................................................................... 1  
8 Future Directions ................................................................................................................................ 2  
Appendix A: Deployment Considerations ............................................................................................ A-1  
Appendix B: Terms and Definitions .................................................................................................... B-1  
Appendix C: Acronyms ........................................................................................................................ C-1  
Appendix D: Resources and References ............................................................................................... D-1  

Figures

Figure 1. NIST Big Data Reference Architecture .................................................................................. 7  
Figure A-1: Big Data Framework Deployment Options ......................................................................... A-1
Executive Summary

This *NIST Big Data Technology Roadmap Volume 6: Reference Architecture* was prepared by the NBD-PWG’s Reference Architecture Subgroup to provide a vendor-neutral, technology- and infrastructure-agnostic conceptual model and examine related issues. The conceptual model is based on the analysis of public Big Data material and inputs from the other NBD-PWG subgroups. The NIST Big Data Reference Architecture (NBDRA) was crafted by examining publicly available Big Data architectures representing various approaches and products. It is applicable to a variety of business environments, including tightly-integrated enterprise systems, as well as loosely-coupled vertical industries that rely on the cooperation by independent stakeholders. The NBDRA captures the two known Big Data economic value chains: the information flow, where the value is created by data collection, integration, analysis, and applying the results to data-driven services; and the IT industry, where the value is created by providing networking, infrastructure, platforms, and tools, in support of vertical data-based applications.

The other volumes that make up the NIST Big Data Roadmap are:

- Volume 1: Definitions
- Volume 2: Taxonomies
- Volume 3: Use Cases and General Requirements
- Volume 4: Security and Privacy Requirements
- Volume 5: Architectures White Paper Survey
- Volume 7: Technology Roadmap Summary

The authors emphasize that the information in these volumes represents a work in progress and will evolve as time goes on and additional perspectives are available.
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 our networked, digitized, sensor-laden, information-driven world. The availability of vast data resources carries the potential to answer questions previously out of reach, including the following:

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

However, 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 the 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. 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 our 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’s 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 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 overwhelming participation from 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 a consensus on definitions, taxonomies, secure reference architectures, security and privacy requirements, and a technology 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.

1.2 Scope and Objectives of the Reference Architectures Subgroup

Reference architectures provide “an authoritative source of information about a specific subject area that guides and constrains the instantiations of multiple architectures and solutions.” Reference architectures generally serve as a foundation for solution architectures and may also be used for comparison and alignment purposes.

The goal of the NBD-PWG Reference Architecture Subgroup was to develop a Big Data, open reference architecture that achieves the following objectives:

- Provide a common language for the various stakeholders
- Encourage adherence to common standards, specifications, and patterns
- Provide consistent methods for implementation of technology to solve similar problem sets
- Illustrate and improve understanding of the various Big Data components, processes, and systems, in the context of vendor and technology agnostic Big Data conceptual model
- Provide a technical reference for U.S. Government departments, agencies, and other consumers to understand, discuss, categorize, and compare Big Data solutions
- Facilitate the analysis of candidate standards for interoperability, portability, reusability, and extendibility

The reference architecture is intended to facilitate 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 NIST Big Data Reference Architecture does not address the following:

- Detailed specifications for any organization’s operational systems
- Detailed specifications of information exchanges or services
- Recommendations or standards for integration of infrastructure products

1.3 Report Production

There is a wide spectrum of Big Data architectures that have been explored and developed from various industries, academics, and government initiatives. The approach for developing the NIST Big Data Reference Architecture involved five steps:

1. Announce the NBD-PWG Reference Architecture Subgroup is open to the public in order to attract and solicit a wide array of subject matter experts and stakeholders in government, industry, and academia
2. Gather publicly available Big Data architectures and materials representing various stakeholders, different data types, and different use cases
3. Examine and analyze the Big Data material to better understand existing concepts, usage, goals, objectives, characteristics, and key elements of the Big Data, and then document the findings using NIST’s Big Data taxonomies model (presented in *NIST Big Data Technology Roadmap: Volume 2 Taxonomies*)
4. Develop an open reference architecture based on the analysis of Big Data material and the inputs from the other NBD-PWG subgroups
5. Produce this report to document the findings and work of the NBD-PWG Reference Architecture Subgroup

1.4 Report Structure
The organization of this document roughly follows the process used by the NBD-PWG to develop the NIST Big Data Reference Architecture (NBDRA). The remainder of this document is organized as follows:

- Section 2 contains high-level requirements relevant to the design of the NBDRA and discusses the development of these requirements
- Section 3 presents the generic NBDRA system
- Section 4 discusses the five main functional components of the NBDRA
- Section 5 describes the system and lifecycle management considerations
- Section 6 addresses security and privacy
- Section 7 outlines Big Data taxonomy with respect to security and privacy
- Appendix A summarizes deployment considerations
- Appendix B lists the terms and definitions
- Appendix C defines the acronyms used in this document
- Appendix D lists general resources and the references used in this document
2 High Level Reference Architecture Requirements

To begin the process of developing a reference architecture, publically available information was collected for various Big Data architectures used in several broad areas (application domains). The information on these use cases was collected from participants in the NBD-PWG Use Case and Requirements Subgroup and other interested parties. The NBD-PWG used a template for collection of the information to standardize the responses, which made it easier to analyze and compare the use cases. However, varied levels of detail and quantitative or qualitative information were received for each use case template section.

The NIST Big Data Technology Roadmap, Volume 3: Use Cases and General Requirements document presents the original use cases, an analysis of the compiled information, and the various levels of requirements extracted from the use cases. The original use cases can also be downloaded from the NIST document library (http://bigdatawg.nist.gov/usecases.php).

Building a list of requirements for the reference architecture involved two steps:

1. Extract requirements specific to each use case within the following characterization categories:
   - Data sources (e.g., data size, file formats, rate of growth, at rest or in motion)
   - Data transformation (e.g., data fusion, analytics)
   - Capabilities
   - Data consumer
   - Security and privacy
   - Lifecycle management (e.g., curation, conversion, quality check, pre-analytic processing)
   - Other requirements

2. Aggregate the use case-specific requirements into high-level, generalized requirements which are vendor- and technology-neutral

The high-level, generalized Big Data requirements were allocated to one of seven categories, which mirror the components defined in the NBDRA presented in Section 3. The categories and requirements are as follows:

**Data Provider Requirements**
- Reliable real time, asynchronous, streaming, and batch processing to collect data from centralized, distributed, and cloud data sources, sensors, or instruments
- Slow, bursty, and high throughput data transmission between data sources and computing clusters
- Diversified data content ranging from structured and unstructured text, documents, graphs, web sites, geospatial, compressed, timed, spatial, multimedia, simulation, and instrumental (i.e., system managements and monitoring) data

**Big Data Application Provider Requirements**
- Diversified compute intensive, analytic processing and machine learning techniques
- Batch and real time analytic processing
- Processing large diversified data content and modeling
- Processing data in motion (e.g., streaming, fetching new content, data tracking, traceability, data change management, and data boundaries)

**Big Data Framework Provider Requirements**
- Legacy software and advanced software packages
- Legacy and advanced computing platforms
- Legacy and advanced distributed computing clusters, co-processors, I/O processing
• Advanced networks (e.g., Software Defined Networks) and elastic data transmission, including fiber, cable, and wireless networks, LAN, WAN, MAN and WiFi
• Legacy, large, virtual and advanced distributed data storage
• Legacy and advanced programming executable, applications, tools, utilities, and libraries

Data Consumer Requirements
• Fast searches (~0.1 seconds) from processed data with high relevancy, accuracy, and high recall
• Diversified output file formats for visualization, rendering, and reporting
• Visual layout for results presentation
• Rich user interface for access using browser, visualization tools
• High resolution multi-dimension layer of data visualization
• Streaming results to clients

Security and Privacy Requirements
• Protect and preserve security and privacy on sensitive data
• Ensure multi-tenant, multi-level policy-driven, sandbox, access control, authentication on protected data is in line with accepted governance, risk, and compliance (GRC) and confidentiality, integrity, and availability (CIA) best practices

Lifecycle Management Requirements
• Data quality curation including pre-processing, data clustering, classification, reduction, format transformation
• Dynamic updates on data, user profiles, and links
• Data lifecycle and long-term preservation policy including data provenance
• Data validation
• Human annotation for data validation
• Prevention of data loss or corruption
• Multi-site (including cross-border, geographically dispersed) archival
• Persistent identifier and data traceability
• Standardization, aggregation, and normalization of data from disparate sources

Other Requirements
• Rich user interface from mobile platforms to access processed results
• Performance monitoring on analytic processing from mobile platforms
• Rich visual content search and rendering from mobile platforms
• Mobile device data acquisition and management
• Security across mobile devices and other smart devices such as sensors
3 Conceptual Model

The NBDRA shown in Figure 1 represents a technology-independent Big Data system comprised of logical functional components connected by interoperability interfaces (a.k.a., services). The components represent functional roles in the Big Data ecosystem and are called Providers to indicate that they provide or implement a specific technical function within the system.

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 stakeholders or actors, so the roles can either reside within the same business entity or can be implemented by different business entities. As such, the NBDRA is applicable to a variety of business environments, including tightly-integrated enterprise systems, as well as 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 this 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 is organized around two axes representing the two Big Data value chains: the information flow (horizontal axis) and the IT integration (vertical axis). Along the information flow 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 and operating of the Big Data in support of required data applications. Note that the Big Data Application Provider component is at the intersection of both axes, indicating that data analytics and its implementation are of special value to Big Data stakeholders in both value chains.

The five main NBDRA components, shown in Figure 1 and discussed in detail in Section 4, represent different technical roles that exist in every Big Data system. These functional components are as follows:

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

Two additional components (i.e., Security and Privacy, Management) provide services and functionality to the rest of the system components in the areas specific to Big Data. These two components, shown in Figure 1 as the fabrics enclosing the five functional components, are crucial to any Big Data solution.

The NBDRA supports 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 DATA arrows show the flow of data between the system’s main components. The 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.
4 Functional Components of the NBDRA

Five main functional components defined in the NBDRA represent the different technical roles within every Big Data system. The functional components are as follows:

- **Data Provider**: introduces new data or information feeds into the Big Data system
- **Big Data Application Provider**: executes a life cycle to meet security and privacy requirements as well as System Orchestrator-defined requirements
- **Big Data Framework Provider**: establishes a computing fabric in which to execute certain transformation applications while protecting the privacy and integrity of data
- **Data Consumer**: includes end users or other systems who utilize the results of the Big Data Application Provider
- **System Orchestrator**: defines and integrates the required data application activities into an operational vertical system

4.1 Data Provider

The Data Provider introduces new data or information feeds into the Big Data system for discovery, access, and transformation. New data feeds are distinct from the data already in use within the system and residing in the various system repositories, although similar technologies can be used to access both.

One important characteristic of a Big Data system is the ability to import and use data from a variety of data sources. Data sources can include internal and public records, online or offline applications, tapes, images, audio recordings, videos, Web logs, system and audit logs, HTTP cookies, humans, machines, sensors, and Internet technologies.

In its role, the Data Provider creates an abstraction of data sources. For example, in the case of raw data sources, a Data Provider can cleanse, correct, and store the data in a format that is accessible to the Big Data system via the Big Data Application Provider. Frequently, the role of Data Provider and Big Data Application Provider would belong to different organizations, unless the organization implementing the Big Data Application Provider owns the data sources. Consequently, data from different sources may have different security and privacy considerations.

The Data Provider can also provide an abstraction of data previously transformed by another system, which can be either a legacy system or another Big Data system. In this case, the Data Provider would represent a Data Consumer of that other system. For example, a Data Provider can generate a big streaming data source from another Big Data set at rest.

Data Provider activities include the following:

- Collect data
- Persist data
- Create metadata describing the data source(s), usage policies/access rights, and other relevant attributes
- Publish the availability of the information and the means to access it
- Make data accessible by other NBDRA components using a suitable programmable interface
- Provide push or pull access mechanisms
- Enforce access rights on data access
- Establish formal or informal contract for data access authorizations
- Provide transformation functions for data scrubbing of implicit or explicit personally identifiable information
The Data Provider will expose a collection of interfaces (or services) to facilitate discovery of and access to the data. These services typically include a registry so that the applications can locate a Data Provider; identify what data of interest it contains; understand what types of access are allowed and what types of analysis are supported; identify where the data source is located and how to access the data; and understand security requirements and privacy requirements for the data. As such, the interface would include the means for registering the data source, for querying the registry, and a standard set of data contained by the registry.

Because the data can be too large to economically move across the network, the interface could also allow the submission of analysis requests (as a software code implementing a certain algorithm for execution) with the results returned to the requestor. Subject to data characteristics (e.g., volume, velocity, and variety) and system design considerations, interfaces for exposing and accessing data vary in their complexity and may include both push and pull software mechanisms. These mechanisms can include subscription to events, listening to data feeds, querying for specific data properties or content, and the ability to submit a code for execution to process the data in situ. Data access may not always be automated, but rather might involve a human role logging into the system and providing directions to where the new data should be transferred (e.g., FTP).

### 4.2 Big Data Application Provider

The Big Data Application Provider executes a specific set of data lifecycle processes to meet the requirements established by the System Orchestrator, as well as the security and privacy requirements. The data lifecycle refers to the conversion of raw data into organized information and then into actionable knowledge through the processes of data collection, preparation, analytics and action. Big Data Application Provider activities are typically specific to the application and therefore are not candidates for standardization. However, especially when the application represents a vertical industry, the metadata and the policies defined and exchanged between the application’s sub-components could be standardized.

As the data spreads through the ecosystem, it is being processed and transformed in different ways to extract the value from the information. Each activity within the Big Data Application Provider can be implemented by independent stakeholders and deployed as stand-alone services. As such, the Big Data Application Provider can be a single application provider or a collection of more granulate application providers, each implementing different steps in the data lifecycle.

Big Data Application Provider activities include:

- **Collection**
  - Obtain connection to Data Provider APIs to connect into local system or to access dynamically when requested.
  - At the initial collection stage, sets of data (e.g., data records) of similar structure are collected (and combined), resulting in uniform security considerations and policies. Initial metadata (e.g., subjects with keys are identified) facilitate subsequent aggregation or lookup method(s).

- **Preparation**
  - Prepare the data through cleansing, outlier removal, and standardization for the ingestion and storage processes.
  - Aggregate data from different data providers with easily correlated metadata (e.g., identical keys) into a single data set. As a result, the information about each object is enriched or the number of objects in the collection grows. Security considerations and policies concerning the resultant data set are typically similar to the original data.
  - Match data from different data providers with disparate metadata (e.g., keys) into a single data set. As a result, the information about each object is enriched. The security
considerations and policies concerning the resultant data set may differ from the original policies.
  o Optimize data by determining the appropriate data manipulations and indexing to optimize subsequent transformation processes.

- **Analytics**
  o Extract knowledge from the data based on the requirements of the data scientist, who has specified the algorithms to process the data to produce new insights that will address the technical goal.

- **Visualization**
  o Ensure relevant visualization tools are integrated into data lifecycle system.
  o Format and present data to optimally communicate meaning and knowledge.
  o Develop appropriate statistical charts and diagrams to reflect analysis results.

- **Access**
  o Identify and store data in persistent repositories for use, sharing, and re-use.
  o Ensure descriptive, administrative, and preservation metadata and metadata schemes are developed and used.
  o Ensure secure access to data and information.
  o Curate storage media, file formats, etc., over time to account for changes and ensure data access for specified periods.

While many of these tasks have traditionally existed in data processing systems, the volume, velocity, and variety present in Big Data systems radically changes their implementation. The algorithms and mechanisms should be re-written and optimized to create applications that can respond to ever growing data collections.

Each of the activities can run on separate Big Data Framework Providers or all can use a common Big Data Framework Provider. The considerations behind these different system approaches would depend on (potentially different) technological needs, business and/or deployment constraints, including privacy and other policy considerations. The NBDRA does not show the underlying technologies, business considerations, and deployment constraints, thus making it applicable for any kind of system approaches and deployments.

For example, the Big Data Application Provider’s own infrastructure would be represented as one of the Big Data Framework Providers. If the Big Data Application Provider uses external/outsourced infrastructure as well, it will be represented as another Big Data Framework Provider in the NBDRA. Framework Providers are shown as many in the NBDRA, indicating that there can be as many as are being used for a single Big Data Application Provider.

The Big Data Application Provider will expose a collection of interfaces (or services) for consuming the results of the data processing performed by the Big Data system. These interfaces can be called and composed by 3rd party applications that have the permissions to consume the data. While the specifics of these interfaces may be application-specific, commonality will exist in a number of areas, including:

- **Data Transfer**: Facilitates secure transfer of data between different repositories and/or between the Big Data Application Provider and the Big Data Framework Providers, as well as transfer of data from a Framework Provider to a Data Consumer.
- **Identity Management and Authorization**: Individual vertical Big Data Application Providers may implement their own schemes for usage of their services. Identity management enables Big Data Application Providers to implement charging schemes, provide levels of differentiated service, and protect confidential services from unauthorized access.
• **Discovery**: Data Consumers require a directory that defines the Big Data Application Provider’s services.

• **Code execution services**: A Big Data Application Provider may allow Data Consumers to push analytics in the form of a code to execute on the Big Data system. The usage services will define the precise form that these requests support, for example the software languages that can be used, the constraints (e.g., execution time) on the code provided to the service, and how the results will be delivered to the Data Consumer.

• **Charging**: Big Data Application Providers may implement charging schemes to generate revenue from Data Consumers. The usage services will enable users to discover the amounts they are being charged, monitor usage, and make payments.

### 4.3 Big Data Framework Provider

The Big Data Framework Provider ensures a computing fabric (e.g., system hardware, network, storage, virtualization, computing platform) to enable execution of certain transformation applications while protecting the privacy and integrity of data. The computing fabric facilitates a mix-and-match of traditional and state-of-art computing features from software, platforms, and infrastructures based on the needs of each application. The computing fabric consists of the following:

- Processing Frameworks
- Platforms
- Infrastructures
- Physical and Virtual Resources

The Framework Provider consists of one or more instances of components typically hierarchically organized based on the IT Value Chain as shown in the RA diagram. There is no requirement that all instances at a given level in the hierarchy be of the same technology and in fact most Big Data implementations are hybrids combining multiple technology approaches in order to provide flexibility or meet the complete range of requirements which are driven from the Application Provider Framework. The four component areas that make up the Framework provider are Resources, Infrastructures, Platforms, and Processing. The following paragraphs briefly describe each of these component areas and more detailed discussions can be found in other NIST Big Data publications including the Technology Roadmap document and the Definitions and Taxonomy document.

#### 4.3.1 Physical and Virtual Resources

This component represents the raw storage and computing power that will be leveraged by the other framework components. While at some level all resources have a physical representation, the Big Data Framework components may be deployed directly on physical resources or on virtual resources. Frequently Big Data Framework deployments will employ a combination of physical and virtual resources. Virtual resources may be used to deploy either components that require additional elasticity (e.g., can change size dynamically) or that have specific high availability requirements that are easier to meet using virtual resource managers. Physical, resources in turn are frequently used to deploy horizontally scalable components that will be duplicated across a large number of physical nodes. Environmental resources such as power and heating, ventilation, and air conditioning (a.k.a., HVAC) are also part of this component layer since they are in fact finite resources that must be managed by the overall framework.

#### 4.3.2 Infrastructures

The infrastructure component defines how the physical and virtual resources and organized and connected. The three key sub elements of this are the cluster/computing, storage, and networking infrastructures. The cluster/computing infrastructure logical distribution may vary from a dense grid of physical commodity machines in a rack, to a set of virtual machines running on a cloud service provider,
to a loosely coupled set of machines distributed around the globe providing access to un-used computing resources. This infrastructure also frequently includes the underlying operating systems and associated services used to interconnect the cluster resources. The storage infrastructure may be likewise be organized as anything from isolated local disks to Storage Area Networks (SANs) or global distributed object stores.

The cluster/compute and storage components are in turn interconnected via the network infrastructure. The volume and velocity of Big Data often is a driving factor in the implementation of this component of the architecture. For example, if the implementation requires frequent transfers of large multi-gigabyte files between cluster nodes then high speed and low latency links are required. Depending on the availability requirements, redundant and fault tolerant links may be required. Other aspects of the network infrastructure include name resolution (e.g. DNS) and encryption along with firewalls and other perimeter access control capabilities. Finally, this layer may also include automated deployment/provisioning capabilities/agents and infrastructure wide monitoring agents that are leveraged by the management elements to implement a specific model.

4.3.3 Platforms
The platform element consists of the logical data organization and distribution combined with the associated access APIs or methods. That organization may range from simple delimited flat files to fully distributed relational or columnar data stores. Accordingly, the access methods may range from file access APIs to query languages such as SQL. A typical Big Data framework implementations would support both: a basic file system style storage and one or more indexed storage approaches. This logical organization may or may not be distributed across a cluster of computing resources based on the specific Big Data system considerations.

The platform may also include data registry and metadata services along with semantic data descriptions such as formal ontologies or taxonomies.

4.3.4 Processing Frameworks
Processing frameworks define how the computation/processing of the data is organized. Processing frameworks are generally divided along the lines of batch oriented and streaming oriented. However, depending on the specific data organization and platform, many Big Data frameworks span a range from high latency to near real time processing. Overall, many Big Data systems use multiple frameworks depending on the nature of the application.

Big Data applications rely on various platforms and technologies to meet their challenges of scalable data analytics and operation. The specific technological approaches vary widely; different vertical application domains will use a variety of technologies to meet their functional and cost requirements.

A broad categorization of the IT services that will be supported in a Big Data system are as follows:

- **Data Services**: A Big Data system exposes its data resources through a set of services that can be invoked by a Big Data Application Provider. The nature and granularity of these services vary, but generally they provide conventional create/read/update/delete (CRUD) functionality. The services are designed to efficiently support application requests, and commonly one service will invoke a cascade of internal IT service invocations to access a multitude of individual Big Data collections. As Big Data is often replicated, data services might expose functions that enable an application to explicitly trade off consistency and latency to more efficiently satisfy a request at the risk of obtaining stale data or performing inconsistent updates.

- **Security and Privacy Services**: The Big Data Framework Provider exposes services to perform identity management and provide authentication and authorization of the data and processing resources that are encompassed. This ensures resources are protected from unauthorized access.
and protects from tampering. This can be a particularly challenging area for Big Data systems that integrate heterogeneous data resources and/or execute on cloud platforms.

- **Automation Services:** A Big Data system will have many parts that operate dynamically. Automation is a fundamental principle in building Big Data systems, and hence automation services are an integral component of the provided services. Automation services range from virtual machine (VM) deployment and recovery, to fine grained monitoring of system performance and detecting and diagnosing faults.

- **Test Services:** A unique characteristic of Big Data systems is that it is impossible to fully test application changes before deployment, as the scale of the data and processing environment precludes exhaustive testing in an isolated environment. For this reason, Big Data systems often provide services to support testing of new features in a production environment. Techniques such as canary testing and A/B testing are used, require the ability to reconfigure the Big Data platform to direct percentages of live requests to test components, and provide detailed information and logging from the components under test.

- **Processing Services:** Supporting *in situ* processing allows Big Data Application Providers to push analytics to be performed by Big Data Framework Providers. To achieve this, the Big Data Application Provider provides services to receive the code for the analytics, execute the code in a protected environment, and return the results to the user. The latter is typically achieved asynchronously as many such analytics will be long running tasks that process many millions of data items.

### 4.4 Data Consumer

The Data Consumer is the role performed by end users or other systems who use the interfaces (or services) of Big Data Application Providers to get access to information of interest. These services can include data reporting, data retrieval, and data rendering.

Data Consumer activities include:

- Data search, query, and retrieval
- Exploring data using data visualization software
- Creating reports and organized drill-down using business intelligence software
- Ingesting data
- Putting data to work for the business, for example, by converting knowledge produced by the Big Data applications into a business rule transformation
- Conversion of data into additional data-derived products

The Data Consumer can also play the role of the Data Provider to the same system or to another system. The Data Consumer can provide requirements to the Big Data system as a user of the output of the system, whether initially or in a feedback loop.

### 4.5 System Orchestrator

The System Orchestrator defines and integrates the required data applications’ activities into an operational vertical system. Typically, the System Orchestrator would represent a collection of more specific roles performed by one or more actors, which manages and orchestrates the operation of the Big Data system.

The NBDRA represents a broad range of Big Data systems—from tightly-coupled enterprise solutions (integrated by standard or proprietary interfaces) to loosely-coupled verticals maintained by a variety of stakeholders or authorities bounded by agreements and standard or standard-de-facto interfaces.
In an enterprise environment, the System Orchestrator role is typically centralized and can be mapped to the traditional role of System Governor, which provides the overarching requirements and constraints that the system must fulfill (e.g., policy, architecture, resources, business requirements). The System Governor works with a collection of other roles (e.g., Data Manager, Data Security, and System Manager) to implement the requirements and the system’s functionality.

In a loosely-coupled vertical, the System Orchestrator role is typically decentralized. Each independent stakeholder is responsible for its system management, security, and integration. In this situation, each stakeholder is responsible for integration within the Big Data distributed system using the interfaces provided by other stakeholders.

In both tightly and loosely coupled cases, System Orchestrators assume the following responsibilities:

- Translate business goal(s) into technical requirements
- Supply and integrate with both external and internal Data Providers
- Oversee evaluation of data available from Data Providers
- Define requirements for the collection, preparation, and analysis of data
- Establish system architecture requirements
- Audit data application activities for compliance with requirements
- Define data dictionaries and data storage models
5 Management Component of the NBDRA

Big Data volume, velocity, and variety characteristics demand a versatile management platform for storing, processing and managing complex data. Big Data management involves system, data, security and privacy management considerations, with the goal of ensuring a high level of data quality and secure accessibility.

As discussed above, the NBDRA represents a broad range of Big Data systems—from tightly-coupled enterprise solutions (integrated by standard or proprietary interfaces) to loosely-coupled verticals maintained by a variety of stakeholders or authorities bounded by agreements and standard or standard-de-facto interfaces. Therefore different considerations and technical solutions would be applicable for different cases.

5.1 System Management

The complexity and volume of Big Data poses system management challenges on traditional management platforms. To efficiently capture, store, process, analyze, and distribute complex and large sized datasets arriving or leaving with high velocity, a resilient system management is needed.

In an enterprise environment, the management platform would typically provide enterprise-wide monitoring and administration of the Big Data distributed components. Network management, fault management, configuration management, system accounting, performance and security management are also of vital importance.

In a loosely-coupled vertical, each independent stakeholder is responsible for its system management, security, and integration. In this situation, each stakeholder is responsible for integration within the Big Data distributed system using the interfaces provided by other stakeholders.

5.2 Lifecycle Management

Lifecycle Management is responsible for monitoring data coming into the system, residing within the system, and going out of the system for application usage. In other words, the role of Lifecycle Management is to ensure that the data are being handled correctly by other NBDRA components in each process within the data lifecycle; from the moment they are ingested into the system by the Data Provider until the data are removed from the system. Because in the era of Big Data, the task of lifecycle management is distributed among different organization and/or individuals, this coordination of data processes between NBDRA components has greater difficulty in complying with policies, regulations, and security requirements.

Lifecycle Management activities include:

- **Metadata Management**: Metadata management enables Lifecycle Management, since metadata are used to store information that governs the lifecycle management of the data within the system. Metadata also contains critical information such as persistent identification of the data, the fixity, and the access rights.

- **Accessibility Management**:
  - Data masking for security privacy: Privacy information may need to be anonymized prior to the data analytics process. For instance, demographic data can be aggregated and analyzed to reveal data trends, but specific personally identifiable information (PII) with names and social security numbers should be masked. This masking managed by Lifecycle Management
depends on the type of application usage and the authorization usage specified by security and privacy.

- Accessibility of data may change over time. For instance, census data can be made available to the public after 75 years. In that case, Lifecycle Management is responsible for triggering the update of the accessibility of the data or sets of data according to the policy and legal requirements. Normally, data accessibility information is stored in the metadata.

- **Data Recovery**: Data management should also include recovering data that were lost due to disaster, or system/storage fault. Traditionally, this data recovery can be achieved using backup and restore mechanisms. To cope with the large volume of Big Data, however, this should be embedded in the architectural design, and the exploitation of modern technologies within the Big Data Framework Provider.

- **Preservation Management**: At the basic level, the system needs to ensure the integrity of the data so that the veracity and velocity of the analytics process are fulfilled. Due to the extremely large volume of Big Data, Preservation Management is responsible to disposition-aged data contained in the system. Depending on the retention policy, these aged data can be deleted or migrated to archival storage. On the other hand, in the case where data need to be retained for years, decades, and even centuries, a preservation strategy will be needed so the data can be accessed by the Provider Components if required. This will invoke the so-called long-term digital preservation that can be performed by Big Data Application Providers using the resources of the Big Data Framework Provider.

In the context of Big Data, Lifecycle Management has to deal with the three ‘V’ characteristics of volume, velocity, and variety. As such, Lifecycle Management and its components interact with other components of the NBDRA. Below are examples of such functions:

- **Data Provider**: manage the metadata from the entry of data into the Big Data system
- **Big Data Application Provider**: perform data masking and format transformations for preservation purpose
- **Big Data Framework Provider**: perform basic bit-level preservation and data recovery
- **Security and Privacy**: keep the data management up to date according to new security policy and regulations

In the other direction, Security and Privacy also uses information coming from Lifecycle Management with respect to data accessibility. Assuming that Security and Privacy controls access to the functions and data usage produced by the Big Data system, this data access control can be informed by the metadata managed and updated by Lifecycle Management.
6 Security and Privacy Component of the NBDRA

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 1 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 and ensure verifiable compliance with both governance, risk, and compliance (GRC); and confidentiality, integrity, and availability (CIA) regulations, standards, and best practices. This way, the role of Security and Privacy is depicted in the right relation to the components and at the same time does not explode into finer details, which may be more accurate but are best relegated to a more detailed security reference architecture.

In addition to the Big Data Application and Framework Providers components, the Data Provider and Data Consumer are enveloped in the security fabric since, at a minimum, the Data Provider and Data Consumer should agree on the security protocols and mechanisms in place.

In the NIST Big Data Roadmap Volume 4: Security and Privacy, Section 7 (Reference Architecture), additional guidance and best practices regarding these critical and sensitive areas of individual privacy and corporate security are discussed.
7 Reference Architecture Taxonomy

<Insert section from Definitions and Taxonomy Work Group with high-level taxonomy relevant to the design of Reference Architecture.>

1
8 Future Directions

The group of stakeholders preparing this volume set out to achieve the following:

- Illustrate the various Big Data components, processes, and systems to establish a common language for the various stakeholders
- Provide a technical reference for the industry to understand, discuss, categorize, and compare Big Data solutions
- Encourage adherence to common standards, specifications, and patterns by facilitating the analysis of candidate standards for interoperability, portability, reusability, and extendibility

Volume 6 represents consensus on the top-level reference architecture and the descriptions for its key components. The resultant NBDRA represents a vendor-neutral, technology- and infrastructure-agnostic conceptual model.

Future versions of this roadmap could include the definition of conceptual scenarios based on the real-life use cases examined and documented in the NIST Big Data Roadmap: Volume 3 Use Cases and General Requirements document. These scenarios should then be mapped to the NBDRA with the goal of identifying commonalities between the different scenarios, particularly in the interfaces between the NBDRA main components. Mapping the scenarios will help to validate the direction of defining abstract interfaces applicable across different scenarios and use cases. Actual implementation of scenarios will significantly help with correct mapping to the NBDRA and serve as a proof of concept of the NBDRA.
Appendix A: Deployment Considerations

The NIST Big Data Reference Architecture is applicable to a variety of business environments and technologies. As a result, possible deployment models are not part of the core concepts discussed in the main body of this document. However, this appendix demonstrates the application of the Big Data Framework Provider functional component to the two major deployment configurations: Cloud Computing environment and the traditional on-premise deployment model. Figure A-1 illustrates the application of the functional component to these two deployment configurations. The Big Data Framework Provider functional component implements various functionalities in order to support Big Data applications. These include infrastructures (e.g., VM clusters, storage, networking), platforms (e.g., programming languages and runtimes, databases), and applications (e.g., analytic tools). Traditional data frameworks have been implemented as dedicated, custom-made, on-premise systems. Recently, Big Data analytics and processing has been flourishing thanks to the adoption and use of cloud computing.

**Figure A-1: Big Data Framework Deployment Options**

**Cloud Service Providers**

Recent data analytics solutions use algorithms that can utilize and benefit from the frameworks of the cloud computing systems. Cloud computing has essential characteristics such as rapid elasticity and scalability, multi-tenancy, on-demand self-service and resource pooling, which together facilitate the realization of Big Data implementations.
The Cloud Service Provider (CSP) implements and delivers cloud services. Processing of a service invocation is done by means of an instance of the service implementation, which may involve the composition and invocation of other services as determined by the design and configuration of the service implementation.

**Cloud Service Component**

The cloud service component contains the implementation of the cloud services provided by a CSP. It contains and controls the software components that implement the services (but not the underlying hypervisors, host operating systems, device drivers, etc.)

Cloud services can be described in terms of cloud capability types and service categories. The cloud capability types describe the resources provided by the cloud service and are the following:

- Application capabilities
- Platform capabilities
- Infrastructure capabilities

Cloud services are also grouped into categories, where each service category is characterized by qualities that are common between the services within the category. The services in these categories may include capabilities from one or more of the capabilities types above. Some common cloud service categories include:

- Infrastructure as a Service (IaaS)
- Platform as a Service (PaaS)
- Software as a Service (SaaS)
- Network as a Service (NaaS)

The capabilities types inside the cloud service component are significant. The cloud service component offers its services via a service interface whose capabilities are in turn graphically represented by the inverted L diagrams. Each L represents the service interface(s) associated with the respective cloud capability type that is implemented by the cloud service category (e.g. IaaS, PaaS and SaaS).

The relative positioning of the L diagrams are also significant. The implication of the L shapes is that application capabilities can be implemented using platform capabilities, or not (at the choosing of the cloud service provider), and that platform capabilities can be implemented using infrastructure capabilities, or not.

**Resource Abstraction and Control Component**

The Resource Abstraction and Control component is used by cloud service providers to provide access to the physical computing resources through software abstraction. Resource abstraction needs to ensure efficient, secure, and reliable usage of the underlying physical resources. The control feature of the component enables the management of the resource abstraction features.

The Resource Abstraction and Control component enables a cloud service provider to offer qualities such as rapid elasticity, resource pooling, on-demand self-service and scale-out. The Resource Abstraction and Control component can include software elements such as hypervisors, virtual machines, virtual data storage, and time-sharing.

The Resource Abstraction and Control component enables control functionality. For example, there may be a centralized algorithm to control, correlate, and connect various processing, storage, and networking units in the physical resources so that together they deliver an environment where IaaS, PaaS or SaaS cloud service categories can be offered. The controller might decide which CPUs/racks contain which virtual machines executing which parts of a given cloud workload, and how such processing units are
connected to each other, and when to dynamically and transparently reassign parts of the workload to new units as conditions change.

**Traditional On-Premise Frameworks**

Traditional frameworks needed to support data analytics require custom-designed computing resources in order to satisfy the heavy demand on computing, storage, and database systems. Such on-premise systems usually use large, high performing custom-design databases designed to address unique demands of the data analytics workloads. Traditional solutions offer software, platform and cluster frameworks that can be used separately, or in conjunction with each other. The frameworks could be implemented in a layered fashion, or could be built vertically. Regardless, the implementations of such frameworks will require physical resources such as CPUs, storage devices, networking equipment, cooling and electrical facilities, etc.

**Physical Resources**

This layer represents physical resources needed for supporting the cloud computing system. Examples include CPUs, storage devices, networking equipment, and cooling and electrical facilities.
Appendix B: Terms and Definitions

NBDRA COMPONENTS

- **Big Data Engineering**: Advanced techniques that harness independent resources for building scalable data systems when the characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis.
- **Data Provider**: Organization or entity that introduces information feeds into the Big Data system for discovery, access, and transformation by the Big Data system.
- **Big Data Application Provider**: Organization or entity that executes a generic vertical system data lifecycle, including: (a) data collection from various sources, (b) multiple data transformations being implemented using both traditional and new technologies, (c) diverse data usage, and (d) data archiving.
- **Big Data Framework Provider**: Organization or entity that provides a computing fabric (such as system hardware, network, storage, virtualization, and computing platform) to execute certain Big Data applications, while maintaining security and privacy requirements.
- **Data Consumer**: End users or other systems that use the results of data applications.
- **System Orchestrator**: Organization or entity that defines and integrates the required data transformations components into an operational vertical system.

OPERATIONAL CHARACTERISTICS

- Extendability:
- Interoperability: The capability to communicate, to execute programs, or to transfer data among various functional units under specified conditions.
- Portability: The ability to transfer data from one system to another without being required to recreate or reenter data descriptions or to modify significantly the application being transported.
- Privacy: The assured, proper, and consistent collection, processing, communication, use and disposition of data associated with personal information (PI) and personally-identifiable information (PII) throughout its lifecycle.
- Reusability:
- Security: Protecting data, information, and systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide:
  - Integrity: guarding against improper data modification or destruction, and includes ensuring data nonrepudiation and authenticity
  - Confidentiality: preserving authorized restrictions on access and disclosure, including means for protecting personal privacy and proprietary data
  - Availability: ensuring timely and reliable access to and use of data

PROVISION MODELS

- Infrastructure as a Service (IaaS): The capability provided to the consumer to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls). (Source: NIST CC Definition)
- Platform as a Service (PaaS): The capability provided to the consumer to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages
and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations. (Source: NIST CC Definition)

- Software as a Service (SaaS): The capability provided to the consumer to use applications running on a cloud infrastructure. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings. (Source: NIST CC Definition)
Appendix C: Acronyms
Appendix D: Resources and References

GENERAL RESOURCES

The following resources provide additional information related to Big Data architecture.


DOCUMENT REFERENCES
