NIST Special Publication 1500-2

DRAFT NIST Big Data Interoperability Framework:

Volume 2, Big Data Taxonomies

NIST Big Data Public Working Group
Definitions and Taxonomies Subgroup

Draft Version 1
April 6, 2015

http://dx.doi.org/10.6028/NIST.SP.1500-2

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

NIST Special Publication 1500-2
Information Technology Laboratory

DRAFT NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies

Draft Version 1

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

April 2015

U. S. Department of Commerce
Penny Pritzker, Secretary

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

National Institute of Standards and Technology Special Publication 1500-2
32 pages (April 6, 2015)

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.

Public comment period: April 6, 2015 through May 21, 2015

Comments on this publication may be submitted to Wo Chang

National Institute of Standards and Technology
Attn: Wo Chang, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8900) Gaithersburg, MD 20899-8930
Email: SP1500comments@nist.gov

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at NIST promotes the U.S. economy and public welfare by providing technical leadership for the Nation’s measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. 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.

Abstract

Big Data is a term used to describe the new deluge of data in our networked, digitized, sensor-laden, information-driven world. While great opportunities exist with Big Data, it can overwhelm traditional technical approaches and its growth 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 questions related to Big Data. The results are reported in the *NIST Big Data Interoperability Framework* series of volumes. This volume, Volume 2, contains the Big Data taxonomies developed by the NBD-PWG. These taxonomies organize the reference architecture components, fabrics, and other topics to lay the groundwork for discussions surrounding Big Data.

Keywords

Big Data, Data Science, Reference Architecture, System Orchestrator, Data Provider, Big Data Application Provider, Big Data Framework Provider, Data Consumer, Security and Privacy Fabric, Management Fabric, Big Data taxonomy, use cases, Big Data characteristics

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: Definitions and Taxonomies Subgroup led by Nancy Grady (SAIC), Natasha Balac (SDSC), and Eugene Luster (R2AD); 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 SP1500-2, 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:

Natasha Balac
University of California, San Diego, Supercomputer Center

Chaitan Baru
University of California, San Diego, Supercomputer Center

Deborah Blackstock
MITRE Corporation

Pw Carey
Compliance Partners, LLC

Wo Chang
NIST

Yuri Demchenko
University of Amsterdam

Nancy Grady
SAIC

Karen Guertler
Consultant

Christine Hawkinson
U.S. Bureau of Land Management

Pavithra Kenjige
PK Technologies

Orit Levin
Microsoft

Eugene Luster
U.S. Defense Information Systems Agency/R2AD LLC

Bill Mandrick
Data Tactics

Robert Marcus
ET-Strategies

Gary Mazzaferro
AlloyCloud, Inc.

William Miller
MaCT USA

Sanjay Mishra
Verizon

Rod Peterson
U.S. Department of Veterans Affairs

John Rogers
HP

William Vorhies
Predictive Modeling LLC

Mark Underwood
Krypton Brothers LLC

Alicia Zuniga-Alvarado
Consultant

The editors for this document were Nancy Grady and Wo Chang.

Notice to Readers

NIST is seeking feedback on the proposed working draft of the *NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies*. Once public comments are received, compiled, and addressed by the NBD-PWG, and reviewed and approved by NIST internal editorial board, Version 1 of this volume will be published as final. Three versions are planned for this volume, with Versions 2 and 3 building on the first. Further explanation of the three planned versions and the information contained therein is included in Section 1.5 of this document.

Please be as specific as possible in any comments or edits to the text. Specific edits include, but are not limited to, changes in the current text, additional text further explaining a topic or explaining a new topic, additional references, or comments about the text, topics, or document organization. These specific edits can be recorded using one of the two following methods.

1. **TRACK CHANGES**: make edits to and comments on the text directly into this Word document using track changes
2. **COMMENT TEMPLATE**: capture specific edits using the Comment Template (http://bigdatawg.nist.gov/_uploadfiles/SP1500-1-to-7_comment_template.docx), which includes space for Section number, page number, comment, and text edits

Submit the edited file from either method 1 or 2 to SP1500comments@nist.gov with the volume number in the subject line (e.g., Edits for Volume 2.)

Please contact Wo Chang (wchang@nist.gov) with any questions about the feedback submission process.

Big Data professionals continue to be welcome to join the NBD-PWG to help craft the work contained in the volumes of the NIST Big Data Interoperability Framework. Additional information about the NBD-PWG can be found at http://bigdatawg.nist.gov.

Table of Contents

FIGURES

Executive Summary

This *NIST Big Data Interoperability Framework: Volume 2, Taxonomies* was prepared by the NIST Big Data Public Working Group (NBD-PWG) Definitions and Taxonomy Subgroup to facilitate communication and improve understanding across Big Data stakeholders by describing the functional components of the NIST Big Data Reference Architecture (NBDRA). The top-level roles of the taxonomy are System Orchestrator, Data Provider, Big Data Application Provider, Big Data Framework Provider, Data Consumer, Security and Privacy, and Management. The actors and activities for each of the top-level roles are outlined in this document as well. The NBDRA taxonomy aims to describe new issues in Big Data systems but is not an exhaustive list. In some cases, exploration of new Big Data topics includes current practices and technologies to provide needed context.

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 as follows:

- 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:

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 framework should define and prioritize Big Data requirements, including interoperability, portability, reusability, extensibility, data usage, analytics, and technology infrastructure. In doing so, the framework 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 standardsroadmap. 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 as follows:

- 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:

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 NIST Big Data Reference Architecture (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.

1.2 SCOPE AND OBJECTIVES OF THE DEFINITIONS AND TAXONOMIES SUBGROUP

The NBD-PWG Definitions and Taxonomy Subgroup focused on identifying Big Data concepts, defining terms needed to describe this new paradigm, and defining reference architecture terms. This taxonomy provides a hierarchy of the components of the reference architecture. It is designed to meet the needs of specific user groups, as follows:

For **managers**, the terms will distinguish the categorization of techniques needed to understand this changing field.

For **procurement officers**, it will provide the framework for discussing organizational needs and distinguishing among offered approaches.

For **marketers**, it will provide the means to promote Big Data solutions and innovations.

For the **technical community**, it will provide a common language to better differentiate Big Data's specific offerings.

1.3 REPORT PRODUCTION

This document derives from discussions in the NBD-PWG Definitions and Taxonomy Subgroup and with interested parties. This volume provides the taxonomy of the components of the NBDRA. This taxonomy was developed using a mind map representation, which provided a mechanism for multiple inputs and easy editing.

It is difficult to describe the new components of Big Data systems without fully describing the context in which they reside. The Subgroup attempted to describe only what has changed in the shift to the new Big Data paradigm, and only the components needed to clarify this shift. For example, there is no attempt to create a taxonomy of analytics techniques as these pre-date Big Data. This taxonomy will be a work in progress to mature as new technologies are developed and the patterns within data and system architectures are better understood.

In addition to the reference architecture taxonomy, the Subgroup began the development of a data hierarchy

1.4 REPORT STRUCTURE

This document provides multiple hierarchical presentations related to Big Data.

The first presentation is the taxonomy for the NBDRA. This taxonomy provides the terminology and definitions for the components of technical systems that implement technologies for Big Data. Section 2 introduces the NBDRA using concepts of actors and roles and the activities each performs. In the NBDRA presented in *NIST Big Data Interoperability Framework Volume 6: Reference Architecture*, there are two roles that span the activities within the other roles: Management, and Security and Privacy. These two topic areas will be addressed further in future versions of this document. The NBDRA components are more fully described in the *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture* and the *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy* documents. Comparing the related sections in these two documents will give the reader a more complete picture of the consensus of the working groups.

The second presentation is a hierarchical description about the data itself. For clarity, a strict taxonomy is not followed; rather, data is examined at different groupings to better describe what is new with Big Data. The grouping-based description presents data elements, data records, datasets, and multiple datasets. This examination at different groupings provides a way to easily identify the data characteristics that have driven the development of Big Data engineering technologies, as described in the *NIST Big Data Interoperability Framework: Volume 1, Definitions.*

Within the following sections, illustrative examples are given to facilitate understanding of the role/actor and activity of the NBDRA. There is no expectation of completeness in the components; the intent is to provide enough context to understand the specific areas that have changed because of the new Big Data paradigm. Likewise, the data hierarchy only expresses the broad overview of data at different levels of granularity to highlight the properties that drive the need for Big Data architectures.

For descriptions of the future of Big Data and opportunities to use Big Data technologies, the reader is referred to the *NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap*. Finally, to understand how these systems are architected to meet users' needs, the reader is referred to *NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements*.

1.5 FUTURE WORK ON THIS VOLUME

As mentioned in the previous section, the Subgroup is continuing to explore the changes in both Management and in Security and Privacy. As changes in the activities within these roles are clarified, the taxonomy will be developed further. In addition, a fuller understanding of Big Data and its technologies should consider the interactions between the characteristics of the data and the desired methods in both technique and time window for performance. These characteristics drive the application and the choice of tools to meet system requirements. Investigation of the interfaces between data characteristics and technologies is a continuing task for the NBD-PWG Definitions and Taxonomy Subgroup and the NBD-PWG Reference Architecture Subgroup. Finally, societal impact issues have not yet been fully explored. There are a number of overarching issues in the implications of Big Data, such as data ownership and data governance, which need more examination. Big Data is a rapidly evolving field, and the initial discussion presented in this volume must be considered a work in progress.

2 REFERENCE ARCHITECTURE TAXONOMY

This section focuses on a taxonomy for the NBDRA, and is intended to describe the hierarchy of actors and roles and the activities the actors perform in those roles. There are a number of models for describing the technologies needed for an application, such as a layer model of network, hardware, operating system, application. For elucidating the taxonomy, a hierarchy has been chosen to allow placing the new technologies within the context previous technologies. As this taxonomy is not definitive, it is expected that the taxonomy will mature as new technologies emerge and increase understanding of how to best categorize the different methods for building data systems.

2.1 ACTORS AND ROLES

In system development, actors and roles have the same relationship as in the movies. The roles are the parts the actors play in the overall system. One actor can perform multiple roles. Likewise, a role can be played by multiple actors, in the sense that a team of independents entities—perhaps from independent organizations—may be used to satisfy end-to-end system requirements. System development actors can represent individuals, organizations, software, or hardware. Each activity in the taxonomy can be executed by a different actor. Examples of actors include the following:

- Sensors
- Applications
- Software agents
- Individuals
- Organizations
- Hardware resources
- Service abstractions

In the past, data systems tended to be hosted, developed, and deployed with the resources of only one organization. Currently, roles may be distributed, analogous to the diversity of actors within a given cloud-based application. Actors in Big Data systems can likewise come from multiple organizations.

Developing the reference architecture taxonomy began with a review of the NBD-PWG analyses of the use cases and reference architecture survey provided in *NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements* and *NIST Big Data Interoperability Framework: Volume 5, Reference Architecture Survey*, respectively. From these analyses, several commonalities between Big Data architectures were identified and formulated into five general architecture components, and two fabrics interwoven in the five components, as shown in Figure 1.

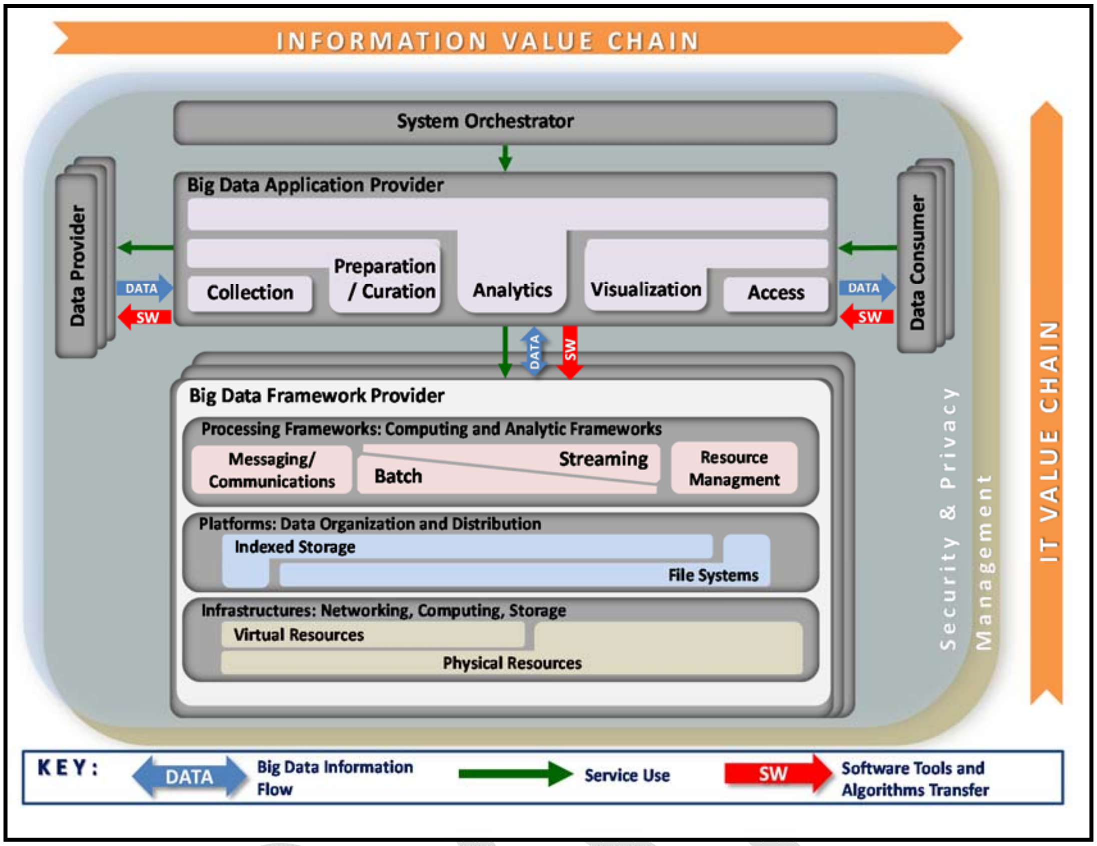

INFORMATION VALUE CHAIN
System Orchestrator
Big Data Application Provider
Collection
Preparation / Curation
Analytics
Visualization
Access
Data Provider
Data Consumer
DATA
SW
Big Data Framework Provider
Processing Frameworks: Computing and Analytic Frameworks
Messaging/ Communications
Batch
Streaming
Resource Managment
Platforms: Data Organization and Distribution
Indexed Storage
File Systems
Infrastructures: Networking, Computing, Storage
Virtual Resources
Physical Resources
Security & Privacy
Management
IT VALUE CHAIN
KEY:
DATA
Big Data Information Flow
Service Use
SW
Software Tools and Algorithms Transfer

Figure 1: NIST Big Data Reference Architecture

These seven items— five main architecture components and two fabrics interwoven in them—form the foundation of the reference architecture taxonomy.

The five main components, which represent the central roles, are summarized below and discussed in this section (Section 2).

- **System Orchestrator:** Defines and integrates the required data application activities into an operational vertical system
- **Data Provider:** Introduces new data or information feeds into the Big Data system
- **Big Data Application Provider:** Executes a lifecycle to meet security and privacy requirements as well as System Orchestrator-defined requirements
- **Big Data Framework Provider:** Establishes a computing framework in which to execute certain transformation applications while protecting the privacy and integrity of data
- **Data Consumer:** Includes end users or other systems who use the results of the Big Data Application Provider

The two fabrics, which are discussed separately in Sections 3 and 4, are as follows:

- **Security and Privacy Fabric**
- **Management Fabric**

Figure 2 outlines potential actors for the seven items listed above. The five central roles are explained in greater detail in the following subsections.

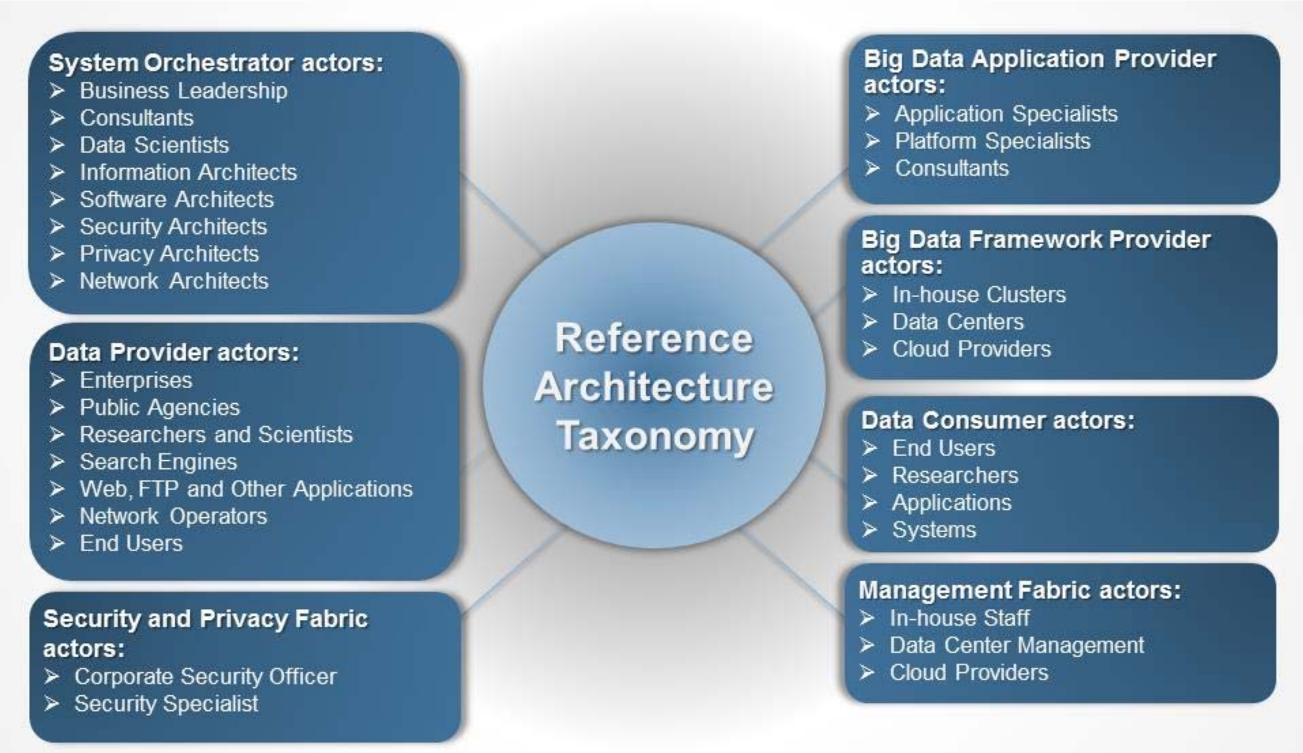


Figure 2: Roles and a Sampling of Actors in the NBDRA Taxonomy

2.2 SYSTEM ORCHESTRATOR

The System Orchestrator provides the overarching requirements that the system must fulfill, including policy, governance, architecture, resources, and business requirements, as well as monitoring or auditing activities to ensure the system complies with those requirements.

The System Orchestrator role includes defining and integrating the required data application activities into an operational vertical system. The System Orchestrator role provides system requirements, high-level design, and monitoring for the data system. While the role pre-dates Big Data systems, some related design activities have changed within the Big Data paradigm.

Figure 3 lists the actors and activities associated with the System Orchestrator, which are further described below.

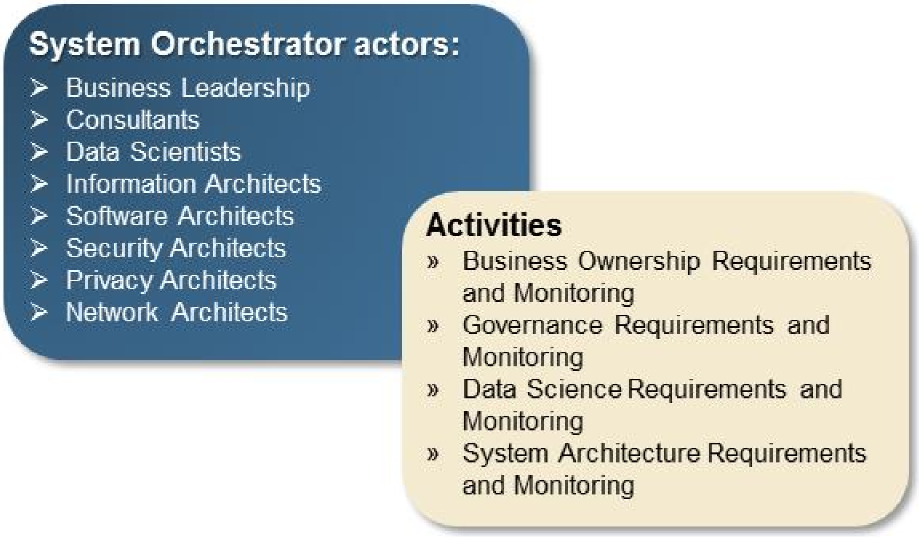

System Orchestrator actors:
➢ Business Leadership
➢ Consultants
➢ Data Scientists
➢ Information Architects
➢ Software Architects
➢ Security Architects
➢ Privacy Architects
➢ Network Architects
Activities
» Business Ownership Requirements and Monitoring
» Governance Requirements and Monitoring
» Data Science Requirements and Monitoring
» System Architecture Requirements and Monitoring

Figure 3: System Orchestrator Actors and Activities

A. Business Ownership Requirements and Monitoring

As the business owner of the system, the System Orchestrator oversees the business context within which the system operates, including specifying the following:

- Business goals
- Targeted business action
- Data Provider contracts and service-level agreements (SLAs)
- Data Consumer contracts and SLAs
- Negotiation with capabilities provider
- Make/buy cost analysis

A number of new business models have been created for Big Data systems, including Data as a Service (DaaS), where a business provides the Big Data Application Provider role as a service to other actors. In this case, the business model is to process data received from a Data Provider and provide the transformed data to the contracted Data Consumer.

B. Governance Requirements and Monitoring

The System Orchestrator establishes all policies and regulations to be followed throughout the data lifecycle, including the following:

- Policy compliance requirements and monitoring
- Change management process definition and requirements
- Data stewardship and ownership

Big Data systems potentially interact with processes and data being provided by other organizations, requiring more detailed governance and monitoring between the components of the overall system.

C. Data Science Requirements and Monitoring

The System Orchestrator establishes detailed requirements for functional performance of the analytics for the end-to-end system, translating the business goal into data and analytics design, including:

- Data source selection (e.g., identifying descriptions, location, file types, and provenance)
- Data collection and storage requirements and monitoring
- Data preparation requirements and monitoring
- Data analysis requirements and monitoring

- Analytical model choice (e.g., search, aggregation, correlation and statistics, and causal modeling)
- Data visualization requirements and monitoring
- Application type specification (e.g., streaming, real-time, and batch)

A number of the design activities have changed in the new paradigm. In particular, a greater choice of data models now exists beyond the relational model. Choosing a non-relational model will depend on the data type. Choosing the data fields that are used to decide how to distribute the data across multiple nodes will depend on the organization's data analysis needs, and on the ability to use those fields to distribute the data evenly across resources.

D. System Architecture Requirements and Monitoring

The System Orchestrator establishes detailed architectural requirements for the data system, including the following:

- Data process requirements
- Software requirements
- Hardware requirements
- Logical data modeling and partitioning
- Data export requirements
- Scaling requirements

The system architecture has changed in the Big Data paradigm due to the potential interplay between the different actors. The coordination between the five functional NBDRA components is more complex, with additional communications and interconnectivity requirements among the independently operated component activities. Maintaining the needed performance can lead to a very different architecture from that used prior to the new distribution of data across system nodes.

2.3 DATA PROVIDER

A Data Provider makes data available to itself or to others. The actor fulfilling this role can be part of the Big Data system, from another system, or internal or external to the organization orchestrating the system. Once the data is within the local system, requests to retrieve the needed data will be made by the Big Data Application Provider and routed to the Big Data Framework Provider. Data Provider actors include those shown in Figure 4.

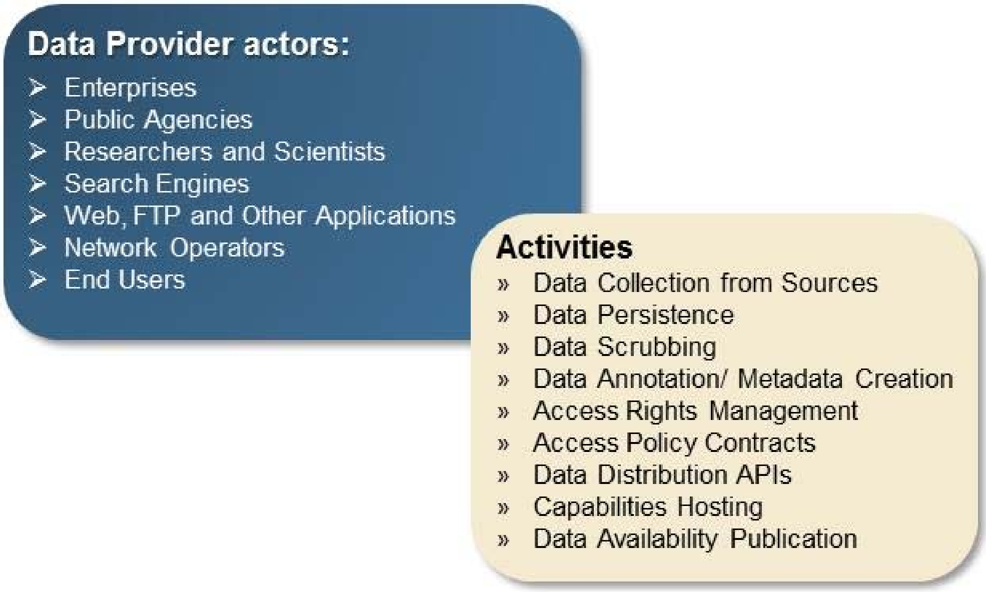


Figure 4: Data Provider Actors and Activities

While the concept of a Data Provider is not new, the greater data collection and analytics capabilities have opened up new possibilities for providing valuable data. The U.S. government's Open Data Initiative advocates that Federal agencies which are stewards of public data also serve the role of Data Provider.

The nine possible Data Provider activities outlined in Figure 4 are discussed further below.

A. Data Capture from Sources

The Data Provider captures data from its own sources or others. This activity could be described as the capture from a data producer, whether it is a sensor or an organizational process. Aspects of the data sources activity include both online and offline sources. Among possible online sources are the following:

- Web browsers
- Sensors
- Deep packet inspection devices (e.g., bridge, router, border controller)
- Mobile devices

Offline sources can include the following:

- Public records
- Internal records

While perhaps not theoretically different from what has been in use before, data capture from sources is an area that is exploding in the new Big Data paradigm. New forms of sensors are now providing not only a number of sources of data, but also data in large quantities. Smartphones and personal wearable devices (e.g., exercise monitors, household electric meters) can all be used as sensors. In addition, technologies such as radio frequency identification (RFID) chips are sources of data for the location of shipped items. Collectively, all the data-producing sensors are known as the Internet of Things (IoT). The subset of personal information devices are often referred to as "wearable tech", with the resulting data sometimes referred to as "digital exhaust".

B. Data Persistence

The Data Provider stores the data in a repository from which the data can be extracted and made available to others. The stored data is subject to a data retention policy. The data can be stored (i.e., persisted) in the following ways:

- Internal hosting
- External hosting
- Cloud hosting (a different hosting model whether internal or external)

Hosting models have expanded through the use of cloud computing. In addition, the data persistence is often accessed through mechanisms such as web services that hide the specifics of the underlying storage. DaaS is a term used for this kind of data persistence that is accessed through specific interfaces.

C. Data Scrubbing

Some datasets contain sensitive data elements that are naturally collected as part of the data production process. Whether for regulatory compliance or sensitivity, such data elements may be altered or removed. As one example of data scrubbing for Personally Identifiable Information (PII), the Data Provider can:

- Remove PII
- Perform data randomization

The latter obscures the PII to remove the possibility of directly tracing the data back to an individual, while maintaining the value distributions within the data. In the era of Big Data, data scrubbing requires greater diligence. While individual sources may not contain PII, when combined with other data sources, the risk arises that individuals may be identified from the integrated data.

D. Data Annotation and Metadata Creation

The Data Provider maintains information about the data and its processing, called metadata, in their repository, and also maintains the data itself. The metadata, or data annotation, would provide information about the origins and history of the data, in sufficient detail to enable proper use and interpretation of the data. The following approaches can be used to encode the metadata:

- In an ontology: a semantic description of the elements of the data
- Within a data file: in any number of formats

With the push for open data where data is repurposed to draw out additional value beyond the initial reason for which it was generated, it has become even more critical that information about the data be encoded to clarify the data’s origins and processing. While the actors that collected the data will have a clear understanding of the data history, repurposing data for other uses is open to misinterpretations when other actors use the data at a later time.

E. Access Rights Management

The Data Provider determines the different mechanisms that will be used to define the rights of access, which can be specified individually or by groupings such as the following:

- Data sources: the collection of datasets from a specific source
- Data producer: the collection of datasets from a given producer
- PII access rights: as an example of restrictions on data elements

F. Access Policy Contracts

The Data Provider defines policy for others’ use of the accessed data, as well as what data will be made available. These contracts specify:

- Policies for primary and secondary rights
- Agreements

To expand this description, the contracts specify acceptable use policies and any specific restrictions on the use of the data, as well as ownership of the original data and any derivative works from the data.

G. Data Distribution Application Programming Interfaces

Technical protocols are defined for different types of data access from data distribution application programming interfaces (APIs), which can include:

- File Transfer Protocol (FTP) or streaming
- Compression techniques (e.g., single compressed file, split compressed file,)
- Authentication methods
- Authorization

H. Capabilities Hosting

In addition to offering data downloads, the Data Provider offers several capabilities to access the data, including the following:

- Providing query access without transferring the data
- Allowing analytic tools to be sent to operate on the data sets

For large volumes of data, it may become impractical to move the data to another location for processing. This is often described as moving the processing to the data, rather than the data to the processing.

I. Data Availability Publication

The Data Provider makes available the information needed to know what data or data services they offer. Such publication can consist of the following:

- Web description
- Services catalog
- Data dictionaries
- Advertising

A number of third-party locations also currently publish a list of links to available datasets (e.g., U.S. Government's Open Data Initiative[2].)

2.4 BIG DATA APPLICATION PROVIDER

The Big Data Application Provider executes the manipulations of the data lifecycle to meet requirements established by the System Orchestrator, as well as meeting security and privacy requirements. This is where the general capabilities within the Big Data framework are combined to produce the specific data system. Figure 5 lists the actors and activities associated with the Big Data Application Provider.

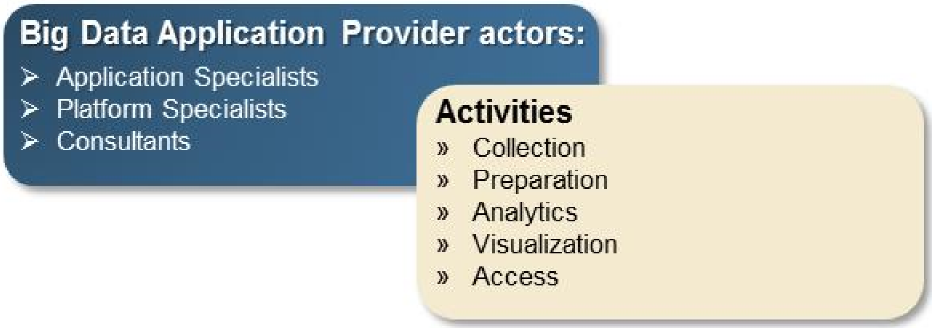


Figure 5: Big Data Application Provider Actors and Activities

While the activities of an application provider are the same whether the solution being built concerns Big Data or not, the methods and techniques have changed for Big Data because the data and data processing is parallelized across resources.

A. Collection

The Big Data Application Provider must establish the mechanisms to capture data from the Data Provider. These mechanisms include the following:

- Transport protocol and security
- Data format
- Metadata

While the foregoing transport mechanisms predate Big Data, the resources to handle the large volumes or velocities do result in changes in the way the processes are resourced.

B. Preparation

Whether processes are involved before or after the storage of raw data, a number of them are used in the data preparation activity, analogous to current processes for data systems. Preparation processes include the following:

- Data validation (e.g., checksums/hashes, format checks)
- Data cleansing (e.g., eliminating bad records/fields, deduplication)
- Outlier removal
- Data conversion (e.g., standardization, reformatting, and encapsulating)
- Calculated field creation and indexing
- Data aggregation and summarization
- Data partition implementation
- Data storage preparation
- Data virtualization layer

Just as data collection may require a number of resources to handle the load, data preparation may also require new resources or new techniques. For large data volumes, data collection is often followed by storage of the data in its raw form. Data preparation processes then occur after the storage and are handled by the application code. This technique of storing raw data first and applying a schema upon interaction with the data is commonly called “schema on read”, and is a new area of emphasis in Big Data due to the size of the datasets. When storing a new cleansed copy of the data is prohibitive, the data is stored in its raw form and only prepared for a specific purpose when requested.

Data summarization is a second area of added emphasis due to Big Data. With very large datasets, it is difficult to render all the data for visualization. Proper sampling would need some *a priori* understanding of the distribution of the entire dataset. Summarization techniques can characterize local subsets of the data, and then provide these characterizations for visualization as the data is browsed.

C. Analytics

The term data science is used in many ways. While it can refer to the end-to-end data lifecycle, the most common usage focuses on the steps of discovery (i.e., rapid hypothesis-test cycle) for finding value in big volume datasets. This rapid analytics cycle (also described as agile analytics) starts with quick correlation or trending analysis, with greater effort spent on hypotheses that appear most promising.

The analytics processes for structured and unstructured data have been maturing for many years. There is now more emphasis on the analytics of unstructured data because of the greater quantities now available. The knowledge that valuable information resides in unstructured data promotes a greater attention to the analysis of this type of data.

While analytic methods have not changed with Big Data, their implementation has changed to accommodate parallel data distribution across a cluster of independent nodes and data access methods. For example, the overall data analytic task may be broken into subtasks that are assigned to the independent data nodes. The results from each subtask are collected and compiled to achieve the final full dataset analysis. Furthermore, data often resided in simple tables or relational databases. With the introduction of new storage paradigms, analytics techniques should be modified for different types of data access.

Some considerations for analytical processes used for Big Data or small data are the following:

- Metadata matching processes
- Analysis complexity considerations (e.g., computational, machine learning, data extent, data location)
- Analytics latency considerations (e.g., real-time or streaming, near real-time or interactive, batch or offline)
- Human-in-the-loop analytics lifecycle (e.g., discovery, hypothesis, hypothesis testing)

While these considerations are not new to Big Data, implementing them can be tightly coupled with the specifics of the data storage and the preparation step. In addition, some of the preparation tasks are done during the analytics phase (the schema-on-read discussed above).

D. Visualization

While visualization (or the human in the loop) is often placed under analytics, the added emphasis due to Big Data warrants separate consideration of visualization. The following are three general categories of data visualization:

- Exploratory data visualization for data understanding (e.g., browsing, outlier detection, boundary conditions)
- Explicatory visualization for analytical results (e.g., confirmation, near real-time presentation of analytics, interpreting analytic results)
- Explanatory visualization to “tell the story” (e.g., reports, business intelligence, summarization)

Data science relies on the full dataset type of discovery or exploration visualization from which the data scientist would form a hypothesis. While clearly predating Big Data, a greater emphasis now exists on exploratory visualization, as it is immensely helpful in understanding large volumes of repurposed data because the size of the datasets requires new techniques.

Explanatory visualization is the creation of a simplified, digestible visual representation of the results, suitable for assisting a decision or communicating the knowledge gained. Again, while this technique has long been in use, there is now greater emphasis to “tell the story”. Often this is done through simple visuals or “infographics”. Given the large volumes and varieties of data, and the data’s potentially complex relationships, the communication of the analytics to a non-analyst audience requires careful visual representation to communicate the results in a way that can be easily consumed.

E. Access

The Big Data Application Provider gives the Data Consumer access to the results of the data system, including the following:

- Data export API processes (e.g., protocol, query language)
- Data charging mechanisms
- Consumer analytics hosting
- Analytics as a service hosting

The access activity of the Big Data Application Provider should mirror all actions of the Data Provider, since the Data Consumer may view this system as the Data Provider for their follow-on tasks. Many of the access-related tasks have changed with Big Data, as algorithms have been rewritten to accommodate for and optimize the parallelized resources.

2.5 BIG DATA FRAMEWORK PROVIDER

The Big Data Framework Provider has general resources or services to be used by the Big Data Application Provider in the creation of the specific application. There are many new technologies from which the Big Data Application Provider can choose in using these resources and the network to build the specific system. Figure 6 lists the actors and activities associated with the Big Data Framework Provider.

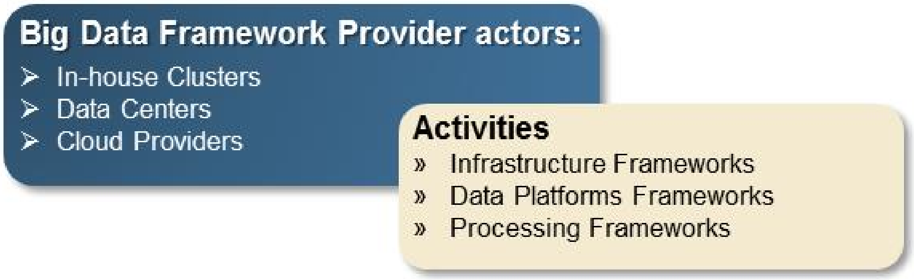

Big Data Framework Provider actors:
➢ In-house Clusters
➢ Data Centers
➢ Cloud Providers
Activities
» Infrastructure Frameworks
» Data Platforms Frameworks
» Processing Frameworks

Figure 6: Big Data Framework Provider Actors and Activities

The Big Data Framework Provider role has seen the most significant changes with the introduction of Big Data. The Big Data Framework Provider consists of one or more instances of the three subcomponents or activities: infrastructure frameworks, data platform frameworks, and processing frameworks. 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. These provide flexibility and can meet the complete range of requirements that are driven from the Big Data Application Provider. Due to the rapid emergence of new techniques, this is an area that will continue to need discussion. As the Subgroup continues its discussion into patterns within these techniques, different orderings will no doubt be more representative and understandable.

A. Infrastructure Frameworks

Infrastructure frameworks can be grouped as follows:

- Networking: These are the resources that transfer data from one resource to another (e.g., physical, virtual, software defined)
- Computing: These are the physical processors and memory that execute and hold the software of the other Big Data system components (e.g., physical resources, operating system, virtual implementation, logical distribution)
- Storage: These are resources which provide persistence of the data in a Big Data system (e.g., in-memory, local disk, hardware/software [HW/SW] redundant array of independent disks [RAID], Storage Area Networks [SAN], network-attached storage [NAS])
- Environmental: These are the physical plant resources (e.g., power, cooling) that must be accounted for when establishing an instance of a Big Data system

The biggest change under the Big Data paradigm is the cooperation of horizontally scaled, independent resources to achieve the desired performance.

B. Data Platform Frameworks

This is the most recognized area for changes in Big Data engineering, and given rapid changes, the hierarchy in this area will likely change in the future to better represent the patterns within the techniques. The data platform frameworks activity was expanded into the following logical data organization and distribution approaches to provide additional clarity needed for the new approaches of Big Data.

- Physical storage (e.g., distributed and non-distributed file systems and object stores)
- File systems (e.g., centralized, distributed)

- Logical storage
    - Simple tuple (e.g., relational, non-relational or not only SQL [NoSQL] tables both row and column)
    - Complex tuple (e.g., indexed document store, non-indexed key-value or queues)
    - Graph (e.g., property, hyper-graph, triple stores)

The logical storage paradigm has expanded beyond the "flat file" and relational model paradigms to develop new non-relational models. This has implications for the concurrency of the data across nodes within the non-relational model. Transaction support in this context refers to the completion of an entire data update sequence and the maintenance of eventual consistency across data nodes. This is an area that needs more exploration and categorization.

C. Processing Frameworks

Processing frameworks provide the software support for applications which can deal with the volume, velocity, variety, and variability of data. Some aspects related to processing frameworks are the following:

- Data type processing services (e.g., numeric, textual, spatial, images, video)
- Schema information or metadata (e.g., on demand, pre-knowledge)
- Query frameworks (e.g., relational, arrays)
- Temporal frameworks
    - Batch (e.g., dense linear algebra, sparse linear algebra, spectral, N-body, structured grids, unstructured grids, Map/Reduce, Bulk Synchronous Parallel [BSP])
    - Interactive
    - Real-time/streaming (e.g., event ordering, state management, partitioning)
- Application frameworks (e.g., automation, test, hosting, workflow)
- Messaging/communications frameworks
- Resource management frameworks (e.g., cloud/virtualization, intra-framework, inter-framework)

Both the Big Data Application Provider activities and the Big Data Framework Provider activities have changed significantly due to Big Data engineering. Currently, the interchange between these two roles operates over a set of independent, yet coupled, resources. It is in this interchange that the new methods for data distribution over a cluster have developed. Just as simulations went through a process of parallelization (or horizontal scaling) to harness massive numbers of independent process to coordinate them to a single analysis, Big Data services now perform the orchestration of data processes over parallel resources.

2.6 DATA CONSUMER

The Data Consumer receives the value output of the Big Data system. In many respects, the Data Consumer receives the same functionality that the Data Provider brings to the Big Data Application Provider. After the system adds value to the original data sources, the Big Data Application Provider then offers that same functionality to the Data Consumer. There is less change in this role due to Big Data, except, of course, in the desire for Consumers to extract extensive datasets from the Big Data Application Provider. Figure 7 lists the actors and activities associated with the Data Consumer.

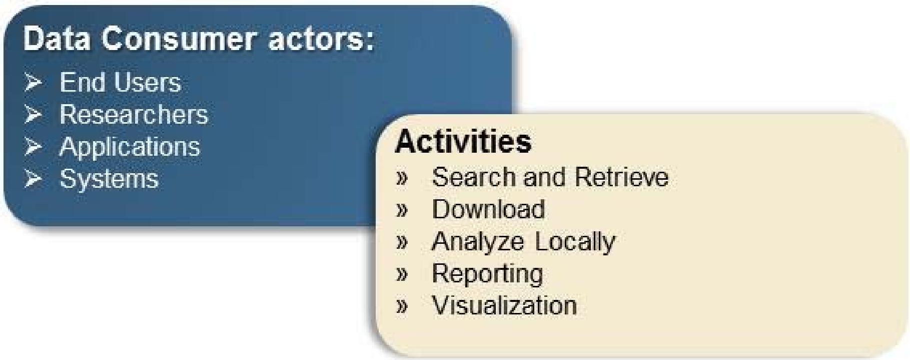


Figure 7: Data Consumer Actors and Activities

The activities listed in Figure 7 are explicit to the Data Consumer role within a data system. If the Data Consumer is in fact a follow-on application, then the Data Consumer would look to the Big Data Application Provider for the activities of any other Data Provider. The follow-on application's System Orchestrator would negotiate with this application's System Orchestrator for the types of data wanted, access rights, and other requirements. The Big Data Application Provider would thus serve as the Data Provider, from the perspective of the follow-on application.

A. Search and Retrieve

The Big Data Application Provider could allow the Data Consumer to search across the data, and query and retrieve data for its own usage.

B. Download

All the data from the Data Provider could be exported to the Data Consumer for download.

C. Analyze Locally

The Data Provider could allow the Data Consumer to run their own application on the data.

D. Reporting

The data can be presented according to the chosen filters, values, and formatting.

E. Visualization

The Data Consumer could be allowed to browse the raw data, or the data output from the analytics.

2.7 MANAGEMENT FABRIC

The Big Data characteristics of volume, velocity, variety, and variability demand a versatile management platform for storing, processing, and managing complex data. Management of Big Data systems should handle both system and data related aspects of the Big Data environment. The Management Fabric of the NBDRA encompasses two general groups of activities: system management and Big Data lifecycle management. System management includes activities such as provisioning, configuration, package management, software management, backup management, capability management, resources management, and performance management. Big Data lifecycle management involves activities surrounding the data lifecycle of collection, preparation/curation, analytics, visualization, and access. More discussion about the Management Fabric is needed, particularly with respect to new issues in the management of Big Data and Big Data engineering. This section will be developed in Version 2of this document.

Figure 8 lists an initial set of activities associated with the Management role of the NBDRA.

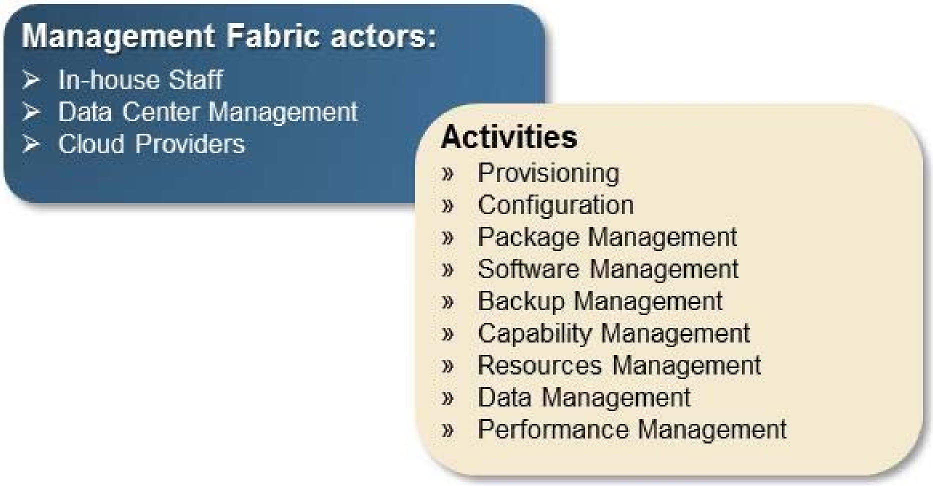


Figure 8: Big Data Management Actors and Activities

2.8 SECURITY AND PRIVACY FABRIC

Security and privacy issues affect all other components of the NBDRA, as depicted by the encompassing Security and Privacy box in Figure 1. A Security and Privacy Fabric could interact with the System Orchestrator for policy, requirements, and auditing and also with both the Big Data Application Provider and the Big Data Framework Provider for development, deployment, and operation. These ubiquitous security and privacy activities are described in the *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy* document. Figure 9 lists representative actors and activities associated with the Security and Privacy Fabric of the NBDRA. Security and privacy actors and activities will be further developed in Version 2 of *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy* document and summarized in this volume.

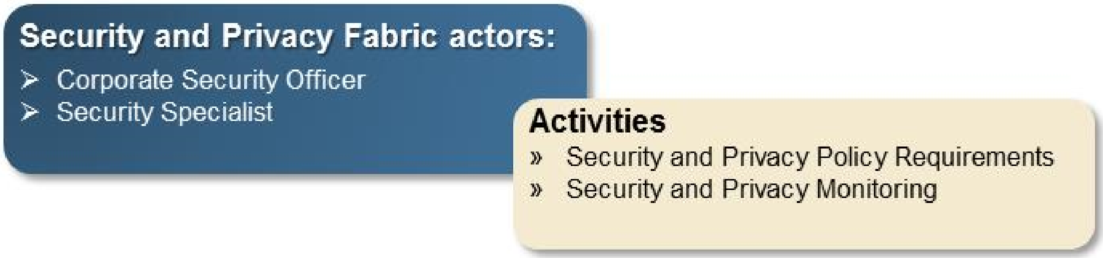
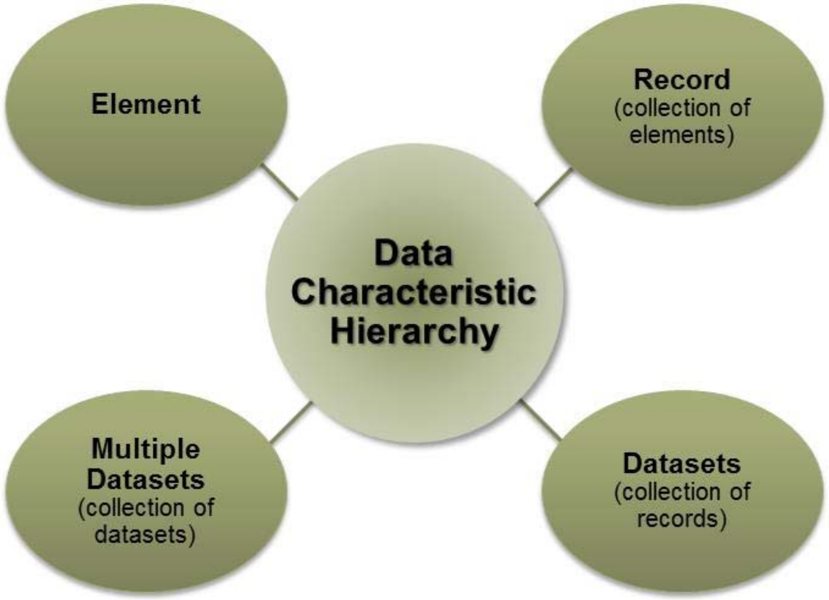

Security and Privacy Fabric actors:
Corporate Security Officer
Security Specialist
Activities
Security and Privacy Policy Requirements
Security and Privacy Monitoring

Figure 9: Big Data Security and Privacy Actors and Activities

3 DATA CHARACTERISTIC HIERARCHY

Equally important to understanding the new Big Data engineering that has emerged in the last ten years, is the need to understand what data characteristics have driven the need for the new technologies. In Section 2 of this document, a taxonomy was presented for the NBDRA, which is described in *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*. The NBDRA taxonomy has a hierarchy of roles/actors, and activities. To understand the characteristics of data and how they have changed with the new Big Data Paradigm, it is illustrative to look at the data characteristics at different levels of granularity. Understanding what characteristics have changed with Big Data can best be done by examining the data scales of data elements, of related data elements grouped into a record that represents a specific entity or event, of records collected into a dataset, and of multiple datasets—all in turn, as shown in Figure 10. Therefore, this section does not present a strict taxonomy, breaking down each element into parts, but provides a description of data objects at a specific granularity, attributes for those objects, and characteristics and subcharacteristics of the attributes. The framework described will help illuminate areas where the driving characteristics for Big Data can be understood in the context of the characteristics of all data.

Figure 10: Data Characteristic Hierarchy


Element
Record (collection of elements)
Data Characteristic Hierarchy
Multiple Datasets (collection of datasets)
Datasets (collection of records)

3.1 DATA ELEMENTS

Individual data elements have naturally not changed in the new Big Data paradigm. Data elements are understood by their data type and additional contextual data, or metadata, which provides history or additional understanding about the data.

A. Data Format

Data formats are well characterized through International Organization for Standardization (ISO) standards such as ISO 8601: 2004 Data elements and interchange formats—Information interchange—Representation of dates and times.[3] The data formats have not changed for Big Data.

B. Data Values and Vocabulary

The data element is characterized by its actual value. This value is restricted to its defined data type (e.g., numeric, string, date) and chosen data format. Sometimes the value is restricted to a specific standard vocabulary for interoperability with others in the field, or to a set of allowed values.

C. Metadata and Semantics

Metadata is sometimes simplistically described as “data about data.” Metadata can refer to a number of categories of contextual information, including the origins and history of the data, the processing times, the software versions, and other information. In addition, data can be described semantically to better understand what the value represents, and to make the data machine-operable. Both metadata and semantic data are not specific to Big Data. [a]

D. Quality And Veracity

Veracity is one of the Big Data characteristics used in describing Big Data, but the accuracy of the data is not a new concern. Data quality is another name for the consideration of the reliability of the data. Again this topic predated Big Data and is beyond the scope of this volume.[b]

3.2 RECORDS

Data elements are grouped into records that describe a specific entity or event or transaction. At the level of records, new emphasis for Big Data begins to be seen.

A. Record Format

Records have structure and formats. Record structures are commonly grouped as structured, semi-structured, and unstructured. Structured data was traditionally described through formats such as comma separated values, or as a row in a relational database. Unstructured refers to free text, such as in a document or a video stream. An example of semi-structured is a record wrapped with a markup language such as XML or HTML, where the contents within the markup can be free text.

These categories again predate Big Data, but two notable changes have occurred with Big Data. First, structured and unstructured data can be stored in one of the new non-relational formats, such as a key-value record structure, a key-document record, or a graph. Second, a greater emphasis is placed on unstructured data due to increasing amounts on the Web (e.g., online images and video.)

B. Complexity

Complexity refers to the interrelationship between data elements in a record, or between records (e.g., in the interrelationships in genomic data between genes and proteins.) Complexity is not new to Big Data.

C. Volume

Records themselves have an aspect of volume in the emerging data sources, such as considering an entire DNA on an organism as a record.

D. Metadata and Semantics

The same metadata categories described for data elements can be applied to records. In addition, relationships between data elements can be described semantically in terms of an ontology.

3.3 DATASETS

Records can be grouped to form datasets. This grouping of records can reveal changes due to Big Data.

[a] Further information about metadata and semantics can be found in: ISO/IEC 11179 Information Technology–Metadata registries; W3C’s work on the Semantic Web.

[b] Further information about data quality can be found in ISO 8000 Data Quality.

Quality and Consistency

A new aspect of data quality for records focuses on the characteristic of consistency. As records are distributed horizontally across a collection of data nodes, consistency becomes an issue. In relational databases, consistency was maintained by assuring that all operations in a transaction were completed successfully, otherwise the operations were rolled back. This assured that the database maintained its internal consistency.[c] For Big Data, with multiple nodes and backup nodes, new data is sent in turn to the appropriate nodes. However, constraints may or may not exist to confirm that all nodes have been updated when the query is sent. The time delay in replicating data across nodes can cause an inconsistency. The methods used to update nodes are one of the main areas in which specific implementations of non-relational data storage methods differ. A description of these patterns is a future focus area for this NBD-PWG.

3.4 MULTIPLE DATASETS

The primary focus on multiple datasets concerns the need to integrate or fuse multiple datasets. The focus here is on the variety characteristic of Big Data. Extensive datasets cannot always be converted into one structure (e.g., all weather data being reported on the same spatio-temporal grid). Since large volume datasets cannot be easily copied into a normalized structure, new techniques are being developed to integrate data as needed.

Personally Identifiable Information

An area of increasing concern with Big Data is the identification of individuals from the integration of multiple datasets, even when the individual datasets would not allow the identification. For additional discussion, the reader is referred to *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy*.

[c] For additional information on this concept the reader is referred to the literature on ACID properties of databases.

4 SUMMARY

Big Data and data science represent a rapidly changing field due to the recent emergence of new technologies and rapid advancements in methods and perspectives. This document presents a taxonomy for the NBDRA, which is presented in *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*. This taxonomy is a first attempt at providing a hierarchy for categorizing the new components and activities of Big Data systems. This initial version does not incorporate a breakdown of either the Management or the Security and Privacy roles within the NBDRA as those areas need further discussion within the NBD-PWG. In addition, a description of data at different scales was provided to place concepts being ascribed to Big Data into their context. The NBD-PWG will further develop the data characteristics and attributes in the future, in particular determining whether additional characteristics related to data at rest or in-motion should be described. The Big Data patterns related to transactional constraints such as ACID (Atomicity, Consistency, Isolation, Durability—a set of properties guaranteeing reliable processing of database transactions) have not been described here, and are left to future work as the interfaces between resources is an important area for discussion. This document constitutes a first presentation of these descriptions, and future enhancements should provide additional understanding of what is new in Big Data and in specific technology implementations.

Appendix A: Acronyms

ACID	Atomicity, Consistency, Isolation, Durability
APIs	application programming interfaces
BSP	Bulk Synchronous Parallel
DaaS	Data as a Service
FTP	File Transfer Protocol
HW/SW RAID	hardware/software redundant array of independent disks
IoT	Internet of Things
ISO	International Organization for Standardization
ITL	Information Technology Laboratory
NARA	National Archives and Records Administration
NAS	network-attached storage
NASA	National Aeronautics and Space Administration
NBD-PWG	NIST Big Data Public Working Group
NBDRA	NIST Big Data Reference Architecture
NIST	National Institute of Standards and Technology
NoSQL	not only SQL
NSF	National Science Foundation
PII	Personally Identifiable Information
RFID	radio frequency identification
SAN	Storage Area Networks
SLAs	service-level agreements

Appendix B: References

DOCUMENT REFERENCES

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

[2] General Services Administration, “The home of the U.S. Government’s open data,” *Data.gov*, http://www.data.gov/

[3] ISO 8601: 2004, “Data elements and interchange formats -- Information interchange -- Representation of dates and times,” *International Organization of Standardization*, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=40874