**NIST Special Publication XXX-XXX**

**DRAFT NIST Big Data Interoperability Framework:**

**Volume 2, Big Data Taxonomies**

NIST Big Data Public Working Group

Definitions and Taxonomies Subgroup

Draft Release 1

November 11, 2014

http://dx.doi.org/10.6028/NIST.SP.XXX



NIST Special Publication xxx-xxx

Information Technology Laboratory

**DRAFT NIST Big Data Interoperability Framework:**

**Volume 2, Big Data Taxonomies**

**Release 1**

NIST Big Data Public Working Group (NBD-PWG)

Definitions and Taxonomies Subgroup

National Institute of Standards and Technology

Gaithersburg, MD 20899

November 2014



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*

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 8900) Gaithersburg, MD 20899-8930

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.

**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: 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 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:

|  |  |
| --- | --- |
| 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, National Institute of Standards and Technology  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.

Table of Contents

Executive Summary viii

1 Introduction 1

1.1 Background 1

1.2 Scope and Objectives of the Definitions and Taxonomies Subgroup 2

1.3 Report Production 2

1.4 Report Structure 3

1.5 Future Work of this Volume 3

2 Reference Architecture Taxonomy 3

2.1 Actors and Roles 4

2.2 Data Provider 6

2.2.1 Data Capture from Sources 6

2.2.2 Data Persistence 7

2.2.3 Data Scrubbing 7

2.2.4 Data Annotation and Metadata Creation 7

2.2.5 Access Rights Management 7

2.2.6 Access Policy Contracts 8

2.2.7 Data Distribution Application Programming Interfaces 8

2.2.8 Capabilities Hosting 8

2.2.9 Data Availability Publication 8

2.3 Big Data Application Provider 9

2.3.1 Data Collection Processes 9

2.3.2 Data Preparation Processes 9

2.3.3 Data Analytics Processes 10

2.3.4 Visualization 10

2.3.5 Access 11

2.4 Big Data Framework Provider 12

2.4.1 Infrastructures 12

2.4.2 Platforms 13

2.4.3 Processing Frameworks 14

2.5 Data Consumer 16

2.5.1 Current Application as Data Provider 16

2.5.2 Search and Retrieve 16

2.5.3 Download 16

2.5.4 Analyze Locally 16

2.5.5 Reporting 16

2.5.6 Visualization 16

2.5.7 Data to Use for Their Own Processes 16

2.6 System Orchestrator 18

2.6.1 Business Ownership Requirements and Monitoring 18

2.6.2 Governance Requirements and Monitoring 18

2.6.3 Data Science Requirements and Monitoring 19

2.6.4 System Architecture Requirements and Monitoring 19

3 Management 21

3.1 Provisioning 21

3.2 Configuration 21

3.3 Package Management 21

3.4 Software Management 21

3.5 Backup Management 21

3.6 Capability Management 21

3.7 Resources Management 22

3.8 Data Management 22

3.9 Big Data Lifecycle Management 22

3.10 Administration 22

3.11 Performance Management 22

3.12 Security and Policy Management 22

3.13 Security and Privacy Fabric 22

4 Security and Privacy Policy Requirements 22

4.1.1 Security and Privacy Monitoring 23

4.1.2 Security Protection Requirements and Monitoring 23

4.1.3 Data Provenance Monitoring 23

4.1.4 Data Privacy Requirements and Compliance Monitoring 23

5 Security and Privacy Taxonomy (New Section) 24

5.1 Conceptual Taxonomy of Security and Privacy Topics 24

5.1.1 Privacy 24

5.1.2 Provenance 24

5.1.3 System Health 24

5.2 Operational Taxonomy of Security and Privacy Topics 24

5.2.1 Device and Application Registration 25

5.2.2 Identity and Access Management 25

5.2.3 Data Governance 25

5.2.4 Infrastructure Management 25

5.2.5 Risk and Accountability 25

5.3 Roles Taxonomy of Security and Privacy Topics 25

5.3.1 Infrastructure Technology 25

5.3.2 Governance, Risk Management, and Compliance 25

5.3.3 Information Worker 25

6 Data Characteristic Taxonomy 26

6.1 Data Elements 26

6.2 Data Records 26

6.3 Datasets 26

6.4 Multiple Datasets 26

7 Summary 26

Appendix A: Index of Terms 1

Appendix B: Acronyms 1

Appendix C: References 1

Figures

[Figure 1: NIST Big Data Reference Architecture 5](#_Toc389169650)

[Figure 2: Roles and a Sampling of Actors 5](#_Toc389169651)

[Figure 3: Data Provider Actors and Activities 6](#_Toc389169652)

[Figure 4: Big Data Application Provider Actors and Activities 9](#_Toc389169653)

[Figure 5: Big Data Framework Provider Actors and Activities 12](#_Toc389169654)

[Figure 6: Data Consumer Actors and Activities 16](#_Toc389169655)

[Figure 7: System Orchestrator Actors and Activities 18](#_Toc389169656)

[Figure 8: Big Data Security and Privacy Actors and Activities 21](#_Toc389169657)

[Figure 9: Management Actors and Activities 23](#_Toc389169658)

Executive Summary

This ***NIST Big Data Interoperability Framework Volume 2: Taxonomies*** was prepared by the NBD-PWG’s Definitions and Taxonomy Subgroup to facilitate better communication and understanding across the participants in this field by expanding on the functional components of the reference architecture. The top-level roles of the taxonomy are Data Provider, System Orchestrator, Data Consumer, Big Data Application Provider, Big Data Framework Provider, Security and Privacy, and Management. The activities within each of these roles are specified to see where the functional components reside. This taxonomy is not meant to be an exhaustive list; instead, it describes what is new in Big Data systems. In some cases this requires also listing current practices and technologies in the same category.

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. In addition to this volume, the other volumes are as follows:

* Volume 1, Definitions
* Volume 3, Use Cases and General Requirements
* Volume 4, Security and Privacy Requirements
* Volume 5, Architectures White Paper Survey
* Volume 6, Reference Architectures
* Volume 7, Technology Roadmap

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.

# Introduction

## 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.[[1]](#endnote-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 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.

The *Draft* *NIST Big Data Interoperability Framework* containsthe following seven volumes:

* Volume 1, Definitions
* Volume 2, Taxonomies (this volume)
* Volume 3, Use Case and General Requirements
* Volume 4, Security and Privacy Requirements
* Volume 5, Architectures White Paper Survey
* Volume 6, Reference Architectures
* Volume 7, Technology Roadmap

## 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.

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

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

For **marketers**, this document 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.

## Report Production

This document derives from the discussions in the Definitions and Taxonomy Subgroup of the NBD-PWG. The Definitions and Taxonomy Subgroup produced two volumes in the *NIST Big Data Interoperability Framework*. *Volume 1, Definitions* provides terms and definitions for the important concepts related to Big Data. *Volume 2, Taxonomy* (this volume) provides the taxonomy of technology and processes related to Big Data and the components of the NIST Big Data Reference Architecture. This taxonomy was developed using a mindmap 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 Definitions and Taxonomy Subgroup attempted to describe only what has changed in the shift to the new Big Data paradigm, and only the components needed to clarify the this shift. There is, for example, no attempt to create a taxonomy of analytics techniques, as these pre-date Big Data. This taxonomy should remain a work in progress to mature as new technologies are developed and the patterns within data and system architectures are better understood.

## Report Structure

This document provides the several hierarchical presentations related to Big Data. The first is the NIST Big Data Reference Architecture (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, then the activities the actor or role performs. The components and subcomponents within a given activity are listed underneath that activity in a hierarchical fashion. In the NGDRA, as shown in Volume 6: Reference Architecture, there are two roles that span the activities within the other roles. These roles are Management and Security and Privacy. These two topic areas are placed in their own sections as the working group feels they need more attention. These two roles will be addressed 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 explain what has changed with the introduction of Big Data Engineering technologies, as described in the NBDRA Volume 1: Definitions.

Within the following sections, illustrative examples are given to understand the role/actor and activity. There is no expectation of completeness in the components, the intent is to provide enough context for understanding the specific areas that have changed because of the new Big Data paradigm.

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, Technology Roadmap*. Finally, to understand how these systems are architected to meet users’ needs, the reader is referred to the *NIST Big Data Interoperability Framework: Volume 3, Use Cases and Requirements* document.

## Future Work of this Volume

As mentioned in the previous section, the working group is continuing to explore the changes in both Management and in Security and Privacy. As the changes in the activities within these roles are better clarified, the taxonomy will be developed in these areas. In addition, a fuller understanding of Big Data and its technologies should focus on the interactions between the characteristics of the data and the desired analytics in both technique and time window for performance. These characteristics drive the application, and the choice of tools to meet the system requirements. An investigation of the interfaces between the data characteristics and the technologies is a continuing task within the definitions and taxonomy subgroup, and the reference architecture subgroup. Finally, societal impact issues have not been fully explored. There are a number of over-arching issues in the implications of Big Data, such as data ownership and data governance that need more examination.

# Reference Architecture Taxonomy

The first taxonomy is for the Reference Architecture, and is intended to describe the hierarchy of actors and roles, the activities performed by those roles, and the components and sub-components within that activity. There are a number of models for describing the technologies needed for an application, including for example a layer model of network, hardware, operating system, application, etc. For elucidating the taxonomy we have chosen a hierarchy that allows us to place the new technologies within a context of what came before. This is certainly not the definitive taxonomy, and we expect this to mature as new technologies emerge and help us understand how better to categorize the different methods for building data systems.

## Actors and Roles

In system development, actors and roles have the same relationship as in the movies but system development actors can represent individuals, organizations, software, or hardware. Each element 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 single resources of only one organization. Currently, roles may be distributed, analogous to the diversity of actors within a given solution spurred by the spread of cloud solutions. Actors in Big Data systems can also come from multiple organizations.

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 the end-to-end system requirements.

The taxonomy development began by reviewing the NBD-PWG analyses of the use cases and reference architecture survey. Details of the use cases are provided in *NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements*. Details of the reference architecture survey are provided in *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture Survey*. 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, the NIST Big Data Reference Architecture.



Figure 1: NIST Big Data Reference Architecture

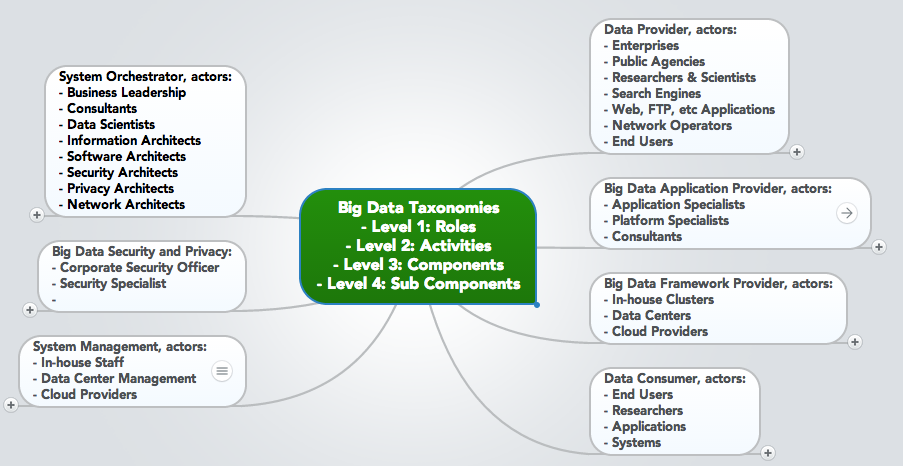
These seven items form the foundation of the reference architecture taxonomy. The five central roles are as follows:

* **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

The two fabrics discussed separately in Sections 3 and 4 are:

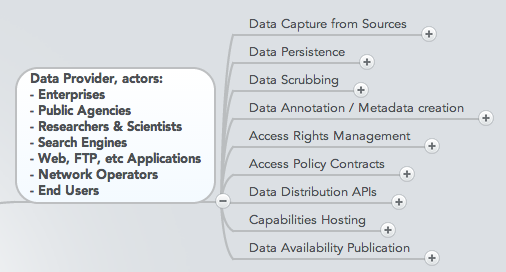
* **Security and Privacy Fabric**
* **Management Fabric**

Figure 2 outlines potential actors for the seven roles listed above. In the following sub-sections, the first five roles will each be explored, in turn, with their activities.

Figure 2: Roles and a Sampling of Actors

## Data Provider

A ***Data Provider*** makes data available to themselves or to others. The actor fulfilling this role can be part of the Big Data system, internal to the organization in another system, 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 can include those shown in Figure 3.

Figure 3: 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 for Federal agencies that are stewards of public data to also serve the role of Data Provider.

Nine possible Data Provider activities are outlined in Figure 3 and discussed further in the following subsections.

### Data Capture from Sources

The Data Provider captures data from their 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. Components of the data sources activity include both online and offline sources. Online sources can include 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, this 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, personal wearable devices, such as exercise monitors and 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.’ The subset of personal information devices are often referred to as ‘wearable tech,’ with the resulting data sometimes referred to as ‘digital exhaust.’

### Data Persistence

The Data Provider stores the data into a repository from which the data can be extracted and made available to others. The stored data would be subject to a data retention policy. The data can be persisted (i.e., stored) 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. Data-as-a-Service (DaaS) is a term used for this kind of data persistence that is accessed through specific interfaces.

### 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 for implicit PII

The latter obscures the PII to remove the possibility of directly tracing the data back to an individual. In the era of Big Data, data scrubbing is an area that 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.

### Data Annotation and Metadata Creation

The Data Provider would maintain information about the data, called metadata, in their repository, and also maintain 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. While the actors that collected the data will have a clear understanding of the processes, repurposing data for other uses is open to misinterpretations when other actors subsequently use the data at a later time.

### Access Rights Management

The Data Provider will determine 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

### Access Policy Contracts

The Data Provider will define policy for others’ use of the accessed to the data, and what data will be made available. These contracts specify:

* Policies for primary and secondary rights
* Agreements

The contracts will specify acceptable use policies, as well as any specific restrictions on the use of the data, as well as the ownership of the original data and any derivative works from the data.

### 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
* Authorization

### Capabilities Hosting

In addition to offering data downloads, the Data Provider could offer 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 can 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.

### 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:

* 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]](#endnote-2).)

## Big Data Application Provider

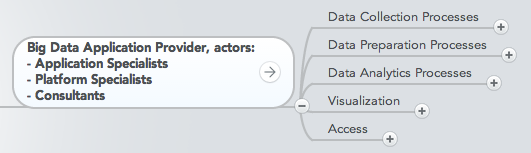
The Big Data Application Provider executes the manipulations of the data lifecycle to meet the requirements established by the System Orchestrator. This is where the general capabilities within the Big Data Framework are combined to produce the specific data system. Figure 4 lists the actors and activities associated with the Big Data Application Provider.

Figure 4: Big Data Application Provider Actors and Activities

### Data Collection Processes

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 these 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.

### Data Preparation Processes

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

* Data cleansing
* Data conversion
* Calculated field creation
* Data summarization
* Data partition implementation
* Data storage preparation
* Data virtualization layer
* Schema-on-read

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 due to the size of the datasets. When it is prohibitive store a new cleansed copy of the data, it is stored in it’s raw form and only prepared for what data is needed when it is needed.

A second area of additional emphasis because of Big Data is in the area of data summarization. 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 provide the characteristics of a number of local subsets of the data, and then provide these characteristics for visualization as the data is browsed to develop greater understanding.

### Data Analytics Processes

The term data science is used in many ways. While this term can refer to the end-to-end data lifecycle, the most common usage focuses on the discovery, or rapid hypothesis-test cycle, for finding value in big volume datasets. This rapid analytics cycle is also described as ‘agile analytics,’ which 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 greater 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 the distribution of the data across a cluster of independent nodes and its method of access. 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.

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

* Human-in-the-loop analytics lifecycle
* Discovery
* Hypothesis
* Hypothesis testing
* Alternate methods
* Metadata matching processes
* Analysis Complexity Considerations
  + Computational complexity
  + Machine Learning technique
  + Data extent complexity
  + Data Location complexity
* Analytics Latency Considerations
  + Real-time or streaming
  + Near real-time or interactive
  + Batch or off-line

While these considerations are not new to Big Data, their implementations can be tightly coupled with the specifics of the data storage and the preparation step.

### Visualization

While visualization or the human-in-the-loop is often placed into the category “analytics”, it is useful to identify categories of visualizations.

* Exploratory Data Visualization for data understanding
* Data browsing
* Outlier detection
* Boundary conditions
* Explicatory visualization for analytical results
  + Confirmation of validity or results
  + Near real-time presentation for rapid analytics cycles
  + Understanding analytics-derived features
* Explanatory visualization to “tell the story”
* Reports
* Business intelligence
* Customer summarization presentation

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 re-purposed data, and because of 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.

### Access

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

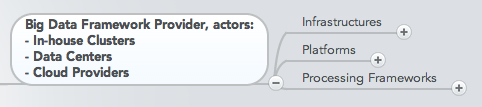
* Data export API processes
* API
* Protocol
* Query language
* Data charging mechanisms
* Consumer analytics hosting
* Analytics-as-a-service hosting

The access activity of the Big Data Application Provider should, in fact, mirror everything the Data Provider does, 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 and be optimized for the horizontally distributed resources.

## 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 components here from which the Big Data Application Provider can choose in utilizing these resources and the network to build the specific system. Figure 5 lists the actors and activities associated with the Big Data Framework Provider.

Figure 5: Big Data Framework Provider Actors and Activities

This is the role that has seen the most significant changes because of Big Data. The Big Data Framework Provider consists of one or more instances of the three subcomponents: Processing Frameworks, Data Platforms, and Infrastructure 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 in order to provide flexibility or 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 NIST working group continues its discussion into patterns within these techniques, different orderings will no doubt be more representative and understandable.

### Infrastructure Frameworks

Infrastructures can be grouped as

* Cluster Implementation
* Operating System
* Physical Resource Implementation (server, cluster, multi-site cluster)
* Virtual Resource Implementation
* Logical Distribution (Peer-to-Peer, Master-Slave, Multiple masters with slave nodes)
* Storage implementation
  + In memory
  + Local disks
  + HW/SW RAID
  + SAN
  + NAS
  + Distributed file systems
  + Distributed object stores
* Networking implementation
  + Physical network
  + Virtual network
  + Fault tolerance
* Security/encryption
* Virtual network
* Protocol
* Fault tolerance
* Security/encryption

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

### Data Platforms

This is the most recognized area for changes in Big Data Engineering, and given the rapid changes in this area, the hierarchy in this area will likely change in the future to better represent the patterns within the techniques. We expand the sub-components as this is an area that clarifies the new approaches of Big Data and because this is an area that needs additional clarification.

* Logical Storage
  + File Systems
  + Simple Tuple
    - Relational
    - Non-relational or NoSQL tables (row or column-oriented)
  + Complex Tuple
    - Indexed (NoSQL document Store)
    - Non-indexed (NoSQL key-value, queues)
  + Graph
    - Property graph
    - Hyper-graph
    - Triple-stores
* Physical Storage
  + Distributed filesystems
  + Distributed Object 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.

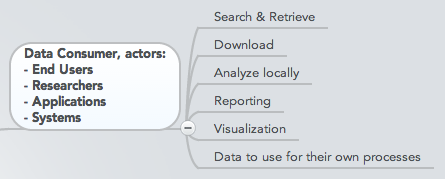
### Processing Frameworks

* Data Type Processing Services
* Textual
* Documents
* Spatial
* Images
* Aerial
* Faces
* Fingerprints
* Astronomical
* Video
* Schema Information (metadata)
* Queryable
* Schema on demand
* Pre-knowledge
* Query Frameworks
  + Relational
  + Arrays
  + Distributed
    - Batch
    - Interactive
    - Real-time
    - In-memory
    - Distributed transactional SQL
* Application Framework
* Automation services
* Test services
* Analytics hosting services
* Data quality service
* Workflow services
* Batch Processing Frameworks
* MapReduce
* Linear Algebra
* Sparse
* Dense
* Spectral methods
* N-Body methods
* Grids
* Structured
* Unstructured
* Combinational Logic
* Graph Traversal
* Backtrack and Branch-and-Bound
* Dynamic Programming
* Graphical Models
* Finite State Machines
* Stream Processing Frameworks
* Complex Event Processing

Both the Big Data Application Provider activities and the Big Data Framework Provider activities have changed significantly due to Big Data Engineering. Now 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, now Big Data services perform the orchestration of data processes over horizontally-scaled resources.

## Data Consumer

The Data Consumerreceives the value output of the Big Data system. In many respects, they are the recipients of 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. Figure 6 lists the actors and activities associated with the Data Consumer.

Figure 6: Data Consumer Actors and Activities

The activities listed in Figure 6 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 (described below) would negotiate with this application’s System Orchestrator for the types of data wanted, access rights, etc. The Big Data Application Provider would thus need to serve as the Data Provider, from the perspective of the follow-on application.

### Search and Retrieve

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

### Download

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

### Analyze Locally

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

### Reporting

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

### Visualization

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

## 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 the monitoring or auditing activities to ensure the compliance of the system with respect to those requirements.

This role provides the system requirements, high-level design, and monitoring for the data system. While the roles pre-date Big Data systems, there are design activities that have changed within the Big Data paradigm.

Figure 7 lists the actors and activities associated with the System Orchestrator.

Figure 7: System Orchestrator Actors and Activities

### 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
* Capabilities provider negotiation
* 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.

### 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.

### Data Science Requirements and Monitoring

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

* Data source selection
* Data description
* Data location
* File types
* File attributes
* Data provenance evaluation
* Data collection requirements and monitoring
* Data preparation requirements and monitoring
* Data analysis requirements and monitoring
* Analytical model choice
* Data visualization requirements and monitoring
* Application type specification
* Streaming
* Aggregation
* Integration
* Transfer
* Search
* Statistics
* Real-time analytics
* Batch analytics
* Interactive annotation
* Others

A number of the design activities have changed in the new paradigm. In particular, a greater choice of data models now exists beyond just the relational model. Choosing a non-relational model will depend on the data type. Choosing keys will depend on the organization’s data analysis needs. The best choice for the data element to use when dividing the data evenly among nodes depends on the data quantities within the chosen ranges for the data element.

### System Architecture Requirements and Monitoring

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

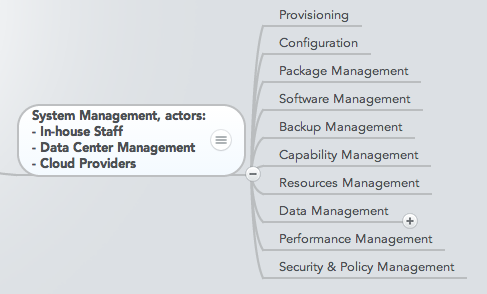
* Data process requirements
* Software component determination
* Hardware component determination
* Logical data modeling and partitioning
* Data export requirements
* Scaling requirements

The system architecture has changed in the Big Data paradigm due to the potential interplay between the different actors. The coordination between functional components is more complex with additional communications and inter-connectivity requirements among the independently operated components. Maintaining the needed performance can lead to a very different architecture from what would have been used prior to the new distribution of data across system nodes.

# Management Fabric

In the NBDRA, Management is the overarching control of the execution of the system, its deployment of the system, and its operational maintenance. This is an area that needs more discussion in the relationships to what is new in the management of Big Data and Big Data Engineering. This section will be developed in the next version of the Taxonomy document.

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

Figure 8: Management Actors and Activities

# Security and Privacy Fabric

Security and privacy issues affect all other components of the NBDRA, as depicted by the encompassing Security and Privacy grey box in Figure 1. The Security and Privacy fabric interacts 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 Requirements* document. Figure 9 lists the actors and activities associated with the Security and Privacy component of the NBDRA.

Figure 9: Big Data Security and Privacy Actors and Activities

# 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, a taxonomy was presented for the reference architecture, which is described in Volume 6: Reference Architecture. This taxonomy presented a hierarchy of roles, activities, components, and subcomponents. To understand the characteristics of data and how they have changed with the new Big Data Paradigm, it is more illustrative to look at the data characteristics at different levels of granularity. We can best understand what characteristics have changed with Big Data by examining in turn the data elements, related data elements grouped into a record that represents a specific entity or event, records collected into a dataset, and then multiple datasets, as shown in Figure 10. This section will therefore not present a strict taxonomy, breaking down each elements in component parts, but with present a hierarchical description of data objects at a specific granularity, attributes for those objects, then characteristics and sub-characteristics for the attributes. This framework will provide an opportunity to illuminate the areas where the driving characteristics for Big Data can be understood in the context of characteristics of all data.

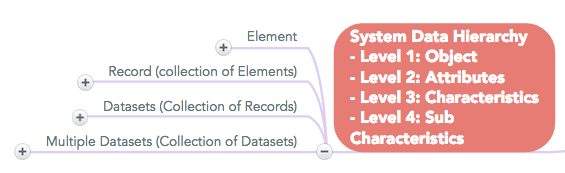


Figure 10: System Data Hierarchy

## 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, that provides history or additional understanding about the data.

### Data Format

Data Formats are well characterized through ISO standards such as

ISO 8601 : 2008, and its title is "Data elements and interchange formats - Information interchange - Representation of dates and times".

### Data Values and Vocabulary

Naturally the data element is characterized by its actual value. This value is restricted to it’s defined data type (numeric, string, date, etc) and chosen format (as mentioned above). Sometimes the value is restricted to a specific standard vocabulary for interoperability with others in the field, or to set of allowed values.

### 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, processing information, etc. 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, and further information can be found in:

* ISO/IEC 11179 Information Technology – Metadata registries
* W3C’s work on the Semantic Web.

### Quality And Veracity

Veracity is one of the “V’s” often referred to when describing the characteristics of Big Data, but the accuracy of the data is not a new concern. Data Quality is the other name for the consideration of the reliability of the data. Again this topic predated Big Data and is beyond the scope of this volume. The reader is referred to:

* ISO 8000 Data Quality

## Data Records

Data elements are grouped into records that describe a specific entity or event or transaction. At the level of records we begin to see some new emphasis for Big Data.

### Record Format

Records also have structure and formats. Record structure is commonly referred to as Structured, Semi-Structured, and Unstructured. Structured data was traditionally described through formats such as common separated values, or as a row in a relational database. Unstructured refers to for example free text in a document or a video stream. Semi-structured refers to for example a record wrapped with a markup language such as XML or HTML, where the contents within the market can be free text.

These categories again predate Big Data, but there are two changes of note. The first is that structured and unstructured data can be stored one of the new non-relational formats, such as a key-value record structure, or as a key-document record, or as a graph. In additional there is a greater emphasis on unstructured data due to its increasing amounts through the Web or online images and video.

### Complexity

Complexity refers to the inter-relationship between elements in a record, or between records. This is not new, but is present in for example genomic data relating genes and proteins.

### Volume

Records themselves have an aspect of volume in the emerging data sources, for example if you consider an entire genome on an organism as a record.

### Metadata and Semantics

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

## Datasets

Collections of records form datasets, and we see changes due to Big Data.

### 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, there is an issue of consistency. In relational databases consistency was maintained by ensuring that all operations in a transaction were completed successfully, otherwise the operations were rolled back. This ensured that the database maintained its internal consistency. For additional information on this concept the reader is referred to the literature on ACID properties of databases. For Big Data with multiple nodes and backup nodes, new data is sent to the appropriate node(s), but there is not necessarily a constraint that when you query the data all the nodes have been updated. The time delay in replicating data across nodes can cause an inconsistency. The methods used to update nodes is 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 the working group.

## 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 (for example 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 on the fly.

### Implicit Personally Identifiable Information

An area of increasing concern with Big Data is the integration of multiple datasets may allow the identification of individuals, even when the individual datasets would not. This area of implicit PII is an area of growing concern. For additional discussion, the reader is referred to Volume 4: Security and Privacy.

# Summary

In this document we have presented a Taxonomy for the NIST Big Data Reference Architecture (NBDRA), as provided in Volume 6: Reference Architecture. The taxonomy given here is a first attempt at providing a hierarchy for categorizing the new components of Big Data. In addition, a hierarchical description of data was provided, to place concepts being ascribed to Big Data into their context. This initial version does not incorporate a breakdown of either Management or Security and Privacy roles within the NBDRA. These are areas that need more development. The Big Data patterns related to transactional constraints such as ACID have not been described here, and are left to future work. Finally there are other areas that may need development, such as a hierarchical description of the “states” of data, such as at-rest or in-motion that could be explored as a means of again providing additional understanding of what is new in Big Data and of specific technology implementations.

Appendix A: Index of Terms

Appendix B: Acronyms

DaaS Data-as-a-Service

DoS denial of service

PII personally identifiable information

RFID radio frequency identification

SLAs service-level agreements

Appendix C: References

General Resources

Suggestions for general resources could include the following document types that discuss:

* U.S. Government’s Open Data Initiative

Document References

1. “Big Data is a Big Deal”, The White House, Office of Science and Technology Policy. <http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal> (accessed February 21, 2014) [↑](#endnote-ref-1)
2. U.S. Government’s Open Data Initiative, <http://www.data.gov/> [↑](#endnote-ref-2)