**NIST Special Publication 1500-1**

**DRAFT NIST Big Data Interoperability Framework: Volume 1, Definitions**

NIST Big Data Public Working Group

Definitions and Taxonomies Subgroup

Version 2, Draft 2

March 20, 2017

[http://dx.doi.org/10.6028/-------1](http://dx.doi.org/10.6028/NIST.SP.1500-1)



NIST Special Publication 1500-1

Information Technology Laboratory

**DRAFT NIST Big Data Interoperability Framework:**

**Volume 1, Definitions**

**Version 2, Draft 2**

NIST Big Data Public Working Group (NBD-PWG)

Definitions and Taxonomies Subgroup

National Institute of Standards and Technology

Gaithersburg, MD 20899

[http://dx.doi.org/10.6028/------1](http://dx.doi.org/10.6028/NIST.SP.1500-1)

March 2017



U. S. Department of Commerce

*Wilbur L. Ross, Jr., Secretary*

National Institute of Standards and Technology

*Dr. Kent Rochford, Acting Under Secretary of Commerce for Standards and Technology*

*and Acting NIST Director*

**National Institute of Standards and Technology (NIST) Special Publication 1500-1**

37 pages (March 20, 2017)

NIST Special Publication series 1500 is intended to capture external perspectives related to NIST standards, measurement, and testing-related efforts. These external perspectives can come from industry, academia, government, and others. These reports are intended to document external perspectives and do not represent official NIST positions.

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 publications are available at <http://www.nist.gov/publication-portal.cfm>.

**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](mailto:SP1500comments@nist.gov)

Request for Contributions

The NIST Big Data Public Working Group (NBD-PWG) requests contributions to this draft version 2 of the *NIST Big Data Interoperability Framework Volume 1, Definitions.* All contributions are welcome, especially comments or additional content for the current draft.

The NBD-PWG is actively working to complete version 2 of the set of NBDIFdocuments. The goals of version 2 are to enhance the version 1 content, define general interfaces between the NIST Big Data Reference Architecture (NBDRA) components by aggregating low-level interactions into high-level general interfaces, and demonstrate how the NBDRA can be used.

To contribute to this document, please follow the steps below as soon as possible but no later than May 1, 2017.

1. Register as a user of the NIST Big Data Portal (https://bigdatawg.nist.gov/newuser.php)
2. Record comments and/or additional content in one of the 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
3. Submit the edited file from either method above to SP1500comments@nist.gov with the volume number in the subject line (e.g., Edits for Volume 1).
4. Attend the weekly virtual meetings on Tuesdays for possible presentation and discussion of your submission. Virtual meeting logistics can be found at https://bigdatawg.nist.gov/program.php

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.

The comments and additional content will be reviewed by the subgroup co-chair responsible for the volume in question. Comments and additional content may be presented and discussed by the NBD-PWG during the weekly virtual meetings on Tuesday.

Three versions are planned for the NBDIF set of documents, 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 each NBDIF document.

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

Big Data professionals are always welcome to join the NBD-PWG to help craft the work contained in the volumes of the NBDIF. Additional information about the NBD-PWG can be found at <http://bigdatawg.nist.gov>. Information about the weekly virtual meetings on Tuesday can be found at https://bigdatawg.nist.gov/program.php.

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 large amount of data in the networked, digitized, sensor-laden, information-driven world. While opportunities exist with Big Data, the data can overwhelm traditional technical approaches and the growth of data is outpacing scientific and technological advances in data analytics. To advance progress in Big Data, the NIST Big Data Public Working Group (NBD-PWG) is working to develop consensus on important, fundamental concepts related to Big Data. The results are reported in the *NIST Big Data Interoperability Framework* series of volumes. This volume, Volume 1, contains a definition of Big Data and related terms necessary to lay the groundwork for discussions surrounding Big Data.

Keywords

Big Data; Big Data Definitions; Big Data Application Provider; Big Data Characteristics; Big Data Framework Provider; Big Data taxonomy; Data Consumer; Data Provider; Data Science; Management Fabric; Reference Architecture; Security and Privacy Fabric; System Orchestrator; use cases.

Acknowledgements

This document reflects the contributions and discussions by the membership of the NBD-PWG, co-chaired by Wo Chang (NIST ITL), Nancy Grady (SAIC), Geoffrey Fox (University of Indiana), Arnab Roy (Microsoft), Mark Underwood (Krypton Brothers), David Boyd (InCadence Corp), Russell Reinsch (Loci), and Gregor von Laszewski (University of Indiana).

The document contains input from members of the NBD-PWG Definitions and Taxonomies Subgroup, led by Nancy Grady (SAIC).

NIST SP1500-1, Version 2 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[[1]](#footnote-2) to this volume by the following NBD-PWG members.

A list of contributors to version 2 of this volume will be added here.

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

Table of Contents

[Executive Summary ix](#_Toc478543465)

[1 Introduction 1](#_Toc478543466)

[1.1 Background 1](#_Toc478543467)

[1.2 Scope and Objectives of the Definitions and Taxonomies Subgroup 2](#_Toc478543468)

[1.3 Report Production 3](#_Toc478543469)

[1.4 Report Structure 3](#_Toc478543470)

[1.5 Future Work on this Volume 3](#_Toc478543471)

[2 Big Data 4](#_Toc478543472)

[2.1 Big Data Definitions 4](#_Toc478543473)

[2.2 When do you have a big data problem? 5](#_Toc478543474)

[2.2.1 Volume 6](#_Toc478543475)

[2.2.2 Velocity 6](#_Toc478543476)

[2.2.3 Variety 6](#_Toc478543477)

[2.2.4 Variability 6](#_Toc478543478)

[2.3 Other Usage of the Term Big Data 7](#_Toc478543479)

[3 Big Data Engineering (Frameworks) 9](#_Toc478543480)

[3.1 Horizontal Infrastructure Scaling 9](#_Toc478543481)

[3.2 Scalable Logical Data Storage 11](#_Toc478543482)

[3.3 Relationship to other Technological Innovations 12](#_Toc478543483)

[3.3.1 High Performance Computing 12](#_Toc478543484)

[3.3.2 Cloud 12](#_Toc478543485)

[3.3.3 Internet of Things 12](#_Toc478543486)

[3.3.4 Embedded Computing – Cyber-Physical Systems 13](#_Toc478543487)

[3.3.5 Blockchain 13](#_Toc478543488)

[4 Data Science Applications 14](#_Toc478543489)

[4.1 When is it Data Science? 14](#_Toc478543490)

[4.2 Data Science Characteristics 15](#_Toc478543491)

[4.2.1 Veracity 15](#_Toc478543492)

[4.2.2 Validity 16](#_Toc478543493)

[4.2.3 Volatility 16](#_Toc478543494)

[4.2.4 Visualization 16](#_Toc478543495)

[4.2.5 Value 16](#_Toc478543496)

[4.2.6 Metadata 16](#_Toc478543497)

[4.2.7 Data Types 17](#_Toc478543498)

[4.2.8 Complexity 17](#_Toc478543499)

[4.2.9 Latency 17](#_Toc478543500)

[4.3 Big Data Science 17](#_Toc478543501)

[4.3.1 Data Cleansing 18](#_Toc478543502)

[4.3.2 Sampling 18](#_Toc478543503)

[4.3.3 Machine Learning 18](#_Toc478543504)

[4.4 Emergent Behavior 19](#_Toc478543505)

[4.5 Data Scientists 19](#_Toc478543506)

[4.6 Big Data Metrics and Benchmarks 20](#_Toc478543507)

[5 Big Data Security and Privacy 21](#_Toc478543508)

[6 Big Data Management 22](#_Toc478543509)

[6.1 Orchestration 22](#_Toc478543510)

[6.2 Data Governance 22](#_Toc478543511)

[6.3 Data Ownership 22](#_Toc478543512)

[6.4 Societal Implications 23](#_Toc478543513)

[Appendix A: Terms and Definitions A-1](#_Toc478543514)

[Appendix B: Acronyms B-1](#_Toc478543515)

[Appendix C: References C-1](#_Toc478543516)

Figure

[Figure 1: Skills Needed in Data Science 18](#_Toc477251086)

Table

[Table 1: Sampling of Definitions Attributed to Big Data 6](#_Toc477251087)

Executive Summary

The summary below will be updated when the document content is near finalization.

The NIST Big Data Public Working Group (NBD-PWG) Definitions and Taxonomy Subgroup prepared this *NIST Big Data Interoperability Framework: Volume 1, Definitions* to address fundamental concepts needed to understand the new paradigm for data applications, collectively known as Big Data, and the analytic processes collectively known as Data Science. While Big Data has been defined in a myriad of ways, the shift to a Big Data paradigm occurs when the scale of the data leads to the need for parallelization through a cluster of computing and storage resources to enable cost-effective data management. Data science combines various technologies, techniques, and theories from various fields, mostly related to computer science, linguistics, and statistics, to obtain useful knowledge from data. This report seeks to clarify the underlying concepts of Big Data and Data Science to enhance communication among Big Data producers and consumers. By defining concepts related to Big Data and Data Science, a common terminology can be used among Big Data practitioners.

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 development stages of the NBD-PWG work. The three stages aim to achieve the following with respect to the NIST Big Data Reference Architecture (NBDRA).

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

Potential areas of future work for the Subgroup during stage 3 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.

# 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 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:

* How is Big Data defined?
* What attributes define Big Data solutions?
* What is the significance of possessing Big Data?
* 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.[[2]](#endnote-2) 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 Standards Roadmap. Forum participants noted that this roadmap should define and prioritize Big Data requirements, including interoperability, portability, reusability, extensibility, data usage, analytics, and technology infrastructure. In doing so, the roadmap would accelerate the adoption of the most secure and effective Big Data techniques and technology.

On June 19, 2013, the NIST Big Data Public Working Group (NBD-PWG) was launched with extensive participation by industry, academia, and government from across the nation. The scope of the NBD-PWG involves forming a community of interests from all sectors—including industry, academia, and government—with the goal of developing consensus on definitions, taxonomies, secure reference architectures, security and privacy, and—from these—a standards roadmap. Such a consensus would create a vendor-neutral, technology- and infrastructure-independent framework that would enable Big Data stakeholders to identify and use the best analytics tools for their processing and visualization requirements on the most suitable computing platform and cluster, while also allowing value-added from Big Data service providers.

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

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

On September 16, 2015, seven volumes *NIST Big Data Interoperability Framework* V1.0 documents were published (http://bigdatawg.nist.gov/V1\_output\_docs.php), 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

Currently the NBD-PWG is working on Stage 2 with the goals to enhance the version 1 content, define general interfaces between the NBDRA components by aggregating low-level interactions into high-level general interfaces, and demonstrate how the NBDRA can be used. As a result, the following two additional volumes have been identified.

* Volume 8, Reference Architecture Interfaces
* Volume 9, Adoption and Modernization

Potential areas of future work for each volume during Stage 3 are highlighted in Section 1.5 of each volume. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

## Scope and Objectives of the Definitions and Taxonomies Subgroup

This volume was prepared by the NBD-PWG Definitions and Taxonomy Subgroup, which focused on identifying Big Data concepts and defining related terms in areas such as data science, reference architecture, and patterns.

The aim of this volume is to provide a common vocabulary for those involved with Big Data. For managers, the terms in this volume will distinguish the concepts needed to understand this changing field. For procurement officers, this document will provide the framework for discussing organizational needs and distinguishing among offered approaches. For marketers, this document will provide the means to promote solutions and innovations. For the technical community, this volume will provide a common language to better differentiate the specific offerings.

## Report Production

This section will be revised.

*Big Data* and *Data Science* are being used as buzzwords and are composites of many concepts. To better identify those terms, the NBD-PWG Definitions and Taxonomy Subgroup first addressed the individual concepts needed in this disruptive field. Then, the two over-arching buzzwords⎯Big Data and Data Science⎯and the concepts they encompass were clarified.

To keep the topic of data and data systems manageable, the Subgroup attempted to limit discussions to differences affected by the existence of Big Data. Expansive topics such as data type or analytics taxonomies and metadata were only explored to the extent that there were issues or effects specific to Big Data. However, the Subgroup did include the concepts involved in other topics that are needed to understand the new Big Data methodologies.

Terms were developed independent of a specific tool or implementation, to avoid highlighting specific implementations, and to stay general enough for the inevitable changes in the field.

The Subgroup is aware that some fields, such as legal, use specific language that may differ from the definitions provided herein. The current version reflects the breadth of knowledge of the Subgroup members. During the comment period, the broader community is requested to address any domain conflicts caused by the terminology used in this volume.

To achieve technical and high quality document content, this document will go through public comments period along with NIST internal review.

## Report Structure

This section will be revised.

This volume seeks to clarify the meanings of the broad terms Big Data and Data Science. Big Data and Big Data Engineering are discussed in Sections 2 and 3. Section 4 explores concepts of Data Science, Section 5 explores Security and Privacy, and Section 6 discusses Management concepts. This second version of *NIST Big Data Interoperability Framework: Volume 1, Definitions* describes some of the fundamental concepts that will be important to determine categories or functional capabilities that represent architecture choices.

Tightly coupled information can be found in the other volumes of the *NIST Big Data Interoperability Framework. Volume 2, Taxonomies* provides a description of the more detailed components of the NIST Big Data Reference Architecture (NBDRA) presented in *Volume 6, Reference Architecture*. Security- and privacy-related concepts are described in detail in *Volume 4, Security and Privacy*. To understand how these systems are architected to meet users’ needs, the reader is referred to *Volume 3, Use Cases and General Requirements*. *Volume 7, Standards Roadmap* recaps the framework established in Volumes 1 through 6 and discusses NBDRA-related standards. Comparing related sections in these volumes will provide a more comprehensive understanding of the consensus of the NBD-PWG.

## Future Work on this Volume

This section will be written when the document content is near finalization.

# Big Data

The rate of growth of data generated and stored has been increasing exponentially. In a 1965 paper,[[3]](#endnote-3) Gordon Moore estimated that the density of transistors on an integrated circuit board was doubling every two years. Known as “Moore’s Law,” this rate of growth has been applied to all aspects of computing, from clock speeds to memory. The growth rates of data volumes are estimated to be faster than Moore’s Law, with data volumes more than doubling every eighteen months. This data explosion is creating opportunities for new ways of combining and using data to find value, as well as providing significant challenges due to the size of the data being managed and analyzed. One significant shift is in the amount of unstructured data. Historically, structured data has typically been the focus of most enterprise analytics, and has been handled through the use of the relational data model. Recently, the quantity of unstructured data, such as micro-texts, web pages, relationship data, images and videos, has exploded, and the trend indicates an increase in the incorporation of this unstructured data to generate value. The central benefit of Big Data analytics is the ability to process large amounts and various types of information. Big Data does not imply that the current data volumes are simply “bigger” than before, or bigger than current techniques can efficiently handle. The need for greater performance or efficiency happens on a continual basis. However, Big Data represents a fundamental change in the architecture needed to efficiently handle current datasets.

In the evolution of data systems, there have been a number of times when the need for efficient, cost-effective data analysis has forced a revolution in existing technologies. For example, the move to a relational model occurred when methods to reliably handle changes to structured data led to the shift toward a data storage paradigm that modeled relational algebra. That was a fundamental shift in data handling. The current revolution in technologies referred to as Big Data has arisen because the relational data model can no longer efficiently handle all the current needs for analysis of large and often unstructured datasets. It is not just that data is bigger than before, as it has been steadily getting larger for decades. The Big Data revolution is instead a one-time fundamental shift in architecture, just as the shift to the relational model was a one-time shift. As relational databases evolved to greater efficiencies over decades, so too will Big Data technologies continue to evolve. Many of the conceptual underpinnings of Big Data have been around for years, but the last decade has seen an explosion in their maturation and application to scaled data systems.

The term Big Data has been used to describe a number of concepts, in part because several distinct aspects are consistently interacting with each other. To understand this revolution, the interplay of the following four aspects must be considered: the characteristics of the datasets, the analysis of the datasets, the performance of the systems that handle the data, and the considerations of cost-effectiveness.

In the following sections, the three broad concepts, Big Data, Big Data Engineering and Data Science, are broken down into specific individual terms and concepts.

## Big Data Definitions

Big Data refers to the inability of traditional data architectures to efficiently handle the new datasets. Characteristics of Big Data that force new architectures are as follows:

* ***Volume*** (i.e., the size of the dataset);
* ***Variety*** (i.e., data from multiple repositories, domains, or types);
* ***Velocity*** (i.e., rate of flow); and
* ***Variability*** (i.e., the change in velocity or structure).

These characteristics—volume, variety, velocity, and variability—are known colloquially as the ‘Vs’ of Big Data and are further discussed in Section 2.2. While many other V’s have been attributed to Big Data, only the above four drive the shift to new scalable architectures for data-intensive applications, in order to achieve cost-effective performance. These Big Data characteristics dictate the overall design of a Big Data system, resulting in different data system architectures or different analytics life cycle process orderings to achieve needed efficiencies.

**Big Data** consists of extensive datasets⎯primarily in the characteristics of volume, variety, velocity, and/or variability⎯that require a scalable architecture for efficient storage, manipulation, and analysis.

Note that this definition contains the interplay between the characteristics of the data and the need for a system architecture that can scale to achieve the needed performance and cost efficiency. Scalable refers to the use of additional resources to handle additional load. Ideally if you have twice the amount of data you would have linear scaling – meaning you could process the data in the same amount of time if you had twice as many processing nodes in your cluster. There are two fundamentally different methods for system scaling, often described metaphorically as “vertical” or “horizontal” scaling.

***Vertical Scaling*** *is increasing the performance of data processing through improvements to processors, memory, storage, or connectivity.*

Vertical scaling has be the core method for increasing performance. Again following Moore’s Law, the historical trend is to achieve greater system performance through improved components. Then as theoretical limits to physical capability were approached in a technology, completely new technologies have been developed to continue the scaling of Moore’s Law.

***Horizontal Scaling*** *is increasing the performance of distributed data processing through the addition of nodes in a cluster for additional resources.*

The core of the new data-intensive processing technologies of extensive datasets has been the maturation of techniques for distributed processing across independent resources, or nodes in a cluster. This distributed processing can in fact be parallel processing for one of the first techniques known as scatter-gather, which will be discussed in Section 3 on Big Data Engineering. It is this horizontal scaling, which we will refer to as parallel processing that is at the heart of the Big Data revolution.

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

While Big Data strictly speaking should apply to the characteristics of the data, the term also refers to this paradigm shift that suggests that a system is a Big Data system when the scale of the data causes the management of the data to be a significant driver in the design of the system architecture. The selection of the 4 V’s of volume, velocity, variety, and variability as fundamental for their importance in selecting processing architectures will be discussed in Section 2.2.

While the methods to achieve efficient scalability across resources will continually evolve, this paradigm shift is a one-time occurrence. The same shift to parallelism for compute-intensive systems occurred in the late 1990’s for large-scale science and engineering simulations. The simulation community went through the same shift in techniques for parallel processing analogously to the more recent shift in data-intensive processing techniques for handing Big Data.

The next question is when do you have a Big Data Problem?

## When do you have a big data problem?

One main issue is when a problem is considered within the realm of Big Data. Developing universal criteria for the determination that you have a Big Data problem is difficult since the choice to use the parallel Big Data architecture is based on the interplay of performance, cost and time constraints on the end-to-end system processing. Designation of a problem as a Big Data problem depends on an analysis of the application’s requirements. There are four fundamental drivers that determine if you have a big data problem:

### Volume

The most commonly recognized characteristic of Big Data is the presence of extensive datasets - representing the large amount of data available for analysis to extract valuable information. There is an implicit assumption that greater value results from processing more of the data. There are many examples of this, known as the network effect, where data models improve with greater amounts of data. Much of the advances from machine learning arise from those techniques that process more data. As an example, object recognition in images significantly improved when the numbers of images went from thousands into millions. The time and expense required to process massive datasets was one of the original drivers for distributed processing. Volume drives the need for computation and storage parallelism and its management during processing of large data sets.

### Velocity

Velocity is a measure of the rate of arrival of the data in a dataset. Traditionally, high velocity systems have been described as ***streaming data***. While these aspects are new for some industries, other industries (e.g., telecommunications) have processed high volume and short time interval data for years. Data in motion is processed and analyzed in real time, or near real time, and has to be handled in a very different way than data at rest (i.e., persisted data). Data in motion tends to resemble event-processing architectures, and focuses on real-time or operational intelligence applications. The need for real-time data processing, even in the presence of large data volumes, drives a different type of architecture where the data is not stored, but is processed typically in memory. Note that time constraints for real-time processing can present the need for distributed processing even when the datasets are relatively small—a scenario often present in the Internet of Things.

### Variety

The variety characteristic represents the need to analyse data from a number of domains and a number of data types. The variety of data from multiple domains was previously handled through transformations or pre-analytics to identify features that would allow fusion with other data. Automated data fusion relies on semantic metadata, where the understanding of the data through the metadata allows it to be machine operable. The integrated data from the variety of datasets was then placed into a database or data warehouse. Analytics across a range of data types, logical models, timescales, and semantics complicates the development of analytics that can span this variety of data. Distributed processing allows individual pre-analytics on different types of data, followed by different analytics to span these interim results.

### Variability

High variability refers to changes in dataset, whether data flow rate, format/structure, semantics, and/or quality that impact the analytics application. Impacts can include the need to refactor architectures, interfaces, processing/algorithms, integration/fusion, storage, applicability, or even the ability to use the data. The data processing is often tied up with the automatic provisioning of additional virtualized resources in a cloud environment. Detailed discussions of the techniques used to process data can be found in other industry publications that focus on operational cloud architectures.[[4]](#endnote-4) [[5]](#endnote-5) Variability in data volumes implies the need to scale-up or scale-down virtualized resources to efficiently handle the additional processing load, one of the advantages capabilities of cloud computing. This keeps systems efficient rather than having to design and resource to the expected peak capacity—resulting in a system that at most times sits idle. It should be noted that this variability refers to dataset characteristics, whereas the term volatility refers to the changing values of individual data elements. Since the latter does not affect the architecture—only the analytics—it is only variability that affects the architectural design.

## Other Usage of the Term Big Data

A number of Big Data definitions have been suggested as efforts have been made to understand the extent of this new field. Several Big Data concepts, discussed in previous sections, were observed in a sample of definitions taken from blog posts and magazine articles.[[6]](#endnote-6), [[7]](#endnote-7), [[8]](#endnote-8), [[9]](#endnote-9), [[10]](#endnote-10), [[11]](#endnote-11), [[12]](#endnote-12), [[13]](#endnote-13), [[14]](#endnote-14) The sample of formal and informal definitions offer a sense of the spectrum of concepts applied to the term Big Data. The sample of Big Data concepts and definitions are aligned in Table 1. The NBD-PWG’s definition is closest to the Gartner definition, with additional emphasis that the horizontal scaling is the element that provides the cost efficiency. The Big Data definitions in Table 1 are not comprehensive, but rather illustrate the interrelated concepts attributed to the catch-all term Big Data.

Table 1: Sampling of Definitions Attributed to Big Data

| Concept | Author | Definition |
| --- | --- | --- |
| 3Vs | Gartner | “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” |
| Volume | Techtarget**.** | “Although Big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data.” |
| Oxford English Dictionary (OED) | “big data n. Computing (also with capital initials) data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data.” |
| “Bigger” Data | Annette Greiner**.** | “Big data is data that contains enough observations to demand unusual handling because of its sheer size, though what is unusual changes over time and varies from one discipline to another.” |
| Not Only Volume | Quentin Hardy | “What’s ‘big’ in big data isn’t necessarily the size of the databases, it’s the big number of data sources we have, as digital sensors and behavior trackers migrate across the world.” |
| Chris Neumann | “…our original definition was a system that (1) was capable of storing 10 TB of data or more … As time went on, diversity of data started to become more prevalent in these systems (particularly the need to mix structured and unstructured data), which led to more widespread adoption of the “3 Vs” (volume, velocity, and variety) as a definition for big data.” |
| Big Data Engineering | IDC | “Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis.” |
| Hal Varian | “Big data means data that cannot fit easily into a standard relational database.” |
| McKinsey | “Big Data refers to a dataset whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.” |
| Less Sampling | John Foreman | “Big data is when your business wants to use data to solve a problem, answer a question, produce a product, etc., crafting a solution to the problem that leverages the data without simply sampling or tossing out records.” |
| Peter Skomoroch | “Big data originally described the practice in the consumer Internet industry of applying algorithms to increasingly large amounts of disparate data to solve problems that had suboptimal solutions with smaller datasets.” |
| New Data Types | Tom Davenport | “The broad range of new and massive data types that have appeared over the last decade or so.” |
| Mark van Rijmenam | “Big data is not all about volume, it is more about combining different data sets and to analyze it in real-time to get insights for your organization. Therefore, the right definition of big data should in fact be: mixed data.” |
| Analytics | Ryan Swanstrom | “Big data used to mean data that a single machine was unable to handle. Now big data has become a buzzword to mean anything related to data analytics or visualization.” |
| Data Science | Joel Gurin | “Big data describes datasets that are so large, complex, or rapidly changing that they push the very limits of our analytical capability.” |
| Josh Ferguson | “Big data is the broad name given to challenges and opportunities we have as data about every aspect of our lives becomes available. It’s not just about data though; it also includes the people, processes, and analysis that turn data into meaning.” |
| Value | Harlan Harris | “To me, ‘big data’ is the situation where an organization can (arguably) say that they have access to what they need to reconstruct, understand, and model the part of the world that they care about.” |
| Jessica Kirkpatrick | “Big data refers to using complex datasets to drive focus, direction, and decision making within a company or organization.” |
| Hilary Mason | “Big data is just the ability to gather information and query it in such a way that we are able to learn things about the world that were previously inaccessible to us.” |
| Gregory Piatetsky-Shapiro | “The best definition I saw is, “Data is big when data size becomes part of the problem.” However, this refers to the size only. Now the buzzword “big data” refers to the new data-driven paradigm of business, science and technology, where the huge data size and scope enables better and new services, products, and platforms.” |
| Cultural Change | Drew Conway | “Big data, which started as a technological innovation in distributed computing, is now a cultural movement by which we continue to discover how humanity interacts with the world—and each other—at large-scale.” |
| Daniel Gillick | “‘Big data’ represents a cultural shift in which more and more decisions are made by algorithms with transparent logic, operating on documented immutable evidence. I think ‘big’ refers more to the pervasive nature of this change than to any particular amount of data.” |
| Cathy O’Neil | “‘Big data’ is more than one thing, but an important aspect is its use as a rhetorical device, something that can be used to deceive or mislead or overhype.” |

# Big Data Engineering (Frameworks)

Some definitions for Big Data, as described in Section 2.3, focus on the systems engineering innovations required because of the characteristics of Big Data. The NBD-PWG has adopted the following definition of Big Data engineering.

**Big Data engineering** includes advanced techniques that harness independent resources for building scalable data systems when the characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis.

Once again the definition is coupled, so that Big Data engineering is used when the volume, velocity, variety, or variability characteristics of the data require it. New engineering techniques in the data layer have been driven by the growing prominence of datasets that cannot be handled efficiently in a traditional relational model. The need for scalable access in structured data has led to software built on the key-value pair paradigm. The rise in importance of document analysis has spawned a document-oriented database paradigm, and the increasing importance of relationship data has led to efficiencies through the use of graph-oriented data storage.

## Horizontal Infrastructure Scaling

Section 2.1 defines horizontal scaling—adding additional resources (processing nodes) as the amount of data to be processed increases. Horizontal scaling can occur at a number of points in the infrastructure.

* ***Shared-disk File Systems*:** Approaches, such as Storage Area Networks (SANs) and Network Attached Storage (NAS)*,* use a single storage pool, which is accessed from multiple computing resources. While these technologies solved many aspects of accessing very large datasets from multiple nodes simultaneously, they suffered from issues related to data locking and updates and, more importantly, created a performance bottleneck (from every input/output [I/O] operation accessing the common storage pool) that limited their ability to scale up to meet the needs of many Big Data applications. These limitations were overcome through the implementation of fully *distributed file systems.*
* ***Distributed File Systems***: In distributed file storage systems, multi-structured (object) datasets are distributed across the computing nodes of the server cluster(s). The data may be distributed at the file/dataset level, or more commonly, at the block level, allowing multiple nodes in the cluster to interact with different parts of a large file/dataset simultaneously. Big Data frameworks are frequently designed to take advantage of data locality to each node when distributing the processing, which avoids any need to move the data between nodes. In addition, many distributed file systems also implement file/block level replication where each file/block is stored multiple times on different machines for both reliability/recovery (data is not lost if a node in the cluster fails), as well as enhanced data locality. Any type of data and many sizes of files can be handled without formal extract, transformation, and load conversions, with some technologies performing markedly better for large file sizes.
* ***Distributed Computing:*** The popular framework for distributed computing consists of a storage layer and processing layer combination that implements a multiple-class, algorithm-programming model. Low-cost servers supporting the distributed file system that stores the data can dramatically lower the storage costs of computing on a large scale of data (e.g., web indexing). ***Map/Reduce*** is the default processing component in data-distributed computing where the query is “scattered” across the processors and the results are “gathered” into a central processor. Processing results are typically then loaded into an analysis environment.

The use of inexpensive servers is appropriate for slower, batch-speed Big Data applications, but do not provide good performance for applications requiring low latency processing. The use of basic Map/Reduce for processing places limitations on updating or iterative access to the data during computation. Bulk Synchronous Parallelism systems or newer Map/Reduce developments can be used when repeated updating is a requirement. Improvements and “generalizations” of Map/Reduce have been developed that provide additional functions lacking in the older technology, including fault tolerance, iteration flexibility, elimination of middle layer, and ease of query.

A more general description of scatter/gather processes is often referred to as *moving the processing to the data, not the data to the processing.*

**Computational portability** is the movement of the computation to the location of the data.

The implication is that data is too extensive to be queried and moved into another resource for analysis, so the analysis program is instead distributed (scattered) to the data-holding resources, with only the results being aggregated (gathered) on a remote resource. This concept of data locality is a new aspect of parallel data architectures that was not a critical issue before storage leveraged multiple independent nodes. Additional system concepts are the *interoperability* (ability for tools to work together), *reusability* (ability to apply tools from one domain to another), and *extendibility* (ability to add or modify existing tools for new domains). These system concepts are not specific to Big Data, but their presence in Big Data can be understood in the examination of a Big Data reference architecture, which is discussed in *NIST Big Data Interoperability Framework:* *Volume 6, Reference Architecture* of this series.

***Resource Negotiation:*** The common distributed computing system has little in the way of built-in data management capabilities. In response, several technologies have been developed to provide the necessary support functions, including operations management, workflow integration, security, and governance. Of special importance to resource management development are new features for supporting additional processing models (other than Map/Reduce) and controls for multi-tenant environments, higher availability, and lower latency applications.

In a typical implementation, the resource manager is the hub for several node managers. The client or user accesses the resource manager which in turn launches a request to an application master within one or many node managers. A second client may also launch its own requests, which will be given to other application masters within the same or other node managers. Tasks are assigned a priority value allocated based on available CPU and memory, and provided the appropriate processing resource in the node.

Data movement is normally handled by transfer and application program interface (API) technologies other than the resource manager. In rare cases, peer-to-peer (P2P) communications protocols can also propagate or migrate files across networks at scale, meaning that technically these P2P networks are also distributed file systems. The largest social networks, arguably some of the most dominant users of Big Data, move binary large objects (BLOBs) of over 1 gigabyte (GB) in size internally over large numbers of computers via such technologies. The internal use case has been extended to private file synchronization, where the technology permits automatic updates to local folders whenever two end users are linked through the system.

In external use cases, each end of the P2P system contributes bandwidth to the data movement, making this currently the fastest way to leverage documents to the largest number of concurrent users. For example, this technology is used to make 3GB images available to the public, or to allow general purpose home computers to devote compute power for scientific challenges such as protein folding. However, any large bundle of data (e.g., video, scientific data) can be quickly distributed with lower bandwidth cost.

There are additional aspects of Big Data that are changing rapidly and are not fully explored in this document, including cluster management and other mechanisms for providing communication among the cluster resources holding the data in the non-relational models. Discussion of the use of multiple tiers of storage (e.g., in-memory, cache, solid state drive, hard drive, network drive) in the newly emerging software-defined storage can be found in other industry publications. Software-defined storage is the use of software to determine the dynamic allocation of tiers of storage to reduce storage costs while maintaining the required data retrieval performance.

***Concurrency***: A definition and discussion of concurrency is needed.

## Scalable Logical Data Storage

Essentially, Big Data refers to the extensibility of data repositories and data processing across resources working in parallel, in the same way that the compute-intensive simulation community embraced massively parallel processing two decades ago. By working out methods for communication among resources, the same scaling is now available to data-intensive applications. From the user’s perspective, however, the data is stored in a single logical structure.

The new non-relational model database paradigms are typically referred to as ***NoSQL*** (Not Only or No Structured Query Language [SQL]) systems. The problem with identifying Big Data storage paradigms as NoSQL is, first, that it describes the storage of data with respect to a set theory-based language for query and retrieval of data, and second, that there is a growing capability in the application of the SQL query language against the new non-relational data repositories. While NoSQL is in such common usage that it will continue to refer to the new data models beyond the relational model, it is hoped that the term itself will be replaced with a more suitable term, since it is unwise to name a set of new storage paradigms with respect to a query language currently in use against that storage.

**Non-relational models**, frequently referred to as **NoSQL**, refer to logical data models that do not follow relational algebra for the storage and manipulation of data.

Data elements are collected into records that describe a particular observation, event, or transaction. Previously, most of the data in business systems was ***structured***data, where each record was consistently structured and could be described efficiently in a ***relational model***. Records are conceptualized as the rows in a table where data elements are in the cells. Unstructured data types, such as text, image, video, and relationship data, have been increasing in both volume and prominence. While modern relational databases tend to have support for these types of data elements, their ability to directly analyze, index, and process them has tended to be both limited and accessed via nonstandard SQL extensions. The need to analyze ***unstructured*** or ***semi-structured*** data has been present for many years. However, the Big Data paradigm shift has increased the emphasis on extracting the value from unstructured or relationship data, and also on different engineering methods that can handle data more efficiently.

Big Data Engineering refers to the new ways that data is stored in records. In some cases, the records are still in the concept of a table structure. One storage paradigm is a key-value structure, with a record consisting of a key and a string of data together in the value. The data is retrieved through the key, and the non-relational database software handles accessing the data in the value. This can be viewed as a subset/simplification of a relational database table with a single index field and column. A variant on this is the document store, where the document has multiple value fields, any of which can be used as the index/key. The difference from the relational table model is that the set of documents do not need to have all the same value fields.

Another type of new Big Data record storage is in a graphical model. A graphical model represents the relationship between data elements. The data elements are nodes, and the relationship is represented as a link between nodes. Graph storage models represent each data element as a series of subject, predicate, and object triples. Often, the available types of objects and relationships are described via ontologies as discussed above.

There is another variant of scalable databases, this one being described as NewSQL. This term refers to databases that follow a number of principles of relational database transactions. The reader is referred to the extensive literature to learn about the CAP theorem, ACID and BASE.

## Relationship to other Technological Innovations

Big Data is related to a number of other domains where computational innovations are being developed.

### High Performance Computing

Subsection Scope: Enhance discussion. Add references to other HPC documents.

As stated above, fundamentally, the Big Data paradigm is a shift in data system architectures from monolithic systems with vertical scaling (i.e., adding more power, such as faster processors or disks, to existing machines) into a parallelized, “horizontally scaled”, system (i.e., adding more machines to the available collection) that uses a loosely coupled set of resources in parallel. This type of parallelization shift began over 20 years ago for compute-intensive applications in High Performance Computing (HPC) communities, when scientific simulations began using massively parallel processing (MPP) systems.

**Massively parallel processing** refers to a multitude of individual processors working in parallel to execute a particular program.

In different combinations of splitting the code and data across independent processors, computational scientists were able to greatly extend their simulation capabilities. This, of course, introduced a number of complications in such areas as message passing, data movement, latency in the consistency across resources, load balancing, and system inefficiencies, while waiting on other resources to complete their computational tasks.

The Big Data paradigm, likewise, is undergoing the shift to parallelism for data-intensive applications. Data systems need a level of extensibility that matches the scaling in the data. To get that level of extensibility, different mechanisms are needed to distribute data and for data retrieval processes across loosely coupled resources.

The obvious next focus area will be in systems that are a blend of compute-intensive and data-intensive processing. New work to add more data-intensive capabilities to HPC systems is known as High Performance Data Analytics (HPDA). Some problems need more extensive computation than the ‘embarrassingly parallel’ distribution of data onto separate nodes in Map/Reduce systems. Data-intensive machine learning (which also requires significant computation) is known as Big Data Analytics (BDA). Deep Learning is an example of this type of machine learning – needing lots of data and lots of computation.

### Cloud

Subsection Scope: Enhance discussion. Add references to other cloud documents.

Early Big Data systems built by Internet search providers and others were frequently deployed on bare metal to achieve the best efficiency at distributing I/O across the clusters and multiple storage devices. While cloud (i.e., virtualized) infrastructures were frequently used to test and prototype Big Data deployments, there are recent trends, due to improved efficiency in I/O virtualization infrastructures, of production solutions being deployed on cloud or Infrastructure-as-a-Service (IaaS) platforms. A high-velocity system with high variability may be deployed on a cloud infrastructure, because of the cost and performance efficiency of being able to add or remove nodes to handle the peak performance. Being able to release those resources when they are no longer needed provides significant cost savings for operating this type of Big Data system. Very large implementations and in some cases cloud providers are now implementing this same type of elastic infrastructure on top of their physical hardware (known as “bare metal”). This is especially true for organizations that already have extensive infrastructure but simply need to balance resources across application workloads that can vary.

### Internet of Things

Subsection Scope: Enhance discussion. Add references to other IoT documents.

The Internet of Things (IoT) refers to the connection of physical devices to the Internet that can produce data and communicate with other devices. IoT represents the other extreme from high volume Big Data that represents one or a few very large datasets. Conversely IoT represents a very large number of potentially very small streams of data. IoT devices are designed to handle the tradeoffs of generating transmitting all of the data or providing some on-board analytics to reduce the data needing to be transmitted – critical in low bandwidth scenarios. Given the typical simplicity of IoT devices, IoT presents a number of challenges for security and privacy.

### Embedded Computing – Cyber-Physical Systems

Subsection Scope: This subsection will discuss the scope of cyber-physical systems and implications for Big Data. Smart cities/cars/buildings/etc. could be included. Add references to other CPS documents.

### Blockchain

Subsection scope: This subsection will discuss the definition of blockchain and the relation to Big Data. Blockchain refers to a distributed database that maintains a continuously-growing list of ordered records called blocks.

# 

# Data Science Applications

In its purest form, Data Science is the *fourth paradigm* of science, following experiment, theory, and computational sciences. The fourth paradigm is a term coined by Dr. Jim Gray in 2007.[[15]](#endnote-15) Data-intensive science, shortened to data science, refers to the conduct of data analysis as an empirical science, learning directly from data itself. This can take the form of collecting data followed by open-ended analysis without preconceived hypothesis (sometimes referred to as discovery or data exploration). The second empirical method refers to the formulation of a hypothesis, the collection of the data—new or preexisting—to address the hypothesis, and the analytical confirmation or denial of the hypothesis (or the determination that additional information or study is needed.) In both methods, the conclusions are based on the data. In many data science projects, the original data is browsed first, which informs a hypothesis, which is then investigated. As in any experimental science, the end result could be that the original hypothesis itself needs to be reformulated. The key concept is that data science is an empirical science, performing the scientific process directly on the data. Note that the hypothesis may be driven by a need, or can be the restatement of a need in terms of a technical hypothesis.

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

Data science can be understood as the activities happening in the processing layer of the system architecture, against data stored in the data layer, in order to extract knowledge from the data.

## When is it Data Science?

The term Data Science arose in the context of Big Data, where the size of the data became part of the problem. Any analysis of large volume or high velocity streaming data has to consider the distribution of the data and the engineering methods being used for management and processing. Data Science is the extension of the analytics lifecycle to encompass Big Data.

The term **analytics** refers to the discovery of meaningful patterns in data, and is one of the steps in the data analytics life cycle.

The **data** **analytics life cycle** is the set of processes in an application that transform data into useful knowledge.

The end-to-end data analytics life cycle consists of four fundamental steps:

1. ***Collection***: This stage gathers and stores data in its original form (i.e., raw data.).
2. ***Preparation***: This stage involves the collection of processes that convert raw data into cleansed, organized information.
3. ***Analysis***: This stage involves the techniques that produce synthesized knowledge from organized information.
4. ***Visualization***: A description for this new step is needed.
5. ***Action***: This stage involves processes that use the synthesized knowledge to generate value for the enterprise.

Analytics is used to refer to the methods, their implementations in tools, and the results of the use of the tools as interpreted by the practitioner.

The **analytics** **process** is the synthesis of knowledge from information.

Analytics strictly speaking refers to a specific process in the complete data analytics life cycle, whereas Data Science involves all steps in the data analytics life cycle including when Big Data is being processed.

Analytic methods were classically developed for simple data tables. When storage became expensive, relational databases and methods were developed. With the advent of less expensive storage and Big Data, new strategies to manage large volumes don’t necessarily require the use of relational methods. These new strategies for Big Data Engineering involve rearrangement of the data, parallel storage, parallel processing, and revisions to algorithms. With the new Big Data paradigm, analytics implementations can no longer be designed independent of the data storage design, as could be previously assumed if the data was already cleansed and stored in a relational model. In the Data Science realm, analytics implementations must now be designed at the same time that the data distribution is being designed for storage – the techniques are no longer independent and separable as they were before.

Data science across the entire data analytics life cycle incorporates principles, techniques, and methods from many disciplines and domains including data cleansing, data management, analytics, visualization, engineering, and in the context of Big Data, now also includes Big Data Engineering.

**Data science applications** implement data transformation processes from the data analytics life cycle including Big Data Techniques.

In the traditional data warehouse, the data handling process followed the order above (i.e., collection, preparation, storage, and analysis.) The data warehouse was designed in a way that optimized the intended analytics.

The different Big Data characteristics have influenced changes in the ordering of the data handling processes – in particular when the data is stored. Different Big Data use cases can be characterized in terms of the dataset characteristics and in terms of the time limits for the end-to-end analytics processing. Dataset characteristics change the analytics life cycle processes in different ways, for example:

* ***Data warehouse***: Persistent storage occurs after data preparation.
* ***Big Data volume system***: Data is stored immediately in raw form before preparation; preparation occurs on read, and is referred to as ‘schema on read.’
* ***Big Data velocity application***: The collection, preparation, and analytics (alerting) occur on the fly, and possibly includes some summarization or aggregation prior to storage.

Just as simulations split the analytical processing across clusters of processors, data processes are redesigned to split data transformations across data nodes. Because the data may be too big to move, the transformation code may be sent in parallel across the data persistence nodes, rather than the data being extracted and brought to the transformation servers.

## Data Science Characteristics

Subsection Scope: The discussions below need to be enhanced. Add an analogy to the who, what, when, why, and how to our V words. So we explain that he V words are telling part of the story of Big Data. They help flesh out the storytelling of the Big Data scenario. The more precise understanding is through the BD Characteristics – computational parallelism. Talk about data hubris, missing data, bias, data confounding, and data quality.

In addition to volume, velocity, variety, and variability, several terms, many beginning with V, have been used in connection with Big Data but specifically refer to the analytical processing of the data rather than the characteristic of the datasets themselves. Some of these terms strongly relate to analytics on the data.

### Veracity

Veracity refers to the completeness and accuracy of the data and relates to the vernacular “garbage-in, garbage-out” description for data quality issues in existence for a long time. If the analytics are causal, then the quality of every data element is extremely important. If the analytics are correlations or trending over massive volume datasets, then individual bad elements could be lost in the overall counts and the trend will still be accurate. As mentioned in Section 2.2, many people debate whether “more data is superior to better algorithms,” but that is a topic better discussed elsewhere.

### Validity

Subsection Scope: One paragraph is needed to describe validity and the relation to Big Data. Validity addresses if the data is appropriate to the question being asked.

### Volatility

### Visualization

Subsection Scope: One paragraph is needed to describe visualization and the relation to Big Data. In addition to other items, the subsection could discuss the following:

\* Presentation and aggregation of Big Data sets (visualization)

\* The need to recognize the parallelization needs of presenting visualizations of large datasets.

### Value

Subsection Scope: One paragraph is needed to describe value and the relation to Big Data.

### Metadata

Metadata is descriptive data about objects. Metadata describes additional information about the data such as how and when data was collected and how it has been processed. Metadata should itself be viewed as data with all the requirements for tracking, change management, and security. Many standards are being developed for metadata, for general metadata coverage (e.g., ISO/IEC 11179-x[[16]](#endnote-16)) and discipline-specific metadata (e.g., ISO 19115-x[[17]](#endnote-17) for geospatial data).

Metadata that describes the history of a dataset is called its *provenance*. As *open data* (data available to others) and *linked data* (data that is connected to other data) become the norm, it is increasingly important to have information about how data was collected, transmitted, and processed. Provenance type of metadata guides users to correct data utilization when the data is repurposed from its original collection process in an effort to extract additional value.

*Semantic metadata*, another type of metadata,refers to the description of a data element to assist with proper interpretation. An *ontology* can be conceptualized as a graphic model, representing a semantic relationship between entities. Ontologies are semantic models constrained to follow different levels of logic models. Ontologies and semantic models predated Big Data and are not discussed in depth this document. Ontologies can be very general or extremely domain-specific in nature. A number of mechanisms exist for implementing these unique descriptions, and the reader is referred to the World Wide Web Consortium (W3C) efforts on the semantic web[[18]](#endnote-18) [[19]](#endnote-19) for additional information. Semantic data is important in the new Big Data Paradigm since the Semantic Web represents a Big Data attempt to provide cross-cutting meanings for terms. Again, semantic metadata is especially important for linked data efforts.

*Taxonomies* represent in some sense metadata about data element relationships. Taxonomy is a hierarchical relationship between entities, where a data element is broken down into smaller component parts. While these concepts are important, they predated the Big Data paradigm shift.

### Data Types

Individual data elements have not changed with Big Data and are not discussed in detail in this document. For additional information on data types, readers are directed to the ISO standard ISO/IEC 11404:2007 General Purpose Datatypes,[[20]](#endnote-20) and, as an example, its extension into healthcare information data types in ISO 21090:2011 Health Informatics.[[21]](#endnote-21)

### Complexity

Another data element relationship concept that is not new in the Big Data paradigm shift is the presence of *complexity* between the data elements. There are systems where data elements cannot be analyzed outside the context of other data elements. This is evident, for example, in the analytics for the Human Genome Project, where it is the relationship between the elements and their position and proximity to other elements that matters. The term *complexity* is often attributed to Big Data, but it refers to this interrelationship between data elements or across data records, independent of whether the dataset has the characteristics of Big Data.

### Latency

Subsection Scope: One or more paragraphs are needed to describe latency and the relation to Big Data.

## Big Data Science

Data science is tightly linked to Big Data, and refers to the management and execution of the end-to-end data processes, including the behaviors of the data system. As such, data science includes all of analytics, but analytics does not include all of data science.

Analytic processes are often characterized as ***discovery*** for the initial hypothesis formulation, ***development*** for establishing the analytics process for a specific hypothesis, and ***applied*** for the encapsulation of the analysis into an operational system. While Big Data has touched all three types of analytic processes, the majority of the changes is observed in development and applied analytics. New Big Data engineering technologies change the types of analytics that are possible, but do not result in completely new types of analytics. However, given the retrieval speeds, analysts are able to interact with their data in ways that were not previously possible. Traditional statistical analytic techniques downsize, sample, or summarize the data before analysis. This was done to make analysis of large datasets reasonable on hardware that could not scale to the size of the dataset. Big Data analytics often emphasize the value of computation across the entire dataset, which gives analysts better chances to determine causation, rather than just correlation. Correlation, though, is still useful when knowing the direction or trend of something is enough to take action. Today, most analytics in statistics and data mining focus on causation—being able to describe why something is happening. Discovering the cause aids actors in changing a trend or outcome. Actors, which in system development can represent individuals, organizations, software, or hardware, are discussed in *NIST Big Data Interoperability Framework: Volume 2, Taxonomy*. Big Data solutions make it more feasible to implement causation type of complex analytics for large, complex, and heterogeneous data.

Another analytics consideration is the speed of interaction between the analytics processes and the person or process responsible for delivering the actionable insight. Analytic data processing speed can fall along a continuum between batch and streaming-oriented processing. Although the processing continuum existed prior to the era of Big Data, the desired location on this continuum is a large factor in the choice of architectures and component tools to be used. Given the greater query and analytic speeds within Big Data due to the scaling across a cluster, there is an increasing emphasis on interactive (i.e., real-time) processing. Rapid analytics cycles allow an analyst to do exploratory discovery on the data, browsing more of the data space than might otherwise have been possible in any practical time frame. The processing continuum is further discussed in *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*.

### Data Cleansing

Subsection Scope: Tell the story of what is required to clean up the data along various strategies (e.g., data cleansing (correcting sematic issues such as Baltimore, HI should be Baltimore, MD), data preening/formatting (Street vs. St.), anything that will make the data better). Possibly discuss fake data. Mention major concepts that can make data dirty (e.g., Wikipedia vs. chatbots fight).

### Sampling

Several issues are currently being debated within the data science community. Two prominent examples are data sampling, and the idea that more data is superior to better algorithms.

Data sampling, a central concept of statistics, involves the selection of a subset of data from the larger data population. Provided that the subset is adequately representative of the larger population, the subset of data can be used to explore the appropriateness of the data population for specific hypothesis tests or questions. For example, it is possible to calculate the data needed to determine an outcome for an experimental procedure (e.g., during a pharmaceutical clinical trial).

When the data mining community began, the emphasis was typically on repurposed data (i.e., data used to train models was sampled from a larger dataset that was originally collected for another purpose). The often-overlooked critical step was to ensure that the analytics were not prone to over-fitting (i.e., the analytical pattern matched the data sample but did not work well to answer questions of the overall data population). In the new Big Data paradigm, it is implied that data sampling from the overall data population is no longer necessary since the Big Data system can theoretically process all the data without loss of performance. However, even if all of the available data is used, it still may only represent a population subset whose behaviors led them to produce the data, which might not be the true population of interest. For example, studying Twitter data to analyze people’s behaviors does not represent all people, as not everyone uses Twitter. While less sampling may be used in data science processes, it is important to be aware of the implicit data sampling when trying to address business questions.

The assertion that more data is superior to better algorithms implies that better results can be achieved by analyzing larger samples of data rather that refining the algorithms used in the analytics. The heart of this debate states that a few bad data elements are less likely to influence the analytical results in a large dataset than if errors are present in a small sample of that dataset. If the analytics needs are correlation and not causation, then this assertion is easier to justify. Outside the context of large datasets in which aggregate trending behavior is all that matters, the data quality rule remains “garbage-in, garbage-out,” where you cannot expect accurate results based on inaccurate data.

### Machine Learning

Subsection Scope: One or more paragraphs are needed to describe machine learning and the relation to Big Data. In addition to other items, the subsection could discuss the following:

\* network behavior

\* where the analytics becomes more accurate due to the sheer volume of data, such as deep learning techniques of image classification>

## Emergent Behavior

Subsection Scope: A description of emergent behavior and the relation to Big Data are needed. This subsection could contain the following (along with other items): Emergent behavior (inference) “Matrix Effect” (Mosaic effect) where for example now have PII concerns. Volunteers for this subsection include:

Cavan – share thoughts/outline specific contribution (Mosaic effect)

Ann – write up something with the “inference is emerging”

## Data Scientists

Data scientists and data science teams solve complex data problems by employing deep expertise in one or more of these disciplines, in the context of business strategy, and under the guidance of domain knowledge. Personal skills in communication, presentation, and inquisitiveness are also very important given the complexity of interactions within Big Data systems.

A **data scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs; domain knowledge; analytical skills; and software and systems engineering to manage the end-to-end data processes in the analytics life cycle.

While this full collection of skills can be present in a single individual, it is also possible that these skills, as shown in Figure 1, are covered in the members of a team. A Venn diagram is often shown to describe the overlapping skills needed for data science. In the early days of experiments, experts in a particular domain would perform the data analysis. With the advent of computers for analysis, additional skills in statistics or machine learning were needed for more sophisticated analysis, or domain experts would work with software and system engineers to build customized analytical applications. With the increase in complexity of compute-intensive simulations across parallel processors, computational science techniques were needed to implement the algorithms on these architectures. For data-intensive applications, all of these skill groups are needed to distribute both the data and the computation across systems of resources working in parallel. While data scientists seldom have strong skills in all these areas, they need to have an understanding of all areas to deliver value from data-intensive applications and work in a team that spans all of these areas.

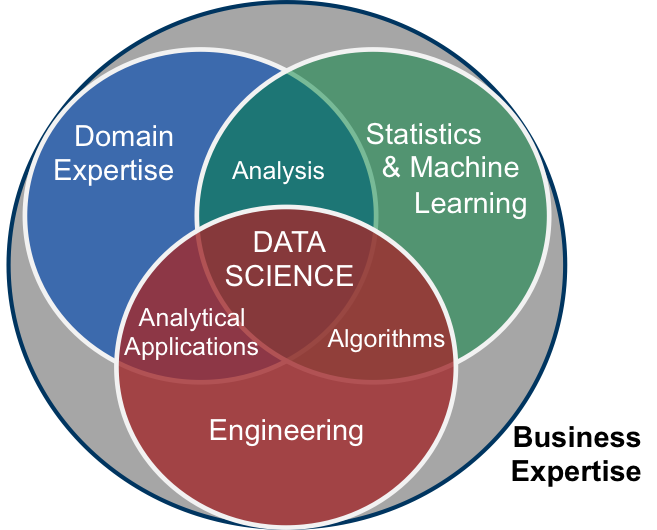


Figure 1: Skills Needed in Data Science

Data science is not solely concerned with analytics, but also with the end-to-end life cycle, where the data system is essentially the scientific equipment being used to develop an understanding of a real-world process. The implication is that the data scientist must be aware of the sources and provenance of the data, the appropriateness and accuracy of the transformations on the data, the interplay between the transformation algorithms and processes, and the data storage mechanisms. This end-to-end overview role ensures that everything is performed correctly to explore the data and create and validate hypotheses. Data science is increasingly used to influence business decisions. These analytics concepts are discussed further in Section 3.4.

## Big Data Metrics and Benchmarks

Subsection Scope: Additional text is needed to round out the discussion of Big Data metrics and benchmarks. Look at other Consortia’s work

Initial considerations in the use of Big Data engineering include the determination, for a particular situation, of the size threshold after which data should be considered Big Data. Multiple factors must be considered in this determination, and the outcome is particular to each application. As described in Section 2.1, Big Data characteristics lead to use of Big Data engineering techniques that allow the data system to operate affordably and efficiently. Whether a performance or cost efficiency can be attained for a particular application requires a design analysis, which is beyond the scope of this report.

There is a significant need for metrics and benchmarking to provide standards for the performance of Big Data systems. While there are a number of standard metrics used in benchmarking, only the ones relevant to the new Big Data Paradigm would be within the scope of this work. This topic is being addressed by the Transaction Processing Performance Council TCP-xHD Big Data Committee, and available information from their efforts may be included in future versions of this report.

# 

# Big Data Security and Privacy

Subsection Scope: This section needs to be expanded and aligned with NBDIF Volume 4: Big Data Security and Privacy. Collaboration with Rose/MITRE team may be possible by sharing thoughts and outlining specific contribution. Additional topics to be discussed in this section and in Volume 4 could be the following:

* Confidentiality
* Encryption
* Data Integrity
* Models (present a base model to later discuss uncertainty models; This is more than metadata, and it's a necessary precondition for the domain model we introduce in Vol 4. This thread has been taken in up academia: A. Kolovos, L. M. Smith, A. Schwab-McCoy, S. Gengler, and H.-L. Yu, "Emerging patterns in multi-sourced data modeling uncertainty," Spatial Statistics, vol. 18, Part A, pp. 300-317, 2016, spatial Statistics Avignon: Emerging Patterns. [Online]. Available: [www.sciencedirect.com/science/article/pii/S2211675316300239](http://www.sciencedirect.com/science/article/pii/S2211675316300239)

Security and privacy have also been affected by the emergence of the Big Data paradigm. A detailed discussion of the influence of Big Data on security and privacy is included in *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy*. Some of the effects of Big Data characteristics on security and privacy summarized below:

* **Variety:** Retargeting traditional relational database security to non-relational databases has been a challenge. An emergent phenomenon introduced by Big Data variety that has gained considerable importance is the ability to infer identity from anonymized datasets by correlating with apparently innocuous public databases.
* **Volume:** The volume of Big Data has necessitated storage in multitiered storage media. The movement of data between tiers has led to a requirement of systematically analyzing the threat models and research and development of novel techniques.
* **Velocity:** As with non-relational databases, distributed programming frameworks such as Hadoop were not developed with security as a primary objective.
* **Variability:** Security and privacy requirements can shift according to the time-dependent nature of roles that collected, processed, aggregated, and stored it. Governance can shift as responsible organizations merge or even disappear

Privacy concerns, and frameworks to address these concerns, predate Big Data. While bounded in comparison to Big Data, past solutions considered legal, social, and technical requirements for privacy in distributed systems, very large databases, and in High Speed Computing and Communications (HPCC). The addition of variety, volume, velocity, and variability to the mix has amplified these concerns to the level of a national conversation, with unanticipated impacts on privacy frameworks.

# Big Data Management

Section Scope: Text is needed to discuss the organization, administration, governance, etc. of Big Data. What new aspects has Big Data brought to data management?

## Orchestration

Subsection Scope: Text is needed to discuss the Orchestrator or Orchestration in Big Data Management. Refer to additional detail in the MindMap (from a previous discussion). Previous discussions on this topic included orchestrator as role, but also as conceptual underpinning for enterprise (“business”) process in the Reference Architecture. The text could also discuss contracts and negotiation among multiple organizations

## Data Governance

Subsection Scope: The text in this section needs to be expanded

Data governance is a fundamental element in the management of data and data systems.

**Data governance** refers to administering, or formalizing, discipline (e.g., behavior patterns) around the management of data.

The definition of data governance includes management across the complete data life cycle, whether the data is at rest, in motion, in incomplete stages, or transactions. To maximize its benefit, data governance must also consider the issues of privacy and security of individuals of all ages, individuals as companies, and companies as companies.

Data governance is needed to address important issues in the new global Internet Big Data economy. For example, many businesses provide a data hosting platform for data that is generated by the users of the system. While governance policies and processes from the point of view of the data hosting company are commonplace, the issue of governance and control rights of the data providers is new. Many questions remain including the following: Do they still own their data, or is the data owned by the hosting company? Do the data producers have the ability to delete their data? Can they control who is allowed to see their data?

The question of governance resides between the value that one party (e.g., the data hosting company) wants to generate versus the rights that the data provider wants to retain to obtain their own value. New governance concerns arising from the Big Data Paradigm need greater discussion, and will be discussed during the development of the next version of this document.

## Data Ownership

Subsection Scope: A description of data ownership and how it is related to Big Data is needed. This subsection could include the following items:

Data Ownership in Big Data

Data is emergent—appears—from any subject. The subject does not require engineering for data to appear. In this way it can be said that the subject is the a priori necessary and sufficient condition, the cause of data. That is, data owes its existence to the data subject. The subject’s existential value is expressed and stored in data. In this way data ownership lies existentially in the data subject.

Data, emergent and apparent from a subject, is the primary common attribute among all subjects. The emergent appearance of data from any subject is more or less a priori compelling for any other subject. Each datum and the emergent data from a subject is existentially valuable to the subject itself.

The stored value of data is realized through—is inculcated in—the understanding and well-being of the subject. The subject’s well-being can be measured through economic, social, physical and many other beneficial attributes.

Since the subject is the cause of the data, the subject must need be the primary but not the sole or exclusive beneficiary owner of the data.

Devices require purpose driven engineering to interact with data from subjects.

For More see [Human Trust Experience](https://www.humantrustexperience.net) and [Big Data Governance Standards](https://bigdatagovernancestandards.org)

* Data ownership (sharing??) and privacy
  + What does data ownership mean?
  + Users can contribute content but company will own them
  + Data ownership involves with intellectual property (licensing, etc.)
  + Cavan: Data ownership: what and who allows access…
* Tim - shared his outline doc submitted to reflector(?) 2-16: Governance Provenance Curation. Sourced Morningstar. Mark suggests [pillar] section #6 is probably the most important. Next / action item for group - Take contributions on relationship to previous work by NARA, maybe something from Quyen
* Quyen: work with Russell’s feedback for Vol. 6 Section 5 on Big Data Lifecycle
* Rose/MITRE team – share thoughts/outline specific contribution
* Governance – architecture view (from velocity, volume, etc.); big data inside/outside ??, Rose

## Societal Implications

Section Scope: Should this section be included here or left to the technical report.

1. Terms and Definitions

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

**Big Data** consists of extensive datasets⎯primarily in the characteristics of volume, variety, velocity, and/or variability⎯that require a scalable architecture for efficient storage, manipulation, and analysis.

**Big Data Engineering** includes advanced techniques that harness independent resources for building scalable data systems when the characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis.

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

**Computational portability** is the movement of the computation to the location of the data.

**Data governance** refers to the overall management of the availability, usability, integrity, and security of the data employed in an enterprise.

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

**Data science** is the empirical synthesis of actionable knowledge from raw data through the complete data life cycle process.

The **data science paradigm** is extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and hypothesis testing.

A d**ata scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes through each stage in the data life cycle.

***Distributed computing*** is a computing system in which components located on networked computers communicate and coordinate their actions by passing messages.

***Distributed file systems*** contain multi-structured (object) datasets that are distributed across the computing nodes of the server cluster(s).

A **federated database system** is a type of meta-database management system, which transparently maps multiple autonomous database systems into a single federated database.

***Horizontal scaling*** implies the coordination of individual resources (e.g., server) that are integrated to act in parallel as a single system (i.e., operate as a cluster).

***Latency*** refers to the delay in processing or in availability.

**Massively parallel processing** refers to a multitude of individual processors working in parallel to execute a particular program.

**Non-relational models**, frequently referred to as NoSQL, refer to logical data models that do not follow relational algebra for the storage and manipulation of data.

***Resource negotiation*** consists of built-in data management capabilities that provide the necessary support functions, such as operations management, workflow integration, security, governance, support for additional processing models, and controls for multi-tenant environments, providing higher availability and lower latency applications.

**Shema on read** is the application of a data schema through preparation steps such as transformations, cleansing, and integration at the time the data is read from the database.

***Shared-disk file systems***, such as Storage Area Networks (SANs) and Network Attached Storage (NAS), use a single storage pool, which is accessed from multiple computing resources.

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

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

***Variability*** refers to the change in other data characteristics.

***Variety*** refers to data from multiple repositories, domains, or types.

***Velocity*** refers to the rate of data flow.

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

***Vertical scaling*** implies increasing the system parameters of processing speed, storage, and memory for greater performance.

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

***Volume*** refers to the size of the dataset.

1. Acronyms

API application programming interface

BLOB binary large object

GB gigabyte

I/O input/output

ISO International Organization for Standardization

ITL Information Technology Laboratory

JTC1 Joint Technical Committee 1

MPP massively parallel processing

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 (or no) Structured Query Language

NSF National Science Foundation

OED Oxford English Dictionary

P2P peer-to-peer

SAN storage area network

SQL Structured Query Language

W3C World Wide Web Consortium

1. References

Document References

1. “Contributors” are members of the NIST Big Data Public Working Group who dedicated great effort to prepare and gave substantial time on a regular basis to research and development in support of this document. [↑](#footnote-ref-2)
2. The White House Office of Science and Technology Policy, “Big Data is a Big Deal,” *OSTP Blog*, accessed February 21, 2014, <http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal>. [↑](#endnote-ref-2)
3. Gordon Moore, "Cramming More Components Onto Integrated Circuits," *Electronics*, Volume 38, Number 8 (1965), pages 114-117. [↑](#endnote-ref-3)
4. Lee Badger, David Bernstein, Robert Bohn, Frederic de Vaulx, Mike Hogan, Michaela Iorga, Jian Mao, John Messina, Kevin Mills, Eric Simmon, Annie Sokol, Jin Tong, Fred Whiteside, and Dawn Leaf, “US Government Cloud Computing Technology Roadmap Volume I: High-Priority Requirements to Further USG Agency Cloud Computing Adoption; and Volume II: Useful Information for Cloud Adopters,” *National Institute of Standards and Technology*, October 21, 2014, <http://dx.doi.org/10.6028/NIST.SP.500-293>. [↑](#endnote-ref-4)
5. Lee Badger, Tim Grance, Robert Patt-Corner, and Jeff Voas, “Cloud Computing Synopsis and Recommendations,” *National Institute of Standards and Technology*, May 2012, <http://csrc.nist.gov/publications/nistpubs/800-146/sp800-146.pdf>. [↑](#endnote-ref-5)
6. ISO/IEC JTC 1 Study Group on Big Data (SGBD), “N0095 Final SGBD Report to JTC1,” September 3, 2014, <http://jtc1bigdatasg.nist.gov/_uploadfiles/N0095_Final_SGBD_Report_to_JTC1.docx>. [↑](#endnote-ref-6)
7. Gartner IT Glossary, “Big Data” (definition), *Gartner.com*, accessed November 17, 2014, <http://www.gartner.com/it-glossary/big-data>. [↑](#endnote-ref-7)
8. Jenna Dutcher, “What is Big Data,” *Data Science at Berkeley Blog*, September 3, 2014, <http://datascience.berkeley.edu/what-is-big-data/>. [↑](#endnote-ref-8)
9. Oxford English Dictionary, “Big Data” (definition), *OED.com*, accessed November 17, 2014, <http://www.oed.com/view/Entry/18833#eid301162178>. [↑](#endnote-ref-9)
10. John Gantz and David Reinsel, “Extracting Value from Chaos,” *IDC iView sponsored by EMC Corp*, accessed November 17, 2014, <http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf>. [↑](#endnote-ref-10)
11. James Manyika et al., “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, May 2011. [↑](#endnote-ref-11)
12. Tom Davenport, “Big Data@Work,” Harvard Business Review Press, February 25, 2014. [↑](#endnote-ref-12)
13. Emerging Technology From the arXiv (Contributor), “The Big Data Conundrum: How to Define It?” MIT Technology Review, October 3, 2013, <http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/>. [↑](#endnote-ref-13)
14. Gil Press (Contributor), “12 Big Data Definitions: What’s Yours?” *Forbes.com*, accessed November 17, 2014, <http://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/>. [↑](#endnote-ref-14)
15. Microsoft Research, “Jim Gray on eScience: A Transformed Scientific Method,” accessed June 1, 2015, <http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf>. [↑](#endnote-ref-15)
16. ISO/IEC 11179-2004, Information technology – “Metadata registries (MDR) – Part 1: Framework,” *International Organization for Standardization*, <http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=35343>. [↑](#endnote-ref-16)
17. ISO 19115-2014, “Geographic information – Metadata – Part 1: Fundamentals,” *International Organization for Standardization*, <http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53798>. [↑](#endnote-ref-17)
18. Phil Archer, “W3C Data Activity Building the Web of Data,” *W3C*, <http://www.w3.org/2013/data/>. [↑](#endnote-ref-18)
19. Dan Brickley and Ivan Herman, “Semantic Web Interest Group,” *W3C*, June 16, 2012, <http://www.w3.org/2001/sw/interest/>. [↑](#endnote-ref-19)
20. ISO/IEC 11404:2007, “Information technology -- General-Purpose Datatypes (GPD),” *International Organization for Standardization*, <http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=39479>. [↑](#endnote-ref-20)
21. ISO 21090:2011, “Health informatics -- Harmonized data types for information interchange,” *International Organization for Standardization*, <http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=35646>. [↑](#endnote-ref-21)