NIST Special Publication 1500-1

DRAFT NIST Big Data Interoperability Framework:
Volume 1, Definitions

NIST Big Data Public Working Group
Definitions and Taxonomies Subgroup

Draft Version 1
April 6, 2015
http://dx.doi.org/10.6028/NIST.SP.1500-1

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

NIST Special Publication 1500-1
Information Technology Laboratory

DRAFT NIST Big Data Interoperability Framework: Volume 1, Definitions

Draft Version 1

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

April 2015

U. S. Department of Commerce
Penny Pritzker, Secretary

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

National Institute of Standards and Technology NIST Special Publication 1500-1
33 pages (April 6, 2015)

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

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

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at NIST promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in Federal information systems. This document reports on ITL's research, guidance, and outreach efforts in Information Technology and its collaborative activities with industry, government, and academic organizations.

Abstract

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

Keywords

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

Acknowledgements

This document reflects the contributions and discussions by the membership of the NBD-PWG, co-chaired by Wo Chang of the NIST ITL, Robert Marcus of ET-Strategies, and Chaitanya Baru, University of California San Diego Supercomputer Center.

The document contains input from members of the NBD-PWG Definitions and Taxonomies Subgroup, led by Nancy Grady (SAIC), Natasha Balac (SDSC), and Eugene Luster (R2AD).

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

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

Deborah Blackstock
MITRE Corporation

David Boyd
L3 Data Tactics

Pw Carey
Compliance Partners, LLC

Wo Chang
NIST

Yuri Demchenko
University of Amsterdam

Frank Farance
Consultant

Geoffrey Fox
University of Indiana

Ian Gorton
CMU

Nancy Grady
SAIC

Karen Guertler
Consultant

Keith Hare
JCC Consulting, Inc.

Christine Hawkinson
U.S. Bureau of Land Management

Thomas Huang
NASA

Philippe Journeau
ResearXis

Pavithra Kenjige
PK Technologies

Orit Levin
Microsoft

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

Ashok Malhotra
Oracle

Bill Mandrick
L3 Data Tactics

Robert Marcus
ET-Strategies

Lisa Martinez
Consultant

Gary Mazzaferro
AlloyCloud, Inc.

William Miller
MaCT USA

Sanjay Mishra
Verizon

Bob Natale
Mitre

Rod Peterson
U.S. Department of Veterans Affairs

Ann Racuya-Robbins
World Knowledge Bank

Russell Reinsch
Calibrum

John Rogers
HP

Arnab Roy
Fujitsu

Mark Underwood
Krypton Brothers LLC

William Vorhies
Predictive Modeling LLC

Tim Zimmerman
Consultant

Alicia Zuniga-Alvarado
Consultant

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

Notice to Readers

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

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

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

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

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

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

Table of Contents

FIGURE

TABLE

Executive Summary

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 a cluster of computing and storage resources to provide cost-effective data management. Data science combines various technologies, techniques, and theories from various fields, mostly related to computer science and statistics, to obtain actionable 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 stages of the NBD-PWG work. The three stages aim to achieve the following:

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

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

1 INTRODUCTION

1.1 BACKGROUND

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

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

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

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

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

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

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

Motivated by the White House initiative and public suggestions, the National Institute of Standards and Technology (NIST) has accepted the challenge to stimulate collaboration among industry professionals to further the secure and effective adoption of Big Data. As one result of NIST's Cloud and Big Data Forum held on January 15–17, 2013, there was strong encouragement for NIST to create a public working group for the development of a Big Data 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* consists of seven volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The seven volumes are as follows:

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

The *NIST Big Data Interoperability Framework* will be released in three versions, which correspond to the three stages of the NBD-PWG work. The three stages aim to achieve the following:

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

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

1.2 SCOPE AND OBJECTIVES OF THE DEFINITIONS AND TAXONOMIES SUBGROUP

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.

1.3 REPORT PRODUCTION

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.

1.4 REPORT STRUCTURE

This volume seeks to clarify the meanings of the broad terms Big Data and data science, which are discussed at length in Section 2. The more elemental concepts and terms that provide additional insights are discussed in Section 3. Section 4 explores several concepts that are more detailed. This first 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.

1.5 FUTURE WORK ON THIS VOLUME

This volume represents the beginning stage of the NBD-PWG's effort to provide order and clarity to an emerging and rapidly changing field. Big Data encompasses a large range of data types, fields of study, technologies, and techniques. Distilling from the varied viewpoints a consistent, core set of definitions to frame the discussion has been challenging. However, through discussion of the varied viewpoints a greater understanding of the Big Data paradigm will emerge. As the field matures, this document will also need to mature to accommodate innovations in the field. To ensure the concepts are accurate, future NBD-PWG tasks will consist of the following:

- Defining the different patterns of communications between Big Data resources to better clarify the different approaches being taken
- Updating Volume 1 taking into account the efforts of other working groups such as International Organization for Standardization (ISO) Joint Technical Committee 1 (JTC 1) and the Transaction Processing Performance Council.
- Improve the discussions of governance and data ownership
- Develop the Management section
- Develop the Security and Privacy section
- Add a discussion of the value of data

2 BIG DATA AND DATA SCIENCE DEFINITIONS

The rate of growth of data generated and stored has been increasing exponentially. In a 1965 paper[2], 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 considered 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 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 change 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 business considerations of cost effectiveness.

In the following sections, the two broad concepts, Big Data and data science, are broken down into specific individual terms and concepts.

2.1 BIG DATA DEFINITIONS

Big Data refers to the inability of traditional data architectures to efficiently handle the new datasets. Characteristics of Big Data that force new architectures are ***volume*** (i.e., the size of the dataset) and ***variety*** (i.e., data from multiple repositories, domains, or types), and the data in motion characteristics of ***velocity*** (i.e., rate of flow) and ***variability*** (i.e., the change in other characteristics). These characteristics—volume, variety, velocity, and variability—are known colloquially as the 'Vs' of Big Data and are further discussed in Section 3. Each of these characteristics influences the overall design of a Big Data system, resulting in different data system architectures or different data lifecycle 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. There are two fundamentally different methods for system scaling, often described metaphorically as "vertical" or "horizontal" scaling. ***Vertical scaling*** implies increasing the system parameters of processing speed, storage, and memory for greater performance. This approach is limited by physical capabilities whose improvements have been described by Moore's Law, requiring ever more sophisticated elements (e.g., hardware, software) that add time and expense to the implementation. The alternate method is to use ***horizontal scaling***, to make use of a cluster of individual (usually commodity) resources integrated to act as a single system. It is this horizontal scaling 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.*

This new paradigm leads to a number of conceptual definitions that suggest Big Data exists when the scale of the data causes the management of the data to be a significant driver in the design of the system architecture. This definition does not explicitly refer to the horizontal scaling in the Big Data paradigm.

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 in the simulation community, 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 of today is similar. 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 data retrieval processes across loosely coupled resources.

While the methods to achieve efficient scalability across resources will continually evolve, this paradigm shift (in analogy to the prior shift in the simulation community) is a one-time occurrence. Eventually, a new paradigm shift will likely occur beyond this distribution of a processing or data system that spans multiple resources working in parallel. That future revolution will need to be described with new terminology.

Big Data focuses on the self-referencing viewpoint that data is big because it requires scalable systems to handle it. Conversely, architectures with better scaling have come about because of the need to handle Big Data. It is difficult to delineate a size requirement for a dataset to be considered Big Data. Data is usually considered "big" if the use of new scalable architectures provides a cost or performance efficiency over the traditional vertically scaled architectures (i.e., if similar performance cannot be achieved in a traditional, single platform computing resource.) This circular relationship between the characteristics of

the data and the performance of data systems leads to different definitions for Big Data if only one aspect is considered.

Some definitions for Big Data focus on the systems innovations required because of the characteristics of Big Data.

> ***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 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 in the use of graph-oriented data storage.

The new non-relational model database paradigms are typically referred to as ***NoSQL*** (Not Only or No Structured Query Language [SQL]) systems, which are further discussed in Section 3. 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 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.*

Another related engineering technique is the federated database system, which is related to the variety characteristic of Big Data.

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

A federated database is thus a database system comprised of underlying database systems. Big Data systems can likewise pull a variety of data from many sources, but the underlying repositories do not all have to conform to the relational model.

Note that for systems and analysis processes, the Big Data paradigm shift also causes changes in the traditional data lifecycle processes. One description of the end-to-end data lifecycle categorizes the process steps as collection, preparation, analysis, and action. Different Big Data use cases can be characterized in terms of the dataset characteristics and in terms of the time window for the end-to-end data lifecycle. Dataset characteristics change the data lifecycle processes in different ways, for example in the point in the lifecycle at which the data is placed in persistent storage. In a traditional relational model, the data is stored after preparation (for example, after the extract-transform-load and cleansing processes). In a high velocity use case, the data is prepared and analyzed for alerting, and only then is the data (or aggregates of the data) given a persistent storage. In a volume use case, the data is often stored in the raw state in which it was produced—before being cleansed and organized (sometimes referred to as extract-load-transform). The consequence of persistence of data in its raw state is that a schema or model for the data is only applied when the data is retrieved for preparation and analysis. This Big Data concept is described as schema-on-read.

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

299 Another concept of Big Data is often referred to as *moving the processing to the data, not the data to the*
300 *processing.*

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

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

310 Additional concepts used in reference to the term Big Data refer to changes in analytics, which will be
311 discussed in Section 2.2. A number of other terms (particularly terms starting with the letter V) are also
312 used, several of which refer to the data science process or its benefit, instead of new Big Data
313 characteristics. Some of these additional terms include ***veracity*** (i.e., accuracy of the data), ***value*** (i.e.,
314 value of the analytics to the organization), ***volatility*** (i.e., tendency for data structures to change over
315 time), and ***validity*** (i.e., appropriateness of the data for its intended use). While these characteristics and
316 others—including quality control, metadata, and data provenance—long pre-dated Big Data, their impact
317 is still important in Big Data systems. Several of these terms are discussed with respect to Big Data
318 analytics in Section 3.4.

319 Essentially, Big Data refers to the extensibility of data repositories and data processing across resources
320 working in parallel, in the same way the compute-intensive simulation community embraced massively
321 parallel processing two decades ago. By working out methods for communication among resources, the
322 same scaling is now available to data-intensive applications.

323 2.2 DATA SCIENCE DEFINITIONS

324 In its purest form, data science is the *fourth paradigm* of science, following theory, experiment, and
325 computational science. The fourth paradigm is a term coined by Dr. Jim Gray in 2007. It refers to the
326 conduct of data analysis as an empirical science, learning directly from data itself. Data science as a
327 paradigm would refer to the formulation of a hypothesis, the collection of the data—new or pre-
328 existing—to address the hypothesis, and the analytical confirmation or denial of the hypothesis (or the
329 determination that additional information or study is needed.) In many data science projects, the raw data
330 is browsed first, which informs a hypothesis, which is then investigated. As in any experimental science,
331 the end result could be that the original hypothesis itself needs to be reformulated. The key concept is that
332 data science is an empirical science, performing the scientific process directly on the data. Note that the
333 hypothesis may be driven by a business need, or can be the restatement of a business need in terms of a
334 technical hypothesis.

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

337 Data science can be understood as the activities happening in the processing layer of the system
338 architecture, against data stored in the data layer, in order to extract knowledge from the raw data through
339 the complete data lifecycle.

*The **data lifecycle** is the set of processes that transform raw data into actionable knowledge.*

Traditionally, the term analytics has been used as one of the steps in the data lifecycle of collection, preparation, analysis, and action.

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

With the new Big Data paradigm, analytics are no longer separable from the data model and the distribution of that data across parallel resources. When structured data was almost exclusively stored as organized information in a relational model, the analytics could be designed for this structure. While the working definition of the data science paradigm refers to learning directly from data, in the Big Data paradigm this learning must implicitly involve all steps in the data lifecycle, with analytics being only a subset.

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

Data science across the entire data lifecycle now incorporates principles, techniques, and methods from many disciplines and domains, including the analytics domains of mathematics, data mining (specifically machine learning and pattern recognition), statistics, operations research, and visualization, along with the domains of systems, software, and network engineering. 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 through each stage in the data lifecycle.*

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.

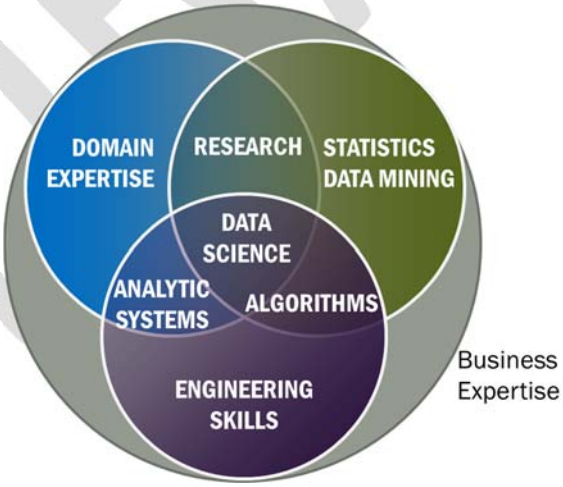

Figure 1: Skills Needed in Data Science

Data science is not solely concerned with analytics, but also with the end-to-end experimental lifecycle, where the data system is essentially the scientific equipment. 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 meaningfully address the hypothesis. These analytics concepts are discussed further in Section 3.4.

Data science is increasingly used to influence business decisions. In Big Data systems, identifying a correlation is often sufficient for a business to take action. As a simple example, if it can be determined that using the color blue on a website leads to greater sales than using green, then this correlation can be used to improve the business. The reason for the preference is not needed—it is enough to determine correlation.

Several issues are currently being debated within the data science community, two of which 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. The subset of data can be used as input for analytical processes, to determine methodology to be used for experimental procedures, or to address 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 re-purposed 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 only represents 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.

For descriptive purposes, analytics activities can be broken into different categories, including discovery, exploratory analysis, correlation analysis, predictive modeling, and machine learning. Again, these analytics categories are not specific to Big Data, but some have gained more visibility due to their greater application in 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.

2.3 OTHER BIG DATA DEFINITIONS

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 [3] [4] [5] [6]. 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 concepts and definitions in Table 1 are not comprehensive, but rather illustrate the inter-related concepts attributed to the catch-all term Big Data.

Table 1: Sampling of Concepts Attributed to Big Data

Concept	Author	Definition
4Vs (Volume, Variety, Velocity, and Variability) and Engineering	Gartner[7,8]	"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 [9]	"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)[10]	"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[9]	"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[9]	"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[9]	"…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[11] [16]	"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[9]	"Big data means data that cannot fit easily into a standard relational database."
	McKinsey[12]	"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[9]	"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[9]	"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[13]	"The broad range of new and massive data types that have appeared over the last decade or so."
	Mark van Rijmenam[9]	"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."

Concept	Author	Definition
Analytics	Ryan Swanstrom[9]	"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[9]	"Big data describes datasets that are so large, complex, or rapidly changing that they push the very limits of our analytical capability."
	Josh Ferguson[9]	"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[9]	"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[9]	"Big data refers to using complex datasets to drive focus, direction, and decision making within a company or organization."
	Hilary Mason[9]	"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[9]	"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[9]	"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[9]	"'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[9]	"'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."

422

423

3 BIG DATA FEATURES

The diversity of Big Data concepts discussed in Section 2 is similarly reflected in the discussion of Big Data features in Section 3. Some Big Data terms and concepts are discussed in Section 3 to understand new aspects brought about by the Big Data paradigm in the context of existing data architecture and analysis context.

3.1 DATA ELEMENTS AND METADATA

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[14], and, as an example, its extension into healthcare information data types in ISO 21090:2011 Health Informatics[15].

One important concept to Big Data is metadata, which is often described as "data about data." 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]) and discipline specific metadata (e.g., ISO 19115-x[17] for geospatial data).

Metadata that describes the history of a dataset is called its *provenance*, which is discussed in Section 3.6. 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 definitional 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 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 definitional descriptions, and the reader is referred to the World Wide Web Consortium (W3C) efforts on the semantic web[18][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.

3.2 DATA RECORDS AND NON-RELATIONAL MODELS

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 non-standard 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 the value of unstructured or relationship data, and also on different engineering methods that can handle data more efficiently.

Again, semantic metadata is, Big Data Engineering refers to the new ways 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 all need to have 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.

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 inter-relationship between data elements or across data records, independent of whether the dataset has the characteristics of Big Data

3.3 DATASET CHARACTERISTICS AND STORAGE

Data records are grouped into datasets, which can have the Big Data characteristics of volume, velocity, variety, and variability. Dataset characteristics can refer to the data itself, or *data at rest*, while characteristics of the data that is traversing a network or temporarily residing in computer memory to be read or updated is referred to as *data in motion*, which is discussed in Section 3.4.

Data at Rest: Typical characteristics of data at rest that are notably different in the era of Big Data are volume and variety. Volume is the characteristic of data at rest that is most associated with Big Data. Estimates show that the amount of data in the world doubles every two years.[20] Should this trend continue, by 2020 there would be 500 times the amount of data as existed in 2011. The sheer volume of the data is colossal. The data volumes have stimulated new ways for scalable storage across a collection of horizontally coupled resources, as described in Section 2.1.

The second characteristic of data at rest is the increasing need to use a variety of data, meaning the data represents a number of data domains and a number of data types. Traditionally, a variety of data was handled through transformations or pre-analytics to extract features that would allow integration with other data. The wider range of data formats, logical models, timescales, and semantics, which is desirous to use in analytics, complicates the integration of the variety of data. For example, data to be integrated could be text from social networks, image data, or a raw feed directly from a sensor source. To deal with a wider range of data formats, a federated database model was designed as a database across the underlying databases. Data to be integrated for analytics could now be of such volume that it cannot be moved to integrate, or it may be that some of the data is not under control of the organization creating the data system. In either case, the variety of Big Data forces a range of new Big Data engineering solutions to efficiently and automatically integrate data that is stored across multiple repositories, in multiple formats, and in multiple logical data models.

Big Data engineering has spawned data storage models that are more efficient for unstructured data than the traditional relational model, causing a derivative issue for the mechanisms to integrate this data. New scalable techniques have arisen to manage and manipulate Big Data not stored in traditional expensive high-performance "vertically" scaled systems, but rather spread across a number of less expensive resources. For example, the document store was developed specifically to support the idea of storing and indexing heterogeneous data in a common repository for analysis. New types of non-relational storage for data records are discussed below.

***Shared-disk File Systems*:** These 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). ***MapReduce*** is the default processing component in data-distributed computing. 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 MapReduce for processing places limitations on updating or iterative access to the data during computation. Bulk Synchronous Parallelism systems or newer MapReduce developments can be used when repeated updating is a requirement. Improvements and "generalizations" of MapReduce 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.

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 MapReduce) 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, NASA (U.S. National Aeronautics and Space Administration) uses this technology to make 3GB images available to the public. 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.

3.4 DATA IN MOTION

Another important characteristic of Big Data is the time window in which the analysis can take place. 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.

Typical characteristics of data in motion that are significantly different in the era of Big Data are velocity and variability. The velocity is the rate of flow at which the data is created, stored, analyzed, and visualized. Big Data velocity means a large quantity of data is being processed in a short amount of time. In the Big Data era, data is created and passed on in real time or near real time. Increasing data flow rates create new challenges to enable real- or near real-time data usage. Traditionally this concept has 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. However, the new in-parallel scaling approaches do add new Big Data engineering options for efficiently handling this data.

The second characteristic for data in motion is variability, which refers to any change in data over time, including the flow rate, the format, or the composition. Given that many data processes generate a surge in the amount of data arriving in a given amount of time, new techniques are needed to efficiently handle this 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.[21] [22] 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. This is especially true for organizations that already need extensive infrastructure but simply need to balance resources across application workloads that can vary.

3.5 DATA SCIENCE LIFECYCLE MODEL FOR BIG DATA

As was introduced in Section 2.1, the data lifecycle consists of the following four stages:

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. ***Action***: This stage involves processes that use the synthesized knowledge to generate value for the enterprise.

In the traditional data warehouse, the data handling process followed the order above (i.e., collection, preparation, storage, and analysis.) The relational model 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. Examples of these changes are as follows:

- ***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.

3.6 BIG DATA ANALYTICS

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.

In addition to volume, velocity, variety, and variability, several terms, many beginning with V, have been used in connection with Big Data requirements for the system architecture. Some of these terms strongly relate to analytics on the data. Veracity and provenance are two such terms and are discussed below.

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.

As discussed in Section 3.1, the provenance, or history of the data, is increasingly an essential factor in Big Data analytics, as more and more data is being repurposed for new types of analytics in completely different disciplines from which the data was created. As the usage of data persists far beyond the control of the data producers, it becomes ever more essential that metadata about the full creation and processing history is made available along with the data. In addition, it is vital to know what analytics may have produced the data, since there are always confidence ranges, error ranges, and precision/recall limits associated with analytic outputs.

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*.

3.7 BIG DATA METRICS AND BENCHMARKS

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 to 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. 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.

3.8 BIG DATA SECURITY AND PRIVACY

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 multi-tiered 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.
- **Veracity:** Complex challenges have been introduced in protecting data integrity as well as maintaining privacy policies as data moves across individual boundaries to groups, communities of interest, state, national, and international boundaries.
- **Volatility:** 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 HPCC. The addition of variety, volume, velocity, veracity, volatility, and value to the mix has amplified these concerns to the level of a national conversation, with unanticipated impacts on privacy frameworks.

3.9 DATA GOVERNANCE

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 lifecycle, 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.

4 BIG DATA ENGINEERING PATTERNS (FUNDAMENTAL CONCEPTS)

To define the differences between Big Data technologies, different 'scenarios' and 'patterns' are needed to illustrate relationships between Big Data characteristics (Section 2.1) and between the NBDRA components found in *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*. The scenarios would describe the high-level functional processes that can be used to categorize and, therefore, provide better understanding of the different use cases presented in *NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements*, as well as help to clarify the differences in specific implementations of components listed in the *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture.*

The topics surrounding the relaxation of the principles of a relational model in non-relational systems are very important. These topics are discussed in industry publications on concurrency, and will be addressed more fully in of future additions to this document.

Appendix A: Index of Terms

Appendix B: 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 lifecycle process.

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

A ***Latency*** 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 lifecycle.

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.

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

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

774 ***validity*** refers to appropriateness of the data for its intended use

775 ***value*** refers to the inherent wealth, economic and social, embedded in any data set

776 ***variability*** refers to the change in other data characteristics

777 ***variety*** refers to data from multiple repositories, domains, or types

778 ***velocity*** refers to the rate of data flow

779 ***veracity*** refers to the accuracy of the data

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

782 ***volatility*** refers to the tendency for data structures to change over time

783 ***volume*** refers to the size of the dataset

784

Appendix C: Acronyms

API	application program interface
BLOBs	binary large objects
GB	gigabyte
I/O	input/output
ISO	International Organization for Standardization
ITL	Information Technology Laboratory
JTC 1	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
NSF	National Science Foundation
OED	Oxford English Dictionary
P2P	peer-to-peer
SANs	Storage Area Networks
SQL	Structured Query Language
NoSQL	Not Only or No Structured Query Language
W3C	World Wide Web Consortium

Appendix D: References

DOCUMENT REFERENCES

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

[2] Gordon Moore, "Cramming More Components Onto Integrated Circuits," *Electronics*, Volume 38, Number 8 (1965), pages 114-117.

[3] Jenna Dutcher, "What is Big Data," *Data Science at Berkeley Blog*, September 3, 2014, http://datascience.berkeley.edu/what-is-big-data/.

[4] 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/.

[5] 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.

[6] 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/.

[7] 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.

[8] Gartner IT Glossary, "Big Data" (definition), *Gartner.com*, accessed November 17, 2014, http://www.gartner.com/it-glossary/big-data.

[9] Jenna Dutcher, "What is Big Data," *Data Science at Berkeley Blog*, September 3, 2014, http://datascience.berkeley.edu/what-is-big-data/.

[10] Oxford English Dictionary, "Big Data" (definition), *OED.com*, accessed November 17, 2014, http://www.oed.com/view/Entry/18833#eid301162178.

[11] 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.

[12] James Manyika et al., "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, May 2011.

[13] Tom Davenport, "Big Data@Work," Harvard Business Review Press, February 25, 2014.

[14] 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.

[15] 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.

[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.

[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.

[18] Phil Archer, "W3C Data Activity Building the Web of Data," *W3C*, http://www.w3.org/2013/data/.

[19] Dan Brickley and Ivan Herman, "Semantic Web Interest Group," *W3C*, June 16, 2012, http://www.w3.org/2001/sw/interest/.

[20] EMC2, "Digital Universe," *EMC*, accessed February 21, 2014, http://www.emc.com/leadership/programs/digital-universe.htm.

[21] 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.

[22] 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.