NIST Special Publication 1500-1r1

NIST Big Data Interoperability Framework: Volume 1, Definitions

NIST Big Data Public Working Group
Definitions and Taxonomies Subgroup

Version 2
June 2018

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-1r1

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

NIST Special Publication 1500-1r1

NIST Big Data Interoperability Framework: Volume 1, Definitions

Version 2

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

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-1r1

June 2018

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

National Institute of Standards and Technology (NIST) Special Publication 1500-1r1
Natl. Inst. Stand. Technol. Spec. Publ. 1500-1r1, 43 pages (June 2018) CODEN: NSPUE2

This publication is available free of charge from: https://doi.org/10.6028/NIST.SP.1500-1r1

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 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 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

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. The growth of data is outpacing scientific and technological advances in data analytics. Opportunities exist with Big Data to address the volume, velocity and variety of data through new scalable architectures. 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* (NBDIF) 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 Application Provider; Big Data Characteristics; Big Data Definitions; Big Data Framework Provider; Big Data Taxonomy; cloud; Data Consumer; Data Provider; data science; Internet of Things (IoT); machine learning; 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, cochaired by Wo Chang (NIST ITL), Bob Marcus (ET-Strategies), and Chaitan Baru (San Diego Supercomputer Center; National Science Foundation [NSF]). For all versions, the Subgroups were led by the following people: Nancy Grady (SAIC), Natasha Balac (SDSC), and Eugene Luster (R2AD) for the Definitions and Taxonomies Subgroup; Geoffrey Fox (Indiana University) and Tsegereda Beyene (Cisco Systems) for the Use Cases and Requirements Subgroup; Arnab Roy (Fujitsu), Mark Underwood (Krypton Brothers; Synchrony Financial), and Akhil Manchanda (GE) for the Security and Privacy Subgroup; David Boyd (InCadence Strategic Solutions), Orit Levin (Microsoft), Don Krapohl (Augmented Intelligence), and James Ketner (AT&T) for the Reference Architecture Subgroup; and Russell Reinsch (Center for Government Interoperability), David Boyd (InCadence Strategic Solutions), Carl Buffington (Vistronix), and Dan McClary (Oracle), for the Standards Roadmap Subgroup.

The editors for this document were the following:

- ***Version 1***: Nancy Grady (SAIC) and Wo Chang (NIST)
- ***Version 2***: Nancy Grady (SAIC) and Wo Chang (NIST)

Laurie Aldape (Energetics Incorporated) and Elizabeth Lennon (NIST) provided editorial assistance across all NBDIF volumes.

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 to this volume, during Version 1 and/or Version 2 activities, by the following NBD-PWG members:

Claire Austin
Environment Cananda

Deborah Blackstock
MITRE Corporation

David Boyd
InCadence Strategic Services

Cavan Capps
U.S. Census Bureau

Pw Carey
Compliance Partners, LLC

Wo Chang
NIST

Yuri Demchenko
University of Amsterdam

Frank Farance
Consultant

Geoffrey Fox
Indiana University

Ian Gorton
CMU

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

Bob Natale
Mitre Corporation

Quyen Nguyen
U.S. Census Bureau

Rod Peterson
U.S. Department of Veterans Affairs

Ann Racuya-Robbins
World Knowledge Bank

Russell Reinsch
Center for Government Interoperability

John Rogers
HP

Arnab Roy
Fujitsu

Jim Tyson
SAIC

Mark Underwood
Krypton Brothers; Synchrony Financial

Nancy Grady
SAIC

Karen Guertler
Consultant

Keith Hare
JCC Consulting, Inc.

Lisa Martinez
Consultant

Sanjay Mishra
Verizon

Gary Mazzaferro
AlloyCloud, Inc.

William Miller
MaCT USA

William Vorhies
Predictive Modeling LLC

Tim Zimmerlin
Automation Technologies Inc.

Alicia Zuniga-Alvarado
Consultant

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 (NBDIF): 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 characteristics of the data lead 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 NBDIF consists of nine volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The nine NBDIF volumes, which can be downloaded from https://bigdatawg.nist.gov/V2_output_docs.php, are as follows:

- Volume 1, Definitions (this volume)
- Volume 2, Taxonomies [1]
- Volume 3, Use Cases and General Requirements [2]
- Volume 4, Security and Privacy [3]
- Volume 5, Architectures White Paper Survey [4]
- Volume 6, Reference Architecture [5]
- Volume 7, Standards Roadmap [6]
- Volume 8, Reference Architecture Interfaces [7]
- Volume 9, Adoption and Modernization [8]

The *NBDIF* 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).

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

Potential areas of future work for the Definitions and Taxonomy 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.

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:

- How is Big Data defined?
- What attributes define Big Data solutions?
- What is new in Big Data?
- What is the difference between Big Data and *bigger data* that has been collected for years?
- 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, secure Big Data solutions?

Within this context, on March 29, 2012, the White House announced the Big Data Research and Development Initiative. [9] 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 analysts' 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 added value from Big Data service providers.

The *NIST Big Data Interoperability Framework* (NBDIF) 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).

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

On September 16, 2015, seven NBDIF Version 1 volumes 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 (this volume)
- Volume 2, Taxonomies [1]
- Volume 3, Use Cases and General Requirements [2]
- Volume 4, Security and Privacy [3]
- Volume 5, Architectures White Paper Survey [4]
- Volume 6, Reference Architecture [5]
- Volume 7, Standards Roadmap [6]

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 of the Stage 2 work, the following two additional NBDIF volumes have been developed.

- Volume 8, Reference Architecture Interfaces [7]
- Volume 9, Adoption and Modernization [8]

Version 2 of the NBDIF volumes, resulting from Stage 2 work, can be downloaded from the NBD-PWG website (https://bigdatawg.nist.gov/V2_output_docs.php). 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.

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 clarify concepts and 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. To achieve technical and high-quality document content, this document went through a public comments period along with NIST internal review. During the public comment period, the broader community was 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. Terms and definitions of concepts integral to Big Data are presented as a list in Section 2 and in the sections with relevant discussions. Big Data characteristics and Big Data engineering are discussed in Sections 3 and 4, respectively. Section 5 explores concepts of data science. Section 6 provides a summary of security and privacy concepts. Section 7 discusses Management concepts. This second version of *NBDIF: 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 *NBDIF*. *Volume 2, Taxonomies* provides a description of the more detailed components of the 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 designed 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. *Volume 8, Reference Architecture Interface* explores a set of interfaces, defined through example, which are used to create schema-based definitions of objects that can be manipulated through Big Data design patterns. *Volume 9, Adoption and Modernization* examines the adoption and barriers to adoption of Big Data systems, maturity of Big Data technology, and considerations for implementation of Big Data systems. Comparing related sections in these volumes will provide a more comprehensive understanding of the work of the NBD-PWG.

1.5 FUTURE WORK ON THIS VOLUME

A number of topics have not been discussed and clarified sufficiently to be included in Version 2. Topics that remain to be addressed in Version 3 of this document include better characterization of implementation differences in instances of NoSQL (Not Only or No Structured Query Language [SQL]) or NewSQL data platforms; guidance on system and data flow metrics; discussion of changes in algorithms and statistical measures; discussions of new programming languages; and expanding relationships to machine learning and virtual reality.

2 TERMS AND DEFINITIONS

The following definitions were created by the working group during the development of this document, and are collected here for convenience.

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.

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.

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

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

Data locality refers to the data processing occurring at the location of the data storage.

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

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.

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).

Extensibility is the ability to add or modify existing tools for new domains.

Fabric represents the presence of activities and components throughout a computing system.

Fault-tolerance is the ability of a system to continue to run when a component of the system fails.

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 ***Horizontal Data Scientist*** is a generalist who understands enough of the multiple disciplines of mission need, the domain processes that produced the data, math and statistics, and computer science (or software and systems engineering).

Horizontal scaling is increasing the performance of distributed data processing through the addition of nodes in the cluster.

Interoperability is the ability for tools to work together.

Latency refers to the delay in processing or in availability.

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

Metadata is data about data or data elements, possibly including their data descriptions, data about data ownership, access paths, access rights, and data volatility.

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.

Reusability is the ability to apply tools from one domain to another.

Schema 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.

Small Data refers to the limits in the size of datasets that analysts can fully evaluate and understand.

SQL is the SQL query language standard used to query relational databases.

Structured data is data that has a predefined data model or is organized in a predefined way.

Unstructured data is data that does not have a predefined data model or is not organized in a predefined way.

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 changes in dataset, whether data flow rate, format/structure, semantics, and/or quality that impact the analytics application.

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.

A ***Vertical Data Scientist*** is a subject matter expert in specific disciplines involved in the overall data science process.

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

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

Volume refers to the size of the dataset.

3 BIG DATA CHARACTERISTICS

The rate of growth of data generated and stored has been increasing exponentially. In a 1965 paper, [10] 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, prompting the desire to incorporate this unstructured data to generate additional value. The central benefit of Big Data analytics is the ability to process large amounts and various types of information. The need for greater performance or efficiency happens on a continual basis. However, Big Data represents a fundamental shift to parallel scalability 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 to parallel architectures, 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 mainstream application to scaled data systems.

The terms Big Data and data science have 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 architectures of the systems that handle the data, the analysis of the datasets, and the considerations of cost-effectiveness.

In the following sections, the three broad concepts, Big Data (remainder of Section 3), Big Data engineering (Section 4), and data science (Section 5), are broken down into specific individual terms and concepts.

3.1 BIG DATA DEFINITIONS

Big Data refers to the need to parallelize the data handling in data-intensive applications. The characteristics of Big Data that force new architectures are as follows:

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

These characteristics—volume, velocity, variety, and variability—are known colloquially as the Vs of Big Data and are further discussed in Section 3.2. While many other Vs have been attributed to Big Data, only the above four characteristics 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, velocity, variety, 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 want to process the data in the same amount of time using twice the resources. 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 been 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 were 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.*

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 4.1. It is this horizontal scaling, which we will refer to as data-intensive parallel processing, that is at the heart of the Big Data revolution. Ideally, this parallelization achieves linear scalability in that twice the data can be processed in the same amount of time using twice the number of data nodes.

While Big Data strictly speaking should apply only 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—forcing parallel processing. The selection of the 4 Vs of volume, velocity, variety, and variability as fundamental for their importance in selecting processing architectures will be discussed in Section 3.2.

> *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 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 1990s for large-scale science and engineering simulations known as high performance computing (HPC), which will be discussed further in Section 4.3.1. The simulation community went through the same shift then in techniques for massively parallel, compute-intensive processing—analogous to the more recent shift in data-intensive processing techniques for handing Big Data.

3.2 BIG DATA CHARACTERISTICS

Defining universal criteria for determining that Big Data characteristics require a scalable solution is difficult, since the choice to use parallel Big Data architectures 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. The four fundamental drivers that determine if a Big Data problem exists are volume, velocity, variety, and variability—the Big Data characteristics listed in Section 3.1.

3.2.1 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 here 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 that could be analyzed went from thousands into millions through the use of scalable techniques. The time and expense required to process massive datasets was one of the original drivers for distributed processing. Volume drives the need for processing and storage parallelism, and its management during processing of large datasets.

3.2.2 Velocity

Velocity is a measure of the rate of data flow. Traditionally, high-velocity systems have been described as ***streaming data***. While these aspects are new for some industries, other industries (e.g., telecommunications and credit card transactions) have processed high volumes of data in short time intervals for years. Data in motion is processed and analyzed in real time, or near real time, and must 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 create the need for distributed processing even when the datasets are relatively small—a scenario often present in the Internet of Things (IoT).

3.2.3 Variety

The variety characteristic represents the need to analyze data from multiple repositories, domains, or types. The variety of data from multiple domains was previously handled through the identification of features that would allow alignment of datasets, and their fusion into a data warehouse. Automated data fusion relies on semantic metadata, where the understanding of the data through the metadata allows it to be integrated. A range of data types, domains, 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. Note that while volume and velocity allow faster and more cost-effective analytics, it is the variety of data that allows analytic results that were never possible before. “Business benefits are frequently higher when addressing the variety of data than when addressing volume.” [11]

3.2.4 Variability

Variability is a slightly different characteristic than volume, velocity, and variety, in that it refers to a change in a dataset rather than the dataset or its flow directly. Variability refers to changes in a dataset, whether in the data flow rate, format/structure, and/or volume, that impacts its processing. Impacts can include the need to refactor architectures, interfaces, processing/algorithms, integration/fusion, or storage. 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 advantageous capabilities of cloud computing. Detailed discussions of the techniques used to process data can be found in other industry publications that focus on operational cloud or virtualized architectures. [12], [13] Dynamic scaling keeps systems efficient, rather than having to design and resource to the expected peak capacity (where the system at most times sits idle). It should be noted that this variability refers to changes in dataset characteristics, whereas the term volatility (Section 5.4.3) 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.

3.2.5 STRUCTURED AND UNSTRUCTURED DATA TYPES

The data types for 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 International Organization for Standardization (ISO) standard ISO/ International Electrotechnical Commission (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]

The type of data does, however, relate to the choice of storage platform (see *NBDIF: Volume 6, Reference Architecture* of this series). 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 datasets, such as text, image, or video, do not have a predefined data model or are not organized in a predefined way. Unstructured datasets have been increasing in both volume and prominence (e.g., with the generation of web pages and videos.) While modern relational databases tend to have support for these types of data elements through text or binary objects, their ability to directly analyze, index, and process them has tended to be both limited and accessed via nonstandard extensions to the SQL query language. Semi-structured data refers to datasets in formats such as XML (eXtensible Markup Language) or JavaScript Object Notation (JSON) that has an overarching structure, but with individual elements that are unstructured. 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 on different engineering methods that can handle these types of datasets more efficiently.

3.3 BIG DATA SCIENCE AND ENGINEERING — A FIELD OF STUDY

Big Data has been described conceptually using the “V” terms in common usage. These common terms are conceptual, referring to the different qualitative characteristics of datasets. They help to understand, in general, when the handling of the data benefits from the use of the new scalable technologies.

Big Data Science and Engineering concerns a new *field of study*, just as chemistry, biology, statistics, and terminology are fields of study. This interdisciplinary field of study deals with the convergence of problems in the subfields of new data structures, parallelism, metadata, flow rate, and visual communication. Note that the “V” word for each subfield is provided to help align the field of study with the concepts discussed previously in Section 3.2.

3.3.1 DATA STRUCTURES

The field of Big Data is concerned with irregular or heterogeneous data structures, their navigation, query, and data typing—relating to the variety concept. Note that Big Data is not necessarily about a large amount of data because many of the above concerns can be demonstrated with small (less than one gigabyte [GB]) datasets. Big Data concerns typically arise in processing large amounts of data because some or all of the four main characteristics (i.e., irregularity, parallelism, real-time metadata, presentation/ visualization) are unavoidable in such large datasets. *Irregularity* means that the structure and shape are not known or agreed-upon in advance. The data structure can be described as *structured* (both datatyping and navigation are known in advance), *semi-structured* (either datatyping or navigation are known, but not both), or *unstructured* (neither datatyping nor navigation are known). To know the datatyping means to know the datatype as defined in ISO/IEC 11404 General Purpose Datatypes. To know the *navigation* means to know the path to complete the reference to the data. For example, a uniform resource identifier (URI) is a kind of navigation, as is the following:

```
employee[17].home_address.address_line[2]
```

The *shape* refers to the dimensions on a multidimensional array (e.g., 4-by-5-by-6 array is a three-dimensional array, its rank is 3). A scalar is zero-dimensional array (i.e., its rank is 0, its dimensions are the empty vector {}). NoSQL techniques are used for data storage of such irregular data. JSON and XML are examples of serializing such irregular data.

3.3.2 PARALLELISM

A Big Data subfield is new engineering for computational and storage parallelism and its management during processing of large datasets (i.e., volume). Computational parallelism issues concern the unit of processing (e.g., thread, statement, block, process, and node), contention methods for shared access, and begin-suspend-resume-completion-termination processing. *Parallelism* involves the distribution of a task into multiple subtasks with associated partition of resources, the coordination and scheduling of the subtasks including access/contention of the related resources, and the consolidation of the subtask results. A *task* is conceived in its broadest sense—a computer, a process, a subroutine, a block of code, and a single programming language statement are all examples of a task in this sense. The MapReduce technique is an example of parallelism, but certainly not the only method in common use.

3.3.3 METADATA

Big Data Science requires descriptive data and self-inquiry about objects for real-time decision making—validity and veracity. Descriptive data is metadata, and self-inquiry is known as reflection or introspection in some programming paradigms. *Metadata* is descriptive data (e.g., describing an interface in JSON, ISO/IEC 11179 metadata describing data structures, Dublin Core describing resources). *Real-time decision making* is necessary because of the *irregularity* of data. For example, rather than processing each element of data in the same way, the irregularity means that some real-time decisions are made concerning how the data is processed. One illustration could be as a medical record is searched for particular X-ray images in a series, subfolders must be reviewed to determine which ones are relevant to the query/question/study in progress. Real-time changes in processing may be necessary based upon the datatype (e.g., numeric versus string processing), the way in which the data element was observed, the quality of data (e.g., accumulating quantities of differing precision/accuracy), or other data irregularities.

3.3.4 FLOW RATE

The Big Data field is also concerned with the rate of arrival of the data—velocity. Streaming data is the traditional description for high-velocity data. This has long been a specialization in such domains as telecommunications or in the credit industry. In both domains, incidents of fraud needed to be detected in near real time to take a mitigating action as quickly as possible.

3.3.5 VISUAL COMMUNICATION

Additional challenges in Big Data projects include the presentation and aggregation of such datasets—visualization. The visual limitations consist of the amount of information a human can usefully process on a single display screen or sheet of paper. For example, the presentation of a connection graph of 500 nodes might require more than 20 rows and columns, along with the connections (i.e., relationships) among each of the pairs. Typically, this is too much for a human to comprehend in a useful way. Big Data presentation/visualization issues concern reformulating the information in a way that can be presented for convenient human consumption.

3.4 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. The term has been used to describe a number of topic areas including the following:

- Dataset characteristics
- More data than before
- Unstructured data
- The new scalable data processing techniques
- The analysis of extensive datasets
- The generation of value
- The loss of privacy
- The impact on society

Several Big Data concepts were observed in a sample of definitions taken from blog posts and magazine articles as shown in Table 1. The sample of formal and informal definitions offer a sense of the spectrum of concepts applied to the term Big Data. The NBD-PWG's definition is closest to the Gartner definition, with additional emphasis that the innovation is the horizontal scaling 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 [16], [17]	"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 [18]	"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 [19]	"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 [18]	"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 [18]	"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 [18]**Error! Bookmark not defined.**	"…our original definition was a system that (1) was capable of storing 10 TB [terabyte] 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 [20]**Error! Bookmark not defined.**	"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 [18]**Error! Bookmark not defined.**	"Big data means data that cannot fit easily into a standard relational database."

Concept	Author	Definition
	McKinsey[21]Error! Bookmark not defined.	“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 [18]	“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 [18]	“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 [22]	“The broad range of new and massive data types that have appeared over the last decade or so.”
	Mark van Rijmenam [18]	“Big data is not all about volume, it is more about combining different datasets 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 [18]	“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 [18]	“Big data describes datasets that are so large, complex, or rapidly changing that they push the very limits of our analytical capability.”
	Josh Ferguson [18]	“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 [18]	“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 [18]	“Big data refers to using complex datasets to drive focus, direction, and decision making within a company or organization.”
	Hilary Mason [18]	“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 [18]	“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 [18]	“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 [18]	“‘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 [18]	“‘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.”

4 BIG DATA ENGINEERING (FRAMEWORKS)

Some definitions for Big Data, as described in Section 3.3, 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.*

Big Data engineering is used when the volume, velocity, variety, or variability characteristics of the data require scalability for processing efficiency or cost-effectiveness. New engineering techniques in the storage layer have been driven by the growing prominence of datasets that cannot be handled efficiently in a traditional relational model (e.g., unstructured text and video.) The need for scalable access in structured data has led, for example, 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.

4.1 HORIZONTAL INFRASTRUCTURE SCALING

Section 3.1 defines horizontal scaling as increasing the performance of distributed data processing through the addition of nodes in the cluster. Horizontal scaling can occur at a number of points in the infrastructure.

4.1.1 SHARED-DISK FILE SYSTEMS

Approaches, such as SANs and 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.*

4.1.2 DISTRIBUTED FILE SYSTEMS

In distributed file storage systems, multi-structured 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 on each node when distributing the processing, which avoids the 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.

4.1.3 DISTRIBUTED DATA PROCESSING

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* was an early processing technique in data-intensive distributed computing where the query was “scattered” across the processors and the results were “gathered” into a central compute node.

***MapReduce** is the technique to allow queries to run across distributed data nodes.*

In relation to the compute-intensive simulations in HPC (discussed further in Section 4.3.1), scatter-gather is an embarrassingly parallel technique—meaning the same query and analysis code is sent to each data node to process the portion of the dataset on that node. For embarrassingly parallel techniques, there is no processing that relies on the data in more than one node.

The use of inexpensive servers is appropriate for slower, batch-speed Big Data applications, but does 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—other methods should 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.

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

***Data locality** refers to the data processing occurring at the location of the data storage.*

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 help to better understand the performance of Big Data systems, such as the *interoperability* (ability for tools to work together), *reusability* (ability to apply tools from one domain to another), and *extensibility* (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 *NBDIF: Volume 6, Reference Architecture* of this series.

4.1.4 RESOURCE NEGOTIATION

Several technologies have been developed for the management of a distributed computing system 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, which is allocated based on available central processing unit (CPU) and memory, and then are provided the appropriate processing resource in the node.

4.1.5 DATA MOVEMENT

Data movement is normally handled by transfer and application programming interface (API) technologies rather 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 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.

4.1.6 CONCURRENCY

Concurrency in data storage has traditionally referred to the ability to have multiple users making changes in a database, while restricting the ability for multiple users to change the same record. In other words, this means dealing with many things at the same time. This can be confused with parallelism, which is doing something on separate resources at the same time. An example may help to clarify the concept of concurrency. Non-relational systems are built for *fault-tolerance*, meaning it is assumed that some system components will fail. For this reason, data is not only distributed across a set of nodes, known as master nodes, but is also replicated to additional nodes known as slave nodes. If a master data node fails, the system can automatically switch over and begin to use one of the slave nodes. If the data on a master node is being updated, it will take time before it is updated on a slave node. Thus, a program running against the data nodes has the possibility of obtaining inconsistent results if one portion of the application runs against a master node, and another portion runs against a slave node that had not yet been updated. This possibility of obtaining different results based on the timing of events is an additional complication for Big Data systems.

4.1.7 TIERS OF STORAGE

Data storage can make use of multiple tiers of storage (e.g., in-memory, cache, solid state drive, hard drive, network drive) to optimize between performance requirements and cost. 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. Software-defined storage is among several emerging techniques for dynamic control of hardware (such as software-defined networks), and a detailed discussion is beyond the scope of this document.

4.1.8 CLUSTER MANAGEMENT

There are additional aspects of Big Data that are changing rapidly and are beyond the scope of this document, including the techniques for managing a cluster and mechanisms for providing communication among the cluster resources holding the data.

4.2 SCALABLE LOGICAL DATA PLATFORMS

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 MPP 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.

4.2.1 RELATIONAL PLATFORMS

The data management innovation in the 1960s and 1970s was the development of relational databases that followed a relational algebra (an algebra operating on sets) for data storage and query. The additional control that these platforms provided, helped drive innovations in the management of data in business systems. Relational Database Management Systems (RDBMS) have been at the heart of analytic systems ever since.

The properties of RDBMS are beyond the scope of this volume. The reader is referred to the extensive literature to learn about concepts such as Atomicity, Consistency, Isolation, Durability (ACID); Basic Availability, Soft-State, and Eventual Consistency (BASE); and the Consistency, Availability, and Partition (CAP) Tolerance theorem.

4.2.2 NON-RELATIONAL PLATFORMS (NOSQL)

The new non-relational model databases are typically referred to as ***NoSQL*** (Not Only or No 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 extensions 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.*

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. A variant on this is the *document store*, where the document in the value field can be indexed and searched, not just the key. Another type of new Big Data record storage is in a *graphical database*. 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 essentially represent each data element as a series of subject, predicate, and object triples. Often, the available types of objects and relationships are described via controlled vocabularies or ontologies.

4.2.3 NON-RELATIONAL PLATFORMS (NEWSQL)

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 including maintaining the ACID guarantees of a traditional RDBMS. Three categories of NewSQL databases that have been described are those having shared-nothing data nodes, those that are optimized as SQL storage engines, and those that leverage a sharding middleware. *Sharding* is the splitting of a dataset or database across nodes.

4.3 RELATIONSHIP TO OTHER TECHNOLOGICAL INNOVATIONS

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

4.3.1 HIGH PERFORMANCE COMPUTING

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 HPC communities, when scientific simulations began using 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 could greatly extend their simulation capabilities. This, of course, introduced a number of complexities 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. Some simulations may be *embarrassingly parallel* in that the computations do not require information from other compute nodes. When inter-node communication is required, libraries such as openMPI[a] have enabled the communications between compute processors. Simplistically, there are two basic types of HPC architectures. The first is classified as *shared nothing* (i.e., the data needed resides on each compute node), which is composed of all compute nodes. The second type of HPC architecture is *shared everything* where the data that each compute processor needs resides on one data node. These two scenarios are illustrated in Figure 1[b] where HPC systems either have one data node or none.

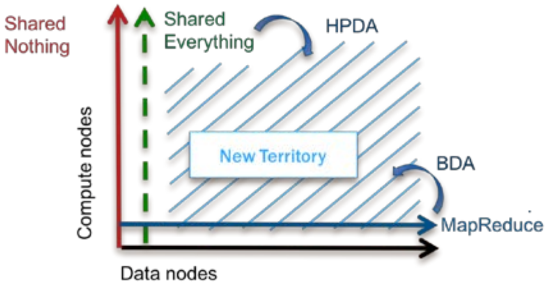

Figure 1: Compute and Data-Intensive Architectures

The Big Data paradigm, likewise, is the shift to parallelism for data-intensive applications, with implementations discussed in Section 4.1 with respect to infrastructure and in Section 4.2 with respect to platform considerations. To get the needed level of scaling, different mechanisms distribute data across nodes, and then process the data across those nodes. Recall that MapReduce is an *embarrassingly parallel* processing of data distributed across nodes, since the same query/analytics request is sent in parallel to each data node. The results are gathered back on the compute node, so MapReduce represents the use of one compute node and any number of data nodes, as shown in Figure 1.

The obvious next focus area will be in systems that are a blend of compute-intensive and data-intensive processing, represented by the hatched area in Figure 1. New work to add more data-intensive capabilities to HPC systems is known as high performance data analytics (HPDA). Some data-intensive applications are not as embarrassingly parallel as MapReduce. Machine learning techniques, such as deep learning, not only deal in large input datasets, but also require graphic processing units (GPUs) to get the compute scalability needed for the models to learn. The addition of parallel computation methods to Big Data applications is known as Big Data analytics. It is currently a research area, exploring optimal ways to blend nodes for both compute- and data-intensive applications.

4.3.2 CLOUD COMPUTING

The NIST Cloud Computing Standards Roadmap Working Group developed the following definition [23] for Cloud Computing:

> ***Cloud computing*** *is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model promotes availability and is composed of five essential characteristics, three service models, and four deployment models.* [24]

Cloud computing (usually referred to simply as cloud) typically refers to external providers of virtualized systems known as *public cloud*. Similarly, virtualized environments can also be deployed on premise—known as *private clouds*. Management services now exist to blend the two environments seamlessly into one environment—understandably referred to as *hybrid clouds*. Cloud deals primarily with the infrastructure, or infrastructure-as-a-service (IaaS). Other offerings include platform-as-a-service (PaaS) and software-as-a-service (SaaS). These three

[a] https://www.open-mpi.org/

[b] Figure 1 provided courtesy of Nancy Grady.

offerings align with the NBDRA's framework providers. NIST has published cloud-related documents in the Special Publications series for Cloud Reference Architecture [25], Security Reference Architecture [26], Technology Roadmap Volumes 1-3 [13], [27], [28], and Standards Roadmap [24], and is currently working on metrics descriptions.

While Big Data systems are commonly built on virtualized environments, many implementations by Internet search providers, and others, are frequently deployed on *bare metal* (deploying directly on the physical hardware) to achieve the best efficiency at distributing I/O across the clusters and multiple storage devices. Often cloud infrastructures are used to test and prototype Big Data applications. A system with high variability will typically 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. The ability to release those resources when they are no longer needed provides significant operational cost savings for this type of Big Data system.

The term *cloud* is often used to reference an entire implementation. Strictly speaking, cloud refers to the infrastructure framework (e.g., compute resources and network as described in the NBDRA). Big Data is primarily concerned with the platform and processing layers. Given that either virtualized or bare metal resources can be used, Big Data is a separate domain from cloud computing.

4.3.3 INTERNET OF THINGS / CYBER-PHYSICAL SYSTEMS

Cyber-physical systems (CPS) are smart systems that include engineered interacting networks of physical and computational components. CPS and related systems (including the IoT and the Industrial Internet) are widely recognized as having great potential to enable innovative applications and impact multiple economic sectors in the worldwide economy [29]. CPS focus areas include smart cities, energy smart grid, transportation, and manufacturing. The IoT refers to the connection of physical devices to the Internet—producing data and communicating with other devices. Big Data systems are often characterized by high volume—representing one or a few very large datasets. IoT is the other extreme representing a very large number of potentially very small streams of data. A subset of IoT may also represent high-velocity data. The overarching view of IoT consists of tiered systems where edge devices produce data—with a trade-off in what analytics are done on-board the edge devices versus what data is transmitted to local systems. End-to-end IoT represents a tiered system, different from the tiered storage discussed in Section 4.1.7. In this context, the IoT tiers represent coupled independent and distributed systems, generally described as a sensor tier, edge computing tier, regional tier, and global tier. For example, a smart watch may produce data and communicate with a smart phone, which then can connect to a local laptop and/or a cloud data collector. IoT thus represents a system-of-systems architecture, with the correspondingly greater optimization and security complexity.

4.3.4 BLOCKCHAIN

Blockchain technology was first proposed as a solution to the well-known problem of achieving consistency in distributed databases when the reliability of participating nodes could not be guaranteed. This solution provided a foundation for the bitcoin digital currency and payment system. The phrase *distributed ledger technology* (DLT) is often used as a synonym for blockchain, as the blockchain may be viewed as a shared, distributed, write-once ledger of verified transactions.

The term *blockchain* has come to refer to a type of distributed database that has certain characteristics, as follows:

- Data is fully replicated across participating network nodes;
- Data can be added by any participating node, by submitting a transaction to the network;
- Transactions are time-stamped and cryptographically signed using asymmetric or public-key cryptography to provide proof-of-origin;
- Transactions are verified by each participating node, and assembled into blocks;
- Blocks are assembled by participating nodes using cryptographic hash functions to build a hash tree of the transactions within the block, and then to generate a secure hash of the entire block;

- Blocks are validated by a distributed consensus mechanism, and valid blocks are added to the chain; and
- Each new block includes a link to the previous block—the data blocks so form a linked list or chain, hence the term *blockchain*.

Blockchain transactions are write-once; they are appended to blocks, and never subsequently updated or deleted. If the underlying data recorded as part of a transaction does change, a new transaction is written. Therefore, the full history of any information written to the blockchain can be traced by examining each related transaction and its associated timestamp.

The use of secure hashes in assembling blocks is intended, in part, to make it computationally infeasible for a participating node to alter the content of any transaction once it has been recorded to the blockchain. As a result, blockchains are typically characterized as immutable. The use of hash trees for transactions within blocks also reduces the time required to verify that a given transaction is in fact recorded within a block.

Blockchain, or DLT, has been proposed as a *disruptive* solution in a wide range of use cases beyond supporting digital currencies—including finance, identity management, supply chain management, healthcare, public records management, and more. Blockchain has also been widely proposed as an efficient means to securely record and manage the provenance of research data; to simplify data acquisition and sharing; to improve data quality and data governance; and even to facilitate the *monetization* of data at any scale that may be of interest to researchers.

More information can be found in the draft version of *NIST Interagency Report (IR) 8202, Blockchain Technology Overview*. [29]

4.3.5 NEW PROGRAMMING LANGUAGES

The need for distributed data and component management has led to the development of a number of new open source languages. New software assists in the provisioning of resources, addition of resources to the cluster, simpler functions for data movement and transformation, and new analytic frameworks to isolate the analytics from the underlying storage platform. Many Big Data databases have developed SQL-like query languages for data retrieval, or extensions to the standard SQL itself. In addition, existing analytics programming languages have all created modules to leverage the new distributed file systems, or data stored in the new distributed databases.

5 DATA SCIENCE

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. [30] 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 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 is tightly linked to the analysis of Big Data, and refers to the management and execution of the end-to-end data processes, including the behaviors of the components of the data system. As such, data science includes all of analytics, but analytics does not include all of data science. As discussed, data science contains different approaches to leveraging data to solve mission needs. While the term *data science* can be understood as the activities in any analytics pipeline that produces knowledge from data, the term is typically used in the context of Big Data.

5.1 DATA SCIENCE, STATISTICS, AND DATA MINING

The term *data science* has become a catch-all term, similar to the catch-all usage for the term Big Data. The definition given above could in fact be applied to any end-to-end data analytics pipeline.

The original data analytics field was dominated by statistics. In this discipline, the design of experiments determined the precise input data that was necessary and sufficient to definitively address a hypothesis. This field of analytics remains critical for providing verifiable results (e.g., for the analysis of clinical trials for the drug approval process in the pharmaceutical industry). Data sampling, a central concept of statistical analysis, 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., for a medical analysis to determine whether the treatment should prove effective). Care is taken to cleanse the data, and ensure the input data sample contains no external bias that would skew the results. The discovery process, or browsing data for something interesting, has been described pejoratively as data dredging, which does not result in definitive answers to a hypothesis.

In the late 1990s, a new analytics specialization emerged, known as data mining or knowledge discovery in databases. It became apparent that large datasets that had been collected could potentially add insights in areas different to the purpose for which the data was collected. This analysis of repurposed data still required careful sampling to remove bias and data cleansing to remove errors, but the machine learning or data mining techniques could generate approximate models to a wider range of data problems. The critical step in data mining is to ensure that the models have not been over-fitted (i.e., the analytical pattern matched the training data sample but did not provide accurate results on a separate testing data sample). In addition to the math and statistics skills, data mining required knowledge of the domain to ensure the repurposed data was being properly used. This is represented in the two upper circles of the Venn diagram in Figure 2[c]. The statistically meaningful results from modeling the data provided approximate (but not definitive) answers to a hypothesis. data mining is considered by some as a generalization of statistics, in that the class of problems that can be addressed is broader than those accessible by traditional statistics. data mining required not only math and statistics skills, but also required domain understanding—understanding how the data was produced, and how it should be appropriately used. While involved in automated systems, the initial focus for data mining encompassed a single analyst addressing a specific mission problem, selecting data internal to an organization, processing the data on their own local system, and delivering the results through presentations to mission leaders.

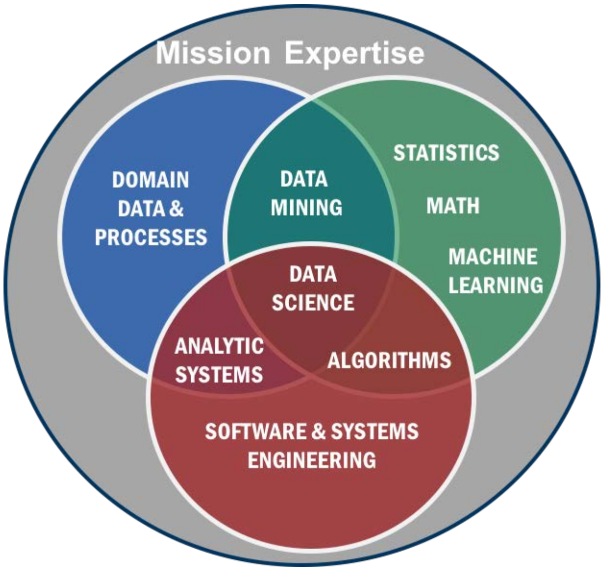

Figure 2: Data Science Sub-disciplines

Data science became a commonly used term in the mid-2000s as new techniques for handling Big Data began to emerge. The term *data science* was originally applied in the context of Big Data systems that were processing very large datasets—where the size of the data become a problem of its own. This additional complexity required the addition of computer science skills, as shown in Figure 2, to understand how to deploy large volume datasets across multiple data nodes, and how to alter query and analytics techniques to address the distributed data.

Data science is thus a super-set of the fields of statistics and of data mining and machine learning to include the analysis of Big Data. Data science often relaxes some of the concerns in data mining.

- Both statistical and data mining analysis required a careful sampling of data to ensure that the complete data population was being properly represented. In data science, typically all the data is processed and analyzed through scaling.
- In some problems, it is assumed that in the sheer volume of data, small errors will tend to cancel out, thereby reducing or eliminating the need to cleanse the data.
- Given the large datasets, sometimes the simplest algorithms can yield acceptable results. This has led to the debate in some circumstances whether *more data is superior to better algorithms*.
- Many hypotheses can be difficult to analyze, so data science also focuses on determining a surrogate question that does not address the original hypothesis, but whose analytical result can be applied to the original mission concern.
- The richness of data sources has increased the need to explore data to determine what might be of interest. As opposed to the data dredging of statistics or data mining, broader understanding of the data leads to either discovery of insights, or the formulation of hypotheses for testing.

Several issues are currently being debated within the data science community. Two prominent issues are data sampling, and the idea that more data is superior to better algorithms. 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 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 representative of the true population of interest. For example, studying social media data to analyze people's

[c] A reimagining provided by Nancy Grady of the Venn diagram description originally published by Drew Conway, http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, accessed June 7, 2017.

behaviors does not represent all people, as not everyone uses social media. 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—where you cannot expect accurate results based on inaccurate data.

5.2 DATA SCIENTISTS

Data scientists and data science teams solve complex data problems by employing deep expertise in one or more of the disciplines of math, statistics, and computer engineering, in the context of mission strategy, and under the guidance of domain knowledge. Personal skills in communication, presentation, and inquisitiveness are also very important given the need to express 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 2, are covered by different members of a team. 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 enough understanding of all areas to deliver value from data-intensive applications and work in a team whose skills spans these areas.

Similar to the term *Big Data*, data science and data scientist have also come to be used in multiple ways. Data scientist is applied essentially to anyone who performs any activity that touches data. To provide some specificity, the term *data scientist* should refer to the generalist that understands not only the mission needs, but the end-to-end solution to meet those needs. This can be described conceptually as a horizontal data scientist.

> *A **horizontal data scientist** is a generalist who understands enough of the multiple disciplines of mission need, the domain processes that produced the data, math and statistics, and computer science (or software and systems engineering).*

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 and analysis of a real-world process. The implication is that the horizontal 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, the analytic methods 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.

Conversely, those who specialize in a particular technique in the overall data analytics system should be more narrowly referred to as a subject matter expert.

> *A **vertical data scientist** is a subject matter expert in specific disciplines involved in the overall data science process.*

Examples of vertical data scientists could include the following: machine learning experts; Big Data engineers who understand distributed data storage platforms; data modelers who understand methods of organizing data to be appropriate for analytics; data engineers who transform data to prepare for analytics; or semantic specialists or ontologists who deal in metadata models. There is no consensus in the use of the term *data scientist*, but the specific use of the term to represent the generalist overseeing the end-to-end process and the use of a particular subject matter expert category for specialists would greatly improve the understanding of the skills needed to solve Big Data problems.

5.3 DATA SCIENCE PROCESS

Data science is focused on the end-to-end data processing life cycle to encompass Big Data.

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

The end-to-end data science life cycle consists of five fundamental steps:

1. ***Capture***: gathering and storing data, typically in its original form (i.e., raw data);
2. ***Preparation***: processes that convert raw data into cleansed, organized information;
3. ***Analysis***: techniques that produce synthesized knowledge from organized information;
4. ***Visualization***: presentation of data or analytic results in a way that communicates to others; and
5. ***Action***: processes that use the synthesized knowledge to generate value for the enterprise.

Analytics refer to a specific step 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. The data science life cycle encompasses many more activities beyond these five fundamental steps, including policy and regulation, governance, operations, data security, master data management, meta-data management, and retention/destruction.

Analytic processes are often characterized as ***discovery*** for the initial hypothesis formulation, ***development*** for establishing the analytics process for a specific hypothesis, and ***application*** 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 are 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 can 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. 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. Big Data science emphasizes the value from computation across the entire dataset. Determining correlation (and not necessarily causation) can be useful when knowing the direction or trend of something is enough to take action. Big Data solutions make it more feasible to implement causation type of complex analytics for large, complex, and heterogeneous data.

Another data science life cycle consideration is the speed of interaction between the analytics processes and the person or process responsible for delivering the actionable insight. Different Big Data use cases can be characterized further in terms of the time limits for the end-to-end analytics processing, at real time, near real time, or batch. 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 *NBDIF: Volume 6, Reference Architecture*.

5.3.1 DATA PERSISTENCE DURING THE LIFE CYCLE

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 do not 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 database.

In the traditional data warehouse, the data handling process followed the life cycle order given in Section 5.3 above (i.e., capture to staging, preparation and storage into a data warehouse, possibly query into data marts, and then 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 persisted. Dataset characteristics change the analytics life cycle processes in different ways. The following scenarios provide examples of changes to the points in the life cycle when data is stored as a function of dataset characteristics:

- ***Data warehouse***: Persistent storage occurs after data preparation.
- ***Big Data volume system***: Data is stored immediately in original 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 optionally include some summarization or aggregation of the streaming data prior to storage.

5.4 DATA CHARACTERISTICS IMPORTANT TO DATA SCIENCE

In addition to volume, velocity, variety, and variability, several terms, many beginning with V, have been used in connection with Big Data. Some refer to the characteristics important to achieving accurate results from the data science process.

5.4.1 *Veracity*

Veracity refers to the 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 very 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 would still be accurate. Data quality concerns, for the most part, are still vitally important for Big Data analytics. This concept is not new to Big Data, but remains important.

5.4.2 *Validity*

While the data may have high veracity (accurate representation of the real-world processes that created it), there are times when the data is no longer valid for the hypothesis being asked. For example, in a fast-changing environment such as the stock market, while historical price data has high veracity, it is not valid for certain types of analysis that rely upon real-time pricing information. In many cases, there is a time window before which the data is no longer valid for analysis. This concept is not new to Big Data, but remains important.

5.4.3 *Volatility*

Volatility refers to the change in the data values over time. Equipment can degrade, and biases or shifts be introduced, such as measurements from satellites. To analyze data accumulated over time, it becomes critical to understand changes in the production of the data to correct for drift or volatility and to enable historical analysis across all the data. This concept is not new to Big Data, but remains important.

5.4.4 *Visualization*

Visualization is an important step in any data science application to allow human understanding of the data, the analytics, or the results. There are three very different types of visualization that vary in techniques and in purpose.

Exploratory visualization refers to the techniques needed to browse original datasets to gain an understanding of distributions in data element values (e.g., across cases or across geography). This type of visualization has become more important with Big Data, and can require additional techniques for data summarization or aggregation to make the data accessible to a particular presentation format or device.

Evaluative visualization refers to the visualization that enables an evaluation and understanding of the performance and accuracy of a particular analytic or machine learning method. This visualization need is not new to Big Data since it refers to the results of the analysis—which typically is not a large amount of data. Small data is a new term has been coined to refer to the inability of analysts to properly evaluate results if the data is too large and complex.

> ***Small data*** *refers to the limits in the size of datasets that analysts can fully evaluate and understand directly.*

The evaluation of the results of analytic methods is vital to ensure that the results will indeed meet the mission need for the analysis.

Explanatory visualization, previously described as information visualization, concerns the presentation of complex data in a way easily understood by decision makers. Communication of the probabilistic accuracy of analytic results becomes a vital part of explanatory visualization as the production of Big Data analytic results becomes more complicated. Explanatory visualization could, for example, present the probabilities that a particular result will occur (e.g., in weather forecasts). Expressive techniques that enable cognition in humans are critical so that those who were not involved in the data collection or analytics will still understand the accuracy and applicability of the results.

5.4.5 VALUE

Value is the measure of gain, achieved by an organization or individual, as a result of the data science process implemented. Value is also used as a measure of the inherent potential in datasets—should they be fully analyzed. In the new information economy, value quantification and usage of calculated values for a particular application have not been standardized. For example, how should a company add the value (or potential value) of data to their asset balance sheet? This concept is not new to Big Data, but remains important.

5.4.6 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 was 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 [31]) and discipline-specific metadata (e.g., ISO 19115-x [32] for geospatial data). Metadata is not new to Big Data, but there is an increasing emphasis on proper creation and curation of metadata to improve automated data fusion. The following are examples of types of metadata.

Provenance type of metadata provides the history of the data so users can correctly interpret the data, especially when the data is repurposed from its original collection process to extract additional value. As data sharing becomes common practice, it is increasingly important to have information about how data was collected, transmitted, and processed. O***pen data*** is data that has been made publicly available to others outside the original data producer team. A more detailed discussion of the types of metadata is beyond the scope of this document, but can be found in the ISO standards documents. For continuing the "V" word theme, we could describe part of metadata as ***Vocabulary***, and part of keeping track of dataset changes with ***Versioning***.

Semantic metadata, another type of metadata, refers to the description of a data element that assists with proper interpretation of the data element. Semantic relationships are typically described using the Resource Description Framework[d] where you have a triple of noun-verb-subject (or entity-relationship-second Entity). Semantic

[d] See for example https://www.w3.org/RDF/

relationships 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 [33], [34] for additional information. Semantic metadata is important in the new Big Data paradigm since the Semantic Web represents a Big Data attempt to provide cross-cutting meanings for terms but is not new with Big Data.

Linked data[e] is data that is described according to a standard metadata relationship. With common data elements, datasets can be **aligned** along those elements, then the two datasets can be fused into a new dataset. This is the same process done in the integration of data in a data warehouse, where each dataset has primary keys that align with foreign keys in another dataset. Linked data arose from the semantic web, and the desire to do *data mashups*. The reuse of metadata element definitions has continued to expand, and is required for automated integration of disparate datasets.

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. Taxonomies help us to develop a conceptual model of an area, and allow comparison between conceptual models. Taxonomies will be discussed further in the *NBDIF: Volume 2, Taxonomies* on Big Data[2].

While these metadata concepts are important, they predate the Big Data paradigm shift.

5.4.7 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. Complexity is, however, important for data science in a number of disciplines. There are domains where data elements cannot be analyzed without understanding their relationship to other data elements. The data element relationship concept is evident, for example, in the analytics for the Human Genome Project, where it is the relationship between the elements (e.g., base pairs, genes, proteins) and their position and proximity to other elements that matters. The term *complexity* is often attributed to Big Data, but it refers to the interrelationship between data elements or across data records, independent of whether the dataset has the characteristics of Big Data.

5.4.8 OTHER C WORDS

There are a number of other characteristics of analytics that predated Big Data science. They can be described with a set of “C” words—concepts such as ***Completeness***, ***Cleanliness***, ***Comprehensiveness***, ***Consistency***, ***Concurrency***. These are characteristics on datasets in relationship to the full population they are intended to represent. They are all critical in the performance of analytics. All these concepts are not new to Big Data, and are discussed extensively in the data management and data analytics literature.

5.5 EMERGENT BEHAVIOR

There are four topics worth noting that have emerged due to Big Data science.

5.5.1 NETWORK EFFECT—DEEP LEARNING

The scaling of data management and processing has changed the nature of analytics in a number of ways. Obtaining significantly new or more accurate results from analytics based on larger amounts of data is known as the ***network effect***. One significant example is in a type of machine learning known as deep learning. Deep Learning is a computational approach to building models for pattern recognition, following loosely the concepts of how the brain is able to learn patterns and predict outcomes. Analytic techniques such as neural networks have been used for decades, but were typically restricted to small or “toy” datasets. The advances in computing and Big Data have changed the analytic capacity of such learning techniques significantly. Instead of processing a couple

[e] See, for example, LinkedData.org and lod-cloud.net.

thousand images or documents, deep learning can train on millions of examples –fundamentally changing analytics in a number of disciplines. Deep learning implementations have revolutionized many areas such as speech recognition, machine translation, object detection and identification in images and video, facial recognition, computer vision, and self-driving cars. Deep learning is an example analytic technique that has demonstrated the network effect of increasing accuracy due to the computational and data processing resources to analyze very large datasets.

5.5.2 MOSAIC EFFECT

Previously, it was difficult to obtain datasets from different domains. This led to what has been called *security by obscurity*. The costs involved in obtaining datasets ensured some level of privacy. With the greater availability of different types of datasets, it is now possible to easily create *mashups* (i.e., the integration of multiple datasets). One consequence of this is known as the Mosaic Effect, where the integration of multiple datasets now allows the identification of individuals, whereas this was not possible when examining the individual datasets alone. This has significantly changed the concerns over privacy. Information aggregators have integrated a number of datasets that provide a complete profile of an individual, typically monetizing this data by selling it to marketing firms. Similarly, the greater quantity of data available about an individual can also lead to privacy concerns. For example, the availability of cell-tower or geo-location tracking on mobile devices has made it possible to determine personal information such as where a person lives, where they work, or if they are on travel. Both data mashups and the availability of granular data have led to emergent concerns over protecting privacy.

5.5.3 IMPLICATIONS FOR DATA OWNERSHIP

Digitization has increased concerns over data ownership (the control over the usage of data), due to the ease of collection and replication of data. There are a number of roles in handling Big Data, including the data creator, the data subject (entity the data is describing), the data provider, and the repository owner where the data is stored. In some cases, these roles are filled by separate individuals or organizations. While clearly a legal question, we recognize that Big Data has made the question of data ownership more acute. Two examples can illustrate the emerging issues.

As an example of the changes with Big Data, the ubiquitous usage of smart phones and of social platforms has led to greater complexity in the concerns over data ownership. When data is generated through the use of social apps, the user is the data subject, but now the application provider has unlimited rights to use, reuse, analyze, and potentially sell that data to others. The individual user no longer has control over the use of this data, nor can they request that the data be deleted. In some ways, this is not a new phenomenon, since mailing lists have always been shared with other companies. The Big Data challenge arises from the substantial increase in information known about an individual from just their address. The aggregation of data by credit bureaus and marketing firms, as well as the sale of data collected through social apps increases the concerns over data ownership and control of data where you are the data subject.

Data subject, data ownership, and reuse control issues are also emerging in healthcare, in particular with the enhanced usage of electronic health records (EHRs). The data creator is the hospital or lab, the data subject is the patient. There are others who have an interest in this data, including insurance companies that are data consumers and health data exchanges which are data repository owners. Ownership and control in EHRs is a Big Data arena with many roles interacting with the data.

5.6 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 3.1, Big Data characteristics lead to the 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[f], and available information from their efforts may be included in future versions of this report.

[f] See www.tpc.org/tpch

6 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 *NBDIF: 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, as discussed in Section 5.5.2.
- **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 performance computing and communications (HPCC). The addition of new techniques to handle the variety, volume, velocity, and variability has amplified these concerns to the level of a national conversation, with unanticipated impacts on privacy frameworks.

Security and Privacy concerns are present throughout any Big Data system. In the past, security focused on a perimeter defense, but now it is well understood that defense-in-depth is critical. The term *security and privacy fabric* in the context of the NBDRA (see *NBDIF Volume 6: Reference Architecture*, Section 3) conceptually describes the presence of security and privacy concerns in every part of a Big Data system.

> ***Fabric** conceptually represents the presence of activities and components throughout a computing system.*

Security standards define a number of controls at each interface and for each component. Likewise, privacy is a concern for Big Data systems, where additional privacy concerns can be created through the fusion of multiple datasets, or the granularity of the data being collected.

7 BIG DATA MANAGEMENT

Given the presence of management concerns and activities throughout all components and activities of Big Data systems, management is represented in the NIST reference architecture as a *fabric*, similar to its usage for security and privacy. The primary change to managing Big Data systems naturally centers around the distribution of the data. Sizing a set of data nodes is a new skill in Big Data engineering, since data on a node is typically replicated across two slave nodes (for failover). This increases the capacity needed to handle a specific amount of data. The choice must be made up front what data values in a field to use to split up the data across nodes (known as *sharding*). This choice may not be the best one in terms of the eventual analytics, so the distribution is monitored and potentially reallocated to optimize systems. At the infrastructure level, since many applications run in virtualized environments across multiple servers, the cluster management portion is not new, but the complexity of the data management has increased.

7.1 ORCHESTRATION

The Orchestration role for Big Data systems is discussed in the *NBDIF: Volume 6, Reference Architecture*. This role focuses on all the requirements generation, and compliance monitoring on behalf of the organization and the system owner. One major change is in the negotiation of data access and usage rights with external data providers as well as the system's data consumers. This includes the need to coordinate the data exchange software and data standards.

7.2 DATA GOVERNANCE

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

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

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 in transactions. To maximize its benefit, data governance must also consider the issues of privacy and security of individuals of all ages, individuals as organizations, and organizations as organizations. Additional discussion of governance with respect to security and privacy can be found in the *NBDIF: Volume 4, Security and Privacy*.

Data governance is needed to address important issues in the new global Internet Big Data economy. One major change is that an organization's data is being accessed and sometimes updated from other organizations. This has happened before in the exchange of business information, but has now expanded beyond the exchange between direct business partners. Just as cloud-based systems now involve multiple organizations, Big Data systems can now involve external data over which the organization has no control.

Another example of change is that 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 further during the development of the next version of this document.

Appendix A: Acronyms

ACID	Atomicity, Consistency, Isolation, Durability
API	Application Programming Interface
BASE	Basic Availability, Soft-State, and Eventual Consistency
BLOBs	Binary Large Objects
CAP	Consistency, Availability, and Partition
CPS	Cyber-Physical Systems
CPU	central processing unit
DLT	distributed ledger technology
EHR	electronic health records
GB	gigabyte
GPU	graphic processing units
HPC	high performance computing
HPCC	high performance computing and communications
HPDA	High Performance Data Analytics
I/O	Input/Output
IaaS	Infrastructure-as-a-Service
IEC	International Electrotechnical Commission
IoT	Internet of Things
ISO	International Organization for Standardization
ITL	Information Technology Laboratory (within NIST)
JSON	JavaScript Object Notation
MPP	massively parallel processing
NARA	National Archives and Records Administration
NAS	network-attached storage
NASA	National Aeronautics and Space Administration
NBDIF	NIST Big Data Interoperability Framework
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
P2P	Peer-to-Peer
PaaS	Platform-as-a-Service
RDBMS	Relational Database Management Systems
SaaS	Software-as-a-Service
SANs	Storage Area Networks
SP	Special Publication
SQL	Structured Query Language
TB	terabyte
TCP	Transaction Processing Performance Council
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

Appendix B: References

[1] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies (SP1500-2),” 2015.

[2] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements (SP1500-3),” 2015.

[3] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 4, Security and Privacy (SP1500-4),” 2015.

[4] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 5, Architectures White Paper Survey (SP1500-5),” 2015.

[5] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 6, Reference Architecture (SP1500-6),” 2015.

[6] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap (SP1500-7),” 2015.

[7] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 8, Reference Architecture Interface (SP1500-9),” 2017.

[8] W. Chang and NIST Big Data Public Working Group, “NIST Big Data Interoperability Framework: Volume 9, Adoption and Modernization (SP1500-10),” 2017.

[9] T. White House Office of Science and Technology Policy, “Big Data is a Big Deal,” *OSTP Blog*, 2012. [Online]. Available: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal. [Accessed: 21-Feb-2014].

[10] G. E. Moore, “Cramming More Components Onto Integrated Circuits, Electronics, April 19, 1965,” *Electronics*, vol. 38, no. 8, pp. 82–85, 1965.

[11] M. A. Beyer and D. Laney, “The Importance of Big Data: A Definition,” Jun. 2012.

[12] L. Badger, T. Grance, R. Patt Corner, and J. Voas, “NIST Cloud Computing Synopsis and Recommendations,” 2012.

[13] L. Badger *et al.*, “US Government Cloud Computing Technology Roadmap, Volume I,” *Spec. Publ. 500-293*, vol. Volume I:, p. 42, 2014.

[14] International Organization for Standardization, *ISO/IEC 11404:2007, Information technology -- General-Purpose Datatypes (GPD)*. 2007.

[15] International Organization for Standardization, *ISO 21090:2011, Health informatics -- Harmonized data types for information interchange*. International Organization for Standardization, 2011.

[16] *N0095 Final SGBD Report to JTC1*. ISO/IEC JTC 1 Study Group on Big Data (SGBD), 2014.

[17] Gartner, “IT Glossary.” [Online]. Available: http://www.gartner.com/it-glossary/big-data. [Accessed: 17-Nov-2014].

[18] J. Dutcher, “What is Big Data,” *Data Science at Berkeley Blog*, 2014. [Online]. Available: http://datascience.berkeley.edu/what-is-big-data/. [Accessed: 17-Nov-2014].

[19] “Big Data (definition),” *Oxford English Dictionary*. [Online]. Available: http://www.oed.com/view/Entry/18833#eid301162178. [Accessed: 17-Nov-2014].

[20] J. Gantz and D. Reinsel, "Extracting Value from Chaos State of the Universe: An Executive Summary," *IDC iView*, no. June, pp. 1–12, 2011.

[21] McKinsey & Company, "Big data: The next frontier for innovation, competition, and productivity," *McKinsey Glob. Inst.*, no. June, p. 156, 2011.

[22] T. H. Davenport, *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Harvard Business Review Press, 2014.

[23] P. Mell and T. Grance, "NIST SP 800-145: The NIST Definition of Cloud Computing," 2011.

[24] M. D. Hogan, F. Liu, A. W. Sokol, T. Jin, and NIST Cloud Computing Standards Roadmap Working Group, "NIST Cloud Computing Standards Roadmap," 2011.

[25] NIST Cloud Computing Reference Architecture and Taxonomy Working Group, "NIST Cloud Computing Service Metrics Description," 2015.

[26] National Institute of Standards and Technology (NIST), "NIST Cloud Computing Security Reference Architecture (NIST SP 500-299)," *Spec. Publ. 500-299*, 2013.

[27] L. Badger *et al.*, "US Government Cloud Computing Technology Roadmap, Volume II," *Spec. Publ. 500-293*, vol. Volume II, p. 97, 2014.

[28] National Institute of Standards and Technology (NIST), "US Government Cloud Computing Technology Roadmap, Volume III (DRAFT)," vol. Volume III, no. Technical considerations for USG Cloud Computing Deployment Decisions, 2011.

[29] D. Yaga, P. Mell, N. Roby, and K. Scarfone, "Draft Blockchain Technology Overview (NISTIR-8202)," Jan. 2018.

[30] J. Grey, "Jim Gray on eScience: A Transformed Scientific Method," *The Fourth Paradigm*, 2009. [Online]. Available: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf. [Accessed: 01-Jun-2015].

[31] International Organization for Standardization, *ISO/IEC 11179-1:2015, Information technology - Metadata registries (MDR) – Part 1: Framework*. International Organization for Standardization, 2015.

[32] International Organization for Standardization, *ISO 19115-1:2014, Geographic information – Metadata – Part 1: Fundamentals*. International Organization for Standardization, 2014.

[33] Archer Phil, "W3C DATA ACTIVITY Building the Web of Data," *W3C*, 2013. [Online]. Available: https://www.w3.org/2013/data/.

[34] Dan Brickley and I. Herman, "Semantic Web Interest Group," *W3C*. [Online]. Available: https://www.w3.org/2001/sw/interest/.