Authority

This publication has been developed by National Institute of Standards and Technology (NIST) to further its statutory responsibilities …

Nothing in this publication should be taken to contradict the standards and guidelines made mandatory and binding on Federal agencies ….

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

There may be references in this publication to other publications currently under development by NIST in accordance with its assigned statutory responsibilities. The information in this publication, including concepts and methodologies, may be used by Federal agencies even before the completion of such companion publications. Thus, until each publication is completed, current requirements, guidelines, and procedures, where they exist, remain operative. For planning and transition purposes, Federal agencies may wish to closely follow the development of these new publications by NIST.

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. All NIST Information Technology Laboratory publications, other than the ones noted above, are available at http://www.nist.gov/publication-portal.cfm.

Comments on this publication may be submitted to:
National Institute of Standards and Technology
Attn: Information Technology Laboratory
100 Bureau Drive (Mail Stop 8900) Gaithersburg, MD 20899-8930
Reports on Computer Systems Technology

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

National Institute of Standards and Technology Special Publication XXX-series
xxx pages (April 23, 2014)

DISCLAIMER

This document has been prepared by the National Institute of Standards and Technology (NIST) and describes issues in Big Data computing.

Certain commercial entities, equipment, or material may be identified in this document in order to describe a concept adequately. Such identification is not intended to imply recommendation or endorsement by the NIST, nor is it intended to imply that these entities, materials, or equipment are necessarily the best available for the purpose.
Acknowledgements

This document reflects the contributions and discussions by the membership of the NIST Big Data Public Working Group (NBD-PWG), co-chaired by Wo Chang of the NIST Information Technology Laboratory, Robert Marcus of ET-Strategies, and Chaitanya Baru, University of California San Diego Supercomputer Center.

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

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

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

Natasha Balac, University of California, San Diego, Supercomputer Center
Chaitan Baru, University of California, San Diego, Supercomputer Center
Deborah Blackstock, MITRE Corporation
Pw Carey, Compliance Partners, LLC
Wo Chang, National Institute of Standards and Technology
Yuri Demchenko, University of Amsterdam
Nancy Grady, SAIC
Karen Guertler, Consultant
Christine Hawkinson, U.S. Bureau of Land Management
Pavithra Kenjige, PK Technologies

Orit Levin, Microsoft
Eugene Luster, U.S. Defense Information Systems Agency/R2AD LLC
Bill Mandrick, Data Tactics
Robert Marcus, ET-Strategies
Gary Mazzaferro, AlloyCloud, Inc.
William Miller, MaCT USA
Sanjay Mishra, Verizon
Rod Peterson, U.S. Department of Veterans Affairs
Mark Underwood, Krypton Brothers LLC
William Vorhies, Predictive Modeling LLC
Alicia Zuniga-Alvarado, Consultant

The editors for this document were Nancy Grady and Wo Chang.
# Table of Contents

Executive Summary ........................................................................................................... 1

1 Introduction .......................................................................................................................... 2
   1.1 Background .................................................................................................................. 2
   1.2 Scope and Objectives of the Definitions and Taxonomies Subgroup ......................... 3
   1.3 Report Production ..................................................................................................... 3
   1.4 Report Structure ........................................................................................................ 3

2 Big Data Definitions ........................................................................................................... 5
   2.1 Big Data Definitions .................................................................................................. 5
   2.2 Data Science Definitions .......................................................................................... 7

3 Big Data Elements ............................................................................................................ 10
   3.1 Data Elements .......................................................................................................... 10
   3.2 Dataset at Rest ......................................................................................................... 10
   3.3 Dataset in Motion .................................................................................................... 11
   3.4 Analytics within Data Science ................................................................................ 12
   3.5 Big Data Metrics ...................................................................................................... 12
   3.6 Big Data Security and Protection ............................................................................ 13

4 Big Data Patterns ............................................................................................................. 14
   4.1 Data Process ............................................................................................................ 14
   4.2 Data Process Ordering Changes ............................................................................. 14
   4.3 Transactions ............................................................................................................. 14
      4.3.1 ACID Transactions .......................................................................................... 15
      4.3.2 BASE Transactions ......................................................................................... 15
      4.3.3 Brewer’s CAP Theorem ................................................................................ 16
      4.3.4 Read Versus Write Transactions ................................................................... 16
   4.4 Data Analytics Time Window ................................................................................ 16
   4.5 Storage Medium Changes ....................................................................................... 16

5 Future Directions ............................................................................................................ 17

Appendix A: Definitions Study Source Materials .............................................................. A-1
Appendix B: Index of Terms .............................................................................................. B-1
Appendix C: Acronyms ..................................................................................................... C-1
Appendix D: References .................................................................................................... D-1

Figure

Figure 1: Skills Needed in Data Science ............................................................................. 8
Executive Summary

This *NIST Big Data Interoperability Framework: Volume 1, Definitions* was prepared by the NIST Big Data Public Working Group (NBD-PWG) Definitions and Taxonomy Subgroup 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. New technologies tend to emerge with a lot of hype, but it can take some time to tell what is new and different. While Big Data has been defined in a myriad of ways, the heart of the Big Data paradigm is the horizontal scaling of resources to achieve near linear scalability. The term data science takes into account the methods for handling Big Data, but is mostly used for the trend toward a more agile analytics cycle with ever more refined results. This report seeks to clarify the underlying concepts of Big Data and data science. Specifically by facilitating the use of the same term for the same concept, this report aims to enhance communication among Big Data producers and consumers.

The other volumes contained in the NIST Big Interoperability Framework are:
- Volume 2: Taxonomies
- Volume 3: Use Cases and General Requirements
- Volume 4: Security and Privacy Requirements
- Volume 5: Architectures White Paper Survey
- Volume 6: Reference Architectures
- Volume 7: Technology Roadmap

The authors emphasize that the information in these volumes represents a work in progress and will evolve in the future and as additional perspectives are available.
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 our networked, digitized, sensor-laden, information-driven world. The availability of vast data resources carries the potential to answer questions previously out of reach, including the following:

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

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

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

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

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

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

Motivated by the White House’s initiative and public suggestions, the National Institute of Standards and Technology (NIST) has accepted the challenge to stimulate collaboration among industry professionals to further the secure and effective adoption of Big Data. As one result of NIST’s Cloud and Big Data Forum held January 15–17, 2013, there was strong encouragement for NIST to create a public working group for the development of a Big Data Technology 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 overwhelming participation from industry, academia, and government from across the nation. The scope of the NBD-PWG involves forming a community of interests from all sectors—including industry, academia, and government—with the goal of developing a consensus on definitions, taxonomies, secure reference architectures, security and privacy requirements, and a technology roadmap. Such a consensus would create a vendor-neutral, technology- and infrastructure-independent framework that would enable Big Data stakeholders to identify and use the best analytics tools for their processing and visualization requirements on the most suitable computing platform and cluster, while also allowing value-added from Big Data service providers.

1.2 Scope and Objectives of the Definitions and Taxonomies Subgroup

The Definitions and Taxonomy Subgroup focused on identifying Big Data concepts and defining related terms in areas such as data science, reference architecture, and patterns.

The aim of this document is to provide a common base for those involved with Big Data. For managers, the terms 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, it 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 were not explored. However, the Subgroup did include the concepts involved in other methodologies 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, but also to stay general enough for the inevitable changes in the field.

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

1.4 Report Structure

This document seeks to clarify the meanings of the broad terms Big Data and data science, which are discussed at length in Section 2. The more elemental concepts and terms that shed additional insights are discussed in Section 3. Section 4 explores several more detailed concepts. This first version of Volume 1, Descriptions describes some of the fundamental concepts that will be important to determine categories or functional capabilities that represent architecture choices. By understanding the underlying communication and storage patterns, the strengths and weaknesses of different approaches for different needs will become clearer.
Tightly coupled information can be found in the other volumes of the NIST Big Data Interoperability Framework. Volume 2, Taxonomy provides a description of the more detailed components of the NIST Big Data Reference Architecture (NBDRA) described in Volume 6, Reference Architecture. The concepts related to security and privacy are more fully described in Volume 4, Security and Privacy. To understand how these systems are architected to meet users’ needs, the reader is referred to Volume 3, Use Cases and General Requirements. Descriptions of the future of Big Data and opportunities to use Big Data technologies are included in Volume 7, Technology Roadmap. Comparing the related sections in these volumes will provide a more complete picture of the consensus of the NBD-PWG.
2 Big Data Definitions

There are two fundamental concepts in the emerging discipline of Big Data that have been used to represent multiple concepts. These two concepts, Big Data and data science, are broken down into individual terms and concepts in the following subsections.

2.1 Big Data Definitions

Big Data refers to the inability of traditional data architectures to efficiently handle new data sets.

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

Data set characteristics that force a new architecture are: 1) the dataset at rest characteristics of **volume** and **variety** (i.e., data from multiple repositories, domains, or types), and 2) from the data in motion characteristics of **velocity** (i.e., rate of flow) and **variability** (i.e., the change in velocity.) These characteristics, known as the ‘V’s’ of Big Data, will be explained in Section 3. Each of these characteristics result in different data system architectures or different data lifecycle process orderings to achieve needed efficiencies. A number of other terms (particularly anything that can be expressed using a term starting with the letter ‘V’) are also used, several of which refer to the analytics process instead of new Big Data technologies.

The new Big Data paradigm occurs when the scale of the data—whether at rest or in motion—causes the management of the data to be a significant driver in the design of the system architecture. 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 horizontally scaled (i.e., adding more machines to the available collection) system that uses a loosely coupled set of resources in parallel. This type of parallelization shift began over 20 years ago, in the simulation community, when scientific simulations began using massively parallel processing (MPP) systems.

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

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

The Big Data paradigm of today is similar. Data systems need a level of extensibility that matches the scaling in the data. To get that level of extensibility, different mechanisms are needed to distribute data and data retrieval processes across loosely-coupled resources.

**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 (in analogy to the prior shift in the simulation community) is a one-time occurrence. Eventually, though, a new paradigm shift will occur beyond this ‘crowdsourcing’ of a processing or data system across multiple horizontally-coupled resources.
A difficult question is what makes Big Data big, or how large does a dataset have to be for it to be called Big Data? The answer is an unsatisfying “it depends.” Data is considered big if the use of the new scalable architectures provides a cost or performance efficiency over the relational data model. In other words, the functionality cannot be achieved in a traditional, single platform relational database.

Big data essentially focuses on the self-referencing viewpoint that data is big because it requires scalable systems to handle it. Conversely, architectures with better scaling have come about because of the need to handle Big Data.

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

New engineering techniques in the data layer have been driven by the growing prominence of data types that cannot be handled efficiently in a traditional relational model. The need for scalable access in structured data has led to software built on the name-value pair or non-relational table paradigms. The rise in importance of document analysis has spawned a document-oriented database paradigm, and the increasing importance of relationship data has led to efficiencies in the use of graph-oriented data storage.

The new non-relational model database paradigms are typically referred to as NoSQL systems, alternately defined as “no SQL” or “not only SQL” (see the concept discussions in section 3). The problem with identifying Big Data storage paradigms as NoSQL is first that it describes the storage of data with respect to a language for query and retrieval of data, and second, that there is a growing capability in the application of the SQL query language against the new non-relational data repositories. While this term will continue to refer to the new data models beyond the relational model, the term itself will hopefully be replaced with a more suitable term, since it is misplaced to name a set of new storage paradigms with respect to a query language that is now being used.

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

A term that is closely related to the variety characteristic of Big Data is federation. A **federated database system** is a type of meta²-database management system (DBMS), which transparently maps multiple autonomous database systems into a single federated database.

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

The Big Data paradigm has other implications beyond the technical innovations in logical data models. The changes are in the parallel distribution of data and code in the physical file system and also in the direct queries against this storage.

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

A third concept of Big Data is often referred to as moving the processing to the data, not the data to the processing.

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

The implication is that data is too extensive to be queried and moved into another resource for analysis, so the analysis program is instead distributed to the data-holding resources, with only the results being aggregated on a remote resource.

A number of additional uses of the buzzword Big Data actually refer to changes in analytics as a consequence of the extensiveness of the datasets, which will be discussed in the next section on data science.

At its heart, Big Data refers to the extensibility of data repositories and data processing across horizontally-scaled resources, much in the same way the compute-intensive simulation community embraced massively parallel processing two decades ago. By working out methods for communication among resources, the same scaling is now available to data-intensive applications.

2.2 Data Science Definitions

In its purest form, data science is the fourth paradigm of science, following theory, experiment, and computational science. The fourth paradigm is a term coined by Dr. Jim Gray in 2007 to refer to the conduct of data analysis as an empirical science, learning directly from data itself. Data science as a paradigm would refer to the formulation of a hypothesis, the collection of the data—new or pre-existing—to address the hypothesis, and the analytical confirmation or denial of the hypothesis (or the determination that additional information or study is needed). As in any experimental science, the end result could in fact 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 business need, or can be the restatement of a business need in terms of a technical hypothesis.

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

Data science can be understood as the activities happening in the “data layer” of the system architecture to extract knowledge from the raw data.

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

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

Analytics is the extraction of knowledge from information.

What has changed with the new Big Data paradigm is that analytics is no longer separable from the data model and the distribution of that data across horizontally scaled resources. When structured data was almost exclusively stored as organized information in a relational model, the analytics could be designed for this structure. While our definition of the data science paradigm refers to learning directly from data, in the Big Data paradigm this learning must now implicitly involve all steps in the data lifecycle, with analytics being only a subset.
Data science is the construction of actionable knowledge from raw data through the complete data lifecycle process.

Data science across the entire data lifecycle now incorporates principles, techniques and methods from many disciplines and domains including the analytics domains of mathematics, data mining (and, more specifically, machine learning and pattern recognition), statistics, operations research and visualization, along with the domains of systems, software and network engineering. Data scientists or data science teams, solve complex data problems by employing deep expertise in one or more of these disciplines, in the context of business strategy and under the guidance of domain knowledge. Personal skills in communication, presentation and inquisitiveness are also very important given the complexity of interactions within Big Data systems.

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

While this full collection of skills can be present in a single individual, it is also possible that these skills, as shown in Figure 1, are covered in the members of a team.

![Figure 1: Skills Needed in Data Science](image)

Data science is not solely concerned with analytics, but also with the end-to-end experimental lifecycle, where the data system is the “scientific equipment.” The implication is that the data scientist must be aware of the sources and provenance of the data, the appropriateness and accuracy of the transformations on the data, the interplay between the transformation algorithms and processes and the data storage mechanisms, etc. This end-to-end overview role is to ensure that everything is being done right to meaningfully address the hypothesis.

Data science has also been used as a buzzword to mean additional concepts beyond those given above.

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

Data sampling is another debated concept within data science. A brief digression into history is useful here. Statistics has a branch of study on the computer design of experiments, in which one calculates the necessary data to determine an outcome—for example, in a pharmaceutical clinical trial. When the data mining community began, the emphasis was typically on re-purposed data (i.e., data used to train models
was sampled from a larger dataset that was originally collected for another purpose). The often overlooked critical step was to ensure that the analytics were not prone to over-fitting—meaning the pattern matched the data sample chosen for training, but did not work well to answer questions of the overall data population. In the new Big Data paradigm, it is implied that data sampling is no longer necessary since the greater system performance can theoretically process all the data. However, even if all the available data is used, it still only represents those who had behaviors that led them to produce the data, which might not be the true population of interest. For example, studying Twitter data to analyze people’s behaviors doesn’t represent all people, as not everyone uses Twitter.

Another data science debate is the assertion that **more data beats better algorithms**. The heart of this debate says that a few bad data elements are less likely to influence the analytical results in a large dataset than if errors are present in a small sample of that dataset. If the analytics needs are correlation and not causation, then this assertion is easier to justify. Outside the context of large datasets in which aggregate trending behavior is all that matters, the data quality rule remains “garbage-in, garbage-out.”

Finally, Big Data is given a number of additional characteristics that actually refer to data science, including veracity (accuracy of the data), value (of the analytics to the organization), volatility (tendency for data structures to change over time), and validity (quality and accuracy of the data). These characteristics and others, including all quality control, metadata, and data provenance, are already present in data analytics, and are not new to Big Data. Section 3 contains additional discussions of these terms.

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

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

The rate of growth in the amount of data generated and stored has been increasing exponentially. Data growth rates are considered to be more than Moore’s Law would indicate if applied to data volumes - implying a doubling in volume every three months. This data explosion is creating opportunities for new ways of combining and using data to find value. One of the significant shifts is in the amount of unstructured data. Structured data has typically been the focus of most enterprise analytics, and has been handled through use of the relational data model. Micro-texts, relationship data, images and videos have seen such an explosion in quantity that the trend is to try to incorporate this data to generate value. The central benefit of Big Data analytics is the ability to process large amounts and types of information. Big Data To understand the components of Big Data, and to characterize the terms that are associated with Big Data and data science, the concepts will be addressed by category.

3.1 Data Elements

A thorough description of data elements and their data types is beyond the scope of this work. Previously, most of the data in business systems was structured data that could be described efficiently in a relational model. Unstructured data types such as text, image, video, and relationship data have been increasing in both volume and prominence. The need to analyze unstructured or semi-structured data has been present for many years, so while it is important to a discussion of storage and analytics, a data type taxonomy will not be discussed here, as it is outside the scope. There are no new types of data in the Big Data paradigm shift, only a change in emphasis on the value of unstructured or relationship data, and in the engineering of the different ways data can be more efficiently handled.

An additional concept that is again not new in the paradigm shift is the presence of complexity in the data elements. There are systems where data elements cannot be addressed independently. This is evident, for example, in analytics for the Human Genome, where it is the relationship between the elements, their position and proximity to other elements that matters. The term complexity refers to this inter-relationship between data elements or data records.

Also of note is the concept of metadata, or data about data. As we move into an era of open data and linked data, it becomes ever more important to have information about how data was collected, transmitted, and processed. This ensures that it will be used correctly when repurposed from its original collection process.

A companion concept that is outside the scope of Big Data is none-the-less worth noting here. Semantic data refers to the definitional description of a data element to ensure it is properly interpreted. There are a number of mechanisms for implementing these unique definitional descriptions which are outside our scope. There are also taxonomies or ontologies that represent information about entities and their relationships. The hierarchy of the components of the NBDRA is described in the NIST Big Data Interoperability Framework, Volume 2: Taxonomy and Volume 6: Reference Architecture.

Big Data comes in two distinct states: at rest and in motion. Data at rest is typically used in a batch process planned for a later time (e.g., the end of the day, month, quarter). Data in motion concerns the data as it is in transit from one persistence to another. This distinction enables us to clarify some of the characteristics of Big Data.

3.2 Dataset at Rest

Data at rest sometimes refers to all data in computer storage, while excluding data that is traversing a network or temporarily residing in computer memory to be read or updated. Data at rest can be archival or
reference files that are changed rarely or never; data at rest can also be data that is subject to regular but not constant change. Examples include vital corporate files stored on the hard drive of an employee's laptop, files on an external backup medium, files on the servers of a storage area network (SAN), or files on the servers of an offsite backup service provider. While there are monolithic systems that competently process data at rest, there are a number of options to process it by spreading it across a number of less expensive resources. Typical characteristics of data at rest that are significantly different in the era of Big Data are the volume and variety.

Volume is the characteristic of data at rest that is most associated with Big Data. Estimates show that the amount of data in the world doubles every two years. Should this trend continue, by 2020, there will be 500 times the amount of data as existed in 2011. The sheer volume of the data is colossal—and the era of a trillion sensors is upon us. These data volumes has stimulated new ways for scalable storage across a collection of horizontally coupled resources, as described in Section 2.1, Big Data Engineering.

Briefly, the traditional relational model has been relaxed for the persistence of newly prominent data types. These non-relational data models, typically lumped together as NoSQL, can currently be classified as Table, Name-Value, Document and Graphical models. A discussion to compare and contrast these logical data models was not part of the scope for this document.

The second characteristic of data at rest is the increasing need to use a variety of data, meaning the data represents a number of data domains and a number of data types. Traditionally, a variety of data was handled through transformations or pre-analytics to extract features that would allow integration with other data through a relational model. Alternately, a federated database was constructed as a relational database across the underlying relational databases. Given the wider range of data formats, logical models, timescales and semantics that are desirous to use in analytics, the integration of the variety of data becomes more complex. This challenge arises as data to be integrated could be text from social networks, image data, or a raw feed directly from a sensor source. Big Data engineering has spawned data storage models that are more efficient for unstructured data types than a relational model, causing a derivative issue for the mechanisms to integrate this data. It is possible that the data to be integrated for analytics may be of such volume that it cannot be moved in order to integrate, or it may be that some of the data is not under control of the organization creating the data system. In either case, the variety of Big Data forces a range of new Big Data engineering to efficiently and automatically integrate data that is stored across multiple repositories, in multiple formats, and in multiple logical data models.

There are additional aspects of Big Data at rest that will not be fully explored in this document, including the range of persistence mechanisms (flatfiles, RDB, markup, NoSQL models), and the mechanisms for providing the communication among the horizontally coupled resources holding the data in the NoSQL or non-relational models. This discussion relates to the relaxation of the principles of a relational model.

While very important, this was considered out-of-scope for this document as it needs much more thought and discussion. Future versions of this document may further investigate this topic. Likewise any discussion of the use of multiple tiers of storage (in-memory, cache, solid state drive, hard drive, network drive, etc.) of software defined storage for storage efficiency is likewise deferred for a later discussion.

Software Defined Storage is the use of software to determine the dynamic allocation of tiers of storage to reduce storage costs while maintaining the required data retrieval performance.

### 3.3 Dataset in Motion

Big Data in motion is processed and analyzed in real time, or near-real time, and therefore it has to be handled in a very different way than persisted data. Big Data in motion tends to resemble event processing architectures, and focuses on real-time or operational intelligence applications.

The velocity is the rate of flow at which the data is created, stored, analyzed and visualized. Simplistically, this means a large quantity of data is being processed in a short amount of time. In the Big
Data era, data is created, and passed on, in real time or near real time. Data flow rates are increasing with enormous speeds and variability, creating new challenges to enable real- or near real-time data usage. Traditionally this concept has been described as **streaming data**. For some companies, like those in telecommunications, who have been sifting through high volume and short-time interval data for years, these aspects are not new. The new horizontal scaling approaches do, however, add new Big Data engineering options for efficiently handling this data.

The second concept for data in motion is variability. **Variability** can refer to a change in the rate of data flow. Given that many data processes generate a surge in the amount of data arriving in a given amount of time, new techniques are needed to efficiently handle this data. The data processing is often tied up with the automatic provisioning of additional virtualized resources in a cloud environment. The techniques used here are again outside our scope, and more appropriate to a discussion of operational cloud architectures.

### 3.4 Analytics within Data Science

The new Big Data engineering technologies change the types of analytics that are possible, but do not result in completely new types of analytics. However, given the retrieval speeds, analysts are able to interact with their data in ways that were not previously possible. The analytic processes are often characterized as **discovery** for the initial hypothesis formulation, **developmental** for establishing the analytics process for a specific hypothesis, or **applied** to refer to the encapsulation of the analysis into an operational system.

In addition, greater emphasis is placed on the value of correlation. Most traditional analytics has focused on causation—being able to describe why something is happening. In some cases, though, knowing the direction of a trend is enough to take action.

With Big Data systems, some analytics techniques downsize or summarize the analytics before you can use them.

While analytics have otherwise not changed, analytics tools have to adapt to run against the horizontally distributed non-relational data repositories. One of the mechanisms for this is a “divide and conquer” algorithm known as MapReduce. MapReduce is a method of splitting a query and subsequent analytics tasks into code that runs on the individual data nodes. A corresponding method combines the results from each node into the final result of the query and analytics.

While a number of “V”s have been proposed to relate to analytics in the Big Data realm, most of these attributes are not different from what they have always been.

**Veracity** refers to the completeness and accuracy of the data. This relates to the “garbage-in, garbage-out” issue that has been with us for a long time. If the analytics are causal, then the quality of every data element is critically important. If the analytics are correlations or trending over massive volume datasets, then individual bad elements will be lost in the overall counts and the trend will still be accurate. It’s worth noting, though, that many people debate whether “more data beats better algorithms.”

The **provenance**, or history of the data, is becoming more critical in Big Data analytics, as more and more data is being re-purposed for new types of analytics in completely different disciplines from which the data was created. As the usage of data persists far beyond the control of the data producers, it becomes ever more critical that metadata about the full creation and processing history is made available along with the data.

### 3.5 Big Data Metrics

One of our most fundamental questions was actually the first one asked: “how big does data have to be in order to be Big Data” The unsatisfactory response is that it depends. The basic answer is whenever the
data system needs to span horizontally scaled systems in order to handle the data, then you have moved into needing Big Data technology. Another way this is often expressed is that you have a Big Data problem when the size of the data is itself a significant part of the problem. Additional consideration and development of this topic may be included in future version of this roadmap.

### 3.6 Big Data Security and Protection

The security and privacy components and concerns are discussed in the *NIST Big Data Interoperability Framework, Volume 4: Security and Privacy Requirements.*
4  Big Data Patterns

To provide definitions of the differences in Big Data technologies, different 'scenarios' and 'patterns' are described that illustrate methods related to the data characteristics described in Section 3. The scenarios describe the high-level functional processes that can be used to categorize and, therefore, better understand the different use cases presented in NIST Big Data Interoperability Framework, Volume 3: Use Cases and General Requirements. The scenarios will be developed in future versions of NBDWG documents as a way to harmonize across the seven volumes and drive one level down in understanding the NBDRA.

The patterns refer to the characteristics of the components used inside the NBDRA. The patterns will be explored in later versions of the Taxonomy document to better characterize the hierarchy of the components of the NBDRA. While these patterns will be explored more fully in later stages of the NBD-PWG discussions, some of the fundamental concepts are introduced here.

4.1  Data Process

The data lifecycle consists of four stages:

1. **Collection**: results in “raw” data, or data in its original form.
2. **Preparation**: the collection of processes that take raw data and turn it into cleansed, organized information.
3. **Analysis**: the techniques that take organized information and produce synthesized knowledge.
4. **Action**: the processes that take the synthesized or created knowledge and put them to use in the generation of value for the enterprise.

4.2  Data Process Ordering Changes

In the traditional data warehouse, data was collected, prepared, and then stored. The relational model was designed in a way that optimized the intended analytics. Given the different Big Data characteristics, the ordering of the data handling processes changes as follows:

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

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

4.3  Transactions

*Transaction processing* is a style of computing that divides work into individual, indivisible operations, called transactions. The implementations of relational database management systems (RDBMS) have relied on the enforcement of specific rules on data transactions—that all the steps that update the different tables and relations in a transaction are completed or each step that was completed is reversed so that the database remains in its original state. To maintain the integrity of the database, the transaction either completes all steps or is not applied at all.
Transactions are at the heart of the difference between relational models and non-relational models. While a more thorough analysis and discussion of the patterns in non-relational models will be done in later versions of this Definition document, the relational concepts for transactions are being included in this initial version to frame the on-going discussion.

### 4.3.1 ACID Transactions

Relational databases have traditionally supported the ACID transaction model. ACID transactions are:

- **Atomic**: Either all of the actions in a transaction are completed (i.e., transaction is committed) or none of them are completed (i.e., transaction is rolled back).
- **Consistent**: The transaction must begin and end with the database in a consistent state and must comply with all protocols (i.e., rules) of the database.
- **Isolated**: The transaction will behave as if it is the only operation being performed upon the database.
- **Durable**: The results of a committed transaction can survive system malfunctions.

The SQL standard defines four transaction isolation levels (i.e., read uncommitted, read committed, repeatable read, and serializable) in terms of three phenomena that could occur between two concurrent transactions, T1 and T2. The three phenomena are:

- **Dirty Reads**: T1 reads data modified by T2 but not yet committed
- **Unrepeatable reads**: T1 rereads data and see effects of data T2 has modified or deleted AND committed
- **Phantom reads**: T1 rereads data and sees data T2 has inserted AND committed

The following chart shows for each theoretical isolation level which phenomena are possible.

<table>
<thead>
<tr>
<th>Isolation Level</th>
<th>Dirty Read</th>
<th>Unrepeatable Read</th>
<th>Phantom Read</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read uncommitted</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Read committed</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Repeatable read</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Serializable</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

SQL implementations support transactions and isolation levels using a variety of mechanisms. These mechanisms typically require some amount of overhead. This overhead is often viewed as an impediment to highly scalable databases.

### 4.3.2 BASE Transactions

The BASE acronym is often used to describe the types of transactions typically supported by non-relational databases, although its origins reside in early mobile computing. BASE is specifically contrived to be the opposite of ACID. A **BASE System** is described in contrast to an ACID-compliant systems as

- **Basically Available**, **Soft state**, and **Eventually Consistent**

While ACID transactions must be consistent at the end of the transaction, BASE transactions allow a database to be in a temporarily inconsistent state that will eventually be resolved. This eventual consistency is an important concept in the overall state of a dataset that is distributed across resources.

BASE transactions are the other end of a continuum from ACID transactions where the continuum is in part described by Brewer’s CAP theorem.
4.3.3 Brewer’s CAP Theorem
Distributed Big Data is subject to Brewer’s CAP Theorem, which states that a distributed system can support only two of the following three characteristics:

- **Consistency** The client perceives that a set of operations has occurred all at once.
- **Availability** Every operation must terminate in an intended response.
- **Partition tolerance** Operations will complete, even if individual components are unavailable.

These definitions of consistency, availability, and partition tolerance do not reflect the various interpretations of each term since Brewer’s presentation of the theorem in 2000, or the complexities encountered when actually applying the theorem to a distributed system or database.

4.3.4 Read Versus Write Transactions
In many database applications, the ACID transaction characteristics are critical for transactions that write (or modify) one or more rows in one or more tables. For some applications, the ACID transaction properties are also critical for transactions that only read data. For these applications, it is critical for the data to effectively remain unchanged across the life of the transactions.

However, many applications can tolerate changes to the underlying data being read during a transaction. A statistical analysis of a billion rows, documents, etc. is unlikely to be significantly affected by the addition or modification of a small percentage of the underlying data. For these applications, it may be reasonable to operate in a read uncommitted mode.

4.4 Data Analytics Time Window
One area that needs more differentiation is how the analytics or data lifecycle time window affect the possible architectures. The Big Data characteristic velocity was discussed earlier, and refers to the rate at which information is flowing into and through the system. An additional time span for consideration is the speed of interaction between the analytics processes and the person or process responsible for delivering the actionable insight. While the three broadest categories of batch (or offline) processing, online or real-time processing, and interactive processing are not new, they are a large factor in the choice of architecture and component tools to be used. Given the greater query and analytic speeds within Big Data, due to the horizontal scaling, there is an increasing emphasis on the interactive category. 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.

4.5 Storage Medium Changes
Another area to be explored in the patterns approach to explaining different methods and methodologies is the use of a broad range of storage capacities. Big Data storage could range from using the spectrum from in-memory techniques, to caching techniques, to local disk, to remote disk, to archival storage. As mentioned in Section 3.2, this is an area that needs to be described with reference to the patterns that are used and the benefits of each.
5 Future Directions

This document has explored the 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. In future working group discussions, greater detail will be given to many of the features of Big Data systems to describe the abstract scenarios that can be combined into use cases, and the patterns that are instantiated in different Big Data component implementations to achieve greater capabilities and performance.
Appendix A: Definitions Study Source Materials
## Appendix B: Index of Terms

| A | analytics, 7 |
| B | Big Data, 5, 6 |
|   | Big Data engineering, 5 |
|   | Big Data paradigm, 4 |
|   | Big Data velocity application, 14 |
|   | Big Data volume system, 14 |
| C | complexity, 10 |
|   | Computational portability, 6 |
| D | data lifecycle, 6 |
|   | data sampling, 8 |
|   | data science, 8 |
|   | data science paradigm, 6 |
|   | data scientist, 7 |
|   | data warehouse, 14 |
| F | federated database system, 5 |
|   | fourth paradigm, 6 |
| M | massively parallel processing, 4 |
|   | metadata, 10 |
| N | non-relational models, NoSQL, 5 |
| O | ontologies, 10 |
| P | provenance, 13 |
| R | relational model, 10 |
| S | Schema-on-read, 6 |
|   | semantic data, 10 |
|   | semi-structured data, 10 |
|   | streaming data, 12 |
|   | structured data, 10 |
| T | taxonomies, 10 |
| U | unstructured data, 10 |
| V | validity, 8 |
|   | value, 8 |
|   | variability, 4 |
|   | variety, 4 |
|   | velocity, 10 |
|   | veracity, 8 |
|   | volatility, 8 |
|   | volume, 4 |
Appendix C: Acronyms
Appendix D: References

General Resources

Document References


2 http://en.wikipedia.org/wiki/Meta-