NIST Special Publication XXX-XXX

DRAFT NIST Big Data Interoperability Framework:

Volume 4, Security and Privacy Requirements

NIST Big Data Public Working Group
Security and Privacy Subgroup

Draft Version 1
April 23, 2014
http://dx.doi.org/10.6028/NIST.SP.XXX

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

NIST Special Publication xxx-xxx
Information Technology Laboratory

DRAFT NIST Big Data Interoperability Framework: Volume 4, Security and Privacy Requirements

Version 1

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

Month 2014

U. S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Patrick D. Gallagher, Under Secretary of Commerce for Standards and Technology and Director

Authority

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

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

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

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

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

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

Reports on Computer Systems Technology

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

National Institute of Standards and Technology Special Publication XXX-series
Natl. Inst. Stand. Technol. Spec. Publ. XXX-series, xxx pages (April 23, 2014)

DISCLAIMER

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

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

Acknowledgements

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

The document contains input from members of the NBD-PWG: Security and Privacy Subgroup, led by Arnab Roy (Fujitsu), Mark Underwood (Krypton Brothers), and Akhil Manchanda (GE); and the Reference Architecture Subgroup, led by Orit Levin (Microsoft), Don Krapohl (Augmented Intelligence), and James Ketner (AT&T).

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

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

Pw Carey, Compliance Partners, LLC

Wo Chang, National Institute of Standards and Technology

Brent Comstock, Cox Communications

Roy D'Souza, AlephCloud Systems, Inc.

Eddie Garcia, Gazzang, Inc.

David Harper, Johns Hopkins University/ Applied Physics Laboratory

Pavithra Kenjige, PK Technologies

Orit Levin, Microsoft

Yale Li, Microsoft

Akhil Manchanda, General Electric

Marcia Mangold, General Electric

Serge Mankovski, CA Technologies

Robert Marcus, ET-Strategies

Lisa Martinez, Northbound Transportation and Infrastructure, US

William Miller, MaCT USA

Sanjay Mishra, Verizon

Arnab Roy, Fujitsu

Anh-Hong Rucker, Jet Propulsion Laboratory

Paul Savitz, ATIS

John Schiel, CenturyLink, Inc.

Mark Underwood, Krypton Brothers LLC

Alicia Zuniga-Alvarado, Consultant

The editors for this document were Arnab Roy, Mark Underwood, and Wo Chang.

DRAFT

Table of Contents

Executive Summary **1**
1 Introduction **2**
1.1 Background 2
1.2 Scope and Objectives of the Security and Privacy Subgroup 3
1.3 Report Production 3
1.4 Report Structure 3
2 Big Data Security and Privacy **5**
2.1 Introduction 5
2.2 Big Data Governance, Risk Management, and Compliance (GRC) Challenges 6
2.3 Effects of Big Data Characteristics on Security and Privacy 7
2.3.1 Variety 7
2.3.2 Volume 7
2.3.3 Velocity 7
2.3.4 Veracity and Validity 7
2.3.5 Volatility 8
2.4 Other Signposts and Guidelines 8
2.5 Big Data Security Concerns 9
2.6 Actors 9
2.7 Specialized Security and Privacy Topics 9
2.8 Semantic Web and Domain-Specific Security 10
3 Example Use Cases for Security and Privacy **12**
3.1 Retail/Marketing 12
3.1.1 Consumer Digital Media Usage 12
3.1.2 Nielsen Homescan: Project Apollo 13
3.1.3 Web Traffic Analytics 13
3.2 Healthcare 14
3.2.1 Health Information Exchange 14
3.2.2 Genetic Privacy 15
3.2.3 Pharma Clinical Trial Data Sharing 15
3.3 Cybersecurity 16
3.3.1 Network Protection 16
3.4 Government 17
3.4.1 Military: Unmanned Vehicle Sensor Data 17
3.4.2 Education: Common Core Student Performance Reporting 17
3.5 Industrial: Aviation 18
3.5.1 Sensor Data Storage and Analytics 18
3.6 Transportation 19
3.6.1 Cargo Shipping 19
4 Taxonomy of Security and Privacy Topics **20**
4.1 Taxonomy – Conceptual Axis 20
4.1.1 Privacy 20
4.1.2 Provenance 21
4.1.3 System Health and Resilience 21
4.2 Taxonomy – Operational Axis 21

DRAFT

4.2.1 Registration, Security Model, and Policy Enforcement 22
4.2.2 Identity and Access Management 22
4.2.3 Data Governance 23
4.2.4 Visibility and Infrastructure Management 23
4.2.5 Risk and Accountability: 24
4.3 Taxonomy – Roles Axis – Infrastructure Technology, GRC, and Information Worker 25
4.3.1 Infrastructure Technology (IT) 25
4.3.2 GRC 25
4.3.3 Information Worker (IW) 25
4.4 Relationships Between S&P Concepts and Roles 26
4.4.1 Data Privacy 26
4.4.2 Data Veracity (Provenance) 26
4.4.3 System Health 26
4.5 Uncategorized Topics 27
4.5.1 Provisioning, Metering, and Billing 27
4.5.2 Data Syndication 27
5 Security Reference Architecture 28
5.1 Interface of Data Providers → Big Data Application Provider 29
5.2 Interface of Big Data Application Provider → Data Consumer 29
5.3 Interface of Application Provider ↔ Big Data IT Provider 29
5.4 Internal to Big Data IT Provider 29
5.5 General Considerations 30
6 Mapping Use Cases to NBDRA 31
6.1 Consumer Digital Media Use 31
6.2 Nielsen Homescan: Project Apollo 32
6.3 Web Traffic Analytics 33
6.4 Health Information Exchange (HIE) 33
6.5 Genetic Privacy 35
6.6 Pharma Clinical Trial Data Sharing 35
6.7 Network Protection 36
6.8 Military: Unmanned Vehicle Sensor Data 37
6.9 Education: Common Core Student Performance Reporting 38
6.10 Sensor Data Storage and Analytics 39
6.11 Cargo Shipping 39
7 Future Directions 41
Appendix A. Specialized Security and Privacy Topics A-1
Appendix B: Internal Security Practices: Internal Access Control Rules for General Industry B-1
Appendix C: Terms and Definitions C-1
Appendix D: Acronyms D-1
Appendix E: References E-1

Figures

Figure 1: Cargo Shipping Scenario. 19
Figure 2: Conceptual Axis of Big Data Security and Privacy Taxonomy. 20
Figure 3: Big Data Security and Privacy Taxonomy. 22

DRAFT

Figure 4: The General Big Data Reference Architecture. 28
Figure 5: Big Data Security Reference Architecture. 29
Figure A-1: Top Perceived Downsides of the Internet of Everything. 3
Figure B.1: Composite Cloud Ecosystem Security Architecture. 1

Tables

Table 1. Preliminary Big Data Topic and Subtopic List 10
Table 2: Mapping Consumer Digital Media Usage to the Security Reference Architecture 31
Table 3: Mapping Nielsen Homescan to the Security Reference Architecture 32
Table 4: Mapping Web Traffic Analytics to the Security Reference Architecture 33
Table 5: Mapping to the Security Reference Architecture 34
Table 6: Mapping Pharma Clinical Trial Data Sharing to the Security Reference Architecture 35
Table 7: Mapping Network Protection to the Security Reference Architecture 36
Table 8: Mapping Military Unmanned Vehicle Sensor Data to the Security Reference Architecture. 37
Table 9: Mapping Common Core K–12 Student Reporting to the Security Reference Architecture 38
Table 10: Mapping Cargo Shipping to the Security Reference Architecture 39

DRAFT

Executive Summary

This ***NIST Big Data Interoperability Framework Volume 4: Security and Privacy Requirements*** was prepared by the NBD-PWG's Security and Privacy Subgroup to identify security and privacy issues particular to Big Data. Big Data application domains include health care, drug discovery, finance and many others from both the private and public sectors. Among the scenarios within these application domains are health exchanges, clinical trials, mergers and acquisitions, device telemetry, and international anti-piracy. Security technology domains include identity, authorization, audit, network and device security, and federation across trust boundaries.

Clearly, the advent of Big Data has necessitated paradigm shifts in the understanding and enforcement of security and privacy (S&P) requirements. Significant changes are evolving, notably in scaling existing solutions to meet the volume, variety, and velocity of Big Data, and retargeting security solutions amid shifts in technology infrastructure, e.g., distributed computing systems and non-relational data storage. In addition, as diverse datasets become ever-easier to access, many are increasingly personal in nature. Thus, a whole new set of emerging issues must be addressed, including balancing privacy and utility, enabling analytics and governance on encrypted data, and reconciling authentication and anonymity.

With the key Big Data characteristics of variety, volume, and velocity in mind, the subgroup gathered use cases from volunteers, developed a consensus security and privacy taxonomy and reference architecture, and validated it by mapping the use cases to the reference architecture.

The other volumes that make up the NIST Big Data Roadmap are:

- Volume 1: Definitions
- Volume 2: Taxonomies
- Volume 3: Use Cases and General Requirements
- Volume 5: Architectures White Paper Survey
- Volume 6: Reference Architectures
- Volume 7: Technology Roadmap

The authors emphasize that the information in these volumes represents a work in progress and will evolve as time goes on and additional perspectives are available.

DRAFT

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.[1] The initiative's goals include helping to accelerate the pace of discovery in science and engineering, strengthening national security, and transforming teaching and learning by improving our ability to extract knowledge and insights from large and complex collections of digital data.

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

Motivated by the White House's initiative and public suggestions, the National Institute of Standards and Technology (NIST) has accepted the challenge to stimulate collaboration among industry professionals to further the secure and effective adoption of Big Data. As one result of NIST's Cloud and Big Data Forum held January 15–17, 2013, there was strong encouragement for NIST to create a public working group for the development of a Big Data Interoperability Framework. Forum participants noted that this roadmap should define and prioritize Big Data requirements, including interoperability, portability, reusability, extensibility, data usage, analytics, and technology infrastructure. In doing so, the roadmap would accelerate the adoption of the most secure and effective Big Data techniques and technology.

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

1.2 Scope and Objectives of the Security and Privacy Subgroup

The focus of the NBD-PWG Security and Privacy Subgroup is to form a community of interest from industry, academia, and government with the goal of developing a consensus, secure reference architecture to handle security and privacy issues across all stakeholders. This includes understanding what standards are available or under development, as well as identifying which key organizations are working on these standards.

Subgroup tasks include the following:

- Gather input from all stakeholders regarding security and privacy concerns in Big Data processing, storage, and services
- Analyze/prioritize a list of challenging security and privacy requirements that may delay or prevent adoption of Big Data deployment
- Develop a Security and Privacy Reference Architecture that supplements the General Big Data Reference Architecture

Subgroup deliverables include the following:

- Produce a working draft for the Big Data Security and Privacy Requirements Document
- Produce a working draft of the Big Data Security and Privacy Reference Architecture

1.3 Report Production

The NBD-PWG Security and Privacy Subgroup explored various facets of Big Data security and privacy to compose this report. The approach for developing this report involved the following activities:

- Announce the NBD-PWG Security and Privacy Subgroup is open to the public in order to attract and solicit a wide array of subject matter experts and stakeholders in government, industry, and academia
- Identify use cases specific to Big Data security and privacy
- Develop a detailed security and privacy taxonomy
- Expand the security component of the NBDRA and detail security and privacy concerns related to NBDRA components
- Map the identified security and privacy use cases to the NBDRA

1.4 Report Structure

The remainder of this document is organized as follows:

- Chapter 2 discusses security and privacy issues particular to Big Data
- Chapter 3 presents use cases
- Chapter 4 discusses abstraction of requirements
- Chapter 5 describes security practices

- Chapter 6 presents taxonomy of security and privacy topics
- Chapter 7 describes the security reference architecture
- Chapter 8 maps the use cases to the reference architecture
- Appendix A discusses special security and privacy topics
- Appendix B contains information about cloud technology
- Appendix C lists the terms and definitions appearing in the taxonomy
- Appendix D contains the acronyms used in this document
- Appendix E lists the references used in the document

DRAFT

2 Big Data Security and Privacy

The NBD-PWG Security and Privacy Subgroup identified security and privacy issues particular to Big Data. To fully address security and privacy concerns, Veracity and Volatility are added to the three primary V's: Variety, Volume and Velocity.

Other sections of this document define Big Data. Security and Privacy in Big Data projects can be different from traditional implementations in a number of ways. While not all concepts apply all of the time, these seven principles are believed to be representative of a larger set of differences:

1. Big Data projects often encompass heterogeneous components in which a single security scheme has not been designed from the outset.
2. Most security and privacy methods have been designed for batch or online transaction processing systems. Big Data projects increasingly involve one or more streamed data sources, used in conjunction with data at rest, creating unique security and privacy scenarios.
3. The use of multiple Big Data sources not originally intended to be used together can compromise privacy, security, or both. Approaches to de-identify personally identifiable information (PII) that were satisfactory prior to Big Data may no longer be adequate.
4. An increased reliance on sensor streams, such as those anticipated with the Internet of Things—e.g., smart medical devices, smart cities, smart homes—can create vulnerabilities that were more easily managed before amassed to Big Data scale.
5. Certain types of data thought to be too big for analysis, such as geospatial and video imaging, will become commodity Big Data sources. These uses were not anticipated, and/or may not have implemented security and privacy measures.
6. Issues of veracity, provenance, and jurisdiction are greatly magnified in Big Data. Multiple organizations, stakeholders, legal entities, governments and far more members of the citizenry will find data about themselves included in Big Data analytics.
7. Volatility is significant because Big Data scenarios envision that data is permanent *by default*. Security is a fast-moving field with multiple attack vectors and countermeasures. Data may be preserved beyond the lifetime of the security measures designed to protect it.

2.1 Introduction

Security and privacy measures are becoming ever more important as the generation and utilization of Big Data increases, and as the storage and availability of the data is increasingly public.

As the generation, access, and utilization of Big Data grow, so does the importance of security and privacy measures. Data generation is expected to double every two years to about 40,000 exabytes in 2020. It is estimated that over one third of the data in 2020 could be valuable if analyzed[2]. Less than a third of data needed protection in 2010 but more than 40% of data will need protection in 2020[3].

Security and privacy measures for Big Data involve a different approach than traditional systems. Big Data is increasingly stored on public cloud infrastructure built by various hardware, operating systems, and analytical software. Traditional security approaches usually addressed small scale systems holding static data on firewalled and semi-isolated networks. The surge in streaming cloud technology necessitates extremely rapid responses to security issues and threats.[4]

Big Data is increasingly generated and utilized across diverse industries such as health care, drug discovery, and finance. Effective communication across these diverse industries will require standardization of the usage of terms related to security and compliance. The NBD-PWG Security and

Privacy Subgroup aims to encourage participation in the global Big Data security discussion without losing sight of the complex and difficult security and privacy issues particular to Big Data.

The NBD-PWG identified nine application domains to frame discussions of Big Data topics: Government Operation, Commercial, Defense, Healthcare and Life Sciences, Deep Learning and Social Media, The Ecosystem for Research, Astronomy and Physics, Earth, Environmental, and Polar Science, and Energy. Examples of scenarios within these application domains include health exchanges, clinical trials, mergers and acquisitions, device telemetry, and international anti-piracy. Security technology domains include identity, authorization, audit, network and device security, and federation across trust boundaries. For example, when using cloud service providers and federating across trust boundaries, there is a clear need for other services such as cryptography to strengthen security

2.2 Big Data Governance, Risk Management, and Compliance (GRC) Challenges

The NBD-PWG Security and Privacy Subgroup identified several challenges related to Big Data GRC. The intention was not to create an exhaustive list of GRC challenges, but rather to contribute to the discussion of Big Data security and privacy. As the NBD-PWG continues its work and as the Big Data realm develops, this list of challenges will evolve. The Big Data GRC challenges are as follows:

- Difficult to develop and maintain a common security vocabulary
- Inappropriate security mechanism for clouds and federations
- Undefined legal standing of data
- Difficult to resolve security, technology and GRC requirements
- Increased security cost from inattention to GRC requirements
- Heightened Regulatory Scrutiny

Difficult to develop and maintain a common security vocabulary : Within one group of stakeholders, a common security vocabulary may take hold because of underlying governance, risk management, and compliance (GRC) requirements. However, there are disparate regulatory bodies across sovereign boundaries; often the entities that certify are distinct from those that audit or adjudicate. When methods to accommodate GRC can grow organically, there is a trend toward consistent blueprints for policy authoring, decision, and enforcement, and for federation of demands, claims, and obligations across organizations. However, consistency is difficult to maintain as variety increases dramatically.

Inappropriate security mechanism for clouds and federations: Big Data may not be implemented using cloud technology, but most expect that the cloud will play a big role. Clouds and federations tend to introduce complications for application and the technology domains, and security mechanisms are often a previous generation of enterprise solutions that are being repurposed inappropriately for a radically different threat model.

Undefined legal standing of data: Also, the legal standing of data can change when it moves from an enterprise data center to the data center of a service provider; therefore, laws pertaining to discovery can be interpreted in unpredictable ways. The variety of data sources could lead to data providers and service providers being in separate legal jurisdictions.

Difficult to resolve security, technology and GRC requirements: For some projects, Big Data efforts could entail trade-offs between security and compliance. Even if all participants were to have state-of-the-art security, it is possible that they would be in violation of compliance requirements if they did not federate identity, authorization, and audits in a manner that would enable each party to address compliance obligations. Big Data participants must communicate technical security requirements as well as their broader GRC intent.

Increased security cost from inattention to GRC requirements: Sometimes there is reduced investment in GRC requirements at project launch. This can later lead to the need to retrofit security using cryptographic techniques that can provide content-level security (e.g., format preserving encryption). In addition, recent news about government surveillance could catalyze changes in how Big Data systems are governed.

Heightened Regulatory Scrutiny: The Health Insurance Portability and Accountability Act (HIPAA), the Health Information Technology for Economic and Clinical Health (HITECH) Act,[5] and the draft Federal Trade Commission (FTC) Privacy Bill of Rights specify consumer access to data, as well as recourse to remedy data errors that are of consequence. These requirements need to shape some facets of Big Data, such as consumer-facing portals, management of health information ecommerce, and chain of custody for personally identifiable information (PII).

2.3 Effects of Big Data Characteristics on Security and Privacy

Variety, volume, and velocity , the 3 V's, are key elements of Big Data. Where possible, these properties directed the NBD-PWG Security and Privacy Subgroup's attention. While the 3 V's is a useful shorthand that has entered the public discourse about Big Data, there are other important characteristics of Big Data that affect security and privacy, such as veracity, validity, and voracity. These elements are discussed below with respect to their impact on Big Data security and privacy.

2.3.1 Variety

Variety describes the organization of the data—whether the data is structured, semi-structured, or unstructured. Retargeting traditional relational database security to non-relational databases has been a challenge[6]. These systems were not designed with security in mind, and security is usually relegated to middleware. Traditional encryption technology also hinders organization of data based on semantics. The aim of standard encryption is to provide semantic security, which means that the encryption of any value is indistinguishable from the encryption of any other value. So once encryption is applied, any organization of the data which depends on any property of the data values themselves are rendered ineffective, whereas organization of the metadata, which may be unencrypted, may still be effective.

2.3.2 Volume

The volume of Big Data describes how much data is coming in and usually ranges from gigabytes to petabytes. The volume of Big Data has necessitated storage in multi-tiered storage media. The movement of data between tiers has led to a requirement of systematically analyzing the thread models and research and development of novel techniques. The thread model for network-based, distributed, auto-tier systems include the following major scenarios[7]: confidentiality and integrity, provenance, availability, consistency, collusion attacks, roll-back attacks and recordkeeping disputes.

A flip side of having volumes of data is that analytics can be performed to detect security breach events. This is an instance where Big Data technologies can fortify security.

2.3.3 Velocity

Velocity describes the speed at which data is processed. The data usually arrives in batches or is streamed continuously. As with non-relational databases, distributed programming frameworks such as Hadoop were not developed with security in mind[8]. Malfunctioning computing nodes might leak confidential data. Partial infrastructure attacks could compromise a significantly large fraction of the system due to high levels of connectivity and dependency. If the system does not enforce strong authentication among geographically distributed nodes, rogue nodes can be added that can eavesdrop on confidential data.

2.3.4 Veracity and Validity

Big Data Veracity and Validity encompass several subcharacteristics.

Provenance—or what some have called veracity, in keeping with "V" theme—is important for both data quality and for protecting security and maintaining privacy policies. Big Data frequently moves across individual boundaries to group, community of interest, state, national, and international boundaries. Provenance addresses the problem of understanding the data's original source, such as through metadata—though the problem extends beyond metadata maintenance. Various approaches have been tried, such as for glycoproteomics,[9] but no clear guidelines yet exist.

Some experts consider the challenge of defining and maintaining metadata to be the overarching principle, rather than provenance. The two concepts, though, are clearly interrelated.

Provenance also encompasses information assurance for the methods through which information was collected. For example, when sensors are used, traceability to calibration, version, sampling and device configuration are needed.

Security and privacy can be compromised through unintentional lapses or malicious attacks on data integrity. Managing data integrity for Big Data presents additional challenges related to all the "V" components, but especially for PII. While there are technologies available to develop methods for de-identification, some experts caution that equally powerful methods can leverage Big Data to re-identify personal information; the availability of as-yet unanticipated data sets could make this possible.

2.3.5 Volatility

Volatility of data—how its management changes over time—directly impacts provenance. Big Data is transformational in part because systems may produce indefinitely persisting data—data that outlives the instruments on which it was collected; the architects who designed the software that acquired, processed, aggregated, and stored it; and the sponsors who originally identified the project's data consumers.

Roles are time-dependent in nature. Security and privacy requirements can shift accordingly. Governance can shift as responsible organizations merge or even disappear.

While research has been conducted into how to manage temporal data (e.g., in e-science for satellite instrument data),[10] there are few standards beyond simplistic timestamps and even fewer common practices available as guidance. To manage security and privacy for long-lived Big Data, data temporality should be taken into consideration.

2.4 Other Signposts and Guidelines

Where no clear guidelines have yet emerged, practitioners can leverage other resources:

- Current efforts in a discipline's Communities of Practice (CoPs)
- Current practice by allied CoPs, such as practitioners looking at radiology Big Data or imagery analyst on geospatial or defense imagery projects
- Surveys of significant user populations, such as levels of patient engagement with electronic medical records, use of social media data, and awareness of citizen rights and responsibilities
- More mature security and privacy practices in other sectors (e.g., military distributed systems) and their applications (e.g., the U.S. Navy's Global Command and Control System – Maritime, the Army Distributed Common Ground System,[11] and emerging Big Data uses for National Information Exchange Model [NIEM])
- Transparent systems specification and deployment processes (e.g., Incremental Commitment Spiral Model of software development),[12] which include design-time and prototyping-time outside reviews of security and privacy practices and are supplemented with penetration testing, use of software assurance tools, and concurrent development of multiple approaches to arrive at the best security approach
- Identify the highest-risk scenarios to prioritize efforts; for example, the National Security Agency (NSA) and insider threats or accurate patient identification in acute care facilities.[13]

2.5 Big Data Security Concerns

In addition, we need to understand how Big Data security concerns arise out of the defining characteristics of Big Data, and how they are differentiated from traditional security concerns.

Big Data may be gathered from diverse end points. There may be more types of actors than just providers and consumers—primarily, data owners, such as mobile users and social network users. Some "actors" may be devices that ingest data streams for still different data consumers.

Data aggregation and dissemination must be secured inside the context of a formal, understandable framework. This process should be an explicit part of a data consumer's contract with the data owner. The availability of data and its current status to data consumers is often an important aspect of Big Data. In some settings, this may dictate a need for public or closed-garden portals and ombudsman-like roles for data at rest.

Data search and selection can lead to privacy or security policy concerns. It is unclear what capabilities are provided by the provider in this respect. A combination of user competency and system protections is likely needed, including the exclusion of databases that enable re-identification.

Privacy-preserving mechanisms are needed for Big Data, such as for PII. Because there may be disparate processing steps between the data owner, provider, and data consumer, the integrity of data coming from end points must be ensured. End-to-end information assurance practices for Big Data—for example, for verifiability—are not dissimilar from other systems, but must be designed on a larger scale.

2.6 Actors

Individuals assume roles as Big Data actors through a variety of touch points in retail, government, education, finance, insurance, employment, and health care. A number of concrete examples are presented in Appendix B.

Big Data system topologies may benefit from more explicit roles than those in traditional systems, due to issues of provenance, a wider potential user audience, and increased use of PII.

Big Data system designers and custodians will face challenges as various as Big Data itself. While some Big Data systems will be self-contained and operate machine-to-machine with comparatively minimal human involvement, these may be the exception rather than the rule. Even with machine-to-machine systems, the designers of instrumentation and data consumers are human and thereby subject to disturbances, technology changes, and other factors.

Actor Literacy User communities evolve at different speeds, but in general, users evolve more rapidly than in the past. More rapid evolution can, among other issues, create a digital divide between digitally facile and digitally challenged populations. Regarding security and privacy, the role-specific competency of administrators, maintenance personnel, and data consumers cannot be overlooked. Big Data managers should monitor the performance of individuals and, when necessary, enhance their performance through training and compliance programs.

2.7 Specialized Security and Privacy Topics

The set of four topics below was initially adapted from the scope of the Cloud Security Alliance (CSA) Big Data Working Group (BDWG) charter, organized according to the classification in the CSA BDWG's *Top 10 Challenges in Big Data Security and Privacy*.[14]

Under each topic, a preliminary list of subtopics has been developed to help identify where Big Data security and privacy challenges could exist. A brief outline of the topics and subtopics are presented in Table XX to represent the current state of community discussion and are by no means complete. Developing subtopics with adequate supporting explanation is a large undertaking. However, the

Subgroup believes a detailed look at these topics and subtopics will be of interest to a larger community. Ongoing work, including revisions to these lists, by the NBD-PWG Security and Privacy Subgroup is presented in Appendix A. Later versions of this roadmap will refine the lists as needed.

Table 1. Preliminary Big Data Topic and Subtopic List

Big Data Security Topics	Subtopics
Infrastructure Security	Review of technologies and frameworks High availability issues
Data Privacy	Impact of the social data revolution Potential for violence and abuse Data protection
Governance	Data discovery and classification Policy management Data masking technologies Cross-border regulation Government access to data Data deletion Computing on encrypted data
Data Management	Securing data stores Attack surface reduction Key management and data ownership
Integrity and Security Intelligence	Security intelligence Large-scale analytics Streaming data analytics Event detection Forensics Security of analytics results

2.8 Semantic Web and Domain-Specific Security

Big Data systems may rely upon combinations of general and specialized knowledge frameworks. When formally organized, these are ontologies, "structural frameworks for organizing information and are used in artificial intelligence, the Semantic Web, systems engineering, software engineering, biomedical informatics, library science, enterprise bookmarking, and information architecture as a form of knowledge representation about the world or some part of it" (Wikipedia https://en.wikipedia.org/wiki/Ontology_(information_science)). Identifying roles, for example, within a specific knowledge framework, can require a detailed understanding of the ontology. Standards and recommendations for work in this area are being developed by Oasis and the Ontolog Forum http://ontolog.cim3.net/cgi-bin/wiki.pl?WikiHomePage#nidB and others.

Developing domain-specific security and privacy methods that work well for a specific application require a similar level of granularity. The methods may be specified within the domain. For example, the U.S. Centers for Medicare & Medicaid Services (CMS) specifies some requirements for U.S. patients whose care is funded in part by the federal government. Department of Defense contractors have specific protocols that must be followed when an individual is "read in" to a project at a specified secrecy level.

Security and privacy architects should understand, and if necessary, flesh out portions of the domain ontology as needed. Some of these ontologies exist in machine-readable form, such as the LHNCBC Medical Ontology Research Project http://mor.nlm.nih.gov/. Other efforts, such as NIEM (www.niem.gov) primarily entail sharing controlled vocabularies across justice, human services and healthcare. Security and privacy elements may be carried in these frameworks, and when omitted, must be integrated in sensible ways that reflect GRC and project constraints.

Variety, if for no other reason, means that no single security and privacy model will fit all Big Data systems. An informal expression of variety is reflected in the semantics of each particular Big Data domain. Security models for the genomics domain are undergoing rapid change, whereas in other domains, such as private-sector physical plant security (seismic sensor, guard post, video camera, etc.), they are less developed.

Security and Privacy for Fused Data: Big Data scenarios that entail fusion of previously disparate data sources are of particular concern because these systems straddle multiple domains. The semantic web may enable these sources to be connected, whether or not the sources intended them to be connected. System architects may wish to study work that has been performed on information fusion and situation awareness.[15] It is not usual for these projects, conducted over the past several decades, to include both human and sensor information, which makes them unusually relevant for Big Data. For historical reasons, these projects utilize linked data less frequently.[16]

Linked data offers Big Data projects the opportunity to achieve a more canonical taxonomy for security, privacy, roles, and "entities" than one-off, idiosyncratic efforts.

IT Infrastructure Considerations In addition to these domain-specific concerns, a system's particular infrastructure may also dictate security and privacy requirements. For instance, a private cloud operated within a company's locus of control could have different security concerns and taxonomies than a public cloud or a machine-to-machine Internet-of-Things (IoT) project. Outsourced infrastructure will involve Service Level Agreements (SLAs) that address security and privacy issues—even liability for loss. Information assurance, including resilience, should also be part of service level objectives.

Because different IT infrastructure entails potentially large differences in security fabric, Appendix _A addresses specialized security topics that are situation-dependent.

DRAFT

3 Example Use Cases for Security and Privacy

It was decided that the general Big Data use cases developed elsewhere in the larger document were insufficient for fleshing out security and privacy needs in a Big Data framework. Accordingly, a different set of use cases was developed. Some of these use cases are legacy applications no longer active, but which expose relevant security and privacy concerns.

An overview of the use cases developed follows.

- **Retail:** Nielsen Homescan Project Apollo data involved family-level retail transactions. Data was in-house but shared with customers who had partial access to data partitions. Others only received reports that include aggregate data, but which could be drilled down for a fee. Access security is traditional group policy, implemented at the field level using a database engine. There was considerable PII data. Survey participants are compensated in exchange for giving up segmentation data, demographics, etc.
- **Healthcare:** An employer should never have access to an employee's medical history, nor the medical records of an employee's family members. There is a need for HIPAA-capable cryptographic controls and key management. Doximity, which is already adopted, is a secure way for doctors to share research, clinical trial data, and patient records in the cloud. Data ownership issues for genetic data should also be investigated. Usage audit and security is required.
- **Media and Communications:** Collection and monitoring of PII data should be investigated. Currently, policy-based access control is enforced. Cybersecurity apps run at high levels of security and thus require separate audit and security measures.
- **Military:** Insider threat and leakage of military data, such as those collected by unmanned vehicles.
- **Education:** The privacy of student performance data needs to be preserved. Access control policy among teachers, students, and parents should be specified. However, this data is useful for predictive analytics for enhancing learning and decision making.
- **Marketing:** Visit-level webserver logs are used in correlation with other data sources. Opt-in consent is in place for tracking non-EU visitors. EU regulations are stricter. IP addresses reveal geographical data. MAC address tracking can be construed as PII.

3.1 Retail/Marketing

3.1.1 Consumer Digital Media Usage

Scenario Description: Consumers, with the help of smart devices, have become very conscious of price, convenience, and access before they decide on a purchase. Content owners license data for use by consumers through presentation portals, such as Netflix, iTunes, and others.

Comparative pricing from different retailers, store location and/or delivery options, and crowd-sourced rating have become common factors for selection. Retailers, to compete, are keeping a close watch on consumer locations, interests, and spending patternsto dynamically create deals and sell products that consumers do not yet know they want.

Current Security and Privacy: Individual data is collected by several means, such as smartphone GPS (global positioning system)/location, browser use, social media, apps on smart devices, etc.

- Privacy:
  - Most means described above offer weak privacy controls. In addition, consumer unawareness and oversight allow third parties to "legitimately" capture information. Consumers can have limited to no expectation of privacy in this scenario.
- Security:
  - Controls are inconsistent and/or not established appropriately to achieve the following:
    - Isolation, containerization, and encryption of data
    - Monitoring and detection of threats
    - Identification of users and devices for data feed
    - Interfacing with other data sources
    - Anonymization of users. Some data collection and aggregation uses anonymization techniques; however, individual users can be re-identified by leveraging other public Big Data pools
    - Original digital rights management (DRM) not built to scale to meet demand for the forecasted use for the data

Current Research: There is limited research in enabling privacy and security controls that protect individual data (whether anonymized or non-anonymized).

3.1.2 Nielsen Homescan: Project Apollo

Scenario Description: Nielsen Homescan is a subsidiary of Nielsen that collects family-level retail transactions. Project Apollo was a project designed to better unite advertising content exposure to purchase behavior among Nielsen panelists. Project Apollo did not proceed beyond a limited trial, but reflects a Big Data intent. The description is a best-effort general description and is not an official perspective from Nielsen, Arbitron or the various contractors involved in the project. The information provided here should be taken as illustrative rather than as a historical record.

A general retail transaction has a checkout receipt that contains all SKUs (stock keeping units) purchased, time, date, store location, etc. Nielsen Homescan collected purchase transaction data using a statistically randomized national sample. As of 2005, this data warehouse was already a multi-terabyte data set. The warehouse was built using structured technologies but was built to scale many terabytes. Data was maintained in house by Homescan but shared with customers who were given partial access through a private web portal using a columnar database, Sybase IQ. Additional analytics was possible through the use of Ab Initio. Other customers would only receive reports that include aggregated data, but greater granularity could be purchased for a fee.

Then-Current (2005-6) Security and Privacy:

- Privacy: There was a considerable amount of PII data. Survey participants are compensated in exchange for giving up segmentation data, demographics, etc.
- Security: There is traditional access security with group policy, implemented at the field level using the database (DB) engine, component-level application security and physical access controls.
- There were audit methods in place, but were only available to in-house staff. Opt-out data scrubbing was minimal.

3.1.3 Web Traffic Analytics

Scenario Description: Visit-level webserver logs are high-granularity and voluminous. To be useful, log data must be correlated with other (potentially Big Data) data sources, including page content (buttons, text, navigation events), and marketing-level events such as campaigns, media classification, etc. There are discussions—if not deployment—of plans for traffic analytics using complex event processing (CEP)

in real time. One nontrivial problem is segregating traffic types, including internal user communities, for which collection policies and security are different.

Current Security and Privacy:

- Non-European Union (EU): Opt-in defaults are relied upon to gain visitor consent for tracking. Internet Protocol (IP) address logging enables some analysts to identify visitors down to the level of a city block.
- Media access control (MAC) address tracking enables analysts to identify IP devices, which is a form of PII.
- Some companies allow for purging of data on demand, but most are unlikely to expunge previously collected webserver traffic.
- The EU has stricter regulations regarding collection of such data, which is treated as PII. Such web traffic is to be scrubbed (anonymized) or reported only in aggregate, even for multinationals operating in the EU but based in the United States.

3.2 Healthcare

3.2.1 Health Information Exchange

Scenario Description: Health Information Exchanges (HIEs) aspire to facilitate sharing of healthcare information that might include electronic health records (EHRs) so that the information is accessible to relevant covered entities, but in a manner that enables patient consent.

HIEs tend to be federated, where the respective covered entity retains custodianship of its data. This poses problems for many scenarios, such as emergencies, for a variety of reasons that include technical (such as interoperability), business, and security concerns.

Cloud enablement of HIEs, through strong cryptography and key management, that meets the HIPAA requirements for protected health information (PHI)—ideally without requiring the cloud service operator to sign a business associate agreement (BAA)—would provide several benefits, including patient safety, lowered healthcare costs, and regulated accesses during emergencies that might include break-the-glass and Centers for Disease Control and Prevention (CDC) scenarios.

The following are some preliminary scenarios that have been proposed:

- Break-the-Glass: There could be situations where the patient is not able to provide consent due to a medical situation, or a guardian is not accessible, but an authorized party needs immediate access to relevant patient records. Cryptographically enhanced key life cycle management can provide a sufficient level of visibility and nonrepudiation that would enable tracking violations after the fact.
- Informed Consent: When there is a transfer of EHRs between covered entities and business associates, it would be desirable and necessary for patients to be able to convey their approval, as well as to specify what components of their EHR can be transferred (e.g., their dentist would not need to see their psychiatric records.) Through cryptographic techniques, one could leverage the ability to specify the fine-grain cipher text policy that would be conveyed.
- Pandemic Assistance: There will be situations when public health entities, such as the CDC and perhaps other nongovernmental organizations that require this information to facilitate public safety, will require controlled access to this information, perhaps in situations where services and infrastructures are inaccessible. A cloud HIE with the right cryptographic controls could release essential information to authorized entities through authorization and audits in a manner that facilitates the scenario requirement.

Project Current and/or Proposed Security and Privacy:

- Security:
  - Lightweight but secure off-cloud encryption: There is a need for the ability to perform lightweight but secure off-cloud encryption of an EHR that can reside in any container that ranges from a browser to an enterprise server, and that leverages strong symmetric cryptography.
  - Homomorphic encryption.
  - Applied cryptography: Tight reductions, realistic threat models, and efficient techniques.
- Privacy:
  - Differential privacy: Techniques for guaranteeing against inappropriate leakage of PII.
  - HIPAA.

3.2.2 Genetic Privacy

Scenario Description: A consortium of policy makers, advocacy organizations, individuals, academic centers, and industry has formed an initiative, **Free the Data!**, to fill the public information gap caused by the lack of available genetic information for the BRCA1 and BRCA2 genes. The consortium also plans to expand to provide other types of genetic information in open, searchable databases, including the National Center for Biotechnology Information's database, ClinVar. The primary founders of this project include Genetic Alliance, the University of California San Francisco, InVitae Corporation, and patient advocates.

This initiative invites individuals to share their genetic variation on their own terms and with appropriate privacy settings in a public database so that their family, friends, and clinicians can better understand what the mutation means. Working together to build this resource means working toward a better understanding of disease, higher-quality patient care, and improved human health.

Current Security and Privacy:

- Security:
  - SSL (Secure Sockets Layer)-based authentication and access control. Basic user registration with low attestation level
  - Concerns over data ownership and custody upon user death
  - Site administrators may have access to data—strong encryption and key escrow are recommended
- Privacy:
  - Strict privacy that lets users control who can see information, and for what purpose
  - Some have expressed concerns over veracity and volatility: data ownership and custody upon user death

Researchers are currently investigating the circumstances under which data can be shared with the private sector and with government.

3.2.3 Pharma Clinical Trial Data Sharing[17]

Scenario Description: Companies routinely publish their clinical research, collaborate with academic researchers, and share clinical trial information on public websites, atypically at three different stages: the time of patient recruitment, after new drug approval, and when investigational research programs have been discontinued. Access to clinical trial data is limited, even to researchers and governments, and no uniform standards exist.

PhRMA, the Pharmaceutical Research and Manufacturers of America, represents the country's leading biopharmaceutical researchers and biotechnology companies. In July 2013, PhRMA joined with the

European Federation of Pharmaceutical Industries and Associations (EFPIA) in adopting joint Principles for Responsible Clinical Trial Data Sharing. According to the agreement, companies will apply these Principles as a common baseline on a voluntary basis, and PhRMA encouraged all medical researchers, including those in academia and government, to promote medical and scientific advancement by adopting and implementing the following commitments:

- Enhancing data sharing with researchers
- Enhancing public access to clinical study information
- Sharing results with patients who participate in clinical trials
- Certifying procedures for sharing trial information
- Reaffirming commitments to publish clinical trial results

Current and Proposed Security and Privacy:

PhRMA does not directly address security and privacy, but these issues were identified either by PhRMA or reviewers of the proposal.

- Security:
    - Longitudinal custody beyond trial disposition is unclear, especially after firms merge or dissolve
    - Standards for data sharing are unclear
    - There is a need for usage audit and security
    - Publication restrictions: Additional security will be required to ensure the rights of publishers; for example, Elsevier or Wiley
- Privacy:
    - Patient-level data disclosure—elective, per company
    - ThePhRMA mentions anonymization ("re-identification"), but mentions issues with small sample sizes
    - Study-level data disclosure—elective, per company

Current Research: TBD

3.3 Cybersecurity

3.3.1 Network Protection

Scenario Description: Network protection includes a variety of data collection and monitoring. Existing network security packages monitor high-volume data sets, such as event logs, across thousands of workstations and servers, but they are not yet able to scale to Big Data. Improved security software will include physical data correlates (access card usage for devices as well as building entrance/exit) and likely be more tightly integrated with applications, which will generate logs and audit records of previously undetermined types or sizes. Big Data analytics systems will be required to process and analyze this data to deliver meaningful results. These systems could also be multi-tenant, catering to more than one distinct company.

This scenario highlights two sub-scenarios:

- Security for Big Data
- Big Data for security

Current Security and Privacy:

- Security in this area is mature; privacy concepts less so.

  - Traditional policy-type security prevails, though temporal dimension and monitoring of policy modification events tends to be nonstandard or unaudited.
  - Cybersecurity apps run at high levels of security and thus require separate audit and security measures.
  - No cross-industry standards exist for aggregating data beyond operating system collection methods.
  - Implementing Big Data cybersecurity should include data governance, encryption/key management, and tenant data isolation/containerization.
  - Volatility should be considered in the design of backup and disaster recovery for Big Data cybersecurity. The useful life of logs may extend beyond the lifetime of the devices which created them.

- Privacy:
  - Enterprise authorization for data release to state/national organizations.
  - Protection of PII data.

Currently vendors are adopting Big Data analytics for mass-scale log correlation and incident response, such as for Security Information and Event Management (SIEM).

3.4 Government

3.4.1 Military: Unmanned Vehicle Sensor Data

Scenario Description: Unmanned vehicles ("drones") and their onboard sensors (e.g., streamed video) can produce petabytes of data that should be stored in nonstandardized formats. These streams are often not processed in real time, but the U.S. Department of Defense (DOD) is buying technology to make this possible. Because correlation is key, GPS, time, and other data streams must be co-collected. The Bradley Manning leak situation is one security breach use case.

Current Security and Privacy:

- Separate regulations for agency responsibility apply.
  - For domestic surveillance—the Federal Bureau of Investigation (FBI).
  - For overseas surveillance—multiple agencies, including the Central Intelligence Agency (CIA) and various DOD agencies.
- Not all uses will be military; for example, the National Oceanic and Atmospheric Administration.
- Military security classifications are moderately complex and determined on need to know basis.
- Information assurance practices are rigorously followed, unlike in some commercial settings.

Current Research:

- Usage is audited where audit means are provided, software is not installed/deployed until "certified," and development cycles have considerable oversight; for example, the U.S. Army's Army Regulation 25-2.[18]
- Insider threats (e.g., Edward Snowden, Bradley Manning, and spies) are being addressed in programs such as the Defense Advanced Research Projects Agency's (DARPA) Cyber-Insider Threat (CINDER) program. This research and some of the unfunded proposals made by industry may be of interest.

3.4.2 Education: Common Core Student Performance Reporting

Scenario Description: Forty-five states have decided to unify standards for K–12 student performance measurement. Outcomes are used for many purposes, and the program is incipient, but it will obtain

longitudinal Big Data status. The data sets envisioned include student-level performance across students' entire school history and across schools and states, as well as taking into account variations in test stimuli.

Current Security and Privacy:

- Data is scored by private firms and forwarded to state agencies for aggregation. Classroom, school, and district identifiers remain with the scored results. The status of student PII is unknown; however, it is known that teachers receive classroom-level performance feedback. The extent of students'/parents' access to test results is unclear.
- Privacy-related disputes surrounding education Big Data are illustrated by the reluctance of states to participate in the InBloom initiative (http://lohud.us/1mV9U2U).
- According to some reports,[19] parents can opt students out of state tests, so opt-out records must also be collected and used to purge ineligible student records.

Current Research:

- Longitudinal performance data would have value for program evaluators if data scales up.
- Data-driven learning[20] will involve access to students' performance data, probably more often than at test time, and also at higher granularity, thus requiring more data. One example enterprise is Civitas Learning's[22] predictive analytics for student decision making.

3.5 Industrial: Aviation

3.5.1 Sensor Data Storage and Analytics

Scenario Description: Most commercial airlines are equipped with hundreds of sensors to constantly capture engine and/or aircraft health information during a flight. For a single flight, the sensors may collect multiple gigabytes of data and transfer this data stream to Big Data analytics systems. Several companies manage these Big Data analytics systems, such as parts/engine manufacturers, airlines, and plane manufacturers, and data may be shared across these companies. The aggregated data is analyzed for maintenance scheduling, flight routines, etc. One common request from airline companies is to secure and isolate their data from competitors, even when data is being streamed to the same analytics system. Airline companies also prefer to control how, when, and with whom the data is shared, even for analytics purposes. Most of these analytics systems are now being moved to infrastructure cloud providers.

Current and Proposed Security and Privacy:

- Encryption at rest: Big Data systems should encrypt data stored at the infrastructure layer so that cloud storage administrators cannot access the data.
- Key management: The encryption key management should be architected so that end customers (airliners) have sole/shared control on the release of keys for data decryption.
- Encryption in motion: Big Data systems should ensure that data in transit at the cloud provider is also encrypted.
- Encryption in use: Big Data systems will desire complete obfuscation/encryption when processing data in memory (especially at a cloud provider).
- Sensor validation and unique identification (device identity management).

Researchers are currently investigating the following security enhancements:

- Virtualized infrastructure layer mapping on a cloud provider.
- Homomorphic encryption.
- Quorum-based encryption.
- Multi-party computational capability.
- Device public key infrastructure (PKI).

3.6 Transportation

3.6.1 Cargo Shipping

The following use case outlines how the shipping industry (e.g., FedEx, UPS, DHL) regularly uses Big Data. Big Data is used in the identification, transport, and handling of items in the supply chain. The identification of an item is important to the sender, the recipient, and all those in between with a need to know the location of the item while in transport and the time of arrival. Currently, the status of shipped items is not relayed through the entire information chain. This will be provided by sensor information, GPS coordinates, and a unique identification schema based on the new International Organization for Standardization (ISO) 29161 standards under development within the ISO technical committee ISO JTC1 SC31 WG2. The data is updated in near real time when a truck arrives at a depot or when an item is delivered to a recipient. Intermediate conditions are not currently known, the location is not updated in real-time, and items lost in a warehouse or while in shipment represent a potential problem for homeland security. The records are retained in an archive and can be accessed for xx days.

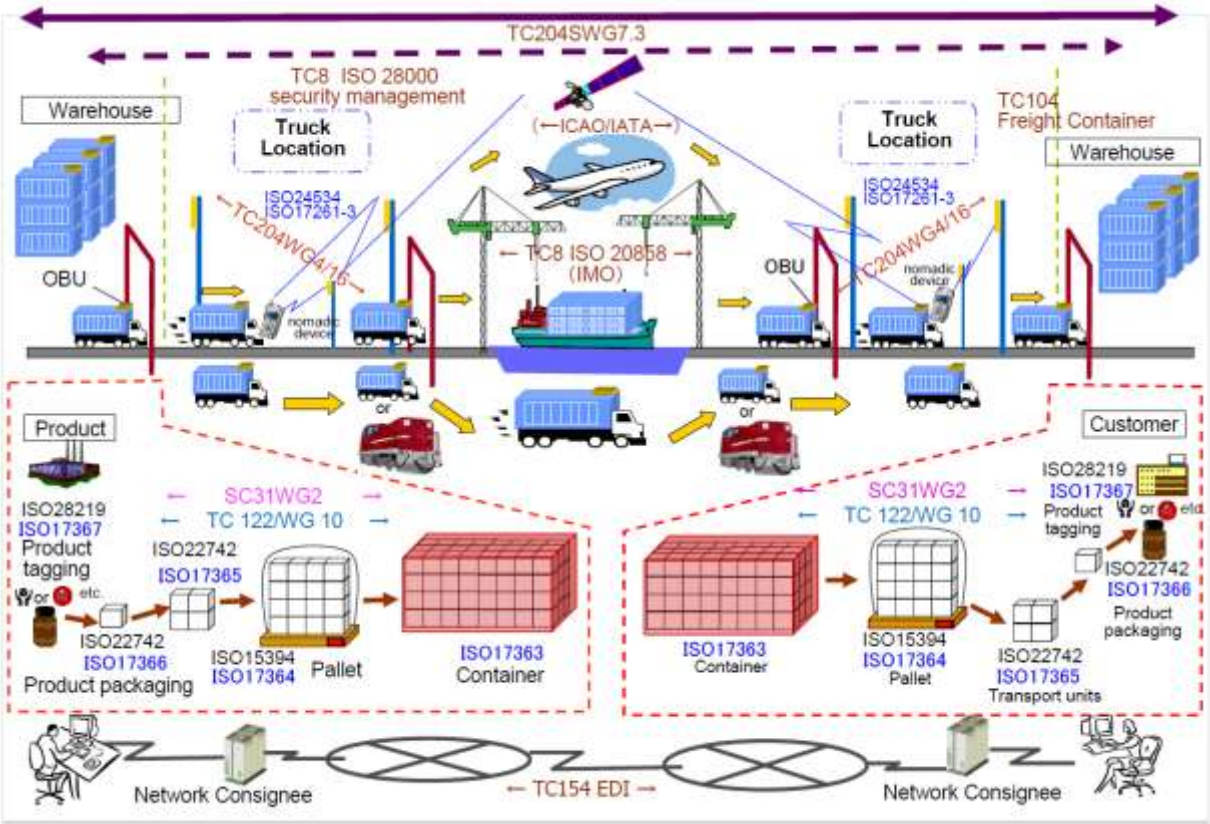

Figure 1: Cargo Shipping Scenario.

4 Taxonomy of Security and Privacy Topics

4.1 Taxonomy – Conceptual Axis

<Introductory material will be placed here>

Privacy	Communication Privacy		
	Data Confidentiality	Access Policies	Systems Crypto Enforced
	Computing on Encrypted Data	Searching and Reporting	
		Fully Homomorphic Encryption	
	Secure Data Aggregation		
	Key Management		
Provenance	End-point Input Validation	Syntactic Validation	
		Semantic Validation	
	Communication Integrity		
	Authenticated Computations on Data	Trusted Platforms	
		Crypto-Enforced	
	Granular Audits		
	Control of Valuable Assets	Lifecycle Management	
		Retention Disposition Hold	
		Digital Rights Management	
System Health	Security Against Denial of Service (DoS)	Construction of cryptographic protocols proactively resistant to DoS	
	Big Data for Security	Analytics for Security Intelligence	
		Data-Driven Abuse Detection	
		Event Detection	
		Forensics	

Figure 2: Conceptual Axis of Big Data Security and Privacy Taxonomy.

4.1.1 Privacy

- Communication Privacy: Confidentiality of data in transit, enforced, for example, by using Transport Layer Security (TLS).
- Confidentiality: Confidentiality of data at rest—policies to access data based on credentials.
    - Systems: Policy enforcement by using systems constructs such as Access Control Lists (ACLs) and Virtual Machine (VM) boundaries.
    - Crypto-Enforced: Policy enforcement by using cryptographic mechanisms, such as PKI and identity/attribute-based encryption.
- Computing on Encrypted Data
    - Searching and reporting: Cryptographic protocols that support searching and reporting on encrypted data—any information about the plaintext not deducible from the search criteria is guaranteed to be hidden.
    - Fully homomorphic encryption: Cryptographic protocols that support operations on the underlying plaintext of an encryption—any information about the plaintext is guaranteed to be hidden.
- Secure data Aggregation: Aggregating data without compromising privacy.
- Key management

4.1.2 Provenance

- End-Point Input Validation: A mechanism to validate whether input data is coming from an authenticated source, such as digital signatures.
    - Syntactic: Validation at a syntactic level.
    - Semantic: Semantic validation is an important concern. Generally, semantic validation would validate typical business rules such as a due date. Intentional or unintentional violation of semantic rules can lock up an application. This could also happen when using data translators that do not recognize the particular variant. Protocols and data formats may be altered by a vendor using, for example, a reserved data field that will allow their products to have capabilities that differentiate them from other products. This problem can also arise in differences in versions of systems for consumer devices, including mobile devices.
      The semantics of a message and the data to be transported should be validated to ensure, at minimum, conformity with any applicable standards. The use of digital signatures will be important to provide assurance that the data has been verified using a validator or data checker. This may be important to provide assurance that data from a sensor or from the data provider. This capability is important, particularly if the data is to be transformed or involved in the curation of the data. If the data fails to meet the requirements, it may be discarded, and if the data continues to present a problem, the source may be restricted in its ability to submit the data. These types of errors would be logged and prevented from being disseminated to consumers.
    - The Big Data system is a special case where use of digital signatures will be very important.
- Communication Integrity: Integrity of data in transit, enforced, for example, by using TLS.
- Authenticated Computations on Data: Ensuring that computations taking place on critical fragments of data are indeed the expected computations.
    - Trusted Platforms: Enforcement through the use of trusted platforms, such as Trusted Platform Modules (TPMs).
    - Crypto-Enforced: Enforcement through the use of cryptographic mechanisms.
- Granular Audits: Enabling audit at high granularity.
- Control of Valuable Assets:
    - Life Cycle Management
    - Retention, Disposition, and Hold
    - DRM

4.1.3 System Health and Resilience

- Security Against DoS: Construction of Cryptographic Protocols Proactively Resistant to DoS
- Big Data for Security
    - Analytics for Security Intelligence
    - Data-Driven Abuse Detection
    - Large-Scale and Streaming Data Analysis
    - Event Detection
    - Forensics

4.2 Taxonomy – Operational Axis

In the proposed taxonomy, broad considerations of privacy, provenance, and systems health appear as recurring features. For example, privacy of communications applies to governance of data at rest and access management, but it is also part of a security metadata model.[23]

The taxonomy will overlap with small data taxonomies while drawing attention to specific issues with Big Data.[24]

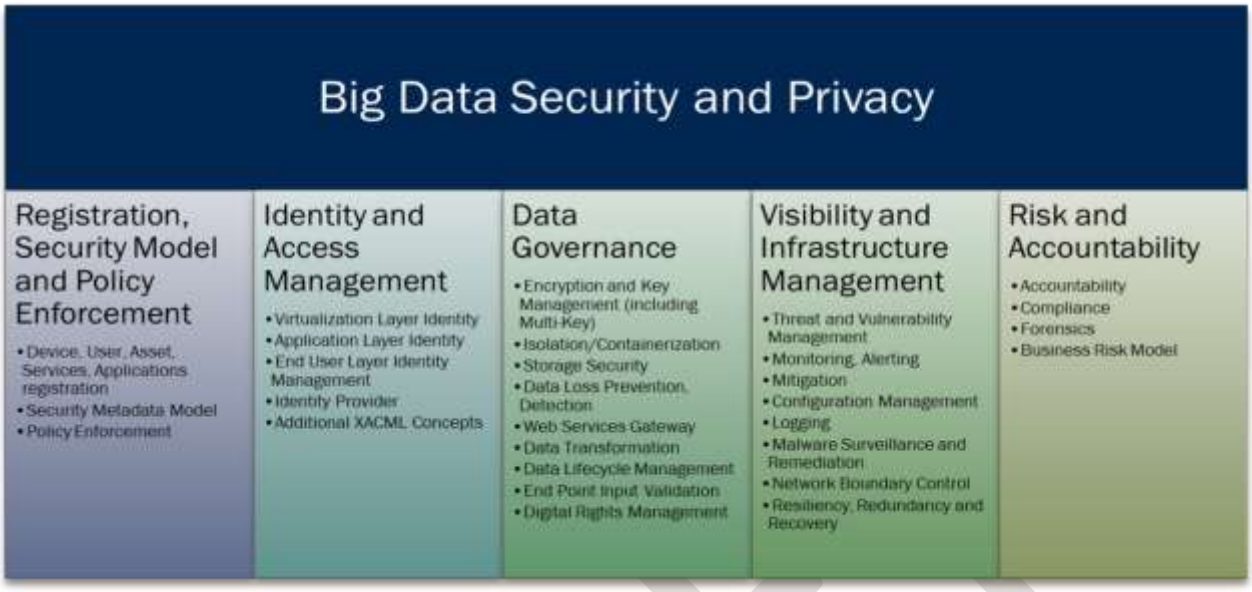

Figure 3: Big Data Security and Privacy Taxonomy.

4.2.1 Registration, Security Model, and Policy Enforcement

- Device, User, Asset, Services, and Applications Registration: Includes registration of devices in Machine to Machine (M2M) and Internet of Things (IoT) networks, DRM-managed assets, services, applications, and user roles.
- Security Metadata Model:
    - The metadata model maintains relationships across all elements of a secured system. It maintains linkages across all underlying repositories. Big Data often needs this added complexity due to its longer life cycle, broader user community, or other aspects.
    - A Big Data model must address aspects such as data velocity, as well as temporal aspects of both data and the life cycle of components in the security model.
- Policy Enforcement
    - Environment build
    - Deployment policy enforcement
    - Governance model
    - Granular policy audit
    - Role-specific behavioral profiling

4.2.2 Identity and Access Management

- Virtualization layer identity (e.g., cloud console, platform as a service [PaaS]): Trusted platforms
- Application layer Identity
- End-user layer identity management: Roles
- Identity Provider (IdP):
    - An IdP is defined in Security Assertion Markup Language (SAML) (2005). In a Big Data ecosystem of data providers, orchestrators, resource providers, framework providers, and data consumers, a scheme such as the SAML/Security Token Service (STS) or eXtensible Access Control Markup Language (XACML) is seen as one helpful—but not proscriptive—way to decompose the elements in the security taxonomy.

    - Big Data may have *multiple* IdPs. An IdP may issue identities (and roles) to access data from a resource provider. In the SAML framework, trust is shared via SAML/Web Services (WS) mechanisms at the registration phase.
    - In Big Data, due to the density of the data, the user "roams" to data (whereas in conventional virtual private network [VPN]-style scenarios, users "roam" across trust boundaries). Therefore, the conventional Authentication/Authorization (AuthN/AuthZ) model needs to be extended because the relying party is no longer fully trusted—they are custodians of somebody else's data. Data is potentially aggregated from multiple resource providers.
    - One approach is to extend the claims-based methods of SAML to add security and privacy guarantees.
- Additional XACML Concepts:
    - XACML introduces additional concepts that may be useful for Big Data security. In Big Data, parties are not just sharing claims, but also sharing policies about what is authorized. There is a policy access point at every data ownership and authoring location, and a policy enforcement point at the data access. A policy enforcement point calls a designated policy decision point for an auditable decision. In this way, the usual meaning of non-repudiation and trusted third parties is extended in XACML. Big Data presumes an abundance of policies, "points," and identity issuers, as well as data:
        - Policy authoring points
        - Policy decision points
        - Policy enforcement point
        - Policy access points

4.2.3 Data Governance

- Encryption and Key Management (Including Multi Key)
    - At rest
    - In memory
    - In transit
    - New: use case of privacy
- Isolation/Containerization
- Storage Security
- Data Loss Prevention and Detection
- WS Gateway
- Data Transformation
    - Aggregated data management
    - Authenticated computations
    - Computations on encrypted data
- Data Life Cycle Management
    - Disposition, migration, and retention policies
    - PII microdata as "hazardous" (Kum et al. 2013)
    - De-identification and anonymization
    - Re-identification risk management
- End-Point Validation
- DRM

4.2.4 Visibility and Infrastructure Management

- Threat and Vulnerability Management:
    - DoS-resistant cryptographic protocols
- Monitoring and Alerting:

    - As noted in the Critical Infrastructure Cybersecurity Framework (CIICF), Big Data affords new opportunities for large-scale security intelligence, complex event fusion, analytics, and monitoring.
- Mitigation:
    - Breach mitigation planning for Big Data may be qualitatively or quantitatively different.
- Configuration Management:
    - Configuration management is one aspect of preserving system and data integrity. It can include the following:
    - Patch Management
    - Upgrades
- Logging:
    - Big Data must produce and manage more logs of greater diversity and velocity. For example, profiling and statistical sampling may be required on an ongoing basis.
- Malware Surveillance and Remediation:
    - This is a well-understood domain, but Big Data can cross traditional system ownership boundaries. Review of NIST's "Identify, Protect, Detect, Respond, and Recover" framework may uncover planning unique to Big Data.
- Network Boundary Control:
    - NEW: Establishes a data-agnostic connection for a secure channel.
        - Shared services network architecture
        - Zones/cloud network design (including connectivity)
- Resilience, Redundancy, and Recovery:
    - Resilience:
        - The security apparatus for a Big Data system may be comparatively fragile in comparison to other systems.
    - Redundancy:
        - < To be detailed in later versions of this effort.>
    - Recovery:
        - Recovery for Big Data security failures may require considerable advance provisioning beyond that required for small data. Response planning and communications with users may be on a similarly large scale.

4.2.5 Risk and Accountability:

- Accountability
    - Information, process, and role behavior accountability can be achieved through various means, including:
        - Transparency portals and inspection points.
        - Forward- and reverse-provenance inspection.
- Compliance:
    - Big Data compliance can span multiple aspects of the security and privacy taxonomy, including privacy, reporting, and nation-specific law.
- Forensics:
    - Forensics techniques enabled by Big Data.
    - Forensics used in Big Data security failure scenarios.
- Business Risk Level:
    - Big Data risk assessments should be mapped to each element of the taxonomy.[25] Business risk models can incorporate privacy considerations.

4.3 Taxonomy – Roles Axis – Infrastructure Technology, GRC, and Information Worker

Documents for review on Big Data security should be accessible to a diverse audience, including individuals who specialize in cryptography, security, compliance, or information technology. In addition, there are domain experts and corporate decision makers who should understand the costs and impact of these controls. Ideally, these documents would be prefaced by information that would help specialists find the content relevant to them. The specialists could then provide feedback on those sections.

Organizations typically contain diverse roles and workflows for participating in a Big Data ecosystem. Therefore, this document proposes a pattern to help identify the "axis" of an individual's roles and responsibilities, as well as classify the security controls in a similar manner to make these more accessible to each class.

4.3.1 Infrastructure Technology (IT)

Typically, the individual role axis contains individuals and groups who are responsible for technical reviews before their organization is on-boarded in a data ecosystem. After the on-boarding, they are usually responsible for addressing defects and security issues.

When IT personnel work across organizational boundaries, they have to accommodate diverse technologies, infrastructures, and workflows that need to be integrated. For Big Data security, these include identity, authorization, access control, and log aggregation.

Their backgrounds and practices, as well as the terminologies they use, tend to be uniform, and they face similar pressures within their organizations to constantly do more with less. 'Save Money' is the underlying theme, and IT usually faces pressure when problems arise.

4.3.2 GRC

Typically, GRC is a function that draws participation from multiple areas of the organization, such as Legal, HR, IT, and Compliance. However, increasingly, GRC departments have their own heads. There tends to be a strong focus on compliance, often in isolation from technologies.

Similar to IT, GRC tends to have uniform backgrounds; leverage a common terminology; and have similar processes and workflows within a vertical, which typically has marquees that influence other organizations within that vertical or sector.

GRC within an organization is under pressure to protect the company from risks that might arise from loss of IP, legal risks due to actions by individuals within the organization, and compliance risks specific to its vertical. "Stay out of jail" is one way to describe GRC's underlying theme. GRC is also under pressure to prevent, then preserve and protect.

4.3.3 Information Worker (IW)

IWs are the individuals and groups who actually operate on the content that spans generation, transformation, and consumption. Due to the nascent nature of the technologies and related businesses, they tend to use common terms at the technical level; however, their roles and responsibilities, and the related workflows, do not always align across organizational boundaries. For example, a data scientist has deep specialization in the content and its transformation, but typically will only care about security and cryptography when it adds friction to his or her ability to transfer or access relevant information.

IWs are being exposed to a great number of products and services. They are under pressure from their organizations to deliver concrete business value from these new Big Data analytics capabilities by monetizing available data, monetizing the capability to transform by becoming a service provider, or optimizing and enhancing business by consuming third-party data.

4.4 Relationships Between S&P Concepts and Roles

To leverage these three axes and to facilitate collaboration and education, a *stakeholder* can be defined as an individual or group within an organization who is directly impacted by the selection and deployment of a Big Data solution. A *ratifier* is defined as an individual or group within an organization who is tasked with assessing the candidate solution before it is selected and deployed. For example, a third-party security consultant may be deployed by an organization as a ratifier, and an internal security specialist with an organization's IT department might serve as both a ratifier and a stakeholder if tasked with ongoing monitoring, maintenance, and audits of the security.

The next sections cover the three current components of the taxonomy: privacy, veracity, and system health. The upcoming sections also explore potential gaps that would be of interest to the anticipated stakeholders and ratifiers who reside on these three new conceptual axes.

4.4.1 Data Privacy

IT specialists who address cryptography should understand the relevant definitions, threat models, assumptions, security guarantees, and core algorithms and protocols. These individuals will likely be ratifiers, rather than stakeholders.

IT specialists who address end-to-end security should have an abbreviated view of the cryptography, as well as a deep understanding of how the cryptography would be integrated into their existing security infrastructures and controls.

GRC should reconcile the vertical requirements—such as, perhaps, HIPAA requirements related to EHRs—and the assessments by the ratifiers that address cryptography and security. GRC managers would in turn be ratifiers to communicate their interpretation of the needs of their vertical. Persons in these roles also serve as stakeholders due to their participation in internal and external audits and other workflows.

4.4.2 Data Veracity (Provenance)

Data veracity (or provenance) is similar to data privacy, but it might introduce Information Workers (IWs) as ratifiers because businesses may need to protect their intellectual property (IP) from direct leakage or from indirect exposure during subsequent Big Data analytics. IWs would need to work with the ratifiers from cryptography and security to convey the business need, as well as understand how the available controls may apply.

Similarly, when an organization is obtaining and consuming data, IWs may need to ensure that the data provenance guarantees some degree of information integrity that addresses data that is incorrect, fabricated, or cloned before the data is presented to an organization.

There could also be Governance, Risk and Compliance (GRC) risks to an organization if one of its data suppliers does not demonstrate the appropriate degree of care in filtering or labeling its data. For example, the organization may not have signed a Basic Agreement (BAA), and the organization's GRC department's interpretation of the HIPAA Omnibus Rule might indicate that it could be at risk if the supplier has access to Electronic Health Records (EHRs) and presumably has signed a Basic Agreement (BAA).

4.4.3 System Health

System health is typically the domain of IT, and IT managers will be ratifiers and stakeholders of technologies, protocols, and products that are used for system health. IT managers will also design how the responsibilities would be shared across the partners who provide data, analytics or services—an area commonly known as operations systems support (OSS) in the telecom industry, which has significant experience in syndication of services.

Security and cryptography specialists should scrutinize the system health to spot potential gaps in the operational architectures. The likelihood of gaps increases when a system infrastructure includes diverse technologies and products.

4.5 Uncategorized Topics

There are additional areas that have not been carefully scrutinized, and it is not clear whether these would fold into existing categories or if new categories for security and privacy concerns would need to be identified and showcased. The following are a few candidates.

4.5.1 Provisioning, Metering, and Billing

Commercial pipelines for Big Data can be constructed and monetized more readily if these systems are agile in offering services, metering access suitably, and integrating with billing systems. While this process can be manual for a small number of participants, it can become complex very quickly when there are many suppliers, consumers, and service providers. IWs and IT professionals who are involved with existing business processes would be candidate ratifiers and stakeholders. Assuring privacy and security of provisioning and metering data may or may not have already been designed into these systems. The scope of metering and billing data will explode, so potential uses and risks have likely not been fully explored.

There are both veracity and validity concerns with these systems.

4.5.2 Data Syndication

Similar to service syndication, a data ecosystem is most valuable if any participant can have multiple roles, which could include supplying, transforming, or consuming Big Data. Therefore, a need exists to consider what types of data syndication models should be enabled; again, IWs and IT professionals are candidate ratifiers and stakeholders. Syndication involves transfer of risk and responsibility for security and privacy.

DRAFT

5 Security Reference Architecture

Security and privacy considerations are a fundamental aspect of the NBDRA. This relationship is depicted in Figure 4 by the presence of the security and privacy fabric around all of the functional components. In Figure 4, the role of security and privacy is depicted in the correct relation to the components. At the same time, it does not expand into finer details. In addition to the application and framework providers, NBD-PWG decided to include data providers and data consumers into the fabric because these entities should, at the very least, agree on the security protocols and mechanisms in place.

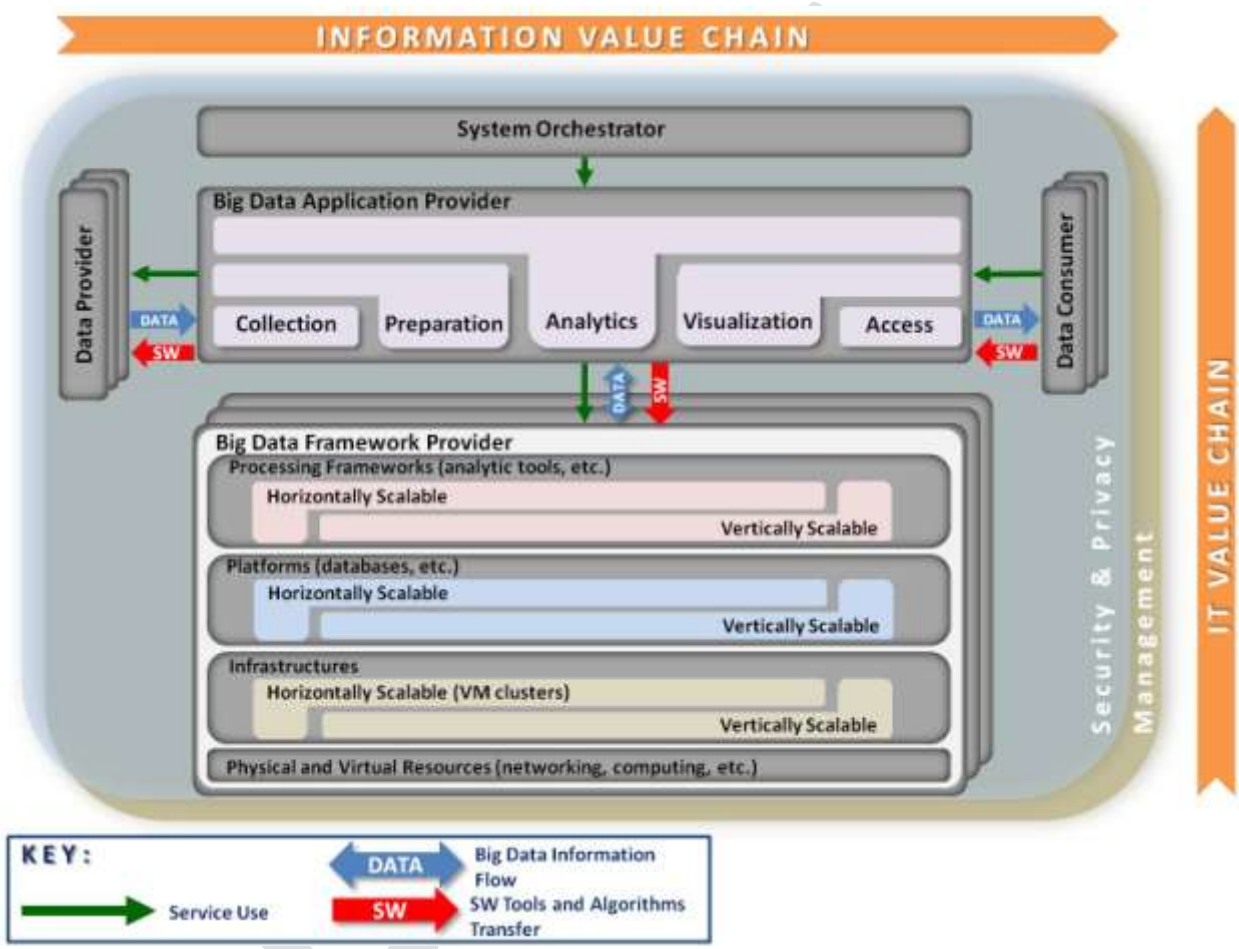

Figure 4: The General Big Data Reference Architecture.

This section proposes the NIST Big Data Security Reference Architecture (Figure 5) and provides details on the security and privacy considerations at the interface of and internal to the NBDRA components.

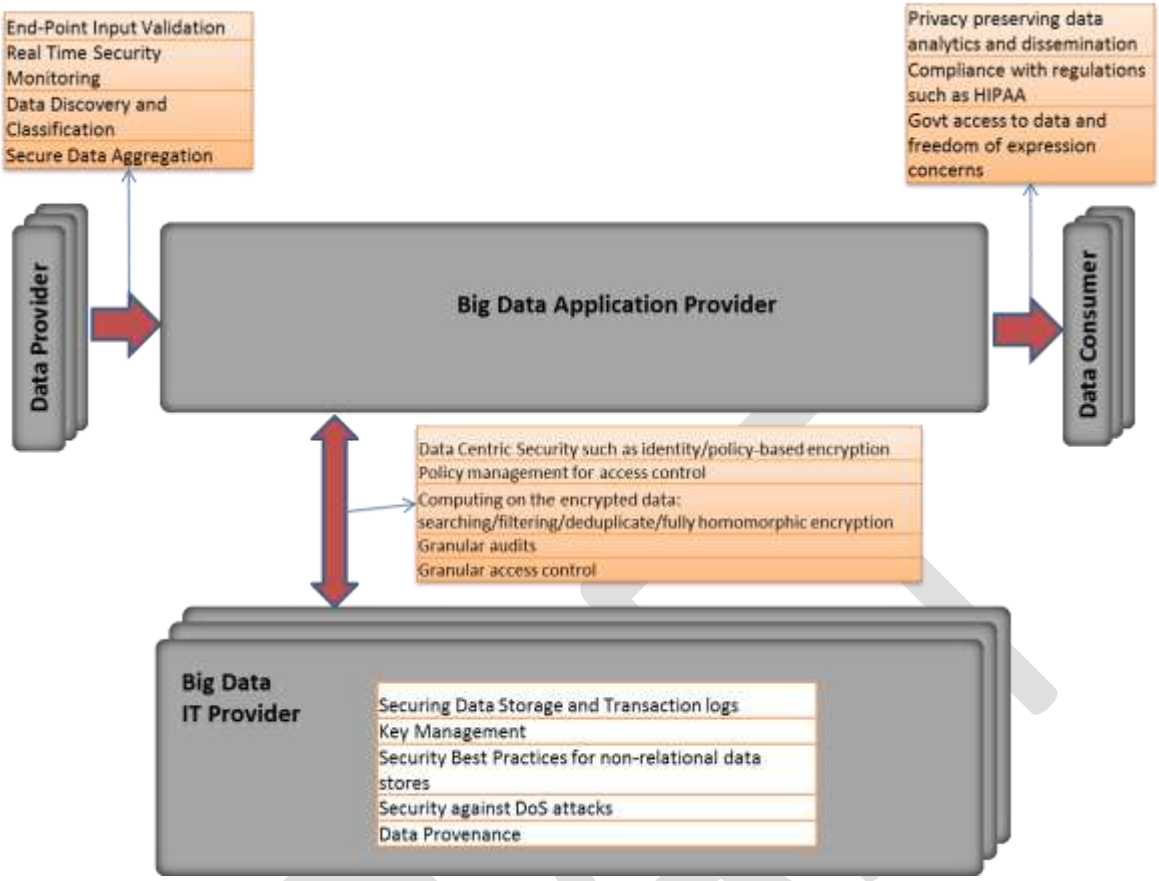

Figure 5: Big Data Security Reference Architecture.

5.1 Interface of Data Providers → Big Data Application Provider

Data coming in from data providers may have to be validated for integrity and authenticity. Incoming traffic may be maliciously used for launching Denial of Service (DoS) attacks or for exploiting software vulnerabilities on premise. Therefore, real-time security monitoring is useful. Data discovery and classification should be performed in a manner that respects privacy.

5.2 Interface of Big Data Application Provider → Data Consumer

Data or aggregate results going out to data consumers must preserve privacy. Data accessed by third parties or other entities should follow legal regulations such as HIPAA. Concerns are access to sensitive data by the government and potential undermining of freedom of expression.

5.3 Interface of Application Provider ↔ Big Data IT Provider

Data can be stored and retrieved under encryption. Access control policies should be in place to ensure that data is only accessed at the required granularity with proper credentials. Sophisticated encryption techniques can allow applications to have rich policy-based access to the data as well as enable searching, filtering on the encrypted data, and computations on the underlying plaintext.

5.4 Internal to Big Data IT Provider

Data at rest and transaction logs should be kept secured. Key management is essential to control access and keep track of keys. Non-relational databases should have a layer of security measures. Data

provenance is essential to having proper context for security and function of the data at every stage. DoS attacks should be mitigated to ensure availability of the data.

5.5 General Considerations

Big Data frameworks can also be used for strengthening security. Big Data analytics can be used for security intelligence, event detection, and forensics.

DRAFT

6 Mapping Use Cases to NBDRA

6.1 Consumer Digital Media Use

Content owners license data for use by consumers through presentation portals, such as Netflix, iTunes, etc. Consumers' digital media use creates Big Data, including both demographics at the user level and patterns of use such as play sequence, recommendations, and content navigation.

Table 2: Mapping Consumer Digital Media Usage to the Security Reference Architecture

RA Component	Security and Privacy Topic	Use Case Mapping
Sources → Transformation	End-point input validation	Varies and is vendor dependent. Spoofing is possible. For example, protections afforded by securing Microsoft Rights Management Services.[26] Secure/Multipurpose Internet Mail Extensions (S/MIME).
	Real-time security monitoring	Content creation security.
	Data discovery and classification	Discovery/classification is possible across media, populations, and channels.
	Secure data aggregation	Vendor-supplied aggregation services—security practices are opaque.
Transformation → Uses	Privacy-preserving data analytics	Aggregate reporting to content owners.
	Compliance with regulations	PII disclosure issues abound.
	Government access to data and freedom of expression concerns	Various issues; for example, playing terrorist podcast and illegal playback.
Transformation ↔ Data Infrastructure	Data-centric security such as identity/policy-based encryption	Unknown
	Policy management for access control	User, playback administrator, library maintenance, and auditor.
	Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption	Unknown
	Audits	Audit DRM usage for royalties.
Data Infrastructure	Securing data storage and transaction logs	Unknown
	Key management	Unknown
	Security best practices for non-relational data stores	Unknown
	Security against DoS attacks	N/A
	Data provenance	Traceability to the right entities to be preserved. (Additional use case: Wikipedia privacy issues when distributing data to researchers).
General	Analytics for security intelligence	Machine intelligence for unsanctioned use/access.
	Event detection	"Playback" granularity defined.
	Forensics	Subpoena of playback records in legal disputes.

6.2 Nielsen Homescan: Project Apollo

Nielsen Homescan involves family-level retail transactions and associated media exposure using a statistically valid national sample. A general description[27] is provided by the vendor. This project description is based on the 2006 architecture.

Table 3: Mapping Nielsen Homescan to the Security Reference Architecture

RA Component	Security and Privacy Topic	Use Case Mapping
Sources → Transformation	End-point input validation	Device-specific keys from digital sources; receipt sources scanned internally and reconciled to family ID. (Role issues)
	Real-time security monitoring	None
	Data discovery and classification	Classifications based on data sources (e.g., retail outlets, devices, and paper sources).
	Secure data aggregation	Aggregated into demographic crosstabs. Internal analysts had access to PII.
Transformation → Uses	Privacy-preserving data analytics	Aggregated to (sometimes) product-specific, statistically valid independent variables.
	Compliance with regulations	Panel data rights secured in advance and enforced through organizational controls.
	Government access to data and freedom of expression concerns	N/A
Transformation ↔ Data Infrastructure	Data-centric security such as identity/policy-based encryption	Encryption not employed in place; only for data-center-to-data-center transfers. XML (Extensible Markup Language) cube security mapped to Sybase IQ and reporting tools.
	Policy management for access control	Extensive role-based controls.
	Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption	N/A
	Audits	Schematron and process step audits.
Data Infrastructure	Securing data storage and transaction logs	Project-specific audits secured by infrastructure team.
	Key management	Managed by project chief security officer (CSO). Separate key pairs issued for customers and internal users.
	Security best practices for non-relational data stores	Regular data integrity checks via XML schema validation.
	Security against DoS attacks	Industry-standard webhost protection provided for query subsystem.
	Data provenance	Unique.
General	Analytics for security intelligence	No project-specific initiatives.
	Event detection	N/A
	Forensics	Usage, cube-creation, and device merge audit records were retained for forensics and billing.

DRAFT

6.3 Web Traffic Analytics

Visit-level webserver logs are of high-granularity and voluminous. Web logs are correlated with other sources, including page content (buttons, text, and navigation events) and marketing events such as campaigns and media classification.

Table 4: Mapping Web Traffic Analytics to the Security Reference Architecture

RA Component	Security and Privacy Topic	Use Case Mapping
Sources → Transformation	End-point input validation	Device-dependent. Spoofing is often easy.
	Real-time security monitoring	Webserver monitoring.
	Data discovery and classification	Some geospatial attribution.
	Secure data aggregation	Aggregation to device, visitor, button, web event, and others.
Transformation → Uses	Privacy-preserving data analytics	IP anonymizing and timestamp degrading. Content-specific opt-out.
	Compliance with regulations	Anonymization may be required for EU compliance. Opt-out honoring.
	Government access to data and freedom of expression concerns	Yes.
Transformation ↔ Data Infrastructure	Data-centric security such as identity/policy-based encryption	Varies depending on archivist; for example, Adobe Omniture.
	Policy management for access control	System- and application-level access controls.
	Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption	Unknown
	Audits	Customer audits for accuracy and integrity are supported.
Data Infrastructure	Securing data storage and transaction logs	Storage archiving—this is a big issue.
	Key management	CSO and applications.
	Security best practices for non-relational data stores	Unknown
	Security against DoS attacks	Standard.
	Data provenance	Server, application, IP-like identity, page point-in-time Document Object Model (DOM), and point-in-time marketing events.
General	Analytics for security intelligence	Access to web logs often requires privilege elevation.
	Event detection	Can infer; for example, numerous sales, marketing, and overall web health events.
	Forensics	See the SIEM use case.

6.4 Health Information Exchange (HIE)

HIE data is aggregated from various data providers, which might include covered entities such as hospitals and contract research organizations (CROs) identifying participation in clinical trials. The data consumers would include emergency room personnel, the CDC, and other authorized health (or other)

organizations. Because any city or region might implement its own HIE, these exchanges might also serve as data consumers and data providers for each other.

Table 5: Mapping to the Security Reference Architecture

RA Component	Security and Privacy Topic	Use Case Mapping
Sources → Transformation	End-point input validation	Strong authentication, perhaps through X.509v3 certificates, potential leverage of SAFE (Secure Access for Everyone) bridge in lieu of general PKI.
	Real-time security monitoring	Validation of incoming records to ensure integrity through signature validation and to ensure HIPAA privacy through ensuring PHI is encrypted. May need to check for evidence of informed consent.
	Data discovery and classification	Leverage Health Level Seven (HL7) and other standard formats opportunistically, but avoid attempts at schema normalization. Some columns will be strongly encrypted while others will be specially encrypted (or associated with cryptographic metadata) for enabling discovery and classification. May need to perform column filtering based on the policies of the data source or the HIE service provider.
	Secure data aggregation	Clear text columns can be deduplicated, perhaps columns with deterministic encryption. Other columns may have cryptographic metadata for facilitating aggregation and deduplication. Retention rules are assumed, but disposition rules are not assumed in the related areas of compliance.
Transformation → Uses	Privacy-preserving data analytics	Searching on encrypted data and proofs of data possession. Identification of potential adverse experience due to clinical trial participation. Identification of potential professional patients. Trends and epidemics, and co-relations of these to environmental and other effects. Determination of whether the drug to be administered will generate an adverse reaction, without breaking the double blind. Patients will need to be provided with detailed accounting of accesses to, and uses of, their EHR data.
	Compliance with regulations	HIPAA security and privacy will require detailed accounting of access to EHR data. Facilitating this, and the logging and alerts, will require federated identity integration with data consumers.
	Government access to data and freedom of expression concerns	CDC, law enforcement, subpoenas and warrants. Access may be toggled based on occurrence of a pandemic (e.g., CDC) or receipt of a warrant (e.g., law enforcement).

RA Component	Security and Privacy Topic	Use Case Mapping
Transformation ↔ Data Infrastructure	Data-centric security such as identity/policy-based encryption	Row-level and column-level access control.
	Policy management for access control	Role-based and claim-based. Defined for PHI cells.
	Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption	Privacy-preserving access to relevant events, anomalies, and trends for CDC and other relevant health organizations.
	Audits	Facilitate HIPAA readiness and HHS audits.
Data Infrastructure	Securing data storage and transaction logs	Need to be protected for integrity and privacy, but also for establishing completeness, with an emphasis on availability.
	Key management	Federated across covered entities, with the need to manage key life cycles across multiple covered entities that are data sources.
	Security best practices for non-relational data stores	End-to-end encryption, with scenario-specific schemes that respect min-entropy to provide richer query operations without compromising patient privacy.
	Security against DDoS attacks	A mandatory requirement: systems must survive DDoS attacks.
	Data provenance	Completeness and integrity of data with records of all accesses and modifications. This information could be as sensitive as the data and is subject to commensurate access policies.
General	Analytics for security intelligence	Monitoring of informed patient consent, authorized and unauthorized transfers, and accesses and modifications.
	Event detection	Transfer of record custody, addition/modification of record (or cell), authorized queries, unauthorized queries, and modification attempts.
	Forensics	Tamper-resistant logs, with evidence of tampering events. Ability to identify record-level transfers of custody and cell-level access or modification.

6.5 Genetic Privacy

Mapping is under development.

6.6 Pharma Clinical Trial Data Sharing

Under an industry trade group proposal, clinical trial data for new drugs will be shared outside intra-enterprise warehouses. Regulatory submissions commonly exceed "millions of pages."

Table 6: Mapping Pharma Clinical Trial Data Sharing to the Security Reference Architecture

RA Component	Security & Privacy Topic	Use Case Mapping
Sources → Transformation	End-point input validation	Opaque—company-specific.
	Real-time security monitoring	None.

	Data discovery and classification	Opaque—company-specific.
	Secure data aggregation	Third-party aggregator.
Transformation → Uses	Privacy-preserving data analytics	Data to be reported in aggregate but preserving potentially small-cell demographics.
	Compliance with regulations	Responsible developer and third-party custodian.
	Government access to data and freedom of expression concerns	Limited use in research community, but there are possible future public health data concerns. Clinical study reports only, but possibly selectively at the study- and patient-levels.
Transformation ↔ Data Infrastructure	Data-centric security such as identity/policy-based encryption	TBD
	Policy management for access control	Internal roles; third-party custodian roles; researcher roles; participating patients' physicians.
	Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption	TBD
	Audits	Release audit by a third party.
Data Infrastructure	Securing data storage and transaction logs	TBD
	Key management	Internal varies by firm; external TBD.
	Security best practices for non-relational data stores	TBD
	Security against DoS attacks	Unlikely to become public.
	Data provenance	TBD—critical issue.
General	Analytics for security intelligence	TBD
	Event detection	TBD
	Forensics	

6.7 Network Protection

Security Information and Event Management (SIEM) is a family of tools used to defend and maintain networks.

Table 7: Mapping Network Protection to the Security Reference Architecture

RA Component	Security and Privacy Topic	Use Case Mapping
Sources → Transformation	End-point input validation	Software-supplier specific; for example, for Microsoft.[28]
	Real-time security monitoring	---
	Data discovery and classification	Varies by tool, but classified based on security semantics and sources.
	Secure data aggregation	Aggregates by subnet, workstation, and server.
Transformation → Uses	Privacy-preserving data analytics	Platform-specific; for example, Windows groups.
	Compliance with regulations	Applicable, but regulated events are not readily visible to analysts.
	Government access to data and	NSA and FBI have access on demand.

	freedom of expression concerns	
Transformation ↔ Data Infrastructure	Data-centric security such as identity/policy-based encryption	Usually a feature of the operating system.
	Policy management for access control	For example, a Windows group policy for an event log.
	Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption	Vendor and platform-specific.
	Audits	Complex—audits are possible throughout.
Data Infrastructure	Securing data storage and transaction logs	Vendor and platform-specific.
	Key management	Chief Security Officer and SIEM product keys.
	Security best practices for non-relational data stores	TBD
	Security against DoS attacks	N/A
	Data provenance	For example, how to know an intrusion record was actually associated with a specific workstation.
General	Analytics for security intelligence	Feature of current SIEMs
	Event detection	Feature of current SIEMs
	Forensics	Feature of current SIEMs

6.8 Military: Unmanned Vehicle Sensor Data

Unmanned vehicles (drones) and their onboard sensors (e.g., streamed video) can produce petabytes of data that should be stored in nonstandard formats. The U.S. Government is pursuing capabilities to expand storage capabilities for Big Data such as streamed video. For more information, refer to the Defense Information Systems Agency (DISA) large data object contract[29] for exabytes in the DOD private cloud.

Table 8: Mapping Military Unmanned Vehicle Sensor Data to the Security Reference Architecture.

RA Component	Security and Privacy Topic	Use Case Mapping
Sources → Transformation	End-point input validation	Need to secure the sensor (e.g., camera) to prevent spoofing/stolen sensor streams. There are new transceivers and protocols in the DOD pipeline. Sensor streams will include smartphone and tablet sources.
	Real-time security monitoring	Onboard and control station secondary sensor security monitoring.
	Data discovery and classification	Varies from media-specific encoding to sophisticated situation-awareness enhancing fusion schemes.
	Secure data aggregation	Fusion challenges range from simple to complex. Video streams may be used[30] unsecured or unaggregated.
Transformation → Uses	Privacy-preserving data analytics	Geospatial constraints: cannot surveil beyond Universal Transverse Mercator (UTM). Military secrecy: target and point of origin privacy.
	Compliance with regulations	Numerous. There are also standards issues.

	Government access to data and freedom of expression concerns	For example, the Google lawsuit over Street View.
Transformation ↔ Data Infrastructure	Data-centric security such as identity/policy-based encryption	Policy-based encryption, often dictated by legacy channel capacity/type.
	Policy management for access control	Transformations tend to be made within DOD/contractor-devised system schemes.
	Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption	Sometimes performed within vendor-supplied architectures, or by image-processing parallel architectures.
	Audits	CSO and Inspector General (IG) audits.
Data Infrastructure	Securing data storage and transaction logs	The usual, plus data center security levels are tightly managed (e.g., field vs. battalion vs. headquarters).
	Key management	CSO—chain of command.
	Security best practices for non-relational data stores	Not handled differently at present; this is changing in DOD.
	Security against DoS attacks	DOD anti-jamming e-measures.
	Data provenance	Must track to sensor point in time configuration and metadata.
General	Analytics for security intelligence	DOD develops specific field of battle security software intelligence—event driven and monitoring—that is often remote.
	Event detection	For example, target identification in a video stream, infer height of target from shadow. Fuse data from satellite infrared with separate sensor stream.
	Forensics	Used for after action review (AAR)—desirable to have full playback of sensor streams.

6.9 Education: Common Core Student Performance Reporting

Cradle-to-grave student performance metrics for every student are now possible—at least within the K-12 community, and probably beyond. This could include every test result ever administered.

Table 9: Mapping Common Core K–12 Student Reporting to the Security Reference Architecture

RA Component	Security and Privacy Topic	Use Case Mapping
Sources → Transformation	End-point input validation	Application-dependent. Spoofing is possible.
	Real-time security monitoring	Vendor-specific monitoring of tests, test-takers, administrators, and data.
	Data discovery and classification	Unknown
	Secure data aggregation	Typical: Classroom-level.
Transformation → Uses	Privacy-preserving data analytics	Various: For example, teacher-level analytics across all same-grade classrooms.
	Compliance with regulations	Parent, student, and taxpayer disclosure and privacy rules apply.
	Government access to data and freedom of expression concerns	Yes. May be required for grants, funding, performance metrics for teachers, administrators, and districts.

Transformation ↔ Data Infrastructure	Data-centric security such as identity/policy-based encryption	Support both individual access (student) and partitioned aggregate.
	Policy management for access control	Vendor (e.g., Pearson) controls, state-level policies, federal-level policies; probably 20-50 different roles are spelled out at present
	Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption	Unknown
	Audits	Support third-party audits by unions, state agencies, responses to subpoenas.
Data Infrastructure	Securing data storage and transaction logs	Large enterprise security, transaction level controls—classroom to the federal government.
	Key management	CSOs from the classroom level to the national level.
	Security best practices for non-relational data stores	Unknown
	Security against DoS attacks	Standard.
	Data provenance	Traceability to measurement event requires capturing tests at point in time.
General	Analytics for security intelligence	
	Event detection	
	Forensics	

6.10 Sensor Data Storage and Analytics

< To be detailed in later versions of this effort.>

6.11 Cargo Shipping

This use case provides an overview of a Big Data application related to the shipping industry for which standards may emerge in the near future.

Table 10: Mapping Cargo Shipping to the Security Reference Architecture

RA Component	Security and Privacy Topic	Use Case Mapping
Sources → Transformation	End-point input validation	Ensuring integrity of data collected from sensors.
	Real-time security monitoring	Sensors can detect abnormal temperature/environmental conditions for packages with special requirements. They can also detect leaks/radiation.
	Data discovery and classification	---
	Secure data aggregation	Securely aggregating data from sensors.
Transformation → Uses	Privacy-preserving data analytics	Sensor-collected data can be private and can reveal information about the package and geo-information. The revealing of such information needs to preserve privacy.
	Compliance with regulations	---
	Government access to data and freedom of expression concerns	The U.S. Department of Homeland Security may monitor suspicious packages moving

		into/out of the country.
Transformation ↔ Data Infrastructure	Data-centric security such as identity/policy-based encryption	---
	Policy management for access control	Private, sensitive sensor data and package data should only be available to authorized individuals. Third-party companies such as LoJack have low-level access to the data.
	Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption	---
	Audits	---
Data Infrastructure	Securing data storage and transaction logs	Logging sensor data is essential for tracking packages. Sensor data at rest should be kept in secure data stores.
	Key management	For encrypted data.
	Security best practices for non-relational data stores	The diversity of sensor types and data types may necessitate the use of non-relational data stores.
	Security against DoS attacks	---
	Data provenance	Metadata should be cryptographically attached to the collected data so that the integrity of origin and progress can be ensured.
General	Analytics for security intelligence	Anomalies in sensor data can indicate tampering/fraudulent insertion of data traffic.
	Event detection	Abnormal events such as cargo moving out of the way or being stationary for unwarranted periods can be detected.
	Forensics	Analysis of logged data can reveal details of incidents after they occur.

DRAFT

7 Future Directions

This document is the result of 14 weeks of intense collaboration between experts from the industry, government, and academia to make inroads in this complex and sophisticated task.

This effort began with enunciating the scope of S&P requirements for Big Data by categorization and discussion of technologies, actors, and specialized S&P topics. The use cases provided by volunteers and described herein were the basis for abstractions used to develop taxonomies delineated by conceptual and operational classifications, respectively. Subsequently, we develop a security reference architecture and map the use cases to the reference architecture.

It is difficult to align schemas and protocols within implementation in any technical domain, and it is extremely difficult to standardize, align, or provide a mapping between terms across a variety of application and technology domains. This effort is a beginning. To effectively communicate across domains, there is a need to continually socialize the usage of terms related to security and compliance.

DRAFT

Appendix A. Specialized Security and Privacy Topics

The following set of topics was initially adapted from the scope of the Cloud Security Alliance (CSA) BDWG charter and organized according to the classification in CSA BDWG's *Top 10 Challenges in Big Data Security and Privacy*.[31] Security and privacy concerns are classified in four categories:

- Infrastructure Security
- Data Privacy
- Data Management
- Integrity and Reactive Security

Rather than a prescriptive document at this stage, this text reproduces community discussion. Later versions will refine and organize as needed.

Infrastructure Security

- Review of technologies and frameworks that have been primarily developed for performance, scalability, and availability; for example, Apache Hadoop, Massively Parallel Processing (MPP) databases, and others.
- High-availability
    - Security against denial-of-service (DoS) attacks.

Data Privacy

- The impact of the social data revolution on the security and privacy of Big Data implementations. [*discussion*]
    - Unknowns of innovation – When a perpetrator, abuser, or stalker misuses technology to target and harm a victim, there are various criminal and civil charges that might be applied to ensure accountability and promote victim safety. A number of U.S. federal and state/territory/tribal laws might apply. To support the safety and privacy of victims, it is important to take technology-facilitated abuse and stalking seriously. This includes assessing all ways that technology is being misused to perpetrate harm, and considering all charges that could or should be applied.
    - Identify laws that address violence and abuse. Identify where they explicitly or implicitly include the use of technology and electronic communications:
        - Stalking and cyberstalking (e.g., felony menacing by, via electronic surveillance, and others).
        - Harassment, threats, and assault.
        - Domestic violence, dating violence, sexual violence, and sexual exploitation.
        - Sexting and child pornography: electronic transmission of harmful information to minors, providing obscene material to a minor, inappropriate images of minors, and lascivious intent.
        - Bullying and cyberbullying.
        - Child abuse.
    - Identify possible criminal or civil charges related to technology, communications, privacy, and confidentiality:
        - Unauthorized access, unauthorized recording/taping, illegal interception of electronic communications, illegal monitoring of communications, surveillance, eavesdropping, wiretapping, and unlawful party to call.
        - Computer and internet crimes: fraud and network intrusion.
        - Identity theft, impersonation, and pretexting.
        - Financial fraud and telecommunications fraud.
        - Privacy violations.

        - Consumer protection laws.
        - Violation of no contact, protection, and restraining orders.
        - Technology misuse: Defamatory libel, slander, economic or reputational harms, and privacy torts.
        - Burglary, criminal trespass, reckless endangerment, disorderly conduct, mischief, and obstruction of justice.
- Data-centric security to protect data no matter where it is stored or accessed (e.g., attribute-based encryption and format-preserving encryption).
- Big data privacy and governance.
    - Data discovery and classification.
    - Policy management for accessing and controlling Big Data (e.g., new policy language frameworks specific to Big Data architectures).
    - Data masking technologies: Anonymization, rounding, truncation, hashing, and differential privacy;
        - It is important to consider how these approaches degrade performance or hinder delivery all together. Often these solutions are proposed and then cause an outage at the time of the release, forcing the removal of the option. [*discussion*]
    - Data monitoring.
    - Compliance with regulations such as the Health Insurance Portability and Accountability Act (HIPAA), European Union (EU) data protection regulations, Asia-Pacific Economic Cooperation (APEC) Cross-Border Privacy Rules (CBPR) requirements, and country-specific regulations.
        - Regional data stores enable regional laws to be enforced
            - Cybersecurity Executive Order 1998—assumed data and information would remain within the region
        - People-centered design makes the assumption that private-sector stakeholders are operating ethically and respecting the freedoms and liberties of all Americans. [*discussion*]
            - Litigation, including class action suits, could follow increased threats to Big Data security, when compared to other systems.
                - People before profit must be revisited to understand the large number of Executive Orders overlooked. [*discussion*]
                - People before profit must be revisited to understand the large number of domestic laws overlooked. [*discussion*]
            - Indigenous and aboriginal people and the privacy of all associated vectors and variables must be excluded from any Big Data store in any case in which a person must opt in. [*discussion*]
                - All tribal land is an exclusion from any image capture and video streaming or capture.
                - Human rights.
    - Government access to data and freedom of expression concerns.
        - Polls show that U.S. citizens are less concerned about the loss of privacy than Europeans, but both are concerned about data misuse and their inability to govern private- and public-sector use.
            - In Cisco's Internet of Everything—a project directly dependent on Big Data—a survey shows respondents worry over "threats to data (loss) and fear for physical safety."

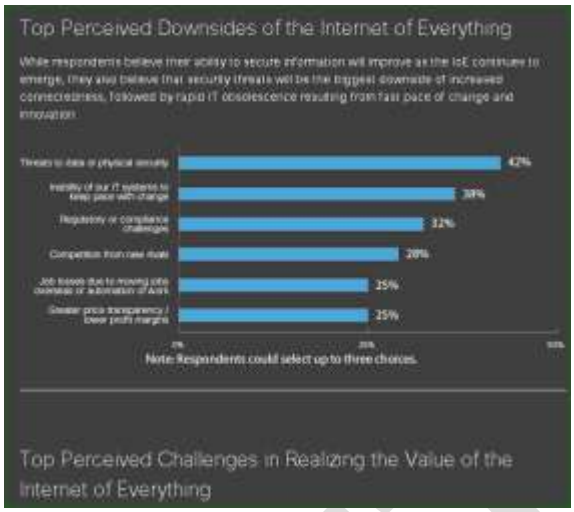

Figure A-1: Top Perceived Downsides of the Internet of Everything.

- Potentially unintended/unwanted consequences or uses
  - Appropriate uses of data collected or data aggregation and problem management capabilities must be enabled [*discussion*]
  - Mechanisms for the appropriate secondary or subsequent data uses
    - Filtered upon entry processed and presented in the inbound framework
- Issues surrounding permission to collect data, consent, and privacy
  - If Facebook or Google permissions are marked "ONLY MY FRIENDS," "ONLY ME," or "ONLY MY CIRCLES," the assumption must be that the person believes that setting in Facebook and Google controls all content presented through Google and Facebook. How should this problem be addressed? Is it a Big Data issue? [*discussion*]
    - Permission based on clear language and not forced by preventing users to access their online services
    - People do not believe the government would allow businesses to take advantage of their rights [*discussion*]
- Data deletion: Responsibility to purge data based on certain criteria and/or events
  - Examples include legal rulings that affect an external data source. For example, if Facebook were to lose a legal challenge and required to purge its databases of certain private information. Is there then a responsibility for downstream data stores to follow suit and purge their copies of the same data? The provider, producer, collector or social media supplier, or host absolutely must inform and remove all versions. Enforcement? Verification? *[discussion]*
- Computing on encrypted data
  - Deduplication of encrypted data
  - Searching and reporting on the encrypted data
  - Fully homomorphic encryption
  - Anonymization of data (no linking fields to reverse identify)
  - De-identification of data (individual centric)
  - Non-identifying data (individual and context centric)
- Secure data aggregation
- Data loss prevention
- Fault tolerance—recovery for zero data loss

        - Aggregation in end-to-end scale of resilience, record, and operational scope for integrity and privacy in a secure or better risk management strategy
        - Fewer applications will require fault tolerance with clear distinction around risk and scope of the risk

Data Management

- Securing data stores
    - Communication protocols
        - Database links (DBLINKS)
        - Access control list (ACL)
        - Application programming interface (API)
        - Channel segmentation
        - Federated (eRate) migration to cloud
    - Attack surface reduction
- Key management and ownership of data
    - Providing full control of the keys to the data owner
    - Transparency of data life cycle process: Acquisition, uses, transfers, dissemination, and destruction
    - Maps to aid non-technical people determine who is using their data and how their data is being used, including custody over time

INTEGRITY AND REACTIVE SECURITY

- Big Data analytics for security intelligence (identifying malicious activity) and situational awareness (understanding the health of the system)
    - Large-scale analytics
        - The largest audience with a "true" competency to make use of large-scale analytics is no more than 5% of the private sector
        - Need assessment of the public sector
    - Streaming data analytics
        - This could require, for example, segregated virtual machines and secure channels
        - This is a low-level requirement (e.g., Phase iii)
            - Roadmap
            - Priority of security and return on investment must be done to move to this degree of maturity
- Event detection
    - Respond to data risk events trigger by application-specific analysis of user and system behavior patterns
    - Data-driven abuse detection
- Forensics
- Security of analytics results

Appendix B: Internal Security Practices: Internal Access Control Rules for General Industry

Any strategy to achieve Access Control & Security (AC&S) within a Big Data, cloud ecosystem enterprise architecture for industry must address the complexities associated with cloud-specific security requirements triggered by the cloud characteristics, including, but limited to, the following:

- Broad network access.
- Decreased visibility and control by consumer.
- Dynamic system boundaries and comingled roles/responsibilities between consumers and providers.
- Multi-tenancy.
- Data residency.
- Measured service.
- Order-of-magnitude increases in scale (on demand), dynamics (elasticity and cost optimization), and complexity (automation and virtualization).

These cloud computing characteristics often present different security risks to an agency than the traditional information technology solutions, altering the agency's security posture.

To preserve the security-level post migration of their data to the cloud, agencies need to identify all cloud-specific risk-adjusted security controls or components in advance and request from the cloud service providers through contractual means and service-level agreements to have all identified security components and controls fully and accurately implemented.

The complexity of multiple interdependencies is best illustrated by Figure 1.

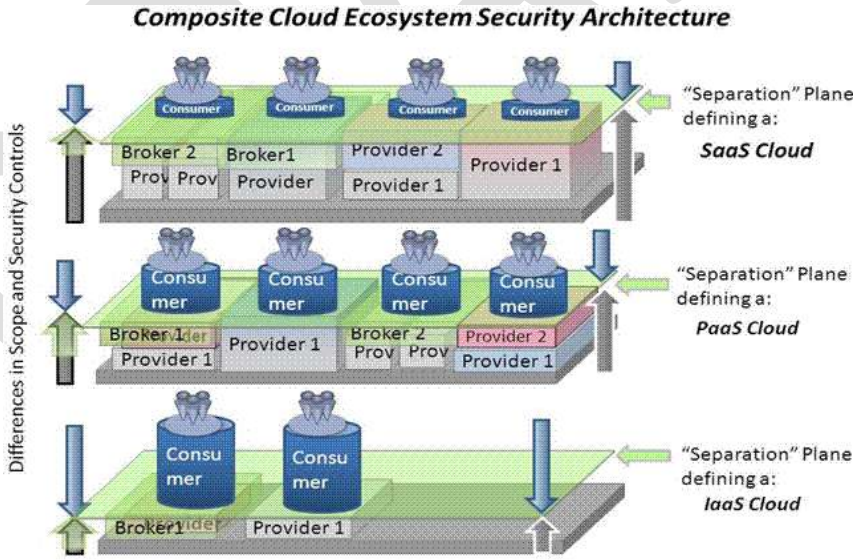

Figure B.1: Composite Cloud Ecosystem Security Architecture.

When unraveling the complexity of multiple interdependencies, it is important to note that enterprise-wide access controls fall within the purview of a well-thought-out Big Data and cloud ecosystem risk management strategy for end-to-end enterprise AC&S, via the following five constructs:

1. Categorize the data value and criticality of information systems and the data custodian's duties and responsibilities to the organization, demonstrated by the data custodian's choice of either a

discretionary access control policy or a mandatory access control policy that is more restrictive; this choice is determined by addressing the specific organizational requirements, such as, but not limited to the following:
   a. Laws and regulatory obligations.
   b. Directives, policy guidelines, strategic goals and objectives, information security requirements, priorities, and resources available (filling in any gaps).
2. Select the appropriate level of security controls required to protect data and to defend information systems
3. Implement access security controls and modify them upon analysis assessments.
4. Authorize appropriate information systems.
5. Monitor access security controls at a minimum of once a year.

To meet the GRC and confidentiality, integrity, and availability regulatory obligations required from the responsible data custodians—and which are directly tied to demonstrating a valid, current, and up-to-date AC&S policy—one of the better strategies is to implement a layered approach to AC&S, comprised of multiple access control gates, including, but not limited to, the following infrastructure AC&S via:

- Physical security/facility security, equipment location, power redundancy, barriers, security patrols, electronic surveillance, and physical authentication.
- Information Security and residual risk management
- Human resources (HR) security, including, but not limited to, employee codes of conduct, roles and responsibilities, job descriptions, and employee terminations.
- Database, end point, and cloud monitoring.
- Authentication services management/monitoring.
- Privilege usage management/monitoring.
- Identify management/monitoring.
- Security management/monitoring.
- Asset management/monitoring.

The following section, Access Control, revisits the traditional access control framework. The traditional framework identifies a standard set of attack surfaces, roles and tradeoffs. These principles are an important part of the CISSP Body of Knowledge. Adopting this framework for Big Data is a reasonable goal for later versions of this NIST effort.

Access Control

Access control is one of the most important areas of Big Data. There are multiple entities, such as mandates, policies, and laws that govern the access of data. The overarching rule is that the highest classification of any data element or string governs the protection of the data. In addition, access should only be granted on a need-to-know/-use basis that is reviewed periodically in order to control the access.

Access control for Big Data covers more than accessing data. The security of the account that is used for access needs to be considered. Most accounts are shared between different systems and environments; therefore, the possibility and opportunity that access control can be compromised is ever present. Data can be accessed via multiple channels, networks, and platforms—including laptops, cell phones, smart phones, tablets, and even fax machines—that are connected to internal networks, remote mobile networks, the internet, or all of the above. With this reality in mind, the same data may be accessed by a user, administrator, another system, etc., and it may be accessed via a remote connection/access point as well as internally via unsecured ports. Therefore, knowing who is accessing the data is critical in protecting same. The trade-offs between strict data access control versus conducting business requires answers to the following questions:

- How important/critical is the data to the life blood and sustainability of the organization?

- What are you responsible for (e.g., all nodes, components, boxes, and machines within the Big Data/cloud ecosystem)?
- Where are they located?
- Who should have access to them?
- within the organization

Very restrictive measures to control accounts are difficult to implement, much less maintain, so this strategy can be considered impractical in most cases. However, there are best practices, such as protection based on classification of the data, least privilege, three-tier authentication, and separation of duties that can help reduce the risks.

The following measures are often included in Best Practices lists for security and privacy. Some require adaptation or expansion for Big Data systems.

- Least privilege—access to data within a Big Data/cloud ecosystem environment should be based on providing an individual with the minimum access rights and privileges to perform his/her job (no more/no less).
- If one of the data elements is protected because of its classification (e.g., PII, HIPAA, payment card industry [PCI]), then all of the data that it is sent with it inherits that classification, retaining the original data's security classification. If the data is joined to and/or associated with other data that may cause a privacy issue, then all data should be protected; this requires due diligence on the part of the data custodian(s) to ensure that this secure and protected state remains throughout the entire end-to-end data flow.
- If data is accessed from, transferred to, or transmitted to the cloud, internet, or another external entity, then the data should be protected based on its classification.
- There should be an indicator/disclaimer on the display of the user if private or sensitive data is being accessed or viewed.
- All accounts (except for system-related accounts) should be reviewed annually, at a minimum, to ensure that they are still required.
- All accounts (except for system-related accounts) that have not been used within 180 days should be deleted. If the system will not allow deletion of an account, then the account should be disabled.
- Access to PII data should be logged. The minimum logging requirements should be a timestamp and the account number.
- Role-based access to Big Data should be based on roles. Each role should be assigned the fewest privileges needed to perform the functions of that role.
- Roles should be reviewed at least every two years to ensure that they are still valid and that the accounts assigned to them are still valid.

User Access Controls

- Each user should have his or her personal account. Shared accounts should not be
- A user account should not be a multipurpose account. For example, a user account should not be used as an administrative account or to run production jobs.

System Access Controls

- There should not be shared accounts in cases of system-to-system access
- Access for a system that contains Big Data needs to be approved by the data owner or his/her representative. The representative should not be the system administrator, because that may cause a separation of duties issue.
- The same type of data stored on different systems should use the same classifications and rules for access controls to ensure that it has the same level of protection.

Administrative Account Controls

- System administrators should maintain a separate user account that is not used for administrative purposes. In addition, an administrative account should not be used as a user account.
- The same administrative account should not be used for access to the production and non-production (e.g., test, development, and quality assurance) systems.

DRAFT

Appendix C: Terms and Definitions

Actors

Big Data systems can comprise simple machine-to-machine "actors," or complex combinations of persons and machines that are systems of systems.

A common meaning of "actor" assigns roles to a person in a system. From a citizen's perspective, a person can have relationships with many applications and sources of information in a Big Data system.

The following list describes a number of roles as well as how roles can shift over time. For some systems, roles are only valid for a specified point in time.

- A retail organization refers to a person as a consumer or prospect before a purchase; afterwards, the consumer becomes a customer.
- A person has a customer relationship with a financial organization for banking services.
- A person may have a car loan with a different organization or the same financial institution.
- A person may have a home loan with a different bank or the same bank.
- A person may be "the insured" on health, life, auto, homeowners, or renters insurance.
- A person may be the beneficiary or future insured person by a payroll deduction in the private sector, or via the employment development department in the public sector.
- A person may have attended one or more public or private schools.
- A person may be an employee, temporary worker, contractor, or third-party employee for one or more private or public enterprises.
- A person may be underage and have special legal or other protections.
- One or more of these roles may apply concurrently.

For each of these roles, system owners should ask themselves whether users can achieve the following:

- Identify which systems their PII has entered.
- Identify how, when, and what type of de-identification process was applied.
- Verify integrity of their own data and correct errors, omissions, and inaccuracies.
- Request to have information purged and have an automated mechanism to report and verify removal.
- Participate in multilevel opt-out systems, such as will occur when Big Data systems are federated.
- Verify that data has not crossed regulatory (e.g., age-related), governmental (e.g., a state or nation), or expired ("I am no longer a customer") boundaries.

DRAFT

Appendix D: Acronyms

DRAFT

Appendix E: References

GENERAL RESOURCES

DOCUMENT REFERENCES

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

[2] http://www.emc.com/leadership/digital-universe/iview/executive-summary-a-universe-of.htm

[3] http://www.emc.com/leadership/digital-universe/iview/information-security-2020.htm

[4] Cloud Security Alliance, *Expanded Top Ten Big Data Security and Privacy Challenges*, April 2013. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf

[5] Health Information Technology for Economic and Clinical Health (HITECH) Act, Pub. L. No. 111-5, 123 Stat. 226 (2009), accessed December 1, 2013, www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/hitechact.pdf.

[6] Cloud Security Alliance Big Data Working Group, *Top 10 Challenges in Big Data Security and Privacy* (Singapore: Cloud Security Alliance, 2012).

[7] Cloud Security Alliance Big Data Working Group, *Top 10 Challenges in Big Data Security and Privacy* (Singapore: Cloud Security Alliance, 2012).

[8] Cloud Security Alliance Big Data Working Group, *Top 10 Challenges in Big Data Security and Privacy* (Singapore: Cloud Security Alliance, 2012).

[9] S. S. Sahoo, A. Sheth, and C. Henson, "Semantic provenance for eScience: Managing the deluge of scientific data," *Internet Computing, IEEE* 12, no. 4 (2008): 46–54, http://dx.doi.org/10.1109/MIC.2008.86.

[10] P. Chen, B. Plale, and M. S. Aktas, "Temporal representation for mining scientific data provenance," *Future Generation Computer Systems*, in press, http://dx.doi.org/10.1016/j.future.2013.09.032.

[11] U.S. Army, "Distributed Common Ground System – Army," U.S. Army, http://dcgsa.apg.army.mil/.

[12] B. Boehm, J. A. Lane, S. Koolmanojwong, and R. Turner, *The Incremental Commitment Spiral Model: Principles and Practices for Successful Systems and Software*. (Indianapolis, IN: Addison-Wesley Professional, 2014).

[13] Linda Fishman, Correspondence from American Hospital Association to Patrick Gallagher (NIST), December 11, 2013, http://assets.fiercemarkets.com/public/newsletter/fiercehealthit/131211-cl-cybersecurity.pdf.

[14] Cloud Security Alliance Big Data Working Group, *Top 10 Challenges in Big Data Security and Privacy* (Singapore: Cloud Security Alliance, 2012).

[15] M. Kokar and G. Ng, "High-level information fusion and situation awareness," *Information Fusion* 10, no. 1 (2009): 2–5. http://dx.doi.org/10.1016/j.inffus.2008.05.003.

[16] S. Auer, "The emerging web of linked data," in *Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications* (New York: Association for Computing Machinery, 2011), http://dx.doi.org/10.1145/1980822.1980823.

[17] Phrma.org, "Principles for Responsible Clinical Trial Data Sharing" (Washington, DC: PhRMA, July 18, 2013), http://phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf.

[18] U.S. Army, *Army Regulation 25-2* (Ft. Belvoir, VA: U.S. Army Publishing Directorate, October 27, 2007), www.apd.army.mil/jw2/xmldemo/r25_2/main.asp.

[19] Lisa Fleisher, "Before Tougher State Tests, Officials Prepare Parents," *Wall Street Journal*, April 15, 2013, http://blogs.wsj.com/metropolis/2013/04/15/before-tougher-state-tests-officials-prepare-parents/.

[20] Debra Donston-Miller, "Common Core Meets Aging Education Technology," *InformationWeek*, July 22, 2013, www.informationweek.com/big-data/news/common-core-meets-aging-education-techno/240158684.

[22] Civitas Learning, "About," Civitas Learning, www.civitaslearning.com/about/.

[23] Consider following some schemes in the NIST Preliminary Critical Infrastructure Cybersecurity Framework (CIICF) of October 2013. A taxonomy is implicit, though the framework contains a reference to the lack of taxonomies of Big Data (**see Section C.5**).

[24] Section references: SAML (2005), Security Token Service (WS-Trust STS), CERT Taxonomy of Operational Cybersecurity Risks (2010). **[unclear what these refer to]**

[25] For an example, see Appendix B of *Taxonomy of Operational Cyber Security Risks* (James J. Cebula and Lisa R. Young, *Taxonomy of Operational Cyber Security Risks*, CMU/SEI-2010-TN-028 [Pittsburgh, PA: Carnegie Mellon University, Software Engineering Institute, December 2010]).

[26] Microsoft, "Deploying Windows Rights Management Services at Microsoft," Microsoft, http://technet.microsoft.com/en-us/library/dd277323.aspx.

[27] Nielsen, "Consumer Panel and Retail Measurement," Nielsen, www.nielsen.com/us/en/nielsen-solutions/nielsen-measurement/nielsen-retail-measurement.html.

[28] Microsoft, "How to set event log security locally or by using Group Policy in Windows Server 2003," Microsoft, http://support.microsoft.com/kb/323076.

[29] Kathleen Hickey, "DISA plans for exabytes of drone, satellite data," *GCN*, April 12, 2013, http://gcn.com/articles/2013/04/12/disa-plans-exabytes-large-data-objects.aspx.

[30] DefenseSystems, "UAV video encryption remains unfinished job," *DefenseSystems*, October 31, 2012, http://defensesystems.com/articles/2012/10/31/agg-drone-video-encryption-lags.aspx.

[31] Cloud Security Alliance Big Data Working Group, *Top 10 Challenges in Big Data Security and Privacy* (Singapore: Cloud Security Alliance, 2012).