**NIST Special Publication XXX-XXX**

**DRAFT NIST Big Data Interoperability Framework:**

**Volume 4, Security and Privacy**

NIST Big Data Public Working Group

Security and Privacy Subgroup

Draft Version 1

January 16, 2015

http://dx.doi.org/10.6028/NIST.SP.XXX



NIST Special Publication xxx-xxx

Information Technology Laboratory

**DRAFT NIST Big Data Interoperability Framework:**

**Volume 4, Security and Privacy**

**Version 1**

NIST Big Data Public Working Group (NBD-PWG)

Security and Privacy Subgroup

National Institute of Standards and Technology

Gaithersburg, MD 20899

November 2014



U. S. Department of Commerce

*Penny Pritzker, Secretary*

National Institute of Standards and Technology

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

Authority

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

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

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

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

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

**Comments on this publication may be submitted to:**

National Institute of Standards and Technology

Attn: Information Technology Laboratory

100 Bureau Drive (Mail Stop 8900) Gaithersburg, MD 20899-8930

Reports on Computer Systems Technology

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

**National Institute of Standards and Technology Special Publication XXX-series**

xxx pages (June 2, 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  Michele Drgon, Data Probity  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  Ann Racuya-Robbins, World Knowledge Bank  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.

Table of Contents

[Executive Summary vii](#_Toc409131989)

[1 Introduction 1](#_Toc409131990)

[1.1 Background 1](#_Toc409131991)

[1.2 Scope and Objectives of the Security and Privacy Subgroup 2](#_Toc409131992)

[1.3 Report Production 2](#_Toc409131993)

[1.4 Report Structure 3](#_Toc409131994)

[1.5 Future Work of this Volume 3](#_Toc409131995)

[2 Big Data Security and Privacy 4](#_Toc409131996)

[2.1 Overview 4](#_Toc409131997)

[2.2 Effects of Big Data Characteristics on Security and Privacy 6](#_Toc409131998)

[2.2.1 Variety 6](#_Toc409131999)

[2.2.2 Volume 6](#_Toc409132000)

[2.2.3 Velocity 6](#_Toc409132001)

[2.2.4 Veracity 6](#_Toc409132002)

[2.2.5 Volatility 7](#_Toc409132003)

[2.3 Relation to Cloud 7](#_Toc409132004)

[3 Example Use Cases for Security and Privacy 9](#_Toc409132005)

[3.1 Retail/Marketing 9](#_Toc409132006)

[3.1.1 Consumer Digital Media Usage 9](#_Toc409132007)

[3.1.2 Nielsen Homescan: Project Apollo 10](#_Toc409132008)

[3.1.3 Web Traffic Analytics 10](#_Toc409132009)

[3.2 Healthcare 11](#_Toc409132010)

[3.2.1 Health Information Exchange 11](#_Toc409132011)

[3.2.2 Genetic Privacy 11](#_Toc409132012)

[3.2.3 Pharma Clinical Trial Data Sharing 12](#_Toc409132013)

[3.3 Cybersecurity 13](#_Toc409132014)

[3.3.1 Network Protection 13](#_Toc409132015)

[3.4 Government 14](#_Toc409132016)

[3.4.1 Military: Unmanned Vehicle Sensor Data 14](#_Toc409132017)

[3.4.2 Education: Common Core Student Performance Reporting 14](#_Toc409132018)

[3.5 Industrial: Aviation 15](#_Toc409132019)

[3.5.1 Sensor Data Storage and Analytics 15](#_Toc409132020)

[3.6 Transportation 15](#_Toc409132021)

[3.6.1 Cargo Shipping 15](#_Toc409132022)

[4 Taxonomy of Security and Privacy Topics 17](#_Toc409132023)

[4.1 Conceptual Taxonomy of Security and Privacy Topics 17](#_Toc409132024)

[4.1.1 Privacy 17](#_Toc409132025)

[4.1.2 Provenance 18](#_Toc409132026)

[4.1.3 System Health and Resilience 19](#_Toc409132027)

[4.2 Operational Taxonomy of Security and Privacy Topics 19](#_Toc409132028)

[4.2.1 Registration, Security Model, and Policy Enforcement 20](#_Toc409132029)

[4.2.2 Identity and Access Management 20](#_Toc409132030)

[4.2.3 Data Governance 21](#_Toc409132031)

[4.2.4 Visibility and Infrastructure Management 22](#_Toc409132032)

[4.2.5 Risk and Accountability 23](#_Toc409132033)

[4.3 Roles Taxonomy of Security and Privacy Topics 23](#_Toc409132034)

[4.3.1 Infrastructure Management 24](#_Toc409132035)

[4.3.2 GRC 24](#_Toc409132036)

[4.3.3 Information Worker 24](#_Toc409132037)

[4.4 Relationships Between Security and Privacy Concepts and Roles 25](#_Toc409132038)

[4.4.1 Data Privacy 25](#_Toc409132039)

[4.4.2 Data Veracity (Provenance) 25](#_Toc409132040)

[4.4.3 System Health 25](#_Toc409132041)

[4.5 Uncategorized Topics 26](#_Toc409132042)

[4.5.1 Provisioning, Metering, and Billing 26](#_Toc409132043)

[4.5.2 Data Syndication 26](#_Toc409132044)

[5 Security and Privacy Fabric 27](#_Toc409132045)

[5.1 Interface of Data Providers Big Data Application Provider 29](#_Toc409132046)

[5.2 Interface of Big Data Application Provider Data Consumer 29](#_Toc409132047)

[5.3 Interface of Application Provider Big Data Framework Provider 30](#_Toc409132048)

[5.4 Internal to Big Data Framework Provider 30](#_Toc409132049)

[5.5 System Orchestrator 30](#_Toc409132050)

[5.6 Privacy by Design 30](#_Toc409132051)

[5.7 General Considerations 30](#_Toc409132052)

[5.8 Relation of the Big Data Security Operational Taxonomy to the NBDRA 30](#_Toc409132053)

[5.8.1 Conceptual Taxonomy 30](#_Toc409132054)

[5.8.2 Security Operational Taxonomy 30](#_Toc409132055)

[5.8.3 Roles Taxonomy 32](#_Toc409132056)

[6 Mapping Use Cases to NBDRA 33](#_Toc409132057)

[6.1 Consumer Digital Media Use 33](#_Toc409132058)

[6.2 Nielsen Homescan: Project Apollo 34](#_Toc409132059)

[6.3 Web Traffic Analytics 35](#_Toc409132060)

[6.4 Health Information Exchange (HIE) 36](#_Toc409132061)

[6.5 Genetic Privacy 38](#_Toc409132062)

[6.6 Pharma Clinical Trial Data Sharing 38](#_Toc409132063)

[6.7 Network Protection 39](#_Toc409132064)

[6.8 Military: Unmanned Vehicle Sensor Data 40](#_Toc409132065)

[6.9 Education: Common Core Student Performance Reporting 42](#_Toc409132066)

[6.10 Sensor Data Storage and Analytics 42](#_Toc409132067)

[6.11 Cargo Shipping 43](#_Toc409132068)

[Appendix A: Candidate Security and Privacy Topics for Big Data Adaptation A-1](#_Toc409132069)

[Appendix B: Internal Security Considerations within Cloud Ecosystems B-1](#_Toc409132070)

[Appendix C: Big Data Actors and Roles: Adaptation to Big Data Scenarios C-1](#_Toc409132071)

[Appendix D: Acronyms D-1](#_Toc409132072)

[Appendix E: References E-1](#_Toc409132073)

Figures

[Figure 1: Cargo Shipping Scenario 16](#_Toc409131964)

[Figure 2: Security and Privacy Conceptual Taxonomy 17](#_Toc409131965)

[Figure 3: Security and Privacy Operational Taxonomy 20](#_Toc409131966)

[Figure 4: NIST Big Data Reference Architecture 28](#_Toc409131967)

[Figure 5: Notional Security and Privacy Fabric Overlay to the NBDRA 29](#_Toc409131968)

[Figure A-1: Top Perceived Downsides of the Internet of Everything. A-3](#_Toc409131969)

[Figure B-1: Composite Cloud Ecosystem Security Architecture B-2](#_Toc409131970)

Tables

[Table 1: Draft Security Operational Taxonomy Mapping to the NBDRA Components 31](#_Toc409131977)

[Table 2: Mapping Consumer Digital Media Usage to the Reference Architecture 33](#_Toc409131978)

[Table 3: Mapping Nielsen Homescan to the Reference Architecture 34](#_Toc409131979)

[Table 4: Mapping Web Traffic Analytics to the Reference Architecture 35](#_Toc409131980)

[Table 5: Mapping HIE to the Reference Architecture 36](#_Toc409131981)

[Table 6: Mapping Genetic Privacy to the Reference Architecture 38](#_Toc409131982)

[Table 7: Mapping Pharma Clinical Trial Data Sharing to the Reference Architecture 39](#_Toc409131983)

[Table 8: Mapping Network Protection to the Reference Architecture 39](#_Toc409131984)

[Table 9: Mapping Military Unmanned Vehicle Sensor Data to the Reference Architecture 40](#_Toc409131985)

[Table 10: Mapping Common Core K–12 Student Reporting to the Reference Architecture 42](#_Toc409131986)

[Table 11: Mapping Sensor Data Storage and Analytics to the Reference Architecture 42](#_Toc409131987)

[Table 12: Mapping Cargo Shipping to the Reference Architecture 43](#_Toc409131988)

Executive Summary

This ***NIST Big Data Interoperability Framework: Volume 4, Security and Privacy*** document was prepared by the NIST Big Data Public Working Group (NBD-PWG) 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 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 *NIST Big Data Interoperability Framework* consists of seven volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. In addition to this volume, the other volumes are as follows:

* 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 in the future and with the availability of additional perspectives.

# Introduction

## Background

There is broad agreement among commercial, academic, and government leaders about the remarkable potential of Big Data to spark innovation, fuel commerce, and drive progress. Big Data is the common term used to describe the deluge of data in 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]](#endnote-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.

The *Draft* *NIST Big Data Interoperability Framework* containsthe following seven volumes:

* Volume 1, Definitions
* Volume 2, Taxonomies
* Volume 3, Use Case and General Requirements
* Volume 4, Security and Privacy Requirements (this volume)
* Volume 5, Architectures White Paper Survey
* Volume 6, Reference Architectures
* Volume 7, Technology Roadmap Summary

## 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 NIST Big Data Reference Architecture (NBDRA)

The Subgroup produced a working draft of

* 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

## 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
* Preliminary mapping of identified security and privacy use cases to the NBDRA

## Report Structure

The remainder of this document is organized as follows:

* Section 2 discusses security and privacy issues particular to Big Data
* Section 3 presents examples of security and privacy related use cases
* Section 4 offers a preliminary taxonomy for security and privacy
* Section 5 introduces the details of a draft NIST Big Data security and privacy reference architecture in relation to the overall NBDRA
* Section 6 maps the use cases presented in Section 3 to the reference architecture
* Section 7 discusses the future directions
* 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 and

## Future Work of this Volume

The NBD-PWG Security and Privacy Subgroup plans to further develop several topics for the subsequent version (Version 2) of this document. These topics include the following:

* A closer examination of other templates. These templates may be adapted to the Big Data security and privacy fabric to address gaps in Version 1 and to bridge the efforts of this Subgroup with the work of others.
* Further developing the Security and Privacy Taxonomy
* Enhancing the connection between the Security and Privacy Taxonomy and the NBDRA components.
* Developing the connection between the Security and Privacy fabric and the NBDRA.
* Expanding the privacy discussion within the scope of Volume 4
* Exploring governance with respect to security and privacy
* Mapping the identified security and privacy use cases to the NBDRA
* Contextualize the content of Appendix B in the NBDRA

# Big Data Security and Privacy

The NBD-PWG Security and Privacy Subgroup discussed security and privacy issues particular to Big Data. From these discussions, a number of ways that security and Privacy in Big Data projects can be different from traditional implementations were identified. While not all concepts apply all of the time, these seven principles are believed to be representative of a larger set of differences:

* Big Data projects often encompass heterogeneous components in which a single security scheme has not been designed from the outset.
* 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.
* 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.
* An increased reliance on sensor streams, such as those anticipated with the Internet of Things (IoT; e.g., smart medical devices, smart cities, smart homes) can create vulnerabilities that were more easily managed before amassed to Big Data scale.
* 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.
* 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.
* 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.

## Overview

Security and privacy measures are becoming ever more important as the generation and utilization of Big Data increase, and as the data storage and availability 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]](#endnote-2). Less than a third of data needed protection in 2010 but more than 40% of data will need protection in 2020[[3]](#endnote-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]](#endnote-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.

There is large body of work in security and privacy spanning decades of academic study and commercial solutions. Much of that work is not conceptually distinct from Big Data, yet may have been produced under different assumptions. Sometimes these assumptions were explicit, sometimes not. Accordingly, the subgroup concluded that one of its objectives is to understand how Big Data security concerns arise out of the defining characteristics of Big Data, and how these concerns are differentiated from traditional security concerns.

What follows is not an exhaustive list of what’s new in Big Data systems. Instead it is a representative list of differences from the concerns that informed earlier big systems security and privacy.

* **Big Data may be gathered from diverse end points.** Actors include more types than just traditional 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. This alone is not new, but the mix of human and device types is on a scale that is unprecedented. The resulting combination of available protection mechanisms and threat vectors for both privacy and security is new.
* **Data aggregation and dissemination must be secured inside the context of a formal, understandable framework.** The availability of data and its current status to data consumers is an important aspect of Big Data, but Big Data systems may be fully operational outside formal, readily understood frameworks, such as those designed by a single team of architects with a clearly defined set of objectives. In some settings, where such frameworks are absent or have been unsystematically joined, there may be a need for public or closed-garden portals and ombudsman-like roles for data at rest. These system combinations and unforeseen combinations call for a renewed Big Data framework.
* **Data search and selection can lead to privacy or security policy concerns.** It is unclear what capabilities are provided by a data provider[[5]](#footnote-1) in this respect. A combination of user competency and system protections may be needed, including the exclusion of databases that can be foreseen as enabling re-identification. If a key feature of Big Data is, as one analyst called it, “the ability to derive differentiated insights from advanced analytics on data at any scale,”[[6]](#endnote-5) the search and selection aspects of analytics will accentuate security and privacy concerns.
* **Privacy-preserving mechanisms are needed for Big Data, such as for Personally Identifiable Information (PII).** Because there may be disparate, potentially unanticipated 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.
* **Big Data is pushing beyond traditional definitions for information trust, openness and responsibility.** Governance, previously consigned to static roles typically found in larger organizations, is increasingly an intrinsic design consideration for Big Data systems.
* **Information Assurance (IA) and Disaster Recovery (DR) for Big Data Systems may require distinctly different practices.** Because of its extreme scalability, Big Data IA and DR present challenges that were not previously addressed in a systematic way. Traditional backup methods, for example, may be impractical for Big Data systems. Test, verification and provenance assurance for Big Data replicas, for example, may not complete in time to meet temporal requirements that were readily accommodated in smaller systems.
* **Big Data creates potential targets of increased value.** The effort required to consummate system attacks will be scaled to meet the opportunity value presented by targets. Big Data targets may represent concentrated, high value targets to adversaries. As Big Data becomes ubiquitous, such targets become more numerous, which is itself a new information technology scenario.

## Effects of Big Data Characteristics on Security and Privacy

Variety, volume, and velocity are key characteristics of Big Data and commonly referred to as the 3 V’s 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 volatility. These elements are discussed below with respect to their impact on Big Data security and privacy.

### 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[[7]](#endnote-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.

### Volume

The volume of Big Data describes how much data is coming in. In Big Data parlance, this typically ranges from gigabytes to exabytes. As a result, 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 cataloging threat models and a surveying of novel techniques. The threat model for network-based, distributed, auto-tier systems include the following major scenarios[[8]](#endnote-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 help detect security breach events. This is an instance where Big Data technologies can fortify security. This document addresses both facets of Big Data security.

### Velocity

Velocity describes the speed at which data is processed. The data usually arrives in batches or is streamed continuously. As with certain other non-relational databases, distributed programming frameworks such as Hadoop were not developed with security in mind.[[9]](#endnote-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.

### Veracity

Big Data Veracity and Validity encompass several sub-characteristics.

Provenance—or what some have called veracity, in keeping with the “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,[[10]](#endnote-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.

**Veracity** (in some circles also called Provenance, though the two terms are not identical) 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 is 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.

**Validity** refers to the accuracy and correctness of data. Traditionally this referred to data quality. In the Big Data security scenario, validity refers to a host of assumptions about data from which analytics are being applied. For example, continuous and discrete measurements have different properties. The field “gender” can be coded as 1=Male, 2=Female, but 1.5 does not mean halfway between male and female. In the absence of such constraints, an analytical tool can make inappropriate conclusions. There are many types of validity whose constraints are far more complex. By definition, Big Data allows for aggregation and collection across disparate data sets in ways not envisioned by system designers.

Several examples of “invalid” uses for Big Data have been cited. Click fraud has been cited[[11]](#endnote-10) as the cause of perhaps $11.6 billion in wasted ad spending. Despite initial enthusiasm, some trend producing applications that use social media to predict the incidence of flu has been called into question. A study by Lazer et. al.[[12]](#endnote-11) suggested that one application overestimated the prevalence of flu for 100 of 108 weeks studied. Careless interpretation of social media is possible when attempts are made to characterize or even predict consumer behavior using imprecise meanings and intentions for “like” and “follow.”

These examples show that what passes for “valid” Big Data can be innocuously lost in translation, interpretation or intentionally corrupted to malicious intent.

### 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),[[13]](#endnote-12) 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.

## Relation to Cloud

Many, though not all, Big Data systems will be designed using cloud architectures. Any strategy to achieve Access Control and 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, organizations 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.

Detailed descriptions can be found in Appendix B. In future versions of this document we plan to contextualize the content of Appendix B in the NBDRA.

# Example Use Cases for Security and Privacy

There are significant Big Data challenges in science and engineering. Many of these are convincingly presented use cases in *NIST Big Data Interoperability Framework: Volume 3, Use Cases and Requirements*. However, the use cases identified were unintentionally skewed toward science and engineering applications for which security and privacy were secondary concerns—if the latter had any impact on system architecture at all. Consequently, a different set of use cases was developed specifically to expose issues ripe for security and privacy discussions. Some of these use cases are no longer active, legacy applications, but were selected because they represent characteristic security / privacy design patterns. The use cases selected for security and privacy are presented in the following subsections. The groupings of the use cases (e.g., Retail/Marketing) were created based on the use cases received and do not necessarily represent the entire spectrum of industries affected by Big Data security and privacy.

## Retail/Marketing

### 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 patterns to 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, including smartphone GPS (global positioning system) or location, browser use, social media, and apps on smart devices

* 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) techniques were not built to scale to meet demand for the forecasted use for the data. “Digital Rights Management (DRM) refers to a broad category of access control technologies aimed at restricting the use and copy of digital content on a wide range of devices.”[[14]](#endnote-13) DRM can be compromised, diverted to unanticipated purposes, defeated, or fail to operate in Big Data V environments—especially Velocity and aggregated Volume.

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

### 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. Additional analytics was possible through the use of 3rd party software. 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 was traditional access security with group policy, implemented at the field level using the database 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.

### 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.

## Healthcare

### 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 by the NBD PWG:

* **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

### 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:
* Transparent, logged, policy-governed controls over access to genetic information
* Full lifecycle data ownership and custody controls

### Pharma Clinical Trial Data Sharing[[15]](#endnote-14)

**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
* The PhRMA mentions anonymization (re-identification), but mentions issues with small sample sizes
* Study-level data disclosure—elective, per company

## Cybersecurity

### 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 (e.g., 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).

## Government

### Military: Unmanned Vehicle Sensor Data

**Scenario Description**: Unmanned vehicles (or 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 U.S. Federal Bureau of Investigation (FBI)
* For overseas surveillance—multiple agencies, including the U.S. 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.[[16]](#endnote-15)
* 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.

### 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[[17]](#endnote-16).
* According to some reports,[[18]](#endnote-17) 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[[19]](#endnote-18) 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[[20]](#endnote-19) predictive analytics for student decision making.

## Industrial: Aviation

### 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 (e.g., 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 (e.g., device identity management)

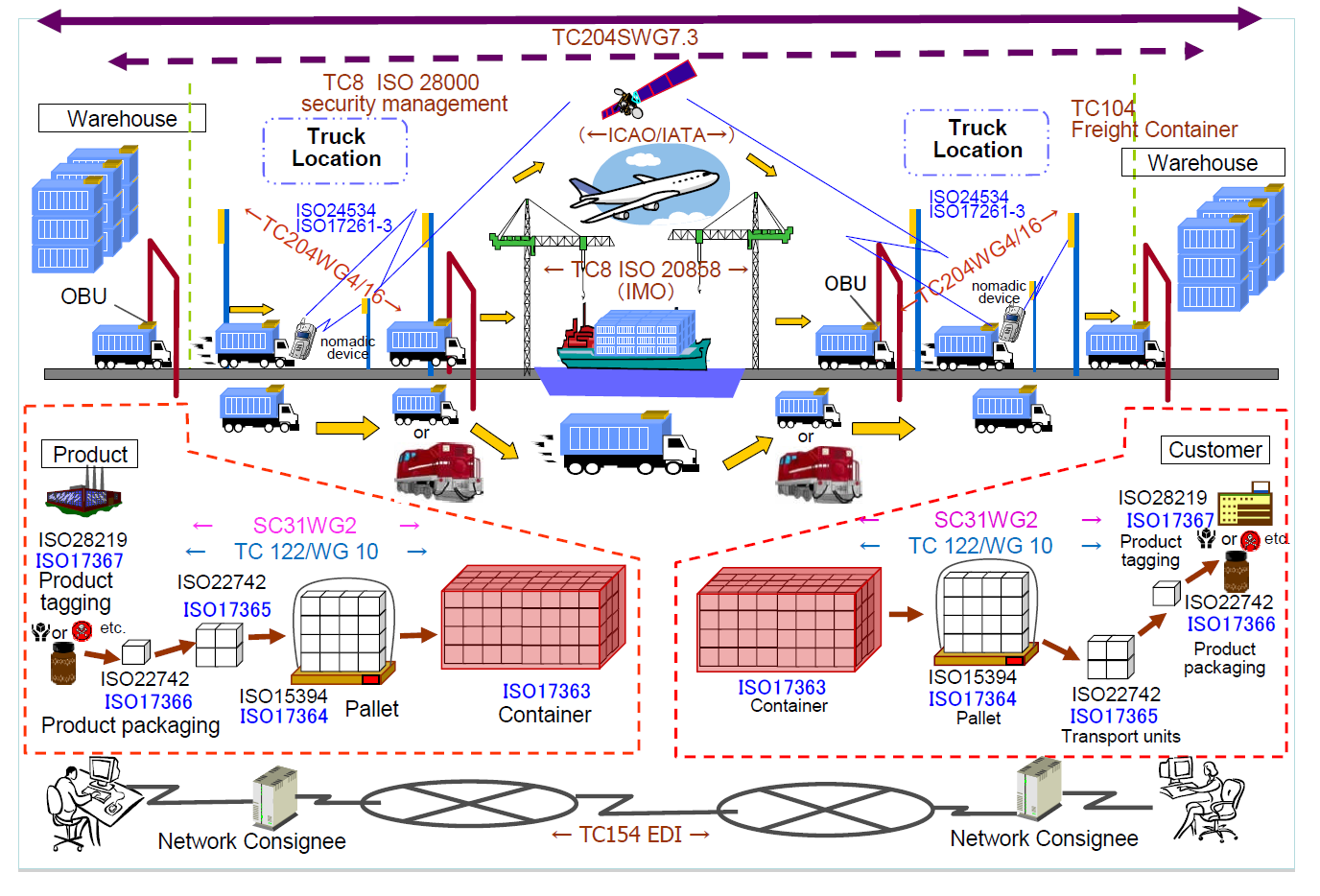
Researchers are currently investigating the following security enhancements:

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

## Transportation

### 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 system-determined number of days.

Figure 1: Cargo Shipping Scenario

# Taxonomy of Security and Privacy Topics

In developing the Security and Privacy Taxonomies, we started with a candidate set of topics from the CSA BDWG article *Top Ten Challenges in Big Data Security and Privacy Challenges[[21]](#endnote-20)*. These candidate topics and the Working Group discussion on the topics are provided in Appendix A as a reference.

## Conceptual Taxonomy of Security and Privacy Topics

The conceptual taxonomy, presented in Table 2, identifies three main considerations: privacy, provenance and system health. This has broad correspondence with the traditional classification of Confidentiality, Integrity and Availability (CIA), retargeted to parallel considerations in Big Data

Figure 2: Security and Privacy Conceptual Taxonomy

### 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
* As noted by Chandramouli and Iorga[[22]](#endnote-21), cloud security for cryptographic keys, an essential building block for security and privacy, takes on "additional complexity," which can be rephrased for Big Data settings: (1) greater Variety due to more cloud consumer-provider relationships, and (2) greater demands and Variety of infrastructures "on which both the Key Management System and protected resources are located.”
* Big Data systems are not purely cloud systems, but as is noted elsewhere in this document, the two are closely related. The security and privacy fabric proposed must be applied in what Chandramouli and Iorga identify as cloud service models (IaaS, PaaS, SaasS) and deployment modes (i.e., public, private, community and hybrid.)
* Challenges for Big Data Key Management Systems (KMS) reflect demands imposed by Big Data V's. For example, leisurely key creation and workflow associated with legacy—and often fastidious—data warehouse key creation is insufficient for Big Data systems deployed quickly and scaled up using massive resources. The lifetime for a Big Data KMS will likely outlive the period of employment of the Big Data system architects who designed it. Designs for location, scale, ownership, custody, provenance and audit for Big Data key management is an aspect of a security and privacy fabric.

### 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 a 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 is valid. 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.
* Digital signatures will be very important in the Big Data system.
* 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

### 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

## Operational Taxonomy of Security and Privacy Topics

Current practice for securing Big Data systems is diverse, employing widely disparate approaches often not part of a unified conceptual framework. The practical methods listed in Table 3 are classified as “operational” because they address specific vulnerabilities or risk management challenges. At this point in the standards development process, these techniques are not part of a cohesive security fabric. They are potentially valuable checklist-style elements that can solve specific security or privacy needs.

In the proposed taxonomy, broad considerations of the conceptual taxonomy 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]](#endnote-22)

The taxonomy will overlap with small data taxonomies while drawing attention to specific issues with Big Data.[[24]](#endnote-23) [[25]](#endnote-24)

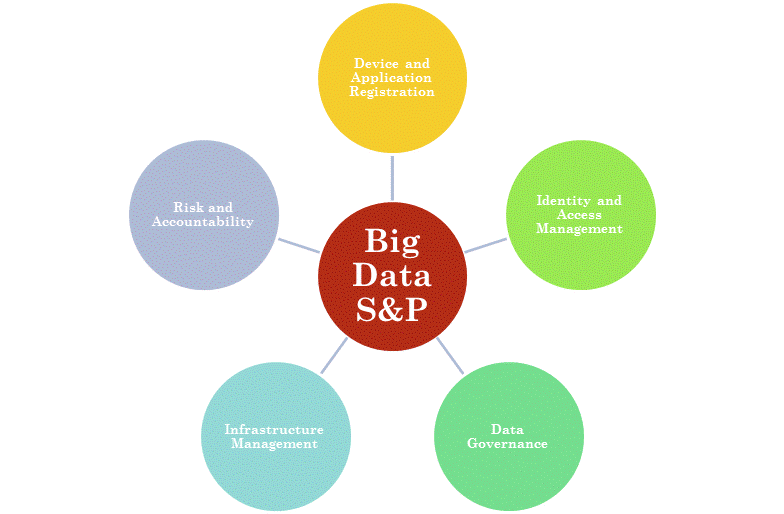


Figure 3: Security and Privacy Operational Taxonomy

### **Registration, Security Model, and Policy Enforcement**

* Device, User, Asset, Services, and Applications Registration: Includes registration of devices in Machine to Machine (M2M) and 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

### **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
* IdP:
* An IdP is defined in the Security Assertion Markup Language.[[26]](#endnote-25) 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

### **Data Governance**

However large and complex Big Data becomes in terms of data volume, velocity, variety and variability, Big Data Governance will in some important conceptual and actual dimensions be much larger. Data Governance will need to persist across the data lifecycle; at rest, in motion, in incomplete stages and transactions all the while serving the privacy and security of the young and the old, individuals as companies and companies as companies—to be an emergent force for good. It will need to insure economy, and innovation; enable freedom of action and individual and public welfare. It will need to rely on standards governing things we do not yet know while integrating the human element from our humanity with strange new interoperability capability. Data Governance will require new kinds and possibilities of perception yet accept that our current techniques are notoriously slow. For example, even as of today we have not yet scoped-in data types.

Big Data without Big Data Governance will be less likely to be a force for good. It may come to be said that the best use of Big Data is Big 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” [[27]](#endnote-26)
* De-identification and anonymization
* Re-identification risk management
* End-Point Validation
* DRM
* Trust
* Openness
* Fairness and Information Ethics [[28]](#endnote-27)

### **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:
* Establishes a data-agnostic connection for a secure channel
* Shared services network architecture, such as those specified as “secure channel use cases and requirements” in the ETSI TS 102 484 Smart Card[[29]](#endnote-28) specification.
* 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:
  + Redundancy within Big Data systems presents challenges at different levels. Replication to maintain intentional redundancy within a Big Data system takes place at one software level. At another level, entirely redundant systems designed to support failover, resilience or reduced data center latency may be more difficult due to Velocity, Volume or other aspects of Big Data.
* 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.

### 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.[[30]](#endnote-29) Business risk models can incorporate privacy considerations.

## Roles Taxonomy of Security and Privacy Topics

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.

### Infrastructure Management

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 infrastructure technology personnel work across organizational boundaries, they accommodate diverse technologies, infrastructures, and workflows and the integration of these three elements. 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 infrastructure technology usually faces pressure when problems arise.

### GRC

Typically, GRC is a function that draws participation from multiple areas of the organization, such as legal, human resources (HR), information technology (IT), and compliance. However, increasingly, GRC departments have their own heads. In some industries and agencies, there may be a strong focus on compliance, often in isolation from technologies. Big Data exacerbates this.

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

GRC within an organization is under pressure to protect the company from risks that might arise from loss of intellectual property, 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.

GRC responsibilities often entail roles assigned in legal, marketing, accounting departments or staff positions connected to the CIO. Internal and external auditors are often involved.

Smaller organizations may create, own or process Big Data, yet may not have GRC systems and practices in place. This is a new scenario. Prior to Big Data, GRC roles in smaller organizations received little attention.

### Information Worker

Information workers 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.

Information workers 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 data by becoming a service provider, or optimizing and enhancing business by consuming third-party data.

## Relationships Between Security and Privacy 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.

### 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.

### Data Veracity (Provenance)

Data veracity (or provenance) is similar to data privacy, but it might introduce information workers as ratifiers because businesses may need to protect their intellectual property 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 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 Ordering Agreement (BOA), 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 EHRs and presumably has signed a BOA.[[31]](#footnote-2)

### 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 to maintain system health would be shared across the organizations 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.

System Health is an umbrella concept that emerges at the intersection of information worker and infrastructure management. As with human health, monitoring nominal conditions for Big Data systems may produce Big Data Volume and Velocity – to cite two of the V’s. Following the human health analogy, some of those potential signals reflect defensive measures such as white cell count. Others could reflect compromised health, such as high blood pressure. Similarly, Big Data systems may employ applications like Security Information and Event Management (SIEM) or Big Data analytics more generally to monitor system health.

What is different about system health for Big Data? Volume, Variety and Velocity of Big Data systems health make it different. Existing systems health tools and design patterns are likely insufficient to handle Big Data – including Big Data security and privacy. At least one commercial web services provider has reported that its internal accounting and systems management tool uses more resources than any other single application. The Volume of system events and the complexity of event interactions is a challenge that demands Big Data solutions to defend Big Data systems.

For example, one aspect motivated by the DevOps movement is the rapid launch, reconfiguration, redeployment and distribution of Big Data systems. Tracking intended vs. accidental or malicious configuration changes is increasingly a Big Data challenge.

## 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.

### 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. Information workers 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. GRC considerations, such as audit and recovery, may overlap with provisioning and metering.

### 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, information workers and IT professionals are candidate ratifiers and stakeholders, For some domains, more complex models may be required to accommodate PII, provenance and governance. Syndication involves transfer of risk and responsibility for security and privacy.

# Security and Privacy Fabric

Security and privacy considerations are a fundamental aspect of the NBDRA. Brainstorming sessions were carried out over two years with the participants in the Subgroup to create the preliminary lists of security and privacy topics. The resulting proposal was for a security and privacy fabric, which is described in the next paragraph.

**Security and Privacy Fabric:** Security and Privacy considerations form a fundamental aspect of the NBDRA. This is geometrically depicted by having a Security and Privacy fabric around and through the reference architecture components, since it touches all of the components. This way the role of Security and Privacy is depicted in the right relation to the components and at the same time does not explode into finer details, which may be more accurate but are best relegated to a more detailed Security Reference Architecture. In addition to the Application and Framework Providers, we also decided to include the Data Provider and Data Consumer into the fabric since at the least they have to agree on the security protocols and mechanisms in place. The Security and Privacy Fabric is an approximate representation that alludes to the intricate interconnected-ness and ubiquity of Security and Privacy in the NBDRA, while projecting a broad abstraction level in coherence with the rest of the architectural elements.

The concept of a security and privacy fabric has precedent in the hardware world, where the notion of a “fabric” of interconnected nodes in a distributed computing environment was introduced. Computing fabrics were invoked as part of cloud computing, grid computing as well as commercial offerings from both hardware and software manufacturers.

Figure 4: NIST Big Data Reference Architecture



The Subgroup is keenly aware that the explanations as to how the proposed fabric concept is implemented across each NBDRA component are cursory. They are more suggestive than prescriptive. Despite this drawback, the Subgroup believes that a template emerges which will form a sound basis for later, more detailed iterations.

This pervasive dimension is depicted in Figure 4 by the presence of the security and privacy fabric surrounding all of the functional components. In addition to the application and framework providers, NBD-PWG decided to include data providers and data consumers into the Security and Privacy fabric because these entities should agree on the security protocols and mechanisms in place. The *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture* document discusses in detail the other components of the NBDRA.

Figure 4 introduces two new concepts that are particular to security and privacy considerations: Information Value Chain and IT Value Chain.

**Information Value Chain:** While it does not apply to all domains, there may be an implied processing progression through which information value is increased, decreased, refined, defined or otherwise transformed. Application of provenance-preservation and other security mechanisms at each stage may be conditioned by the state-specific contributions to information value.

**IT Value Chain** Platform-specific considerations apply to Big Data systems when scaled “up” or “out.” In the process of scaling, specific security, privacy or GRC mechanism or practices may need to be invoked.

Figure 5 provides an overview of several security and privacy topics with respect to some of the components and interfaces. The figure represents a beginning to the elaboration of the interwoven nature of the Security and Privacy Fabric with the NBDRA components. It is not anticipated that Figure 5 will be further developed for Version 2 of this document. However, the relationships between the Security and Privacy Fabric and the NBDRA and the Security and Privacy Taxonomy and the NBDRA will be investigated for Version 2 of this document.

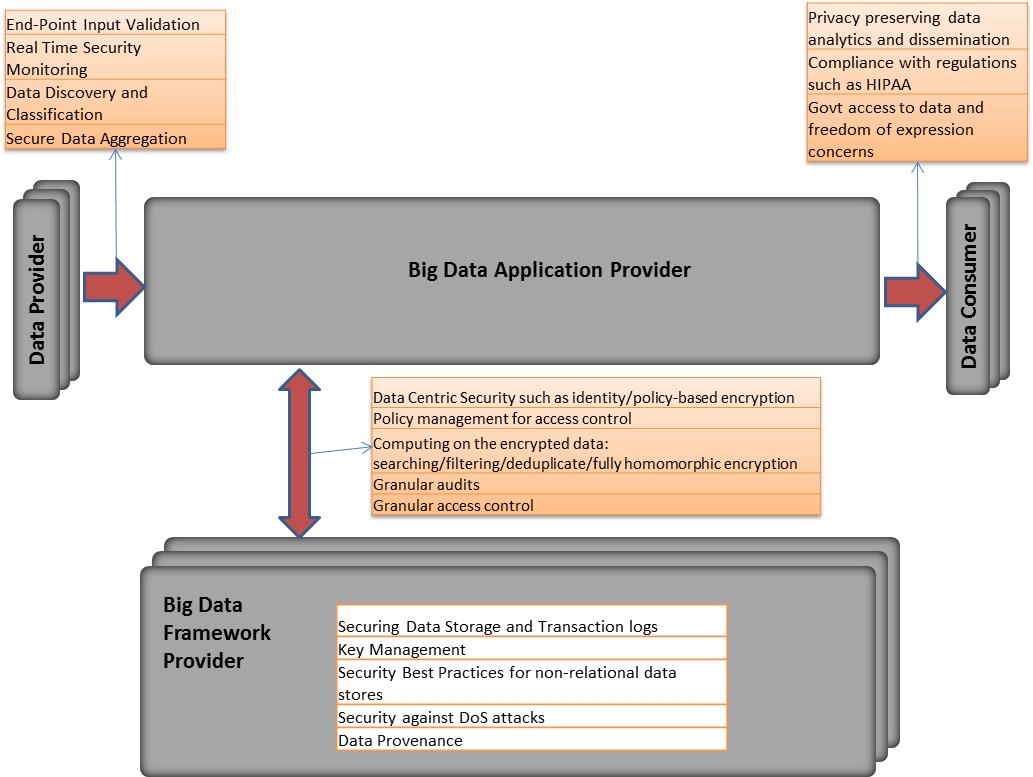


Figure 5: Notional Security and Privacy Fabric Overlay to the NBDRA

## 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.

## 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.

## Interface of Application Provider Big Data Framework 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.

## Internal to Big Data Framework 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.

## System Orchestrator

A System Orchestrator may play a critical role in identifying, managing, auditing and sequencing Big Data processes across the components. For example, a workflow that moves data from a Collection stage to further Preparation may implement aspects of security or privacy.

Orchestrators present an additional, attractive attack surface for adversaries. Orchestrators often require permanent or transitory elevated permissions. Orchestrators present opportunities to both implement security mechanisms, to monitor provenance, to access systems management tools, provide audit points, as well as to inadvertently subjugate privacy or other information assurance measures.

## Privacy by Design

Big Data security and privacy should leverage existing standards and practices. In the privacy arena, the subgroup has identified the foundational principles of [Privacy by Design](http://www.privacybydesign.ca/index.php/about-pbd/7-foundational-principles/) as relevant guidelines to consider when adapting security and privacy practices to Big Data scenarios. At this stage of the subgroup’s efforts, the Privacy by Design template, consisting of seven foundational principles, is identified by the subgroup as potentially helpful, sometimes essential guidance for Big System architects. When working with PII, or with more broadly interpreted

## 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.

## Relation of the Big Data Security Operational Taxonomy to the NBDRA

### Conceptual Taxonomy

### Security Operational Taxonomy

Table Draft Security Operational Taxonomy Mapping to the NBDRA Components

| Activities | Description |
| --- | --- |
| System Orchestrator | |
| * Policy Enforcement * Security Metadata Model * Data Loss Prevention, Detection * Data Lifecycle Management * Threat and Vulnerability Management * Mitigation * Configuration Management * Monitoring, Alerting * Malware Surveillance and Remediation * Resiliency, Redundancy and Recovery * Accountability * Compliance * Forensics * Business Risk Model | Several security functions have been mapped to the System Orchestrator Block as they require architectural level decisions and awareness. Aspects of these functionalities are strongly related to the Security Fabric and thus touch the entire architecture at various points in different forms of operational details.  Such security functions include nation-specific compliance requirements, vastly expanded demand for forensics, and domain-specific, privacy-aware business risk models. |
| Data Provider | |
| * Device, User, Asset, Services, Applications Registration * Application Layer Identity * End User Layer Identity Management * End Point Input Validation * Digital Rights Management * Monitoring, Alerting | Data Providers are subject to guaranteeing authenticity of data and in turn require that sensitive/copyrighted/valuable data is adequately protected. This leads to operational aspects of entity registration and identity ecosystems. |
| Data Consumer | |
| * Application Layer Identity * End User Layer Identity Management * Web Services Gateway * Digital Rights Management * Monitoring, Alerting | Data Consumers exhibit a duality with Data Providers in terms of obligations and requirements – only they face the access/visualization aspects of the Application Provider. |
| Application Provider | |
| * Application Layer Identity * Web Services Gateway * Data Transformation * Digital Rights Management * Monitoring, Alerting | Application Provider interfaces between the Data Provider and Data Consumer. It takes part in all the secure interface protocols with these blocks as well as maintains secure interaction with the Framework Provider. |
| Framework Provider | |
| * Virtualization Layer Identity * Identity Provider * Encryption and Key Management * Isolation/Containerization * Storage Security * Network Boundary Control * Monitoring, Alerting | Framework Provider is responsible for the security of data/computations for a significant portion of the lifecycle of the data. This includes security of data at rest through encryption and access control; security of computations via isolation/virtualization; and security of communication with the Application Provider. |

### Roles Taxonomy

# Mapping Use Cases to NBDRA

In this section, the security and privacy related use cases, presented in Section 3, are mapped to the NBDRA components and interfaces explored in Figure 5.

## Consumer Digital Media Use

Content owners license data for use by consumers through presentation portals. The use of consumer digital media generates 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 Reference Architecture

| NBDRA Component and Interfaces | Security and Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | End-point input validation | Varies and is vendor dependent. Spoofing is possible. For example, protections afforded by securing Microsoft Rights Management Services.[[32]](#endnote-30) 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. |
| Application Provider → Data Consumer | 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. |
| Data Provider ↔  Framework Provider | 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. |
| Framework Provider | 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 data owners, producers, consumers is preserved. |
| Fabric | Analytics for security intelligence | Machine intelligence for unsanctioned use/access. |
| Event detection | “Playback” granularity defined. |
| Forensics | Subpoena of playback records in legal disputes. |

## Nielsen Homescan: Project Apollo

Nielsen Homescan involves family-level retail transactions and associated media exposure using a statistically valid national sample. A general description[[33]](#endnote-31) is provided by the vendor. This project description is based on a 2006 Project Apollo architecture. (Project Apollo did not emerge from its prototype status.)

Table 3: Mapping Nielsen Homescan to the Reference Architecture

| NBDRA Component and Interfaces | Security and Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | 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. |
| Application Provider → Data Consumer | 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 |
| Data Provider ↔  Framework Provider | 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. |
| Framework Provider | 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. |
| Fabric | 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. |

## 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 Reference Architecture

| NBDRA Component and Interfaces | Security and Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | 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. |
| Application Provider → Data Consumer | 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. |
| Data Provider ↔  Framework Provider | Data-centric security such as identity/policy-based encryption | Varies depending on archivist |
| 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. |
| Framework Provider | 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. |
| Fabric | 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. |

## 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 HIE to the Reference Architecture

| NBDRA Component and Interfaces | Security and Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | End-point input validation | Strong authentication, perhaps through X.509v3 certificates, potential leverage of SAFE (Signatures & Authentication for Everything[[34]](#endnote-32)) 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. |
| Application Provider → Data Consumer | 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). |
| Data Provider ↔  Framework Provider | 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. |
| Framework Provider | 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. |
| Fabric | 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. |

## Genetic Privacy

Mapping is under development.

Table 6: Mapping Genetic Privacy to the Reference Architecture

| NBDRA-SP Component | Security and Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | End-point input validation |  |
| Real-time security monitoring |  |
| Data discovery and classification |  |
| Secure data aggregation |  |
| Application Provider → Data Consumer | Privacy-preserving data analytics |  |
| Compliance with regulations |  |
| Government access to data and freedom of expression concerns |  |
| Data Provider ↔  Framework Provider | Data-centric security such as identity/policy-based encryption |  |
| Policy management for access control |  |
| Computing on the encrypted data: searching/ filtering/ deduplicate/ fully homomorphic encryption |  |
| Audits |  |
| Framework Provider | Securing data storage and transaction logs |  |
| Key management |  |
| Security best practices for non-relational data stores |  |
| Security against DoS attacks |  |
| Data provenance |  |
| Fabric | Analytics for security intelligence |  |
| Event detection |  |
| Forensics |  |

## 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 7: Mapping Pharma Clinical Trial Data Sharing to the Reference Architecture

| NBDRA Component and Interfaces | Security & Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | 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. |
| Application Provider → Data Consumer | 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. |
| Data Provider ↔  Framework Provider | 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. |
| Framework Provider | 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. |
| Fabric | Analytics for security intelligence | TBD |
| Event detection | TBD |
| Forensics |  |

## Network Protection

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

Table 8: Mapping Network Protection to the Reference Architecture

| NBDRA Component and Interfaces | Security and Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | End-point input validation | Software-supplier specific; refer to commercially available end point validation.[[35]](#endnote-33) |
| 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. |
| Application Provider → Data Consumer | Privacy-preserving data analytics | Platform-specific. |
| Compliance with regulations | Applicable, but regulated events are not readily visible to analysts. |
| Government access to data and freedom of expression concerns | NSA and FBI have access on demand. |
| Data Provider ↔  Framework Provider | Data-centric security such as identity/policy-based encryption | Usually a feature of the operating system. |
| Policy management for access control | For example, a 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. |
| Framework Provider | 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 DDoS attacks | Big Data application layer DDoS attacks can be mitigated using combinations of traffic analytics, correlation analysis |
| Data provenance | For example, how to know an intrusion record was actually associated with a specific workstation. |
| Fabric | Analytics for security intelligence | Feature of current SIEMs |
| Event detection | Feature of current SIEMs |
| Forensics | Feature of current SIEMs |

## 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[[36]](#endnote-34) for exabytes in the DOD private cloud.

Table 9: Mapping Military Unmanned Vehicle Sensor Data to the Reference Architecture

| NBDRA Component and Interfaces | Security and Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | 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[[37]](#endnote-35) unsecured or unaggregated. |
| Application Provider → Data Consumer | 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. |
| Data Provider ↔  Framework Provider | 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. |
| Framework Provider | 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. |
| Fabric | 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. |

## 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 10: Mapping Common Core K–12 Student Reporting to the Reference Architecture

| NBDRA Component and Interfaces | Security and Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | 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. |
| Application Provider → Data Consumer | 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. |
| Data Provider ↔  Framework Provider | 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 | Proposed [[38]](#endnote-36) |
| Audits | Support both internal and third-party audits by unions, state agencies, responses to subpoenas. |
| Framework Provider | 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 | --- |
| Security against DDoS attacks | Standard. |
| Data provenance | Traceability to measurement event requires capturing tests at a point in time,which may itself require a Big Data platform |
| Fabric | Analytics for security intelligence | Various commercial security applications |
| Event detection | Various commercial security applications |
| Forensics | Various commercial security applications |

## Sensor Data Storage and Analytics

Mapping is under development.

Table 11: Mapping Sensor Data Storage and Analytics to the Reference Architecture

| NBDRA Component and Interfaces | Security and Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | End-point input validation |  |
| Real-time security monitoring |  |
| Data discovery and classification |  |
| Secure data aggregation |  |
| Application Provider → Data Consumer | Privacy-preserving data analytics |  |
| Compliance with regulations |  |
| Government access to data and freedom of expression concerns |  |
| Data Provider ↔  Framework Provider | Data-centric security such as identity/policy-based encryption |  |
| Policy management for access control |  |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption |  |
| Audits |  |
| Framework Provider | Securing data storage and transaction logs |  |
| Key management |  |
| Security best practices for non-relational data stores |  |
| Security against DoS attacks |  |
| Data provenance |  |
| Fabric | Analytics for security intelligence |  |
| Event detection |  |
| Forensics |  |

## 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 12: Mapping Cargo Shipping to the Reference Architecture

| NBDRA Component and Interfaces | Security and Privacy Topic | Use Case Mapping |
| --- | --- | --- |
| Data Provider → Application Provider | 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. |
| Application Provider → Data Consumer | 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. |
| Data Provider ↔  Framework Provider | 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 commercial offerings may implement low-level access to the data. |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | See above section on “Transformation” |
| Audits | --- |
| Framework Provider | 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. Complete preservation of provenance will sometimes mandate a separate Big Data application. |
| Fabric | 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. |

Appendix A: Candidate Security and Privacy Topics for Big Data Adaptation

The following set of topics was initially adapted from the scope of the CSA BDWG charter and organized according to the classification in CSA BDWG’s *Top 10 Challenges in Big Data Security and Privacy*.[[39]](#endnote-37) 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, the text below lists Big Data topics that the NBD-PWG Security and Privacy Subgroup identified to potentially update for Big Data systems. A complete rework of these topics is beyond the scope of this document.

This material will be refined and organized as needed in the future.

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
  + Use of Big Data to enhance defenses against denial-of-service (DDoS) attacks.
* DevOps Security

Data Privacy

* System architects should consider the impact of the social data revolution on the security and privacy of Big Data implementations. Some systems not designed to include social data could be connected to those data systems by third parties, or by other project sponsors within an organization.
  + 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 or 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:
    - 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 laws applicable related to Big Data 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 Security is addressed elsewhere in this document. Data-centric security may be needed to protect certain types of data no matter where it is stored or accessed (e.g., attribute-based encryption and format-preserving encryption). There are domain-specific particulars that should be considered when addressing encryption tools available to system users.
* Big data privacy and governance
  + Data discovery and classification
  + Policy management for accessing and controlling Big Data.
    - Is new policy language frameworks specific to Big Data architectures needed?
  + 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 – *for Big Data systems in particular*. Often these solutions are proposed and then cause an outage at the time of the release, forcing the removal of the option.
  + 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.
      * 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.
        + People before profit must be revisited to understand the large number of domestic laws overlooked.
      * 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.
        + 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.”

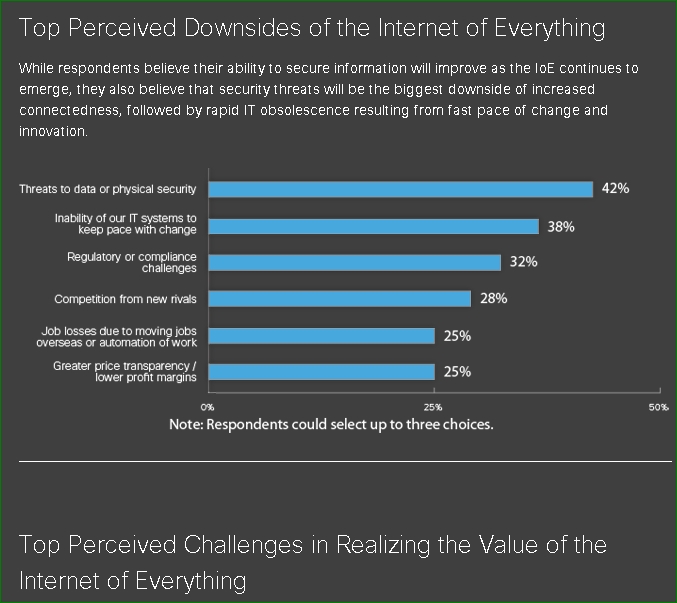


Figure A-1: Top Perceived Downsides of the Internet of Everything.

* + Potentially unintended/unwanted consequences or uses
    - Appropriate uses of data collected or data aggregation and problem management capabilities must be enabled
    - 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?
      * 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
  + 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?
  + 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
    - 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
      * 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 Considerations within Cloud Ecosystems

Many, though not all Big Data systems will be designed using cloud architectures. 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, organizations 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 B-1.

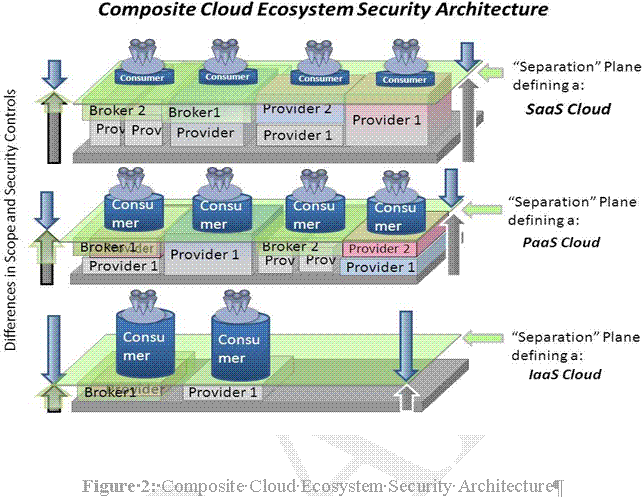
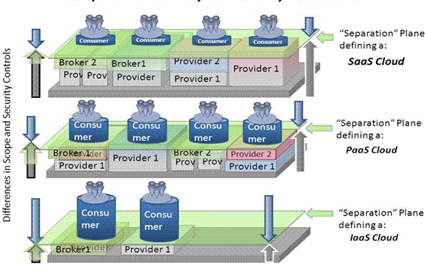


Figure B-1: Composite Cloud Ecosystem Security Architecture[[40]](#endnote-38)

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:
   1. GRC
   2. 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 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 appear in some existing best practices guidelines. For instance, they are an important part of the CISSP Body of Knowledge.[[41]](#endnote-39)

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 factors, 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.. 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, mobile devices, 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. Therefore, visibility as to who is accessing the data is critical in protecting the data. The trade-offs between strict data access control versus conducting business requires answers to questions such as:

* How important/critical is the data to the life blood and sustainability of the organization?
* What is the organization responsible for (e.g., all nodes, components, boxes, and machines within the Big Data/cloud ecosystem)?
* Where are the resources and data located?
* Who should have access to the resources and data?
* Have GRC considerations been given due attention?

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 – perhaps all of them --*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
* 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. Variations on this theme may be required for domain-specific combinations of public and private data hosted by Big Data applications.
* 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. Openness, trust and transparency considerations may require more specific actions, depending on GRC or other broad considerations of how the Big Data system is being used.
* All system roles (“accounts”) should be subjected to periodic meaningful audits to ensure that they are still required.
* All accounts (except for system-related accounts) that have not been used within 180 days should be deactivated.
* Access to PII data should be logged. Role-based access to Big Data should be enforced. Each role should be assigned the fewest privileges needed to perform the functions of that role.
* Roles should be reviewed periodically to ensure that they are still valid and that the accounts assigned to them are still appropriate.

User Access Controls

* Each user should have his or her personal account. Shared accounts should not be the default practice in most settings.
* A user role should match the system capabilities for which it was intended. For example, a user account intended only for information access or to manage an Orchestrator should not be used as an administrative account or to run unrelated production jobs.

System Access Controls

* There should not be shared accounts in cases of system-to-system access. “Meta-accounts” that operate across systems may be an emerging Big Data concern.
* 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 infrastructure support personnel (e.g., a system administrator), because that may cause a separation of duties issue.
* Ideally, 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. In practice, Big Data systems may not follow this practice, and different techniques may be needed to map roles across related but dissimilar components or even across Big Data systems.

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.

Appendix C: Big Data Actors and Roles: Adaptation to Big Data Scenarios

Service-Oriented Architectures (SOA) were a widely discussed paradigm through the early 2000’s. While the concept is employed less often, SOA has influenced systems analysis processes, and perhaps to a lesser extent, systems design. As noted by Patig[[42]](#endnote-40) and Lopez-Sanz et al.[[43]](#endnote-41), actors and roles were incorporated into Unified Modeling Language (UML) so that these concepts could be represented within and well as across services. Big Data calls for further adaptation of these concepts. While actor/role concepts have not been fully integrated into the proposed security fabric, the subgroup felt it important to emphasize to Big Data system designers how these concepts may need to be adapted from legacy and SOA usage.

Similar adaptations[[44]](#endnote-42) from Business Process Execution Language (BPEL), Business Process Model and Notation (BPMN) frameworks offer additional patterns for Big Data security and privacy fabric standards. Baresi et. al (2011) suggest how adaptations might proceed from SOA, but Big Data systems offer somewhat different challenges.

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. Reconsidering temporal aspects of actor security is salient for Big Data systems as some will be architected without explicit archive or deletion policies.

* 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.

Appendix D: Acronyms

ACLs Access Control Lists

AuthN/AuthZ Authentication/Authorization

BAA business associate agreement

CDC U.S. Centers for Disease Control and Prevention

CEP complex event processing

CIA U.S. Central Intelligence Agency

CIICF Critical Infrastructure Cybersecurity Framework

CINDER DARPA Cyber-Insider Threat

CMS U.S. Centers for Medicare & Medicaid Services

CoP communities of practice

CSA Cloud Security Alliance

CSA BDWG Cloud Security Alliance Big Data Working Group

CSP Cloud Service Provider

DARPA Defense Advanced Research Projects Agency’s

DOD U.S. Department of Defense

DoS denial of service

DRM

DRM digital rights management

EFPIA European Federation of Pharmaceutical Industries and Associations

EHRs electronic health records

EU European Union

FBI U.S. Federal Bureau of Investigation

FTC Federal Trade Commission

GPS global positioning system

GRC governance, risk management, and compliance

HIEs Health Information Exchanges

HIPAA Health Insurance Portability and Accountability Act

HITECH Act Health Information Technology for Economic and Clinical Health Act

HR human resources

IdP Identity Provider

IoT internet of things

IP Internet Protocol

IT information technology

LHNCBC Lister Hill National Center for Biomedical Communications

M2M machine to machine

MAC media access control

NBD-PWG NIST Big Data Public Working Group

NBDRA NIST Big Data Reference Architecture

NBDRA-SP NIST Big Data Security and Privacy Reference Architecture

NIEM National Information Exchange Model

NIST National Institute of Standards and Technology

NSA U.S. National Security Agency

OSS operations systems support

PaaS platform as a service

PHI protected health information

PII personally identifiable information

PKI public key infrastructure

SAML Security Assertion Markup Language

SIEM Security Information and Event Management

SKUs stock keeping units

SLAs Service Level Agreements

STS Security Token Service

TLS Transport Layer Security

VM virtual machine

VPN virtual private network

WS web services

XACML eXtensible Access Control Markup Language

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) [↑](#endnote-ref-1)
2. <http://www.emc.com/leadership/digital-universe/iview/executive-summary-a-universe-of.htm> [↑](#endnote-ref-2)
3. <http://www.emc.com/leadership/digital-universe/iview/information-security-2020.htm> [↑](#endnote-ref-3)
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> [↑](#endnote-ref-4)
5. Reference to reference architecture data provider. [↑](#footnote-ref-1)
6. Subgroup correspondence with James G Kobielus (IBM), August 28, 2014 [↑](#endnote-ref-5)
7. Cloud Security Alliance Big Data Working Group, *Top 10 Challenges in Big Data Security and Privacy* (Singapore: Cloud Security Alliance, 2012). [↑](#endnote-ref-6)
8. Cloud Security Alliance Big Data Working Group, *Top 10 Challenges in Big Data Security and Privacy* (Singapore: Cloud Security Alliance, 2012). [↑](#endnote-ref-7)
9. Cloud Security Alliance Big Data Working Group, *Top 10 Challenges in Big Data Security and Privacy* (Singapore: Cloud Security Alliance, 2012). [↑](#endnote-ref-8)
10. 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. [↑](#endnote-ref-9)
11. Taube, A. Botnets will cause $11.6B in wasted ad spending. Business Insider, January 28, 2014, accessed 1/9/15 http://read.bi/1A6Ut9U. [↑](#endnote-ref-10)
12. Lazer, D., Kennedy, R., King, G., Vespignani, A., Mar. 2014. The parable of google flu: Traps in big data analysis. Science 343 (6176), 1203-1205. URL <http://dx.doi.org/10.1126/science.1248506> [↑](#endnote-ref-11)
13. 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. [↑](#endnote-ref-12)
14. Zhang, X., edited by Jain, R. 2011. A survey of digital rights management technologies (CSE571-11). URL <http://bit.ly/1y3Y1P1>, accessed January 9, 2015. [↑](#endnote-ref-13)
15. Phrma.org, “Principles for Responsible Clinical Trial Data Sharing” (Washington, DC: PhRMA, July 18, 2013), http://phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf. [↑](#endnote-ref-14)
16. 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. [↑](#endnote-ref-15)
17. “Cuomo panel: State should cut ties with inBloom”, <http://lohud.us/1mV9U2U> [↑](#endnote-ref-16)
18. 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/. [↑](#endnote-ref-17)
19. 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. [↑](#endnote-ref-18)
20. Civitas Learning, “About,” Civitas Learning, www.civitaslearning.com/about/. [↑](#endnote-ref-19)
21. Cloud Security Alliance Big Data Working Group, *Top 10 Challenges in Big Data Security and Privacy* (Singapore: Cloud Security Alliance, 2012). [↑](#endnote-ref-20)
22. Chandramouli, R., Iorga, M., Chokhani, S., Sep. 2013. Cryptographic key management issues & challenges in cloud services. Tech. Rep. NISTIR 7956, NIST, Gaithersburg MD. URL <http://dx.doi.org/10.6028/NIST.IR.7956> [↑](#endnote-ref-21)
23. There are multiple templates developed by others to adapt as part of a Big Data security metadata model. For instance, the subgroup has considered schemes offered in the NIST Preliminary Critical Infrastructure Cybersecurity Framework (CIICF) of October 2013, <http://1.usa.gov/1wQuti1>, accessed January 9. 2015. [↑](#endnote-ref-22)
24. OASIS “SAML V2.0 Standard,” March 2005, http://bit.ly/1wQByit accessed January 9, 2015.Section references: SAML (2005), Security Token Service (WS-Trust STS), CERT Taxonomy of Operational Cybersecurity Risks (2010). [↑](#endnote-ref-23)
25. Cebula, Young, Dec. 2010. A taxonomy of operational cyber security risks (Technical Note CMU/SEI-2010-TN-028). URL <http://resources.sei.cmu.edu/asset_files/TechnicalNote/2010_004_001_15200.pdf> [↑](#endnote-ref-24)
26. OASIS “SAML V2.0 Standard,” March 2005, http://bit.ly/1wQByit accessed January 9, 2015. Section references: SAML (2005), Security Token Service (WS-Trust STS), CERT Taxonomy of Operational Cybersecurity Risks (2010). [↑](#endnote-ref-25)
27. Kum, H.-C., & Ahalt, S. (2013). Privacy-by-Design: Understanding Data Access Models for Secondary Data. *AMIA Summits on Translational Science Proceedings*, *2013*, 126–130. [↑](#endnote-ref-26)
28. “Justice as Fairness” A Theory of Justice, John Rawls, Information and Computer Ethics, Luciano Floridi, Helen Nissenbaum, and Martha Nussbaum [↑](#endnote-ref-27)
29. ETSI, Dec. 2007. ETSI TS 102 484 smart cards; secure channel between a UICC and an end-point terminal(Release 7). Tech. Rep. DTS/SCP-T0312, ETSI, Sophia Antipolis, France. URL <http://bit.ly/1x2HSUe> [↑](#endnote-ref-28)
30. 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]). [↑](#endnote-ref-29)
31. See BOA for legal definition: http://www.acqnotes.com/Attachments/FAR%2016.703.pdf [↑](#footnote-ref-2)
32. Microsoft, “Deploying Windows Rights Management Services at Microsoft,” Microsoft, http://technet.microsoft.com/en-us/library/dd277323.aspx. [↑](#endnote-ref-30)
33. Nielsen, “Consumer Panel and Retail Measurement,” Nielsen, www.nielsen.com/us/en/nielsen-solutions/nielsen-measurement/nielsen-retail-measurement.html. [↑](#endnote-ref-31)
34. http://www.safe-biopharma.org/ [↑](#endnote-ref-32)
35. Microsoft, “How to set event log security locally or by using Group Policy in Windows Server 2003,” Microsoft, http://support.microsoft.com/kb/323076. [↑](#endnote-ref-33)
36. 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. [↑](#endnote-ref-34)
37. DefenseSystems, “UAV video encryption remains unfinished job,” *DefenseSystems*, October 31, 2012, http://defensesystems.com/articles/2012/10/31/agg-drone-video-encryption-lags.aspx. [↑](#endnote-ref-35)
38. Citation: Fisher, K. A. G., Broadbent, A., Shalm, L. K., Yan, Z., Lavoie, J., Prevedel, R., Jennewein, T., Resch, K. J., Jan. 2014. Quantum computing on encrypted data 5. URL <http://dx.doi.org/10.1038/ncomms4074> [↑](#endnote-ref-36)
39. Cloud Security Alliance Big Data Working Group, *Top 10 Challenges in Big Data Security and Privacy* (Singapore: Cloud Security Alliance, 2012). [↑](#endnote-ref-37)
40. NIST Cloud Computing Reference Architecture, Special Publication 500-292. <http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505> [↑](#endnote-ref-38)
41. https://www.isc2.org/cissp/default.aspx CISSP is a professional computer security certification administered by (ISC)2 [↑](#endnote-ref-39)
42. Patig, S., 2008. Model-Driven development of composite applications. In: Kutsche, R.-D., Milanovic, N. (Eds.), Model-Based Software and Data Integration. Vol. 8 of Communications in Computer and Information Science. Springer Berlin Heidelberg, pp. 67-78+. URL <http://dx.doi.org/10.1007/978-3-540-78999-4_8> [↑](#endnote-ref-40)
43. López-Sanz, M., Acuña, C. J., Cuesta, C. E., & Marcos, E. (2008). Modelling of Service-Oriented Architectures with UML. Electronic Notes in Theoretical Computer Science, 194(4), 23–37. doi:10.1016/j.entcs.2008.03.097 [↑](#endnote-ref-41)
44. Ardagna, D., Baresi, L., Comai, S., Comuzzi, M., Pernici, B., Mar. 2011. A Service-Based framework for flexible business processes. Software, IEEE 28 (2), 61-67. URL <http://dx.doi.org/10.1109/ms.2011.28> [↑](#endnote-ref-42)