NIST Special Publication 1500-4

DRAFT NIST Big Data Interoperability Framework:

Volume 4, Security and Privacy

NIST Big Data Public Working Group
Security and Privacy Subgroup

Draft Version 1
April 6, 2015

http://dx.doi.org/10.6028/NIST.SP.1500-4

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

NIST Special Publication 1500-4
Information Technology Laboratory

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

Draft Version 1

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

April 2015

U. S. Department of Commerce
Penny Pritzker, Secretary

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

National Institute of Standards and Technology Special Publication 1500-4
71 pages (April 6, 2015)

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

Comments on this publication may be submitted to Wo Chang

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

Reports on Computer Systems Technology

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

Abstract

Big Data is a term used to describe the deluge of data in our networked, digitized, sensor-laden, information-driven world. While great opportunities exist with Big Data, it can overwhelm traditional technical approaches and its growth is outpacing scientific and technological advances in data analytics. To advance progress in Big Data, the NIST Big Data Public Working Group (NBD-PWG) is working to develop consensus on important, fundamental questions related to Big Data. The results are reported in the *NIST Big Data Interoperability Framework* series of volumes. This volume, Volume 4, contains an exploration of security and privacy topics with respect to Big Data. This volume considers new aspects of security and privacy with respect to Big Data, reviews security and privacy use cases, proposes security and privacy taxonomies, presents details of the Security and Privacy Fabric of the NIST Big Data Reference Architecture (NBDRA), and begins mapping the security and privacy use cases to the NBDRA.

Keywords

Big Data security, Big Data privacy, Big Data taxonomy, use cases, Big Data characteristics, security and privacy fabric, Big Data risk management, cybersecurity, computer security, information assurance, information security frameworks, encryption standards, role-based access controls, Big Data forensics, Big Data audit

Acknowledgements

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

The document contains input from members of the NBD-PWG 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 SP1500-4, 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 specific contributions[a] to this volume by the following NBD-PWG members:

Pw Carey
Compliance Partners, LLC

Wo Chang
NIST

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.

[a] "Contributors" are members of the NIST Big Data Public Working Group who dedicated great effort to prepare and substantial time on a regular basis to research and development in support of this document.

Notice to Readers

NIST is seeking feedback on the proposed working draft of the *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy*. Once public comments are received, compiled, and addressed by the NBD-PWG, and reviewed and approved by NIST internal editorial board, Version 1 of this volume will be published as final. Three versions are planned for this volume, with Versions 2 and 3 building on the first. Further explanation of the three planned versions and the information contained therein is included in Section 1.5 of this document.

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

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

Submit the edited file from either method 1 or 2 to SP1500comments@nist.gov with the volume number in the subject line (e.g., Edits for Volume 4.)

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

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

Table of Contents

Figures

Tables

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 that are specific to Big Data.

Big Data application domains include health care, drug discovery, insurance, finance, retail 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, targeted marketing 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, velocity, and variability of Big Data and retargeting security solutions amid shifts in technology infrastructure, e.g., distributed computing systems and non-relational data storage. In addition, diverse datasets are becoming easier to access and increasingly contain personal content. A 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, velocity, and variability in mind, the Subgroup gathered use cases from volunteers, developed a consensus-based security and privacy taxonomy, related the taxonomy to the NIST Big Data Reference Architecture (NBDRA), and validated the NBDRA by mapping the use cases to the NBDRA.

The *NIST Big Data Interoperability Framework* consists of seven volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The seven volumes are as follows:

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

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

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

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

1 INTRODUCTION

1.1 BACKGROUND

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

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

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

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

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

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

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

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

The *NIST Big Data Interoperability Framework* consists of seven volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The seven volumes are as follows:

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

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

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

The NBDRA, created in Stage 1 and further developed in Stages 2 and 3, is a high-level conceptual model designed to serve as a tool to facilitate open discussion of the requirements, structures, and operations inherent in Big Data. It is discussed in detail in *NIST Big Data Interoperability Framework: Volume 6, Reference Architecture*. Potential areas of future work for the Subgroup during stage 2 are highlighted in Section 1.5 of this volume. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

1.2 SCOPE AND OBJECTIVES OF THE 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 consensus on a 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.

The scope of the Subgroup's work includes the following topics, some of which will be addressed in future versions of this Volume:

- Provide a context from which to begin Big Data-specific security and privacy discussions
- 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 NBDRA

- Produce a working draft of this Big Data Security and Privacy document
- Develop Big Data security and privacy taxonomies
- Explore mapping between the Big Data security and privacy taxonomies and the NBDRA
- Explore mapping between the use cases and the NBDRA

While there are many issues surrounding Big Data security and privacy, the focus of this Subgroup is on the technology aspects of security and privacy with respect to Big Data.

1.3 REPORT PRODUCTION

The NBD-PWG Security and Privacy Subgroup explored various facets of Big Data security and privacy to develop this document. The major steps involved in this effort included:

- Announce that 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 and privacy fabric of the NBDRA and identify specific topics related to NBDRA components
- Begin mapping of identified security and privacy use cases to the NBDRA

This report is a compilation of contributions from the PWG. Since this is a community effort, there are several topics covered that are related to security and privacy. While an effort has been made to connect the topics, gaps may come to light that could be addressed in Version 2 of this document.

1.4 REPORT STRUCTURE

Following this introductory section, 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 NBDRA
- 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

1.5 FUTURE WORK ON THIS VOLUME

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

- Examining closely other existing templates[b] in literature: The templates may be adapted to the Big Data security and privacy fabric to address gaps 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 this volume
- Exploring governance, risk management, data ownership, and valuation with respect to Big Data ecosystem, with a focus on security and privacy
- Mapping the identified security and privacy use cases to the NBDRA
- Contextualizing the content of Appendix B in the NBDRA
- Exploring privacy in actionable terms with respect to the NBDRA

Further topics and direction may be added, as warranted, based on future input and contributions to the Subgroup, including those received during the public comments period.

[b] 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.)

2 BIG DATA SECURITY AND PRIVACY

The NBD-PWG Security and Privacy Subgroup began this effort by identifying a number of ways that security and Privacy in Big Data projects can be different from traditional implementations. While not all concepts apply all of the time, the following seven principles were considered representative of a larger set of differences:

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

2.1 OVERVIEW

Security and privacy measures are becoming ever more important with the increase of Big Data generation and utilization and increasingly public nature of data storage and availability.

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

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

Big Data system representations that rely on concepts of actors and roles present a different facet to security and privacy. The Big Data systems should be adapted to the emerging Big Data landscape, which is embodied in many commercial and open source access control frameworks. These security approaches will likely persist for some time and may evolve with the emerging Big Data landscape. Appendix C considers actors and roles with respect to Big Data security and privacy.

Big Data is increasingly generated and used across diverse industries such as health care, drug discovery, finance, insurance, and marketing of consumer-packaged goods. Effective communication across these diverse industries will require standardization of the terms related to security and privacy. The NBD-PWG Security and Privacy Subgroup aims to encourage participation in the global Big Data discussion with due recognition to the complex and difficult security and privacy requirements particular to Big Data.

There is a large body of work in security and privacy spanning decades of academic study and commercial solutions. While much of that work is not conceptually distinct from Big Data, it may have been produced using different assumptions. One of the primary objectives of this document is to understand how Big Data security and privacy requirements arise out of the defining characteristics of Big Data, and how these requirements are differentiated from traditional security and privacy requirements.

The following list is a representative—though not exhaustive—list of differences between what is new for Big Data and the requirements that informed previous big system security and privacy.

- **Big Data may be gathered from diverse end points.** Actors include more types than just traditional providers and consumers—data owners, such as mobile users and social network users, are primary actors in Big Data. Devices that ingest data streams for physically distinct data consumers may also be actors. This alone is not new, but the mix of human and device types is on a scale that is unprecedented. The resulting combination of threat vectors and potential protection mechanisms to mitigate them is new.
- **Data aggregation and dissemination must be secured inside the context of a formal, understandable framework.** The availability of data and transparency of its current and past use by data consumers is an important aspect of Big Data. However, Big Data systems may be 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 composed, there may be a need for public or walled 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.** There is a lack of systematic understanding of the capabilities that should be provided by a data provider in this respect.[c] A combination of well-educated users, well-educated architects, and system protections may be needed, as well as excluding databases or limiting queries that may 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,” the search and selection aspects of analytics will accentuate security and privacy concerns. [5]
- **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 privacy and integrity of data coming from end points should be protected at every stage. End-to-end information assurance practices for Big Data 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 and typically employed in larger organizations, is becoming an increasingly important intrinsic design consideration for Big Data systems.
- **Information assurance and disaster recovery for Big Data Systems may require unique and emergent practices.** Because of its extreme scalability, Big Data presents challenges for

[c] Reference to NBDRA Data Provider.

information assurance (IA) and disaster recovery (DR) practices that were not previously addressed in a systematic way. Traditional backup methods may be impractical for Big Data systems. In addition, test, verification, and provenance assurance for Big Data replicas 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. Big Data systems will present concentrated, high value targets to adversaries. As Big Data becomes ubiquitous, such targets are becoming more numerous—a new information technology scenario in itself.
- **Risks have increased for de-anonymization and transfer of PII without consent traceability**. 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 Big Data characteristics, 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. For example, the availability of unanticipated data sets could make re-identification possible. Even when technology is able to preserve privacy, proper consent and use may not follow the path of the data through various custodians.
- **Emerging Risks in Open Data and Big Science.** Data identification, metadata tagging, aggregation, and segmentation—widely anticipated for data science and open datasets—if not properly managed, may have degraded veracity because they are derived and not primary information sources. Retractions of peer-reviewed research due to inappropriate data interpretations may become more commonplace as researchers leverage third party Big Data.

2.2 EFFECTS OF BIG DATA CHARACTERISTICS ON SECURITY AND PRIVACY

Variety, volume, velocity, and variability are key characteristics of Big Data and commonly referred to as the Vs of Big Data. Where appropriate, these characteristics shaped discussions within the NBD-PWG Security and Privacy Subgroup. While the Vs provide a useful shorthand description, used in 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.

2.2.1 VARIETY

Variety describes the organization of the data—whether the data is structured, semi-structured, or unstructured. Retargeting traditional relational database security to non-relational databases has been a challenge[6]. These systems were not designed with security and privacy in mind, and these functions are 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. Therefore, once encryption is applied, any organization of the data that 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.

An emergent phenomenon introduced by Big Data variety that has gained considerable importance is the ability to infer identity from anonymized datasets by correlating with apparently innocuous public databases. While several formal models to address privacy preserving data disclosure have been proposed,[7] [8] in practice, sensitive data is shared after sufficient removal of apparently unique identifiers by the processes of anonymization and aggregation. This is an ad hoc process that is often based on empirical evidence[9] and has led to many instances of de-anonymization in conjunction with publicly available data.[10]

2.2.2 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 includes the following major scenarios: confidentiality and integrity, provenance, availability, consistency, collusion attacks, roll-back attacks and recordkeeping disputes. [11]

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.

2.2.3 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 were not developed with security and privacy in mind.[12] Malfunctioning computing nodes might leak confidential data. Partial infrastructure attacks could compromise a significantly large fraction of the system due to high levels of connectivity and dependency. If the system does not enforce strong authentication among geographically distributed nodes, rogue nodes can be added that can eavesdrop on confidential data.

2.2.4 VERACITY

Big Data veracity and validity encompass several subcharacteristics:

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 groups and communities of interest, and across 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,[13] but no clear guidelines yet exist.

A common understanding holds that provenance data is metadata establishing pedigree and chain of custody, including calibration, errors, missing data (e.g., time stamp, location, equipment serial number, transaction number, and authority.)

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, calibration, version, sampling, and device configuration is needed.

Curation is an integral concept which binds veracity and provenance to principles of governance as well as to data quality assurance, Curation, for example, may improve raw data by fixing errors, filling in gaps, modeling, calibrating values, ordering data collection.

Validity refers to the accuracy and correctness of data. Traditionally this is 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, conducted on a Big Data scale, but which can be detected using Big Data techniques, has been cited as the cause of perhaps $11.6 billion in wasted advertisement spending. A software executive listed seven different types of online ad fraud, including non-human generated impressions, non-human generated clicks, hidden ads, misrepresented sources, all-advertising sites, malicious ad injections, and policy-violating content such as pornography or privacy violations.[14] Each of these can be conducted at Big Data scale and may require Big Data solutions to detect and combat.

Despite initial enthusiasm, some trend producing applications that use social media to predict the incidence of flu have been called into question. A study by Lazer et al.[15] 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.

2.2.5 VOLATILITY

Volatility of data—how data management changes over time—directly affects 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),[16] there are few standards beyond simplistic timestamps and even fewer common practices available as guidance. To manage security and privacy for long-lived Big Data, data temporality should be taken into consideration.

2.3 RELATION TO CLOUD

Many Big Data systems will be designed using cloud architectures. Any strategy to achieve proper access control and security risk management within a Big Data cloud ecosystem enterprise architecture for industry must address the complexities associated with cloud-specific security requirements triggered by cloud characteristics, including, but not limited to, the following:

- Broad network access
- Decreased visibility and control by consumer
- Dynamic system boundaries and commingled roles and 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 organization than the traditional information technology solutions, altering the organization's security posture.

To preserve security when migrating data to the cloud, organizations need to identify all cloud-specific, risk-adjusted security controls or components in advance. It may be necessary in some situations to requests from the cloud service providers through contractual means and service-level agreements that all require security components and controls to be fully and accurately implemented.

A further discussion of internal security considerations within cloud ecosystems can be found in Appendix B. Future versions of this document will contextualize the content of Appendix B in the NBDRA.

3 EXAMPLE USE CASES FOR SECURITY AND PRIVACY

There are significant Big Data challenges in science and engineering. Many of these are described in the use cases in *NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements*. However, these use cases focused primarily on 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 in the preparation of this document specifically to discover security and privacy issues. Some of these use cases represent inactive or legacy applications, but were selected because they demonstrate characteristic security / privacy design patterns.

The use cases selected for security and privacy are presented in the following subsections. The use cases included are grouped to organize this presentation, as follows: retail/marketing, healthcare, cybersecurity, government, industrial, aviation, and transportation. However, these groups do not represent the entire spectrum of industries affected by Big Data security and privacy.

The use cases were collected when the reference architecture was not mature. The use cases were collected to identify representative security and privacy scenarios thought to be suitably classified as particular to Big Data. An effort was made to map the use cases to the NBDRA. In Version 2, additional mapping of the use cases to the NBDRA and taxonomy will be developed. Parts of this document were developed in parallel and the connections will be strengthened in Version 2.

3.1 RETAIL/MARKETING

3.1.1 CONSUMER DIGITAL MEDIA USAGE

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

Comparative pricing from different retailers, store location and/or delivery options, and crowd-sourced rating have become common factors for selection. To compete, retailers are keeping a close watch on consumer locations, interests, and spending patterns to dynamically create marketing strategies 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 applications (apps) on smart devices.

- Privacy:
  - Most data collection 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: while some data collection and aggregation uses anonymization techniques, 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. "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."[17] DRM can be compromised, diverted to unanticipated purposes, defeated, or fail to operate in environments with Big Data characteristics—especially velocity and aggregated volume

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

3.1.2 NIELSEN HOMESCAN: PROJECT APOLLO

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

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

Then Current (2005-2006) Security and Privacy:

- Privacy: There was a considerable amount of PII data. Survey participants are compensated in exchange for giving up segmentation data, demographics, and other information.
- 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.

3.1.3 WEB TRAFFIC ANALYTICS

Scenario Description: Visit-level webserver logs are high-granularity and voluminous. To be useful, log data must be correlated with other (potentially Big Data) data sources, including page content (buttons, text, navigation events), and marketing-level events such as campaigns, media classification, etc. There are discussions—if not deployment—of plans for traffic analytics using complex event processing (CEP) in real time. One nontrivial problem is segregating traffic types, including internal user communities, for which collection policies and security are different.

Current Security and Privacy:

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

3.2 HEALTHCARE

3.2.1 HEALTH INFORMATION EXCHANGE

Scenario Description: Health Information Exchanges (HIEs) 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 Health Insurance Portability and Accountability Act (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. (For related standards efforts regarding consent, see NIST 800-53, Appendix J, Section IP-1), US DHS Health IT Policy Committee, Privacy and Security Workgroup) and Health Level Seven (HL7) International Version 3 standards for Data Access Consent, Consent Directives)
- **Pandemic Assistance:** There will be situations when public health entities, such as the CDC and perhaps other nongovernmental organizations that require this information to facilitate public safety, will require controlled access to this information, perhaps in situations where services and infrastructures are inaccessible. A cloud HIE with the right cryptographic controls could release essential information to authorized entities through authorization and audits in a manner that facilitates the scenario requirement

Project Current and/or Proposed Security and Privacy:

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

3.2.2 GENETIC PRIVACY

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

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

Current Security and Privacy:

- Security:
  - Secure Sockets Layer (SSL)-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

3.2.3 PHARMA CLINICAL TRIAL DATA SHARING[18]

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.

The Pharmaceutical Research and Manufacturers of America (PhRMA) 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 by 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 protect 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

3.3 CYBERSECURITY

3.3.1 NETWORK PROTECTION

Scenario Description: Network protection includes a variety of data collection and monitoring. Existing network security packages monitor high-volume data sets, such as event logs, across thousands of workstations and servers, but they are not yet able to scale to Big Data. Improved security software will include physical data correlates (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 subscenarios:

- Security for Big Data
- Big Data for security

Current Security and Privacy:

- Security in this area is mature; privacy concepts less so.
   - Traditional policy-type security prevails, though temporal dimension and monitoring of policy modification events tends to be nonstandard or unaudited
   - Cybersecurity apps run at high levels of security and thus require separate audit and security measures
   - No cross-industry standards exist for aggregating data beyond operating system collection methods
   - Implementing Big Data cybersecurity should include data governance, encryption/key management, and tenant data isolation/containerization
   - Volatility should be considered in the design of backup and disaster recovery for Big Data cybersecurity. The useful life of logs may extend beyond the lifetime of the devices which created them
- Privacy:
   - Enterprise authorization for data release to state/national organizations
   - Protection of PII data

Currently vendors are adopting Big Data analytics for mass-scale log correlation and incident response, such as for security information and event management (SIEM).

3.4 GOVERNMENT

3.4.1 MILITARY: UNMANNED VEHICLE SENSOR DATA

Scenario Description: Unmanned vehicles (or drones) and their onboard sensors (e.g., streamed video) can produce petabytes of data that should be stored in nonstandard 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[19]
- Insider threats (e.g., Edward Snowden, Bradley Manning, and spies) are being addressed in programs such as the Defense Advanced Research Projects Agency's (DARPA) Cyber-Insider Threat (CINDER) program. This research and some of the unfunded proposals made by industry may be of interest

3.4.2 EDUCATION: COMMON CORE STUDENT PERFORMANCE REPORTING

Scenario Description: Forty-five states have decided to unify standards for K–12 student performance measurement. Outcomes are used for many purposes, and the program is incipient, but it will obtain longitudinal Big Data status. The data sets envisioned include student-level performance across students' entire school history and across schools and states, as well as taking into account variations in test stimuli.

Current Security and Privacy:

- Data is scored by private firms and forwarded to state agencies for aggregation. Classroom, school, and district identifiers remain with the scored results. The status of student PII is unknown; however, it is known that teachers receive classroom-level performance feedback. The extent of student/parent 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[20]
- According to some reports, parents can opt students out of state tests, so opt-out records must also be collected and used to purge ineligible student records. [21]

Current Research:

- Longitudinal performance data would have value for program evaluators if data scales up
- Data-driven learning[22] will involve access to students' performance data, probably more often than at test time, and at higher granularity, thus requiring more data. One example enterprise is Civitas Learning's[23] predictive analytics for student decision making

3.5 INDUSTRIAL: AVIATION

3.5.1 SENSOR DATA STORAGE AND ANALYTICS

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

Current and Proposed Security and Privacy:

- Encryption at rest: Big Data systems should encrypt data stored at the infrastructure layer so that cloud storage administrators cannot access the data
- Key management: The encryption key management should be architected so that end customers (e.g., airliners) have sole/shared control on the release of keys for data decryption
- Encryption in motion: Big Data systems should verify 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)

3.6 TRANSPORTATION

3.6.1 CARGO SHIPPING

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

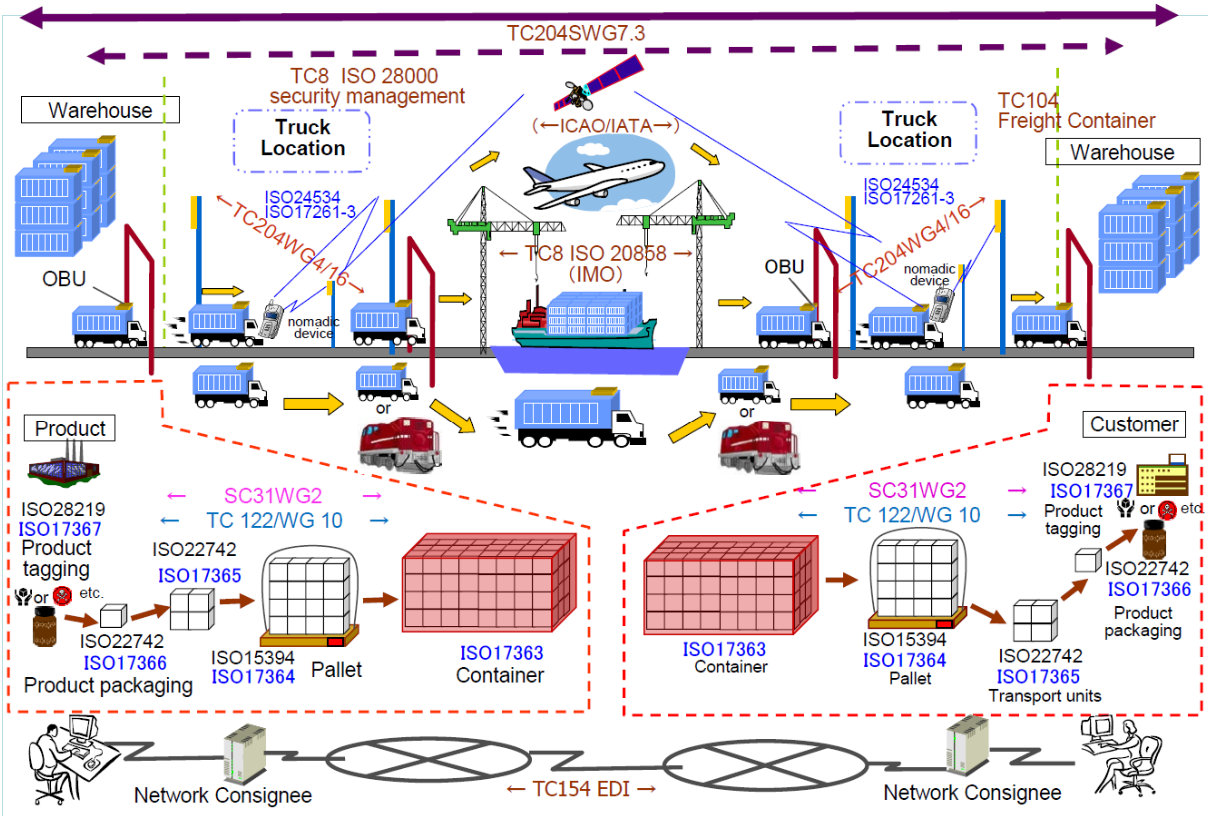

Figure 1: Cargo Shipping Scenario

4 TAXONOMY OF SECURITY AND PRIVACY TOPICS

A candidate set of topics from the Cloud Security Alliance Big Data Working Group (CSA BDWG) article, *Top Ten Challenges in Big Data Security and Privacy Challenges*, was used in developing these security and privacy taxonomies.[24] Candidate topics and related material used in preparing this section are provided for reference in Appendix A.

A taxonomy for Big Data security and privacy should encompass the aims of existing, useful taxonomies. While many concepts surrounding security and privacy exist, the objective in the taxonomies contained herein is to highlight and refine new or emerging principles specific to Big Data.

The following subsections present an overview of each security and privacy taxonomy, along with lists of topics encompassed by the taxonomy elements. These lists are the results of preliminary discussions of the Subgroup and may be developed further in Version 2.

4.1 CONCEPTUAL TAXONOMY OF SECURITY AND PRIVACY TOPICS

The conceptual security and privacy taxonomy, presented in Figure 2, contains four main groups: data confidentiality; data provenance; system health; and public policy, social, and cross-organizational topics. The first three topics broadly correspond with the traditional classification of confidentiality, integrity, and availability (CIA), reoriented to parallel Big Data considerations.

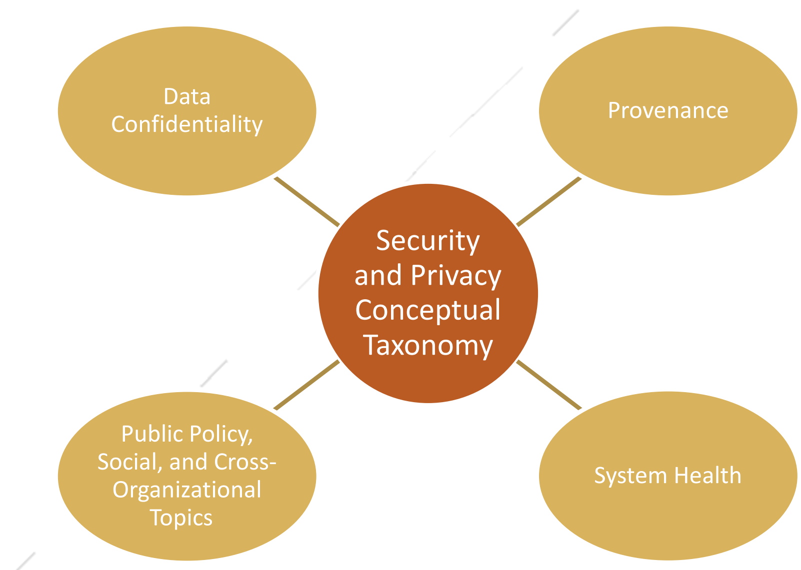

Figure 2: Security and Privacy Conceptual Taxonomy

4.1.1 DATA CONFIDENTIALITY

- Confidentiality of data in transit: For example, enforced by using Transport Layer Security (TLS)
- 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 plain text not deducible from the search criteria is guaranteed to be hidden
    - Homomorphic encryption: Cryptographic protocols that support operations on the underlying plain text of an encryption—any information about the plain text is guaranteed to be hidden
- Secure data aggregation: Aggregating data without compromising privacy
- Data anonymization
    - De-identification of records to protect privacy
- Key management
    - As noted by Chandramouli and Iorga, 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." [25]
    - Big Data systems are not purely cloud systems, but as is noted elsewhere in this document, the two are closely related. One possibility is to retarget the key management framework that Chandramouli and Iorga developed for cloud service models to the NBDRA security and privacy fabric. Cloud models would correspond to the NBDRA and cloud security concepts to the proposed fabric. NIST 800-145 provides definitions for cloud computing concepts, including infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) cloud service models [26]
    - Challenges for Big Data key management systems (KMS) reflect demands imposed by Big Data characteristics (i.e., volume, velocity, variety, and variability). 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

4.1.2 PROVENANCE

- End-point input validation: A mechanism to validate whether input data is coming from an authenticated source, such as digital signatures
    - Syntactic: Validation at a syntactic level
    - Semantic: Semantic validation is an important concern. Generally, semantic validation would validate typical business rules such as a due date. Intentional or unintentional violation of semantic rules can lock up an application. This could also happen when using data translators that do not recognize the particular variant. Protocols and data formats may be altered by a vendor using, for example, a reserved data field that will allow their products to have capabilities that differentiate them from other products. This problem can also arise in differences in versions of systems for consumer devices, including mobile devices. The semantics of a message and the data to be transported should be validated to verify, at a minimum, conformity with any applicable standards. The use of digital signatures will be important to provide assurance that the data from a sensor or data provider has been verified using a validator or data checker and is, therefore, 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 and disposition
  - DRM

4.1.3 SYSTEM HEALTH

- Security against denial-of-service (DoS)
  - Construction of cryptographic protocols proactively resistant to DoS
- Big Data for Security
  - Analytics for security intelligence
  - Data-driven abuse detection
  - Big Data analytics on logs, cyberphysical events, intelligent agents
  - Security breach event detection
  - Forensics
  - Big Data in support of resilience

4.1.4 PUBLIC POLICY, SOCIAL AND CROSS-ORGANIZATIONAL TOPICS

The following set of topics is drawn from an Association for Computing Machinery (ACM) grouping.[27] Each of these topics has Big Data security and privacy dimensions that could affect how a fabric overlay is implemented for a specific Big Data project. For instance, a medical devices project might need to address human safety risks, whereas a banking project would be concerned with different regulations applying to Big Data crossing borders. Further work to develop these concepts for Big Data is anticipated by the Subgroup.

- Abuse and crime involving computers
- Computer-related public / private health systems
- Ethics (within data science, but also across professions)
- Human safety
- Intellectual property rights and associated information management[d]
- Regulation
- Transborder data flows
- Use/abuse of power
- Assistive technologies for persons with disabilities (e.g., added or different security / privacy measures may be needed for subgroups within the population)
- Employment (e.g., regulations applicable to workplace law may govern proper use of Big Data produced or managed by employees)
- Social aspects of ecommerce
- Legal: Censorship, taxation, contract enforcement, forensics for law enforcement

---

[d] For further information, see the frameworks suggested by the Association for Information and Image Management (AIIM; http://www.aiim.org/) and the MIKE 2.0 Information Governance Association (http://mike2.openmethodology.org/wiki/MIKE2.0_Governance_Association)

4.2 OPERATIONAL TAXONOMY OF SECURITY AND PRIVACY TOPICS

Current practice for securing Big Data systems is diverse, employing widely disparate approaches that often are not part of a unified conceptual framework. The elements of the operational taxonomy, shown in Figure 3, represent groupings of practical methodologies. These elements are classified as "operational" because they address specific vulnerabilities or risk management challenges to the operation of Big Data systems. At this point in the standards development process, these methodologies have not been incorporated as part of a cohesive security fabric. They are potentially valuable checklist-style elements that can solve specific security or privacy needs. Future work must better integrate these methodologies with risk management guidelines developed by others (e.g., NIST Special Publication 800-37 Guide for Applying the Risk Management Framework to Federal Information Systems[28] and COBIT Risk IT Framework[29].)

In the proposed operational taxonomy, broad considerations of the conceptual taxonomy appear as recurring features. For example, confidentiality of communications can apply to governance of data at rest and access management, but it is also part of a security metadata model.[30]

The operational taxonomy will overlap with small data taxonomies while drawing attention to specific issues with Big Data.[31] [32]

Figure 3: Security and Privacy Operational Taxonomy

4.2.1 DEVICE AND APPLICATION REGISTRATION

- 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

4.2.2 IDENTITY AND ACCESS MANAGEMENT

- Virtualization layer identity (e.g., cloud console, platform as a service [PaaS])
    - Trusted platforms
- Application layer Identity
- End-user layer identity management
    - Roles
- Identity provider (IdP)
    - An IdP is defined in the Security Assertion Markup Language (SAML). [33] 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 a 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 mechanisms at the registration phase
    - In Big Data, due to the density of the data, the user "roams" to data (whereas in conventional virtual private network [VPN]-style scenarios, users roam across trust boundaries). Therefore, the conventional authentication/authorization (authn/authz) model needs to be extended because the relying party is no longer fully trusted—they are custodians of somebody else's data. Data is potentially aggregated from multiple resource providers
    - One approach is to extend the claims-based methods of SAML to add security and privacy guarantees
- Additional XACML Concepts
    - XACML introduces additional concepts that may be useful for Big Data security. In Big Data, parties are not just sharing claims, but also sharing policies about what is authorized. There is a policy access point at every data ownership and authoring location, and a policy enforcement point at the data access. A policy enforcement point calls a designated policy decision point for an auditable decision. In this way, the usual meaning of non-repudiation and trusted third parties is extended in XACML. Big Data presumes an abundance of policies, "points," and identity issuers, as well as data
        - Policy authoring points
        - Policy decision points
        - Policy enforcement point
        - Policy access points

4.2.3 DATA GOVERNANCE

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. Big Data without Big Data governance may become less useful to its stakeholders. To stimulate positive change, data governance will need to persist across the data lifecycle—at rest, in motion, in incomplete stages, and transactions—while serving the security and privacy of the young, the old, individuals as organizations, and organizations as organizations. It will need to cultivate economic benefits and innovation but also enable freedom of action and foster individual and public welfare. It will need to rely on standards governing technologies and practices not fully understood while integrating the human element. Big Data governance will require new perspectives yet accept the slowness or inefficacy of some current techniques. Some data governance considerations are listed below.

Big Data Apps to Support Governance: The development of new applications employing Big Data principles and designed to enhance governance may be among the most useful Big Data applications on the horizon.

- Encryption and key management
  - At rest
  - In memory
  - In transit
- Isolation/containerization
- Storage security
- Data loss prevention and detection
- Web services 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" [34]
  - De-identification and anonymization
  - Re-identification risk management
- End-point validation
- DRM
- Trust
- Openness
- Fairness and information ethics [35]

4.2.4 INFRASTRUCTURE MANAGEMENT

Infrastructure management involves security and privacy considerations related to hardware operation and maintenance. Some topics related to infrastructure management are listed below.

- 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 European Telecommunications Standards Institute (ETSI) TS 102 484 Smart Card specifications [36]
    - 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. A given security and privacy fabric may be required to consider this. Resilience demands are domain-specific, but could entail geometric increases in Big Data system scale
  - 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

4.2.5 RISK AND ACCOUNTABILITY

Risk and accountability encompass the following topics:

- 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 spans 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.[37] Business risk models can incorporate privacy considerations

4.3 ROLES RELATED TO SECURITY AND PRIVACY TOPICS

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

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

4.3.1 INFRASTRUCTURE 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.

4.3.2 GOVERNANCE, RISK MANAGEMENT, AND COMPLIANCE

Data governance is a fundamental element in the management of data and data systems. Data governance refers to administering, or formalizing, discipline (e.g., behavior patterns) around the management of data. Risk management involves the evaluation of positive and negative risks resulting from the handling of Big Data. Compliance encompasses adherence to laws, regulations, protocols, and other guiding rules for operations related to Big Data. Typically, governance, risk management, and compliance (GRC) is a function that draws participation from multiple areas of the organization, such as legal, human resources (HR), information technology (IT), and compliance. In some industries and agencies, there may be a strong focus on compliance, often in isolation from disciplines.

Professionals working in GRC tend to have similar backgrounds, share a common terminology, and employ similar processes and workflows, which typically influence other organizations within the corresponding vertical market or sector.

Within an organization, GRC professionals aim to protect the organization from negative outcomes that might arise from loss of intellectual property, liability due to actions by individuals within the organization, and compliance risks specific to its vertical market.

In larger enterprises and government agencies, GRC professionals are usually assigned to legal, marketing, or 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, due to the newness of the Big Data scenario to the organization, a lack of resources, or other factors specific to small organizations. Prior to Big Data, GRC roles in smaller organizations received little attention.

A one-person company can easily construct a Big Data application and inherit numerous unanticipated related GRC responsibilities. This is a new GRC scenario.

A security and privacy fabric entails additional data and process workflow in support of GRC, which is most likely under the control of the System Orchestrator component of the NBDRA, as explained in Section 5.

4.3.3 INFORMATION WORKER

Information workers are individuals and groups who work on the generation, transformation, and consumption of content. Due to the nascent nature of the technologies and related businesses in which they work, they tend to use common terms at a technical level within a specialty. 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 may not focus on security or privacy until it adds effort, cost, risk, or compliance responsibilities to the process of accessing domain-specific data or analytical tools.

Information workers may serve as data curators. Some may be research librarians, operate in quality management roles, or be involved in information management roles such as content editing, search indexing, or performing forensic duties as part of legal proceedings.

Information workers are 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.

4.4 RELATION OF ROLES TO THE SECURITY AND PRIVACY CONCEPTUAL TAXONOMY

The next sections cover the four components of the conceptual taxonomy: data confidentiality, data provenance, system health, and public policy, social and cross-organizational topics. 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 affected 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 upcoming sections also explore potential gaps that would be of interest to the anticipated stakeholders and ratifiers who reside on these three new conceptual axes.

4.4.1 DATA CONFIDENTIALITY

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

4.4.2 PROVENANCE

Provenance (or veracity) is related in some ways 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, information workers may need to confirm that the data provenance guarantees some degree of information integrity and address incorrect, fabricated, or cloned data before it is presented to an organization.

Additional risks to an organization could arise if one of its data suppliers does not demonstrate the appropriate degree of care in filtering or labeling its data. As noted in the U.S. Department of Health and Human Services (HHS) press release announcing the HIPAA final omnibus rule:

> *"The changes announced today expand many of the requirements to business associates of these entities that receive protected health information, such as contractors and subcontractors. Some of the largest breaches reported to HHS have involved business associates. Penalties are increased for noncompliance based on the level of negligence with a maximum penalty of $1.5 million per violation."*[38]

Organizations using or sharing health data among ecosystem partners, including mobile apps and SaaS providers, will need to verify that the proper legal agreements are in place to require data veracity and provenance.

4.4.3 SYSTEM HEALTH MANAGEMENT

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 that provide data, analytics, or services—an area commonly known as operations support systems (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—two of the Big Data characteristics. 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.

Volume, velocity, variety, and variability of Big Data systems health make it different from small data system health. Health tools and design patterns for existing systems 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. Managing systems health—including security—will require roles defined as much by the tools needed to manage as by the organizational context. Stated differently, Big Data is transforming the role of the Computer Security Officer.

For example, one aspect motivated by the DevOps movement (i.e., move toward blending tasks performed by applications development and systems operations teams) 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.

4.4.4 PUBLIC POLICY, SOCIAL, AND CROSS-ORGANIZATIONAL TOPICS

Roles in setting public policy related to security and privacy are established in the U.S. by federal agencies such as the Federal Trade Commission, the Food and Drug Administration or the DHHS Office of National Coordinator. DHS is responsible for aspects of domestic U.S. computer security through the activities of US-CERT. Social roles include the influence of NGO's, interest groups, professional organizations and standards development organizations. Cross-organizational roles include design patterns employed across or within certain industries such as pharmaceuticals, logistics, manufacturing, distribution to facilitate data sharing, curation, and even orchestration. Big Data frameworks will impact, and are impacted by cross-organizational considerations, possibly industry-by-industry. Further work to develop these concepts for Big Data is anticipated by the Subgroup.

4.5 ADDITIONAL TAXONOMY TOPICS

Additional areas have been identified but not carefully scrutinized, and it is not yet clear whether these would fold into existing categories or if new categories for security and privacy concerns would need to be identified and developed. Some candidate topics are briefly described below.

4.5.1 PROVISIONING, METERING, AND BILLING

Provisioning, metering and billing are elements in typically commercial systems used to manage assets, meter their use and invoice clients for that usage. 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.

4.5.2 DATA SYNDICATION

A feature of Big Data systems is that data is bought and sold as a valuable asset. That Google Search is free relies on users giving up information about their search terms on a Big Data scale. Google and Facebook can choose to repackage and syndicate that information for use by others for a fee.

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.

5 SECURITY AND PRIVACY FABRIC

Security and privacy considerations are a fundamental aspect of the NBDRA. Using the material gathered for this volume and extensive brainstorming among the NBD-PWG Security and Privacy Subgroup members and others, the following proposal for a security and privacy fabric was developed. [e]

Security and Privacy Fabric: Security and privacy considerations form a fundamental aspect of the NBDRA. This is geometrically depicted in Figure 4 by the Security and Privacy Fabric surrounding the five main components, since all components are affected by security and privacy considerations. Thus, the role of security and privacy is correctly depicted in relation to the components but does not expand into finer details, which may be more accurate but are best relegated to a more detailed security and privacy reference architecture. The Data Provider and Data Consumer are included in the Security and Privacy Fabric since, at the least, they should agree on the security protocols and mechanisms in place. The Security and Privacy Fabric is an approximate representation that alludes to the intricate interconnected nature and ubiquity of security and privacy throughout the NBDRA.

This pervasive dimension is depicted in Figure 4 by the presence of the security and privacy fabric surrounding all of the functional components., NBD-PWG decided to include the Data Provider and Data Consumer as well as the Big Data Application and Framework Providers in 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.

At this time, explanations as to how the proposed fabric concept is implemented across each NBDRA component are cursory—more suggestive than prescriptive. However, it is believed that, in time, a template will evolve and form a sound basis for more detailed iterations.

[e] The concept of a "fabric" for security and privacy 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 and grid computing, as well as for commercial offerings from both hardware and software manufacturers.

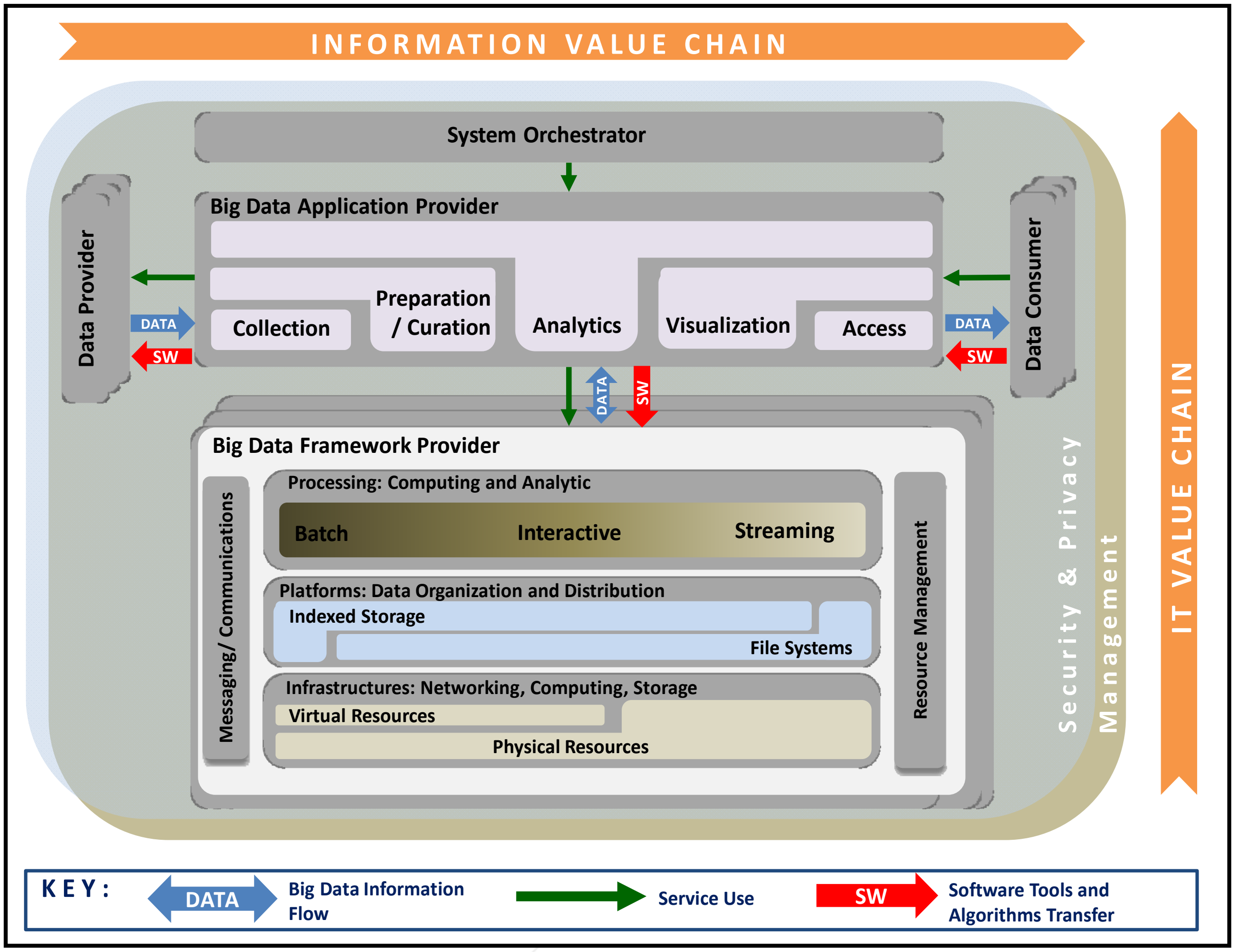

Figure 4: NIST Big Data Reference Architecture

Figure 4 introduces two new concepts that are particularly important 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.

5.1 SECURITY AND PRIVACY FABRIC IN THE NBDRA

Figure 5 provides an overview of several security and privacy topics with respect to some key NBDRA components and interfaces. The figure represents a beginning characterization 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.

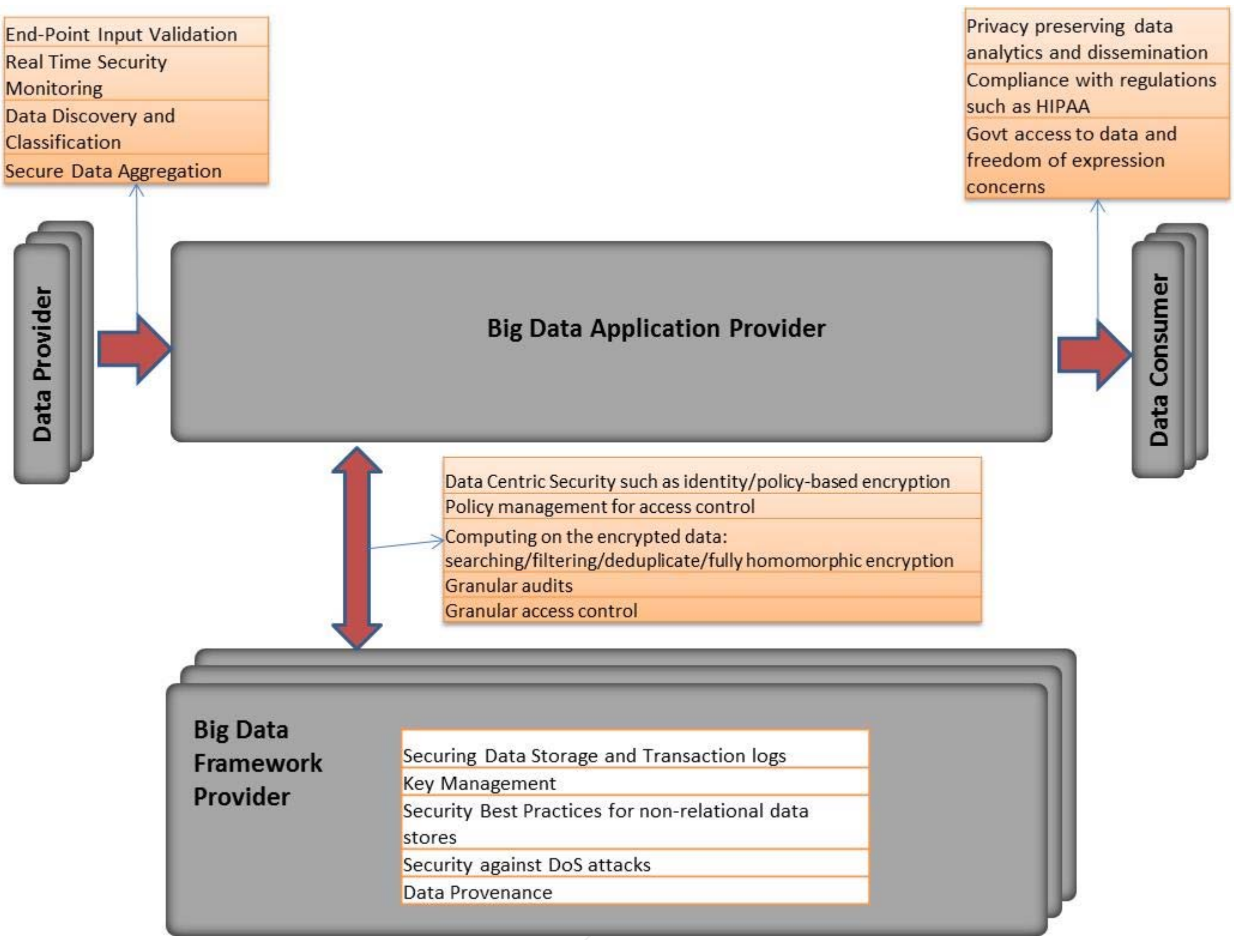

Figure 5: Notional Security and Privacy Fabric Overlay to the NBDRA

The groups and interfaces depicted in Figure 5 are described below.

A. INTERFACE BETWEEN 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 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.

B. INTERFACE BETWEEN BIG DATA APPLICATION PROVIDER → DATA CONSUMER

Data, including aggregate results delivered to data consumers, must preserve privacy. Data accessed by third parties or other entities should follow legal regulations such as HIPAA. Concerns include access to sensitive data by the government.

C. INTERFACE BETWEEN APPLICATION PROVIDER ↔ BIG DATA FRAMEWORK PROVIDER

Data can be stored and retrieved under encryption. Access control policies should be in place to assure 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.

D. 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 assure availability of the data.

E. 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.

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

5.2 PRIVACY ENGINEERING PRINCIPLES

Big Data security and privacy should leverage existing standards and practices. In the privacy arena, a systems approach that considers privacy throughout the process is a useful guideline to consider when adapting security and privacy practices to Big Data scenarios. The Organization for the Advancement of Structured Information Standards (OASIS) Privacy Management Reference Model (PMRM), consisting of seven foundational principles, provides appropriate basic guidance for Big System architects. [39,40] When working with any personal data, privacy should be an integral element in the design of a Big Data system.

Other privacy engineering frameworks are also under consideration.[41] [42] [43] [44] [45] [46]

Related principles include identity management frameworks such as proposed in the National Strategy for Trusted Identities in Cyberspace (NSTIC)[47] and considered in the NIST Cloud Computing Security Reference Architecture.[48] Aspects of identity management that contribute to a security and privacy fabric will be addressed in future versions of this document.

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

5.3 RELATION OF THE BIG DATA SECURITY OPERATIONAL TAXONOMY TO THE NBDRA

Table 1 represents a preliminary mapping of the operational taxonomy to the NBDRA components. The topics and activities listed for each operational taxonomy element (Section 4.2) have been allocated to a NBDRA component under the Activities column in Table 1. The description column provides additional information about the security and privacy aspects of each NBDRA component.

Table 1: 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	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

Activities	Description
• Resiliency, Redundancy and Recovery • Accountability • Compliance • Forensics • Business Risk Model	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, or valuable data be 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.

6 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, Notional Security and Privacy Fabric Overlay to the NBDRA.

6.1 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. [49] 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

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
	Event detection	“Playback” granularity defined
	Forensics	Subpoena of playback records in legal disputes

6.2 NIELSEN HOMESCAN: PROJECT APOLLO

Nielsen Homescan involves family-level retail transactions and associated media exposure using a statistically valid national sample. A general description[50] 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

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
	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

6.3 WEB TRAFFIC ANALYTICS

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

Table 4: Mapping Web Traffic Analytics to the 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	Web server 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

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
	Event detection	Can infer; for example, numerous sales, marketing, and overall web health events
	Forensics	See the SIEM use case

6.4 HEALTH INFORMATION EXCHANGE

Health information exchange (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[51]) bridge in lieu of general PKI
	Real-time security monitoring	Validation of incoming records to assure integrity through signature validation and to assure 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

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
		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 distributed denial of Service (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-

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
		level transfers of custody and cell-level access or modification

6.5 GENETIC PRIVACY

Mapping of genetic privacy is under development and will be included in future versions of this document.

6.6 PHARMACEUTICAL CLINICAL TRIAL DATA SHARING

Under an industry trade group proposal, clinical trial data for new drugs will be shared outside intra-enterprise warehouses.

Table 6: Mapping Pharmaceutical 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 Real-time security monitoring Data discovery and classification Secure data aggregation	Opaque—company-specific None Opaque—company-specific Third-party aggregator
Application Provider → Data Consumer	Privacy-preserving data analytics Compliance with regulations Government access to data and freedom of expression concerns	Data to be reported in aggregate but preserving potentially small-cell demographics Responsible developer and third-party custodian 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 Policy management for access control Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption Audits	TBD Internal roles; third-party custodian roles; researcher roles; participating patients' physicians TBD Release audit by a third party
Framework Provider	Securing data storage and transaction logs Key management Security best practices for non-relational data stores Security against DoS attacks Data provenance	TBD Internal varies by firm; external TBD TBD Unlikely to become public TBD—critical issue
Fabric	Analytics for security intelligence Event detection Forensics	TBD TBD

6.7 NETWORK PROTECTION

SIEM is a family of tools used to defend and maintain networks.

Table 7: 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[52]
	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

6.8 MILITARY: UNMANNED VEHICLE SENSOR DATA

Unmanned vehicles (drones) and their onboard sensors (e.g., streamed video) can produce petabytes of data that should be stored in nonstandard formats. The U.S. government is pursuing capabilities to expand storage capabilities for Big Data such as streamed video. For more information, refer to the Defense Information Systems Agency (DISA) large data object contract for exabytes in the DOD private cloud.[53]

Table 8: 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[54] 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 infers height of target from shadow. Fuse data from satellite infrared with separate sensor stream
	Forensics	Used for after action review (AAR)—desirable to have full playback of sensor streams

6.9 EDUCATION: COMMON CORE STUDENT PERFORMANCE REPORTING

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

Table 9: Mapping Common Core K–12 Student Reporting to the 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 [55]
	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

6.10 SENSOR DATA STORAGE AND ANALYTICS

Mapping of sensor data storage and analytics is under development and will be included in future versions of this document.

6.11 CARGO SHIPPING

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

Table 10: Mapping Cargo Shipping to the 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 assured. 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

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
	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*.[56] Security and privacy concerns are classified in four categories:

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

NBD-PWG Security and Privacy Subgroup identified the Big Data topics below for possible inspection during the preparation of Version 2 of this document. A complete rework of these topics is beyond the scope of this document. This material may be refined and organized if needed in future versions of this document.

Infrastructure Security

- Review of technologies and frameworks that have been primarily developed for performance, scalability, and availability, massively parallel processing (MPP) databases, and others.
- High-availability
  - Use of Big Data to enhance defenses against 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 social 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)
    - 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-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
        - Are 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.
    - 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 are, but both are concerned about data misuse and their inability to govern private- and public-sector use

    - 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, such as filtered upon entry processed and presented in the inbound framework
    - Issues surrounding permission to collect data, consent, and privacy
        - Differences between where the privacy settings are applied in web services and the user's perception of the privacy setting application
        - 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
    - 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
        - 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 Big Data systems will be designed using cloud architectures. Any strategy to implement a mature security and privacy framework within a Big Data cloud ecosystem enterprise architecture must address the complexities associated with cloud-specific security requirements triggered by the cloud characteristics. These requirements could include 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, thereby altering the agency's security posture.

To preserve the security-level after the migration of their data to the cloud, organizations need to identify all cloud-specific, risk-adjusted security controls or components in advance. The organizations must also 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.

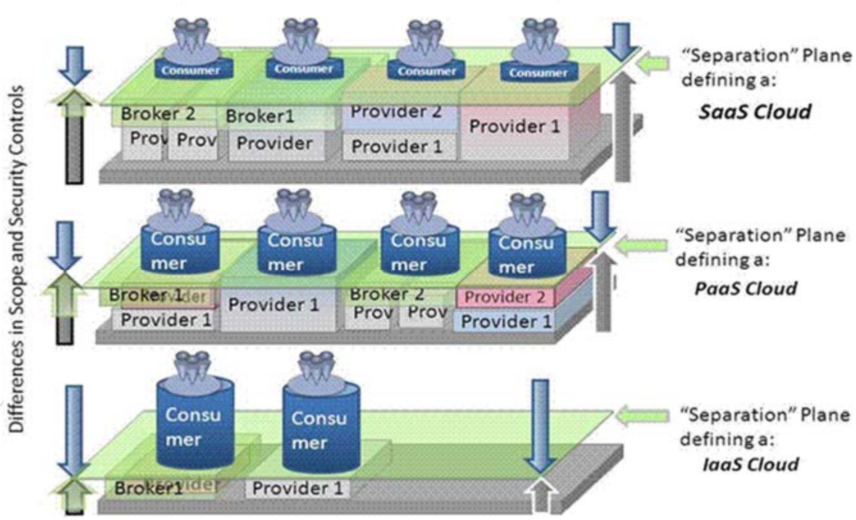

Figure B-1: Composite Cloud Ecosystem Security Architecture[57]

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

To meet GRC and confidentiality, integrity, and availability regulatory obligations required from the responsible data custodians—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 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 Certified Information Systems Security Professional (CISSP) body of knowledge. [f] This framework for Big Data may be adopted during the future work of the NBD-PWG.

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. One 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. 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

[f] CISSP is a professional computer security certification administered by $(ISC)^2$. (https://www.isc2.org/cissp/default.aspx)

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 the following.

- How important/critical is the data to the lifeblood 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, 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[58], 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, and perhaps all, of the measure 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 check 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 check 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 provide 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 and Lopez-Sanz et al., actors and roles were incorporated into Unified Modeling Language so that these concepts could be represented within and well as across services. [59] [60] 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 from Business Process Execution Language, Business Process Model and Notation frameworks offer additional patterns for Big Data security and privacy fabric standards. Ardagna et al. [61] 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 could 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

OPT-IN REVISITED

While standards organizations grapple with frameworks such as the one developed here, and until an individual's privacy and security can be fully protected using such a framework, some observers believe that the following two simple "protocols" ought to govern PII Big Data collection in the meantime.

Suggested Protocol one: An individual can only decide to opt-in for inclusion of their personal data manually, and it is a decision that they can revoke at any time.

Suggested Protocol number two: The individual's privacy and security opt-in process should enable each individual to modify their choice at any time, to access and review log files and reports and establish a self-destruct timeline (similar to the EU's "right to be forgotten".)

Appendix D: Acronyms

AC&S	access control and security
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
DDoS	distributed denial of Service
DOD	U.S. Department of Defense
DoS	denial of service
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

Luciano, Floridi (ed.), *The Cambridge Handbook of Information and Computer Ethics* (New York, NY: Cambridge University Press, 2010).

Julie Lane, Victoria Stodden, Stefen Bender, and Helen Nissenbaum (eds.), *Privacy, Big Data and the Public Good: Frameworks for Engagement* (New York, NY: Cambridge University Press, 2014).

Martha Nussbaum, *Creating Capabilities: The Human Development Approach* (Cambridge, MA: Belknap Press, 2011).

John Rawls, *A Theory of Justice* (Cambridge, MA: Belknap Press, 1971).

Martin Rost and Kirsten Bock, "Privacy by Design and the New Protection Goals," English translation of Privacy By Design und die Neuen Schutzziele, *Datenschutz und Datensicherheit*, Volume 35, Issue 1 (2011), pages 30-35.

DOCUMENT REFERENCES

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

[2] EMC^2, "Digital Universe," *EMC*, accessed February 21, 2014, http://www.emc.com/leadership/programs/digital-universe.htm.

[3] EMC^2, "Digital Universe," *EMC*, accessed February 21, 2014, http://www.emc.com/leadership/programs/digital-universe.htm.

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

[5] Subgroup correspondence with James G Kobielus (IBM), August 28, 2014.

[6] Big Data Working Group, "Top 10 Challenges in Big Data Security and Privacy," *Cloud Security Alliance*, November 2012, http://www.isaca.org/Groups/Professional-English/big-data/GroupDocuments/Big_Data_Top_Ten_v1.pdf.

[7] Benjamin Fung, Ke Wang, Rui Chen, and Philip S. Yu. "Privacy-preserving data publishing: A survey of recent developments", ACM Computing Surveys (CSUR), 42(4):14, 2010.

[8] Cynthia Dwork. "Differential privacy", In Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, editors, ICALP 2006: 33rd International Colloquium on Automata, Languages and Programming, Part II, volume 4052 of Lecture Notes in Computer Science, pages 1-12, Venice, Italy, July 10-14, 2006. Springer, Berlin, Germany.

[9] Latanya Sweeney. "k-anonymity: A model for protecting privacy", International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05):557-570, 2002.

[10] Arvind Narayanan and Vitaly Shmatikov. "Robust de-anonymization of large sparse datasets", In 2008 IEEE Symposium on Security and Privacy, pages 111-125, Oakland, California, USA, May 18-21, 2008. IEEE Computer Society Press.

[11] Big Data Working Group, "Top 10 Challenges in Big Data Security and Privacy," *Cloud Security Alliance*, November 2012, http://www.isaca.org/Groups/Professional-English/big-data/GroupDocuments/Big_Data_Top_Ten_v1.pdf.

[12] Big Data Working Group, "Top 10 Challenges in Big Data Security and Privacy," *Cloud Security Alliance*, November 2012, http://www.isaca.org/Groups/Professional-English/big-data/GroupDocuments/Big_Data_Top_Ten_v1.pdf.

[13] S. S. Sahoo, A. Sheth, and C. Henson, "Semantic provenance for eScience: Managing the deluge of scientific data," *Internet Computing, IEEE,* Volume 12, Issue 4 (2008), pages 46–54, http://dx.doi.org/10.1109/MIC.2008.86.

[14] Ronan Shields, "AppNexus CTO on the fight against ad fraud," *Exchange Wire*, October 29, 2014, https://www.exchangewire.com/blog/2014/10/29/appnexus-cto-on-the-fight-against-ad-fraud/.

[15] David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani, "The parable of google flu: Traps in big data analysis" *Science* Volume 343, Issue 6176 (2014), pages 1203-1205, http://dx.doi.org/10.1126/science.1248506.

[16] Peng Chen, Beth Plale, and Mehmet Aktas, "Temporal representation for mining scientific data provenance," *Future Generation Computer Systems*, Volume 36, Special Issue (2014), pages 363-378, http://dx.doi.org/10.1016/j.future.2013.09.032.

[17] Xiao Zhang, edited by Raj Jain, "A survey of digital rights management technologies," *Washington University in Saint Louis,* accessed January 9, 2015, http://bit.ly/1y3Y1P1.

[18] PhRMA, "Principles for Responsible Clinical Trial Data Sharing," *European Federation of Pharmaceutical Industries and Associations,* July 18, 2013, http://phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf.

[19] U.S. Army, "Army Regulation 25-2," *U.S. Army Publishing Directorate*, October 27, 2007, www.apd.army.mil/jw2/xmldemo/r25_2/main.asp.

[20] Jon Campbell, "Cuomo panel: State should cut ties with inBloom," *Albany Bureau,* March 11, 2014, http://lohud.us/1mV9U2U.

[21] 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/.

[22] 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.

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

[24] Big Data Working Group, "Top 10 Challenges in Big Data Security and Privacy," *Cloud Security Alliance*, November 2012, http://www.isaca.org/Groups/Professional-English/big-data/GroupDocuments/Big_Data_Top_Ten_v1.pdf.

[25] R. Chandramouli, M. Iorga, and S. Chokhani, "Cryptographic key management issues & challenges in cloud services," *National Institute of Standards and Technology*, September 2013, http://dx.doi.org/10.6028/NIST.IR.7956.

[26] Peter Mell and Timothy Grance, "The NIST Definition of Cloud Computing: Recommendations of the National Institute of Standards and Technology," *National Institute of Standards and Technology,* September 2011, http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[27] ACM, Inc., "The ACM Computing Classification System" *Association for Computing Machinery, Inc.*, 1998, http://www.acm.org/about/class/ccs98-html#K.4.

[28] Computer Security Division, Information Technology Laboratory, "Guide for Applying the Risk Management Framework to Federal Information Systems: A Security Life Cycle Approach," *National Institute for Standards and Technology,* February 2010, http://csrc.nist.gov/publications/nistpubs/800-37-rev1/sp800-37-rev1-final.pdf.

[29] ISACA, "The Risk IT Framework," *www.isaca.org,* 2009, http://www.isaca.org/Knowledge-Center/Research/ResearchDeliverables/Pages/The-Risk-IT-Framework.aspx.

[30] Cybersecurity Framework, "Framework for Improving Critical Infrastructure Cybersecurity" *National Institute for Standards and Technology,* accessed January 9, 2015, http://1.usa.gov/1wQuti1.

[31] OASIS "SAML V2.0 Standard," *SAML Wiki*, accessed January 9, 2015, http://bit.ly/1wQByit.

[32] James Cebula and Lisa Young, "A taxonomy of operational cyber security risks," *Carnegie Melon University*, December 2010, http://resources.sei.cmu.edu/asset_files/TechnicalNote/2010_004_001_15200.pdf.

[33] OASIS "SAML V2.0 Standard," *SAML Wiki*, accessed January 9, 2015, http://bit.ly/1wQByit.

[34] H. C. Kum and S. Ahalt, "Privacy-by-Design: Understanding Data Access Models for Secondary Data," *AMIA Summits on Translational Science Proceedings*, *2013*, pages 126–130.

[35] John Rawls, "Justice as Fairness," *A Theory of Justice*, 1985.

[36] ETSI, "Smart Cards; Secure channel between a UICC and an end-point terminal," *etsy.org,* December 2007, http://bit.ly/1x2HSUe.

[37] James Cebula and Lisa Young, "Taxonomy of Operational Cyber Security Risks," (Pittsburgh, PA: Carnegie Mellon University, Software Engineering Institute, 2010).

[38] HHS Press Office, "New rule protects patient privacy, secures health information," *U.S. Department of Health and Human Services,* January 17, 2013, http://www.hhs.gov/news/press/2013pres/01/20130117b.html.

[39] John Sabo, Michael Willet, Peter Brown, and Dawn Jutla, "Privacy Management Reference Model and Methodology (PMRM) Version 1.0," *OASIS,* March 26, 2012, http://docs.oasis-open.org/pmrm/PMRM/v1.0/csd01/PMRM-v1.0-csd01.pdf.

[40] NIST, "National Strategy for Trusted Identities in Cyberspace (NSTIC)," *National Institute for Standards and Technology,* 2015, http://www.nist.gov/nstic/.

[41] Wayne Jansen and Timothy Grance, SP800-144, "Guidelines on Security and Privacy in Public Cloud Computing," *National Institute for Standards and Technology,* December 2011, http://csrc.nist.gov/publications/nistpubs/800-144/SP800-144.pdf.

[42] Wayne Jansen and Timothy Grance, SP 800-144, "Guidelines on Security and Privacy in Public Cloud Computing," *National Institute for Standards and Technology,* December 2011, http://csrc.nist.gov/publications/nistpubs/800-144/SP800-144.pdf.

[43] Carolyn Brodie, Clare-Marie Karat, John Karat, and Jinjuan Feng, "Usable security and privacy: A case study of developing privacy management tools," *Proceedings of the 2005 Symposium on Usable Privacy and Security,* 2005, http://doi.acm.org/10.1145/1073001.1073005.

[44] W. Knox Carey, Jarl Nilsson, and Steve Mitchell, "Persistent security, privacy, and governance for healthcare information," *Proceedings of the 2nd USENIX Conference on Health Security and Privacy,* 2011, http://dl.acm.org/citation.cfm?id=2028026.2028029.

[45] Paul Dunphy, John Vines, Lizzie Coles-Kemp, Rachel Clarke, Vasilis Vlachokyriakos, Peter Wright, John McCarthy, and Patrick Olivier, "Understanding the Experience-Centeredness of privacy and security technologies," *Proceedings of the 2014 Workshop on New Security Paradigms Workshop,* 2014, http://doi.acm.org/10.1145/2683467.2683475.

[46] Ebenezer Oladimeji, Lawrence Chung, Hyo Taeg Jung, and Jaehyoun Kim, "Managing security and privacy in ubiquitous eHealth information interchange," *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication,"* 2011, http://doi.acm.org/10.1145/1968613.1968645.

[47] NIST, "National Strategy for Trusted Identities in Cyberspace (NSTIC)," *National Institute for Standards and Technology,* 2015, http://www.nist.gov/nstic/.

[48] NIST Cloud Computing Security Working Group, "NIST Cloud Computing Security Reference Architecture," *National Institute for Standards and Technology,* May 15, 2013, http://collaborate.nist.gov/twiki-cloud-computing/pub/CloudComputing/CloudSecurity/NIST_Security_Reference_Architecture_2013.05.15_v1.0.pdf.

[49] Microsoft, "Deploying Windows Rights Management Services at Microsoft," *Microsoft*, 2015, http://technet.microsoft.com/en-us/library/dd277323.aspx.

[50] The Nielsen Company, "Consumer Panel and Retail Measurement," *Nielsen*, 2015, www.nielsen.com/us/en/nielsen-solutions/nielsen-measurement/nielsen-retail-measurement.html.

[51] SAFE-BioPharma, "Welcome to SAFE-BioPharma," *SAFE-BioPharma Association,* accessed March 3, 2015, http://www.safe-biopharma.org/.

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

[53] 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.

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

[55] K. A. G. Fisher, A. Broadbent, L. K. Shalm, Z. Yan, J. Lavoie, R. Prevedel, T. Jennewein, and K. J. Resch, "Quantum computing on encrypted data 5," *Nature Communications,* January 2015, http://www.nature.com/ncomms/2014/140121/ncomms4074/full/ncomms4074.html.

[56] Big Data Working Group, "Top 10 Challenges in Big Data Security and Privacy," Cloud Security Alliance, November 2012, http://www.isaca.org/Groups/Professional-English/big-data/GroupDocuments/Big_Data_Top_Ten_v1.pdf.

[57] Fang Liu, Jin Tong, Jian Mao, Robert Bohn, John Messina, Lee Badger, and Dawn Leaf, SP500-292, "NIST Cloud Computing Reference Architecture," *National Institute of Standards and Technology,* September 2011, http://www.nist.gov/customcf/get_pdf.cfm?pub_id=909505.

[58] John Mutch and Brian Anderson, "Preventing Good People From Doing Bad Things: Implementing Least Privilege," (Berkeley, CA: Apress, 2011).

[59] S. Patig, "Model-Driven development of composite applications," *Communications in Computer and Information Science*, 2008, http://dx.doi.org/10.1007/978-3-540-78999-4_8.

[60] M. López-Sanz, C. J. Acuña, C. E. Cuesta, and E. Marcos, "Modelling of Service-Oriented Architectures with UML," *Theoretical Computer Science*, Volume 194, Issue 4 (2008), pages 23–37.

[61] D. Ardagna, L. Baresi, S. Comai, M. Comuzzi, and B. Pernici, "A Service-Based framework for flexible business processes," *IEEE*, March 2011, http://dx.doi.org/10.1109/ms.2011.28.