NIST Special Publication 1500-4r1

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

NIST Big Data Public Working Group
Security and Privacy Subgroup

Version 2
June 2018

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-4r1

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

NIST Special Publication 1500-4r1

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

Version 2

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

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1500-4r1

June 2018

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

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

National Institute of Standards and Technology (NIST) Special Publication 1500-4r1
Natl. Inst. Stand. Technol. Spec. Publ. 1500-4r1, 171 pages (June 2018) CODEN: NSPUE2

This publication is available free of charge from: https://doi.org/10.6028/NIST.SP.1500-4r1

NIST Special Publication series 1500 is intended to capture external perspectives related to NIST standards, measurement, and testing-related efforts. These external perspectives can come from industry, academia, government, and others. These reports are intended to document external perspectives and do not represent official NIST positions.

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

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

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. All NIST publications are available at http://www.nist.gov/publication-portal.cfm.

Comments on this publication may be submitted to Wo Chang

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

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at NIST promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology (IT). 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 IT and its collaborative activities with industry, government, and academic organizations.

Abstract

Big Data is a term used to describe the large amount of data in the networked, digitized, sensor-laden, information-driven world. While opportunities exist with Big Data, the data can overwhelm traditional technical approaches and the growth of data 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 concepts related to Big Data. The results are reported in the *NIST Big Data Interoperability Framework (NBDIF)* series of volumes. This volume, Volume 4, contains an exploration of security and privacy topics with respect to Big Data. The 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 characteristics; Big Data forensics; Big Data privacy; Big Data risk management; Big Data security; Big Data taxonomy, computer security; cybersecurity; encryption standards; information assurance; information security frameworks; role-based access controls; security and privacy fabric; use cases.

Acknowledgements

This document reflects the contributions and discussions by the membership of the NBD-PWG, cochaired by Wo Chang (NIST ITL), Bob Marcus (ET-Strategies), and Chaitan Baru (San Diego Supercomputer Center; National Science Foundation). The document contains input from members of the NBD-PWG Security and Privacy Subgroup. For all versions, the Subgroups were led by the following people: Nancy Grady (SAIC), Natasha Balac (San Diego Supercomputer Center), and Eugene Luster (R2AD) for the Definitions and Taxonomies Subgroup; Geoffrey Fox (Indiana University) and Tsegereda Beyene (Cisco Systems) for the Use Cases and Requirements Subgroup; Arnab Roy (Fujitsu), Mark Underwood (Krypton Brothers; Synchrony Financial), and Akhil Manchanda (GE) for the Security and Privacy Subgroup; David Boyd (InCadence Strategic Solutions), Orit Levin (Microsoft), Don Krapohl (Augmented Intelligence), and James Ketner (AT&T) for the Reference Architecture Subgroup; and Russell Reinsch (Center for Government Interoperability), David Boyd (InCadence Strategic Solutions), Carl Buffington (Vistronix), and Dan McClary (Oracle), for the Standards Roadmap Subgroup.

The editors for this document were the following:

- ***Version 1***: Arnab Roy (Fujitsu), Mark Underwood (Krypton Brothers; Synchrony Financial) and Wo Chang (NIST)
- ***Version 2***: Arnab Roy (Fujitsu), Mark Underwood (Krypton Brothers; Synchrony Financial) and Wo Chang (NIST)

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

NIST SP1500-4, Version 2 has been collaboratively authored by the NBD-PWG. As of the date of this publication, there are over six hundred NBD-PWG participants from industry, academia, and government. Federal agency participants include the National Archives and Records Administration (NARA), National Aeronautics and Space Administration (NASA), National Science Foundation (NSF), and the U.S. Departments of Agriculture, Commerce, Defense, Energy, Health and Human Services, Homeland Security, Transportation, Treasury, and Veterans Affairs.

NIST would like to acknowledge the specific contributions to this volume, during Version 1 and/or Version 2 activities, by the following NBD-PWG members:

Cavan Capps
U.S. Census Bureau

Pw Carey
Compliance Partners, LLC

Wo Chang
NIST

Brent Comstock
Cox Communications

Lee Anne Davies
Agenomics

Jacob Dilles
Mount Airey Group

Michele Drgon
Data Probity

Roy D'Souza
AlephCloud Systems, Inc.

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

Robert Reyling
Wyle Aerospace Corporation

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; Synchrony Financial

Eddie Garcia
Gazzang, Inc.

David Harper
Johns Hopkins University/ Applied Physics Laboratory

Pavithra Kenjige
PK Technologies

Alicia Zuniga-Alvarado
Consultant

TABLE OF CONTENTS

Figures

Tables

EXECUTIVE SUMMARY

This *NIST Big Data Interoperability Framework (NBDIF): 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 healthcare, 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 NBDIF consists of nine volumes, each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The nine volumes, which can be downloaded from https://bigdatawg.nist.gov/V2_output_docs.php, are as follows:

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

The NBDIF will be released in three versions, which correspond to the three stages of the NBD-PWG work. The three stages aim to achieve the following with respect to the NBDRA.

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

Potential areas of future work for the Subgroup during Stage 3 are highlighted in Section 1.5 of this volume. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

Version 2 of *NBDIF: Volume 4, Security and Privacy* is principally informed by the introduction of the NIST Big Data Security and Privacy Safety Levels (NBD-SPSL). Using the NBD-SPSL, organizations can identify specific elements to which their systems conform. Readers are encouraged to study the NBD-SPSL (Appendix A) before launching into the body of this version of the document. Appendix A is designed to be a stand-alone, readily transferred artifact that can be used to share concepts that can improve Big Data security and privacy safety engineering.

By declaring conformance with selected elements from the NBD-SPSL, practitioners in Big Data can voluntarily attest to specific steps they have undertaken to improve Big Data security and privacy in their systems. The NBD-SPSL provides a clear path to implement the recommendations of standards aimed at improving ethical practices (e.g., Institute of Electrical and Electronics Engineers [IEEE] P7000, IEEE P7002, IEEE P7007, International Organization for Standardization [ISO] 27500:2016), as well as methods to integrate security and privacy into Big Data DevOps, (e.g., IEEE P2675).

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 cybersecurity threats be reversed?

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

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

- How is Big Data defined?
- What attributes define Big Data solutions?
- What is new in Big Data?
- What is the difference between Big Data and *bigger data* that has been collected for years?
- How is Big Data different from traditional data environments and related applications?
- What are the essential characteristics of Big Data environments?
- How do these environments integrate with currently deployed architectures?
- What are the central scientific, technological, and standardization challenges that need to be addressed to accelerate the deployment of robust, secure Big Data solutions?

Within this context, on March 29, 2012, the White House announced the Big Data Research and Development Initiative. [9] The initiative's goals include helping to accelerate the pace of discovery in science and engineering, strengthening national security, and transforming teaching and learning by improving analysts' ability to extract knowledge and insights from large and complex collections of digital data.

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

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

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

The *NIST Big Data Interoperability Framework* (NBDIF) will be released in three versions, which correspond to the three stages of the NBD-PWG work. The three stages aim to achieve the following with respect to the NIST Big Data Reference Architecture (NBDRA).

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

On September 16, 2015, seven NBDIF Version 1 volumes were published (http://bigdatawg.nist.gov/V1_output_docs.php), each of which addresses a specific key topic, resulting from the work of the NBD-PWG. The seven volumes are as follows:

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

Currently, the NBD-PWG is working on Stage 2 with the goals to enhance the Version 1 content, define general interfaces between the NBDRA components by aggregating low-level interactions into high-level general interfaces, and demonstrate how the NBDRA can be used. As a result of the Stage 2 work, the following two additional NBDIF volumes have been developed.

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

Version 2 of the NBDIF volumes, resulting from Stage 2 work, can be downloaded from the NBD-PWG website (https://bigdatawg.nist.gov/V2_output_docs.php). Potential areas of future work for each volume during Stage 3 are highlighted in Section 1.5 of each volume. The current effort documented in this volume reflects concepts developed within the rapidly evolving field of Big Data.

1.2 SCOPE AND OBJECTIVES OF THE 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. Early standards work, including the efforts of this Public Working Group, helped to focus attention on emerging risks as well as on the underlying technology.

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;
- 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; and
- 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.

In Version 1, the NBD-PWG introduced the concept of a security and privacy fabric. The fundamental idea is that security and privacy considerations impact all components within the NBDRA. This version of the document extends and amplifies this concept into the NIST Big Data Security and Privacy Safety Levels (NBD-SPSL) set forth in a single artifact (Appendix A). The single broadest objective for this document is to offer a three-level security and privacy safety rating for a Big Data system. This high-medium-low simplification is offered in a list form (Appendix A), though it can be implemented through semi-automated means; the latter are indicated but not proscriptive.

In addition, rather than embracing a maturity model, a safety engineering approach was chosen. The threats to safety and privacy in Big Data are sufficiently grave, and the teams involved in Big Data creation and analytics potentially so small, that a heavyweight, organizationally demanding framework seemed inappropriate for broad use. Other frameworks, both existing and under development, including some at NIST, address that space for Big Data and Internet of Things (IoT).

Since the initial version of this document, recent developments—some refocusing the practice of software engineering on specific components such as scalability, others form part of the steady march of technology—have impacted security and privacy. These recent developments include the following:

- Risks for intentional/unintentional breaches of privacy or discrimination against protected groups through machine learning and algorithmic reasoning;
- Need for decentralization of high-risk data, particularly authenticating resources;
- Adoption and integration of safety engineering practices;
- Security and safety engineering in DevOps (a clipped compound of software DEVelopment and information technology OPerationS) frameworks (DevSecOps);
- Security and privacy practices in agile development;
- Collaborative use of software-defined networks to partition and protect data, application realms, and physical infrastructure;
- Integral use of domain, application, and utility models to guide security and privacy practices;
- Blockchain and higher-granularity dynamic *smart contracts*;
- Cryptography and privacy-preserving methods;
- Big Data forensics frameworks to be concurrently engineered, not constructed after-the-fact;
- Increased use of attribute-based access control [10];
- Providing a broadly usable self-assessment for conformance to Big Data security levels; and

- Microservices, containers, and software-defined network as opportunity areas for security and privacy fabric enhancements.

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 the following:

- Announce that the NBD-PWG Security and Privacy Subgroup is open to the public 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;
- Expand the security and privacy fabric of the NBDRA and identify specific topics related to NBDRA components; and
- Begin mapping of identified security and privacy use cases to the NBDRA.

This report is a compilation of contributions from the NBD-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 3 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 explores details of the NBDRA, Security and Privacy Fabric, cryptographic technologies, risk management, Big Data security modeling and simulation (ModSim), and security and privacy management.
- Section 6 introduces the topic of domain-specific security.
- Section 7 introduces the topic of audit and configuration management.
- Section 8 considers standards, best practices, and gaps with respect to security and privacy.
- Appendix A presents the draft NBD-SPSL.
- Appendix B introduces concepts developed in selected existing standards.
- Appendix C discusses considerations when implementing a mature security and privacy framework within a Big Data cloud ecosystem enterprise architecture.
- Appendix D expands the notion of actors and roles.
- Appendix E maps the security- and privacy-related use cases presented in Section 3 to the NBDRA components.
- Appendix F provides a high-level list of additional topics explored in Version 2.
- Appendix G contains the acronyms used in this document.
- Appendix H lists the references used in the document.

1.5 FUTURE WORK ON THIS VOLUME

A number of topics have not been discussed and clarified sufficiently to be included in Version 2. Topics that remain to be addressed in Version 3 of this document include the following:

- Expand Big Data security and privacy analysis and foster adoption through deeper cross-linking to related standards. Progress was made but work remains.

- Implement explicit phase-specific guidance, which was initiated, but not fully implemented, especially in Appendix A.
- Develop improved guidelines for integrating supporting Big Data systems dedicated to security and privacy (Big Data security and privacy dogfooding).[a] Healthcare is the strongest use case, where risks are best understood.
- Incorporate security and privacy metadata-rich Big Data orchestration processes, enabled by tools such as Rundeck [11].
- Facilitate incorporation of security and privacy models for the software development life cycle. For example, there is a need to describe Big Data security and privacy practices that address the full life cycle of Big Data analytics and machine learning.
- Draft a Big Data-annotated version of the NIST SP 800-53 Privacy Controls Catalog (see (National Institute of Standards and Technology [NIST], 2014), Appendix J).
- Identify Big Data touchpoints for Privacy by Design, Organisation for Economic Co-Operation and Development (OECD), and other external privacy guidelines.
- Integrate models such as Sensing as a Service [13].
- Provide a deeper explanation of Big Data Application Provider security and privacy requirements.
- Develop security and privacy risk frameworks for specific design patterns (apart from cloud), including distributed computing, middleware (enterprise service bus), agent-based, recommendation engines, and web portals fronting legacy applications.
- More clearly identify where Big Data systems management intersects with security and privacy guidelines. The gold standard use case is the use of logging data for both operational intelligence and security and privacy, though the mapping is demonstrably nonorthogonal.
- Depict security and privacy policy and metadata orchestration using descriptions of test beds, such as those developed in the *NBDIF: Volume 8, Reference Architecture Interfaces* document.
- Update or build new frameworks for Big Data that reference existing International Organization for Standardization (ISO) and other standards for Big Data life cycle, audit, configuration management, and privacy-preserving practices.
- Contextualize the content of Appendix C (Internal Security Considerations within Cloud Ecosystems) in the NBDRA.
- Create additional mapping for the use cases to the NBDRA and security and privacy taxonomy.
- Enhance the discussion of infrastructure management, including in-depth exploration of *left shift* and its implications for Big Data security and privacy, implications for infrastructure as code, and relevance to NIST Critical Infrastructure.
- Identify Big Data challenges associated with ownership traceability, custody and curation – especially in light of data volatility introduced by mergers, cross-border flows, and project life cycle (Section 2.3.5).
- Explore the best practices and gaps associated with the SABSA and Zachman Framework

[a] Typically, such supporting security and privacy Big Data is provided as part of a fully integrated Build Phase, but some solutions can implement Security as a Service, with some or all security and privacy resources provided by third parties. Third parties may specialize in security and privacy for specific domains, with machine learning, ontologies, and other specialized resources that may be beyond the capabilities of Build architects.

2 BIG DATA SECURITY AND PRIVACY

Opinions, standards, and analysis on the topics of security and privacy are vast, with intensive work under way in disciplines ranging from law and education to highly specialized aspects of systems engineering. An overarching goal of the current work is to focus as narrowly as possible on Big Data security and privacy concerns, while identifying related work elsewhere that can clarify or strengthen the present undertaking.

2.1 WHAT IS DIFFERENT ABOUT 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 the time, the following 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, while alternative approaches to protecting privacy are made feasible. Although de-identification techniques can apply to data from single sources as well, the prospect of unanticipated consequences from the fusion of multiple datasets exacerbates the risk of compromising privacy.
4. A huge increase in the number of sensor streams for the Internet of Things (e.g., smart medical devices, smart cities, smart homes) creates vulnerabilities in the Internet connectivity of the devices, in the transport, and in the eventual aggregation.
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, context, 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.
8. Data and code can more readily be shared across organizations, but many standards presume management practices that are managed inside a single organizational framework. A related observation is that smaller firms, subject to fewer regulations or lacking mature governance practices, can create valuable Big Data systems. Lack of common data schemas can further inhibit consistent security and privacy practices.

The Security and Privacy Subgroup envisions further work to investigate the following list of potential differences between Big Data projects and traditional implementations with respect to security and privacy.

- Inter-organizational issues (e.g., federation, data licensing—not only for cloud);
- Mobile/geospatial increased risk for deanonymization;
- Change to life cycle processes (no *archive* or *destroy* due to Big Data);
- Related sets of standards are written with large organizational assumptions but currently, Big Data can be created / analyzed with small teams;
- Audit and provenance for Big Data intersects in novel ways with other aspects;
- Big Data as a technology accelerator for improved audit (e.g., blockchain, noSQL, machine learning for information security enabled by Big Data), analytics for intrusion detection, complex event processing;
- Transborder data flows present challenges to Big Data as it moves across national boundaries [14];
- Consent (e.g., smart contracts) frameworks, perhaps implemented using blockchain;
- Impact of real-time Big Data on security and privacy;
- Risk management in Big Data moves the focus to inter-organizational risk and risks associated with analytics versus a simplified four-walls perspective; and
- Of lesser importance, but relevant to how Big Data systems are often built, DevOps and agile processes inform the efforts of small teams (even single-developer efforts) in creation and fusion with Big Data.

2.2 OVERVIEW

Security and privacy measures are becoming ever more important with the increase of Big Data generation and utilization and the 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. (EMC2) Less than a third of data needed protection in 2010, but more than 40 percent of data will need protection in 2020. (EMC2)

Security and privacy measures for Big Data involve a different approach than for 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. [15]

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 healthcare, 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 may be applicable for protection of 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 related emerging technologies, and how these requirements are different 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 security and privacy and those of other big systems:

- 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. 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.[16]
- Privacy-preserving mechanisms are needed for Big Data, such as for PII. The privacy and integrity of data coming from end points should be protected at every stage because there may be disparate, potentially unanticipated processing steps between the data owner, provider, and data consumer. 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.[b]
- Legacy security solutions need to be retargeted to the infrastructural shift due to Big Data. Legacy security solutions address infrastructural security concerns that persist in Big Data, such as authentication, access control, and authorization. These solutions need to be retargeted to the underlying Big Data High Performance Computing (HPC) resources or completely replaced. Oftentimes, such resources can face the public domain, and thus necessitate vigilant security monitoring methods to prevent adversarial manipulation and to preserve integrity of operations.
- Information assurance (IA) and disaster recovery (DR) for Big Data Systems may require unique and emergent practices. Because of its extreme scalability, Big Data presents challenges for IA and DR practices that were not previously addressed in a systematic way. Traditional backup and replication 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.

[b] Reference to NBDRA Data Provider.

- 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 (IT) scenario in itself.
- Risks have increased for deanonymization 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 datasets could make re-identification possible. Even when technology can preserve privacy, proper consent and use may not follow the path of the data through various custodians. Because of the broad collection and set of uses of Big Data, consent for collection is much less likely to be sufficient and should be augmented with technical and legal controls to provide auditability and accountability for use. [17] [18]
- There are emerging risks in open data and Big Data 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.

2.3 SECURITY AND PRIVACY IMPACTS ON BIG DATA CHARACTERISTICS

Volume, velocity, variety, 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.

2.3.1 VOLUME

The volume of Big Data describes the size of the dataset. In Big Data parlance, this typically ranges from gigabytes (GB) to exabytes and beyond. As a result, the volume of Big Data has necessitated storage in multitiered 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. [19]

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.

2.3.2 VELOCITY

Velocity describes the rate of data flow. 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. [19] Malfunctioning computing nodes might leak confidential data. Partial infrastructure attacks could compromise a significantly large fraction of the system due to high levels of connectivity and dependency. If the system does not enforce strong authentication among geographically distributed nodes, rogue nodes can be added that can eavesdrop on confidential data.

2.3.3 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. [15] 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. The inference process is also aided by data volume, but the diversity of data sources is the primary cause here. While several formal models to address privacy-preserving data disclosure have been proposed, [20] [21] in practice, sensitive data is shared after sufficient removal of apparently unique identifiers, and indirectly identifying information by the processes of anonymization and aggregation. This is an ad hoc process that is often based on empirical evidence [22] and has led to many instances of deanonymization in conjunction with publicly available data. [23] Although some laws/regulations recognize only identifiers per se, laws such as the Health Insurance Portability and Accountability Act (HIPAA; the statistician provision), the Family Educational Rights and Privacy Act (FERPA), and 45 Code of Federal Regulations (CFR) 46 recognize that combinations of attributes, even if not the identifiers by themselves, can lead to actionable personal identification, possibly in conjunction with external information.

2.3.4 VERACITY

Big Data veracity and validity encompass several sub-characteristics as described below.

Provenance—or what some have called veracity in keeping with the V theme, though the two terms are not identical—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. Also, as noted before, with respect to privacy policy, additional context is needed to make responsible decisions over collected data, which may include the form of consent, intended use, temporal connotations (e.g., Right to be Forgotten), or broader context of collection. The additional context could be considered a type of provenance, broadly, but goes beyond the range of provenance information typically collected in production information systems. Various approaches have been tried, such as for glycoproteomics, [24] 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 although the two terms are not identical (see [25] for a deeper discussion), 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 are 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, and ordering data collection.

Furthermore, there is a central and broadly recognized privacy principle, incorporated in many privacy frameworks (e.g., the OECD principles, European Union [EU] data protection directive, Federal Trade Commission [FTC] fair information practices), that data subjects must be able to view their own information. Some frameworks stipulate that only correct information about a data subject be collected in the database.

Validity refers to the accuracy and correctness of data for its application. Traditionally, this has been referred to as 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 datasets in ways not envisioned by system designers.

Invalid uses of Big Data can be malicious or unintended. 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 billion in wasted advertisement spending [26]. A software executive listed seven different types of online ad fraud, including nonhuman-generated impressions, nonhuman-generated clicks, hidden ads, misrepresented sources, all-advertising sites, malicious ad injections, and policy-violating content such as pornography or privacy violations. [27] Each of these can be conducted at Big Data scale and may require Big Data solutions to detect and combat.

While not malicious, 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. [28] 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, misinterpreted, or intentionally corrupted to malicious intent.

2.3.5 VOLATILITY

Volatility of data—how data structures change 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), [29] there are few standards beyond simplistic time stamps and even fewer common practices available as guidance. To manage security and privacy for long-lived Big Data, data temporality should be taken into consideration.

2.4 EFFECTS OF EMERGING TECHNOLOGY ON BIG DATA SECURITY AND PRIVACY

2.4.1 CLOUD COMPUTING

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 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 consumers
- Dynamic system boundaries and commingled roles and responsibilities between consumers and providers
- Multi-tenancy;
- Different organizations are responsible for different parts of one system;
- Data residency;
- Measured service; and
- Order-of-magnitude increases in scale (e.g., on demand), dynamics (e.g., elasticity and cost optimization), and complexity (e.g., automation and virtualization).

These cloud computing characteristics often present different security risks to an organization than the traditional IT 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 request from the cloud service providers, through contractual means and service-level agreements, that all required security components and controls be fully and accurately implemented. A further discussion of internal security considerations within cloud ecosystems can be found in Appendix C. Future versions of this document will contextualize the content of Appendix C in the NBDRA.

Even though cloud computing is driving innovation in technologies that support Big Data, some Big Data projects are not in the cloud. However, because of the resurgence of the cloud, considerable work has been invested in developing cloud standards to alleviate concerns over its use. A number of organizations, including NIST, are diligently engaged in standards work around cloud computing. Central among these for Big Data security and privacy is NIST SP 800-144 [30], which included a then-current list of related standards and guides, which is reproduced in Appendix C. In the EU, the European Telecommunications Standards Institute (ETSI) produced the Cloud Standards Coordination Report. [31] More recently, the Defense Information Systems Agency (DISA) at the U.S. Department of Defense (DoD) published its Cloud Security Requirements Guide [32], which covers DoD projects through the secret level.

On the privacy front, when the Federal Chief Information Officer (CIO) Council published recommendations for Digital Privacy Controls [33], Big Data received a mention in a footnote:

> *"The potential for re-identifying, tracing, or targeting individuals may arise from the application of predictive analyses and other "data mining" techniques to "big data" (i.e., the increasing availability of vast amounts of stored and streaming digital information). See, e.g., NIST Data Mining Portal (describing ongoing programs, projects, and workshops), http://www.nist.gov/data-mining-portal.cfm. Agencies should ensure that their PIAs for digital services and programs consider whether data mining could be used to identify, trace or target individuals, and be aware of statutory reporting obligations when engaged in data mining for the detection of criminal or terrorist activities. See GAO, Data Mining; Agencies Have Taken Key Steps to Protect Privacy in*

> *Selected Efforts, but Significant Compliance Issues Remain (Aug. 2005) (noting need for agencies to provide proper notice and perform PIAs), http://www.gao.gov/new.items/d05866.pdf; Federal Agency Data Mining Reporting Act of 2007, 42 U.S.C. 2000ee3 (requiring the reporting to Congress of pattern-based queries, searches, or analyses of one or more databases by or on behalf of the Federal Government to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals) (p. 10).”*

2.4.2 BIG DATA SECURITY AND PRIVACY SAFETY LEVELS

Following the practice of standards work elsewhere, this document offers guidance to enterprises wishing to commit to improving security practices. During work on Version 2, an understanding emerged from discussions within the Security and Privacy Subgroup of the links between safety and security. This link is increasingly noted in the literature. For example, Draeger noted [34]:

> *"The close connection between safety and security has led to a growing interest in a combined handling of these two areas of research … The conditions enabling a combined safety and security analysis are identified and used as starting point of the elaboration. Utilizing these properties, a theoretical framework unifying key aspects of both safety and security is developed, whereby a model-based approach is chosen.”*

The Security and Privacy Subgroup proposes the NIST Big Data Security and Privacy Safety Levels (NBD-SPSL), which contains three levels of conformance to security safety practices for Big Data security and privacy. The initial development work on the NBD-SPSL is presented in Appendix A and contains some Big Data security and privacy elements with details of the three Big Data security and privacy safety levels. When paired with a checklist and recommended practices, organizations can self-designate their systems as conforming to a level of the NBD-SPSL, as identified in this report.

That safety engineering has a clear counterpart in Big Data security and privacy can be seen by considering the fabric of safety that encompasses commercial and military aviation. Aviation is a complex milieu of human, mechanical, and geospatial aspects, yet aviation has achieved extraordinary safety levels.

A closer look at the analogy between the aviation safety fabric and Big Data security and privacy safety considerations is illustrative. Taken as a whole, the aviation industry (e.g., aircraft and engine manufacturers, Federal Aviation Administration [FAA], airports, airline maintenance, airline crews, travel agents, Transportation Security Administration [TSA]) is one the oldest and most mature Big Data verticals. From the earliest days of automaton, aviation has utilized computer networks and the most modern testing equipment as early adopters. Aviation is distributed globally. Every aircraft down to nuts and bolts is registered by tail number and then monitored for safety incidents throughout its life. Every significant line replaceable unit is numbered and tracked during its life cycle, representing comprehensive traceability.[c] Every instrument is recalibrated periodically. Every licensed pilot is periodically checked out medically and for proficiency. Crews are scheduled within strict safety rules. Currently, all the information is stored in computers federated around the globe. Many terabytes stream from commercial aircraft every day, to ground computers. [35] Currently, ground controllers record much flight data. The digital data is stovepiped and networked globally.

These aviation industry concepts and practices of data collection, traceability, parts registration, and safety monitoring can be translated to analogous elements of Big Data systems. The state of the art in

[c] Some historians believe that the Titanic sank because some of the rivets used were substandard, which could be proven by tracing the rivets to their original point of manufacture. http://www.bbem.com/military-hardware-traceability

aviation Big Data for operational analytics is dynamic and expanding. [36] Someday, future Big Data generating elements, functional components, and other pieces of the Big Data ecosystem might be as closely monitored as aircraft, flights, pilots, and air crews. At present, most nascent cyber-physical systems (CPSs), including IoT items, are very far removed from a regulated and enforced Big Data-driven environment. Much work remains before artificial intelligence (AI) systems and Big Data achieve acceptable security safety levels.

Extensive literature surveys have demonstrated an intimate connection between “methods, models, tools and techniques” employed in safety engineering and “transposed to security engineering, or vice versa.” [37] The Piètre-Cambacédès & Bouissou study observed the following.

> *“A careful screening of the literature (this paper contains 201 references) made it possible to identify cross-fertilizations in various fields such as architectural concepts (e.g., defense in depth, security or safety kernels), graphical formalisms (e.g., attack trees), structured risk analyses or fault tolerance and prevention techniques” (p. 110).*

The time for a Big Data security and privacy safety framework has arrived—to protect not only the public but also its practitioners enmeshed in a complex web of engineering and marketing of Big Data. The proposed NBD-SPSL is intended to serve as an accessible first step.

2.4.3 INTERNET OF THINGS AND CPS

The Big Data security and privacy community has identified relevant intersections with work in IoT security and crosswalks to related standards efforts in those communities at NIST [38] and elsewhere.

Methods to secure individual IoT devices fall outside the scope of the NBDRA; however, it is worthwhile to note that IoT devices present unique security challenges due to limited hardware capability, rapid market evolution, and lack of a widely used security standard. While some progress has been made with industrial devices [39], [40], consumer device manufactures have no regulatory or market incentive to secure their devices.

Until IoT hardware reaches sufficient maturity to allow TLS communication and support other cryptographic authentication mechanisms, IoT data required for a BDRA will typically be collected under a single provider per device type or class. Volume and Velocity for an individual IoT device are low, due to power and processing constraints, though in an aggregate provider, very high volumes are easily realized. Veracity of this provider is strongly dependent on hardware and protocol implementation details, which might be opaque to relying Big Data consumers.

IoT aggregate NBDRA Data Providers should authenticate individual IoT device connections prior to accepting data wherever possible. While statistical analytics might detect a security breach, relying on this alone is undesirable as it lacks means to distinguish between individual and compromised devices – resulting in a complete loss of functionality in the event of a breach.

2.4.4 MOBILE DEVICES AND BIG DATA

On its face, mobile devices are simply an evolution of decades-old concepts in distributed computing. While this is undeniable, there are certainly lessons in distributed computing that must be updated for current security concerns. Mobile must be viewed as a critical element of Big Data.

Although mobile spans many facets of computer security, there are several reasons for addressing mobile in any comprehensive Big Data security and privacy checklist, including the following:

- Mobile devices challenge governance and controls for enterprises, especially in BYOD (bring your own device) environments. As a result, specialized security approaches enabling mobile-centric access controls have been proposed. [41]

- Some web-based and desktop applications may be migrated to mobile versions without adequate security and privacy protections.
- Mobile devices are less subject to physical security protection, yet they can access Big Data systems as well as any desktop.
- Many organizations lag in the control of mobile device security, preferring to focus on server and desktop security, which has a longer history and is more profitable for tools suppliers.
- Mobile devices often disclose geospatial data, which can be used in Big Data settings to enrich other datasets, and even to perform deanonymization.

2.4.5 INTEGRATION OF PEOPLE AND ORGANIZATIONS

The Security and Privacy Fabric did not integrate the ways in which people and organizations impact Big Data workflow and contribute to the strength or weakness of a Big Data system's security and privacy.

To communicate across organizations, eXtensible Markup Language (XML)-based solutions should be considered. For example, Lenz and Oberweis suggested using an XML variant of Petri nets. [42] They point out that, "Due to the fast growth of Internet-based electronic business activities, languages for modeling as well as methods for analyzing and executing distributed business processes are becoming more and more important. Efficient inter-organizational business processes in the field of ecommerce require the integration of electronic document interchange and inter-organizational process management." (p. 243) [42]

Similarly, Hypertext Markup Language (HTML) microdata can be used to transfer or house information exchanged across organizational boundaries [43]. Microdata has been extended for use with Resource Description Framework (RDF) [44].

The Security and Privacy Subgroup looked at a body of research that addressed concerns for digital systems sharing across organizations. The scope is considerable. Information sharing is key to exchanges in finance, supply chain, healthcare, emergency services, and defense. [45]

That said, in mature systems such as the Enterprise Data Management (EDM) Council's Financial Industry Business Ontology (FIBO; https://www.edmcouncil.org/financialbusiness), the issues of Big Data security and privacy, despite its regulatory facets, may be understated. Additional work is needed to ensure that such frameworks address security and privacy knowledge representation—thus permitting automated reasoning about some aspects of a Big Data system's level of compliance, as well as facilitating comparisons across Big Data security and privacy frameworks by deployment of a unifying model.

Various Institute of Electrical and Electronics Engineers (IEEE) and ISO standards address organizational, life cycle, and systems development processes (e.g., ISO 15288). It remains as an open task to consider if and how such standards affect Big Data security and privacy and whether improvements are needed to enhance Big Data security and privacy safety.

2.4.6 SYSTEM COMMUNICATOR

Big Data systems that collect, store, manage, or transform data considered in need of protection (e.g., data called out as payment card industry [PCI]) should be designed with accessible portals that enable classes of persons to review their own data, direct its removal or extraction, and to understand how it is being used.

The System Communicator is one of the elements in the NBD-SPSL. Additional work is needed to identify how System Communicator requirements should be crafted to meet both usability objectives (e.g., for public PII) and interoperability requirements to work with legacy as well as greenfield Big Data applications.

By providing a System Communicator capability that can be accessed by all stakeholders—potentially including software agents, as well as human stakeholders—Big Data systems can be made more transparent, responsive to citizen- or stakeholder-initiated data correction, and offer feature continuity for such capabilities as data and code moves between organizations.

2.4.7 ETHICAL DESIGN

Journalists, as well as technologists, have decried the apparent lack of ethical standards in Big Data. The incorporation of ethical, and often technical, guidelines is part of ISO 27500 and a suite of IEEE working groups, especially P7000, P7002, P7003, and P7007. As the work of these teams proceeds, features and capabilities that enhance the Security and Privacy Fabric and add to the NBD-SPSL will surface. The subsections below touch on a few aspects of ethical design. Other topics may be investigated in the future, including relations to systems management.

2.4.7.1 Self-Cleansing Systems

Some reports suggest that as much as 20% of the data in global firms is less than fully reliable. This citation is repeated in a proposal by Khayyat et al. [46], in which the case is made for self-cleansing Big Data systems. The presence of erroneous or misleading information, such as citizens mistakenly placed on terrorist watch lists or falsely connected to other criminal activities, is a Big Data security and privacy problem.

Their work and other research [47] reflect increased attention to data quality, data curation, and its associated risk. Connecting these recommendations to the NBDRA to improve security and privacy is a future need.

2.4.7.2 The Toxic Data Model

In other fields of study, *toxicity* is employed as a construct to help represent risk associated with a material or process. An analogous approach to high-risk data is suggested in Appendix A. Data elements should be assessed based on their toxicity. For example, a U.S. passport number or an HIV diagnosis on an electronic health record could be said to have high toxicity. A standard, based on the well-established Material Safety Data Sheets, should be employed for data elements in Big Data systems.

For instance, the U.S. Department of Labor, Occupational Safety and Health Administration promulgates a standard communication format for chemical hazards (https://www.osha.gov/Publications/OSHA3514.html). Future standards could specify the content and format that should accompany Big Data elements shared across and within enterprises. Recipients of a data communications might decline to accept certain types of Big Data, or recognize what changes would be required in their systems and processes to accommodate *toxic* data. System and process changes, for information-intensive organizations such as the U.S. Census Bureau or social media firms, could prove essential to their mission.

2.4.7.3 Big Data Security Safety Annotation

Risk management and federation of security safety practices are two areas for future study of the Security and Privacy Subgroup. The Subgroup recognized the need to reassess the NIST and ISACA Risk Management frameworks to understand what changes may be needed in these widely used frameworks to better address Big Data security and privacy.

Federation is key to information supply chains. Most of the world's global enterprises and governments rely upon extensive information system supply chains, yet managing these to ensure security and privacy is challenging. A review of currently available approaches is needed. One approach is seen in marketplace notions (e.g., closed clearinghouses, federation as an engineering principle, InCommon, GENI.net, Organization for the Advancement of Structured Information Standards [OASIS] IDTrust). However, sometimes there will also be requirements for out-of-band guest identity, such as for emergencies, regulatory, or other exceptional circumstances.

2.4.7.4 Big Data Trust and Federation

Federation and trust are aspects of information sharing. These are sometimes explicit, sometimes not. The level of detail exchanged between organizations varies wildly. Some limit themselves to a one-off exchange of keys. One research team has suggested the use of *transactional memory* managed through the use of cloud brokers. [48]

The scope of this document is necessarily limited, whereas there are entire disciplines within computing dedicated to various aspects of federation.

Middleware, message-passing, and enterprise service bus remain important concepts for Big Data. For example, in SE-CLEVER investigators wanted to address issues raised by the Cloud Security Alliance in their Extensible Messaging and Presence Protocol (XMPP)-based middleware. [49]

Enterprises large and small will increasingly automate functions and share information, creating new and varied Big Data sources. Even for relatively mature organizations, federation across a supply chain or customer federation multiplies threats while governance, risk management, and compliance (GRC) is weakened. That weakening is a necessary byproduct of cross-organization sharing, but still a risk. While shared standards, mutual open dialog, and other socialization and training techniques matter, systems must be put in place that operate across organizational boundaries.

2.4.7.5 Orchestration in Weak Federation Scenarios

Orchestration design patterns may be needed for weak federation scenarios. How these interact with broad orchestration for Big Data (e.g., Kubernetes, Topology and Orchestration Specification for Cloud Applications [TOSCA]) requires further study.

2.4.7.6 Consent and the Glass-Breaking Scenario

The *glass-breaking* scenario is important to Big Data security and privacy because it identifies the need for systematically framed exceptions to security and privacy hardening.

In healthcare standards such as Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR; http://hl7.org/fhir/), glass-breaking may be needed to save a life in an emergency. The emergency itself occasions a different security and privacy context, which a Big Data application may need to recognize.

The importance of geospatial Big Data for emergency management is acknowledged [50], [51], and the absence of consent to single out disabled individuals in a high-rise fire point to nuanced rules around information access, as well as the velocity of the underlying decision support infrastructure.

An abuse-resistant glass-break mechanism for time-critical situations (such as fires, medical emergencies) across multiple Providers may require machine learning, as policy reconfiguration for even a highly skilled human operator would take too long, or be too easy to bypass. The mechanism must have strong authentication and non-repudiation, with the identity, location, and motive of the initiator preserved permanently through a cryptographic mechanism (such as blockchain).

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 *NBDIF: Volume 3, Use Cases and General Requirements*. However, the primary focus of these use cases was on science and engineering applications, and therefore, security and privacy impacts on system architecture were not highlighted. Consequently, a different set of use cases, presented in this document, was developed specifically to discover security and privacy issues. Some of these use cases represent inactive or legacy applications, but were selected to 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 security and privacy use cases, collected when the reference architecture was not mature, were provided by NBD-PWG members 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.

Additional security and privacy use cases were collected (in the same format as the original security and privacy use cases) during Version 2 work, which have helped guide the development of the NBD-SPSL. However, the need for more specific and standardized use case information lead to the creation of a new use case template.

During Version 2 activities, the Security and Privacy Subgroup collaborated with the Use Cases and Requirements Subgroup to develop the new Use Case Template 2, which is currently being used to collect additional use cases. In addition to questions from the original use case template, the Use Case Template 2 contains questions aimed at providing a comprehensive view of security, privacy, and other topics for each use case. Use cases submitted through the Use Case Template 2 will greatly assist the Subgroups in strengthening future work. To submit a use case, please fill out the PDF form (https://bigdatawg.nist.gov/_uploadfiles/M0621_v2_7345181325.pdf) and email it to Wo Chang (wchang@nist.gov). Use cases will be evaluated as they are submitted and will be accepted until the end of Phase 3 work.

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 to reach customers who would buy their products.

Current Security and Privacy Issues/Practices: 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:
  - Controls are inconsistent and/or not established to appropriately achieve the following objectives:
    - Predictability around the processing of personal information, to give individuals a reliable sense of how their information is processed and enable them to make appropriate determinations for themselves, or prevent problems arising from actions such as unanticipated revelations about individuals
    - Manageability of personal information, to prevent problems arising from actions such as dissemination of inaccurate information
    - Controls may not address the inability of some consumers to access information about themselves that is available to enterprises or governments
    - Unlinkability of information from individuals to prevent actions such as surveillance of individuals
- Security:
  - Controls are inconsistent and/or not established appropriately to achieve the following:
    - Isolation, containerization, and encryption of data
    - Monitoring and detection of threats, as well as incident handling
    - 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.” [52] 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) for consumer digital media usage settings such as these.

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 dataset. 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 Issues/Practices:

- 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 Issues/Practices:

- Opt-in defaults are relied upon in some countries to gain visitor consent for tracking of website visitor IP addresses. In some countries Internet Protocol (IP) address logging can allow analysts to identify visitors down to levels as detailed as latitude and longitude, depending on the quality of the maps and the type of area being mapped.
- 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 in some countries 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. [53]

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 is through strong cryptography and key management to meet the HIPAA requirements for protected health information (PHI). Ideally this does not require the cloud service operator to sign a business associate agreement (BAA). Cloud usage would provide several benefits, including patient safety, lowered healthcare costs, and regulated accesses during emergencies.

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 non-repudiation 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; U.S. 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.
- Cross-government and cross-industry sharing

Current Security and Privacy Issues/Practices:

- 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 is not widely deployed but is anticipated by some experts as a medium-term practice. [54]
  - 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 Issues/Practices:

- Security:

  - Secure Sockets Layer (SSL)/ Transport Layer Security (TLS)-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 life cycle data ownership and custody controls

3.2.3 PHARMA CLINICAL TRIAL DATA SHARING [55]

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 Security and Privacy Issues/Practices:

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 datasets, such as event logs, across thousands of servers. 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.

The roles that Big Data plays in protecting networks can be grouped into two broad categories:

- Security for Big Data: When launching a new Big Data initiative, new security issues often arise, such as a new attack surface for server clusters, user authentication and access from additional locations, new regulatory requirements due to Big Data Variety, or increased use of open source code with the potential for defaulted credentials or other risks. [56]
- Big Data for security: Big Data can be used to enhance network security. For example, a Big Data application can enhance or eventually even replace a traditional Security Information and Event Management (SIEM). [57]

Current Security and Privacy Issues/Practices:

- Security
  - Big Data security in this area is under active research, and maintaining data integrity and confidentiality while data is in-motion and/or at-rest warrants constant encryption/decryption that works well for Small Data, but is still inadequate for Big Data. In addition, privacy concepts are even less mature.
  - 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:
  - Need to consider enterprise practices for data release to external organizations
  - Lack of protection of PII data

Currently vendors are adopting Big Data analytics for mass-scale log correlation and incident response, such as for SIEM.

3.4 GOVERNMENT

3.4.1 UNMANNED VEHICLE SENSOR DATA

Scenario Description: Unmanned Aerial Vehicles (UAVs), also called Remotely Piloted Vehicles (RPVs) or Unmanned Aerial Systems (UAS), can produce petabytes of data, some of it streamed, and often stored in proprietary formats. These streams, which can include what in military circles is referred to as full motion video, are not always processed in real time. UAVs are also used domestically. The Predator drone is used to patrol U.S. border areas, and sometimes flood areas; it allows authorized government workers to see real-time video and radar. [58]

Current Security and Privacy Issues/Practices:

- Military UAV projects are governed by extensive rules surrounding security and privacy guidelines. Security and privacy requirements are further dictated by applicable service (Navy, Army, Air Force, Marines) instructions. [59]
- Not all UAV data uses are military. For example, NASA, National Oceanic and Atmospheric Administration and the FAA may have specific use for UAV data. Issues and practices regarding the use of sensor data gathered non-DoD UAVs is still evolving, as demonstrated by a draft U.S. Department of Justice (DOJ) policy guideline produced by the DOJ Office of Legal Policy. [60] The guideline acknowledges the value of UAS data as “a viable law enforcement tool” and predicts that “UAS are likely to come into greater use.” The draft reiterates that UAS monitoring must be consistent with First and Fourth Amendment guarantees, and that data “may only be used in connection with properly authorized investigations.” Additional guidance addresses PII that has been collected, such that it cannot be retained for more than 180 days except when certain conditions are met. Annual privacy reviews and accountability for compliance with security and privacy regulations are prominent in the draft.
- Collection of data gathered by UAVs outside of the United States is subject to local regulation. For example, in the EU, guidelines are under discussion, which incorporate Remotely Piloted Aircraft Systems in the European Aviation System. The EU sponsored a report addressing potential privacy, data protection, and ethical risks related to civil Remotely Piloted Aircraft System (RPAS) applications (http://ec.europa.eu/enterprise/sectors/aerospace/uas /).

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 datasets 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 Issues/Practices:

- 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. As set forth in the Data Quality Campaign, protecting student data is seen as a state education agency responsibility: to define “the permissible collection and uses of data by external technologies and programs used in classrooms.” This source identifies additional resources for safeguarding student data and communicating with parents and staff about data and privacy rights. [61]
- Privacy-related disputes surrounding education Big Data are illustrated by the reluctance of states to participate in the InBloom initiative. [62]
- 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. [63]

Current Research:

- Longitudinal performance data would have value for program evaluators and educators. Work in this area was proposed by Deakin Crack, Broadfoot & Claxton [64] as a “Lifelong Learning Inventory,” and further by Ferguson, [65] whose reference to data variety observed that

"Increasingly, learners will be looking for support from learning analytics outside the Virtual Learning Environment or Learning Management System, whilst engaged in lifelong learning in open, informal or blended settings. This will require a shift towards more challenging datasets and combinations of datasets, including mobile data, biometric data, and mood data. To solve the problems faced by learners in different environments, researchers will need to investigate what those problems are and what success looks like from the perspective of learners." [65]

- Data-driven learning [66] 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 [67] 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 GB 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. Companies also prefer to control how, when, and with whom the data is shared, even for analytics purposes. Many of these analytics systems are now being moved to infrastructure cloud providers.

Current Security and Privacy Issues/Practices:

- 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)
- Protocols for API security, such as OAuth 2.0

Researchers are currently investigating the following security enhancements:

- Virtualized infrastructure layer mapping on a cloud provider
- Homomorphic encryption
- Quorum-based encryption
- Multiparty 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 ISO 29161 standards under development within the ISO joint technical committee (JTC) ISO JTC1 SC31 WG2. There are likely other standards evolving in parallel. 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.

3.7 ADDITIONAL SECURITY AND PRIVACY USE CASES

The following use cases were collected to further inform the work of the Security and Privacy Subgroup. These use cases were in the initial phases of collection when the need for the Use Case Template 2 arose. Therefore, the use cases have not been as fully developed as the previously presented use cases that were collected during Version 1 work. However, the information provided below contains valuable information that guided Version 2 work, including formation of the NBD-SPSL.

The NBD-PWG invites the public to submit additional use cases to help strengthen future work on the NBDIF. To submit a use case, please fill out the PDF form (https://bigdatawg.nist.gov/_uploadfiles/M0621_v2_7345181325.pdf) and email it to Wo Chang (wchang@nist.gov). Use cases will be evaluated as they are submitted and will be accepted until the end of Phase 3 work. Additional use cases could be submitted for the following topics:

- Object Management Group (OMG) Data Residency Initiative. The Cloud Standards Customer Council and OMG collaboration produced an extensive use case matrix [14] (p.10)
- Emergency management data (XChangeCore interoperability standard)
- Healthcare Consent Flow
- Audit aspects of consent are presented in documents developed by an HL7 community [68]
- “Heart Use Case: Alice Selectively Shares Health-Related Data with Physicians and Others”[d]
- Blockchain for Fintech
- In-Stream PII
- A widely used chat application is moved from the desktop to mobile environment, potentially exposing PII in a different setting.

3.7.1 SEC Consolidated Audit Trail

The SEC Consolidated Audit Trail (CAT) project is forecast to consume 10 terabytes of data daily (SEC Rule 613 https://www.sec.gov/divisions/marketreg/rule613-info.htm). The system’s security requirements, which stemmed from a past system failure with lack of traceability, are considerable. Figure 1 presents the High-Level CAT Security Requirements[e].

[d] https://bitbucket.org/openid/heart/wiki/Alice_Shares_with_Physicians_and_Others_UMA_FHIR

[e] Source: http://www.catnmsplan.com/web/groups/catnms/@catnms/documents/appsupportdocs/cat_nms_security_requirements_032416.pdf

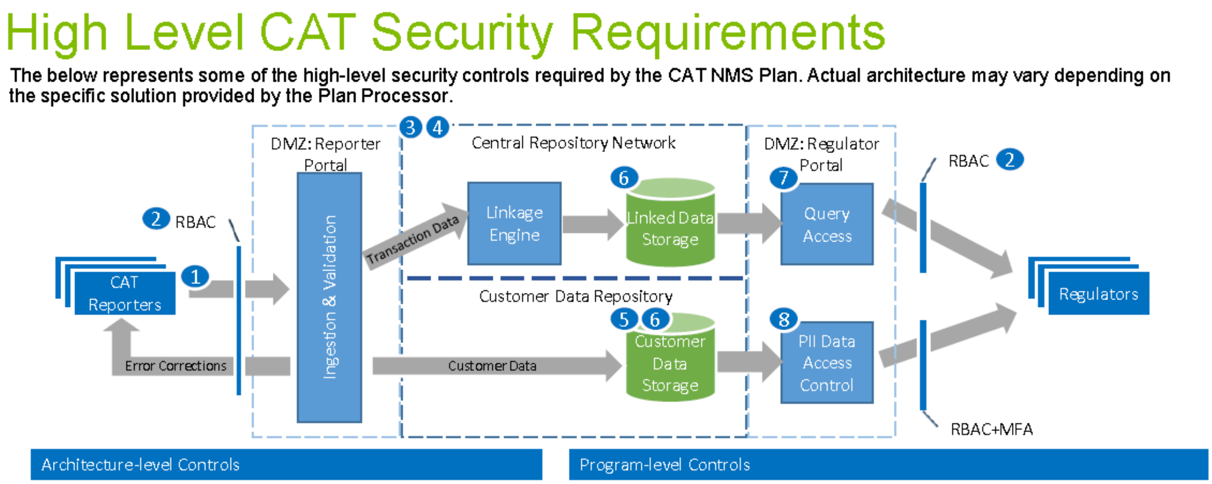


Figure 1: High-Level CAT Requirements

3.7.2 IoT Device Management

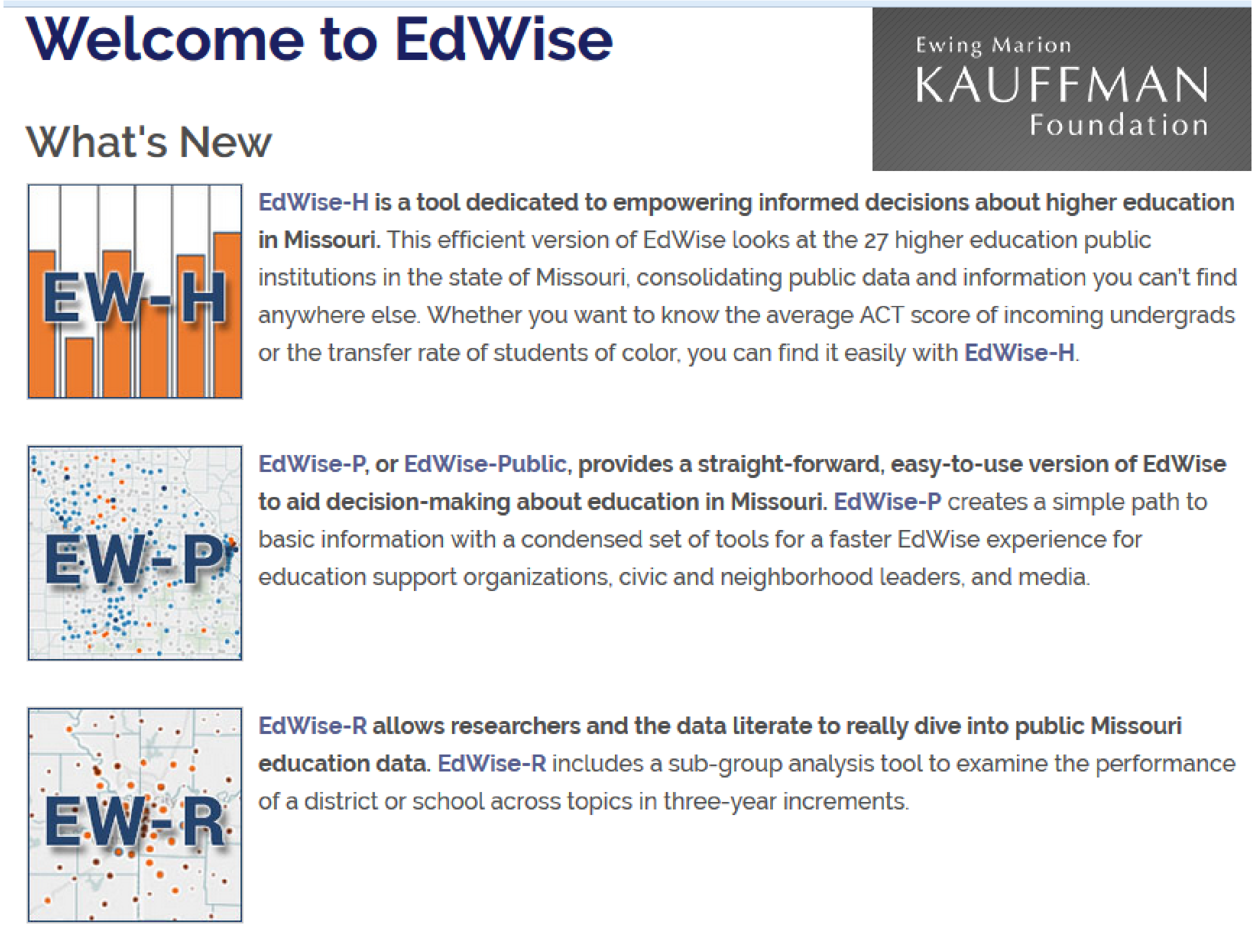
This family of use cases involves the onboarding, decommissioning, and/or quarantining of numerous devices, such as for IoT and CPS. The sheer number of devices and the limited defenses against tampering that low-cost devices can incorporate, put Big Data systems at risk.

Safety systems incorporating voluminous sensor streams represent this family of use cases. Preliminary research addressing IoT safety is already under way [69], [34] and [70]. The latter work was reported during an international conference now more than a decade old, the International Conference on System Safety and Cybersecurity.

One application of IoT is in smart homes. Smart homes allow for remote monitoring through Wi-Fi networks and present new Big Data sources and new attack surfaces for private residences, government facilities, and other entities.

3.7.3 Statewide Education Data Portal

The Kauffman Foundation EdWise web resource provides public access to higher education data for consumers, parents, support organizations, and leaders. It is a data aggregator as well as an analytics portal [71]. The portal attempts to provide anonymized student and institutional performance data for educational decision support.



Welcome to EdWise

What's New

EdWise-H is a tool dedicated to empowering informed decisions about higher education in Missouri. This efficient version of EdWise looks at the 27 higher education public institutions in the state of Missouri, consolidating public data and information you can't find anywhere else. Whether you want to know the average ACT score of incoming undergrads or the transfer rate of students of color, you can find it easily with **EdWise-H**.

EdWise-P, or EdWise-Public, provides a straight-forward, easy-to-use version of EdWise to aid decision-making about education in Missouri. EdWise-P creates a simple path to basic information with a condensed set of tools for a faster EdWise experience for education support organizations, civic and neighborhood leaders, and media.

EdWise-R allows researchers and the data literate to really dive into public Missouri education data. EdWise-R includes a sub-group analysis tool to examine the performance of a district or school across topics in three-year increments.

Figure 2: EdWise Figure

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. [19] Candidate topics and related material used in preparing this section are provided in Appendix C.

A taxonomy for Big Data security and privacy should encompass the aims of existing useful taxonomies. While many concepts surround security and privacy, 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 3. The focus has been predominantly on security and security-related privacy risks (i.e., risks that result from unauthorized access to personally identifiable information). Privacy risks that may result from the processing of information about individuals, and how the taxonomy may account for such considerations, will be explored in greater detail in future versions.

4.1 CONCEPTUAL TAXONOMY OF SECURITY AND PRIVACY TOPICS

The conceptual security and privacy taxonomy, presented in Figure 3, 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.

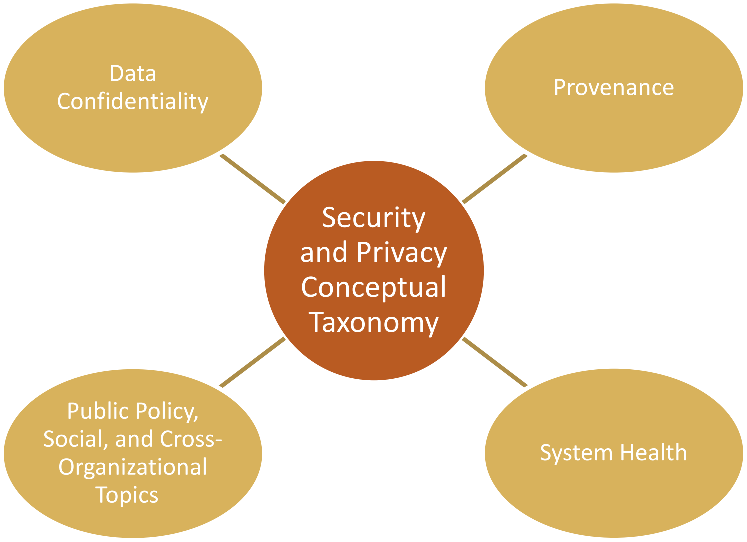

Data Confidentiality
Provenance
Security and Privacy Conceptual Taxonomy
Public Policy, Social, and Cross-Organizational Topics
System Health

Figure 3: 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, such as Functional Encryption [72] 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, [73] 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.” [73]
  - Big Data systems are not purely cloud systems, but as 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. [74]
  - Challenges for Big Data key management systems (KMS) reflect demands imposed by Big Data characteristics (i.e., volume, velocity, variety, and variability). For example, relatively slow-paced 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

In a separate discussion, the interwoven notions of design, development, and management are addressed directly. A Big Data system likely requires additional measures to ensure availability, as illustrated by the unanticipated restore time for a major outage [75].

- System availability is a key element in CIA—Security against denial of service (DoS)
  - Construction of cryptographic protocols (developed with encryption, signatures, and other cryptographic integrity check primitives) proactively resistant to DoS
- System Immunity—Big Data for Security
  - Analytics for security intelligence
  - Data-driven abuse detection
  - Big Data analytics on logs, cyber-physical 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. [76] 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[f]
- 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

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 4, 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 NBDIF 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 Revision 1, *Guide for Applying the Risk Management Framework to Federal Information Systems*, [77] NIST Internal Report (NISTIR) 8062, *An Introduction to Privacy Engineering and Risk Management in Federal Systems*, [78] and COBIT Risk IT Framework. [79]

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. [80]

The operational taxonomy will overlap with small data taxonomies while drawing attention to specific issues with Big Data. [81] [82]

[f] 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)).

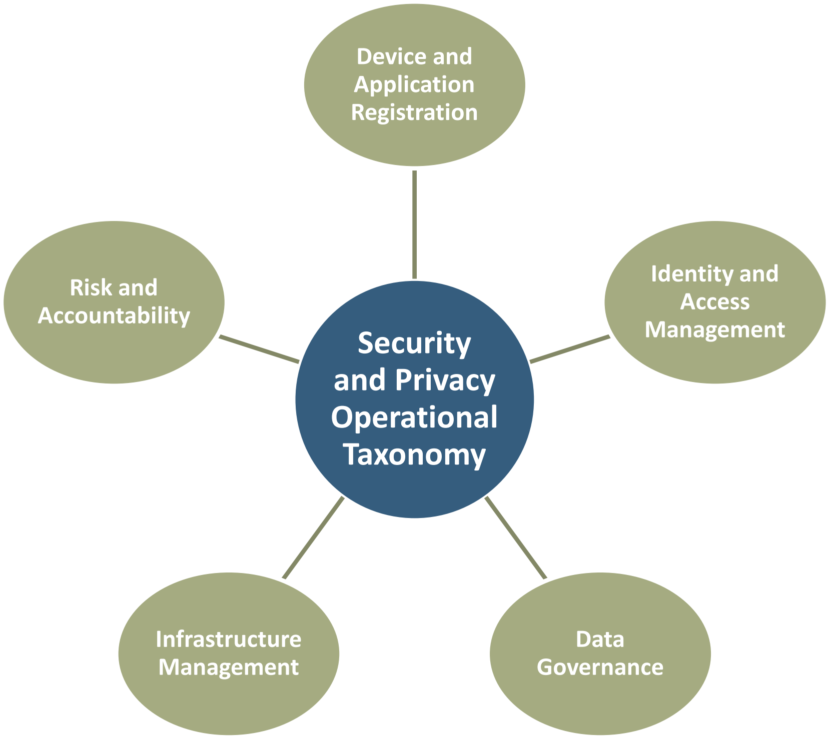

Device and Application Registration
Risk and Accountability
Security and Privacy Operational Taxonomy
Identity and Access Management
Infrastructure Management
Data Governance

Figure 4: 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, 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). [81] 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. Data governance refers to administering, or formalizing, discipline (e.g., behavior patterns) around the management of data. 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 life cycle 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” [83]
    - De-identification and anonymization
    - Re-identification risk management
- End-point validation
- DRM
- Trust
- Openness
- Fairness and information ethics [84]

4.2.3.1 Compliance, Governance and Management as Code

The Fedramp-related initiative Open Control seizes upon the connection between increased use of automation for all facets of today’s systems. Its proponents argue for the following progression:

- Software as code,
- Tests as code,
- Infrastructure as code, and
- Compliance as code.

Just as software-defined network (SDN) can be seen as a way to create and manage infrastructure with reduced manual intervention, Open Control was used by GSA’s lean startup-influenced digital services agency 18F to facilitate *continuous authorization*. Continuous authorization is seen as logically similar to agile’s *continuous deployment*. The 18F team employs YAML to implement a *schema* which is publicly available on GitHub.

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 NIST Critical Infrastructure Cybersecurity Framework, Big Data affords new opportunities for large-scale security intelligence, complex event fusion, analytics, and monitoring.
- Mitigation
    - Breach mitigation planning for Big Data may be qualitatively or quantitatively different.
- Configuration Management
    - Configuration management is one aspect of preserving system and data integrity. It can include the following:
    - Patch management

    - Upgrades
- Logging
    - Big Data must produce and manage more logs of greater diversity and velocity. For example, profiling and statistical sampling may be required on an ongoing basis.
- Malware surveillance and remediation
    - This is a well-understood domain, but Big Data can cross traditional system ownership boundaries. Review of NIST's "Identify, Protect, Detect, Respond, and Recover" framework may uncover planning unique to Big Data.
- Network boundary control
    - Establishes a data-agnostic connection for a secure channel
        - Shared services network architecture, such as those specified as "secure channel use cases and requirements" in the ETSI TS 102 484 Smart Card specifications [85]
        - 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. [82] 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 both within an organization and across supply chains. Access should include individuals who specialize in cryptography, security, compliance, or IT. In addition, the ideal audience includes domain experts and organization decision makers who understand the costs and impact of these controls. Ideally, written guidelines setting forth policy and compliance for Big Data security and privacy would be prefaced by additional 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 onboarding, 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 aspects typically include topics in identity, authorization, access control, and log aggregation. This is not an exhaustive list.

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, GRC is a function that draws participation from multiple areas of the organization, such as legal, human resources (HR), 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 in which Big Data operates.

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. Information workers 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 (DHHS) 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."* [86]

Organizations using or sharing health data among ecosystem partners, including mobile apps and SaaS providers, may need to verify that the proper legal agreements are in place. Compliance may be needed to ensure data veracity and provenance. [87]

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 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 United States by federal agencies such as the FTC, the U.S. Food and Drug Administration (FDA), or the DHHS Office of National Coordinator. Examples of agency responsibilities or oversight are:

- DHS is responsible for aspects of domestic U.S. computer security through the activities of US-CERT (U.S. Computer Emergency Readiness Team). US-CERT describes its role as “[leading] efforts to improve the Nation's cybersecurity posture, coordinate cyber information sharing, and proactively manage cyber risks to the Nation while protecting the constitutional rights of Americans.” [88]
- The Federal Trade Commission offers guidance on compliance with the Children’s Online Privacy Protection Act (COPPA) via a *hot line* (CoppaHotLine@ftc.gov), with website privacy policies, and compliance with the Fair Credit Reporting Act. The Gramm-Leach-Bliley Act, Red Flags Rule, and the US-EU Safe Harbor Framework. [89]
- The DHHS Office of National Coordinator offers guidance and regulations regarding health information privacy, security and health records, including such tools as a Security Risk Assessment, HIPAA rule enforcement, and the embedding of HIPAA privacy and security requirements into Medicare and Medicaid EHR Meaningful Use requirements. [90]
- Increased use of EHRs and smart medical devices has resulted in new privacy and security initiatives at the FDA related to product safety, such as the Cybersecurity of Medical Devices as related to the FDA’s Medical Product Safety Network (MedSun). [91]

Social roles include the influence of nongovernmental organizations, 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 topics have been identified but not 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. Free search engines rely on users giving up information about their search terms on a Big Data scale. Search engines and social media sites 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.

4.5.3 ACM TAXONOMY

Where possible, this document uses the terminology adopted by the ACM Computing Classification System [92] and [93]. The ACM 2012 CCS is accessible online [76] and can be represented in Simple Knowledge Organization System (SKOS) format [94]. A snippet of the Security and Privacy Category from the 2012 CSS is presented below.

- Database and storage security
    - Data anonymization and sanitation
    - Management and querying of encrypted data
    - Information accountability and usage control
    - Database activity monitoring
- Software and application security
    - Software security engineering
    - Web application security
    - Social network security and privacy
    - Domain-specific security and privacy architectures
    - Software reverse engineering
- Human and societal aspects of security and privacy
    - Economics of security and privacy
    - Social aspects of security and privacy
    - Privacy protections
    - Usability in security and privacy

A systematic taxonomy has several benefits for Big Data security and privacy. In addition to tracking new research and guidelines (e.g., software and application security snippet from the list above), standardized terminology can, in some limited contexts, allow for automated reasoning. Automated reasoning, based on cybersecurity ontologies, for example, could enable fine-grained alerts, which could elevate as the need arises, while minimizing false positives and less significant events. One approach extended a malware ontology to include elements of *upper ontologies*, which can add *utility*-domain aspects such as temporal, geospatial, person, events, and network operations [95]. Utility domains form part of the NBD-SPSL.

Other taxonomies may be useful. For example, the NISTIR 8085 draft *Forming Common Platform Enumeration (CPE) Names from Software Identification (SWID) Tags* is designed to “support automated and accurate software asset management [96], p. iii.

4.6 WHY SECURITY ONTOLOGIES MATTER FOR BIG DATA

Suppose an engineer inherits software and/or data from a third party. Whether it's within the organization, or across organizations, it's important to know what security components are present in the inherited system. Yet the terminology and underlying components are rarely described in terms that are readily exchanged between practitioners, much less between analysts, SMEs, testers, and users. However, standardizing the terminology is insufficient.

As noted in the literature [95], systematic use of ontologies could enable information security tools to process standardized information streams from third parties, using methods such as the Security Content Automation Protocol (SCAP). This model could enable automated reasoning to address potential breaches closer to real time, or which have indirect effects on networks or applications which require a mixture of human and machine cognition.

While SCAP is mainly used to facilitate alignment between configuration settings and NIST SP 800-53, this approach was not designed for the velocity or volume of Big Data security information. Attempts to integrate real-time logs with internal and external SCAP feeds are likely to encounter scalability challenges, numerous false positives, and crippling information overload from the human computer interaction (HCI) perspective.

DAEDALUS-VIZ was a research project whose architects felt it necessary to build a "novel real-time 3D visualization engine called DAEDALUS-VIZ that enables operators to grasp visually and in real time a complete overview of alert circumstances" [97]. Scaling these projects to Big Data dimensions would tax even the most gifted security analysts.

SIEM and related tools are today relatively unsophisticated in their reasoning capabilities. Big Data demands a more sophisticated framework for security and privacy frameworks than are currently available. As Obrst et al. explain,

> *"Events are entities that describe the occurrences of actions and changes in the real world. Situations represent histories of action occurrences. In this context at least, situations are not equivalent to states. Events and situations are dynamic and challenging to model in knowledge representation systems. As in the temporal and spatial domains, logic formalisms have been created for representing and reasoning about events and situations. These are the event calculus and situation calculus. Both calculi employ the notion of fluents. A fluent is a condition that can change over time. The main elements of the event calculus are fluents and actions, and for the situation calculus they are fluents, actions and situations."* [95]

An arguably chronic weakness in conventional databases is their ability to manage *point in time* representations. Big Data applications allow for unstructured repositories but do not themselves solve the problem of integrating temporal and spatial elements. If network topologies are analogs or even literal spatial representations, it is clear that reasoning about cyber events and situations will require ontological discipline and Big Data. While visualization is often seen as the cure-all for this, Shabtai et al. [98] referred to the real underlying need as "knowledge-based interpretation, summarization, query, visualization and interactive exploration of time-oriented data." Among other requirements, the researchers cite "a domain-specific knowledge base" as an essential component.

As shown in the proposed NBD-SPSL (Appendix A), ontologies that represent knowledge of applications, domains and utility (so-called *middle* and *upper* ontologies) are likely to comprise the most effective means of processing cybersecurity Big Data. Cloud-centric work by Takahashi et al. [99] demonstrated the feasibility of the approach.

Additional ontologies to support privacy will be needed for some Big Data systems. While it did not result in ontologies, at least one project took a model-based systems engineering (MBSE) approach to produce “a model of private information flow and a graphical notation for visualizing this flow are proposed. An application example of using the notation to identify privacy vulnerabilities is given” [100].

5 BIG DATA REFERENCE ARCHITECTURE AND 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 proposed Security and Privacy Fabric was developed.[g] This is geometrically depicted in Figure 5 by the Security and Privacy Fabric surrounding the five main components, since all components are affected by security and privacy considerations. The role of security and privacy is correctly depicted in relation to the components but does not expand into finer details, which may be 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. The *NBDIF: Volume 6, Reference Architecture* document discusses in detail the other components of the NBDRA.

[g] 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.

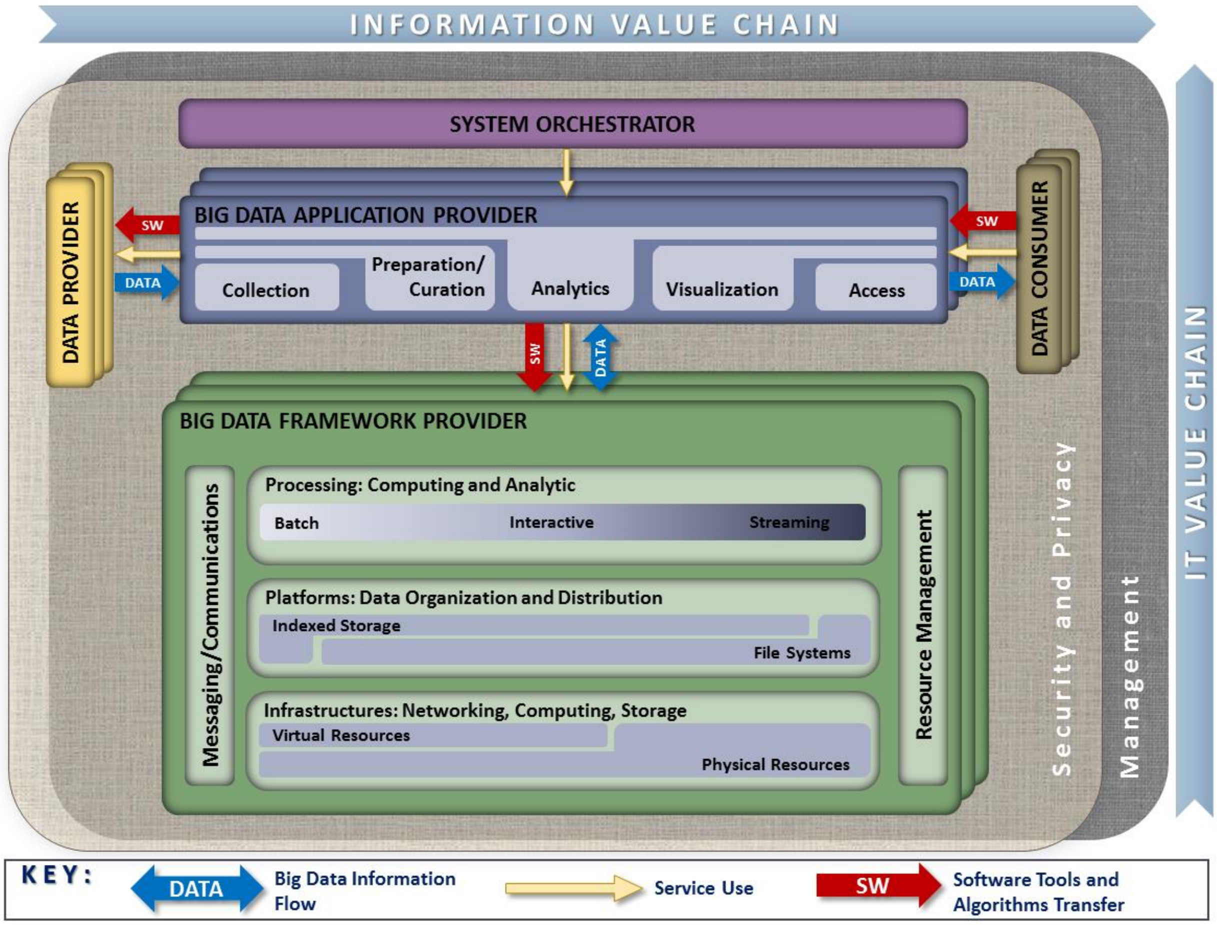


Figure 5: NIST Big Data Reference Architecture

At this time, explanations as to how the proposed security and privacy 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.

Figure 5 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 scaled-out. In the process of scaling, specific security, privacy, or GRC mechanism or practices may need to be invoked.

5.1 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 from the operational taxonomy elements (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 Life Cycle Management • Threat and Vulnerability Management • Mitigation • Configuration Management • Monitoring, Alerting • Malware Surveillance and Remediation • Resiliency, Redundancy, and Recovery • Accountability • Compliance • Forensics • Business Risk Model	Several security functions have been mapped to the System Orchestrator block, as they require architectural level decisions and awareness. Aspects of these functionalities are strongly related to the Security Fabric and thus touch the entire architecture at various points in different forms of operational details. Such security functions include nation-specific compliance requirements, vastly expanded demand for forensics, and domain-specific, privacy-aware business risk models.
Data Provider	
• Device, User, Asset, Services, Applications Registration • Application Layer Identity • End User Layer Identity Management • End Point Input Validation • Digital Rights Management • Monitoring, Alerting	Data Providers are subject to guaranteeing authenticity of data, and in turn require that sensitive, copyrighted, 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 Big Data Application Provider.
Big Data Application Provider	
• Application Layer Identity • Web Services Gateway • Data Transformation • Digital Rights Management • Monitoring, Alerting	The Big Data 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 Big Data Framework Provider.
Big Data Framework Provider	
• Virtualization Layer Identity • Identity Provider • Encryption and Key Management • Isolation/Containerization • Storage Security • Network Boundary Control • Monitoring, Alerting	The Big Data Framework Provider is responsible for the security of data/computations for a significant portion of the life cycle 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 Big Data Application Provider.

5.2 SECURITY AND PRIVACY FABRIC IN THE NBDRA

Figure 6 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 6 will be further developed in future work of this document. However, the relationships between the Security and Privacy Fabric and the NBDRA and the Security and Privacy Taxonomy and the NBDRA could be investigated in future work.

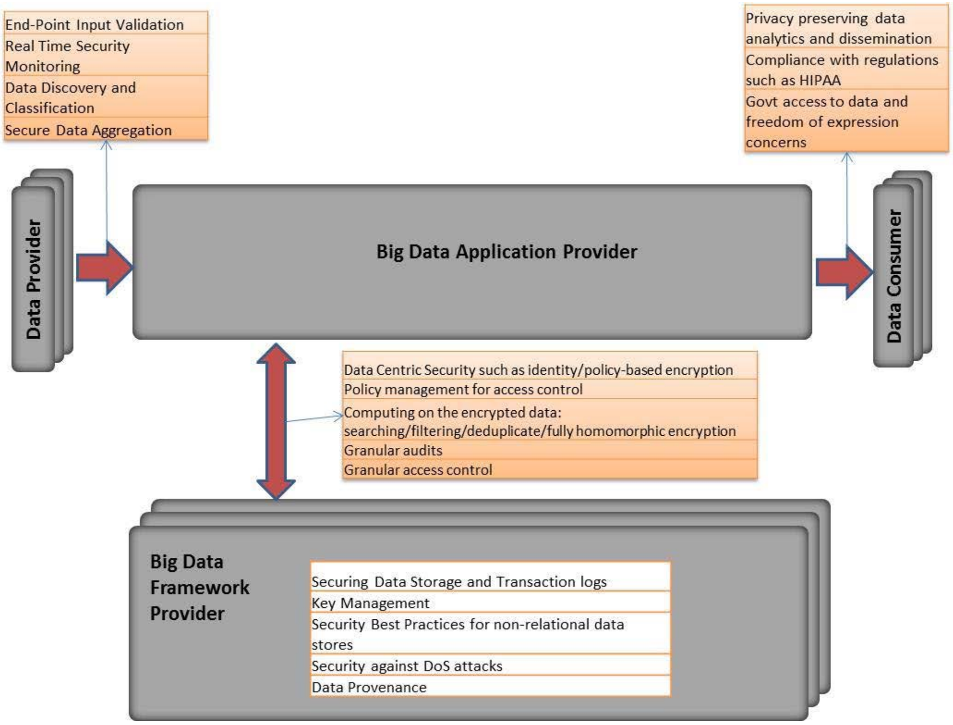

End-Point Input Validation
Real Time Security Monitoring
Data Discovery and Classification
Secure Data Aggregation
Privacy preserving data analytics and dissemination
Compliance with regulations such as HIPAA
Govt access to data and freedom of expression concerns
Data Provider
Big Data Application Provider
Data Consumer
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
Granular audits
Granular access control
Big Data Framework Provider
Securing Data Storage and Transaction logs
Key Management
Security Best Practices for non-relational data stores
Security against DoS attacks
Data Provenance

Figure 6: Notional Security and Privacy Fabric Overlay to the NBDRA

The groups and interfaces depicted in Figure 6 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 INTERFACE WITHIN THE 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. Certifications (not self-signed) should be used to mitigate man-in the-middle attacks.

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.

Appendix E contains mapping of Security and Privacy use cases to the fabric overlay described in Figure 6.

5.3 SECURITY AND PRIVACY FABRIC 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 OASIS Privacy Management Reference Model (PMRM), consisting of seven foundational principles, provides appropriate basic guidance for Big System architects. When working with any personal data, privacy should be an integral element in the design of a Big Data system. Appendix B introduces a comprehensive list of additional security and privacy concepts developed in selected existing standards. There is an intentional emphasis on privacy concepts, reflecting public and enterprise concerns about Big Data security and privacy. Although not all concepts are fully addressed in the current release of this volume, readers may identify particular notions which can focus attention for particular Big Data security and privacy implementations or domain-specific scenarios.

Other privacy engineering frameworks, including the model presented in NISTIR 8062 are also under consideration. [30], [101]–[104]

Related principles include identity management frameworks such as proposed in the National Strategy for Trusted Identities in Cyberspace (NSTIC) [105] and considered in the NIST Cloud Computing Security Reference Architecture. [106] 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.

5.4 SECURITY AND PRIVACY APPROACHES IN ANALYTICS

The introduction to the IEEE P7003 working group notes that "individuals or organizations creating algorithms, largely in regard to autonomous or intelligent systems, [need] certification-oriented methodologies to provide clearly articulated accountability and clarity around how algorithms are targeting, assessing, and influencing the users and stakeholders of said algorithm." (https://standards.ieee.org/develop/project/7003.html)

Big Data analytical and machine learning capabilities are central goals of many Big Data systems, yet not all address the associated security and privacy issues surrounding them. Analysts and the consumers of conclusions reached by Big Data systems require guidance to help interpret and manage visualizations such as dashboards and narratives derived from Big Data systems.

THE CASE OF CRISP-DM

Despite its widespread adoption for Big Data analytics, CRISP-DM has been criticized for its omission of domain-specific processes. For example, Li, et al. [107] point out that even as Big Data has taken hold in hospital information systems, "There are [only] a few known attempts to provide a specialized [CRISP-DM] methodology or process model for applications in the medical domain …" (p. 73).

One of the few cited attempts provides extensions for CRISP-DM, but domain specificity is rare [108]. A result of this lightweight coverage for domain-specific granularity is potentially weak coverage for Big Data security and privacy concerns that emerge from the specifics of that system.

In U.S. healthcare, disclosure of health information associated with HIV/AIDS, alcohol use, or social status is potentially damaging to patients and can put caregivers and analysts at risk, yet CRISP-DM models may not take these issues into account.

Securing intellectual property, reputation, and privacy are concerns for individuals, organizations as well as governments—though the objectives are sometimes in conflict. Risks associated with loss of algorithmic security and lack of transparency are challenges that often are associated with Big Data systems.

Transparency of such systems affects user performance, as a study by Schaffer et al. demonstrated [109]. That said, achieving transparency is not a skill that most developers have attained, and for some domains, transparency has attendant risks that must also be addressed.

5.5 CRYPTOGRAPHIC TECHNOLOGIES FOR DATA TRANSFORMATIONS

Security and privacy of Big Data systems are enforced by ensuring integrity and confidentiality at the datum level, as well as architectural awareness at the fabric level. Diversity of ownership, sensitivity, accuracy, and visibility requirements of individual datum is a defining characteristic of Big Data. This requires cryptographic encapsulation of the right nature at the right levels. Homomorphic, Functional, and Attribute-based Encryption are examples of such encapsulation. Data transactions respecting trust boundaries and relations between interacting entities can be enabled by distributed cryptographic protocols such as Secure MPC and Blockchain. Many of the expensive cryptographic operations can be substituted by hardware primitives with circumscribed roots of trust, but one must be aware that there are inherent limitations and dangers to such approaches.

5.5.1 CLASSIFICATION

Table 2 provides a classification of cryptographic technologies in terms of their relation to the NBDRA, the features they support, and the data visibility they enforce.

Table 2: Classification of Cryptographic Technologies

Technology	Data Provider	Application Provider	Feature	Visibility
Homomorphic Encryption	Encrypts data	Stores encrypted data	Capability to perform computations	Only at Data Provider
Functional Encryption	Encrypts data	Stores encrypted data	Capability to perform computations	Result of allowed computations visible at Application Provider
Access Control Policy-Based Encryption	Encrypts data	Stores encrypted data	No capability to perform computations	Only for entities which have a secret key satisfying the access control policy
Secure Multi-Party Computation	Plaintext data	Stores plaintext data	Collaborative computation among multiple Application Providers	Application Providers do not learn others' inputs. They only learn the jointly computed function.
Blockchain	Plaintext or encrypted data	Decentralized	Immutable decentralized database	Transaction logging in a decentralized, untrusted environment
Hardware primitives for secure computations	Encrypts data	Stores encrypted data	Capability to perform computations. Verified execution.	Controllable visibility at Application Provider.

5.5.2 HOMOMORPHIC ENCRYPTION

> *Scenario: Data Provider has data to be kept confidential. Application Provider is requested to do computations on the data. Data Provider gets back results from Application Provider.*

Consider that a client wants to send all its sensitive data to a cloud—photos, medical records, financial records, and so on. She could send everything encrypted, but this wouldn't be of much use if she wanted the cloud to perform some computations on them, such as calculating the amount she spent on movies last month. With Fully Homomorphic Encryption (FHE), a cloud can perform any computation on the underlying plaintext, all while the results are encrypted. The cloud obtains no information about the plaintext or the results. [110]

Technically, for a cryptographic protocol for computation on encrypted data, the adversary should not be able to identify the corresponding plaintext data by looking at the ciphertext, even if given the choice of a correct and an incorrect plaintext. Note that this is a very stringent requirement because the adversary is able to compute the encryption of arbitrary functions of the encryption of the original data. In fact, a stronger threat model called chosen ciphertext security for regular encryption does not have a meaningful counterpart in this context - search to find such a model continues. [111]

In a breakthrough result [112] in 2009, Gentry constructed the first FHE scheme. Such a scheme allows one to compute the encryption of arbitrary functions of the underlying plaintext. Earlier results [113] constructed partially homomorphic encryption schemes. Gentry's original construction of a FHE scheme used ideal lattices over a polynomial ring. Although lattice constructions are not terribly inefficient, the computational overhead for FHE is still far from practical. Research is ongoing to find simpler constructions [114], [115], efficiency improvements [116], [117], and partially homomorphic schemes [118] that suffice for an interesting class of functions.

5.5.3 FUNCTIONAL ENCRYPTION

> *Scenario: Data Provider has data to be kept confidential. Application Provider or Data Consumer are allowed to do only a priori specified class of computations on the data and see the results.*

Consider a system to receive emails encrypted under the owner's public key. However, the owner does not want to receive spam mails. With plain public key encryption, there is no way to distinguish a legitimate email ciphertext from a spam ciphertext. However, with recent techniques, one can give a *token* to a filter, such that the filter can apply the token to the ciphertext only deducing whether it satisfies the filtering criteria or not. However, the filter does not get any clue about any other property of the encrypted message! [110]

Technically, for a cryptographic protocol for searching and filtering encrypted data, the adversary should not be able to learn anything about the encrypted data beyond whether the corresponding predicate was satisfied. Recent research has also succeeded in hiding the search predicate itself so that a malicious entity learns nothing meaningful about the plaintext or the filtering criteria.

Boneh and Waters [119] construct a public key system that supports comparison queries, subset queries, and arbitrary conjunction of such queries. In a recent paper [120], Cash et al. present the design, analysis, and implementation of the first sub-linear searchable symmetric encryption (SSE) protocol that supports conjunctive search and general Boolean queries on symmetrically-encrypted data and that scales to very large datasets and arbitrarily-structured data including free text search.

While with standard functional encryption, the objective is to compute a function over a single user's encrypted input, multi-input functional encryption (MIFE) is a relatively recent cryptographic primitive which allows restricted function evaluation over independently encrypted values from multiple users. It is possible to realize this primitive over the broadest class of permitted functions with a basic primitive called *indistinguishability obfuscation*, which to this date is prohibitively impractical. However, MIFE for important practical classes of functions such as vector inner products [121], equality and approximation testing and order evaluation are known using practically available tools like elliptic curves and lattices.

5.5.4 ACCESS CONTROL POLICY-BASED ENCRYPTION

> *Scenario: The Infrastructure Provider is part of an organization which employs many people in different roles. The requirement is to encrypt data so that only roles with the right combination of attributes can decrypt the data.*

Traditionally access control to data has been enforced by systems—Operating Systems, Virtual Machines—which restrict access to data, based on some access policy. The data is still in plaintext. There are at least two problems to the systems paradigm: (1) systems can be hacked, and (2) security of the same data in transit is a separate concern. [110]

The other approach is to protect the data itself in a cryptographic shell depending on the access policy. Decryption is only possible by entities allowed by the policy. One might make the argument that keys can also be hacked. However, this exposes a much smaller attack surface. Although covert side-channel attacks [122], [123] are possible to extract secret keys, these attacks are far more difficult to mount and require sanitized environments. Also encrypted data can be moved around, as well as kept at rest, making its handling uniform.

Technically, for a cryptographically-enforced access control method using encryption, the adversary should not be able to identify the corresponding plaintext data by looking at the ciphertext, even if given the choice of a correct and an incorrect plaintext. This should hold true even if parties excluded by the access control policy collude among each other and with the adversary.

Identity-based encryption (IBE) and attribute-based encryption (ABE) methods enforce access control using cryptography. In identity-based systems [124], plaintext can be encrypted for a given identity, and the expectation is that only an entity with that identity can decrypt the ciphertext. Any other entity will be unable to decipher the plaintext, even with collusion. Boneh and Franklin [125] came up with the first IBE using pairing-friendly elliptic curves. Since then, there have been numerous efficiency and security improvements [126]–[128].

ABE extends this concept to attribute-based access control. Sahai and Waters [129] presented the first ABE, in which a user's credentials is represented by a set of string called *attributes* and the access control predicate is represented by a formula over these attributes. Subsequent work [130] expanded the expressiveness of the predicates and proposed two complementary forms of ABE. In Key-Policy ABE, attributes are used to annotate the ciphertexts, and formulas over these attributes are ascribed to users' secret keys. In Ciphertext-Policy ABE, the attributes are used to describe the user's credentials and the formulas over these credentials are attached to the ciphertext by the encrypting party. The first work to explicitly address the problem of Ciphertext-Policy Attribute-Based Encryption was by Bethencourt, Sahai, and Waters [131], with subsequent improvement by Waters. [132]

As an example of Ciphertext-Policy ABE, consider a hospital with employees who have some possible combination of four attributes: *is a doctor*, *is a nurse*, *is an admin*, and *works in Intensive Care Unit (ICU)*. Take for instance a nurse who works in ICU—she will have the attributes *is a nurse* and *works in ICU*, but not the attribute *is a doctor*. The patient can encrypt his data under his access control policy of choice, such as, only a doctor OR a nurse who works in ICU can decrypt his data. Only employees who have the exact attributes necessary can decrypt the data. Even if two employees collude, who together have a permissible set of attributes, but not individually so, should not be able to decrypt the data. For example, an admin who works in the ICU and a nurse who doesn’t work in the ICU should not be able to decrypt data encrypted using the above access control policy.

5.5.5 SECURE MULTI-PARTY COMPUTATIONS

Consider a scenario where a government agency has a list of terrorism suspects and an airline has a list of passengers. For passenger privacy, the airline does not wish to give the list in the clear to the agency, while the agency too does not wish to disclose the name of the suspects. However, both the organizations are interested to know the name of the suspects who are going to travel using the airline. Communicating all the names in each list is a breach of privacy and clearly more information than required by either. On the other hand, knowing the intersection is beneficial to both the organizations.

Secure multi-party computations (MPC) are a class of distributed cryptographic protocols which address the general class of such problems. In an MPC between n entities, each entity P_i has a private input x_i and there is a joint function $f(x_1, \ldots, x_n)$ that everyone wants to know the value of. In the above scenario, the private inputs are the respective list of names and the joint function is the set intersection. The protocol proceeds through communication rounds between the entities, in which each message depends on the entity's own input, the result of some random coin flips and the transcript of all the previous messages. At the end of the protocol, the entities are expected to have enough information to compute f.

What makes such a protocol tricky to construct is the privacy guarantee it provides, which essentially says that each entity just learns the value of the function, and nothing else about the input of the other parties. Of course, given the output of the function, one can narrow down the possibilities for the inputs of the other parties—but, that is the *only* additional knowledge that it is allowed to gain.

Other examples include privacy-preserving collaborative analytics, voting protocols, medical research on private patient data, and so on. The foundations of MPC were given by Yao [133] , with a long line of work described in the survey by Saia and Mahdi [134]. This is a very active area of cryptography research and some practical implementations can be found in the multi-party computation library by Zamani [135].

5.5.6 BLOCKCHAIN

Bitcoin is a digital asset and a payment system invented by an unidentified programmer, or group of programmers, under the name of Satoshi Nakamoto [https://bitcoin.org/bitcoin.pdf]. While Bitcoin has become the most popular cryptocurrency, its core technological innovation, called the blockchain, has the potential to have a far greater impact.

The evidence of possession of a Bitcoin is given by a digital signature. While the digital signature can be efficiently verified by using a public key associated with the source entity, the signature can only be generated by using the secret key corresponding to the public key. Thus, the evidence of possession of a Bitcoin is just the secret key.

Digital signatures are well studied in the cryptographic literature. However, by itself this does not provide a fundamental characteristic of money—one should not be able to spend more than one has. A trusted and centralized database recording and verifying all transactions, such as a bank, is able to provide this service. However, in a distributed network, where many participating entities may be untrusted, even malicious, this is a challenging problem.

This is where blockchain comes in. Blockchain is essentially a record of all transactions ever maintained in a decentralized network in the form of a linked list of blocks. New blocks get added to the blockchain by entities called miners. To add a new block, a miner has to verify the current blockchain for consistency and then solve a hard cryptographic challenge, involving both the current state of the blockchain and the block to be added, and publish the result. When enough blocks are added ahead of a given block collectively, it becomes extremely hard to unravel it and start a different fork. As a result, once a transaction is deep enough in the chain, it's virtually impossible to remove. At a high level, the trust assumption is that the computing power of malicious entities is collectively less than that of the honest participants. The miners are incentivized to add new blocks honestly by getting rewarded with bitcoins.

The blockchain provides an abstraction for public ledgers with eventual immutability. Thus, beyond cryptocurrency, it can also support decentralized record keeping which can be verified and accessed widely. Examples of such applications can be asset and ownership management, transaction logging for audit and transparency, bidding for auctions, and contract enforcement.

While the verification mechanism for the Bitcoin blockchain is tailored specifically for Bitcoin transactions, it can in general be any algorithm such as a complex policy predicate. Recently a number of such frameworks called Smart Contracts, such as Ethereum, have recently come to the fore. The Linux Foundation has instituted a public working group called Hyperledger which is building a blockchain core on which smart contracts, called chain codes, can be deployed.

As specialized blockchain platforms emerge, guidance on blockchain uses and its possible applications in Big Data (and as Big Data) are needed. The WG is monitoring standards work under way in IEEE P2418 (Standard for the Framework of Blockchain use in IoT).

Another potential Big Data blockchain influence could come from the IEEE "Digital Inclusion through Trust and Agency" initiative (http://standards.ieee.org/develop/indconn/digital_inclusion/), whose possible initiative outcomes see distributed ledger (blockchain-like) solutions as facilitating broad social aims:

- Have agency over our data and cyber-identity;
- Provide the capacity to identify ourselves online in a way that protects our privacy, our right to be forgotten, and our off-line ability to have multiple personas;
- Give a voice to the underserved and vulnerable with the creation of standards that are inclusive of their needs;
- Encourage distributed ledger technology (e.g., Blockchain) standards that facilitate financial inclusion and other decentralized data sharing capabilities; and
- Develop a collaborative approach to technology and policy design regarding digital inclusion, trust, personal data, agency, security, and privacy for all demographics.

5.5.7 HARDWARE SUPPORT FOR SECURE COMPUTATIONS

While sophisticated cryptographic technologies like homomorphic and functional encryption work directly on encrypted data without decrypting it, currently practical implementations remain out of reach for most applications. Secure hardware primitives, such as TPM (Trusted Platform Module) and SGX (Software Guard Extensions), provide a middle ground where the central processing unit (CPU) and a dedicated portion of the hardware contain private keys and process data after decrypting the ciphertexts communicated to these components.

The premise is that all communications within a Trusted Computing Base (TCB) is considered sensitive and is carried out using an isolated and protected segment of memory. Communications to and from the TCB with external code and memory spaces are always encrypted. This segregation of a trusted zone and the untrusted environment can be carefully engineered and leveraged to provide higher-level security guarantees.

Verifiable Confidential Cloud Computing (VC3) [136] is a recent work which is aimed at trustworthy data analytics on Hadoop using the SGX primitive. The work addresses the following two objectives in their implemented framework:

1. Confidentiality and integrity for both code and data (i.e., the guarantee that they are not changed by attackers and that they remain secret); and
2. Verifiability of execution of the code over the data (i.e., the guarantee that their distributed computation globally ran to completion and was not tampered with).

VC3's threat model includes malicious adversaries that may control the whole cloud provider's software and hardware infrastructure, except for the SGX-enabled processors. However, DoS attacks, side channels, and traffic analyses are out of scope.

Advantages:

- Secure code runs competitively fast with respect to native execution of the same code.
- The only entity trusted is the CPU itself. Not even the operating system is trusted.

Disadvantages:

- Secure code execution is susceptible to side-channel leakage like timing, electromagnetic and power analysis attacks.
- Once secret keys embedded within the CPU are leaked, the hardware is rendered ineffective for further secure execution. If the leakage is detected, there are revocation mechanisms to invalidate

the public keys for the victim. However, a compromised CPU cannot be re-provisioned with a fresh key.

5.5.8 CRYPTOGRAPHIC KEY ROTATION

To limit leakage of sensitive data, cryptographic keys should be refreshed periodically. The period depends on the security level offered by the scheme (technically, the *security parameter*), level of protection given to storing the key, sensitivity of the data being operated on by the key, and the frequency of usage of the key.

The PCI-DSS (Payment Card Industry Data Security Standard, https://www.pcisecuritystandards.org) standard lists key rotation as a requirement. To quote, it requires “Cryptographic key changes for keys that have reached the end of their cryptoperiod (for example, after a defined period of time has passed and/or after a certain amount of cipher-text has been produced by a given key), as defined by the associated application vendor or key owner, and based on industry best practices and guidelines (for example, NIST Special Publication 800-57).” [137]

NIST Special Publication 800-57 [138] has a very detailed set of recommendations regarding key management in general, with a comprehensive treatment of key rotation. The recommendations are intended for a spectrum of roles in an IT environment and apply to a Big Data system orchestrator when making key management decisions about cryptographic operations to secure the following interfaces and storage:

- Communication interface between Data Consumers and Application Provider;
- Internal storage of sensitive data in the Framework Provider;
- Communication interface between Application Provider and Framework Provider; and
- Communication interface between Application Provider and Data Consumer.

The recommendations span description of cryptographic algorithms for specific goals, different types of keys that are needed, states that the keys cycle through, how long the keys need to be retained, and guidance for audit and accountability.

5.5.9 FEDERAL STANDARD FIPS140-2 ON CRYPTOGRAPHIC SYSTEMS

NIST publication FIPS140-2 [139] describes security requirements for cryptographic modules intended to handle sensitive data, in four increasing levels of stringency. The levels are intended to cater to the degree of data sensitivity required by the applications utilizing a given module. The security levels presented in FIPS 140-2 are as follows:

Security Level 1 is the lowest level which “allows the software and firmware components of a cryptographic module to be executed on a general-purpose computing system using an unevaluated operating system. Such implementations may be appropriate for some low-level security applications when other controls, such as physical security, network security, and administrative procedures are limited or nonexistent.” ([139] p.1)

“Security Level 2 enhances the physical security mechanisms of a Security Level 1 cryptographic module by adding the requirement for tamper-evidence, which includes the use of tamper-evident coatings or seals or for pick-resistant locks on removable covers or doors of the module. Tamper-evident coatings or seals are placed on a cryptographic module so that the coating or seal must be broken to attain physical access to the plaintext cryptographic keys and critical security parameters (CSPs) within the module. Tamper-evident seals or pick-resistant locks are placed on covers or doors to protect against unauthorized physical access. Security Level 2 requires, at a minimum, role-based authentication in which a cryptographic module authenticates the authorization of an operator to assume a specific role and perform a corresponding set of services.” ([139] p. 2)

Security Level 3: "In addition to the tamper-evident physical security mechanisms required at Security Level 2, Security Level 3 attempts to prevent the intruder from gaining access to CSPs [critical security parameters] held within the cryptographic module. Physical security mechanisms required at Security Level 3 are intended to have a high probability of detecting and responding to attempts at physical access, use or modification of the cryptographic module. The physical security mechanisms may include the use of strong enclosures and tamper detection/response circuitry that zeroizes all plaintext CSPs [critical security parameters] when the removable covers/doors of the cryptographic module are opened. Security Level 3 requires identity-based authentication mechanisms, enhancing the security provided by the role-based authentication mechanisms specified for Security Level 2. A cryptographic module authenticates the identity of an operator and verifies that the identified operator is authorized to assume a specific role and perform a corresponding set of services." ([139] p. 2)

"Security Level 4 provides the highest level of security defined in this standard. At this security level, the physical security mechanisms provide a complete envelope of protection around the cryptographic module with the intent of detecting and responding to all unauthorized attempts at physical access. Penetration of the cryptographic module enclosure from any direction has a very high probability of being detected, resulting in the immediate zeroization of all plaintext CSPs [critical security parameters]. Security Level 4 cryptographic modules are useful for operation in physically unprotected environments. Security Level 4 also protects a cryptographic module against a security compromise due to environmental conditions or fluctuations outside of the module's normal operating ranges for voltage and temperature. Intentional excursions beyond the normal operating ranges may be used by an attacker to thwart a cryptographic module's defenses. A cryptographic module is required to either include special environmental protection features designed to detect fluctuations and zeroize CSPs [critical security parameters], or to undergo rigorous environmental failure testing to provide a reasonable assurance that the module will not be affected by fluctuations outside of the normal operating range in a manner that can compromise the security of the module." ([139] p. 3)

These Security Levels provide a spectrum of local assurance of data protection. A consumer of these systems must remain aware that even Security Level 4 is not sufficient to provide security and privacy of sensitive data, unless the complete architecture that handles the data in consideration is analyzed with precise security and privacy guarantees that are intended.

5.6 RISK MANAGEMENT

To manage risk, NIST 800-39 recommends organizing risk across "three tiers of organization, mission/business processes, and information systems." [140] To some extent, this risk framework assumes an organizational monoculture that may not be present for Big Data. Managing risk across organizations may prove to be the norm under certain CPS/ IoT scenarios. As previously mentioned, future work must reassess existing risk frameworks elsewhere in light of Big Data considerations.

5.6.1 PII AS REQUIRING TOXIC SUBSTANCE HANDLING

Treating certain data elements as more toxic than others is necessary to highlight risks for developers, operators, auditors, and forensics. Section 2.4.7.2 discusses toxic data elements. For instance, information associating a patient with a highly contagious disease is important from a public safety perspective, but simultaneously creates privacy risks. Protecting both demands that tagging, traceability, and detailed data communications become more widely practiced in Big Data scenarios.

5.6.2 CONSENT WITHDRAWAL SCENARIOS

After a divorce, some previously provided consent must be withdrawn. In a few scenarios, this could be matter of life and death for an ex-spouse or a child, yet systematic methods for consent withdrawal are often ignored. Consent traceability through one of several means is seen as a Big Data priority for some scenarios.

5.6.3 TRANSPARENCY PORTAL SCENARIOS

How best to create data and algorithmic transparency is an emerging area of specialization in HCI. Several projects [83], [141], [142] are illustrative of attempts in this area, and there is even a recent formulation for an "organizational transparency model" [143]. Big Data systems are more likely to spur transparency model investments for several reasons including the following:

- The element of surprise may occur when citizens realize where and how their data is being used in scenarios seemingly far afield from their original intent. Recently, increased use of automated image identification created new concerns.
- Large scale breaches have occurred.
- Increased reliance on automated systems is forecast for public IoT applications, such as outdoor parking management, environmental monitoring, precision irrigation and monitoring, traffic management, smart metering, and many other areas [144]. This reliance will expose more people to Big Data-driven solutions, as well as to the security and privacy limitations of those systems. For some, engagement will become essential to protect basic services, such as access to healthcare or convenient air travel.
- As federated systems become more common—especially between small- and mid-size enterprises, participants will demand greater process transparency as well as access to data. Transparency may prove essential for collaborative decision making. As noted by Grogan et al., "Design methods for federated systems must consider local incentives and interactive effects among independent decision-makers," [145]. Access to shared Big Data pools is likely to be needed to fully leverage proprietary systems in-house.
- Cross-organizational Risk Management is well understood in construction circles as best governed by "target value design principles" and characterized by "shared risk and reward" [146]. As analogous concepts coalesce in Big Data systems, transparency of algorithms, data, and processes will become as important for participating enterprises as for the sources of data (e.g., consumers, devices, other systems).

5.6.4 BIG DATA FORENSICS AND OPERATIONAL AAR

After Action Review (AAR) is an essential component to effective security in the Big Data era. AAR demands huge volumes of data to support high-fidelity replay and log analytics. Yet most Big Data systems have haphazard or nonexistent support for audit, unless regulatory bodies demand more.

Support for forensics in part derives from the need to build integrated test frameworks for continuous delivery (at least for agile projects). However, forensics scenarios often encompass broad swaths of scenarios, rather than specific test exercises. Accomplishing this in a systematic way is still beyond the reach of Big Data architects. This in turn weakens attempts to protect and anticipate risks to security and privacy.

For many organizations, the starting point may be a reconsideration of logs and dependency models. Is the data needed for AAR being captured? Can scenarios be fully replayed? ModSim may be essential in more complex settings.

5.7 BIG DATA SECURITY MODELING AND SIMULATION (MODSIM)

Penetration testing is accepted as a best practice for security professionals. However, penetration testing cannot detect numerous security problems which arise. As systems become more complex and multi-organizational, unitary penetration is simply not feasible. Instead, a combination of live test, desktop walkthroughs, and simulation are likely to be needed.

The domain, utility, and application models recommended in the NBD-SPSL are helpful preparatory efforts in support of ModSim. The NBD-SPSL includes multiple features which exploit ModSim.

More than a decade ago, Nicol called for increased "emulation, in which real and virtual worlds are combined to study the interaction between malware and systems" [147]. Such methods question the usual assumptions about attack surfaces; red teams typically focus on perimeter attacks. White hat efforts do not have these limitations, but lack the necessary tools to test what-if scenarios internally. ModSim, in addition to code walkthroughs and other methods, allows for security threats to complex systems to be more systematically studied.

In studies focused on specific areas such as equipment maintenance, recent work has shown that Big Data systems call for different ModSim approaches [148]. Future security and privacy Big Data scenarios are likely to include a complex mix of people, legacy software, smartphones, and multi-robotic systems. [149] Dependency models that have been used for critical infrastructure modeling and analysis [150] are equally relevant for planning the *Ops* component of DevOps within the continuous delivery paradigm that is common in Big Data systems.

Machine learning and simulation are increasingly seen as an essential element in situation awareness, leading some analysts to declare these two elements as a key enabler in the win of AlphaGo over a human Go champion [151].

5.8 SECURITY AND PRIVACY MANAGEMENT PHASES

Earlier versions of this document did not clarify design-time, in-situ, and forensic (i.e., after-the-fact) considerations. This version explicitly addresses three phases for managing security and privacy in Big Data. Explicit awareness of these phases is seen as critical for security and privacy models to operate with full situation awareness.

1. ***Build Phase***: The security and privacy Build Phase occurs when a system is being planned, or while under development (in the agile sense). In a straightforward case, the Build Phase takes place in a *greenfield* environment. However, significant Big Data systems will be designed as upgrades to legacy systems. The Build Phase typically incorporates heaviest requirements analysis, relies the most upon application domain-specific expertise, and is the phase during which most architectural decisions are made. [152]
    a. Note: This phase is roughly analogous to NIST SP 800-53 planning controls.
    b. Build phases that incorporate explicit models include the business model canvas. As Scott Shaw argued, "If architecture is the thing you want to get right from the start of your project, you should be modelling the business domain as the sequence of events that occur" [153].
    c. At the build phase, delegated access management approaches should be designed in, using, for example, two-way TLS, OAuth, OpenID, JavaScript Object Notation (JSON) web tokens, hash message authentication code (HMAC) signing, NTLM, or other approaches. Architects must consider compatibility with the Big Data stack of choice.
    d. The design pattern recommended for authorization is stateless, not using sessions or cookies.

2. ***In-Situ Phase:*** This phase reflects a fully deployed, operational system. An in-situ security scenario shares elements with operational intelligence and controls. In a small organization, operations management can subsume security operations. Development may be ongoing, as in an agile environment where code has been released to production. Microservices present "huge challenges with respect to performance of [an] overall integrated system" [154]. Regardless of the predecessor tasks, once released into production, security challenges exist in an arena shared with operations—including issues such as performance monitoring and tuning, configuration management, and other well-understood concepts. This relationship is discussed in more detail in the *NBDIF: Volume 6, Reference Architecture* document in the Management Fabric section.
3. ***Decommissioned Phase***: In its simplest form, this phase reflects a system that is no longer operational. For example, data from a (probably) decommissioned application from a bankrupt company was provided by the bankruptcy court to a third party. There is a more nuanced version of the decommissioned phase as well. *Significant* changes to an existing app could be seen as a decommissioning. Gartner's Structured Data Archiving and Application Requirement [155] contains additional discussion of decommissioning. This phase also includes design for forensics analytics.

In addition to prior work by Ruan et al. [156], the Cloud Security Alliance proposed a Cloud Forensics Capability Maturity Model. As that Model demonstrates, more mature organizations will address phase-specific aspects of Big Data systems, rather than merely focusing on design and post-deployment administration.

MODIFICATIONS FOR AGILE METHODOLOGIES

Agile methods may be particularly well-suited for Big Data projects, though little research has been focused solely on security and privacy aspects. Frankova et al. claim the following:

> *The close cooperation of managers, CIOs, the owners of the product, the development team can … help find the right data, cleanse [data], and they can help in the decision to adopt or reject a hypothesis. In these cases, the agile iterative approach is very important because with Big Data [it] is difficult to predetermine return on investment* [157] *(p. 581).*

Working under the assumption that agile and DevOps are mutually enabling, the IEEE P2675 workgroup is preparing a standard that will improve practices for the development of software for DevOps. The focus of that work is agile methods for building secure systems in DevOps. Integrating Big Data logging, monitoring, traceability, resource management, and safety engineering into DevOps is a challenge that the IEEE P2675 workgroup is seeking to address. Recommendations to be followed from IEEE P2675 development activities may impact the NBD-SPSL.

While its work is still under way, the following are several preliminary conclusions that can be drawn from P2675 deliberations for Big Data Systems Development Life Cycle (SDLC):

- Interlocking, multi-organizational dependency models will demand that Big Data systems scale configuration management upward.
- Continuous security can be built in using any SDLC methodology, but agile may decompose the process.
- Test engineering for Big Data requires additional attention due to the velocity of releases, the Big Data impact on operations and infrastructure, sprint frequency, and the complexity of systems being architected.
- Big Data systems are difficult to manage as well as to build, yet securing these systems requires flexible, powerful administrative capabilities that may not be initially seen as important because the impact of Big Data scale is difficult to assess.

6 DOMAIN-SPECIFIC SECURITY

The importance of domain-specific considerations was a key insight derived from the HL7 FHIR consent workflow use case. Implementers cannot assume that genomic data should be treated using the same practices as electric utility smart meters. Future work of the Security and Privacy Subgroup may study domain-specific security considerations including the following:

- Identify domain-specific workflow,
- Consider domain-specific roles, and
- Investigate domain-specific share policies, content, controls.

Organizations (even including sole proprietorships) must identify which facets of Big Data systems are sharable and to whom. For some organizations, the domain model is not significantly different from that of the profession or industry sector; these models are in some sense, *global* utility models, and nonproprietary. Other aspects of the domain model contain intellectual property, internal roles, execution strategy, branding, and tools deployed; these aspects are shared only selectively.

This can be simplified to public and private *views* [158]. Using this approach, views can evolve (co-evolve with code, or as code itself) over time. When it comes time to federate, a *public* view is available of a NBDRA component.

Consent has emerged as a key Big Data security and privacy element. Implementers may need to take into account consent traceability, withdrawal, and transferal scenarios. Aspects of consent include the following:

- Consent management with respect to domain-specific Big Data security and privacy;
- Consent management in healthcare across provider networks;
- Relation to smart contracts, blockchain, and the law;
- Smart building domain security;
- Domain-specific provenance;
    - Traceability; and
    - Domain-specific reasoning.

7 AUDIT AND CONFIGURATION MANAGEMENT

Auditing fabric topology, including configuration management (CM) changes (taxonomic issues with configuration change data versus audit data). In some Big Data systems, audit, logging, and configuration data—with full history—could become larger than the associated Big Data system itself.

Audit and CM across organizational entities is only lightly covered in other standards. Planning for cross-organizational data transport is a Big Data concern. Of particular concern are the following cross-organizational data transport scenarios:

- Private enterprise → government
- Government agency→ government agency
- Government (e.g., open data resource) → private enterprise
- Private enterprise → external private enterprise

7.1 PACKET-LEVEL TRACEABILITY / REPRODUCIBILITY

An early participant in NBD-PWG proposed that a central Big Data application would keep every Transmission Control Protocol/Internet Protocol (TCP/IP) or User Datagram Protocol (UDP) packet, every binary, or every byte of firmware associated with a system. This exhaustive snapshot of system behavior would represent a fully reproducible dataset that Big Data tools could use for analytics, or if needed, to create an entire execution scenario.

7.2 AUDIT

SIEM applications increasingly rely on extensive log data for analytics. Similarly, log data is essential for many aspects of forensic analysis. Log data itself is increasingly Big Data. In a 2015 presentation, one of the cloud service providers stated that its largest application at the time was its self-monitoring data used for management and billing support. [h]

In 2006, NIST provided a set of recommendations for managing computer logs in order to preserve their integrity [159]. Big Data presents additional challenges for logging and monitoring due to scale and variety. Current InfoSec tools are beginning to take this into account but they lack the capabilities of most Big Data stacks.

Incident response for Big Data has been discussed in literature. In 2006, NIST provided guidance on performing computer and network forensics in the *Guide to Integrating Forensic Techniques into Incident Response* [160]. Future work must explicate the term incident response with respect to Big Data Security Operations Center, and establish how such systems are coordinated with infrastructure management more broadly.

[h] Presentation at a 2015 NYC *Storm* Meetup.

7.3 MONITORING

While monitoring has a conventional place in the security specialist's toolbox, the associated tools may not be sized properly for Big Data systems. For example, in the cloud setting, the following is argued:

> *"Monitoring demonstrates several challenges including gathering metrics from a variety of layers (infrastructure, platform, application), the need for fast processing of this data to enable efficient elasticity and the proper management of this data in order to facilitate analysis of current and past data and future predictions. In this work, we classify monitoring as a big data problem and propose appropriate solutions in a layered, pluggable and extendable architecture for a monitoring component."* [161]

Big Data security and privacy support for audit and logging for monitoring and management is critical, but security operations must be able to scale along with associated Big Data applications. In addition, monitoring must be appropriate for both the utility, domain, and application models involved. This requires a close collaboration between application designers and security and privacy teams that is often not achieved.

8 STANDARDS, BEST PRACTICES AND GAPS

8.1 NIST CYBERSECURITY FRAMEWORK

During 2017, NIST published two drafts of proposed updates to the 2014 Cybersecurity Framework [162]. Since its introduction in 2014, the framework [80] has seen considerable de facto adoption and mention across a variety of industries. In addition to its appearance in the DHS Critical Infrastructure Cyber Community C³ Voluntary Program [163], the NIST Cybersecurity Framework appears in numerous job descriptions. Its appearance in cybersecurity hiring actions and its adaptation for other standards (e.g., SABSA's SENC project [Gonzalez, 2015]) further reflect the importance of the NIST Cybersecurity Framework.

8.2 CONFIGURATION MANAGEMENT FOR BIG DATA

8.2.1 EMERGENCE OF DEVSECOPS

The Point in Time, temporally qualified nature of Big Data configuration management creates numerous challenges for security operations. This has contributed to the development of a movement in industry called DevSecOps, which applies DevOps concepts to security operations (SecOps). Big Data is increasingly part of this, but DevSecOps may also be essential to keep InfoSec tools abreast of fast-moving, fast-changing Big Data.

For instance, one cloud provider "lets sys admins track the state of resources in their account via configuration items. These configuration items can be used in two different ways: They can produce a timeline of events using configuration item states to tell a story about the life cycle of a specific instance. And administrators can report and react to compliance problems using a rule engine called 'Config-Rules,' creating true DevSecOps" [165].

More sophisticated notions of configuration management, and federated CMDB's with semantic web and domain-specific model connections are on the horizon.

A recent *lessons learned* piece by Textor et al. argues for a standards-based ontology as essential to integrating technology with less technical counterparts in risk or cost management:

> *"We present a solution for the semantic information integration of different domain models in the context of automated IT management. For that, we formulate a core ontology based on the COBIT IT governance framework for integration on a conceptual level and discuss features of an extensible knowledge-based runtime system. We present a case study that integrates models from storage management, virtual machine management and a billing model"* [166]

In the meantime, smaller-scale tools are expected to struggle with the pace of change brought about both by Big Data and left shift. This will challenge SecOps. Few SecOps organizations are structured to leverage model-based approaches. Reliance on utility models, such as perimeter threat, has already proven to have diminished usefulness for Big Data applications, or in data centers hosting these apps.

DevSecOps will likely encompass notions that are already part of NIST SP 800-190, *Application Container Security Guide*.

8.2.2 DEPENDENCY MODELS

Dependency models that encompass software bills of resources are less widely used than some standards suggest. In manufacturing, a standard feature of a Bill of Material is the *Where Used* capability, which allows for instant identification of everywhere a part is used, along with its revision level at the time of assembly. Software project management and build / quality management resources such as Apache Maven and other tools attempt to provide similar capabilities, and build tools must provide this at release time. However, Big Data demands a longitudinal perspective on the *Where Used* aspect that preserves all the components of a build for security traceability.

The use of data traceability is even less widely implemented, and the *infrastructure as code, left shift* trend means that data traceability may follow a similar, gradualist path. There are statistical and methodological problems with using some data gathered for one purpose in another setting. Tracing data from its original source, a provenance challenge, is also needed to understand constraints on the contexts where Big Data can be used appropriately.

The format that the dependency model takes and how it is integrated into the development, operations, and forensics setting for Big Data security and privacy requires further research. In HL7, for example, models are exchanged using the Model Interchange Format. Predictive analytical models can be exchanged using the Predictive Model Markup Language (PMML). OMG offers XML Metadata Interchange (XMI) and XML Metadata Interchange Diagram Interchange XMI[DI] as document formats to exchange models and diagrams between applications.

The use of security models and a standardized language to express constraints and access are essential for Big Data scalability and interoperability between organizations.

8.3 BIG DATA SDLC STANDARDS AND GUIDELINES

Today's developers operate under SDLC frameworks including agile [167], waterfall [168], and spiral [169], as well as other models. A significant number of developers operate under less explicit frameworks organized around GitHub practices—and this practice dominates in components used in many a Big Data stack. A convenient method of integrating for instance with the Integrated Development Environment (IDE) tool is essential to foster reuse of libraries, assurance tools, and test environments, yet standards for this have yet to be adopted.

8.3.1 BIG DATA SECURITY IN DEVOPS

The concept of DevSecOps was introduced by Gartner as an emerging principle in DevOps in 2012, shortly before this NIST working group began its work. Progress has been slow. Gartner, in a 2016 report noted the following:

> *"… We estimate that fewer than 20% of enterprise security architects have engaged with their DevOps initiatives to actively and systematically incorporate information security into their DevOps initiatives; and fewer still have achieved the high degrees of security automation required to qualify as DevSecOps"* [170] .

A deeper understanding, with solid technical underpinnings, is needed to specify how DevSecOps teams ought to operate in a Big Data development setting. For example, how should the DevOps pattern described by Cockroft for a major Big Data streaming service be applied to Big Data more generally [171]? This document recognizes the increasing importance of DevOps. DevOps enables small teams to create Big Data systems with much reduced effort—and potentially, much reduced oversight for security and privacy. DevOps does not preclude quality software [172], but it can reduce the importance of traditional checks and balances afforded by others in a larger organization.

The notion of *Infrastructure as Code* is enabled by DevOps and other principally cloud computing technologies [173]. The concept needs additional Big Data treatment to help foster security and privacy best practices in DevOps.

The potential dilution, while not disappearance, of requirements phases and traceability in the agile development paradigm creates challenges for a security-aware SDLC. For instance, while a *technology-agnostic* process termed Secure Development Life Cycle (SDL-IT) was developed at Microsoft to improve its management of security and privacy processes [174], adoption is hardly widespread. Attempts such as Secure-SDLC (S-SDLC) and the Software Assurance Maturity Model (OpenSAMM, which became part of OWASP) are not integrated into IDE in ways that foster secure practices. For Big Data systems, developers rarely receive automated alerts as to practices which could create privacy risks, or which require additional, perhaps standards-based, attention to coding, administrative, and deployment practices.

8.3.1.1 Application Life Cycle Management

Both the application life cycle and the data life cycle must be managed, although they can be delinked in Big Data scenarios as data flows outside an organization. Nolle argues that "DevOps emerged for app developers to communicate deployment and redeployment rules into the operations processes driving application life cycle management." [175]

8.3.1.2 Security and Privacy Events in Application Release Management

Recent focus on release management has been identified as Application Release Management (ARM). Contributions are sought to help identify Big Data ARM practices, especially as they apply to DevOps and agile processes more generally.

8.3.1.3 Orchestration

Nolle insists that DevOps and orchestration are two different concepts in the cloud context, but that orchestration has a loftier aim: "In the long run, what separates DevOps and orchestration may not be their ALM-versus-cloud starting point, but that orchestration is actually a more general and future-proof approach" [175]. Noelle cites TOSCA [176] as leading this charge.

A Big Data adaptation of TOSCA-like concepts is needed that extends beyond cloud computing. *NBDIF: Volume 8, Reference Architecture Implementation* contains further discussion of this topic.

8.3.1.4 API-First

API-first is a concept that was advocated by several industry leaders. In part, it reflected the reality of web practice. Many startups developed business models around which services they would consume, and which they would provide—through Application Programming Interfaces (APIs). Thus, the business model referred to as *API-First* came into existence [177].

API-first also addresses scalability challenges in domains such as healthcare. In the OpenID HEART major use case, the project team writes that, "The architecture of prior provider-to-provider technologies have not been able to scale naturally to patient and consumer environments. This is where an API-first approach has an edge."

In the NBDRA, at the conceptual level, this specifies that application providers and consumers operate through defined APIs which can provide additional safety. A recent example of an API that implements domain-specific resources is the HL7 FHIR Health Relationship Trust Profile for FHIR OAuth 2.0 Scopes. Resources in the scope of this trust profile include patients, medication requests, medication dispensing, medication administration, and clinical observations. This is a design pattern for API-first—API's are designed to operate in tandem with domain-specific resources.

Further work is needed to identify which controls are most effective, but commercial services are already available which monitor API calls and can react to API threats in real time by throttling or closing services.

8.3.2 MODEL DRIVEN DEVELOPMENT

Big Data systems potentially entail multiple models from multiple disciplines implemented across diverse platforms, and often across different organizations. Previous attempts to share information across organizations have not fared well. Sharing of database schemas is a minimal starting point. Appendix A provides a number of citations for this topic.

METAMODEL PROCESSES IN SUPPORT OF BIG DATA SECURITY AND PRIVACY

ISO 33001 offers additional guidance on the use of models and information sharing. Project examples of working domain models include the following:

- OpenBIM, a domain model for construction and facilities management (as in smart buildings) (ISO 16739:2013) Refer to [178];
- The Facility Smart Grid Information Model developed by ASHRAE/NEMA 201;
- HVAC Engineering Standards for Smart Buildings; and
- Automotive engineering models (SPICE).

An approach taken by Atkinson et al. [179] and further developed by Burger offers methods which place domain models firmly inside the SDLC.

> *"This provides a simple metaphor for integrating different development paradigms and for leveraging domain specific languages in software engineering. Development environments that support OSM essentially raise the level of abstraction at which developers interact with their tools by hiding the idiosyncrasies of specific editors, storage choices and artifact organization policies. The overall benefit is to significantly simplify the use of advanced software engineering methods."* [179]

Model-based approaches also provide more elastic approaches to Big Data security and privacy than is available through traditional methods like Role-based Access Control (RBAC) or explicit role-permission assignments (EPA). The authors of one approach, called Contextual Integrity, claim that its:

> *"… norms focus on who personal information is about, how it is transmitted, and past and future actions by both the subject and the users of the information. Norms can be positive or negative depending on whether they refer to actions that are allowed or disallowed. Our model is expressive enough to capture naturally many notions of privacy found in legislation* [180]

Leveraging domain-specific concepts from healthcare, related research demonstrated that EHR privacy policy could be, "… formalized as a logic program [and] used to automatically generate a form of access control policy used in Attribute-Based Encryption [181].

Such recommendations must be carried further to promote security and privacy practices in development. Models such as these are not generally part of the Big Data system architect's apprenticeship.

8.3.3 OTHER STANDARDS THROUGH A BIG DATA LENS

8.3.3.1 ISO 21827:2008 and SSE-CMM

The International Systems Security Engineering Association (ISSEA) promoted a standard referred to as the Systems Security Engineering Capability Maturity Model (SSE-CMM). SSE-CMM was developed in collaboration with more than 40 partner organizations, and is codified in the ISO/IEC 21827:2008 standard. Its roots date to the mid-1990s; it predated Big Data.

8.3.3.2 ISO 27018: Protection of PII in Public Clouds Acting as PII Processors

ISO 27018 is a recent standard that addresses protection of PII for cloud computing. ISO 27018 is based on ISO 27002 and adapted to public cloud considerations. Because much of today's Big Data is cloud-based, this standard addresses concerns that many system owners with *toxic* PII face.

> *Consent: CSPs (Cloud Service Providers) must not use the personal data they receive for advertising and marketing unless expressly instructed to do so by the customer. Moreover, a customer must be able to use the service without submitting to such use of its private information.*
>
> *Control: Customers have explicit control of how their personal data is used.*
>
> *Transparency: CSPs must inform customers where their personal data resides and make clear commitments as to how that data is handled.*
>
> *Accountability: ISO/IEC 27018 asserts that any breach of information security should trigger a review by the service provider to determine if there was any loss, disclosure, or alteration of personal data.*
>
> *Communication: In case of a breach, CSPs should notify customers, and keep clear records of the incident and the response to it.*
>
> *Independent and yearly audit: A successful third-party audit (see e.g., AWS CertifyPoint) of a CSP's compliance documents the service's conformance with the standard, and can then be relied upon by the customer to support their own regulatory obligations. To remain compliant, a CSP must subject itself to yearly third-party reviews.* (Adapted from [182])

8.3.4 BIG DATA TEST ENGINEERING

Techniques such as the ETSI Test Description Language can be employed to exercise an application to test for secure performance under test. For instance, which external sites and URLs should a web application access?

Test engineering is important in software assurance because complex systems cannot be fully tested by developers, or even developer teams without automation assistance. In a recent report, a vice president of product marketing estimated that some 33 exabytes of data had been generated to date. In the same report, a powertrain simulation and tools research leader estimated that their company generates about 500GB of data daily [183].

A fraction of this data is directly relevant to security and privacy, but even at 1%, this represents a daunting challenge.

8.3.5 API-FIRST AND MICROSERVICES

The notion of microservices has evolved from service-oriented architecture (SOA) and object-oriented practices, but is relevant to Big Data because it represents a convergence of several trends. A recent NIST draft NIST SP 800-180 attempts to put forth a standard definition [184]. As explained in the draft, "Applications are decomposed into discrete components based on capabilities as opposed to services and placed into application containers with the resulting deployment paradigm called a Microservices Architecture. This Microservices Architecture, in turn, bears many similarities with SOAs in terms of their modular construction and hence formal definitions for these two terms are also needed in order to promote a common understanding among various stakeholders … " (Preface, p. v)

A full discussion of the approach is presented in greater detail elsewhere [185], but microservices offer applications designers, data center managers, and forensics specialists greater detail and thus control over relevant Big Data security and privacy system events.

At a somewhat higher level in the stack, some have suggested frameworks to support microservices visible to users, as well as lower-level developer-centric services. This was the notion proposed by Versteden et al. in a scheme that supports discovery of semantically interconnected single-page web applications [186].

8.3.6 APPLICATION SECURITY FOR BIG DATA

8.3.6.1 RBAC, ABAC, and Workflow

Initial work by NIST evolved to an ANSI / INCITS standard 369-2004 for RBAC [187]. According to a later report, the "Committee CS1.1 within the International Committee for Information Technology Standards (INCITS) has initiated a revision with the goal of extending its usefulness to more domains, particularly distributed applications" [188]. Kuhn et al. outline potential benefits of an alternative approach, Attribute-Based Access Control (ABAC), though no reference model had emerged. In the same paper, a combination of ABAC and RBAC is suggested.

In 2015, NIST published a description of ABAC in NIST SP 800-162 [10].

Beyond RBAC improvements, Big Data systems must incorporate workflow standards, if not formalisms, to transfer roles and policies along with data (or application / data bundles) between organizations. Previous work has studied ways to extend traditional RBAC to enterprise registries [189], or to include geospatial attributes [190].

Because XACML does not support RBAC directly, Ferrini and Bertino note that while XACML profiles extended the original XACML to include RBAC, "the current RBAC profile does not provide any support for many relevant constraints, such as static and dynamic separation of duty." [191] Ferrini and Bertino recommended expanding the XACML framework to include OWL. More nuanced access control decision processes can be supported by leveraging the reasoning potential of OWL.

> *"It is also important to take into account the semantics of role hierarchies with respect to the propagation of authorizations, both positive and negative, along the role inheritance hierarchies. Supporting such propagation and, at the same time, enforcing constraints requires some reasoning capabilities. Therefore, the main issue with respect to the XACML reference architecture and the engine is how to integrate such reasoning capabilities."* [191]*[p. 145]*

Integrating workflow into the RBAC framework has also been studied. Sun et al. argued that adding workflow to RBAC would better, "support the security, flexibility and expansibility" of RBAC [192]. Workflow-specific as well as time-limited access improves not only controls for audit and forensics, but can help to limit the impact of insider threat.

8.3.6.2 'Least Exposure' Big Data Practices

Just as legacy and software key fobs have rotating authorization keys, Big Data systems should enforce time windows during which data can be created or consumed.

The increased use of massive identity management servers offers economy of scale and improved efficiency and usability through single sign-on. When breached, these datasets are massive losses affecting millions of users. A best practice is obviously to control access to Identity Access Management (IAM) servers, but more importantly to utilize distributed datasets with temporally restricted access.

Big Data should cause system architects to reconsider the entire notion of *admin* and *superuser* in favor of more nuanced domain-specific models. Using those models, Big Data systems can be designed to minimize the size of a breach by segmenting identity, PII and other datasets and limited access to controlled time windows that are *leased.*

8.3.6.3 Logging

The following logging standards are applicable to Big Data security and privacy:

- NIST SP 800-92,
- NIST SP 800-137, and
- DevOps Logging.

Logging standards should be reviewed carefully because some recommendations in existing standards may not scale, or may create untenable risks due to Big Data variety. For Big Data logging to scale properly, domain, application and utility models must come into play. For instance, an array of a thousand IoT sensors sending thousands of samples per second may or may not need to be logged. Logging must often be correlated with other events, which is why complex event processing can be useful for IoT security [193]. Application developers typically have a clearer understanding of the HCI aspects of their logs, but other model considerations also apply. In most cases, IoT security and privacy requires explicit models for sensors and their interfaces [194].

IEEE P2675 is developing a standard that addresses the role of logging in DevOps agile projects. Big Data logs require additional metadata about themselves and the setting in which logs are collected, because the logs may persist far beyond the current infrastructure and could be used in settings outside the current enterprise.

Logs will also be needed to supply data for ModSim, which many think will be key to self-managed infrastructure in the *left shift* movement.

For an example of the scope of today’s thinking about logging, refer to The Art of Monitoring, which devotes more than 500 pages to the subject. Add Big Data and domain-specific models to the mix, and the complexity is no less prevalent [195].

8.3.6.4 Ethics and Privacy by Design

The following standards are related to ethics and privacy by design and could be applicable to Big Data systems:

- IEEE P7000,
- IEEE P7002,
- IEEE P7003,
- IEEE P7007,
- ISO 27500,
- ISO 9241,
- FAIR, and
- NIST IR 8062.

The IEEE initiative to address ethical consideration in systems design, paired with ISO 27500, will provide future guidance in this area important to public consumers of Big Data. As documents are released from the IEEE working groups, this work should be surveyed for the needs of Big Data builders, managers, and consumers.

In an overview of ISO 27500, Tom Stewart summarizes the standard’s goal as: “... ISO 27500 The Human-centered Organization. Aimed at corporate board members, the standard explains the values and beliefs that make an organization human-centered, the significant business and operational benefits that arise, and the policies they need to put in place to achieve this." [196]

Big Data is a part of this larger need to address organizational values and to trace how these are implemented in practice [196].

Some work in this area is motivated by international cooperation around FAIR [197]. Others are driven by regulation [198].

8.4 BIG DATA GOVERNANCE

Big Data Governance is characterized by cross-organizational governance, cross-border considerations, federation, marketplaces, and supply chain frameworks. What is different about Big Data systems in comparison to other systems is that reliance on manual processes is no longer possible. Governance as a separate function of oversight and audit may not always be feasible. Governance must be designed in, hence the need to understand Big Data governance requirements in depth.

Apache Atlas is in incubation as of this writing, but aims to address compliance and governance needs for Big Data applications using Hadoop.

8.5 EMERGING TECHNOLOGIES

8.5.1 NETWORK SECURITY FOR BIG DATA

Protecting virtual machines is the subject of guidelines, such as those in the *NIST Secure Virtual Network Configuration for Virtual Machine (VM) Protection* Special Publication [199]. Virtual machine security also figures in PCI guidelines [200]. Wider adoption may be possible in many data centers, but the technique is currently poorly integrated with developer and asset management capabilities. Refer to the work of IEEE P1915.1 for emerging standards work on secure network function virtualization.

Big data challenges are converging with the 5G wireless standard, which will add velocity and volume stresses on telecommunications infrastructure. Representative of current thinking in this area is work on self-organizing networks (SONs) at a recent systems modeling conference. These investigators proposed, "…. novel Proactive SON methodology based on the Big Data framework to enable the shift in the SON paradigm. In this article, we present a comprehensive Big Data-based SON framework involving innovative Machine Learning techniques which would cater to scalability and programmability of 5G networks with respect to availability, reliability, speed, capacity, security and latency." [201].

Architecture Standards for IoT, such as IEEE P2413, are also of importance for Big Data network security.

8.5.2 MACHINE LEARNING, AI, AND ANALYTICS FOR BIG DATA SECURITY AND PRIVACY

AI and Big Data analytics are critical topics in Big Data, and are the focus of work such as IEEE P7003 IEEE P7007, and ISO 27500. Possible use cases could include conclusions from Medicare End-Stage Renal Disease, Dialysis Facility Compare (ESRD DFC, http://data.medicare.gov/data/dialysis-facility-compare). Additional investigations into machine learning, AI, and analytics with respect to Big Data security and privacy are needed and could include details on the following:

- Risk / opportunity areas for enterprises,
- Risk / opportunity areas for consumers, and
- Risk / opportunities for government.

Appendix A: NIST Big Data Security and Privacy Safety Levels

Version 2 of *NBDIF: Volume 4, Security and Privacy* is principally informed by the introduction of the NIST Big Data Security and Privacy Safety Levels (NBD-SPSL). Using the NBD-SPSL, organizations can identify specific elements to which their systems conform. Readers are encouraged to study the NBD-SPSL, presented in this appendix, before launching into the body of this version of the document. Appendix A is designed to be a stand-alone, readily transferred artifact that can be used to share concepts that can improve Big Data security and privacy safety engineering.

Table A-1: Appendix A: NIST Big Data Security and Privacy Safety Levels

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
"Where-is" monitoring and discovery of human touch points			
System is self-aware of its human touchpoints and is capable of maintaining a persistent safety framework that can identify and monitor where human interactions occur that involve risk for the affected domain.	Traditional "role" artifacts, such as CRT screen or mobile phone UI specifications.	UML, SysML identification of touchpoints within a domain model.	System incorporates awareness of touch points. Automated alerts, escalation when risk profile changes.

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
API-Oriented and API-first Safety			
As was the case with the SOA movement, the definition of clear interfaces is a key element of Big Data systems. Some argue that the numerous cloud-centric applications that have been built in the last decade have increasingly relied on pub-sub design patterns. In particular, designers may consider API characteristics early in the design of Big Data systems. Once established, multiple APIs can enhance security and privacy.	API-first designs in the enterprise take into account safety levels as part of design, management and forensics. APIs are used not just for risk, but also management and creating an ecosystem around the API. Using checklists and other methods, ABAC and RBAC elements are incorporated into APIs. Usage is routinely onboarded into enterprise tools.	Level 2 adds: automated API testing, traceability to SnP design patterns in use within teams and across SnP utility models (e.g., SSO, database encryption, encryption in transit). Third-party and InfoSec tools provide alerts and monitor for scalability and resilience.	Add to Level 2: direct link to domain, app and utility models. Include awareness of dependencies on a potentially increasing pool of third-party APIs. (See Dependency Model).
Selected References L. Xavier, A. Hora, and M. T. Valente, "Why do we break APIs? first answers from developers," in 2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER), Feb. 2017, pp. 392-396. [Online]. Available: http://dx.doi.org/10.1109/SANER.2017.7884640 R. Malcolm, C. Morrison, T. Grandison, S. Thorpe, K. Christie, A. Wallace, D. Green, J. Jarrett, and A. Campbell, "Increasing the accessibility to big data systems via a common services API," in 2014 IEEE International Conference on Big Data (Big Data), Oct. 2014, pp. 883-892. [Online]. Available: http://dx.doi.org/10.1109/BigData.2014.7004319 V. Srivastava, M. D. Bond, K. S. McKinley, and V. Shmatikov, "A security policy oracle: Detecting security holes using multiple API implementations," in Proceedings of the 32Nd ACM SIGPLAN Conference on Programming Language Design and Implementation, ser. PLDI '11. New York, NY, USA: ACM, 2011, pp. 343-354. [Online]. Available: http://doi.acm.org/10.1145/1993498.1993539			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Application Model			
An application model is an abstract description of a system element or application, including its Big Data SnP components. Some version of an application model is a requirement for the BDSQ safety framework. App models can foster transparency and interoperability, essential for long-lived Big Data and potentially, Big Data systems.	Traditional waterfall or agile artifacts, milestone checks with Big Data support	Advanced application design and monitoring tuned to Big Data needs (streaming, IoT, organization bleed-through)	In addition to software-enabled APM, additional In-app workflow implemented as code with explicit model. Full audit and logging to domain model. Model artifacts are produced and consumed inside the Big Data system.
Selected References M. Borek, K. Stenzel, K. Katkalov, and W. Reif, "Abstracting security-critical applications for model checking in a model-driven approach," in 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), Sep. 2015, pp. 11-14. [Online]. Available: http://dx.doi.org/10.1109/ICSESS.2015.7338996			
Authority to collect data			
Long-lived or PII-intensive Big Data systems must capture and maintain transparency for data collection authority. This may be point in time.	XML or equivalent for authority, capture terms of service, legal authorities, versioning information within overall enterprise governance.	Use digital cert associated with collection. Written policies surrounding enterprise handling for PII, but tend to be limited to a single enterprise.	Same as Level 1, but with controls designed for transferability to third parties, especially in supply chain settings. Authority data is tracked using Big Data technologies, detail, audit, traceability.

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Big Data Security Fabric "Communicator"			
A central concern of public institutions and citizenry places Big Data systems in a special, if not unique category. As a consequence of this heightened concern, the safety framework includes a Big Data System Communicator. The System Communicator may include internal artifacts, but its principal audience is a potentially wide spectrum of stakeholders whose concerns it might allay, in part, through transparency and interactivity.	Big Data system implement a portal for users, developers and managers to access system artifacts, FAQs and other relevant information connected to risk, privacy, security and enterprise practices. Content and management is manual.	System Communicator is partially connected to the actual Big Data system SnP apparatus, including partial connectivity with the domain, app and utility models involved. The Communicator hosts resources such as consent management, traceable requirements, limitations, changes in terms of use, and historical tracking.	System Communicator fully integrated: domain model-aware, persists when data moves outside organizations, self-updating. Potentially agent-based or functionally similar to agent-based. Full awareness of data life cycle for PII / PCI components, relevant covenants and consent.
Selected References A. Garcia Frey, "Self-explanatory user interfaces by model-driven engineering," in Proceedings of the 2Nd ACM SIGCHI Symposium on Engineering Interactive Computing Systems, ser. EICS '10. New York, NY, USA: ACM, 2010, pp. 341-344. [Online]. Available: http://doi.acm.org/10.1145/1822018.1822076 J.Preece and Rombach, "A taxonomy for combining software engineering and human-computer interaction measurement approaches: towards a common framework," International Journal of Human-Computer Studies, vol. 41, no. 4, pp. 553-583, Oct. 1994. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S1071581984710731 C. R. Sugimoto, H. R. Ekbia, and M. Mattioli, Big Data and Individuals. MIT Press, 2016. [Online]. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7862699			
Big Data Forensics Playbooks			
Pre-Big Data forensics could fail operate properly at Big Data scale.	Manual playbooks identify both in-house and third-party (e.g., regulator) forensics. Playbooks encompass risk management, transparency, traceability, and whether monitoring is sufficient to support forensics.	Playbooks are directly linked to software releases, with functional capabilities added or removed from playbooks with each release. Playbooks are a well-defined mix of manual and automated processes, and are exercised with periodic forensic "red team" operations.	Add to Level 2: Playbooks are directly linked to domain, app and utility models. Playbooks self-configure based on changes to models. Playbooks are complemented by self-maintaining properties of test frameworks. Red teams operate with real or simulated data to fully exercise playbooks, and are provided with tooling and access to perform these functions.

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Business continuity (BC)			
Business Continuity in the event of Big Data System failure can result in a wide range of scenarios, but could include breaches, lowered privacy shields, or inability to perform customary authentication.	Written BC plan, but most processes are manual. Explicit references to domain and utility models with cross-reference to application models.	Partially automated BC plans which leverage domain, utility and application models.	Fully automated dependency model, transition to/from alternative processing platforms, and support for post-failure forensics. Test, verification, audit systems are pre-instrumented for BC configurations.
Selected References R. Thomas and P. McSharry, Big Data Revolution: What farmers, doctors and insurance agents teach us about discovering big data patterns. Somerset NJ: Wiley, Mar. 2015, Chapter 20. T. Miksa, R. Mayer, M. Unterberger, and A. Rauber, "Resilient web services for timeless business processes," in Proceedings of the 16th International Conference on Information Integration and Web based Applications & Services, ser. iiWAS '14. New York, NY, USA: ACM, 2014, pp. 243-252. [Online]. Available: http://doi.acm.org/10.1145/2684200.2684281			
Capacity management for Security Operations			
Big Data SnP support for audit and logging for monitoring and management is critical, but security operations must be able to scale along with associated Big Data applications.	Big Data SnP framework exists within current platforms as deployed, but with limited ability to sustain attacks across multiple Big Data sources, especially for streaming sources.	Partially scalable implementation of plans to strengthen Security Operations to respond to planned and unplanned surges in Big Data SnP monitoring, management, and mitigation of threats and protective measures.	Failover or other plans, fully tested, for interruptions or pollution of streamed data sources. Typically requires simulations tied to domain and utility models, tied to scalable and resilient infrastructure within and across the infrastructure set of composable services and suppliers.
Selected References M. M. Bersani, D. Bianculli, C. Ghezzi, S. Krstic, and P. S. Pietro, "Efficient Large-Scale trace checking using MapReduce," in 2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE), May 2016, pp. 888-898. [Online]. Available: http://dx.doi.org/10.1145/2884781.2884832 M. Andreolini, M. Colajanni, M. Pietri, and S. Tosi, "Adaptive, scalable and reliable monitoring of big data on clouds," Journal of Parallel and Distributed Computing, vol. 79, pp. 67-79, 2015, special Issue on Scalable Systems for Big Data Management and Analytics. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S074373151400149X			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Consent Interoperability, traceability			
Big Data Systems add a layer of complexity for consent management (think terms of service, for instance, across decades and multiple data custodians). The Big Data Safety Framework recommends a traceable consent management system that addresses both compliance and privacy safety.	Big Data framework for the application includes consent tracking where applicable, with written policies to manage, administer and support forensics.	Adds partial automation with domain models to consent, and supports consent transference and withdrawal through a mix of manual and automated methods.	Consent traceability fully integrated with domain model. "Smart contracts" represent one possible approach to traceability, but specific requirements are domain-specific, automatically resolved by consulting the domain model(s).
Selected References A. T. Gjerdrum, H. D. Johansen, and D. Johansen, "Implementing informed consent as Information-Flow policies for secure analytics on eHealth data: Principles and practices," in 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), Jun. 2016, pp. 107-112. [Online]. Available: http://dx.doi.org/10.1109/CHASE.2016.39 M.Benchoufi, R. Porcher, and P. Ravaud, "Blockchain protocols in clinical trials: Transparency and traceability of consent [version 1; referees: 1 approved, 1 not approved]," F1000Research, vol. 6, no. 66, 2017. [Online]. Available: http://dx.doi.org/10.12688/f1000research.10531.1 E. Luger, "Consent reconsidered; reframing consent for ubiquitous computing systems," in Proceedings of the 2012 ACMConference on Ubiquitous Computing, ser. UbiComp '12. New York, NY, USA: ACM, 2012, pp. 564-567. [Online]. Available: http://doi.acm.org/10.1145/2370216.2370310			
Continuous delivery of SnP components			
As Big Data and its support software shifts and evolves over time, the associated SnP components will also evolve. Continuous Delivery of SnP elements can enhance safety by exposing dynamic aspects of SnP that can rapidly evolve to meet new threats or opportunities to preserve secrecy.	Periodic Big Data dev team reviews, adoption of agile (see IEEE P2675) methods for delivery. No build server integration.	Periodic reviews plus library reuse, continuous delivery, automated test and CMDB with partial domain and utility model integration.	Fully deployed, transparent, continuously deployed SnP microservices on build, test, production servers using agile or spiral delivery and integration with domain and utility models.
Selected References R. Heinrich, A. van Hoorn, H. Knoche, F. Li, L. E. Lwakatare, C. Pahl, S. Schulte, and J. Wettinger, "Performance engineering for microservices: Research challenges and directions," in Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering Companion, ser. ICPE '17 Companion. New York, NY, USA: ACM, 2017, pp. 223-226. [Online]. Available: http://doi.acm.org/10.1145/3053600.3053653 T. Margaria and B. Steffen, "Continuous Model-Driven engineering," Computer, vol. 42, pp. 106-109, 2009. [Online]. Available: http://dx.doi.org/10.1109/MC.2009.315 M. Sicker. (2017, Apr.) why use a microservice architecture. MuSigma. Chicago IL. [Online]. Available: http://musigma.org/architecture/2017/04/20/microservices.html			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Dependency and federation model			
Dependency models for Big Data SnP must take into account variety, volume, and velocity as scalability and diversity stresses on integrity and governance. Sometimes Big Data systems will span organizations, thus requiring related federation standards, which are needed for SnP continuity at scale. A dependency model takes into account the desired safety level; some Big Data systems will be deployed with high risk out of necessity, in which case dependency models are critical.	Implements a dependency model that is largely manual but addresses the mandatory human and computer elements in place to protect this particular Big Data system and deliver the stated safety levels.	Automated dependency model that incorporates interoperating information security tools (e.g., SIEM) and addresses dependencies outside the enterprise, including suppliers of data (cross-industry advisories) and software updates. Limited connectivity to domain and app models.	All capabilities of Level 2, but include greater automation and live connections to domain, app and utility dependencies. Greenfield and maintenance software occurs with dependency constraints provided within IDEs.
Selected References Z. Xu, Z. Wu, Z. Li, K. Jee, J. Rhee, X. Xiao, F. Xu, H. Wang, and G. Jiang, "High fidelity data reduction for big data security dependency analyses," in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 504-516. [Online]. Available: http://doi.acm.org/10.1145/2976749.2978378			
DevOps Pipeline Safety Engineering			
Big Data systems are increasingly built using DevOps pipelines. The Big Data DevOps pipeline incorporates safety concerns.	DevOps teams are provided with indoctrination for enterprise-wide safety frameworks for SnP. Scrum masters and product owners recognize which products and services are typically associated with the safety concerns of the enterprise.	DevOps teams routinely incorporate safety elements in scrums and refer to the Big Data SnP Elements by name. Elements can be tested and releases can be failed by citing safety thresholds by element.	Add to Level 2: DevOps CD pipeline integrates safety constraints, violation detection, monitoring, transparency, operational resource simulation.
Selected References A. Froehlich, "Your big data strategy needs DevOps," Information Week, Feb. 2017. [Online]. Available: http://www.informationweek.com/big-data/your-big-data-strategy-needs-devops/a/d-id/1328184			
Disaster Planning and Information Sharing			
The focus for disaster planning writ in general tends to be returning to full availability. Big Data disaster planning must address the impact of both lost availability and the impact of massive breaches such as the OPM and Yahoo breaches.	Community-level collaboration, such as generator-sharing, carpooling contingencies, and other "manual" plans.	Explicit model for DR and information sharing across domains, especially geospatial. Automation is typically partial, with domain SnP only partially enumerated.	Fully tested environment for digital information sharing, e.g., XchangeCore, but fully integrated with SnP domain and utility models.

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Disaster Recovery (DR)			
Recovering from a Big Data system outage can require measures beyond those required for smaller systems, as demonstrated by a 2017 AWS outage. In addition, DR plans must include remediation of weakened or lost privacy, notification of affected parties, and mandated regulatory actions.	Written DR plan which encompasses human and computing infrastructure. Loosely connected to domain and utility models.	Same as Level 3 but only partially automated.	Complete integration of DR plan with automated connections to resilience apparatus, human and computing infrastructure. Domain and utility models are part of system creation.
Selected References Amazon_Web_Services, "Summary of the amazon s3 service disruption in the northern Virginia (US-EAST-1) region," Amazon Web Services Blog, Mar. 2017. [Online]. Available: https://aws.amazon.com/message/41926/			
Domain model interoperability			
Big Data tends to move across organizational, even national boundaries. Because of this, safety within a domain is strengthened when the domain models minimize idiosyncratic constructs.	Ability to produce SnP metrics, alerts and to consume external intelligence applicable to the domain. Some or all are manual.	Partial automation of domain-specific interoperability exists, e.g., SEC compliance, HIPAA compliance. Explicit policies mandating crosswalk to third party or industry standard domain models (e.g., EHR, FIBO).	Fully automated and standards-based interoperability at the highest level supported by the domain or a fully elaborated scenario, e.g., HL7 FHIR.
Selected References X. Q. Huang, K. D. Zhang, C. Chen, Y. J. Cao, and C. Q. Chen, "Study on the integration architecture of electric power big data based on four kinds of integration patterns," in 10th International Conference on Advances in Power System Control, Operation Management (APSCOM 2015), Nov. 2015, pp. 1-6. [Online]. Available: http://dx.doi.org/10.1049/ic.2015.0234			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Explicit, reusable design patterns for SnP process orchestration			
Big Data systems may employ automated orchestration practices. When used, orchestration is enhanced by SnP design patterns that script, test, and audit orchestration using Big Data infrastructure, often mirroring underlying domain structures.	Enterprise standards are in place to identify how SnP is to be orchestrated when containers or other methods are used to deploy computing resources. Processes are largely manual or checklist-oriented.	Orchestration processes incorporate SnP practices that are integrated with infrastructure management (service management) as well as IDEs. Test engineering verifies compliance post-deployment.	Same as Level 2, but with live references to domain, app and utility models.
Selected References Luo and M. B. Salem, "Orchestration of software-defined security services," in 2016 IEEE International Conference on Communications Workshops (ICC), May 2016, pp. 436-441. [Online]. Available: http://dx.doi.org/10.1109/ICCW.2016.7503826 B. Pariseau, "EBay to bottle its special sauce for kubernetes management," Search Target IT Operations, May 2017. [Online]. Available: http://searchitoperations.techtarget.com/news/450419112/EBayto-bottle-its-special-sauce-for-Kubernetes-management N. Rathod and A. Surve, "Test orchestration a framework for continuous integration and continuous deployment," in 2015 International Conference on Pervasive Computing (ICPC), Jan. 2015, pp. 1-5. [Online]. Available: http://dx.doi.org/10.1109/PERVASIVE.2015.7087120 (repeated above)			
Exposure-limiting risk operations			
While there will be exceptions, the Big Data safety framework eschews the aggregation of PII/PCI in single, massive repositories using Hadoop, SQL or any other technology. This is especially true for identity and authentication support systems.	Closely managed RBAC and ABAC policies used in tandem that limit the scope of access and the duration of access, taking into account levels of risk associated with usage patterns	Same as Level 3 but only partially automated.	Big Data framework for limited access tightly integrated with live, automated connections to domain, utility, application models. IDEs surface risk levels associated with specific application functions to developers and testers.
Selected References W. H. Winsborough, A. C. Squicciarini, and E. Bertino, "Information carrying identity proof trees," in Proceedings of the 2007 ACM Workshop on Privacy in Electronic Society, ser. WPES '07. New York, NY, USA: ACM, 2007, pp. 76-79. [Online]. Available: http://doi.acm.org/10.1145/1314333.1314348			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Fully leveraged network layer SnP, including SDN			
A property of Big Data systems is that they tend to be challenging to back up using the usual methods. Thus, their storage requirements tend to favor network layer isolation practices to enhance SnP. Applications vary, but the method is being studied for 5G networks and vehicular networks, for instance.	Using traditional data center governance, leverages network filtering and DMZ to restrict, monitor, scale, manage access. Limited if any use of SDN itself.	Partial use of SDN to limit access, especially for SnP data elements and when OpenStack is an option. Maturing collaboration between application and infrastructure teams to plan resilience and secure platforms for apps.	SDN microsegmentation fully integrated with SDLC, design, test, resilience, forensics. SDN is leveraged to isolate code and data and is used both by app teams and infrastructure specialists together rather than separately, relying on common domain, app and utility models.
Selected References S. Marek, "How does Micro-Segmentation help security? explanation," SDx Central, 2017. [Online]. Available: https://www.sdxcentral.com/sdn/network-virtualization/definitions/how-does-microsegmentation-help-security-explanation/ L. Cui, F. R. Yu, and Q. Yan, "When big data meets software-defined networking: SDN for big data and big data for SDN," IEEE Network, vol. 30, no. 1, pp. 58-65, Jan. 2016. [Online]. Available: http://dx.doi.org/10.1109/MNET.2016.7389832			
Information Assurance resilience engineering			
Engineering Big Data systems for resilience is required to provide the Assurance dimension of Big Data information safety. For instance, full redundancy may not be affordable or feasible for some systems, whereas other Big Data systems can leverage sharded cloud/premise storage.	Fallback plan(s) with written playbooks for Big Data breaches or loss of service. Plans are principally manual with checklists and not subject to automated test.	Same as Level 3 but only partially automated.	Automated playbooks fully integrated with domain and utility models. Some assurance claims can be tested using continuously deployed test frameworks on Big Data platforms. HCI includes a transparent fully enumerated mix of machine and human test points.

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Integration of domain- and utility SnP models			
Domain models are specific to subjects such as healthcare or education scenarios. Utility models address cross-domain practices such as storage management, passwords, containers, access tokens, keys, certificates, DRM, encryption at rest/in transit. Safety improves as these two types are integrated.	Models used for domain and/or cross-domain utilities (e.g., help desk, SAN representation) but are not cross-linked	Same as Level 3 but only partially automated.	Complete integration of the Big Data safety system with domain and utility models. Advanced systems utilize ontologies or other explicit model representations of security and privacy concepts through methods such as Domain Driven Development, Domain Specific Languages, or other techniques in support of domain-aware safety engineering. Integrated with test and management systems including simulation and DevOps continuous deployment processes for security and privacy frameworks.
Selected References D. Zage, K. Glass, and R. Colbaugh, "Improving supply chain security using big data," in 2013 IEEE International Conference on Intelligence and Security Informatics, Jun. 2013, pp. 254-259. [Online]. Available: http://dx.doi.org/10.1109/ISI.2013.6578830 L. Obrst, P. Chase, and R. Markeloff, "Developing an ontology of the cyber security domain," in Proceedings of the Seventh International Conference on Semantic Technologies for Intelligence, Defense, and Security, P. C. G. Laskey and K. B. Laskey, Eds. CEUR, Oct. 2012, pp. 49-56. [Online]. Available: http://ceur-ws.org/Vol-966/ S. Fenz, "Ontology-based generation of IT-security metrics," in Proceedings of the 2010 ACM Symposium on Applied Computing, ser. SAC '10. New York, NY, USA: ACM, 2010, pp. 1833-1839. [Online]. Available: http://dx.doi.org/10.1145/1774088.1774478			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Integration of IoT scenarios, models			
IoT scenarios vary greatly from smart city designs to wearable medical devices. IoT Big Data, poised to become one of the Biggest of Big Data, requires integration of sensor and processing models.	Using traditional governance frameworks, an IoT model for the system has been designed with separate models for sensors, transducers, relays, protocols, and other elements.	Loosely coupled IoT SnP models allowing for partial integration with domain-specific and utility models for the big data application.	IoT SnP model fully integrated with domain and utility models.
Selected References D. Geneiatakis, I. Kounelis, R. Neisse, I. Nai-Fovino, G. Steri, and G. Baldini, "Security and privacy issues for an IoT based smart home," in 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 2017, pp. 1292-1297. [Online]. Available: http://dx.doi.org/10.23919/MIPRO.2017.7973622 M. A. Underwood, Big Data Complex Event Processing for Internet of Things Provenance: Benefits for Audit, Forensics and Safety. Hoboken NJ: Wiley, Nov. 2016, ch. 8. [Online]. Available: http://www.wiley.com/WileyCDA/WileyTitle/productCd-1119193869,subjectCd-EE23.html M. Underwood, M. Gruninger, L. Obrst, K. Baclawski, M. Bennett, G. Berg-Cross, T. Hahmann, and R. D. Sriram, "Internet of things: Toward smart networked systems and societies." Applied Ontology, vol. 10, no. 3-4, pp. 355-365, 2015. [Online]. Available: http://dblp.uni-trier.de/db/journals/ao/ao10.html#UnderwoodGOBBBH15 C. Jouvray, S. Gerard, F. Terrier, S. Bouaziz, and R. Reynaud, "Smart sensor modeling with the UML for real-time embedded applications," in IEEE Intelligent Vehicles Symposium, 2004, Jun. 2004, pp. 919-924. [Online]. Available: http://dx.doi.org/10.1109/IVS.2004.1336508 N. Foukia, D. Billard, and E. Solana, "PISCES: A framework for privacy by design in IoT," in 2016 14th Annual Conference on Privacy, Security and Trust (PST), Dec. 2016, pp. 706-713. [Online]. Available: http://dx.doi.org/10.1109/PST.2016.7907022			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Integration of key management practices with domain models			
Tokenization and key management practices are frequently central to managing proper access to systems and data, especially across enterprises. The Big Data safety framework advises the use of workflow-specific, domain-specific, least-privilege distributed access patterns, using the temporally restricted ('leased") permissions with full audit and traceability.	Adoption of key management practices to manage federated entities with manual transparency and audit.	Key management is partially integrated with domain, app and utility models.	Fully integrated key management with domain, app and utility models. Testing is automated when continuous deployment is practiced using Big Data frameworks.
Selected References R. Alguliyev and F. Abdullayeva, "Development of risk factor management method for federation of clouds," in 2014 International Conference on Connected Vehicles and Expo (ICCVE), Nov. 2014, pp. 24-29. [Online]. Available: http://dx.doi.org/10.1109/ICCVE.2014.7297548 D. R. dos Santos, S. Ranise, L. Compagna, and S. E. Ponta, Assisting the Deployment of Security-Sensitive Workflows by Finding Execution Scenarios. Cham: Springer International Publishing, 2015, pp. 85-100. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-20810-7_6			
Integration of risk models with CMDB at scale			
By definition, Big Data systems at scale may persist longer and accrue complexity at a faster pace than other computation. Risk models can be integrated with domain and utility models to accommodate configuration changes, especially in federation, key management, resilience strategies.	Mature risk management, mature configuration management automated CMDB practices, but separately maintained from other models.	Deployed CMDB, with semi-automated connectivity / interoperability with domain and utility models	Fully integrated CMDB, risk, domain and utility models across IDE, management, administration, and forensics.
Selected References J. Whyte, A. Stasis, and C. Lindkvist, "Managing change in the delivery of complex projects: Configuration management, asset information and 'big data'," International Journal of Project Management, vol. 34, no. 2, pp. 339-351, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0263786315000393			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Model-based simulation to assess security and risk at Big Data scale			
Big Data safety systems should incorporate simulation capabilities so that SnP considerations with deployment—not excluding HCI—can be simulated.	ModSim is employed to identify issues with usability, scalability, manageability, and interoperability of an app's SnP capabilities.	ModSim is used for both infrastructure planning and management as part of DevOps. Simulations are used to forecast additional requirements for new applications, infrastructure changes, mergers and acquisitions, and staffing reductions.	Simulation processes fully integrated into Phase D and I, and referencing domain and utility models. Interoperability with third-party models for environmental, geospatial, biomedical (e.g., SNOMED) models is practiced.
Selected References S. Schmidt, R. Bye, J. Chinnow, K. Bsufka, A. Camtepe, and S. Albayrak, "Application-level simulation for network security," SIMULATION, vol. 86, no. 5-6, pp. 311-330, May 2010. [Online]. Available: http://dx.doi.org/10.1177/0037549709340730 D. D. Dudenhoeffer, M. R. Permann, and E. M. Sussman, "General methodology 3: a parallel simulation framework for infrastructure modeling and analysis," in WSC '02: Proceedings of the 34th conference on Winter simulation. Winter Simulation Conference, 2002, pp. 1971-1977.			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Model-based systems engineering (MBSE) development practices			
MBSE is an approach to software engineering which relies on abstract representations of code. Security and privacy concepts for Big Data are best integrated with models vs. add-on, sandbox and "perimeter defense" methods—though it does not exclude other software-building methods even within the same system.	Post hoc models of legacy applications, with views created by SMEs. Models are not directly interoperable or communicating.	Hybrid: some legacy, some greenfield microservices design patterns constructed using model-based systems engineering practices. Models are implemented with partial integration across domain, utility, application models.	Defensive, surveillance, other measures fully integrated into domain, utility and application models. Forensics, IDE, test frameworks, SnP fully interoperable and live.
Selected References M. Borek, K. Stenzel, K. Katkalov, and W. Reif, "Abstracting security-critical applications for model checking in a model-driven approach," in 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), Sep. 2015, pp. 11-14. [Online]. Available: http://dx.doi.org/10.1109/ICSESS.2015.7338996 Estefan, J. 2008. Survey of Candidate Model-Based Systems Engineering (MBSE) Methodologies, rev. B. Seattle, WA, USA: International Council on Systems Engineering (INCOSE). INCOSE-TD-2007-003-02. Accessed April 13, 2015 at http://www.omgsysml.org/MBSE_Methodology_Survey_RevB.pdf A. Ross, "Interactive Model-Centric systems engineering," in 6th Annual SERC Sponsor Research Review, Georgetown University. Washington DC: Systems Engineering Institute, Dec. 2014. D. C.Schmidt, "Guest editor's introduction: Model-Driven engineering," Computer, vol. 39, no. 2, pp. 25-31, Feb. 2006. [Online]. Available: http://dx.doi.org/10.1109/mc.2006.58 A. Endert, S. Szymczak, D. Gunning, and J. Gersh, "Modeling in big data environments," in Proceedings of the 2014 Workshop on Human Centered Big Data Research, ser. HCBDR '14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2609876.2609890 R. Perry, M. Bandara, C. Kutay, and F. Rabhi, "Visualising complex event hierarchies using relevant domain ontologies: Doctoral symposium," in Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems, ser. DEBS '17. New York, NY, USA: ACM, 2017, pp. 351-354. [Online]. Available: http://doi.acm.org/10.1145/3093742.3093901			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Models for Big Data provenance			
Whether for machine learning classifiers, data lineage, or other notions of provenance, Big Data systems may require representations that track data sources, transport. Some have proposed that this must encompass retaining the binaries and network traffic for entire collection events.	Provides explicit organizational guidance about the use of ML tools and training datasets.	Provenance is built into the SDLC process through reusable libraries and requirements engineering. Test frameworks check for provenance flow and integrity and exception detection is an objective of Big Data monitoring. Monitoring in this setting applies primarily to SnP elements.	Employ tools such as PROV-O to manage and trace provenance. For IoT, integration with the W3C PROV family of provenance metadata. Directly incorporates domain, app, and utility models where applicable, and leverages results from industry- or domain-wide simulations.
Selected References K. Taylor, R. Woodcock, S. Cuddy, P. Thew, and D. Lemon, A Provenance Maturity Model. Cham: Springer International Publishing, 2015, pp. 1-18. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-15994-2_1 P. Missier, K. Belhajjame, and J. Cheney, "The W3C PROV family of specifications for modelling provenance metadata," in Proceedings of the 16th International Conference on Extending Database Technology, ser. EDBT '13. New York, NY, USA: ACM, 2013, pp. 773-776. [Online]. Available: http://dx.doi.org/10.1145/2452376.2452478 L. Moreau, J. Freire, J. Futrelle, R. Mcgrath, J. Myers, and P. Paulson, "The open provenance model: An overview," 2008, pp. 323-326. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-89965-5_31			
ModSim for security operations scalability			
Use of Modeling and Simulation (ModSim) for assessing the impact of scaling SnP Big Data systems. For DevOps, this has a more specialized meaning.	Occasional use of ModSim to support Big Data security operations.	Plans are deployed which routinely employ ModSim to estimate and forecast security operations as new applications, data centers, and technologies are onboarded.	Same as Level 2, but with live connections to domain, application, and utility models. Application onboarding includes planning for ModSim support infrastructure including HR.
Selected References S. Jain, C. W. Hutchings, Y. T. Lee, and C. R. McLean, "A knowledge sharing framework for homeland security modeling and simulation," in Proceedings of the 2010 Winter Simulation Conference, Dec. 2010, pp. 3460-3471. [Online]. Available: http://dx.doi.org/10.1109/WSC.2010.5679035 J. Kolodziej, H. González-Vélez, and H. D. Karatza, "High-performance modelling and simulation for big data applications," Simulation Modelling Practice and Theory, vol. 76, pp. 1-2, 2017, high-Performance Modelling and Simulation for Big Data Applications. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1569190X17300722			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
PII identification practices			
Transparent, adaptable practices for Big Data identification of PII should address safety by allowing for remediation (misidentification), continuous improvement of identification process, and Big Data records retention.	Provides a user portal for submitting claims of error or misinformation with manual methods for remediation.	Systematic approach to PII error with automated and manual methods to detect error or spillage of misinformation outside system boundaries.	In addition to Level 2, adds self-checking and self-correcting methods with audit. Remediation is supported with forwarding to downstream data consumers.
Selected References R. Herschel and V. M. Miori, "Ethics & big data," Technology in Society, vol. 49, pp. 31-36, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0160791X16301373			
PII vulnerability management			
PII (or "privacy") vulnerability management adopts principles from software vulnerability detection and remediation, plus other techniques, and applies them to protecting PII.	CFO designated with internal privacy controls and guidelines for federated entities. No separate Vulnerability Management for PII resource.	Enterprise has implemented a PII/PCI vulnerability management resource on a par with its traditional VM SecOps and software assurance capabilities.	Using Big Data or other tools to test for PII leakage, including external nonfederated entities. Same as Level 2, but integrated with domain, app, and utility models to accelerate risk detection.
Selected References N. J. King and J. Forder, "Data analytics and consumer profiling: Finding appropriate privacy principles for discovered data," Computer Law & Security Review, vol. 32, no. 5, pp. 696-714, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0267364916300802 B. Austin, "When to use PII discovery in the audit process," Solarwinds MSP, Apr. 2014. [Online]. Available: https://www.solarwindsmsp.com/blog/when-to-use-pii-discovery-in-the-audit-process			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
PII/PCI isolation			
For some Big Data systems, safety engineering requires separation of PII/PCI from other data elements. Separation can be achieved through a variety of technologies, including SDN.	Separate "files" or tables for designated PII data and code.	Separation is integrated with test and assurance frameworks with regular "penetration" testing using Big Data variety techniques. Partial integration with domain models.	Workflow model controls time windows, total exposure to PII using a Geiger counter-style avoidance model. Self-monitoring according to embedded models. Automated testing using domain-specific test and assurance frameworks in continuous deployment. Some advanced safety frameworks may support user-configured privacy protections and notifications.
Selected References M. Li, W. Zang, K. Bai, M. Yu, and P. Liu, "MyCloud: Supporting user-configured privacy protection in cloud computing," in Proceedings of the 29th Annual Computer Security Applications Conference, ser. ACSAC '13. New York, NY, USA: ACM, 2013, pp. 59-68. [Online]. Available: http://doi.acm.org/10.1145/2523649.2523680			
PII/PCI Toxicity orientation and traceability			
The Big Data SnP safety framework positions PII/PCI data to be handled with information systems analog to the chemical industry's Material Safety Data Sheets. Traceability is required, just as chain of custody is traced for certain class of prescription medications.	Written policies and procedures are in place, which treat PII/PCI disclosure as safety risks. Automation is minimal.	PII/PCI toxicity concept is fully integrated into the security culture, but crosswalk to domain, app, and utility models is not automated. MSDS for data elements are integrated into enterprise business glossaries, data catalogs.	Big Data analytics used to "penetration-test" aggregated data with automated alerts. Automated crosswalk of toxic data elements in domain, app, and utility models with MSDS-like processes fully automated.
Selected References M. Benchoufi, R. Porcher, and P. Ravaud, "Blockchain protocols in clinical trials: Transparency and traceability of consent [version 1; referees: 1 approved, 1 not approved]," F1000Research, vol. 6, no. 66, 2017. [Online]. Available: http://dx.doi.org/10.12688/f1000research.10531.1			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Policies for data or performance uncertainty, error and quality management for Big Data			
The ability to ingest massive amounts of data ensures that the absolute number of erroneous or faulty data will also be ingested. The safety framework requires inclusion of policies to address management of uncertainty and error.	Rough measures of uncertainty / error communicated to providers and consumers. Can be integrated with quality management systems. Largely manual, using checklists.	Explicit, software-based alerts for error and data quality assurance. Some level of self-healing processes is in place that operates in tandem with data quality metrics and stewardship.	Automated alerts are raised when tools (e.g., machine learning) attempt to make inferences that violate statistical or regulatory guidelines and are alerted according to protocols and importance determined by domain, app, and utility models delivered in automated format.
Selected References J. Bendler, S. Wagner, T. Brandt, and D. Neumann, "Taming uncertainty in big data," Business & Information Systems Engineering, vol. 6, no. 5, pp. 279-288, Oct. 2014. [Online]. Available: http://dx.doi.org/10.1007/s12599-014-0342-4 J. R. Busemeyer, "Decision making under uncertainty: a comparison of simple scalability, fixed-sample, and sequential-sampling models." J Exp Psychol Learn Mem Cogn, vol. 11, no. 3, pp. 538-564, Jul. 1985. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/3160815			
Safety Orientation			
As with the 1988 Challenger accident, breaches of Big Data systems (especially cloud-based, but IoT systems are likely to suffer a similar fate) should result in investments in a safety engineering culture. The same must be true for Big Data system architects, managers, and users.	Systematic use of safety terminology, personnel orientation, third-party safety standards and remediation planning. Capture failure events related to Big Data analytics, processes. Most processes are manual, using checklists and orientation.	Failure analytics applied to SDLC: e.g., Failure Mode and Effects Analysis (FMEA), Fault Tree Analysis (FTA), Failure Modes Effects and Diagnostic Analysis (FMEDA). Related monitoring and simulation is partially automated.	Safety metrics integrated into IDEs, performance monitoring, simulation, domain models. Agile team peering routinely considers safety engineering. Fully integrated supply chain safety engineering.
Selected References M. Broy, C. Leuxner, and T. Hoare, Eds., Software and Systems Safety - Specification and Verification, ser. NATO Science for Peace and Security Series - D: Information and Communication Security. IOS Press, 2011, vol. 30. [Online]. Available: http://dx.doi.org/10.3233/978-1-60750-711-6			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Semantic Web / Linked Data Awareness			
Some Big Data systems, arguably all, should map their elements to the semantic web using canonical structures such as ontologies. The semantic web supports artificial intelligence through inductive reasoning as well as machine learning. Big Data architects and users should consider safety aspects of these technologies	A knowledge engineering framework, typically manually maintained through tagging or concept trees, is provided to allow for recognition of SnP components. May or may not be implemented using semantic web standards; could be COTS or open source but idiosyncratic.	Adds to Level 1: Use of RDC or OWL to represent SnP and related components. Allows for automated reasoners and other AI tools to be employed to manage knowledge about SnP issues in the Big Data system.	Adds direct links to domain-specific and upper ontologies so that reasoning, for instance, about which test scenarios test which sorts of aspects of the SnP design, can be automatically interrogated and scheduled.
Selected References Y. Pandey and S. Bansal, "Safety check: A semantic web application for emergency management," in Proceedings of The International Workshop on Semantic Big Data, ser. SBD '17. New York, NY, USA: ACM, 2017. [Online]. Available: http://doi.acm.org/10.1145/3066911.3066917			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
SnP for Location-based Services			
Big Data Variety can facilitate deanonymization. Often Variety comes from mobile device-enabled geospatial data sources. Some applications must mitigate and educate regarding the impact of geospatial Big Data. Other applications may require geospatial Big Data as an essential resource, such as Emergency Management.	Checklists and other manual processes are in place to support risks and/or planned usage of geospatial data. Includes Big Data variety and current or potential mobile data sources.	Protections and monitoring capabilities are in place to manage geospatial data sources, including those used by third parties, customers, or partners to perform unauthorized deanonymization.	Geospatial reasoning integrated into Big Data IDE, SDLC with live links to domain, utility, and app models. Proactive detection and advisories identify risk areas for users, developers, and managers through process and automated links to domain, app, and utility models.
Selected References UN-GGIM, "A guide to the role of standards in Geospatial information management," UN Committee of Experts on Global Geospatial Information Management, Aug. 2015. [Online]. UN-GGIM, "A guide to the role of standards in Geospatial information management," UN Committee of Experts on Global Geospatial Information Management, Aug. 2015. [Online]. Available: http://kbros.co/2ulVyQv K. Liu, Y. Yao, and D. Guo, "On managing geospatial big-data in emergency management: Some perspectives," in Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management, ser. EM-GIS '15. New York, NY, USA: ACM, 2015. [Online]. Available: http://doi.acm.org/10.1145/2835596.2835614 S. Sadri, Y. Jarraya, A. Eghtesadi, and M. Debbabi, "Towards migrating security policies of virtual machines in software defined networks," in Proceedings of the 2015 1st IEEE Conference on Network Softwarization (NetSoft), Apr. 2015, pp. 1-9. [Online]. Available: http://dx.doi.org/10.1109/NETSOFT.2015.7116165 E. Bertino, B. Thuraisingham, M. Gertz, and M. L. Damiani, "Security and privacy for geospatial data: Concepts and research directions," in Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS, ser. SPRINGL '08. New York, NY, USA: ACM, 2008, pp. 6-19. [Online]. Available: http://doi.acm.org/10.1145/1503402.1503406			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Support for user annotation, notification, advisories			
To address user and enterprise safety concerns, a Big Data system should support consumer, "user," subscriber natural language annotations, notifications, and explanations. Notifications should be treated by analogy with food recall and safety notices, but survive according to Big Data planning horizons.	Web-based resources with annotation resources which persist across user sessions.	Annotations are connected to the domain and app models. Notifications can be user and system-managed and respond to internal and external SnP threat or warnings.	Annotation capabilities are connected with domain, app, and utility models. Data collected is used for SnP process improvement / refactoring. Notifications and self-managed with support for multiple channels. Must also support consent forwarding, persistence, transfer, withdrawal.
Selected References S. Szymczak, D. J. Zelik, and W. Elm, "Support for big data's limiting resource: Human attention," in Proceedings of the 2014 Workshop on Human Centered Big Data Research, ser. HCBDR '14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2609876.2609887 J. Schaffer, P. Giridhar, D. Jones, T. Höllerer, T. Abdelzaher, and J. O'Donovan, "Getting the message? A study of explanation interfaces for microblog data analysis," in Proceedings of the 20th International Conference on Intelligent User Interfaces, ser. IUI '15. New York, NY, USA: ACM, 2015, pp. 345-356. [Online]. Available: http://dx.doi.org/10.1145/2678025.2701406 E. U. Weber, "Risk attitude and preference," Wiley Interdisciplinary Reviews: Cognitive Science, vol. 1, no. 1, pp. 79-88, 2010. [Online]. Available: http://dx.doi.org/10.1002/wcs.5			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
System "Read-In" Process			
In intelligence circles, being "read into" a program formalizes the training associated with a compartmented program. This feature serves an analogous purpose for Big Data systems: people are read into the Big Data risks and guidelines of the program when they are onboarded to the project.	Persistent, career-long record of individual employee access to Big Data resources. Explicit read-in as part of employee and team member onboarding. Exit interviews include offboarding, such as cautions against unauthorized information sharing.	Level 1 plus: spans multiple employers and tracks roles assigned to employees (e.g., infrastructure, project manager, scrum master, developer, QA) within a Big Data System. Adds "read out" when employees leave that changes the Big Data configuration beyond mere password expiration.	Big Data identity management, RBAC, ABAC fully integrated with "Where used" functionality, use of ML or AI to detect insider threat at the application level. Offboarding process is part of the IDE and app teams regularly build ABAC-aware onboarding and offboarding roles as part of app domain. Domain and utility models are utilized in real time.
Selected References S. Zehra Haidry, K. Falkner, and C. Szabo, "Identifying Domain-Specific cognitive strategies for software engineering," in Proceedings of the 2017 ACM Conference on Innovation and Technology in Computer Science Education, ser. ITiCSE '17. New York, NY, USA: ACM, 2017, pp. 206-211. [Online]. Available: http://doi.acm.org/10.1145/3059009.3059032 S. Link, P. Hoyer, T. Kopp, and S. Abeck, "A Model-Driven development approach focusing human interaction," International Conference on Advances in Computer-Human Interaction, vol. 0, pp. 90-96, 2009. Y. Takahashi, T. Abiko, E. Negishi, G. Itabashi, Y. Kato, K. Takahashi, and N. Shiratori, "An Ontology-Based e-Learning system for network security," AINA, vol. 01, pp. 197-202, 2005.			
System/SW/Fingerprinting (Big Data CM)			
Big Data systems should leverage scale, velocity, and variety to automatically capture event information, such as version and timestamping at the moment of data capture, e.g., the instance of medication dispensing should capture all relevant details, not only patient, drug, and timestamp.	App designs incorporate fingerprinting of key app events, such as adding a new employee to an HR system. Level 1 goes beyond mere logging of database accesses.	Add to Level 1: Automatic connection to CMDB with transparent updating. IDEs include workflow design patterns for key app events that include full Big Data fingerprinting.	Adds live connection to domain and utility models to Level 2 conformance.
Selected References C. Dincer, G. Akpolat, and E. Zeydan, "Security issues of big data applications served by mobile operators," in 2017 25th Signal Processing and Communications Applications Conference (SIU), May 2017, pp. 1-4. [Online]. Available: http://dx.doi.org/10.1109/SIU.2017.7960253			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Temporal authority traceability			
A working assumption for Big Data systems is that data persists, might be never archived, and represents a steady trend toward limitless, low-cost storage. Thus, traceability for Big Data granting authority for design, use, and administrative policies must span infrastructure in ways that non-Big Data systems did not.	No point-in-time traceability for authority, but role auditing is performed.	Integrated point-in-time authority traceability capturing authority metadata and events using Big Data infrastructure.	Full point-in-time and replay capability (may imply full packet and EXE capture). Traceability expands beyond single enterprises, and is integrated with domain, app, and utility models.
Selected References S. Maro, A. Anjorin, R. Wohlrab, and J.-P. Steghöfer, "Traceability maintenance: Factors and guidelines," in Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, ser. ASE 2016. New York, NY, USA: ACM, 2016, pp. 414-425. [Online]. Available: http://doi.acm.org/10.1145/2970276.2970314			
Test Engineering for SnP aspects across Big Data platforms			
Test engineering for Big Data is needed to ensure that SnP measures can scale across both human (taking into account human and enterprise constraints) and computer constraints. (See also Big Data Dev Ops and Continuous Deployment.)	Test engineering for SnP includes manual checklists (e.g., NIST Cybersecurity Framework) plus scripts to test compliance with SnP requirements.	Enterprise-wide SDLC practices support test engineering. Developers routinely create test frameworks for SnP components using both off-the-shelf, reusable components and app-specific tools.	In addition to Level 2, adds ability to automatically create test scripts for SnP elements within the IDE, directly referencing domain, app, and utility models to guide test behavior. Test engineering frameworks are available to support audit and forensics activities.
Selected References J. G. Enr'ıquez, R. Blanco, F. J. Dom'ınguez-Mayo, J. Tuya, and M. J. Escalona, "Towards an MDE-based approach to test entity reconciliation applications," in Proceedings of the 7th International Workshop on Automating Test Case Design, Selection, and Evaluation, ser. A-TEST 2016. New York, NY, USA: ACM, 2016, pp. 74-77. [Online]. Available: http://doi.acm.org/10.1145/2994291.2994303 N. Garg, S. Singla, and S. Jangra, "Challenges and techniques for testing of big data," Procedia Computer Science, vol. 85, pp. 940-948, 2016, international Conference on Computational Modelling and Security (CMS 2016). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050916306354 N. Rathod and A. Surve, "Test orchestration a framework for continuous integration and continuous deployment," in 2015 International Conference on Pervasive Computing (ICPC), Jan. 2015, pp. 1-5. [Online]. Available: http://dx.doi.org/10.1109/PERVASIVE.2015.7087120			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Use ABAC to improve safety			
Expanded use of ABAC, alone or in conjunction with traditional RBAC, as part of domain model integration.	SDLC process explicitly states that ABAC is to be used in conjunction with RBAC. Use of "admin" design is deprecated. ABAC is manually tied to enterprise metadata management catalogs. Insider threat receives only light attention at Level 1 of ABAC implementation.	ABAC is built into IDEs. Developers routinely identify appropriate RBAC metadata for SnP as well as for monitoring and management. ABAC and RBAC are parts of a merging continuum. Level 2 sees a heavy reliance on domain experts to set ABAC requirements. ABAC requirements include some insider threat consideration in requirements development.	Add to Level 2: ABAC is directly linked to domain, app, and utility models. Test frameworks exercise ABAC attribute defense and vulnerabilities. Mature scenarios exist for insider threat which are tied to the use of Big Data systems to detect as well as to mitigate risk.
Selected References V. C. Hu, D. Ferraiolo, R. Kuhn, A. Schnitzer, K. Sandlin, R. Miller, and K. Scarfone, "Guide to attribute based access control (ABAC) definition and considerations," NIST, Gaithersburg, MD, Tech. Rep. SP 800-162, Jan. 2014. [Online]. Available: http://dx.doi.org/10.6028/NIST.SP.800-162D. R. Kuhn, E. J. Coyne, and T. R. Weil, "Adding attributes to Role-Based access control," Computer, vol. 43, no. 6, pp. 79-81, Jun. 2010. [Online]. Available: http://dx.doi.org/10.1109/MC.2010.155 J. Longstaff and J. Noble, "Attribute based access control for big data applications by query modification," in 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), Mar. 2016, pp. 58-65. [Online]. Available: http://dx.doi.org/10.1109/BigDataService.2016.35			

Brief Description	Safety Level 1	Safety Level 2	Safety Level 3
Value Chain Traceability			
In Big Data systems, the value chain should be preserved with the same priority that is given requirements traceability, e.g., the specialized code associated with "under test" scenarios in the VW emissions software should be traceable to the original specifications and specifiers.	Explicit, readily available checklist of values baked into the Big Data system requirements that enable users and managers to trace system features to intentional SnP risks and the levels of protection afforded given the value proposition. For citizens, specific statements of value with a plain explanation of the benefits should inform documents such as Terms of Service.	Add to Level 1: Value Requirements are present within software traceability schemes within the enterprise SDLC, e.g., encryption and intentional aggregation, classifiers in ML are directly traceable to the value proposition so that trade-offs and risks are visible.	Add to Level 2: direct link to domain, app and utility models.
Selected References A. P. J. Mol, "Transparency and value chain sustainability," Journal of Cleaner Production, vol. 107, pp. 154-161, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0959652613007762 Heindl and S. Biffl, "A case study on value-based requirements tracing," in Proceedings of the 10th European Software Engineering Conference Held Jointly with 13th ACM SIGSOFT International Symposium on Foundations of Software Engineering, ser. ESEC/FSE-13. New York, NY, USA: ACM, 2005, pp. 60-69. [Online]. Available: http://doi.acm.org/10.1145/1081706.1081717			

Appendix B: Existing Standards in Relation to the Security and Privacy Fabric

The following table introduces concepts developed in selected existing standards. There is an intentional emphasis on privacy concepts, reflecting public and enterprise concerns about Big Data security and privacy. The third column, *Security and Privacy Fabric*, is directional and notional rather than definitive at this stage of the effort. The objective is to identify Security and Privacy Fabric-specific elements of the standards and the associated concepts cited.

Table B-1: Terms and Standards in Relation to the Security and Privacy Fabric

Term	Sources	Security and Privacy Fabric	Comments
Privacy disassociability	NIST IR 8062	Privacy fabric for purposes of this analysis	Needs refinement. “Enabling the processing of PII or events without association to individuals or devices beyond the operational requirements of the system.”
Privacy subsystem predictability	NISTIR 8062		Needs refinement for Big Data
Privacy subsystem manageability	NISTIR 8062	TBD	Needs refinement for Big Data
Role: privacy subsystem oversight			
Role: privacy subsystem operations			
Role: privacy subsystem design		Architect responsibilities call-out	NISTIR 8062 groups ops & design. Separation is indicated.
Personal information			“For the purpose of risk assessment, personal information is considered broadly as any information that can uniquely identify an individual as well as any other information, events, or behavior that can be associated with an individual. Where agencies are conducting activities subject to specific laws, regulation, or policy, more precise definitions may apply.”
Privacy risk			Roughly, adverse impact X likelihood of occurrence, scoped

Term	Sources	Security and Privacy Fabric	Comments
Privacy controls: administrative			
Privacy controls: technical			
Privacy controls: physical			
Adverse privacy event			
Privacy context: system			
Privacy engineering	NISTIR 8062	Use for narrative only. May not have normative value beyond describing collection of system features, workflow elements. Operationalizing domain-specific privacy is critical.	"A specialty discipline of systems engineering focused on achieving freedom from conditions that can create problems for individuals with unacceptable consequences that arise from the system as it processes PII."
NIST privacy risk model	NISTIR 8062 Section 3.2		
Privacy metasystem issues			Draft NISTIR 8062 used "Summary Issues." "Initial contextual analyses about data actions that may heighten or decrease the assessment of privacy risk."
Privacy attack vector			Attack against Personal Information, a privacy subsystem, role, etc.
Owner/originator			System component, role or individual originating a data element.
Access*	NISTIR 7298r2, NIST SP 800-32	Includes access to workflow, orchestration	
Role: Access authority*	CNSSI-4009		Person or software
Access Control	FIPS 201		
ACL*	FIPS 201, CNSSI-4009	Consider local vs. global Big Data ACLs. How should this be integrated with ABAC?	

Term	Sources	Security and Privacy Fabric	Comments
Access control mechanism*	CNSSI-4009		
Access type*			
Accountability	NISTIR 7298		Grouped subprocesses: traceability, non-repudiation, deterrence, fault isolation, intrusion detection, intrusion prevention, after-action recovery, legal action.
Active content	NISTIR 7298r2		“Electronic documents that can carry out or trigger actions automatically on a computer platform without the intervention of a user. “
Active/passive security testing		Big data exchanges will often entail passively tested, or passive assurance for exchanges between components[i]	
Administrative Safeguards	NISTIR 7298r2		Focus on mobile and inter-organizational safeguards.
Advisory		Big Data may require a “new” grouping of advisories	“Notification of significant new trends or developments regarding the threat to the information systems of an organization. This notification may include analytical insights into trends, intentions, technologies, or tactics of an adversary targeting information systems.”
Privacy agent		Program acting on behalf of person or organization to automate a privacy-related process	There are some commercial startups that use agent-based approaches.

[i] For example, identifying where there is no active testing available (e.g., encryption assurance).

Term	Sources	Security and Privacy Fabric	Comments
Allocation	NIST SP 800-37	Useful for workflow in determining privacy responsibilities: design-time, governance-time	The process an organization employs to determine whether security controls are defined as system-specific, hybrid, or common. The process an organization employs to assign security controls to specific information system components responsible for providing a particular security capability (e.g., router, server, remote sensor).
Application	NIST SP 800-37	How would a NBDRA app be different? Refer to the application model concept in the NBD-SPSL.	
Assessment	NIST SP 800-53A	Apply to NBDRA privacy (also sec?). How different from audit? Refer to audit in the NBD-SPSL.	Grouping of terms: findings, method, object, objective, procedure, Security Control Assessor
Assurance	NIST SP 800-27, NIST SP 800-53A, CNSSI-4009	Is it possible to map to Privacy Assurance (i.e., map to analogous goals?)	"Grounds for confidence that the other four security goals (integrity, availability, confidentiality, and accountability) have been adequately met by a specific implementation. "Adequately met" includes (1) functionality that performs correctly, (2) sufficient protection against unintentional errors (by users or software), and (3) sufficient resistance to intentional penetration or by-pass."
Assurance Case (for privacy)		Is it possible to map to Privacy Assurance (i.e., map to analogous goals?). Also see below.	"A structured set of arguments and a body of evidence showing that an information system satisfies specific claims with respect to a given quality attribute. "
Assured Information sharing		Analogy for privacy sharing	"The ability to confidently share information with those who need it, when and where they need it, as determined by operational need and an acceptable level of security risk."

Term	Sources	Security and Privacy Fabric	Comments
Attack, sensing, warning; attack signature (for privacy)j		Attack signature for privacy is not the same as a general attack	"Detection, correlation, identification, and characterization of intentional unauthorized activity with notification to decision makers so that an appropriate response can be developed. "
Audit, audit data, audit log, reduction tools, audit review, audit trail		Subset created for privacy. Could be a smaller problem to solve, or a larger one, depending.k	
Authentication (various terms)		Could be needed to allow "owner" of privacy data to see or correct their own data.	
Authority		Centralized vs. decentralized authority. See blockchain as a decentralization of authority. See federation. In most applications, highly domain-specific but there are cross-functional "authorities."	
Authenticity			Provenance
Authorization			Time-limited authorization to access, or use privacy data
Authorization to operate			Interop issues for Big Data concerning privacy data
Automated privacy monitoring		To Do	Use of automated procedures to ensure that privacy controls are not circumvented or the use of these tools to track actions taken by subjects suspected of misusing the information system.

[j] Useful: Notion of a privacy attack vector is a useful big data discriminator, and may be highly system-specific.
[k] Audit for privacy could entail audit for a small subset of a larger database, or audit intended to verify that security or privacy controls are being enforced.

Term	Sources	Security and Privacy Fabric	Comments
Back door (privacy)		Use of Big Data variety to circumvent privacy safeguards	
Baseline security (for privacy controls)			The minimum privacy controls required for safeguarding an IT system based on its identified needs for confidentiality, integrity, and/or availability protection.
Behavioral outcome (for privacy fabric training)		Useful for cross-org privacy	
Biometric information		Special concern for privacy in any system?	
Body of Evidence (for security and privacy controls adherence)			“The set of data that documents the information system’s adherence to the security controls applied. The BoE will include a Requirements Verification Traceability Matrix (RVTM) delineating where the selected security and privacy controls are met and evidence to that fact can be found. The BoE content required by an Authorizing Official will be adjusted according to the impact levels selected. Refer to NIST 800-52 Section 2.3 (Rev 4).“
Boundary; boundary protection		Boundaries may need to be clarified in the NBDRA	
Browsing (for identity info)			
Business impact assessment (for privacy fabric)			“An analysis of an information system’s requirements, functions, and interdependencies used to characterize system contingency requirements and priorities in the event of a significant disruption.”
Certificate (esp. identity certificate)	CNSSI-4009	No different meaning vs. security, but perhaps more urgent context?	Certificate management may be different in privacy fabric when individual citizens (including children) are involved.

Term	Sources	Security and Privacy Fabric	Comments
Certification (see also baseline), certifier		Identify a baseline point at which privacy fabric controls were applied & certified as operational	"A comprehensive assessment of the management, operational, and technical security controls in an information system, made in support of security accreditation, to determine the extent to which the controls are implemented correctly, operating as intended, and producing the desired outcome with respect to meeting the security requirements for the system."
Chain of Custody		IoT plus Big Data for privacy	"A process that tracks the movement of evidence through its collection, safeguarding, and analysis life cycle by documenting each person who handled the evidence, the date/time it was collected or transferred, and the purpose for the transfer."
Chain of Evidence		IoT plus Big Data for privacy. Same, but applied to privacy data subset	"A process and record that shows who obtained the evidence; where and when the evidence was obtained; who secured the evidence; and who had control or possession of the evidence. The "sequencing" of the chain of evidence follows this order: collection and identification; analysis; storage; preservation; presentation in court; return to owner."
Chief Privacy Officer		To be adapted from other standards	
Classified information (*privacy subset)	NIST SP 800-60, EO 13292, CNSSI-4009	Adapt meaning from U.S. mil to apply to privacy subset	
Classified (privacy) data spillage			

Term	Sources	Security and Privacy Fabric	Comments
Clearance for access to privacy data or tools (both?)		Useful to identify fabric roles permitted to access privacy data, or to use re-identifying tools. Obvious: Data access, tools access aren't the same. See access, authorization.	"Formal certification of authorization to have access to classified information other than that protected in a special access program (including SCI). Clearances are of three types: confidential, secret, and top secret. A top-secret clearance permits access to top secret, secret, and confidential material; a secret clearance, to secret and confidential material; and a confidential clearance, to confidential material."
Common Control / Security Control Inheritance / Common criteria		Across app and data providers possibly spanning organizations. "Common criteria" is a document for privacy fabric requirements	"A security control that is inherited by one or more organizational information systems."
Common Control Provider (role for privacy)		Role responsible for inherited privacy controls	"An organizational official responsible for the development, implementation, assessment, and monitoring of common controls (i.e., security controls inherited by information systems)."
Common Misuse Scoring System for Privacy		A rough metric for potential privacy fabric weaknesses	"A set of measures of the severity of software feature misuse vulnerabilities. A software feature is a functional capability provided by software. A software feature misuse vulnerability is a vulnerability in which the feature also provides an avenue to compromise the security of a system."
Community of Interest for privacy data		A CoI may be a class of users in the privacy fabric (e.g., tribal, disabled, genetic abnormalities, high medical cost)	"A collaborative group of users who exchange information in pursuit of their shared goals, interests, missions, or business processes, and who therefore must have a shared vocabulary for the information they exchange. The group exchanges information within and between systems to include security domains."

Term	Sources	Security and Privacy Fabric	Comments
Community risk for privacy		Add – privacy fabric	"Probability that a particular vulnerability will be exploited within an interacting population and adversely impact some members of that population."
Compartmentalization (see DHHS meaning)			"A nonhierarchical grouping of sensitive information used to control access to data more finely than with hierarchical security classification alone."
Compromise – As applied to privacy		Especially re-identification	"Disclosure of information to unauthorized persons, or a violation of the security policy of a system in which unauthorized intentional or unintentional disclosure, modification, destruction, or loss of an object may have occurred."
Compromising Emanations (for privacy data)			"Unintentional signals that, if intercepted and analyzed, would disclose the information transmitted, received, handled, or otherwise processed by information systems equipment."
CND		Different for privacy fabric?	
Confidentiality	NIST SP 800-53, NIST SP 800-53A, NIST SP 800-18, NIST SP 800-27, NIST SP 800-60, NIST SP 800-37, FIPS 200, FIPS 199, 44 U.S.C., Section 3542	Traditional meaning for privacy embodied in numerous standards, despite its problems.	"Preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information."
Contamination		Scenario: a de-identified DB is placed into a system containing potentially re-identifying resources	"Type of incident involving the introduction of data of one security classification or security category into data of a lower security classification or different security category."

Term	Sources	Security and Privacy Fabric	Comments
Continuous monitoring (of privacy fabric)			"The process implemented to maintain a current security status for one or more information systems or for the entire suite of information systems on which the operational mission of the enterprise depends. The process includes: 1) the development of a strategy to regularly evaluate selected IA controls/metrics, 2) recording and evaluating IA relevant events and the effectiveness of the enterprise in dealing with those events, 3) recording changes to IA controls, or changes that affect IA risks, and 4) publishing the current security status to enable information-sharing decisions involving the enterprise."
Controlled interface		Control at the NBDRA interface for privacy fabric (different?)	"A boundary with a set of mechanisms that enforces the security policies and controls the flow of information between interconnected information systems."
Covert testing (of privacy fabric)			
Credential, credential service provider			"A trusted entity that issues or registers Subscriber tokens and issues electronic credentials to Subscribers. The CSP may encompass Registration Authorities (RAs) and Verifiers that it operates. A CSP may be an independent third party, or may issue credentials for its own use."
Criticality, criticality level		Not all privacy data elements or tools may be equal	
Cryptographic binding			"Associating two or more related elements of information using cryptographic techniques."
Conformance to privacy fabric XXX			
Data integrity (privacy corruption)		Mis-identification (e.g., TSA list)	
Default classification (for privacy data, or privacy tooling)			

Term	Sources	Security and Privacy Fabric	Comments
Digital forensics		As applied to privacy fabric: still emerging; check academic lit	
End-to-end privacy XXX		TBD	
Ethics in Design	IEEE P7000, IEEE P7002, IEEE P7007, ISO 27500		Traceability of ethics and value chain are seen as no less feasible than requirements tracing, but no more straightforward either.
Event (privacy)	CNSSI-4009	Subset of events appropriate to privacy	"Any observable occurrence in a system and/or network. Events sometimes provide indication that an incident is occurring."
External provider, external network	NIST SP 800-37, NIST SP 800-53	Critical for privacy data/controls preservation in Big Data across clouds, across organizations	"A provider of external information system services to an organization through a variety of consumer-producer relationships, including but not limited to: joint ventures; business partnerships; outsourcing arrangements (i.e., through contracts, interagency agreements, lines of business arrangements); licensing agreements; and/or supply chain exchanges."
False Acceptance		Mis-identification (?)	Biometric domain in 800-76
Hacker – Identity hacker			
Health Information Exchange	NIST IR 7497	Important as a de facto Big Data Variety source for re-identification due to U.S. ubiquity. See also UnitedHealthCare Optum	"A health information organization that brings together healthcare stakeholders within a defined geographic area and governs health information exchange among them for the purpose of improving health and care in that community."
Identification	NIST SP 800-47	TBD – Needs refinement	"The process of verifying the identity of a user, process, or device, usually as a prerequisite for granting access to resources in an IT system."

Term	Sources	Security and Privacy Fabric	Comments
Identifier	FIPS 201, CNSSI-4009	Identifiers can be automated, e.g., biometric theft, or photo recognition	"A data object - often, a printable, non-blank character string - that definitively represents a specific identity of a system entity, distinguishing that identity from all others."
Identity		Note: Review for consistent usage.	"The set of attribute values (i.e., characteristics) by which an entity is recognizable and that, within the scope of an identity manager's responsibility, is sufficient to distinguish that entity from any other entity."
Identity-based Security Policy			
Identity Binding			
Identity-based access control			
Identity proofing			
Identity token			
Identity validation			
Identity verification			
Impact, impact level, impact value	NIST SP 800-60, CNSSI-4009, NIST SP 800-34, NIST SP 800-30	Same concepts but mapped to privacy fabric	
Incident		Same meaning, covered under "confidentiality"	"An occurrence that actually or potentially jeopardizes the confidentiality, integrity, or availability of an information system or the information the system processes, stores, or transmits or that constitutes a violation or imminent threat of violation of security policies, security procedures, or acceptable use policies."
Incident handling for privacy incidents		Subset, but could be different from superset	
Indicator		Recognized signal that an adversary might be attempting to compromise privacy fabric	

Term	Sources	Security and Privacy Fabric	Comments
Information assurance for privacy			"Measures that protect and defend information and information systems by ensuring their availability, integrity, authentication, confidentiality, and non-repudiation. These measures include providing for restoration of information systems by incorporating protection, detection, and reaction capabilities."
Information Domain		Needs to be enlarged for BD privacy fabric	"A three-part concept for information sharing, independent of, and across information systems and security domains that 1) identifies information sharing participants as individual members, 2) contains shared information objects, and 3) provides a security policy that identifies the roles and privileges of the members and the protections required for the information objects."
Information Operations (as applied to identity disruption)	CNSSI-4009		"The integrated employment of the core capabilities of electronic warfare, computer network operations, psychological operations, military deception, and operations security, in concert with specified supporting and related capabilities, to influence, disrupt, corrupt, or usurp adversarial human and automated decision-making process, information, and information systems while protecting our own."
Information owner			
Information sharing environment		Highlight as a potential area for variety-enabled identification	"ISE in its broader application enables those in a trusted partnership to share, discover, and access controlled information."
Information Security Architect (sub: privacy)	NIST SP 800-39	Identifies design-time role. Architecture refers to the design.	
Information Steward (for confidential data, tools)			"An agency official with statutory or operational authority for specified information and responsibility for establishing the controls for its generation, collection, processing, dissemination, and disposal."

Term	Sources	Security and Privacy Fabric	Comments
IS Resilience		Does this notion apply to identity attacks specifically?	
IS Security Risks (privacy subset)			"Information system-related security risks are those risks that arise through the loss of confidentiality, integrity, or availability of information or information systems and consider impacts to the organization (including assets, mission, functions, image, or reputation), individuals, other organizations, and the Nation."
Information Value			"A qualitative measure of the importance of the information based upon factors such as: level of robustness of the Information Assurance controls allocated to the protection of information based upon: mission criticality, the sensitivity (e.g., classification and compartmentalization) of the information, releasability to other countries, perishability/longevity of the information (e.g., short-life data versus long-life intelligence source data), and potential impact of loss of confidentiality and integrity and/or availability of the information."
Insider threat for confidentiality breaches		E.g., access to personnel records, authentication systems, ACLs	
Intellectual property		Especially IP connected to or owned by a person, but also IP treated the same way as "privacy" data. Further study.l	
Interconnection Security Agreement	NIST SP 800-47, CNSSI-4009		

[1] IP protections, defenses, risks are similar but also different from individual human privacy.

Term	Sources	Security and Privacy Fabric	Comments
Interface Control Document		Different for privacy?	
Internal network privacy controls		Use cases are different	
IT privacy awareness and training program			
IT privacy policy (three + types)	NIST SP 800-12	Program policy; issue (context specific) policies; system- or device- or app-specific policies	"1) Program Policy—high-level policy used to create a Program policy - organization's IT security program, define its scope within the organization, assign implementation responsibilities, establish strategic direction, and assign resources for implementation. 2) Issue-Specific Policies—address specific issues of concern to the organization, such as contingency planning, the use of a particular methodology for systems risk management, and implementation of new regulations or law. These policies are likely to require more frequent revision as changes in technology and related factors take place. 3) System-Specific Policies—address individual systems, such as establishing an access control list or in training users as to what system actions are permitted. These policies may vary from system to system within the same organization. In addition, policy may refer to entirely different matters, such as the specific managerial decisions setting an organization's electronic mail (email) policy or fax security policy."
Key terminology: list, loader, management, logger, exchange, escrow, etc.		TBD—Map to confidentiality-specific logging for a specific domain.	See also utility domains, e.g., ubiquitous O.S. logging, or packet capture.
Least trust		Metrics needed for trust components & disclosed to originator/owner	"The principal that a security architecture should be designed in a way that minimizes 1) the number of components that require trust, and 2) the extent to which each component is trusted."

Term	Sources	Security and Privacy Fabric	Comments
Line-of-business privacy guidelines	OMB, NIST SP 800-60, OMB Business Reference Model FEA V2.3	Domain- or discipline-specific privacy best practicesm	Lengthy discussion best framed through HL7 FHR domain model use case.
List-oriented object privacy protection	CNSSI-4009		
Major / Minor application (for privacy)	OMB Circular A-130 Appendix III, NIST SP 800-18	What makes it major / minor in the NBDRA? Not resolved in V2.	
Masquerading privacy data (see identity)	NIST SP 800-19		
Biometric match event	FIPS 201, CNSSI-4009		Possible paradigmatic event exemplar for Big Data
Media (wearable, implanted digital device)	FDA, adapted from NIST SP 800-53		
Memorandum of Understanding for Privacy data (MOUP)	Simple MOU was NIST SP 800-47		Critical for Big Data Variety
Minor application (susceptible to privacy concerns)	NIST SP 800-18		Identify aspect of a larger application that applies to privacy
Mission/business segment*	NIST SP 800-30		Identify segment associated with business processes that collect PII or other privacy data at risk
Multilevel security (for privacy data)	CNSSI-4009		Applies MLS to privacy data subset
Mutual suspicion	CNSSI-4009		As applicable to privacy data, e.g., consider privacy data across organizational boundaries
National security system (US)	FIPS 200		Use to identify possible exclusions or variations from otherwise universal guidelines or practices. Nation-specific.
Need to know determination	CNSSI-4009		Need to know for PII.
Needs assessment for privacy (policy, risk, etc.)	NIST SP 800-50		"The results of a needs assessment can provide justification to convince management to allocate adequate resources to meet the identified awareness and training needs."

[m] LOB or Domain-specific privacy. See also incidents, events, etc. Needs improved definition and examples.

Term	Sources	Security and Privacy Fabric	Comments
Privacy data resilience	Adapted from CNSSI-4009		Ability to sustain business operations after privacy data attack (e.g., partial leak)
Non-organizational user	NIST SP 800-53		
Network sponsor (for privacy components)	CNSSI-4009		"Individual or organization responsible for stating the security policy enforced by the network, designing the network security architecture to properly enforce that policy, and ensuring that the network is implemented in such a way that the policy is enforced."
Non-repudiation (for PII)	CNSSI-4009		As applied to sender/recipient of PII
Operational controls (for PII)	NIST SP 800-53		"The security controls (i.e., safeguards or countermeasures) for an information system that primarily are implemented and executed by people (as opposed to systems)."
Operations Security (OPSEC, for PII)	CNSSI-4009		"Systematic and proven process by which potential adversaries can be denied information about capabilities and intentions by identifying, controlling, and protecting generally unclassified evidence of the planning and execution of sensitive activities. The process involves five steps: identification of critical information, analysis of threats, analysis of vulnerabilities, assessment of risks, and application of appropriate countermeasures."
Organizational information security continuous monitoring	NIST SP 800-137		"Ongoing monitoring sufficient to ensure and assure effectiveness of security controls related to systems, networks, and cyberspace, by assessing security control implementation and organizational security status in accordance with organizational risk tolerance – and within a reporting structure designed to make real-time, data-driven risk management decisions."
Organizational Registration Authority	CNSSI-4009		"Entity within the PKI that authenticates the identity and the organizational affiliation of the users."

Term	Sources	Security and Privacy Fabric	Comments
Overt testing for privacy	NIST SP 800-115		"Security testing performed with the knowledge and consent of the organization's IT staff."
Partitioned security mode	CNSSI-4009		"Information systems security mode of operation wherein all personnel have the clearance, but not necessarily formal access approval and need-to-know, for all information handled by an information system."
Path histories	NIST SP 800-19		"Maintaining an authenticatable record of the prior platforms visited by a mobile software agent, so that a newly visited platform can determine whether to process the agent and what resource constraints to apply."
Pen testing (for variety attacks)	NIST SP 800-53A		Applies principles of pen testing to attempts to re-identify or identify PII
Periods processing	CNSSI-4009		"The processing of various levels of classified and unclassified information at distinctly different times. Under the concept of periods processing, the system must be purged of all information from one processing period before transitioning to the next."
Personal Identity Verification	CNSSI-4009		Applies U.S. Federal ID standard to other organizations
Personal Identity Verification Authorization Official (role)	See related definitions in FIPS 201		Person in an org responsible for issuing identity credentials
PII			"Information which can be used to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc., alone, or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc."
Personnel Registration Manager (role)			"Management role that is responsible for registering human users."

Term	Sources	Security and Privacy Fabric	Comments
PII Confidentiality Impact Level	NIST SP 800-122		"The PII confidentiality impact level—low, moderate, or high—indicates the potential harm that could result to the subject individuals and/or the organization if PII were inappropriately accessed, used, or disclosed."
Policy-based Access, Certifier, etc.	Set of concepts around POA&M		Use broad framework to help organizations identify responsibilities for managing PII policies associated with a system.
Potential (privacy) impact	CNSSI-4009		“"The loss of confidentiality, integrity, or availability that could be expected to have a limited (low) adverse effect, a serious (moderate) adverse effect, or a severe or catastrophic (high) adverse effect on organizational operations, organizational assets, or individuals.”
Privacy	NIST SP 800-32		"Restricting access to subscriber or Relying Party information in accordance with federal law and agency policy."
Privacy Impact Assessment	NIST SP 800-53		"An analysis of how information is handled: 1) to ensure handling conforms to applicable legal, regulatory, and policy requirements regarding privacy; 2) to determine the risks and effects of collecting, maintaining, and disseminating information in identifiable form in an electronic information system; and 3) to examine and evaluate protections and alternative processes for handling information to mitigate potential privacy risks."
Privacy system	CNSSI-4009		"Commercial encryption system that affords telecommunications limited protection to deter a casual listener, but cannot withstand a technically competent cryptanalytic attack."
Privilege Management	NIST IR 7657		"The definition and management of policies and processes that define the ways in which the user is provided access rights to enterprise systems. It governs the management of the data that constitutes the user’s privileges and other attributes, including the storage, organization and access to information in directories."

Term	Sources	Security and Privacy Fabric	Comments
Profiling (of people)	NIST SP 800-61		"Measuring the characteristics of expected activity so that changes to it can be more easily identified."
Proprietary information (owned by people versus organizations)			"Material and information relating to or associated with a company's products, business, or activities, including but not limited to financial information; data or statements; trade secrets; product research and development; existing and future product designs and performance specifications; marketing plans or techniques; schematics; client lists; computer programs; processes; and know-how that has been clearly identified and properly marked by the company as proprietary information, trade secrets, or company confidential information. The information must have been developed by the company and not be available to the government or to the public without restriction from another source."
Pseudonym	NIST SP 800-63		“A name other than a legal name.”
Residual risk (e.g., after PII breach)	NIST SP 800-33		"The remaining potential risk after all IT security measures are applied. There is a residual risk associated with each threat."
Risk	NIST SP 800-53		"Information system-related security risks are those risks that arise from the loss of confidentiality, integrity, or availability of information or information systems and consider the adverse impacts to organizational operations (including mission, functions, image, or reputation), organizational assets, individuals, other organizations, and the Nation."
Risk-Adaptable Access Control	CNSSI-4009		
Risk Analysis	NIST SP 800-27		
Risk Management Framework, Risk Model, Monitoring, Response, Response Measure, Tolerance, Executive	NIST SP 800-30, NIST SP 800-53A, NIST SP 800-37, CNSSI-4009, FIPS 200, NIST SP 800-34, NIST SP 800-82		Suite of risk-related taxonomy

Term	Sources	Security and Privacy Fabric	Comments
Risk Assessor	NIST SP 800-30		"The individual, group, or organization responsible for conducting a risk assessment."
Role	NIST SP 800-95		"A group attribute that ties membership to function. When an entity assumes a role, the entity is given certain rights that belong to that role. When the entity leaves the role, those rights are removed. The rights given are consistent with the functionality that the entity needs to perform the expected tasks."
Role-based Access Control (RBAC)	NIST SP 800-95		
Rule-Based Security (Privacy) Policy	NIST SP 800-33, CNSSI-4009		"A security policy based on global rules imposed for all subjects. These rules usually rely on a comparison of the sensitivity of the objects being accessed and the possession of corresponding attributes by the subjects requesting access. Also known as discretionary access control (DAC)."
Security Category	FIPS 200, FIPS 199, NIST SP 800-18		"The characterization of information or an information system based on an assessment of the potential impact that a loss of confidentiality, integrity, or availability of such information or information system would have on organizational operations, organizational assets, individuals, other organizations, and the Nation."
Security (Privacy) Domain	NIST SP 800-27		"A collection of entities to which applies a single security policy executed by a single authority." – Concept modified to reflect privacy only.
Security (Privacy) Engineering	CNSSI-4009		Need to reconcile with Oasis standard
Security (privacy) filter	CNSSI-4009		"A secure subsystem of an information system that enforces security policy on the data passing through it."
Security (privacy) incident		Fabric-specific	

Term	Sources	Security and Privacy Fabric	Comments
Security (privacy) label	NIST SP 800-53, FIPS 188	Important for provenance	"A marking bound to a resource (which may be a data unit) that names or designates the security attributes of that resource."
Security (privacy) level	FIPS 188	NBDRA adaptation	"A hierarchical indicator of the degree of sensitivity to a certain threat. It implies, according to the security policy being enforced, a specific level of protection."
Security (privacy) marking	NIST SP 800-53		"Human-readable information affixed to information system components, removable media, or output indicating the distribution limitations, handling caveats, and applicable security markings."
Security (privacy) plan	NIST SP 800-53, NIST SP 800-53A, NIST SP 800-37, NIST SP 800-18		"Formal document that provides an overview of the security requirements for an information system or an information security program and describes the security controls in place or planned for meeting those requirements."
Security (privacy) policy		Needs to be greatly enlarged as it includes both practice and colloquial uses	"Set of criteria for the provision of security services."
Security (privacy) posture	CNSSI-4009		"The security status of an enterprise's networks, information, and systems based on IA resources (e.g., people, hardware, software, policies) and capabilities in place to manage the defense of the enterprise and to react as the situation changes."
Security (privacy) impact analysis	CNSSI-4009		
Security (privacy) program plan	CNSSI-4009		
Security (privacy) range	CNSSI-4009		"Highest and lowest security levels that are permitted in or on an information system, system component, subsystem, or network."

Term	Sources	Security and Privacy Fabric	Comments
Security (privacy)-relevant change or event	CNSSI-4009		"Any change to a system's configuration, environment, information content, functionality, or users which has the potential to change the risk imposed upon its continued operations."
Security (privacy) requirements	CNSSI-4009		Mandated privacy requirements
Security (privacy) requirements traceability matrix	CNSSI-4009		
Security (Privacy) Safeguards	CNSSI-4009		
Security (privacy) service	NIST SP 800-27		"A capability that supports one, or many, of the security goals. Examples of security services are key management, access control, and authentication."
Security (privacy) tag	FIPS 188		"Information unit containing a representation of certain security-related information (e.g., a restrictive attribute bit map)."
Security (privacy) test, evaluation, assess, etc.	CNSSI-4009		
Sensitivity (for privacy data) label	CNSSI-4009		"Information representing elements of the security label(s) of a subject and an object. Sensitivity labels are used by the trusted computing base (TCB) as the basis for mandatory access control decisions. See Security Label."
SLA for Privacy		TBD	
Signed data (applied to privacy)	CNSSI-4009		
Privacy Spillage	CNSSI-4009		"Security incident that results in the transfer of classified or CUI information onto an information system not accredited (i.e., authorized) for the appropriate security level."
Status (for privacy components) monitoring	NIST SP 800-137	Person or s/w agent	"Monitoring the information security metrics defined by the organization in the information security ISCM strategy."

Term	Sources	Security and Privacy Fabric	Comments
Suppression measure (applied to privacy)	CNSSI-4009		"Action, procedure, modification, or device that reduces the level of, or inhibits the generation of, compromising emanations in an information system."
Privacy Integrity	NIST SP 800-27		Adapt from System Integrity?
Privacy subsystem Interconnect	NIST SP 800-47, CNSSI-4009	What contexts?	
System of Records	NIST SP 800-122		"A group of any records under the control of any agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual."
Privacy System owner		Adapt from System Owner?	"Person or organization having responsibility for the development, procurement, integration, modification, operation and maintenance, and/or final disposition of an information system."
Technical Privacy Security Controls	CNSSI-4009	See also Technical Reference Model adapted for Privacy	"Security controls (i.e., safeguards or countermeasures) for an information system that are primarily implemented and executed by the information system through mechanisms contained in the hardware, software, or firmware components of the system."
Privacy – Threat definition, analysis, assessment, event, scenario, source	NIST SP 800-27, CNSSI-4009		
Tracking cookie	NIST SP 800-83		
Traffic Analysis	NIST SP 800-24, NIST SP 800-98	Highly applicable to privacy in IoT	"A form of passive attack in which an intruder observes information about calls (although not necessarily the contents of the messages) and makes inferences, e.g., from the source and destination numbers, or frequency and length of the messages."
Trusted Agent TBD	See trusted identification forwarding and related terms	Earliest or most responsible (TBD) direct digital connection to a person whose data is private	

Term	Sources	Security and Privacy Fabric	Comments
Unauthorized disclosure (privacy data)	FIPS 191		
Privacy data not identified as such by a system			
User ID	CNSSI-4009		
User Registration	NIST SP 800-57		
User Representation			
Vulnerability assessment (for privacy)			

Appendix C: 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 C-1 (Fang Liu, 2011).

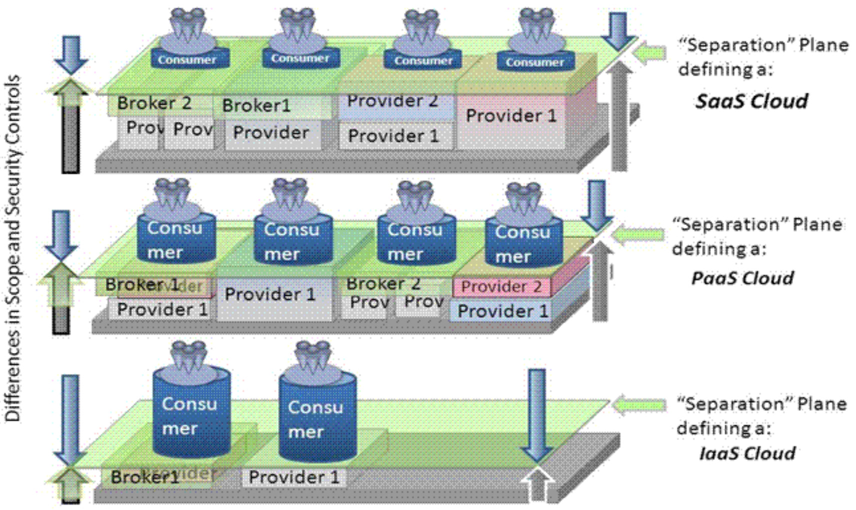


Figure C-1: Composite Cloud Ecosystem Security Architecture

When unraveling the complexity of multiple interdependencies, it is important to note that enterprise-wide access controls fall within the purview of a well thought out Big Data and cloud ecosystem risk management strategy for end-to-end enterprise 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; and
   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 CIA 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

Despite the fact that cloud computing is driving innovation in technologies that support Big Data, some Big Data projects are not in the cloud. However, because of the resurgence of the cloud, considerable work has been invested in developing cloud standards to alleviate concerns over its use. A number of organizations, including NIST, are diligently engaged in standards work around cloud computing. Central among these for Big Data Security and Privacy is NIST SP 800-144 (Jansen & Grance, 2011), which included a then-current list of related standards and guides, which is reproduced in Table C-1.

Table C-1: Standards and Guides Relevant to Cloud Computing

Publication	Title
FIPS 199	Standards for Security Categorization of Federal Information and Information Systems
FIPS 200	Minimum Security Requirements for Federal Information and Information Systems
NIST SP 800-18, Revision 1	Guide for Developing Security Plans for Federal Information Systems

NIST SP 800-34, Revision 1	Contingency Planning Guide for Federal Information Systems
NIST SP 800-37, Revision 1	Guide for Applying the Risk Management Framework to Federal Information Systems: A Security Life Cycle Approach
NIST SP 800-39	Managing Information Security Risk: Organization, Mission, and Information System View
NIST SP 800-53, Revision 4	Recommended Security Controls for Federal Information Systems and Organizations
NIST SP 800-53, Appendix J	Privacy Control Catalog
NIST SP 800-53A, Revision 4	Guide for Assessing the Security Controls in Federal Information Systems
NIST SP 800-60, Revision 1	Guide for Mapping Types of Information and Information Systems to Security Categories
NIST SP 800-61, Revision 2	Computer Security Incident Handling Guide
NIST SP 800-64, Revision 2	Security Considerations in the System Development Life Cycle
NIST SP 800-86	Guide to Integrating Forensic Techniques into Incident Response
NIST SP 800-88, Revision 1	Guidelines for Media Sanitization
NIST SP 800-115	Technical Guide to Information Security Testing and Assessment
NIST SP 800-122	Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)
NIST SP 800-137	Information Security Continuous Monitoring for Federal Information Systems and Organizations

The following section revisits the traditional access control framework. The traditional framework identifies a standard set of attack surfaces, roles, and trade-offs. 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.[n] 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 be granted only 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, smartphones, 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

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

accessed via a remote connection/access point as well as internally. Therefore, visibility as to who is accessing the data is critical in protecting the data. The trade-offs between strict data access control versus conducting business requires answers to questions such as 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, (Anderson, 2011) 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 measures 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 their job.
- If one of the data elements is protected because of its classification (e.g., PII, HIPAA, PCI), then all 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 (i.e., 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 their 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 their 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 D: Big Data Actors and Roles—Adaptation to Big Data Scenarios

SOAs were a widely discussed paradigm through the early 2000s. 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 as well as across services. (Patig, 2008) (M. López-Sanz, 2008) 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. [202] 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; and
- 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 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 to establish a self-destruct timeline (similar to the EU's *right to be forgotten*).

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

E.1 Retail/Marketing

E.1.1 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 E-1: 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. [203] 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/ FHE	Unknown
	Audits	Audit DRM usage for royalties
Framework Provider	Securing data storage and transaction logs	Unknown
	Key management	Unknown

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
	Security best practices for non-relational data stores	Unknown
	Security against DoS attacks	N/A
	Data provenance	Traceability to data owners, producers, consumers is preserved
Fabric	Analytics for security intelligence	Machine intelligence for unsanctioned use/access
	Event detection	“Playback” granularity defined
	Forensics	Subpoena of playback records in legal disputes

E.1.2 Nielsen Homescan: Project Apollo

Nielsen Homescan involves family-level retail transactions and associated media exposure using a statistically valid national sample. A general description [204] 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 E-2: 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 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/FHE	N/A
	Audits	Schematron and process step audits

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
Framework Provider	Securing data storage and transaction logs	Project-specific audits secured by infrastructure team.
	Key management	Managed by project chief security officer (CSO). Separate key pairs issued for customers and internal users.
	Security best practices for non-relational data stores	Regular data integrity checks via XML schema validation
	Security against DoS attacks	Industry-standard webhost protection provided for query subsystem.
	Data provenance	Unique
Fabric	Analytics for security intelligence	No project-specific initiatives
	Event detection	N/A
	Forensics	Usage, cube-creation, and device merge audit records were retained for forensics and billing

E.1.3 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 E-3: 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 time stamp 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/ FHE	Unknown
	Audits	Customer audits for accuracy and integrity are supported

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
Framework Provider	Securing data storage and transaction logs	Storage archiving—this is a big issue
	Key management	CSO and applications
	Security best practices for non-relational data stores	Unknown
	Security against DoS attacks	Standard
	Data provenance	Server, application, IP-like identity, page point-in-time Document Object Model (DOM), and point-in-time marketing events
Fabric	Analytics for security intelligence	Access to web logs often requires privilege elevation.
	Event detection	Can infer; for example, numerous sales, marketing, and overall web health events
	Forensics	See the SIEM use case

E.2 Healthcare

E.2.1 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 E-4: Mapping HIE to the Reference Architecture

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
Data Provider → Application Provider	End-point input validati[205]on	Strong authentication, perhaps through X.509v3 certificates, potential leverage of SAFE (Signatures & Authentication for Everything [205]) 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

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
		filtering based on the policies of the data source or the HIE service provider.
	Secure data aggregation	Combining deduplication with encryption is desirable. Deduplication improves bandwidth and storage availability, but when used in conjunction with encryption, presents particular challenges (*Reference here*). Other columns may require cryptographic metadata for facilitating aggregation and deduplication. The HL7 standards organization is currently studying this set of related use cases. (Weida, 2014)
Application Provider → Data Consumer	Privacy-preserving data analytics	Searching on encrypted data and proofs of data possession. Identification of potential adverse experience due to clinical trial participation. Identification of potential professional patients. Trends and epidemics, and co-relations of these to environmental and other effects. Determination of whether the drug to be administered will generate an adverse reaction, without breaking the double blind. Patients will need to be provided with detailed accounting of accesses to, and uses of, their EHR data.
	Compliance with regulations	HIPAA security and privacy will require detailed accounting of access to EHR data. Facilitating this, and the logging and alerts, will require federated identity integration with data consumers. Where applicable, compliance with U.S. FDA CFR Title 21 Part 56 on Institutional Review Boards is mandated.
	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/ FHE	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

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

E.2.2 Genetic Privacy

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

E.2.3 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 E-5: 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	Opaque—company-specific
	Real-time security monitoring	None
	Data discovery and classification	Opaque—company-specific
	Secure data aggregation	Third-party aggregator
Application Provider → Data Consumer	Privacy-preserving data analytics	Data to be reported in aggregate but preserving potentially small-cell demographics
	Compliance with regulations	Responsible developer and third-party custodian
	Government access to data and freedom of expression concerns	Limited use in research community, but there are possible future public health data concerns. Clinical study reports only, but possibly selectively at the study- and patient-levels

NBDRA Component and Interfaces	Security & Privacy Topic	Use Case Mapping
Data Provider ↔ Framework Provider	Data-centric security such as identity/policy-based encryption	TBD
	Policy management for access control	Internal roles; third-party custodian roles; researcher roles; participating patients' physicians
	Computing on the encrypted data: searching/filtering/deduplicate/ FHE	TBD
	Audits	Release audit by a third party
Framework Provider	Securing data storage and transaction logs	TBD
	Key management	Internal varies by firm; external TBD
	Security best practices for non-relational data stores	TBD
	Security against DoS attacks	Unlikely to become public
	Data provenance	TBD—critical issue
Fabric	Analytics for security intelligence	TBD
	Event detection	TBD
	Forensics	

E.3 Cybersecurity

E.3.1 Network Protection

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

Table E-6: 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. [206]
	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	Ensure that access by law enforcement, state or local agencies, such as for child protection, or to aid locating missing persons, is lawful.
Data Provider ↔	Data-centric security such as identity/policy-based encryption	Usually a feature of the operating system

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
Framework Provider	Policy management for access control	For example, a group policy for an event log
	Computing on the encrypted data: searching/filtering/deduplicate/ FHE	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

E.4 Government

E.4.1 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.

Table E-7: 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 pipeline and elsewhere in federal data systems. 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

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
	Secure data aggregation	Fusion challenges range from simple to complex. Video streams may be used [207] unsecured or unaggregated.
Application Provider → Data Consumer	Privacy-preserving data analytics	Geospatial constraints: cannot surveil beyond Universal Transverse Mercator (UTM). 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 contractor-devised system schemes
	Computing on the encrypted data: searching/filtering/deduplicate/ FHE	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, e.g., see the DoD Cloud Computing Strategy. [208]
	Security against DoS attacks	Anti-jamming e-measures
	Data provenance	Must track to sensor point in time configuration and metadata
Fabric	Analytics for security intelligence	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. [209]
	Forensics	Used for after action review (AAR)—desirable to have full playback of sensor streams

E.4.2 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 E-8: 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/ FHE	Proposed [210]
	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

E.5 Industrial: Aviation

E.5.1 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.

E.6 Transportation

E.6.1 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 E-9: 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. [211]
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/ FHE	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.

NBDRA Component and Interfaces	Security and Privacy Topic	Use Case Mapping
Fabric	Analytics for security intelligence	Anomalies in sensor data can indicate tampering/fraudulent insertion of data traffic.
	Event detection	Abnormal events such as cargo moving out of the way or being stationary for unwarranted periods can be detected.
	Forensics	Analysis of logged data can reveal details of incidents after they occur.

Appendix F: Version 2 Changes and New Topics

The current version of the *NBDIF: Volume 4, Security and Privacy* document reflects changes in the technology environment (e.g., as well as ongoing work within the NBD-PWG). For Version 2, the Security and Privacy Subgroup considered the following topics:

1. See Cryptographic Technologies for Data Transformations. The latest document is updated to reflect recent cryptology practices.
2. The NBD-SPSL is introduced, suitable for use by unaffiliated citizens, Big Data software architects, and IT managers. (Refer to related IEC standards 61508, 61671, 62046, SC22 WG 23.)
3. Provided levels of conformance to Big Data security and privacy practices. Low, medium and high conformance levels were added. (See related work in “Conformity Assessment” of the “NIST Roadmap for Improving Critical Infrastructure Cybersecurity.”) The approach taken is similar to NIST 800-53.
4. Improved descriptions of security and privacy dependency frameworks that interoperate across enterprises, applications, and infrastructure are cited in the NBD-SPSL.
5. The current version reflects the growing importance of security and privacy aspects to the API-first and microservices design pattern.
6. The NBD-SPSL directly addresses security and privacy issues with geospatial and mobile data. [212]
7. The NBD-SPSL includes security hardening through software-defined networks and other virtual network security concepts, as in IEEE P1915.1 and NIST 800-125B. [213]
8. This document now provides references to third-party references on risks, verifiability, and provenance for analytics that affect security and privacy.

Appendix G: Acronyms

AAR	After Action Review
ABAC	Attribute Based Access Control
ABE	Attribute-Based Encryption
AC&S	Access Control and Security
ACL	Access Control List
ACM	Association for Computing Machinery
AI	Artificial Intelligence
API	Application Programming Interface
ARM	Application Release Management
AuthN/AuthZ	Authentication/Authorization
BAA	Business Associate Agreement
BYOD	Bring Your Own Device
CADF	Cloud Auditing Data Federation
CAT	SEC Consolidated Audit Trail
CDC	U.S. Centers for Disease Control and Prevention
CEP	Complex Event Processing
CFR	Code of Federal Regulations
CIA	Confidentiality, Integrity, and Availability
CIO	Chief Information Officer
CISSP	Certified Information Systems Security Professional
CM	Configuration Management
COPPA	Children’s Online Privacy Protection Act
CPE	Common Platform Enumeration
CPS	Cyber-Physical System
CPU	Central Processing Unit
CSA BDWG	Cloud Security Alliance Big Data Working Group
CSP	Cloud Service Provider
DevOps	a clipped compound of software DEVelopment and information technology OPerationS
DevSecOps	Security and Safety Engineering in DevOps
DHHS	U.S. Department of Health and Human Services
DISA	Defense Information Systems Agency

DoD	U.S. Department of Defense
DoS	Denial of Service
DR	Disaster Recovery
DRM	Digital Rights Management
EDM	Enterprise Data Management
EFPIA	European Federation of Pharmaceutical Industries and Associations
EHR	Electronic Health Record
EPA	Explicit role-permission Assignments
ETSI	European Telecommunications Standards Institute
EU	European Union
FAA	Federal Aviation Administration
FDA	U.S. Food and Drug Administration
FERPA	Family Educational Rights and Privacy Act
FHE	Fully Homomorphic Encryption
FHIR	Fast Healthcare Interoperability Resources
FIBO	Financial Industry Business Ontology
FTC	Federal Trade Commission
GPS	Global Positioning System
GRC	Governance, Risk management, and Compliance
HCI	Human Computer Interaction
HIE	Health Information Exchange
HIPAA	Health Insurance Portability and Accountability Act
HPC	High Performance Computing
HR	Human Resources
HTML	HyperText Markup Language
IA	Information Assurance
IaaS	Infrastructure as a Service
IAM	Identity Access Management
IBE	Identity-Based Encryption
IDE	Integrated Development Environment
IdP	Identity provider
IEEE	Institute of Electrical and Electronics Engineers
INCITS	International Committee for Information Technology Standards
IoT	Internet of Things
ISO	International Organization for Standardization

ISSEA	International Systems Security Engineering Association
IT	Information Technology
ITL	Information Technology Laboratory at NIST
KMS	Key Management Systems
M2M	Machine to Machine
MAC	Media Access Control
MBSE	Model-based Systems Engineering
MIFE	Multi-input Functional Encryption
ModSim	Modeling and Simulation
MPC	Multi-party Computations
NBDIF	NIST Big Data Interoperability Framework
NBD-PWG	NIST Big Data Public Working Group
NBDRA	NIST Big Data Reference Architecture
NBD-SPSL	NIST Big Data Security and Privacy Safety Levels
NSTIC	National Strategy for Trusted Identities in Cyberspace
OASIS	Organization for the Advancement of Structured Information Standards
OECD	Organisation for Economic Co-Operation and Development
OMG	Object Management Group
OSS	Operations Support Systems
PaaS	Platform as a Service
PCI	Payment Card Industry
PCI-DSS	Payment Card Industry Data Security Standard
PHI	Protected Health Information
PhRMA	Pharmaceutical Research and Manufacturers of America
PII	Personally Identifiable Information
PKI	Public Key Infrastructure
PMML	Predictive Model Markup Language
PMRM	Privacy Management Reference Model
RBAC	Role-based Access Control
RDF	Resource Description Framework
RPAS	Remotely Piloted Aircraft System
RPV	Remotely Piloted Vehicle
SaaS	Software as a Service
SAML	Security Assertion Markup Language
SCAP	Security Content Automation Protocol

SDLC	Systems Development Life Cycle
SDL-IT	Secure Development Life Cycle
SDN	Software-Defined Network
SEC	U.S. Securities and Exchange Commission
SGX	Software Guard Extensions
SIEM	Security Information and Event Management
SKOS	Simple Knowledge Organization System
SKUs	Stock Keeping Units
SOA	Service-oriented architectures
SON	Self-Organizing Networks
S-SDLC	Secure-SDLC
SSE	Searchable Symmetric Encryption
SSE-CMM	Systems Security Engineering Capability Maturity Model
SSL	Secure Sockets Layer
STS	Security Token Service
SWID	Software Identification
TCB	Trusted Computing Base
TCP/IP	Transmission Control Protocol/Internet Protocol
TLS	Transport Layer Security
TOSCA	Topology and Orchestration Specification for Cloud Applications
TPM	Trusted Platform Module
TSA	Transportation Security Administration
UAS	Unmanned Aerial Systems
UAV	Unmanned Aerial Vehicle
UDP	User Datagram Protocol
US¬CERT	U.S. Computer Emergency Readiness Team
VC3	Verifiable Confidential Cloud Computing
VM	Virtual Machine
VPN	Virtual Private Network
XACML	eXtensible Access Control Markup Language
XML	eXtensible Markup Language
XMPP	Extensible Messaging and Presence Protocol

Appendix H: References

[1] W. Chang and NIST Big Data Public Working Group, "NIST Big Data Interoperability Framework: Volume 1, Definitions (SP1500-1)," 2015.

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

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

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

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

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

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

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

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

[10] V. Hu *et al.*, "NIST SP 800-162: Guide to Attribute Based Access Control (ABAC) Definition and Considerations," *NIST Spec. Publ. 800-162*, vol. 800, no. 162, 2014.

[11] D. Spinellis, "Service orchestration with Rundeck," *IEEE Softw.*, vol. 31, no. 4, pp. 16–18, 2014.

[12] National Institute of Standards and Technology (NIST), "NIST SP 800-53 Rev.4: Security and Privacy Controls for Federal Information Systems and Organizations," 2014.

[13] A. Zaslavsky, C. Perera, and D. Georgakopoulos, "Sensing as a Service and Big Data," *Proc. Int. Conf. Adv. Cloud Comput.*, pp. 21–29, 2012.

[14] M. Abramson *et al.*, "Data Residency Challenges: A Joint Paper with the Object Management Group," Cloud Standards Customer Council, Needham Heights, MA OR - Cloud Standards Customer Council, May 2017.

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

[16] "IBM, Subgroup correspondence with James G Kobielus." 2014.

[17] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman, "Information accountability," *Commun. ACM*, vol. 51, no. 6, pp. 82–87, 2008.

[18] M. Altman, D. O'Brien, S. Vadhan, and A. Wood, "Can You Have Privacy and Big Data Too? — Comments for the White House," *MIT Libraries: Program on Information Science*, 2014.

[Online]. Available: http://informatics.mit.edu/blog/2014/03/can-you-have-privacy-and-big-data-too-—-comments-white-house.

[19] Cloud Security Alliance Big Data Working Group, “Top 10 Challenges in Big Data Security and Privacy,” 2012.

[20] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy-preserving data publishing: A Survey of Recent Developments,” *ACM Comput. Surv.*, vol. 42, no. 4, pp. 1–53, 2010.

[21] C. Dwork, “Differential privacy,” *Proc. 33rd Int. Colloq. Autom. Lang. Program.*, pp. 1–12, 2006.

[22] L. SWEENEY, “k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY,” *Int. J. Uncertainty, Fuzziness Knowledge-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[23] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in *Proceedings - IEEE Symposium on Security and Privacy*, 2008, pp. 111–125.

[24] S. S. Sahoo, A. Sheth, and C. Henson, “Semantic provenance for eScience: Managing the deluge of scientific data,” *IEEE Internet Comput.*, vol. 12, no. 4, pp. 46–54, 2008.

[25] J. Wang, D. Crawl, S. Purawat, M. Nguyen, and I. Altintas, “Big data provenance: Challenges, state of the art and opportunities,” in *2015 IEEE International Conference on Big Data (Big Data)*, 2015, pp. 2509–2516.

[26] G. O’Malley, “Click Fraud Costs Marketers $11B, IAB Issues Key Report,” *MediaPost*, Jan. 2014.

[27] R. Shields, “AppNexus CTO On The Fight Against Ad Fraud,” *Exch. Wire*, vol. October, no. 29, 2014.

[28] D. Lazer, R. Kennedy, G. King, and A. Vespignani, “The Parable of Google Flu: Traps in Big Data Analysis,” *Science (80-.).*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[29] P. Chen, B. Plale, and M. S. Aktas, “Temporal representation for mining scientific data provenance,” *Futur. Gener. Comput. Syst.*, vol. 36, pp. 363–378, 2014.

[30] W. Jansen and T. Grance, “NIST SP 800–144: Guidelines on Security and Privacy in Public Cloud Computing,” Dec. 2011.

[31] ETSI, “Cloud Standards Coordination. Final Report.,” 2013.

[32] DISA, “Department of Defense (DoD) Cloud Computing Security Requirements Guide (SRG),” Fort Meade, MD, 2015.

[33] CIO Council, “Recommendations for Standardized Implementation of Digital Privacy Controls,” Washington, DC, 2012.

[34] J. Draeger, “A roadmap to a unified treatment of safety and security,” in *10th IET System Safety and Cyber-Security Conference 2015*, 2015, pp. 1–6.

[35] M. Finnegan, “Boeing 787s to create half a terabyte of data per flight, says Virgin Atlantic,” *Comput. UK*, Mar. 2013.

[36] T. Larsen, “Cross-platform aviation analytics using big-data methods,” in *2013 Integrated Communications, Navigation and Surveillance Conference (ICNS)*, 2013, pp. 1–9.

[37] L. Piètre-Cambacédès and M. Bouissou, “Cross-fertilization between safety and security engineering,” *Reliab. Eng. Syst. Saf.*, vol. 110, pp. 110–126, Feb. 2013.

[38] J. Voas, “NIST SP 800-183: Networks of ‘Things,’” *NIST Special Publication 800-183*. 2016.

[39] K. Stouffer, J. Falco, and K. Scarfone, “Guide to Industrial Control Systems (ICS) Security (NIST

SP 800-82),” May 2015.

[40] International Electrotechnical Commission and ISA, *IEC 62443x: Industrial Automation and Control Systems Security*. International Electrotechnical Commission.

[41] P. K. Das, A. Joshi, and T. Finin, “Capturing policies for fine-grained access control on mobile devices,” in *Proceedings - 2016 IEEE 2nd International Conference on Collaboration and Internet Computing, IEEE CIC 2016*, 2017, pp. 54–63.

[42] K. Lenz and A. Oberweis, “Inter-organizational business process management with XML nets,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2472, pp. 243–263, 2003.

[43] I. Hickson, “HTML Microdata,” *W3C Work. Gr. Note 29*, pp. 1–29, 2013.

[44] I. Hickson, G. Kellogg, J. Tenisson, and I. Herman, “Microdata to RDF – Second Edition,” *W3C*, 2014. [Online]. Available: http://www.w3.org/TR/microdata-rdf/.

[45] R. Ross, M. McEvilley, and J. C. Oren, “NIST SP 800-160: Systems Security Engineering,” *NIST Special Publication*, Gaithersburg MD, p. 245, Sep-2016.

[46] Z. Khayyat *et al.*, “BigDansing: A System for Big Data Cleansing,” in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data SE - SIGMOD ’15*, 2015, pp. 1215–1230.

[47] L. A. Pachano, T. M. Khoshgoftaar, and R. Wald, “Survey of Data Cleansing and Monitoring for Large-Scale Battery Backup Installations,” in *2013 12th International Conference on Machine Learning and Applications*, 2013, vol. 2, pp. 478–484.

[48] M. Fazio and A. Puliafito, “Virtual Resource Management Based on Software Transactional Memory,” in *2011 First International Symposium on Network Cloud Computing and Applications*, 2011, pp. 1–8.

[49] A. Celesti, M. Fazio, and M. Villari, “SE CLEVER: A secure message oriented Middleware for Cloud federation,” in *Proceedings - International Symposium on Computers and Communications*, 2013, pp. 35–40.

[50] W. Jun, Z. Di, L. Meng, X. Fang, S. Hu-Lin, and Y. Shu-Feng, “Discussion of Society Fire-Fighting Safety Management Internet of Things Technology System,” in *2014 Fifth International Conference on Intelligent Systems Design and Engineering Applications*, 2014, pp. 422–425.

[51] K. Liu, Y. Yao, and D. Guo, “On Managing Geospatial Big-data in Emergency Management: Some Perspectives,” in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management SE - EM-GIS ’15*, 2015.

[52] X. Zhang and R. J. (editor), “A Survey of Digital Rights Management Technologies,” 2015. [Online]. Available: http://www.cse.wustl.edu/~jain/cse571-11/ftp/drm/. [Accessed: 09-Jan-2015].

[53] V. Bael, “European Union: ECJ Confirms that IP Addresses are Personal Data,” *Mondaq*, 2012. [Online]. Available: http://www.mondaq.com/x/162538/Copyright/ECJ+Confirms+That+IP+Addresses+Are+Personal+Data.

[54] Personal Correspondence, “Cloud homomorphic encryption service.” 2015.

[55] Pharma and European Federation of Pharmaceutical Industries and Associations, “Principles for Responsible Clinical Trial Data Sharing,” 2013.

[56] P. Wood, “How to tackle big data from a security point of view,” *ComputerWeekly.com*, 2013. [Online]. Available: http://www.computerweekly.com/feature/How-to-tackle-big-data-from-a-

security-point-of-view.

[57] B. Rossi, "Big security: big data and the end of SIEM," *Information Age*, 29-May-2014. [Online]. Available: http://www.information-age.com/big-security-big-data-and-end-siem-123458055/.

[58] D. Gunderson, "Drone patrol: Unmanned craft find key role in U.S. border security," *Minnesota Public Radio News*, Grand Forks, ND, 19-Feb-2015.

[59] Deputy Under Secretary of the Navy, "Naval security enterprise," *Nav. Secur. Enterp.*, vol. 2nd Quarte, p. 11, 2015.

[60] U.S. Department of Justice, "Guidance on Domestic Use of Unmanned Aircraft Systems." [Online]. Available: https://www.justice.gov/file/441266/download.

[61] Data Quality Campaign, "Roadmap to Safeguarding Student Data," 2015.

[62] J. Campbell, "Cuomo panel: State should cut ties with inBloom," *Albany Bureau, Iohud*, 2014.

[63] L. Fleisher, "Before Tougher State Tests, Officials Prepare Parents," *Wall Str. J.*, vol. April 15, 2013.

[64] R. D. Crick, P. Broadfoot, and G. Claxton, "Developing an Effective Lifelong Learning Inventory: the ELLI Project," *Assess. Educ. Princ. Policy Pract.*, vol. 11, no. 3, pp. 247–272, 2004.

[65] R. Ferguson, "Learning analytics: drivers, developments and challenges," *Int. J. Technol. Enhanc. Learn.*, vol. 4, no. 5/6, p. 304, 2012.

[66] D. Donston-Miller, "Common Core Meets Aging Education Technology," *InformationWeek*, vol. July 22, 2013.

[67] Civitas Learning, "About," 2016. [Online]. Available: https://www.civitaslearning.com/about/.

[68] D. Proud-Madruga, "Project Summary for Privacy, Access and Security Services (PASS) Healthcare Audit Services Conceptual Model," *Health Level Seven International*. 2016.

[69] M. Alam, S. Katsikas, O. Beltramello, and S. Hadjiefthymiades, "Augmented and virtual reality based monitoring and safety system: A prototype IoT platform," *J. Netw. Comput. Appl.*, vol. 89, pp. 109–119, 2017.

[70] M. StJohn-Green, R. Piggin, J. A. McDermid, and R. Oates, "Combined security and safety risk assessment #x2014; What needs to be done for ICS and the IoT," in *10th IET System Safety and Cyber-Security Conference 2015*, 2015, pp. 1–7.

[71] Kauffman_Foundation, "Welcome to EdWise - Education Data for Missouri." Kauffman Foundation, Kansas City, MO, Sep-2016.

[72] D. Boneh, A. Sahai, and B. Waters, "Functional Encryption: Definitions and Challenges," in *Theory of Cryptography: 8th Theory of Cryptography Conference, TCC 2011, Providence, RI, USA, March 28-30, 2011. Proceedings*, Y. Ishai, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 253–273.

[73] R. Chandramouli, M. Iorga, and S. Chokhani, "NIST IR 7956: Cryptographic key management issues & challenges in cloud services," 2013.

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

[75] Anonymous, "Summary of the Amazon S3 Service Disruption in the Northern Virginia (US-EAST-1) Region," *Amaz. Web Serv. Blog*, Mar. 2017.

[76] Association for Computing Machinery, "The 2012 ACM Computing Classification System." Association for Computing Machinery, 2012.

[77] NIST, “NIST SP 800-37: Guide for Applying the Risk Management Framework to Federal Information Systems,” *NIST Spec. Publ. 800-37*, vol. Rev 1, no. February, p. 93, 2010.

[78] S. Brooks, M. Garcia, N. Lefkovitz, S. Lightman, and E. Nadeau, “NIST IR 8062: An Introduction to Privacy Engineering and Risk Management in Federal Systems,” 2017.

[79] ISACA, “The Risk IT Framework,” 2009.

[80] NIST, “Framework for Improving Critical Infrastructure Cybersecurity,” 2014.

[81] OASIS, “SAML V2.0 Standard,” *SAML Wiki*, 2005. [Online]. Available: https://wiki.oasis-open.org/security/FrontPage#SAML_V2.0_Standard. [Accessed: 09-Jan-2015].

[82] J. J. Cebula and L. R. Young, “A Taxonomy of Operational Cyber Security Risks,” *Carnegie-Mellon Univ Pittsburgh Pa Softw. Eng. Inst*, no. December, pp. 1–47, 2010.

[83] H.-C. Kum and S. Ahalt, “Privacy-by-Design: Understanding Data Access Models for Secondary Data.,” *AMIA Jt. Summits Transl. Sci. Proc. AMIA Summit Transl. Sci.*, vol. 2013, pp. 126–30, Jan. 2013.

[84] J. Rawls, “Justice as Fairness: Political not Metaphysical,” *Philos. Public Aff.*, vol. 14, no. 3, pp. 223–251, 1985.

[85] ETSI, “Smart Cards: Secure channel between a UICC and an end-point terminal (Release 7),” 2007.

[86] U.S. Department of Health & Human Services, “New rule protects patient privacy, secures health information,” 17-Jan-2013.

[87] D. F. Sittig and H. Singh, “Legal, Ethical, and Financial Dilemmas in Electronic Health Record Adoption and Use,” *Pediatrics*, vol. 127, no. 4, pp. e1042–e1047, 2011.

[88] US-CERT, “About US-CERT,” 2015. [Online]. Available: https://www.us-cert.gov/about-us. [Accessed: 01-Jan-2015].

[89] U.S. Federal Trade Commission, “Protecting Your Child’s Privacy Online,” *Consumer Information*, Jul-2013. [Online]. Available: https://www.consumer.ftc.gov/articles/0031-protecting-your-childs-privacy-online.

[90] U.S. Department of Health & Human Services, “Health Information Privacy, Security, and your EHR,” *HealthIT.gov, Privacy and Security*, 13-Apr-2015. .

[91] U.S. Food and Drug Administration, “Medical Device Safety Network (MedSun),” *Medical Device Safety*, 08-May-2017. [Online]. Available: https://www.fda.gov/medicaldevices/safety/medsunmedicalproductsafetynetwork/default.htm.

[92] B. Mirkin, S. Nascimento, and L. M. Pereira, “Representing a computer science research organization on the ACM computing classification system,” in *CEUR Workshop Proceedings*, 2008, vol. 354, pp. 57–65.

[93] X. Lin, M. Zhang, H. Zhao, and J. Buzydlowski, “Multi-view of the ACM classification system,” in *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries - JCDL ’12*, 2012, p. 397.

[94] A. Miles and S. Bechhofer, “SKOS Simple Knowledge Organization System Reference,” *W3C Recommendation 18 August 2009*. pp. 1–40, 2009.

[95] L. Obrst, P. Chase, and R. Markeloff, “Developing an Ontology of the Cyber Security Domain,” in *Proceedings of the Seventh International Conference on Semantic Technologies for Intelligence, Defense, and Security*, 2012, pp. 49–56.

[96] D. Waltermire and B. A. Cheikes, "NIST IR8085: Forming Common Platform Enumeration (CPE) Names from Software Identification (SWID) Tags," *NIST Special Publication*, Gaithersburg, MD, Dec-2015.

[97] D. Inoue, M. Eto, K. Suzuki, M. Suzuki, and K. Nakao, "DAEDALUS-VIZ: Novel Real-time 3D Visualization for Darknet Monitoring-based Alert System," in *Proceedings of the Ninth International Symposium on Visualization for Cyber Security SE - VizSec '12*, 2012, pp. 72–79.

[98] A. Shabtai, D. Klimov, Y. Shahar, and Y. Elovici, "An intelligent, interactive tool for exploration and visualization of time-oriented security data," in *VizSEC '06: Proceedings of the 3rd international workshop on Visualization for computer security*, 2006, pp. 15–22.

[99] T. Takahashi, Y. Kadobayashi, and H. Fujiwara, "Ontological Approach Toward Cybersecurity in Cloud Computing," in *Proceedings of the 3rd International Conference on Security of Information and Networks SE - SIN '10*, 2010, pp. 100–109.

[100] G. Yee, "Visualization for privacy compliance," in *VizSEC '06: Proceedings of the 3rd international workshop on Visualization for computer security*, 2006, pp. 117–122.

[101] C. Brodie, C.-M. Karat, J. Karat, and J. Feng, "Usable Security and Privacy: A Case Study of Developing Privacy Management Tools," in *Proceedings of the 2005 symposium on Usable privacy and security - SOUPS '05*, 2005, pp. 35–43.

[102] W. Carey, J. Nilsson, and S. Mitchell, "Persistent security, privacy, and governance for healthcare information," in *Proceedings of the 2nd USENIX Conference on Health Security and Privacy*, 2011.

[103] P. Dunphy *et al.*, "Understanding the Experience-Centeredness of Privacy and Security Technologies," in *Proceedings of the 2014 workshop on New Security Paradigms Workshop - NSPW '14*, 2014, pp. 83–94.

[104] E. A. Oladimeji, L. Chung, H. T. Jung, and J. Kim, "Managing security and privacy in ubiquitous eHealth information interchange," in *Proceedings of the 5th International Confernece on Ubiquitous Information Management and Communication - ICUIMC '11*, 2011, p. 1.

[105] B. Obama, "National Strategy for Trusted Identities in Cyberspace," *The White House*, p. 25, 2011.

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

[107] J.-S. Li, Y.-F. Zhang, and Y. Tian, "Medical Big Data Analysis in Hospital Information System," in *Big Data on Real-World Applications*, 2016.

[108] O. Niakšu, "CRISP Data Mining Methodology Extension for Medical Domain," *Balt. J. Mod. Comput.*, vol. 3, no. 2, pp. 92–109, 2015.

[109] J. Schaffer, H. Tobias, D. Jones, and J. O. Donovan, "Getting the Message ? A Study of Explanation Interfaces for Microblog Data Analysis," *IUI 2015 Proc. 20th Int. Conf. Intell. User Interfaces*, pp. 345–356, 2015.

[110] Cloud Security Alliance, "CSA Big Data Security and Privacy Handbook," 2016.

[111] J. Loftus, A. May, N. P. Smart, and F. Vercauteren, "On CCA-secure somewhat homomorphic encryption," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7118 LNCS, pp. 55–72.

[112] C. Gentry, "A Fully Homomorphic Encryption Scheme," Stanford University, Stanford, CA, USA, 2009.

[113] D. Boneh, E.-J. Goh, and K. Nissim, “Evaluating 2-DNF Formulas on Ciphertexts,” in *Proceedings of the Second International Conference on Theory of Cryptography SE - TCC’05*, 2005, pp. 325–341.

[114] M. Van Dijk, C. Gentry, S. Halevi, and V. Vaikuntanathan, “Fully homomorphic encryption over the integers,” *Adv. Cryptology– EUROCRYPT ’10*, pp. 24–43, 2010.

[115] J.-S. Coron, A. Mandal, D. Naccache, and M. Tibouchi, “Fully Homomorphic Encryption over the Integers with Shorter Public Keys,” in *Advances in Cryptology -- CRYPTO 2011*, 2011, pp. 487–504.

[116] C. Gentry, S. Halevi, and N. P. Smart, “Fully homomorphic encryption with polylog overhead,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7237 LNCS, pp. 465–482.

[117] C. Gentry, S. Halevi, and N. P. Smart, “Homomorphic evaluation of the AES circuit,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7417 LNCS, pp. 850–867.

[118] M. Naehrig, K. Lauter, and V. Vaikuntanathan, “Can homomorphic encryption be practical?,” in *Proceedings of the 3rd ACM workshop on Cloud computing security workshop - CCSW ’11*, 2011, p. 113.

[119] D. Boneh and B. Waters, “Conjunctive, Subset, and Range Queries on Encrypted Data,” *TCC 2007 Theory Cryptogr.*, vol. 4392, pp. 535–554, 2007.

[120] D. Cash, S. Jarecki, C. Jutla, H. Krawczyk, M. C. Roşu, and M. Steiner, “Highly-scalable searchable symmetric encryption with support for Boolean queries,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8042 LNCS, no. Advances in Cryptology-CRYPTO 2013, PART 1, pp. 353–373.

[121] P. Datta, R. Dutta, and S. Mukhopadhyay, “Functional encryption for inner product with full function privacy,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, vol. 9614, pp. 164–195.

[122] C. Percival, “Cache missing for fun and profit,” *BSDCan 2005*, pp. 1–13, 2005.

[123] J. Seifert, Ç. Koç, and O. Aciiçmez, “Predicting Secret Keys Via Branch Prediction,” in *Ct-Rsa*, 2007, vol. 4377, no. October 2016, pp. 225–242.

[124] A. Shamir, “Identity-Based Cryptosystems and Signature Schemes,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1985, vol. 196 LNCS, pp. 47–53.

[125] D. Boneh and M. Franklin, “Identity-Based Encryption from the Weil Pairing,” *SIAM J. Comput.*, vol. 32, no. 3, pp. 586–615, 2003.

[126] B. Waters, “Dual system encryption: Realizing fully secure IBE and HIBE under simple assumptions,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5677 LNCS, pp. 619–636.

[127] J. Chen and H. Wee, “Fully, (almost) tightly secure IBE and dual system groups,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8043 LNCS, no. PART 2, pp. 435–460.

[128] C. S. Jutla and A. Roy, “Shorter Quasi-Adaptive NIZK Proofs for Linear Subspaces,” in *Part I of the Proceedings of the 19th International Conference on Advances in Cryptology - ASIACRYPT 2013 - Volume 8269*, 2013, pp. 1–20.

[129] A. Sahai and B. Waters, "Fuzzy Identity Based Encryption," *Eurocrypt '05*, pp. 457–473, 2005.

[130] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in *Proceedings of the 13th ACM conference on Computer and communications security - CCS '06*, 2006, p. 89.

[131] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-Policy Attribute-Based Encryption," in *Proceedings of the 2007 IEEE Symposium on Security and Privacy SE - SP '07*, 2007, pp. 321–334.

[132] B. Waters, "Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6571 LNCS, pp. 53–70.

[133] A. C. Yao, "Protocols for secure computations," in *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, 1982, pp. 160–164.

[134] J. Saia and M. Zamani, "Recent Results in Scalable Multi-Party Computation," *SOFSEM 2015 Theory Pract. Comput. Sci.*, no. 8939, pp. 24–44, 2015.

[135] M. Zamani, "A Multi-Party Computation Library," *GitHub*, 2015. [Online]. Available: https://github.com/mahdiz/mpclib.

[136] F. Schuster *et al.*, "VC3 : Trustworthy Data Analytics in the Cloud," Mar. 2015.

[137] PCI Security Standards Council, "The Prioritized Approach to Pursue PCI DSS Compliance," *PCI DSS Prioritized Approach PCI DSS 3.2*, 2016.

[138] E. Barker, "Recommendation for Key Management – Part 1: General (Revision 4), NIST Special Publication 800-57," Jan. 2016.

[139] R. Brown and J. Burrows, "FIPS PUB 140-2 Security Requirements For Cryptographic Modules," *Change*, vol. 46, no. 2, p. 69, 2001.

[140] NIST, "NIST Special Publication 800-39, Managing Information Security Risk Organization, Mission, and Information System View," 2011.

[141] T. Pasquier and D. Eyers, "Information Flow Audit for Transparency and Compliance in the Handling of Personal Data," in *2016 IEEE International Conference on Cloud Engineering Workshop (IC2EW)*, 2016, pp. 112–117.

[142] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb, "Leveraging Transparency," *IEEE Softw.*, vol. 30, no. 1, pp. 37–43, Jan. 2013.

[143] K. Benjamin, C. Cappelli, and G. Santos, "Organizational Transparency Maturity Assessment Method," in *Proceedings of the 18th Annual International Conference on Digital Government Research SE - dg.o '17*, 2017, pp. 477–484.

[144] E. Theodoridis, G. Mylonas, and I. Chatzigiannakis, "Developing an IoT Smart City framework," in *IISA 2013*, 2013, pp. 1–6.

[145] P. T. Grogan, K. Ho, A. Golkar, and O. L. de Weck, "Multi-Actor Value Modeling for Federated Systems," *IEEE Syst. J.*, vol. PP, no. 99, pp. 1–10, 2017.

[146] G. Ballard *et al.*, "How to Make Shared Risk and Reward Sustainable," *23rd Annu. Conf. Int. Gr. Lean Constr.*, 2015.

[147] D. M. Nicol, "Modeling and simulation in security evaluation," *Secur. Privacy, IEEE*, vol. 3, no. 5, pp. 71–74, 2005.

[148] V. Volovoi, "Simulation of maintenance processes in the Big Data era," in *2016 Winter Simulation*

Conference (WSC), 2016, pp. 1872–1883.

[149] R. G. Lang, Silva, and R. A. F. Romero, "Development of Distributed Control Architecture for Multi-robot Systems," in *2014 Joint Conference on Robotics: SBR-LARS Robotics Symposium and Robocontrol*, 2014, pp. 163–168.

[150] D. Dudenhoeffer, M. Permann, and E. Sussman, "General methodology 3: a parallel simulation framework for infrastructure modeling and analysis," in *WSC '02: Proceedings of the 34th conference on Winter simulation*, 2002, pp. 1971–1977.

[151] I. Paik, "Situation awareness based on big data analysis," in *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2016, vol. 2, pp. 911–916.

[152] J. Ryoo, R. Kazman, and P. Anand, "Architectural analysis for security," *IEEE Secur. Priv.*, vol. 13, no. 6, pp. 52–59, 2015.

[153] G. Lea, "Notes from YOW! 2014: Scott Shaw on 'Avoiding Speedbumps on the Road to Microservices.'" Graham Lea, p. 1, 02-Mar-2015.

[154] R. Dhall, "Performance Patterns in Microservices based Integrations," *Comput. Now*, 2016.

[155] G. Landers, A. Dayley, and J. Corriveau, "Magic Quadrant for Structured Data Archiving and Application Retirement," *Gartner.com*, 2016. [Online]. Available: https://www.gartner.com/doc/reprints?id=1-39B7753&ct=160613&st=sb.

[156] K. Ruan and J. Carthy, "Cloud Forensic Maturity Model," in *Digital Forensics and Cyber Crime*, M. Rogers and K. C. Seigfried-Spellar, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 22–41.

[157] P. Franková, M. Drahošová, and P. Balco, "Agile Project Management Approach and its Use in Big Data Management," *Procedia Comput. Sci.*, vol. 83, pp. 576–583, 2016.

[158] E. Burger, *Flexible views for view-based model-driven development*. Karlsruhe. Deutschland: KIT Scientific Publishing, 2014.

[159] K. Kent and M. Souppaya, "Guide to Computer Security Log Management," *Natl. Inst. Stand. Technol.*, pp. 1–72, 2006.

[160] K. Kent, S. Chevalier, T. Grance, and H. Dang, "NIST SP 800-86: Guide to Integrating Forensic Techniques Into Incident Response," NIST, Gaithersburg MD OR - NIST, Sep. 2006.

[161] S. Zareian, M. Fokaefs, H. Khazaei, M. Litoiu, and X. Zhang, "A Big Data Framework for Cloud Monitoring," in *Proceedings of the 2Nd International Workshop on BIG Data Software Engineering SE - BIGDSE '16*, 2016, pp. 58–64.

[162] E. Chabrow, "NIST Plans Cybersecurity Framework Update - GovInfoSecurity," *Government Information Security*, 2016. [Online]. Available: http://www.govinfosecurity.com/interviews/nist-considers-cybersecurity-framework-update-i-3199#.V1jIbRyMY7E.twitter.

[163] DHS, "Critical Infrastructure Cyber Community or C^3 Voluntary Program," *US-CERT*, 2014. [Online]. Available: https://www.us-cert.gov/ccubedvp. [Accessed: 14-Aug-2016].

[164] E. Gonzalez, "SENC Project: SABSA Enhanced NIST Cybersecurity Framework," *SABSA*, 2015. [Online]. Available: http://www.sabsa.org/node/176. [Accessed: 15-Aug-2015].

[165] S. Zurier, "6 Things To Know For Securing Amazon Web Services," *Dark Read.*, Aug. 2016.

[166] A. Textor, R. Kroeger, and K. Geihs, "Semantic models for bridging domains in automated IT management: Lessons learned," in *2017 International Conference on Networked Systems (NetSys)*, 2017, pp. 1–8.

[167] E. G. Aydal, R. F. Paige, H. Chivers, and P. J. Brooke, “Security Planning and Refactoring in Extreme Programming,” in *Extreme Programming and Agile Processes in Software Engineering: 7th International Conference, XP 2006, Oulu, Finland, June 17-22, 2006. Proceedings*, P. Abrahamsson, M. Marchesi, and G. Succi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 154–163.

[168] M. Iqbal and M. Rizwan, “Application of 80/20 rule in Software Engineering Waterfall Model,” in *Information and Communication Technologies, 2009. ICICT ’09. International Conference on*, 2009.

[169] B. Boehm, J. A. Lane, S. Koolmanojwong, and R. Turner, *The Incremental Commitment Spiral Model: Principles and Practices for Successful Systems and Software*, 1st ed. Addison-Wesley Professional, 2014.

[170] N. MacDonald and I. Head, “DevSecOps: How to Seamlessly Integrate Security Into DevOps SE - G00315283,” Gartner Group, Stamford CT OR - Gartner Group, Sep. 2016.

[171] A. Cockroft, “Evolution of Microservices (video presentation),” *ACM*. Association for Computing Machinery, New York, NY, 20-Jul-2016.

[172] J. Roche, “Adopting DevOps practices in quality assurance,” *Commun. ACM*, vol. 56, no. 11, pp. 38–43, 2013.

[173] Tom Nolle, “Infrastructure as code complicates hybrid, multiple cloud management (Part 2 of 2),” *Search Cloud Computing*, 2016.

[174] J. Steer and A. Popli, “Building secure business applications at Microsoft,” *Inf. Secur. Tech. Rep.*, vol. 13, no. 2, pp. 105–110, May 2008.

[175] Tom Nolle, “Separating DevOps from the future-driven cloud orchestration,” *Search Cloud Computing*, 2016. [Online]. Available: http://searchcloudcomputing.techtarget.com/tip/Separating-DevOps-from-the-future-driven-cloud-orchestration.

[176] R. Qasha, J. Cala, and P. Watson, “Towards Automated Workflow Deployment in the Cloud Using TOSCA,” in *Proceedings - 2015 IEEE 8th International Conference on Cloud Computing, CLOUD 2015*, 2015, pp. 1037–1040.

[177] P. Chambakara, “API-First Design: Dawn Of New Era In App Development,” *Digital Doughnut*. 2015.

[178] G. Chen and Y. Luo, “A BIM and ontology-based intelligent application framework,” in *2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 2016, pp. 494–497.

[179] C. Atkinson, D. Stoll, and P. Bostan, “Orthographic software modeling: A practical approach to view-based development,” in *Communications in Computer and Information Science*, 2010, vol. 69 CCIS, pp. 206–219.

[180] A. Barth, A. Datta, J. Mitchell, and H. Nissenbaum, “Privacy and Contextual Integrity: Framework and Applications,” in *Proceedings of the 2006 IEEE Symposium on Security and Privacy SE - SP ’06*, 2006, pp. 184–198.

[181] P. Lam, J. Mitchell, A. Scedrov, S. Sundaram, and F. Wang, “Declarative privacy policy: finite models and attribute-based encryption,” in *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium SE - IHI ’12*, 2012, pp. 323–332.

[182] J. Wilson, “THE CLOUD, REGULATIONS, AND PII,” *Iconic*, Jan-2016.

[183] R. Nelson, “Big data analytics becomes strategic test tool,” *Eval. Eng.*, Dec. 2015.

[184] A. Karmel, R. Chandramouli, and M. Iorga, “DRAFT Special Publication 800-180, NIST Definition of Microservices, Application Containers and System Virtual Machines,” *NIST Spec. Publ. 800-180*, vol. 800180, 2016.

[185] S. Newman, “Building microservices : designing fine-grained systems.” O’Reilly Media, Sebastopol CA, 2015.

[186] A. Versteden, E. Pauwels, and A. Papantoniou, “An Ecosystem of User-facing Microservices supported by Semantic Models,” *USEWOD-PROFILES@ESWC*, vol. 1362, pp. 12–21, 2015.

[187] American National Standards Institute, “ANSI INCITS 359-2004 Role Based Access Control Information Technology Industry Council,” 2004.

[188] D. R. Kuhn, E. J. Coyne, and T. R. Weil, “Adding attributes to role-based access control,” *Computer (Long. Beach. Calif).*, vol. 43, no. 6, pp. 79–81, 2010.

[189] D. F. Ferraiolo, G. J. Ahn, R. Chandramouli, and S. I. Gavrila, “The role control center: Features and case studies,” in *Proceedings of ACM Symposium on Access Control Models and Technologies (SACMAT 2002)*, 2003, pp. 12–20.

[190] E. Bertino and B. Catania, “GEO-RBAC: a spatially aware RBAC,” *Proc. tenth ACM Symp. Access Control Model. Technol.*, pp. 29–37, 2005.

[191] R. Ferrini and E. Bertino, “Supporting RBAC with XACML+OWL,” in *Proceedings of the 14th ACM symposium on Access control models and technologies SE - SACMAT ’09*, 2009, pp. 145–154.

[192] Y. Sun, X. Meng, S. Liu, and P. Pan, “An approach for flexible RBAC workflow system,” in *Computer Supported Cooperative Work in Design, 2005. Proceedings of the Ninth International Conference on*, 2005, vol. 1, p. 524–529 Vol. 1.

[193] M. Underwood, “Big Data Complex Event Processing for Internet of Things Provenance: Benefits for Audit, Forensics and Safety,” in *Cyber-Assurance for the Internet of Things*, T. Brooks, Ed. Hoboken NJ: Wiley, 2016.

[194] M. Underwood *et al.*, “Internet of things: Toward smart networked systems and societies,” *Appl. Ontol.*, vol. 10, no. 3–4, pp. 355–365, Sep. 2015.

[195] J. Turnbull, *The Art of Monitoring*. New York, NY: James Turnbull, 2016.

[196] T. Stewart, “Human after all,” *IoSH Mag.*, Jun. 2016.

[197] M. Wilkinson *et al.*, “The FAIR Guiding Principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, p. 160018, Mar. 2016.

[198] C. J. Hoofnagle, “How the Fair Credit Reporting Act Regulates Big Data,” pp. 1–6, Sep. 2013.

[199] R. Chandramouli, “NIST SP 800-125B: Secure Virtual Network Configuration for Virtual Machine (VM) Protection,” *Nist Spec. Publ.*, vol. 800, no. 125B, p. 23, 2016.

[200] PCI Security Standards Council, “PCI DSS Virtualization Guidelines,” 2011.

[201] B. Keshavamurthy and M. Ashraf, “Conceptual design of proactive SONs based on the Big Data framework for 5G cellular networks: A novel Machine Learning perspective facilitating a shift in the SON paradigm,” in *2016 International Conference System Modeling Advancement in Research Trends (SMART)*, 2016, pp. 298–304.

[202] D. Ardagna, L. Baresi, S. Comai, M. Comuzzi, and B. Pernici, “A service-based framework for flexible business processes,” *IEEE Softw.*, vol. 28, no. 2, pp. 61–67, 2011.

[203] Microsoft, “Deploying Windows Rights Management Services at Microsoft,” 2013. [Online].

Available: http://technet.microsoft.com/en-us/library/dd277323.aspx.

[204] The Nielsen Company, “Consumer Panel and Retail Measurement,” 2015. [Online]. Available: www.nielsen.com/us/en/nielsen-solutions/nielsen-measurement/nielsen-retail-measurement.html.

[205] SAFE-BioPharma Association, “Welcome to SAFE-BioPharma.” [Online]. Available: http://www.safe-biopharma.org/.

[206] Microsoft, “How to set event log security locally or by using Group Policy in Windows Server 2003,” 07-Jan-2017. [Online]. Available: http://support.microsoft.com/kb/323076.

[207] DefenseSystems, “UAV video encryption remains unfinished job,” 31-Oct-2012. [Online]. Available: http://defensesystems.com/articles/2012/10/31/agg-drone-video-encryption-lags.aspx.

[208] Department of Defense Memorandum from DoD CIO, “Department of Defense Cloud Computing Strategy,” Jul. 2012.

[209] A. Sanna and F. Lamberti, “Advances in target detection and tracking in forward-looking infrared (FLIR) imagery,” *Sensors (Switzerland)*, vol. 14, no. 11, pp. 20297–20303, 2014.

[210] K. A. G. Fisher *et al.*, “Quantum computing on encrypted data,” *Nat. Commun.*, vol. 5, 2014.

[211] J. Cartledge, “US Lawmakers Pledge to Close Air Cargo Security ‘Loophole,’” *Post and Parcel*, 01-Nov-2010. [Online]. Available: http://postandparcel.info/35115/news/us-lawmakers-pledge-to-close-air-cargo-security-“loophole”/.

[212] S. Captain, “With Mapbox Deal, IBM Watson Will Learn A Lot More About Where Things Are Happening | Fast Company | Business + Innovation,” *Fast Company*, 2016. [Online]. Available: http://www.fastcompany.com/3062635/with-mapbox-deal-ibm-watson-will-know-where-things-are-happening.

[213] R. Chandramouli, “Secure Virtual Network Configuration for Virtual Machine (VM) Protection (NIST SP 800-125B),” 2016.