**NIST Special Publication 1500-4**

**DRAFT: NIST Big Data Interoperability Framework:**

**Volume 4, Security and Privacy**

NIST Big Data Public Working Group

Security and Privacy Subgroup

DRAFT Version 2

May 30, 2017

[http://dx.doi.org/----](http://dx.doi.org/10.6028/NIST.SP.1500-4)



NIST Special Publication 1500-4

Information Technology Laboratory

**DRAFT: NIST Big Data Interoperability Framework:**

**Volume 4, Security and Privacy**

**Draft Version 2**

NIST Big Data Public Working Group (NBD-PWG)

Security and Privacy Subgroup

National Institute of Standards and Technology

Gaithersburg, MD 20899

[http://dx.doi.org/](http://dx.doi.org/10.6028/NIST.SP.1500-4)----

May 2017



U. S. Department of Commerce

*Wilbur L. Ross, Jr., Secretary*

National Institute of Standards and Technology

*Dr. Kent Rochford, Acting Under Secretary of Commerce for Standards and Technology*

*and Acting NIST Director*

**National Institute of Standards and Technology (NIST) Special Publication 1500-4**

113 pages (May 30, 2017)

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

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

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. All NIST 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](mailto:SP1500comments@nist.gov)

Request for Contributions

The NIST Big Data Public Working Group (NBD-PWG) requests contributions to this draft version 2 of the *NIST Big Data Interoperability Framework Volume 4, Security and Privacy.* All contributions are welcome, especially comments or additional content for the current draft.

The NBD-PWG is actively working to complete version 2 of the set of NBDIFdocuments. The goals of version 2 are to enhance the version 1 content, define general interfaces between the NIST Big Data Reference Architecture (NBDRA) components by aggregating low-level interactions into high-level general interfaces, and demonstrate how the NBDRA can be used.

To contribute to this document, please follow the steps below as soon as possible but no later than May 26, 2017.

1. Register as a user of the NIST Big Data Portal (https://bigdatawg.nist.gov/newuser.php)
2. Record comments and/or additional content in one of the following methods:
   1. **TRACK CHANGES**: make edits to and comments on the text directly into this Word document using track changes
   2. **COMMENT TEMPLATE**: capture specific edits using the Comment Template (<http://bigdatawg.nist.gov/_uploadfiles/SP1500-1-to-7_comment_template.docx>), which includes space for Section number, page number, comment, and text edits
3. Submit the edited file from either method above to SP1500comments@nist.gov with the volume number in the subject line (e.g., Edits for Volume 1).
4. Attend the weekly virtual meetings on Tuesdays for possible presentation and discussion of your submission. Virtual meeting logistics can be found at https://bigdatawg.nist.gov/program.php

Please be as specific as possible in any comments or edits to the text. Specific edits include, but are not limited to, changes in the current text, additional text further explaining a topic or explaining a new topic, additional references, or comments about the text, topics, or document organization.

The comments and additional content will be reviewed by the subgroup co-chair responsible for the volume in question. Comments and additional content may be presented and discussed by the NBD-PWG during the weekly virtual meetings on Tuesday.

Three versions are planned for the NBDIF set of documents, with Versions 2 and 3 building on the first. Further explanation of the three planned versions and the information contained therein is included in Section 1.5 of each NBDIF document.

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

Big Data professionals are always welcome to join the NBD-PWG to help craft the work contained in the volumes of the NBDIF. Additional information about the NBD-PWG can be found at <http://bigdatawg.nist.gov>. Information about the weekly virtual meetings on Tuesday can be found at https://bigdatawg.nist.gov/program.php.

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* series of volumes. This volume, Volume 4, contains an exploration of security and privacy topics with respect to Big Data. This volume considers new aspects of security and privacy with respect to Big Data, reviews security and privacy use cases, proposes security and privacy taxonomies, presents details of the Security and Privacy Fabric of the NIST Big Data Reference Architecture (NBDRA), and begins mapping the security and privacy use cases to the NBDRA.

Keywords

Big Data 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, co-chaired by Wo Chang of the NIST ITL, Robert Marcus of ET-Strategies, and Chaitanya Baru, University of California San Diego Supercomputer Center. Subgroup co-chairs are Nancy Grady (SAIC), Geoffrey Fox (University of Indiana), Arnab Roy (Fujitsu), Mark Underwood (Krypton Brothers), David Boyd (InCadence Corp), and Russell Reinsch (Center for Government Interoperability).

The document contains input from members of the NBD-PWG Security and Privacy Subgroup, led by Arnab Roy (Fujitsu) and Mark Underwood (Krypton Brothers); and the Reference Architecture Subgroup, led by David Boyd (InCadence Corp).

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

NIST would like to acknowledge specific contributions[[1]](#footnote-1) to this volume by the following NBD-PWG members:

A list of contributors to version 2 of this volume will be added here.

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

Table of Contents

[Executive Summary xiv](#_Toc483319006)

[1 INTRODUCTION 1](#_Toc483319007)

[1.1 Background 1](#_Toc483319008)

[1.2 Scope And Objectives Of The Security And Privacy Subgroup 2](#_Toc483319009)

[1.3 Report Scope 3](#_Toc483319010)

[1.4 Report Production 4](#_Toc483319011)

[1.5 Report Structure 4](#_Toc483319012)

[1.6 Future Work On This Volume 6](#_Toc483319013)

[2 BIG DATA SECURITY AND PRIVACY 7](#_Toc483319014)

[2.1 What is Different about Big Data Security and Privacy 7](#_Toc483319015)

[2.2 Overview 8](#_Toc483319016)

[2.3 Security And Privacy Impacts On Big Data Characteristics 10](#_Toc483319017)

[2.3.1 Variety 10](#_Toc483319018)

[2.3.2 Volume 11](#_Toc483319019)

[2.3.3 Velocity 11](#_Toc483319020)

[2.3.4 Veracity 11](#_Toc483319021)

[2.3.5 Volatility 12](#_Toc483319022)

[2.4 Effects of Emerging Technology on Big Data Security and Privacy 12](#_Toc483319023)

[2.4.1 Cloud Computing 13](#_Toc483319024)

[2.4.2 Big Data Security Quilt 14](#_Toc483319025)

[2.4.3 Big Data Security Safety Levels 14](#_Toc483319026)

[2.4.4 Internet of Things and CPS 14](#_Toc483319027)

[2.4.5 Mobile Devices and Big Data 14](#_Toc483319028)

[2.4.6 Integration of People and Organizations 15](#_Toc483319029)

[2.4.7 System Communicator 15](#_Toc483319030)

[2.4.8 Ethical Design 15](#_Toc483319031)

[2.4.8.1 Self-Cleansing Systems 15](#_Toc483319032)

[2.4.8.2 The Toxic Data Model 15](#_Toc483319033)

[2.4.8.3 Relation to Systems Management 16](#_Toc483319034)

[2.4.8.4 Big Data Safety Annotation 16](#_Toc483319035)

[2.4.8.5 Big Data Trust and Federation 16](#_Toc483319036)

[2.4.8.6 Orchestration in Weak Federation Scenarios 16](#_Toc483319037)

[2.4.8.7 Consent and the Glass-breaking Scenario 16](#_Toc483319038)

[2.5 Security and Privacy Methodology with Respect to Big Data 16](#_Toc483319039)

[2.5.1 Why is this relevant for big data? 17](#_Toc483319040)

[3 EXAMPLE USE CASES FOR SECURITY AND PRIVACY 18](#_Toc483319041)

[3.1 Retail/Marketing 18](#_Toc483319042)

[3.1.1 Consumer Digital Media Usage 18](#_Toc483319043)

[3.1.2 Nielsen Homescan: Project Apollo 19](#_Toc483319044)

[3.1.3 Web Traffic Analytics 20](#_Toc483319045)

[3.2 Healthcare 20](#_Toc483319046)

[3.2.1 Health Information Exchange 20](#_Toc483319047)

[3.2.2 Genetic Privacy 21](#_Toc483319048)

[3.2.3 Pharma Clinical Trial Data Sharing23 22](#_Toc483319049)

[3.3 Cybersecurity 22](#_Toc483319050)

[3.3.1 Network Protection 22](#_Toc483319051)

[3.4 Government 23](#_Toc483319052)

[3.4.1 Unmanned Vehicle Sensor Data 23](#_Toc483319053)

[3.4.2 Education: Common Core Student Performance Reporting 24](#_Toc483319054)

[3.5 Industrial: Aviation 25](#_Toc483319055)

[3.5.1 Sensor Data Storage And Analytics 25](#_Toc483319056)

[3.6 Transportation 25](#_Toc483319057)

[3.6.1 Cargo Shipping 25](#_Toc483319058)

[3.7 Major Use Case: Sec Consolidated Audit Trail 26](#_Toc483319059)

[3.8 Major Use Case: Iot Device Management 26](#_Toc483319060)

[3.8.1 Smart Home IoT 26](#_Toc483319061)

[3.9 Major Use Case: Omg Data Residency Initiative 26](#_Toc483319062)

[3.9.1 Minor Use Case: Tbd 27](#_Toc483319063)

[3.9.2 Use Case: Emergency management data (XChangeCore interoperability standard). 27](#_Toc483319064)

[3.10 Major Use Case: Health Care Consent Flow 27](#_Toc483319065)

[3.11 Major Use Case: “Heart Use Case: Alice Selectively Shares Health-Related Data With Physicians And Others” 27](#_Toc483319066)

[3.12 Major Use Case Blockchain for Fintech (Arnab) 27](#_Toc483319067)

[3.12.1 Minor Use Case – In-Stream Pii 27](#_Toc483319068)

[3.13 Major Use Case—Statewide Education Data Portal 27](#_Toc483319069)

[4 TAXONOMY OF SECURITY AND PRIVACY TOPICS 29](#_Toc483319070)

[4.1 Conceptual Taxonomy of Security and Privacy Topics 29](#_Toc483319071)

[4.1.1 Data Confidentiality 30](#_Toc483319072)

[4.1.2 Provenance 31](#_Toc483319073)

[4.1.3 System Health 31](#_Toc483319074)

[4.1.4 Public Policy, Social and Cross-Organizational Topics 32](#_Toc483319075)

[4.2 Operational Taxonomy of Security and Privacy Topics 32](#_Toc483319076)

[4.2.1 Device and Application Registration 33](#_Toc483319077)

[4.2.2 Identity and Access Management 34](#_Toc483319078)

[4.2.3 Data Governance 34](#_Toc483319079)

[4.2.3.1 Compliance, Governance and Management as Code 35](#_Toc483319080)

[4.2.4 Infrastructure Management 35](#_Toc483319081)

[4.2.5 Risk and Accountability 36](#_Toc483319082)

[4.3 Roles Related at Security and Privacy Topics 37](#_Toc483319083)

[4.3.1 Infrastructure Management 37](#_Toc483319084)

[4.3.2 Governance, Risk Management, and Compliance 37](#_Toc483319085)

[4.3.3 Information Worker 38](#_Toc483319086)

[4.4 Relation of Roles to the Security and Privacy Conceptual Taxonomy 38](#_Toc483319087)

[4.4.1 Data Confidentiality 39](#_Toc483319088)

[4.4.2 Provenance 39](#_Toc483319089)

[4.4.3 System Health Management 39](#_Toc483319090)

[4.4.4 Public Policy, Social, and Cross-Organizational Topics 40](#_Toc483319091)

[4.5 Additional Taxonomy Topics 40](#_Toc483319092)

[4.5.1 Provisioning, Metering, And Billing 41](#_Toc483319093)

[4.5.2 Data Syndication 41](#_Toc483319094)

[4.5.3 ACM Taxonomy 41](#_Toc483319095)

[4.6 Why Security Ontologies Matter For Big Data 41](#_Toc483319096)

[5 BIG DATA REFERENCE ARCHITECTURE AND SECURITY AND PRIVACY FABRIC 43](#_Toc483319097)

[5.1 Security and Privacy Requirements 43](#_Toc483319098)

[5.2 NIST Big Data Reference Architecture 43](#_Toc483319099)

[5.3 Relation Of The Big Data Security Operational Taxonomy To The NBDRA 45](#_Toc483319100)

[5.4 Mapping Security and Privacy Use Cases to the NBDRA 46](#_Toc483319101)

[5.5 Security and Privacy Fabric in the NBDRA 46](#_Toc483319102)

[5.6 Security and Privacy Fabric Principles 48](#_Toc483319103)

[5.6.1 Related Fabric Concepts 48](#_Toc483319104)

[5.7 Security and Privacy Approaches in Analytics 48](#_Toc483319105)

[5.7.1 CRISP-DM Interop 48](#_Toc483319106)

[5.8 Cryptographic Technologies for Data Transformations 49](#_Toc483319107)

[5.8.1 Classification 49](#_Toc483319108)

[5.8.2 Homomorphic Encryption 50](#_Toc483319109)

[5.8.3 Functional Encryption 51](#_Toc483319110)

[5.8.4 Access Control Policy-Based Encryption 51](#_Toc483319111)

[5.8.5 Secure Multi-Party Computations 52](#_Toc483319112)

[5.8.6 Blockchain 53](#_Toc483319113)

[5.8.7 Hardware Support for Secure Computations 53](#_Toc483319114)

[5.9 Risk Management 54](#_Toc483319115)

[5.9.1 PII as Requiring Toxic Substance Handling 54](#_Toc483319116)

[5.9.2 Consent Withdrawal Scenarios 54](#_Toc483319117)

[5.9.3 Transparency Portal Scenarios 54](#_Toc483319118)

[5.9.4 Cross-organizational Risk Management 54](#_Toc483319119)

[5.9.5 Algorithm-Driven Issues 54](#_Toc483319120)

[5.9.6 Big Data Forensics and Operational AAR 55](#_Toc483319121)

[5.10 Big Data Security Modeling and Simulation (ModSim) 55](#_Toc483319122)

[5.10.1 Safety Systems Modeling 55](#_Toc483319123)

[5.11 Security and Privacy Management Phases 55](#_Toc483319124)

[5.11.1.1 Modifications for Agile Methodologies 56](#_Toc483319125)

[6 Domain-Specific Security 57](#_Toc483319126)

[6.1 Consent Management: Domain-Specific Big Data Security and Privacy 57](#_Toc483319127)

[6.1.1 Consent Management in Health Care 57](#_Toc483319128)

[6.1.1.1 Relation to smart contracts 57](#_Toc483319129)

[6.2 Smart Building Domain Security 57](#_Toc483319130)

[7 Provenance 58](#_Toc483319131)

[*7.1* *IoT Provenance* 58](#_Toc483319132)

[7.1.1 Traceability 58](#_Toc483319133)

[7.1.3 Possible Roles for SnP Ontologies 58](#_Toc483319134)

[7.1.4 Domain-specific Provenance 58](#_Toc483319135)

[8 Audit and Configuration Management 59](#_Toc483319136)

[8.1.1 Packet-Level Traceability / Reproduceability 59](#_Toc483319137)

[8.1.2 Audit 59](#_Toc483319138)

[8.1.3 Big Data Audit and Monitoring 59](#_Toc483319139)

[9 Workflow Models 60](#_Toc483319140)

[*9.1* *Baseline Levels* 60](#_Toc483319141)

[10 Standards, Best Practices and Gaps 61](#_Toc483319142)

[10.1 NIST Cybersecurity Framework 61](#_Toc483319143)

[10.2 SABSA and Zachman Framework 61](#_Toc483319144)

[10.3 Configuration Management for Big Data 61](#_Toc483319145)

[10.3.1 Lineage Provenance 61](#_Toc483319146)

[10.3.2 Dependency Models 61](#_Toc483319147)

[10.4 Encryption Standards 61](#_Toc483319148)

[10.4.1 Blockchain and Extensions 61](#_Toc483319149)

[10.5 Text Introducing Third Party Standards (Temporary) 62](#_Toc483319150)

[10.6 Big Data SDLC Standards and Guidelines 62](#_Toc483319151)

[10.6.1 Big Data Security in DevOps 62](#_Toc483319152)

[10.6.1.1 Application Lifecycle Management 62](#_Toc483319153)

[10.6.1.2 Security and Privacy Events in Application Release Management 62](#_Toc483319154)

[10.6.1.3 Orchestration 63](#_Toc483319155)

[10.6.1.4 API-First 63](#_Toc483319156)

[10.6.1.5 Microservices 63](#_Toc483319157)

[10.6.1.6 Software Security and Reliability in DevOps 63](#_Toc483319158)

[10.6.2 Model Driven Development 63](#_Toc483319159)

[10.6.2.1 Add SI discussion [] 63](#_Toc483319160)

[10.6.2.2 Add Smart Building Examples [] 63](#_Toc483319161)

[10.6.2.3 Metamodel Processes in Support of BD SnP 63](#_Toc483319162)

[10.6.2.4 Cite security ontology work @ Florida 64](#_Toc483319163)

[10.6.2.5 Cite work on Authorization Languages and Contextual Integrity 64](#_Toc483319164)

[10.6.3 Other Standards Through a Big Data Lens 64](#_Toc483319165)

[10.6.3.1 ISO 21827:2008 and SSE-CMM 64](#_Toc483319166)

[10.6.3.2 ISO 12207 and ISO 15504 64](#_Toc483319167)

[10.6.3.3 Process Specifications 64](#_Toc483319168)

[10.6.3.4 ISO 27018 64](#_Toc483319169)

[10.6.4 SnP Quilts for Specific SDLC Methodologies 65](#_Toc483319170)

[10.6.5 Big Data Test Engineering 65](#_Toc483319171)

[10.6.6 API-First and Microservices 65](#_Toc483319172)

[10.6.7 Application Security for Big Data 66](#_Toc483319173)

[10.6.7.1 RBAC, ABAC and Workflow 66](#_Toc483319174)

[10.6.7.2 ‘Least Exposure’ Big Data Practices 66](#_Toc483319175)

[10.6.7.3 Logging 67](#_Toc483319176)

[10.6.7.4 Ethics and Privacy by Design 67](#_Toc483319177)

[10.7 Big Data Governance 67](#_Toc483319178)

[10.7.1 Apache Atlas 67](#_Toc483319179)

[10.7.2 GSA DevOps Open Compliance 67](#_Toc483319180)

[10.8 Infrastructure Management 67](#_Toc483319181)

[10.8.1 Infrastructure as Code 67](#_Toc483319182)

[10.8.2 Particular Issues with Hybrid and Private Cloud 68](#_Toc483319183)

[10.8.3 Relevance to NIST Critical Infrastructure 68](#_Toc483319184)

[10.9 Emerging Technologies 68](#_Toc483319185)

[10.9.1 Blockchain 68](#_Toc483319186)

[10.9.2 DevOps Automation 69](#_Toc483319187)

[10.9.2.1 Application Release Automation 69](#_Toc483319188)

[10.9.3 Network Security for Big Data 69](#_Toc483319189)

[10.9.3.1 Virtual Machines and SDN 69](#_Toc483319190)

[10.9.3.2 Architecture Standards for IoT 69](#_Toc483319191)

[10.9.4 Machine Learning, AI and Analytics for Big Data Security and Privacy 69](#_Toc483319192)

[10.9.4.1 Overview of emerging technologies 69](#_Toc483319193)

[10.9.4.2 Risk / opportunity areas for enterprises 69](#_Toc483319194)

[10.9.4.3 Risk / opportunity areas for consumers 69](#_Toc483319195)

[10.9.4.4 Risk / opportunities for government 70](#_Toc483319196)

[11 Conclusions 71](#_Toc483319197)

[Appendix A: Mapping Use Cases to NBDRA a-1](#_Toc483319198)

[A.1 Retail/Marketing a-1](#_Toc483319199)

[A.1.1 Consumer Digital Media Use A-1](#_Toc483319200)

[A.1.2 Nielsen Homescan: Project Apollo A-2](#_Toc483319201)

[A.1.3 Web Traffic Analytics A-3](#_Toc483319202)

[A.2 Healthcare A-4](#_Toc483319203)

[A.2.1 Health Information Exchange A-4](#_Toc483319204)

[A.2.2 Genetic Privacy A-6](#_Toc483319205)

[A.2.3 Pharmaceutical Clinical Trial Data Sharing A-6](#_Toc483319206)

[A.3 Cybersecurity A-7](#_Toc483319207)

[A.3.1 Network Protection A-7](#_Toc483319208)

[A.4 Government A-8](#_Toc483319209)

[A.4.1 Unmanned Vehicle Sensor Data A-8](#_Toc483319210)

[A.4.2 Education: Common Core Student Performance Reporting A-9](#_Toc483319211)

[A.5 Industrial: Aviation A-10](#_Toc483319212)

[A.5.1 Sensor Data Storage and Analytics A-10](#_Toc483319213)

[A.6 Transportation A-10](#_Toc483319214)

[A.6.1 Cargo Shipping A-10](#_Toc483319215)

[A.7 New Use Cases A-11](#_Toc483319216)

[A.7.1 Major Use Case : SEC Consolidated Audit Trail A-12](#_Toc483319217)

[A.7.2 Major Use Case: IoT Device Management A-12](#_Toc483319218)

[A.7.3 Major Use Case: OMG Data Residency initiative A-12](#_Toc483319219)

[A.7.4 Minor Use Case: TBD A-12](#_Toc483319220)

[A.7.5 Use Case: Emergency management data (XChangeCore interoperability standard ). A-12](#_Toc483319221)

[A.7.6 Major Use Case: Health care consent flow A-12](#_Toc483319222)

[A.7.7 Major Use Case: “HEART Use Case: Alice Selectively Shares Health-Related Data with Physicians and Others” A-12](#_Toc483319223)

[A.7.8 Major Use Case Blockchain for FinTech (Arnab) A-12](#_Toc483319224)

[A.7.9 Minor Use Case – In-stream PII A-12](#_Toc483319225)

[A.7.10 Major Use Case – Statewide Education Data Portal A-12](#_Toc483319226)

[Appendix B: Internal Security Considerations within Cloud Ecosystems B-1](#_Toc483319227)

[Appendix C: Big Data Actors and Roles: Adaptation to Big Data Scenarios C-1](#_Toc483319228)

[Appendix D: Acronyms D-1](#_Toc483319229)

[Appendix E: References E-1](#_Toc483319230)

Figures

[Figure 1: High Level CAT Requirements 26](#_Toc483319231)

[Figure 2: EdWise Figure 28](#_Toc483319232)

[Figure 3: Security and Privacy Conceptual Taxonomy 30](#_Toc483319233)

[Figure 4: Security and Privacy Operational Taxonomy 33](#_Toc483319234)

[Figure 5: NIST Big Data Reference Architecture 44](#_Toc483319235)

[Figure 6: Notional Security and Privacy Fabric Overlay to the NBDRA 47](#_Toc483319236)

[Figure B-1: Composite Cloud Ecosystem Security Architecture [21] B-2](#_Toc483319237)

Tables

[Table 1: Draft Security Operational Taxonomy Mapping to the NBDRA Components 45](#_Toc483319238)

[Table 2: Classification of Cryptographic Technologies 49](#_Toc483319239)

[Table A-1: Mapping Consumer Digital Media Usage to the Reference Architecture A-1](#_Toc483319240)

[Table A-2: Mapping Nielsen Homescan to the Reference Architecture A-2](#_Toc483319241)

[Table A-3: Mapping Web Traffic Analytics to the Reference Architecture A-3](#_Toc483319242)

[Table A-4: Mapping HIE to the Reference Architecture A-4](#_Toc483319243)

[Table A-5: Mapping Pharmaceutical Clinical Trial Data Sharing to the Reference Architecture A-6](#_Toc483319244)

[Table A-6: Mapping Network Protection to the Reference Architecture A-7](#_Toc483319245)

[Table A-7: Mapping Military Unmanned Vehicle Sensor Data to the Reference Architecture A-8](#_Toc483319246)

[Table A-8: Mapping Common Core K–12 Student Reporting to the Reference Architecture A-9](#_Toc483319247)

[Table A-9: Mapping Cargo Shipping to the Reference Architecture A-10](#_Toc483319248)

[Table B-1: Standards and Guides Relevant to Cloud Computing [2] B-3](#_Toc483319249)

Executive Summary

This section will be updated during the finalization of Volume 4.

This *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy*document was prepared by the NIST Big Data Public Working Group (NBD-PWG) Security and Privacy Subgroup to identify security and privacy issues that are specific to Big Data.

Big Data application domains include health care, drug discovery, insurance, finance, retail and many others from both the private and public sectors. Among the scenarios within these application domains are health exchanges, clinical trials, mergers and acquisitions, device telemetry, targeted marketing and international anti-piracy. Security technology domains include identity, authorization, audit, network and device security, and federation across trust boundaries.

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

With the key Big Data characteristics of variety, volume, velocity, and variability in mind, the Subgroup gathered use cases from volunteers, developed a consensus-based security and privacy taxonomy, related the taxonomy to the NIST Big Data Reference Architecture (NBDRA), and validated the NBDRA by mapping the use cases to the NBDRA.

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

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

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

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

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

# INTRODUCTION

## Background

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

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

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

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

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

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

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

Motivated by the White House initiative and public suggestions, the National Institute of Standards and Technology (NIST) has accepted the challenge to stimulate collaboration among industry professionals to further the secure and effective adoption of Big Data. As one result of NIST’s Cloud and Big Data Forum held on January 15–17, 2013, there was strong encouragement for NIST to create a public working group for the development of a Big Data 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 value-added from Big Data service providers.

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

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

On September 16, 2015, seven volumes *NIST Big Data Interoperability Framework* V1.0 documents 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
* Volume 2, Taxonomies
* Volume 3, Use Cases and General Requirements
* Volume 4, Security and Privacy
* Volume 5, Architectures White Paper Survey
* Volume 6, Reference Architecture
* Volume 7, Standards Roadmap

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, the following two additional volumes have been identified.

* Volume 8, Reference Architecture Interfaces
* Volume 9, Adoption and Modernization

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.

## Scope And Objectives Of The Security And Privacy Subgroup

This section will be updated during the finalization of Volume 4.

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

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

* Provide a context from which to begin Big Data-specific security and privacy discussions;
* 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.

## Report Scope

This section will be updated during the finalization of Volume 4.

In Volume 4 Version 1, the NBDPWG introduced the concept of a security and privacy fabric. The fundamental idea is that security and privacy considerations impact all components with the NBDRA. This version of the document extends and amplifies this concept.

In addition, rather than embracing a maturity model, a safety engineering approach is chosen. The threats to safety and privacy in Big Data are sufficiently grave, and 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, address that space for Big Data and IoT.

This document introduces complex topics. Some are new, others are older, but have resurfaced with greater urgency because Big Data amplifies effects and/or risks. Despite that, the single broadest objective is to offer a three-level safety rating for a Big Data system. This high-medium-low simplification is offered in a list form (Appendix []), though it can be implemented through semi-automated means; the latter are indicated but not proscriptive.

Big Data is not yet a mature technology, but it has clearly taken hold. 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.

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

Additional Need: Need a couple of sentences to accompany each bullet below. Background here is to highlight differences from v1.

* 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 in DevOps frameworks (“SecDevOps”)
* 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 models to guide security and privacy practices
* Blockchain and higher-granularity dynamic “smart contracts”
* Cryptography and privacy-preserving methods
* Big Data forensics frameworks must be co-engineered with security measures
* Increased use of attributed based security
* Provide a broadly usable self-assessment for conformance to Big Data security levels
* Microservices and Security

## Report Production

This section will be updated during the finalization of Volume 4.

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

* Announce that the NBD-PWG Security and Privacy Subgroup is open to the public in order to attract and solicit a wide array of subject matter experts and stakeholders in government, industry, and academia;
* Identify use cases specific to Big Data security and privacy;
* 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 PWG. Since this is a community effort, there are several topics covered that are related to security and privacy. While an effort has been made to connect the topics, gaps may come to light that could be addressed in Version 2 of this document.

## Report Structure

This section will be updated during the finalization of Volume 4.

Following this introductory section, the remainder of this document is organized as follows:

* Section 2 discusses security and privacy issues particular to Big Data.
* Section 3 presents examples of security- and privacy-related use cases.
* Section 4 offers a preliminary taxonomy for security and privacy.
* Section 5 introduces the details of a draft NIST Big Data security and privacy reference architecture in relation to the overall NBDRA.
* Section 6 maps the use cases presented in Section 3 to the NBDRA.
* Appendix A discusses special security and privacy topics.
* Appendix B contains information about cloud technology.
* Appendix C lists terms and definitions relevant to Big Data security and privacy.
* Appendix D contains the acronyms used in this document.
* Appendix E lists the references used in the document.

Version Overview

Version 2 of the Big Data Security and Privacy (SnP) document reflects changes in the technology environment as well as ongoing work within the WG.

Specific objectives for this version include:

(The items in this list will be edited and incorporated into the text.)

1. Simplify document usability for novice implementers.
2. Expand the SnP framework depth by deeper cross-linking to related standards. [See possible template for this approach IEEE 1484.1)
3. See Cryptographic Technologies for Data Transformations. The V2 document is updated to reflect recent cryptology practices.
4. A safety framework is introduced, suitable for use by unaffiliated citizens, big data software architects and IT managers. (See IEC standards 61508. 61671, 62046, SC22 WG 23)
5. Explicitly provide for phase-specific guidance.
6. Provide Levels of conformance to Big Data SnP practices across SnP phases. Low, medium and high conformance levels (“Conformity Assessment” in the “NIST Roadmap for Improving Critical Infrastructure Cybersecurity”) are provided (similar to NIST 800-53).
7. Identify guidelines for integrating supporting Big Data systems dedicated to SnP (“Big Data SnP dogfood”).[[2]](#footnote-2) (healthcare as strongest use case)
8. Incorporate SnP metadata-rich Big Data orchestration processes, enabled by tools such as Rundeck [2]. (Paired with the test bed demonstration.)
9. Include SnP dependency frameworks.
10. Reflect the growing importance of SnP aspects to the API-first and microservices design pattern. (Frank has some references from data type, RPCs & possibly API design; Bell Labs on provability of protocols)
11. Facilitate incorporation of SnP models for the software development life cycle. (Is there a LC for analytics?)
12. Draft a Big Data-annotated version of the NIST Privacy Catalog (see NIST 800-53, Appendix J).
13. Identify Big Data touchpoints for Privacy by Design, OECD and other external privacy guidelines.
14. Integrate models such as Sensing as a Service [3]
15. More directly address SnP issues with geospatial and mobile data [4].
16. Include software defined networks and other virtual network security concepts, as in NIST 800-125B [5].
17. Provide a deeper explanation of Application Provider SnP requirements.
18. Provide references to third party references on risks, verifiability and provenance for analytics that affect SnP. (Premise: Big Data amplifies an already known DC risk).
19. SnP risk frameworks for specific design patterns (excludes cloud): distributed computing, middleware (“enterprise service bus”), agent-based, recommendation engines, web portals fronting legacy applications.
20. More clearly identify where Big Data systems management intersects with SnP guidelines. The gold standard use case is the use of logging data for both operational intelligence and SnP, though the mapping is demonstrably nonorthogonal.
21. Identify weaknesses in the current RA and propose updates.
22. Depict policy and metadata orchestration using descriptions of test beds, such as developed at Indiana University.
23. “System Communicator” and FTC
24. API and OpenCL [6]
25. See NIST 800-113 [http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-183.pdf](http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-183.pdf%20%20)

## Future Work On This Volume

This section will be written during the finalization of Volume 4.

# BIG DATA SECURITY AND PRIVACY

Section scope: What is S&P? How is it defined by NBD-PWG? What are emerging technology areas that affect S&P?

## What is Different about Big Data Security and Privacy

Subsection scope: Why a security and privacy fabric?

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

1. Big Data projects often encompass heterogeneous components in which a single security scheme has not been designed from the outset.
2. Most security and privacy methods have been designed for batch or online transaction processing systems. Big Data projects increasingly involve one or more streamed data sources that are used in conjunction with data at rest, creating unique security and privacy scenarios.
3. The use of multiple Big Data sources not originally intended to be used together can compromise privacy, security, or both. Approaches to de-identify personally identifiable information (PII) that were satisfactory prior to Big Data may no longer be adequate, 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 multiple datasets exacerbates the risk of compromising privacy.
4. An increased reliance on sensor streams, such as those anticipated with the Internet of Things (IoT; e.g., smart medical devices, smart cities, smart homes) can create vulnerabilities that were more easily managed before amassed to Big Data scale.
5. Certain types of data thought to be too big for analysis, such as geospatial and video imaging, will become commodity Big Data sources. These uses were not anticipated and/or may not have implemented security and privacy measures.
6. Issues of veracity, 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.

Potential new items for list of differences. These should be converted into separate paragraphs of 3-5 sentences.

1. Inter-organizational (e.g., federation, data licensing -- not only for cloud)
2. Mobile / geospatial increased risk for deanonymization
3. Change to lifecycle processes (no “archive” or “destroy” b/c of big data)
4. Related sets of standards are written with large organizational assumptions; today big data can be created / analyzed with small teams
5. Audit and provenance for big data intersects in novel ways with these other aspects.
6. Big Data AS a technology accelerator for improved audit (e.g., blockchain, noSQL, machine learning for infosec enabled by big data), analytics for intrusion detection, complex event processing
7. Transborder data flows (there is a related OMG initiative)
8. Consent (“smart contracts”) frameworks, perhaps implemented using blockchain
9. Impact of real time big data (e.g., Apache Spark) on security and privacy.
10. Risk Management in big data moves focus to inter-organizational risk and risks associated with analytics vs. four-walls perspective.
11. Lesser importance, but relevant DevOps and Agile processes related to the efforts of small teams (even single-developer effort) in creation and fusion using big data

Overall: Need to build new / update frameworks for Big Data referencing existing ISO and other standards for big data life cycle, audit, configuration management and privacy preserving practices.

## 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. [7] Less than a third of data needed protection in 2010, but more than 40 percent of data will need protection in 2020. [7]

Security and privacy measures for Big Data involve a different approach than traditional systems. Big Data is increasingly stored on public cloud infrastructure built by employing various hardware, operating systems, and analytical software. Traditional security approaches usually addressed small-scale systems holding static data on firewalled and semi-isolated networks. The surge in streaming cloud technology necessitates extremely rapid responses to security issues and threats. [8]

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 is not conceptually distinct from Big Data, it may have been produced using different assumptions. One of the primary objectives of this document is to understand how Big Data security and privacy requirements arise out of the defining characteristics of Big Data and related emerging technologies, and how these requirements are differentiated from traditional security and privacy requirements.

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

* Big Data may be gathered from diverse end points. Actors include more types than just traditional providers and consumers—data owners, such as mobile users and social network users, are primary actors in Big Data. Devices that ingest data streams for physically distinct data consumers may also be actors. This alone is not new, but the mix of human and device types is on a scale that is unprecedented. The resulting combination of threat vectors and potential protection mechanisms to mitigate them is new.
* Data aggregation and dissemination must be secured inside the context of a formal, understandable framework. The availability of data and transparency of its current and past use by data consumers is an important aspect of Big Data. However, Big Data systems may be operational outside formal, readily understood frameworks, such as those designed by a single team of architects with a clearly defined set of objectives. In some settings, where such frameworks are absent or have been unsystematically composed, there may be a need for public or walled garden portals and ombudsman-like roles for data at rest. These system combinations and unforeseen combinations call for a renewed Big Data framework.
* Data search and selection can lead to privacy or security policy concerns. There is a lack of systematic understanding of the capabilities that should be provided by a data provider in this respect.c A combination of well-educated users, well-educated architects, and system protections may be needed, as well as excluding databases or limiting queries that may be foreseen as enabling re-identification. If a key feature of Big Data is, as one analyst called it, “the ability to derive differentiated insights from advanced analytics on data at any scale,” the search and selection aspects of analytics will accentuate security and privacy concerns. [9]
* Privacy-preserving mechanisms are needed for Big Data, such as for Personally Identifiable Information (PII). Because there may be disparate, potentially unanticipated processing steps between the data owner, provider, and data consumer, the privacy and integrity of data coming from end points should be protected at every stage. End-to-end information assurance practices for Big Data are not dissimilar from other systems but must be designed on a larger scale.
* Big Data is pushing beyond traditional definitions for information trust, openness, and responsibility. Governance, previously consigned to static roles and typically employed in larger organizations, is becoming an increasingly important intrinsic design consideration for Big Data systems. c Reference to NBDRA Data Provider.
* Legacy security solutions need to be retargeted to the infrastructural shift due to Big Data. Legacy security solutions address infrastructural security concerns that still 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 methods to prevent adversarial manipulation and preserve integrity of operations.
* Information assurance and disaster recovery for Big Data Systems may require unique and emergent practices. Because of its extreme scalability, Big Data presents challenges for information assurance (IA) and disaster recovery (DR) practices that were not previously addressed in a systematic way. Traditional backup methods may be impractical for Big Data systems. In addition, test, verification, and provenance assurance for Big Data replicas may not complete in time to meet temporal requirements that were readily accommodated in smaller systems.
* Big Data creates potential targets of increased value. The effort required to consummate system attacks will be scaled to meet the opportunity value. Big Data systems will present concentrated, high-value targets to adversaries. As Big Data becomes ubiquitous, such targets are becoming more numerous—a new information technology (IT) scenario in itself.
* Risks have increased for de-anonymization and transfer of PII without consent traceability. Security and privacy can be compromised through unintentional lapses or malicious attacks on data integrity. Managing data integrity for Big Data presents additional challenges related to all the Big Data characteristics, but especially for PII. While there are technologies available to develop methods for de-identification, some experts caution that equally powerful methods can leverage Big Data to re-identify personal information. For example, the availability of unanticipated datasets could make re-identification possible. Even when technology is able to preserve privacy, proper consent and use may not follow the path of the data through various custodians. 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. [10] [11]
* Emerging Risks in Open Data and Big Science. Data identification, metadata tagging, aggregation, and segmentation—widely anticipated for data science and open datasets—if not properly managed, may have degraded veracity because they are derived and not primary information sources. Retractions of peer-reviewed research due to inappropriate data interpretations may become more commonplace as researchers leverage third-party Big Data.

## Security And Privacy Impacts On Big Data Characteristics

Variety, volume, velocity, and variability are key characteristics of Big Data and commonly referred to as the Vs of Big Data. Where appropriate, these characteristics shaped discussions within the NBD-PWG Security and Privacy Subgroup. While the Vs provide a useful shorthand description, used in the public discourse about Big Data, there are other important characteristics of Big Data that affect security and privacy, such as veracity, validity, and volatility. These elements are discussed below with respect to their impact on Big Data security and privacy.

### 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. [12] These systems were not designed with security and privacy in mind, and these functions are usually relegated to middleware. Traditional encryption technology also hinders organization of data based on semantics. The aim of standard encryption is to provide semantic security, which means that the encryption of any value is indistinguishable from the encryption of any other value. Therefore, once encryption is applied, any organization of the data that depends on any property of the data values themselves are rendered ineffective, whereas organization of the metadata, which may be unencrypted, may still be effective.

An emergent phenomenon introduced by Big Data variety that has gained considerable importance is the ability to infer identity from anonymized datasets by correlating with apparently innocuous public databases. While several formal models to address privacy preserving data disclosure have been proposed, [13] 10 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 11 and has led to many instances of de¬anonymization in conjunction with publicly available data.12 Although some laws/regulations recognize only identifiers per se, laws such as HIPAA (the statistician provision), FERPA, and 45 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.

### Volume

The volume of Big Data describes how much data is coming in. In Big Data parlance, this typically ranges from gigabytes 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.13

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

### Velocity

Velocity describes the speed at which data is processed. The data usually arrives in batches or is streamed continuously. As with certain other non-relational databases, distributed programming frameworks were not developed with security and privacy in mind.14 Malfunctioning computing nodes might leak confidential data. Partial infrastructure attacks could compromise a significantly large fraction of the system due to high levels of connectivity and dependency. If the system does not enforce strong authentication among geographically distributed nodes, rogue nodes can be added that can eavesdrop on confidential data.

### Veracity

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

***Provenance***—or what some have called veracity in keeping with the V theme—is important for both data quality and for protecting security and maintaining privacy policies. Big Data frequently moves across individual boundaries to groups and communities of interest, and across state, national, and international boundaries. Provenance addresses the problem of understanding the data’s original source, such as through metadata, though the problem extends beyond metadata maintenance. Also as noted before, with respect to privacy policy, additional context is needed to make responsible decisions over collected data—this may include the form of consent, intended use, temporal connotations (like Right to be Forgotten) or broader context of collection. This 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,15 but no clear guidelines yet exist.

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

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

***Veracity*** (in some circles also called Provenance, though the two terms are not identical) also encompasses information assurance for the methods through which information was collected. For example, when sensors are used, traceability, calibration, version, sampling, and device configuration is needed.

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

Furthermore, there is a central and broadly recognized privacy principle, incorporated in many privacy frameworks (e.g., the OECD principles, EU data protection directive, FTC fair information practices) that data subjects must be able to view and correct information collected about them in a database.

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

Several examples of ‘invalid’ uses for Big Data have been cited. Click fraud, conducted on a Big Data scale, but which can be detected using Big Data techniques, has been cited as the cause of perhaps $11 billion in wasted advertisement spending. 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.16 Each of these can be conducted at Big Data scale and may require Big Data solutions to detect and combat.

Despite initial enthusiasm, some trend-producing applications that use social media to predict the incidence of flu have been called into question. A study by Lazer et al.17 suggested that one application overestimated the prevalence of flu for 100 of 108 weeks studied. Careless interpretation of social media is possible when attempts are made to characterize or even predict consumer behavior using imprecise meanings and intentions for “like” and “follow.”

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

### Volatility

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

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

While research has been conducted into how to manage temporal data (e.g., in e-science for satellite instrument data),18 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.

## Effects of Emerging Technology on Big Data Security and Privacy

Subsection Scope: How does S&P relate to other areas, such as cloud, IoT, etc. How do you know if you have a BD S&P problem and does this report apply to you? Add the technology components from the new topics introduced for v2.

### Cloud Computing

Subsection scope: Discuss the relation between Cloud Computing and Big Data with respect to Security and privacy. What are the challenging problems SnP might face in cloud that are different from on-premises? Keep in mind that the Appendix will have a crosswalk to NIST SnP standards / reference model.

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 consumer
* Dynamic system boundaries and commingled roles and responsibilities between consumers and providers
* Multi-tenancy
* Data residency
* Measured service
* Order-of-magnitude increases in scale (on demand), dynamics (elasticity and cost optimization), and complexity (automation and virtualization)

These cloud computing characteristics often present different security risks to an organization than the traditional 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 require security components and controls to be fully and accurately implemented.

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

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 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 SP 800-144 (Jansen & Grance, 2011), which included a then-current list of related standards and guides, which is reproduced in Appendix B.

In the EU, consider the ETSI Cloud Standards Coordination Report (ETSI, 2013).

More recently, the DISA at the Department of Defense published its Cloud Security Requirements Guide (DISA, 2015), which covers DoD projects through the secret level.

On the privacy front, when the Federal CIO Council published recommendations for Digital Privacy Controls (CIO\_Council, 2012), 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).

### Big Data Security Quilt

In Version 2, the analogy is extended further to the notion of quilt.

The Big Data SnP Quilt (BDSQ) is a working definition for a Big Data SnP package. Implementation of the BD Quilt could be achieved through XML, in a conventional metadata repository, an Excel checklist, a portal, an API, an app, a suite of microservices or a combination of any of these. A BDSQ serves as a container for Big Data fabric descriptions. The container can be inspected, relayed, annotated by different components of the NBDRA.

Participants of the working group considered related design patterns. For example, albeit dissimilar, the DMTF Cloud Auditing Data Federation (CADF), which has been implemented for OpenStack, offers audit prescriptions that can be straightforwardly adapted for the BDRA.

### Big Data Security Safety Levels

Three Big Data Security system safety levels are recommended in this version. When paired with a checklist and recommended practices, organizations can self-designate their systems as conforming to a safety level as identified in this report.

Possibly include text on confounding, missing data, and bias (here and/or Section on data quality etc.).

### Internet of Things and CPS

Section Scope: Discuss internet of things and cyber-physical systems in relation to Big Data security and privacy

This version of the standard identifies connections to IoT security issues and links to related standards efforts in those communities at NIST (Voas, 2016) and elsewhere.

### Mobile Devices and Big Data

Additional need: This section may be revised to make a stronger case

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 dusted off and updated for current security concerns – mobile must be seen as a critical element of Big Data.

Although mobile spans many facets of computer security, there are several reasons for this:

Additional need: Additional items needed in the list for mobile with respect to computer security

1. Mobile devices challenge governance and controls for enterprises, especially in BYOD environments. As a result, specialized security approaches enabling mobile-centric access controls have been proposed (Das, Joshi, & Finin, 2016)
2. Mobile devices often disclose geospatial data which can be used in big data settings to enrich other data sets, and even to perform de-anonymization.
3. [] Continue to list.

### Integration of People and Organizations

Subsection Scope: people and organizations intro. IEEE P7000. See also some ISO series that address organizational aspects “Systems Management” and SysML

The fabric did not integrate roles and organizations into Big Data workflow.

To communicate across organizations, XML-based solutions should be considered. For example, Lenz and Oberweis suggested using an XML variant of Petri nets (Lenz & Oberweis, 2003). 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).

Similarly, HTML microdata can be used to transfer or house information exchanged across organizational boundaries (Hickson, 2013). Microdata has been extended for use with RDF (Hickson, Kellogg, Tenisson, & Herman, 2014).

We 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, defense. [1]

### System Communicator

Big Data systems which collect, store, manage or transform data considered in need of protection (such as what is called out as PCI or PII) 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.

### Ethical Design

Section Scope: discuss ethical design. See the work of IEEE P7000

Section is a shout out to the work of IEEE P7000.

#### Self-Cleansing Systems

Subsection Scope: Describe self-cleaning systems.

#### The Toxic Data Model

Subsection Scope: Describe the toxic data model

#### Relation to Systems Management

Subsection Scope: Discuss the relation of ethical design to systems management. Maybe could include use cases from the press (e.g., Uber, Volkswagen)

#### Big Data Safety Annotation

##### Risk Management

Subsection Scope: What is the Big Data systems take on the NIST and ISACA Risk Management frameworks?

##### Federation of safety practices

Subsection Scope: Possibly using marketplace (closed clearinghouses, etc.; federation as an engineering principle; see InCommon, GENI.net, OASIS IDTrust; see use case of out-of-band guest identity) []

[Tim content]

#### 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” manage through the use of cloud brokers (Fazio & Puliafito, 2011).

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

Middleware, message-passing, enterprise service bus – these concepts remain important for Big Data. For example, in SE-CLEVER, investigators wanted to address issues raised by the Cloud Security Alliance in their XMPP-based middleware (Celesti, Fazio, & Villari, 2013).

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 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. [

#### Orchestration in Weak Federation Scenarios

Subsection Scope: This is a new section. Some academic / industry white paper research could be referenced here.

#### Consent and the Glass-breaking Scenario

Subsection Scope: Insert language on consent. Some text may be available from Tuesday NBDPWG calls and MAU. See also emergency preparedness use case (Frank has Manhattan building evacuation scenario).

## Security and Privacy Methodology with Respect to Big Data

Subsection Scope: Introductory paragraph to explain subsection topic. Reference other standards and set the context for the broader knowledge that we are discussing.

### Why is this relevant for big data?

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 we must be aware that there are inherent limitations and dangers to such approaches.

# EXAMPLE USE CASES FOR SECURITY AND PRIVACY

There are significant Big Data challenges in science and engineering. Many of these are described in the use cases in NIST Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements. However, 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 use cases were collected when the reference architecture was not mature. The use cases were collected from BDWG 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. In Version 2, additional mapping of the use cases to the NBDRA and taxonomy will be developed. Parts of this document were developed in parallel, and the connections will be strengthened in Version 2.

Updated Security and Privacy Use Cases

Emerging use cases have guided the development of the version 2 framework. For instance, while V1 made reference to the use of Big Data systems to support SnP, several have now emerged in regulatory spaces and are influencing design decisions today.

## Retail/Marketing

### Consumer Digital Media Usage

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

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

Current Security and Privacy 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 properties:
* Predictability around the processing of personal information, in order to enable individuals to make appropriate determinations for themselves or prevent problems arising from actions such as unanticipated revelations about individuals.
* Manageability of personal information, in order to prevent problems arising from actions such as dissemination of inaccurate information or taking unfair advantage of individuals based on information asymmetry in the marketplace
* Disassociability of information from individuals in order 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
* 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.”19 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.

### 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.

### 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 web site 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.20
* 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. 21

## Healthcare

### Health Information Exchange

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

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

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

The following are some preliminary scenarios that have been proposed by the NBD PWG:

* Break-the-Glass: There could be situations where the patient is not able to provide consent due to a medical situation, or a guardian is not accessible, but an authorized party needs immediate access to relevant patient records. Cryptographically enhanced key life cycle management can provide a sufficient level of visibility and 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; US DHS Health IT Policy Committee, Privacy and Security Workgroup); and Health Level Seven (HL7) International Version 3 standards for Data Access Consent, Consent Directives)
* Pandemic Assistance: There will be situations when public health entities, such as the CDC and perhaps other nongovernmental organizations that require this information to facilitate public safety, will require controlled access to this information, perhaps in situations where services and infrastructures are inaccessible. A cloud HIE with the right cryptographic controls could release essential information to authorized entities through authorization and audits in a manner that facilitates the scenario requirement.
* 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.22
  + Applied cryptography: Tight reductions, realistic threat models, and efficient techniques • Privacy:
  + Differential privacy: Techniques for guaranteeing against inappropriate leakage of PII
  + HIPAA

### Genetic Privacy

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

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

Current Security and Privacy Issues/Practices:

* Security:
  + Secure Sockets Layer (SSL)-based authentication and access control. Basic user registration with low attestation level
  + Concerns over data ownership and custody upon user death
  + Site administrators may have access to data—strong encryption and key escrow are recommended
* Privacy:
  + Transparent, logged, policy-governed controls over access to genetic information
  + Full life cycle data ownership and custody controls

### Pharma Clinical Trial Data Sharing23

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.

## Cybersecurity

### 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 workstations and servers, but they are not yet able to scale to Big Data. Improved security software will include physical data correlates (e.g., access card usage for devices as well as building entrance/exit) and likely be more tightly integrated with applications, which will generate logs and audit records of previously undetermined types or sizes. Big Data analytics systems will be required to process and analyze this data to deliver meaningful results. These systems could also be multi-tenant, catering to more than one distinct company.

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.24
* 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 Incident and Event Management (SIEM).25

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 security information and event management (SIEM).

## Government

### Unmanned Vehicle Sensor Data

Scenario Description: Unmanned Aerial Vehicles (UAV’s), 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 US border areas, and sometimes flood areas; it allows authorized government workers to see real time video and radar. [[3]](#footnote-3)

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.26
* 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 Justice Department policy guideline produced by the DOJ Office of Legal Policy. [[4]](#footnote-4) The guideline acknowledges the value of Unmanned Aircraft Systems (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 U.S. is subject to local regulation. For example, in the EU, guidelines are under discussion that 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 RPAS applications (http://ec.europa.eu/enterprise/sectors/aerospace/uas /).

### 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.27
* Privacy-related disputes surrounding education Big Data are illustrated by the reluctance of states to participate in the InBloom initiative.28
* 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. 29

Current Research:

* Longitudinal performance data would have value for program evaluators and educators. Work in this area was proposed by Deakin Crack, Broadfoot & Claxton (2004) as a “Lifelong Learning Inventory,” and further by Ferguson (2012), 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. In order 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” (Section 9.2). 30,31
* Data-driven learning 32 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 33 predictive analytics for student decision making.

## Industrial: Aviation

### Sensor Data Storage And Analytics

Scenario Description: Most commercial airlines are equipped with hundreds of sensors to constantly capture engine and/or aircraft health information during a flight. For a single flight, the sensors may collect multiple gigabytes of data and transfer this data stream to Big Data analytics systems. Several companies manage these Big Data analytics systems, such as parts/engine manufacturers, airlines, and plane manufacturers, and data may be shared across these companies. The aggregated data is analyzed for maintenance scheduling, flight routines, etc. 34Companies 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)

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)

## Transportation

### Cargo Shipping

The following use case outlines how the shipping industry (e.g., FedEx, UPS, DHL) regularly uses Big Data. Big Data is used in the identification, transport, and handling of items in the supply chain. The identification of an item is important to the sender, the recipient, and all those in between with a need to know the location of the item while in transport and the time of arrival. Currently, the status of shipped items is not relayed through the entire information chain. This will be provided by sensor information, GPS coordinates, and a unique identification schema based on the new International Organization for Standardization (ISO) 29161 standards under development within the ISO technical committee ISO JTC1 SC31 WG2. (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.

## Major Use Case: [Sec Consolidated Audit Trail](http://www.catnmsplan.com/)

The SEC Consolidated Audit Trail (CAT) project is forecast to consume 10 terabytes of data daily. The system’s security requirements, which stemmed from a past system failure with lack of traceability, are considerable.

High Level CAT Security Requirements[[5]](#footnote-5)

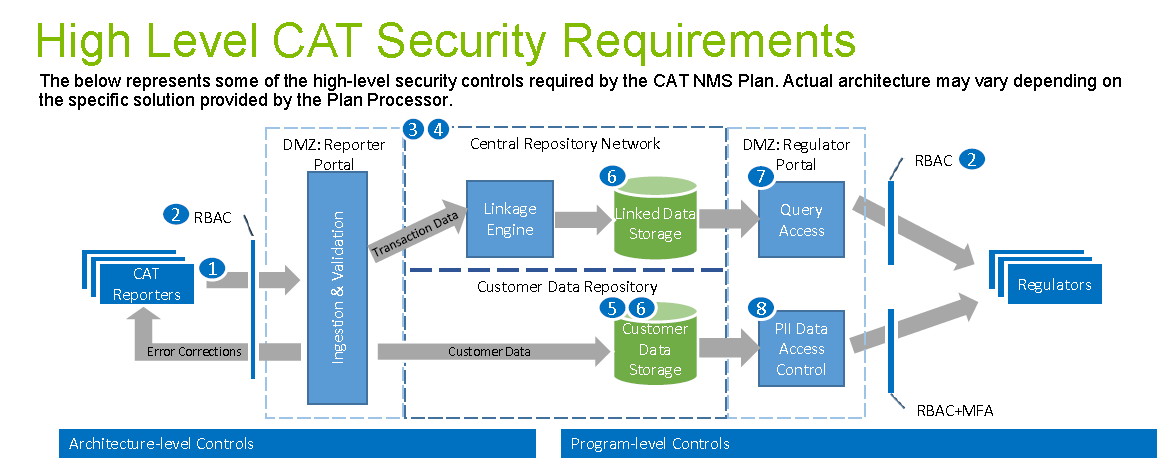


Figure 1: High Level CAT Requirements

## Major Use Case: Iot Device Management

This family of use cases involves the onboarding, decommissioning, quarantining of numerous devices, such as for IoT and cyber-physical systems.

Safety systems incorporating voluminous sensor streams represent this family of use cases.

### Smart Home IoT

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 the like.

## Major Use Case: Omg Data Residency Initiative

(but what about analytics, deep learning, etc.; see analytics in the RA)

### Minor Use Case: TBD

### Use Case: Emergency management data (XChangeCore interoperability standard).

## Major Use Case: Health Care Consent Flow

Audit aspects of consent are presented in documents developed by an HL7 community (Proud-Madruga, 2016).

## Major Use Case: “Heart Use Case: Alice Selectively Shares Health-Related Data With Physicians And Others”[[6]](#footnote-6)

## Major Use Case Blockchain for Fintech (Arnab)

### Minor Use Case – In-Stream PII

Google Hangout shift from desktop to mobile.

## Major Use Case—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 (Kauffman\_Foundation, 2016). The portal attempts to provide anonymized student and institutional performance data for educational decision support.



Figure 2: EdWise Figure

# TAXONOMY OF SECURITY AND PRIVACY TOPICS

Section Scope: This taxonomy is an abstraction of the use cases. The taxonomy items can be considered issues that are faced in BD S&P. Explain taxonomies. Discuss taxonomy with respect to selected use case (side bar?)

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.36 Candidate topics and related material used in preparing this section are provided for reference in Appendix A.

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

The following subsections present an overview of each security and privacy taxonomy, along with lists of topics encompassed by the taxonomy elements. These lists are the results of preliminary discussions of the Subgroup and may be developed further in Version 2. As noted earlier, Version 1 focuses 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.

## 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.

Figure 3: Security and Privacy Conceptual Taxonomy

### 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 [Boneh, Sahai and Waters, “Functional Encryption: Definitions and Challenges,” TCC 2011] that support searching and reporting on encrypted data—any information about the plain text not deducible from the search criteria is guaranteed to be hidden
  + Homomorphic encryption: Cryptographic protocols that support operations on the underlying plain text of an encryption—any information about the plain text is guaranteed to be hidden
* Secure data aggregation: Aggregating data without compromising privacy
* Data anonymization
  + De-identification of records to protect privacy
* Key management
  + As noted by Chandramouli and Iorga, cloud security for cryptographic keys, an essential building block for security and privacy, takes on “additional complexity,” which can be rephrased for Big Data settings: (1) greater variety due to more cloud consumer-provider relationships, and (2) greater demands and variety of infrastructures “on which both the Key Management System and protected resources are located.” 37
  + 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. 38
  + 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.

### 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

### System Health

In a separate discussion, the interwoven notions of [Other Volume Link] [] 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 (Anonymous, "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/).

(INSERT Big Data aspects in [*https://www.techopedia.com/2/27825/security/the-basic-principles-of-it-security*](https://www.techopedia.com/2/27825/security/the-basic-principles-of-it-security)) – Management aspects in summary.

* System Availability is a key element in “C-I-A” (Confidentiality – Integrity – Availability) - 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

### Public Policy, Social and Cross-Organizational Topics

The following set of topics is drawn from an Association for Computing Machinery (ACM) grouping.39 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[[7]](#footnote-7)
* 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

## 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 standards development process, these methodologies have not been incorporated as part of a cohesive security fabric. They are potentially valuable checklist-style elements that can solve specific security or privacy needs. Future work must better integrate these methodologies with risk management guidelines developed by others (e.g., NIST Special Publication 800-37 Revision 1, Guide for Applying the Risk Management Framework to Federal Information Systems 40, draft NIST Internal Report 8062, Privacy Risk Management for Federal Information Systems 41, and COBIT Risk IT Framework 42).

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.43

The operational taxonomy will overlap with small data taxonomies while drawing attention to specific issues with Big Data.44 45

Figure 4: Security and Privacy Operational Taxonomy

### 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

### 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). 46 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

### Data Governance

However large and complex Big Data becomes in terms of data volume, velocity, variety, and variability, Big Data governance will, in some important conceptual and actual dimensions, be much larger. Big Data without Big Data governance may become less useful to its stakeholders. To stimulate positive change, data governance will need to persist across the data 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” 47
  + De-identification and anonymization
  + Re-identification risk management
* End-point validation
* DRM
* Trust
* Openness
* Fairness and information ethics 48

#### Compliance, Governance and Management as Code

In 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 this progression:

* Software as code
* Tests as code
* Infrastructure as code
* Compliance as code [Ed. Italics added]

Just as 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.

### 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 European Telecommunications Standards Institute (ETSI) TS 102 484 Smart Card specifications 49
    - 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.

### 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.50 Business risk models can incorporate privacy considerations.

## Roles Related at 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.

### 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.

### Governance, Risk Management, and Compliance

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

### 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.

## 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.

### 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.

### Provenance

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

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

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

“The changes announced today expand many of the requirements to business associates of these entities that receive protected health information, such as contractors and subcontractors. Some of the largest breaches reported to HHS have involved business associates. Penalties are increased for noncompliance based on the level of negligence with a maximum penalty of $1.5 million per violation.”51

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.52

### System Health Management

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

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

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

Volume, velocity, variety, and variability of Big Data systems health make it different from small data system health. Health tools and design patterns for existing systems are likely insufficient to handle Big Data—including Big Data security and privacy. At least one commercial web services provider has reported that its internal accounting and systems management tool uses more resources than any other single application. The volume of system events and the complexity of event interactions is a challenge that demands Big Data solutions to defend Big Data systems. Managing systems health—including security—will require roles defined as much by the tools needed to manage as by the organizational context. Stated differently, Big Data is transforming the role of the Computer Security Officer.

For example, one aspect motivated by the DevOps movement (i.e., move toward blending tasks performed by applications development and systems operations teams) is the rapid launch, reconfiguration, redeployment, and distribution of Big Data systems. Tracking intended vs. accidental or malicious configuration changes is increasingly a Big Data challenge.

### 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 Federal Trade Commission, the Food and Drug Administration 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.” 53
* 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 web site 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.54
* 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. 55
* 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). 56

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.

## Additional Taxonomy Topics

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

### 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.

### Data Syndication

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

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

### ACM Taxonomy

Subsection Scope:

Where possible, this version of the Big Data SnP standard uses the terminology adopted by the ACM Computing Classification System (Mirkin, Nascimento, & Pereira, 2008) and (Lin, Zhang, Zhao, & J., 2012). The ACM 2012 CCS is accessible online (ACM, n.d.) and can be represented in Simple Knowledge Organization System (SKOS) format (Miles & Bechhofer, 2009).

A systematic taxonomy has several benefits for Big Data SnP. In addition to tracking new research and guidelines (e.g., cryptography index example here), standardized terminology can, in some limited contexts, allow for automated reasoning. Automated reasoning, based on cybersecurity ontologies, for example, could enable fine-grained alerts that elevate when it makes sense to do, 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 (Obrst, Chase, & Markeloff, 2012).

Other taxonomies may be useful. For example, the NIST NISTIR 8085 draft “Forming Common Platform Enumeration (CPE) Names from Software Identification (SWID) Tags” is designed to [] (Cheikes, 2015).

## Why Security Ontologies Matter For Big Data

Subsection Scope:

Suppose you are an engineer who inherits software and/or data from a third party. Whether it’s within your organization, or across organizations, it’s important to know what security components are present in your inheritance.

[] Explain why this matters.

# BIG DATA REFERENCE ARCHITECTURE AND SECURITY AND PRIVACY FABRIC

Section needs: Could use additional review and text to enhance section. Suggestions include the following: Before jumping into SnP Fabric solution, we might want to identify concrete what are the SnP requirements and problems that the SnP fabric is trying to solve. We might want to have a leading section as 3.1 SnP Requirements to extract list of requirements from Section 2 then follow with Section 3.2 on Security and Privacy Fabric in NBDA with sub-sections on 3.2.1 on Security Fabric and Section 3.2.2 on Privacy Fabric. Then Section 3.3 Security and Privacy Approach to Big Data Challenges with 3.3.1 Arnab’s Cryptographic Technologies for Secure Data Transformation 3.3.2 other technologies approach to SnP…

## Security and Privacy Requirements

Subsection Scope: Discuss the security and privacy requirements extracted from the S&P use cases.

## NIST Big Data Reference Architecture

Security and privacy considerations are a fundamental aspect of the NBDRA. Using the material gathered for this volume and extensive brainstorming among the NBD-PWG Security and Privacy Subgroup members and others, the following proposal for a security and privacy fabric was developed.[[8]](#footnote-8)

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

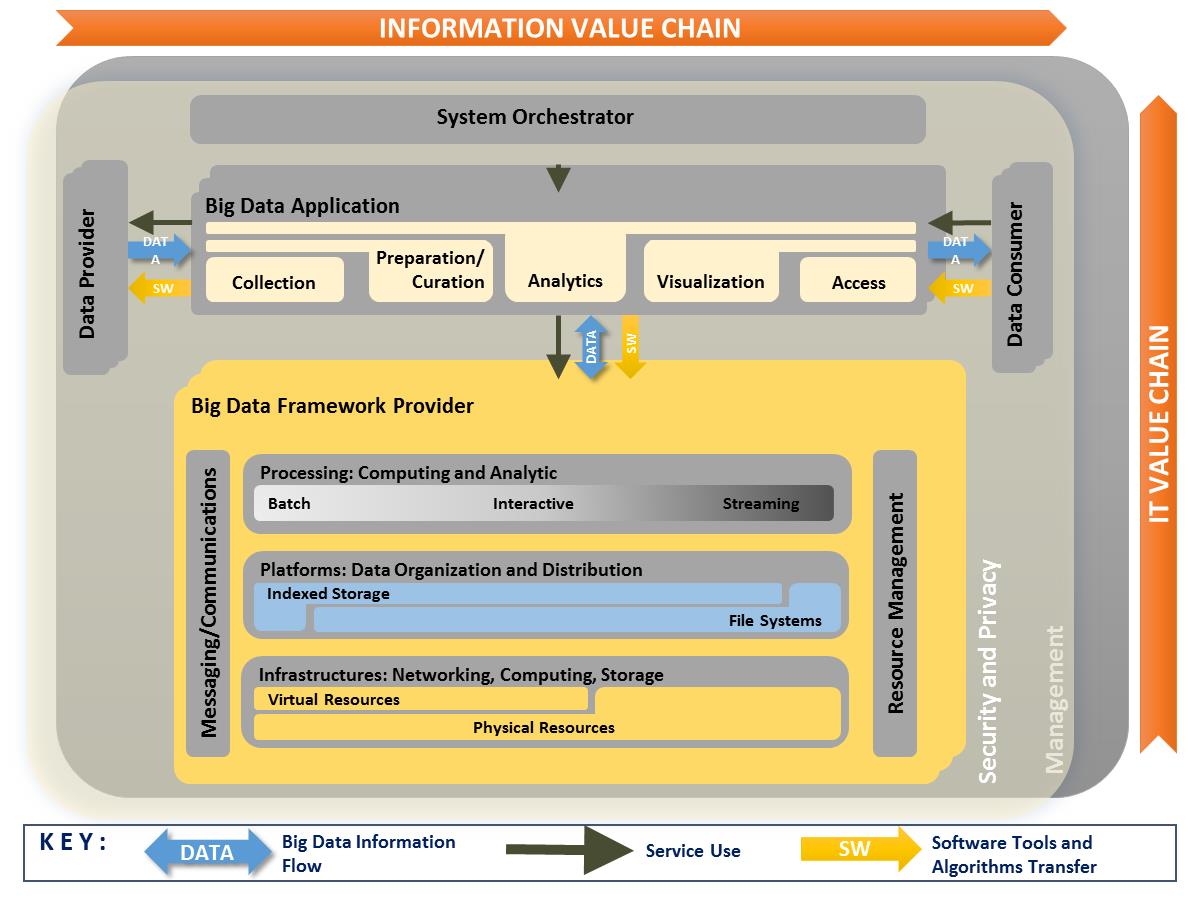


Figure 5: NIST Big Data Reference Architecture

This pervasive dimension is depicted in Figure 5 by the presence of the security and privacy fabric surrounding all of the functional components. NBD-PWG decided to include the Data Provider and Data Consumer as well as the Big Data Application and Framework Providers in the Security and Privacy Fabric because these entities should agree on the security protocols and mechanisms in place. The NIST Big Data Interoperability Framework: Volume 6, Reference Architecture document discusses in detail the other components of the NBDRA.

At this time, explanations as to how the proposed 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 -out. In the process of scaling, specific security, privacy, or GRC mechanism or practices may need to be invoked.

## Relation Of The Big Data Security Operational Taxonomy To The NBDRA

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

Table 1: Draft Security Operational Taxonomy Mapping to the NBDRA Components

| Activities | Description |
| --- | --- |
| System Orchestrator | |
| * Policy Enforcement * Security Metadata Model * Data Loss Prevention, Detection * Data 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 Application Provider. |
| Application Provider | |
| * Application Layer Identity * Web Services Gateway * Data Transformation * Digital Rights Management * Monitoring, Alerting | Application Provider interfaces between the Data Provider and Data Consumer. It takes part in all the secure interface protocols with these blocks as well as maintains secure interaction with the Framework Provider. |
| Framework Provider | |
| * Virtualization Layer Identity * Identity Provider * Encryption and Key Management * Isolation/Containerization * Storage Security * Network Boundary Control * Monitoring, Alerting | Framework Provider is responsible for the security of data/computations for a significant portion of the 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 Application Provider. |

## Mapping Security and Privacy Use Cases to the NBDRA

Subsection Scope: This section will contain a brief summary of the information in Appendix A (Full mapping of use cases to NBDRA). Possibly discuss what the mapping is, overall take away, and maybe run through the example use case.

## 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 for Version 2 of this document. However, the relationships between the Security and Privacy Fabric and the NBDRA and the Security and Privacy Taxonomy and the NBDRA will be investigated for Version 2 of this document.



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

The groups and interfaces depicted in Figure 6 are described below.

1. 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.

1. 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.

1. 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.

1. INTERNAL TO BIG DATA FRAMEWORK PROVIDER

Data at rest and transaction logs should be kept secured. Key management is essential to control access and keep track of keys. Non-relational databases should have a layer of security measures. Data provenance is essential to having proper context for security and function of the data at every stage. DoS attacks should be mitigated to assure availability of the data.

1. 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.

## Security and Privacy Fabric Principles

Subsection Scope:

Big Data security and privacy should leverage existing standards and practices. In the privacy arena, a systems approach that considers privacy throughout the process is a useful guideline to consider when adapting security and privacy practices to Big Data scenarios. The Organization for the Advancement of Structured Information Standards (OASIS) Privacy Management Reference Model (PMRM), consisting of seven foundational principles, provides appropriate basic guidance for Big System architects. 57,58 When working with any personal data, privacy should be an integral element in the design of a Big Data system.

Other privacy engineering frameworks, including the model presented in draft NISTIR 8062, Privacy Risk Management for Federal Information Systems, are also under consideration.59 60 61 62 63 64

Related principles include identity management frameworks such as proposed in the National Strategy for Trusted Identities in Cyberspace (NSTIC)65 and considered in the NIST Cloud Computing Security Reference Architecture.66 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.

### Related Fabric Concepts

Subsection Scope: This section cites related fabric concepts utilized elsewhere. []

## Security and Privacy Approaches in Analytics

Subsection scope: Section could include the following: Technology Usage Challenges – Expose personal behavior from IoT (home thermostat, home lightings, etc.), GPS navigation tools, self-driving vehicles, etc.; Legal and Social Challenges – examples could be from Census data, intrusion detection, criminal detection, etc.; Personal Financial Disclosure Challenges

[] Intro

### CRISP-DM Interop

Despite its widespread adoption for big data analytics, it has been criticized for its omission of domain-specific processes. For example, (Li, Zhang, & Tian, 2016) 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 DM methodology or process model for applications in the medical domain. . . “One of the few cited attempts provides extensions for CRISP-DM, but domain specificity is rare (Niaksu, 2015). 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 US 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 (Schaffer et al., 2015). []

## Cryptographic Technologies for Data Transformations

Subsection Scope: Discuss Cryptographic technologies. Discuss what can and can’t be solved with crypto (maybe reference other documents with more in depth discussions).

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 we must be aware that there are inherent limitations and dangers to such approaches.

### Classification

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. |

### 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 how much did she spend on movies last month? With Fully Homomorphic Encryption (FHE), a cloud can perform any computation on the underlying plaintext, all the while the results are encrypted. The cloud obtains no information about the plaintext or the results. [CSA]

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 [LMSV11].

In a breakthrough result [G09] in 2009, Gentry constructed the first fully homomorphic encryption scheme. Such a scheme allows one to compute the encryption of arbitrary functions of the underlying plaintext. Earlier results [BGN05] constructed partially homomorphic encryption schemes. Gentry’s original construction of a fully homomorphic encryption (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 [DGHV10, CMNT11], efficiency improvements [GHS12b, GHS12a] and partially homomorphic schemes [NLV11] that suffice for an interesting class of functions.

### 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 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! [CSA]

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 [18] construct a public key system that supports comparison queries, subset queries and arbitrary conjunction of such queries. In a recent paper [19], 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 data sets 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) [GGG+13] 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 [BGG+01], which to this date is prohibitively impractical. However, MIFE for important practical classes of functions such as vector inner products [BJK+15, DDM16, TAO16], equality and approximation testing [MR15] and order evaluation [BLN14] are known using practically available tools like elliptic curves and lattices.

### 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, (2) security of the same data in transit is a separate concern. [CSA]

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 [P05] [11] 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 and attribute based encryption methods enforce access control using cryptography. In identity-based systems [S84] (IBE), 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 [12] came up with the first IBE using pairing-friendly elliptic curves. Since then there have been numerous efficiency and security improvements [13] [14] [W09].

Attribute-based encryption (ABE) extends this concept to attribute-based access control. In [15] Sahai and Waters 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 [16]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 [17], with subsequent improvement by Waters [W11].

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 ICU” [LMS+12]. 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.

### 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 has a private input and there is a joint function 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 .

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 was given by [Yao82], with a long line of work described in the survey [JZ15]. This is a very active area of cryptography research and some practical implementations can be found in [MPCLib].

### 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 [Wikipedia]. 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.

### 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, like TPM (Trusted Platform Module) and SGX (Software Guard Extensions), provide a middle ground where the 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 TCB (Trusted Computing Base) 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) [Schuster et al] 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 (quoted from the paper):

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.
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, denial of service (DoS) attacks, side channels and traffic analyses are out of scope.

Pros:

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

Cons:

1. Secure code execution is susceptible to side-channel leakage like timing, electromagnetic and power analysis attacks.
2. 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 particular victim. However, a comprised CPU cannot be re-provisioned with a fresh key.

## Risk Management

To manage risk, NIST 800-39 recommends organizing risk across “three tiers of organization, mission/business processes, and information systems” (NIST, 2011). 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 cyber-physical system / IoT scenarios. []

### PII as Requiring Toxic Substance Handling

Subsection: Need text.

### Consent Withdrawal Scenarios

Subsection: Need text.

### Transparency Portal Scenarios

Subsection: Need text.

### Cross-organizational Risk Management

Subsection: Need text.

### Algorithm-Driven Issues

Subsection: Need text

### Big Data Forensics and Operational AAR

Subsection Scope: Need text. AAR = After Action Review. Describe why big data requires potentially unique preparation for forensics and AAR.

## Big Data Security Modeling and Simulation (ModSim)

Subsection: Needs references to NIST and other ModSim standards.

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,

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” [20]. 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.

### Safety Systems Modeling

Subsection Scope: Need text.

## Security and Privacy Management Phases

Earlier versions of this standard did not clarify design-time, in-situ and forensic (after-the-fact) considerations. This version explicitly addresses three phases for managing security and privacy in big data.

1. Build Phase The SnP 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 “green field” 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 (Ryoo, Kazman, & Anand, 2015).
   1. Note: This phase is roughly analogous to 800-53 Planning controls.
   2. 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” (Lea, 2015).
   3. At the build phase, delegated access management approaches should be designed in, using, for example two-way TLS, OAuth, OpenID, JSON web tokens, HMAC signing, NTLM, etc. Architects must consider compatibility with the big data stack of choice (e.g., some feel that
   4. 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” [21]. 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 Volume 6, “Management Fabric.”
3. Decommissioned Phase In its simplest form, this phase reflects a system that is no longer operational. For example, data from a (probably) decommissioned Radio Shack application 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. See also Gartner’s “Structured Data Archiving and Application Requirement” (Landers, Dayley, & Corriveau, 2016). This Phase also includes design for forensics analytics.

In addition to prior work by Ruan et al. (Ruan & Carthy, 2013) the Cloud Security Alliance proposed a Cloud Forensics Capability Maturity Model. []

#### Modifications for Agile Methodologies

Subsection scope: Need text

# 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.

* Identify domain-specific workflow
* Domain-specific roles
* Domain-specific “share” policies, content, controls.

Organizations (sole proprietorship) must identify which facets of big data systems are sharable and to whom. This is key to understanding how the BDSQ should be applied. For some, the domain model is not significantly different from the profession or industry; these are in some sense, “global” and non-proprietary. Other aspects of the domain model contain intellectual property, internal roles, execution strategy, branding, tools deployed, etc.; these are shared only selectively.

In the BDSQ, this is simplified to public and private “views” (Burger, 2014). 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 a BDRA component.

## Consent Management: Domain-Specific Big Data Security and Privacy

Subsection Scope: To illustrate domain-specific workflow, consider the healthcare HL7 FHIR example.

### Consent Management in Health Care

Subsection Scope: Need text

#### Relation to smart contracts

Subsection Scope: Need text

## Smart Building Domain Security

Subsection Scope: Discuss ISO standard here in one or two paragraphs.

# Provenance

Provenance in Big Data encompasses people and systems.

## *IoT Provenance*

Subsection Scope: Need text

### Traceability

Subsection Scope: Discuss big data traceability, especially for data sent to data lakes, or with automatically appended metadata. What relation to traditional data warehouse? Forensics (see NIST reference).

Representing a big data

* + 1. ***Security/Privacy Reasoning Support***

Subsection Scope: Need text

See Oasis STIX. SnP architectures which employ canonical security descriptions enable automated access to published vulnerabilities as well as emerging Security as a Service offerings. These are not defined in this document, but Big Data components from the RA are suggested for inclusion.

### Possible Roles for SnP Ontologies

Subsection Scope: New section. Insert Obrst discussion. Discuss relevance of continuous security and need for real time automated playbooks in response to insider threat and zero day scenarios.

### Domain-specific Provenance

Subsection Scope: New section. Explain why top conformance level integrates provenance that is mapped to a big data domain model.

# Audit and Configuration Management

Auditing fabric topology, including configuration management (CM) changes (taxonomic issues with configuration change data vs. audit data)

Audit and CM across organizational entities is only lightly covered in other standards. Planning for cross-organizational data transport is a big data concern, in particular:

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

### Packet-Level Traceability / Reproducibility

Subsection Scope: Packet-by-packet dump / restore capability is one approach to developing a Big Data SnP resource. TODO Flesh out

### Audit

Subsection Scope: Text could be enhanced.

Security Intelligence Event Management (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, an Amazon Web Services representative stated that its largest application at the time was its self-monitoring data used for management and billing support (insert briefing reference from Amazon []).

In recommendations released in 2006, NIST provided a set of recommendations for managing computer logs in order to preserve their integrity (Kent & Souppaya, 2006). Big Data presents additional challenges. []

In 2006, NIST also provided guidelines for “Integrating Forensic Techniques into Incident Response” (Kent, Chevalier, Grance, & Dang, 2006). Incident response for Big Data [].

### Big Data Audit and Monitoring

Subsection Scope: New section. Discuss suggestions from recent texts on Big Data monitoring in the Apache stack. Design challenges in building big data alert systems that follow good HCI principles and supports visualization where that can be shown to be efficacious.

# Workflow Models

Section Scope: Describe the workflow models. Opening paragraph to define what a workflow model is, list the workflow models to be discussed, and how they fit into the NBDRA.

Tools such as RunDeck []

Connect some aspects of orchestration to workflow a la BPMN, SysML.

Orchestration can encompass policy automation, such as identifying need-to-know defaults within the application domain model.

New text.

[] Walk through how workflow works in CloudMesh

## *Baseline Levels*

Subsection Scope: Text needed

In Version 1, no distinction was made between high and low compliance levels.

Facilitate scientific and engineering big data (generally low privacy risk without external data)

See NIST 800-53 Rev 4 (three levels) and/or an approach based on treaties

# Standards, Best Practices and Gaps

Section Scope: Discussion of standards related to the topics previously discussed in this document. Either the existing standards or the lack of a standard (i.e., gap) will be will be mentioned for a topic/technology/issue. To show what standards available addressing SnP and what standards NOT available from emerging technologies such as DevOps, etc.

## NIST Cybersecurity Framework

Sometime in 2017, NIST plans an “minor” update to the 2014 Cybersecurity Framework (Eric Chabrow, 2016). Since its introduction in 2014, the framework (NIST National Institute of Standards, 2014) 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 (US Dept of Homeland Security, 2015), NCF appears in numerous published HR position descriptions. Its appearance in cybersecurity hiring actions, as well as adaptation for other standards such as SAFSA SENC (Efrain Gonzalez, 2015) further reflect its importance.

## SABSA and Zachman Framework

Subsection Scope: Needs text.

[] Overview of implications for BDRA

## Configuration Management for Big Data

### Lineage Provenance

Subsection Scope: Needs text.

See AWS Config v. Config Rules http://goo.gl/exs5UB. []

In DoD systems, configuration management is seen as a key role, but the concept is less well integrated in the SDLC.

### Dependency Models

Subsection Scope: Needs text.

►Dependency Models that encompass software bills of resources

## Encryption Standards

[Arnab’s contributions; see outline for merge rules].

### Blockchain and Extensions

Subsection Scope: Needs text.

## Text Introducing Third Party Standards (Temporary)

Subsection Scope: Needs text.

These paragraphs will move.

## Big Data SDLC Standards and Guidelines

Subsection Scope: Objective is to identify the design pattern.

Today’s developers operate under SDLC frameworks including agile (Aydal, Paige, Chivers, & Brooke, 2006), waterfall (Iqbal & Rizwan, 2009) and spiral (Boehm, Lane, Koolmanojwong, & Turner, 2014). A significant number of developers operate under less explicit frameworks organized around GitHub practices. Draping a BDSQ on

### Big Data Security in DevOps

Subsection Scope: Text needs to be revised.

This version of the Big Data Security standard 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 (Roche, 2013), but it can reduce the importance of traditional checks and balances afforded by others in a larger organization.

A certain type of scalability is enabled by DevOps []

The notion of “Infrastructure as Code” is enabled by DevOps and other principally cloud computing technologies (Tom Nolle, 2016a).

The dilution, while not disappearance, of requirements phases and traceability in the agile development paradigm creates challenges for a security-aware SDLC. A “technology-agnostic” process termed Secure Development Life Cycle (SDL-IT) was developed at Microsoft to improve its management of security and privacy processes (Steer & Popli, 2008).

Big Data System SecDevOps

[] Brief sketch of how SecDevOps teams work in a BDRA. Discuss 12 factor and continuous security processes. Role of Ops and infrastructure. Reference to Adrian Cockroft.

#### Application Lifecycle Management

Both the application lifecycle and the data life cycle must be managed, though Big Data scenarios can see them delinked 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 lifecycle management” (Tom Nolle, 2016b).

#### Security and Privacy Events in Application Release Management

Subsection Scope: Opaque reference to new assessment domain at Gartner. Useful for DevOps and agile

#### Orchestration

Nolle insists that DevOps are orchestration are two different things 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” (Tom Nolle, 2016b). Noelle cites TOSCA (Qasha, Cala, & Watson, 2015) as leading this charge.

Additional Need: The following text Needs 3-3 sentences. Conformance or suggestions?

A Big Data adaptation of TOSCA-like concepts extends beyond cloud computing.

#### 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 APIs. Thus, the business model referred to “API-First” came into being (Chambakara, 2015).

API-first also addresses scalability challenges in domains such as healthcare. In the 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 BDRA, at the conceptual level, this specifies that application providers and consumers operate through defined APIs which can provide additional BDSQ controls.

#### Microservices

Subsection Scope: Needs text.

#### Software Security and Reliability in DevOps

Subsection Scope: Needs text. Discussion of IEEE P2675 and related standards. Connections to Big Data concepts only.

### Model Driven Development

Big Data systems potentially entail multiple models from multiple disciplines implemented across diverse platforms, often across different organizations. Previous attempts to share information across organizations have not fared well. Sharing of database schemas

#### Add SI discussion []

Subsection Scope: Needs text.

#### Add Smart Building Examples []

Subsection Scope: Needs text.

#### Metamodel Processes in Support of BD SnP

Subsection Scope: Needs text.

A BDSQ []

An approach taken by Atkinson et al. (Atkinson, Stoll, & Bostan, 2010) 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.”

[]

#### Cite security ontology work @ Florida

Subsection Scope: Needs text

#### Cite work on Authorization Languages and Contextual Integrity

Subsection Scope: Needs text.

### Other Standards Through a Big Data Lens

#### ISO 21827:2008 and SSE-CMM

Subsection Scope: Needs text.

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-90s; it predated “big data.”

[] Refresh!

#### ISO 12207 and ISO 15504

Subsection Scope: Needs text.

#### Process Specifications

Subsection Scope: Needs text. Review of domain-specific as well as cross-cutting process specifications for security/privacy processes.

##### PSL ISO 18629

Subsection Scope: Needs text. Process specification language.

#### ISO 27018

Subsection Scope: Needs text. *The following text from Mark U. Must be curated; may infringe. Walk through impact on the NBDRA.*

Consent: CSPs 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](https://d0.awsstatic.com/certifications/iso_27018_certification.pdf)) 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

### SnP Quilts for Specific SDLC Methodologies

Subsection Scope: Needs text.

Each SDLC approach calls for []

Architectures designed using techniques such as Business Model Canvas (Osterwalder & Pigneur, 2010) or Lean Canvas (Maurya, 2012) incorporate models which

### 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. Speaking of data generated by National Instruments and its broader ecosystem, a VP of product marketing estimated that some 33 exabytes of data had been generated to date. In the same report, The powertrain simulation and tools research leader at Jaguar Land Rover estimated that it was generated about 500GB daily (Nelson, 2015).

A fraction of this data is directly relevant to SnP, but even at 1%, this represents a daunting challenge. []

### API-First and Microservices

The notion of microservices has evolved from SoA and object-oriented practices, but is relevant to Big Data because it represents a convergence of several trends. A recent NIST draft 800-180 attempts to put forth a standard definition (Karmel, Chandramouli, & Iorga, 2016). 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 (Newman, 2015), but microservices offer applications designers, data center managers and forensics specialists greater detail over relevant big data 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 (Versteden, Pauwels, & Papantoniou, 2015). []

### Application Security for Big Data

#### RBAC, ABAC and Workflow

Initial work by NIST evolved to an ANSI / INCITS standard 369-2004 for RBAC (INCITS, 2004). 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” (Kuhn, Coyne, & Weil, 2010). 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.

##### Hybrid RBAC / ABAC

Subsection Scope: text needs to be revised.

In 2015, NIST published a description of ABAC in SP-800-162 (Hu et al., 2014).

Beyond RBAC improvements, Big Data systems must incorporate workflow standards – if not formalisms, in order 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 (Ferraiolo, Chandramouli, Ahn, & Gavrila, 2003), or to include geospatial attributes (Damiani, Bertino, Catania, & Perlasca, 2007).

For Big Data systems, []

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, “. . .the current RBAC profile does not provide any support for many relevant constraints, such as static and dynamic separation of duty.” Ferrini and Bertino recommended expanding the XACML framework to include OWL (Ferrini & Bertino, 2009). 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. [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 (Sun, Meng, Liu, & Pan, 2005). Workflow models [discussed in Section XXX] can further

Why is this a big data issue? Because as the

#### ‘Least Exposure’ Big Data Practices

Just as legacy and software keyfobs 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 data sets with temporally restricted access. [References needed] []

#### Logging

##### NIST Logging Standards

###### NIST SP 800-92

Subsection Scope: Needs text.

###### NIST SP 800-137

Subsection Scope: Needs text.

###### DevOps Logging

Subsection Scope: Needs text.

###### Citation: The Art of Monitoring (Turnbull, 2016),

Subsection Scope: Needs text.

#### Ethics and Privacy by Design

##### IEEE P7000

Subsection Scope: Needs text.

##### NIST IR 8062

Subsection Scope: Needs text.

## Big Data Governance

### Apache Atlas

Subsection Scope: Needs text.

[Apache Atlas](http://atlas.incubator.apache.org/) is in incubation as of this writing, but aims to address compliance and governance needs for Big Data applications using Hadoop

### GSA DevOps Open Compliance

Subsection Scope: Needs text.

[It’s actually named something else.]

## Infrastructure Management

### Infrastructure as Code

Subsection Scope: Needs text.

### Particular Issues with Hybrid and Private Cloud

Subsection Scope: Needs text. Review and cite the CSCC Hybrid Cloud Security document.

TBD – This is an area where standards coverage is discontinuous with initiatives of Cloud Security Alliance and CSCC.

### Relevance to NIST Critical Infrastructure

Subsection Scope: Needs text.

See <https://www.nist.gov/sites/default/files/documents/cyberframework/cybersecurity-framework-021214.pdf>

## Emerging Technologies

### 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 [Wikipedia]. 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.

### DevOps Automation

#### Application Release Automation

Subsection Scope: Needs text.

Industry example: XebiaLabs

### Network Security for Big Data

#### Virtual Machines and SDN

Subsection Scope: Needs text and revision of included notes.

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 (Chandramouli, 2016). Virtual machine security also figures in PCI guidelines (PCI Security Standards Council, 2011).

Cite IEEE P1915.1 []

NIST 800-125 addresses []

Potential advantages

#### Architecture Standards for IoT

Subsection Scope: Needs text.

IEEE P2413

### Machine Learning, AI and Analytics for Big Data Security and Privacy

Subsection Scope: Possibly incorporate use case or conclusions from Medicare End-Stage Renal Disease, Dialysis Facility Compare  (ESRD DFC) <http://data.medicare.gov/data/dialysis-facility-compare> (Liu, CJ)

#### Overview of emerging technologies

Subsection Scope: Needs text.

#### Risk / opportunity areas for enterprises

Subsection Scope: Needs text.

#### Risk / opportunity areas for consumers

Subsection Scope: Needs text.

#### Risk / opportunities for government

Subsection Scope: Needs text.

# Conclusions

This section will be written at a later date.

While Big Data as a concept can drift toward the nebulous, big data risks to security and privacy are tangible and well reported.

Editorial note: Some NIST reports have conclusions that can be summarized (e.g., see this summary of the NIST Cloud Computing Standards roadmap).

1. 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.

* 1. Retail/Marketing
     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 A-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. [11] Secure/Multipurpose Internet Mail Extensions (S/MIME) |
| Real-time security monitoring | Content creation security |
| Data discovery and classification | Discovery/classification is possible across media, populations, and channels. |
| Secure data aggregation | Vendor-supplied aggregation services—security practices are opaque. |
| Application Provider → Data Consumer | Privacy-preserving data analytics | Aggregate reporting to content owners |
| Compliance with regulations | PII disclosure issues abound |
| Government access to data and freedom of expression concerns | Various issues; for example, playing terrorist podcast and illegal playback |
| Data Provider ↔  Framework Provider | Data-centric security such as identity/policy-based encryption | Unknown |
| Policy management for access control | User, playback administrator, library maintenance, and auditor |
| Computing on the encrypted data: searching/ filtering/ deduplicate/ fully homomorphic encryption | Unknown |
| Audits | Audit DRM usage for royalties |
| Framework Provider | Securing data storage and transaction logs | Unknown |
| Key management | Unknown |
| Security best practices for non-relational data stores | Unknown |
| Security against DoS attacks | N/A |
| Data provenance | Traceability to data owners, producers, consumers is preserved |
| Fabric | Analytics for security intelligence | Machine intelligence for unsanctioned use/access |
| Event detection | “Playback” granularity defined |
| Forensics | Subpoena of playback records in legal disputes |

* + 1. Nielsen Homescan: Project Apollo

Nielsen Homescan involves family-level retail transactions and associated media exposure using a statistically valid national sample. A general description [12] 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 A-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 (Extensible Markup Language) cube security mapped to Sybase IQ and reporting tools |
| Policy management for access control | Extensive role-based controls |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | N/A |
| Audits | Schematron and process step audits |
| Framework Provider | Securing data storage and transaction logs | Project-specific audits secured by infrastructure team. |
| Key management | Managed by project chief security officer (CSO). Separate key pairs issued for customers and internal users. |
| Security best practices for non-relational data stores | Regular data integrity checks via XML schema validation |
| Security against DoS attacks | Industry-standard webhost protection provided for query subsystem. |
| Data provenance | Unique |
| Fabric | Analytics for security intelligence | No project-specific initiatives |
| Event detection | N/A |
| Forensics | Usage, cube-creation, and device merge audit records were retained for forensics and billing |

* + 1. 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 A-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/fully homomorphic encryption | Unknown |
| Audits | Customer audits for accuracy and integrity are supported |
| Framework Provider | Securing data storage and transaction logs | Storage archiving—this is a big issue |
| Key management | CSO and applications |
| Security best practices for non-relational data stores | Unknown |
| Security against DoS attacks | Standard |
| Data provenance | Server, application, IP-like identity, page point-in-time Document Object Model (DOM), and point-in-time marketing events |
| Fabric | Analytics for security intelligence | Access to web logs often requires privilege elevation. |
| Event detection | Can infer; for example, numerous sales, marketing, and overall web health events |
| Forensics | See the SIEM use case |

* 1. Healthcare
     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 A-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 validation | Strong authentication, perhaps through X.509v3 certificates, potential leverage of SAFE (Signatures & Authentication for Everything [13]) bridge in lieu of general PKI |
| Real-time security monitoring | Validation of incoming records to assure integrity through signature validation and to assure HIPAA privacy through ensuring PHI is encrypted. May need to check for evidence of informed consent. |
| Data discovery and classification | Leverage Health Level Seven (HL7) and other standard formats opportunistically, but avoid attempts at schema normalization. Some columns will be strongly encrypted while others will be specially encrypted (or associated with cryptographic metadata) for enabling discovery and classification. May need to perform column filtering based on the policies of the data source or the HIE service provider. |
| Secure data aggregation | 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. [14] |
| 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 US 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/fully homomorphic encryption | Privacy-preserving access to relevant events, anomalies, and trends for CDC and other relevant health organizations |
| Audits | Facilitate HIPAA readiness and HHS audits |
| Framework Provider | Securing data storage and transaction logs | Need to be protected for integrity and privacy, but also for establishing completeness, with an emphasis on availability. |
| Key management | Federated across covered entities, with the need to manage key life cycles across multiple covered entities that are data sources |
| Security best practices for non-relational data stores | End-to-end encryption, with scenario-specific schemes that respect min-entropy to provide richer query operations without compromising patient privacy |
| Security against distributed denial of service (DDoS) attacks | A mandatory requirement: systems must survive DDoS attacks |
| Data provenance | Completeness and integrity of data with records of all accesses and modifications. This information could be as sensitive as the data and is subject to commensurate access policies. |
| Fabric | Analytics for security intelligence | Monitoring of informed patient consent, authorized and unauthorized transfers, and accesses and modifications |
| Event detection | Transfer of record custody, addition/modification of record (or cell), authorized queries, unauthorized queries, and modification attempts |
| Forensics | Tamper-resistant logs, with evidence of tampering events. Ability to identify record-level transfers of custody and cell-level access or modification |

* + 1. Genetic Privacy

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

* + 1. 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 A-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 |
| Data Provider ↔  Framework Provider | Data-centric security such as identity/policy-based encryption | TBD |
| Policy management for access control | Internal roles; third-party custodian roles; researcher roles; participating patients’ physicians |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | TBD |
| Audits | Release audit by a third party |
| Framework Provider | Securing data storage and transaction logs | TBD |
| Key management | Internal varies by firm; external TBD |
| Security best practices for non-relational data stores | TBD |
| Security against DoS attacks | Unlikely to become public |
| Data provenance | TBD—critical issue |
| Fabric | Analytics for security intelligence | TBD |
| Event detection | TBD |
| Forensics |  |

* 1. Cybersecurity
     1. Network Protection

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

Table A-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. [15] |
| 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 ↔  Framework Provider | Data-centric security such as identity/policy-based encryption | Usually a feature of the operating system |
| Policy management for access control | For example, a group policy for an event log |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | Vendor and platform-specific |
| Audits | Complex—audits are possible throughout |
| Framework Provider | Securing data storage and transaction logs | Vendor and platform-specific |
| Key management | Chief Security Officer and SIEM product keys |
| Security best practices for non-relational data stores | TBD |
| Security against DDoS attacks | Big Data application layer DDoS attacks can be mitigated using combinations of traffic analytics, correlation analysis. |
| Data provenance | For example, how to know an intrusion record was actually associated with a specific workstation. |
| Fabric | Analytics for security intelligence | Feature of current SIEMs |
| Event detection | Feature of current SIEMs |
| Forensics | Feature of current SIEMs |

* 1. Government
     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 A-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 |
| Secure data aggregation | Fusion challenges range from simple to complex. Video streams may be used [16] 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/fully homomorphic encryption | Sometimes performed within vendor-supplied architectures, or by image-processing parallel architectures |
| Audits | CSO and Inspector General (IG) audits |
| Framework Provider | Securing data storage and transaction logs | The usual, plus data center security levels are tightly managed (e.g., field vs. battalion vs. headquarters) |
| Key management | CSO—chain of command |
| Security best practices for non-relational data stores | Not handled differently at present; this is changing. E.g., see the DoD Cloud Computing Strategy (July 2012). [17] |
| 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. [18] |
| Forensics | Used for after action review (AAR)—desirable to have full playback of sensor streams |

* + 1. 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 A-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/fully homomorphic encryption | Proposed [19] |
| 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 |

* 1. Industrial: Aviation
     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.

* 1. Transportation
     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 A-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. [20] |
| Data Provider ↔  Framework Provider | Data-centric security such as identity/policy-based encryption | --- |
| Policy management for access control | Private, sensitive sensor data and package data should only be available to authorized individuals. Third-party commercial offerings may implement low-level access to the data. |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | See above section on “Transformation.” |
| Audits | --- |
| Framework Provider | Securing data storage and transaction logs | Logging sensor data is essential for tracking packages. Sensor data at rest should be kept in secure data stores. |
| Key management | For encrypted data |
| Security best practices for non-relational data stores | The diversity of sensor types and data types may necessitate the use of non-relational data stores |
| Security against DoS attacks | --- |
| Data provenance | Metadata should be cryptographically attached to the collected data so that the integrity of origin and progress can be assured. Complete preservation of provenance will sometimes mandate a separate Big Data application. |
| Fabric | Analytics for security intelligence | Anomalies in sensor data can indicate tampering/fraudulent insertion of data traffic. |
| 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. |

* 1. New Use Cases

Subsection Scope: The Use cases that are new in Version 2, could be mapped here

* + 1. Major Use Case : SEC Consolidated Audit Trail
    2. Major Use Case: IoT Device Management
    3. Major Use Case: OMG Data Residency initiative
    4. Minor Use Case: TBD
    5. Use Case: Emergency management data (XChangeCore interoperability standard ).
    6. Major Use Case: Health care consent flow
    7. Major Use Case: “HEART Use Case: Alice Selectively Shares Health-Related Data with Physicians and Others”
    8. Major Use Case Blockchain for FinTech (Arnab)
    9. Minor Use Case – In-stream PII
    10. Major Use Case – Statewide Education Data Portal

1. Internal Security Considerations within Cloud Ecosystems

Many Big Data systems will be designed using cloud architectures. Any strategy to implement a mature security and privacy framework within a Big Data cloud ecosystem enterprise architecture must address the complexities associated with cloud-specific security requirements triggered by the cloud characteristics. These requirements could include the following:

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

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

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

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

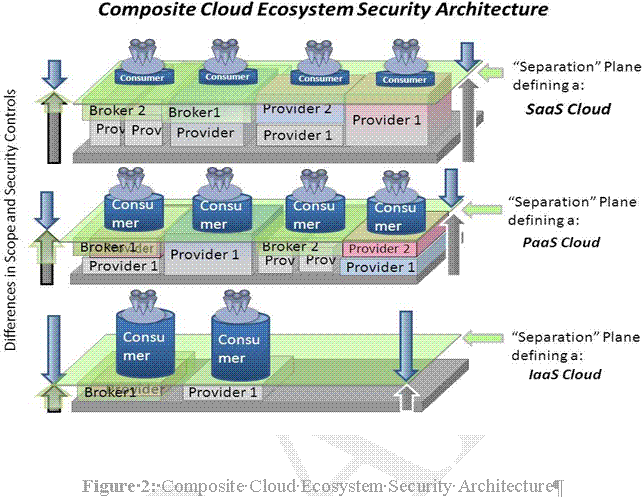


Figure B-1: Composite Cloud Ecosystem Security Architecture [21]

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:
   1. GRC; and
   2. Directives, policy guidelines, strategic goals and objectives, information security requirements, priorities, and resources available (filling in any gaps).
2. Select the appropriate level of security controls required to protect data and to defend information systems.
3. Implement access security controls and modify them upon analysis assessments.
4. Authorize appropriate information systems.
5. Monitor access security controls at a minimum of once a year.

To meet GRC and 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

A brief statement of Cloud Computing Related Standards will be included here to introduce Table B-1, which is from NIST SP 800-144 document.

Table B-1: Standards and Guides Relevant to Cloud Computing [2]

|  |  |
| --- | --- |
| 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 |
| SP 800-18 | Guide for Developing Security Plans for Federal Information Systems |
| SP 800-34, Revision 1 | Contingency Planning Guide for Federal Information Systems |
| SP 800-37, Revision 1 | Guide for Applying the Risk Management Framework to Federal Information Systems |
| SP 800-39 | Managing Information Security Risk |
| SP 800-53, Revision 3 | Recommended Security Controls for Federal Information Systems and Organizations |
| SP 800-53, Appendix J | Privacy Control Catalog |
| SP 800-53A, Revision 1 | Guide for Assessing the Security Controls in Federal Information Systems |
| SP 800-60 | Guide for Mapping Types of Information and Information Systems to Security Categories |
| SP 800-61, Revision 1 | Computer Security Incident Handling Guide |
| SP 800-64, Revision 2 | Security Considerations in the System Development Life Cycle |
| SP 800-86 | Guide to Integrating Forensic Techniques into Incident Response |
| SP 800-88 | Guidelines for Media Sanitization |
| SP 800-115 | Technical Guide to Information Security Testing and Assessment |
| SP 800-122 | Guide to Protecting the Confidentiality of Personally Identifiable Information (PII) |
| 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.[[9]](#footnote-9) This framework for Big Data may be adopted during the future work of the NBD-PWG.

Access Control

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

Access control for Big Data covers more than accessing data. Data can be accessed via multiple channels, networks, and platforms—including laptops, cell phones, 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 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, [23] 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, payment card industry [PCI]), then all of the data that it is sent with it inherits that classification, retaining the original data’s security classification. If the data is joined to and/or associated with other data that may cause a privacy issue, then all data should be protected. This requires due diligence on the part of the data custodian(s) to ensure that this secure and protected state remains throughout the entire end-to-end data flow. Variations on this theme may be required for domain-specific combinations of public and private data hosted by Big Data applications.
* If data is accessed from, transferred to, or transmitted to the cloud, Internet, or another external entity, then the data should be protected based on its classification.
* There should be an indicator/disclaimer on the display of the user if private or sensitive data is being accessed or viewed. Openness, trust, and transparency considerations may require more specific actions, depending on GRC or other broad considerations of how the Big Data system is being used.
* All system roles (“accounts”) should be subjected to periodic meaningful audits to check that they are still required.
* All accounts (except for system-related accounts) that have not been used within 180 days should be deactivated.
* Access to PII data should be logged. Role-based access to Big Data should be enforced. Each role should be assigned the fewest privileges needed to perform the functions of that role.
* Roles should be reviewed periodically to check that they are still valid and that the accounts assigned to them are still appropriate.

User Access Controls

* Each user should have 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.

1. Big Data Actors and Roles: Adaptation to Big Data Scenarios

Section information: This appendix will be edited to discuss hybrid- and access-based security.

Service-oriented architectures (SOA) 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. [24] [25] 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. [26] 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”).

1. Acronyms

The acronym list will be updated when the text has been finalized.

AC&S access control and security

ACL Access Control List

AuthN/AuthZ Authentication/Authorization

BAA business associate agreement

CDC U.S. Centers for Disease Control and Prevention

CEP complex event processing

CIA confidentiality, integrity, and availability

CINDER DARPA Cyber-Insider Threat

CoP communities of practice

CSA Cloud Security Alliance

CSA BDWG Cloud Security Alliance Big Data Working Group

CSP Cloud Service Provider

DARPA Defense Advanced Research Projects Agency’s

DDoS distributed denial of service

DOD U.S. Department of Defense

DoS denial of service

DRM digital rights management

EFPIA European Federation of Pharmaceutical Industries and Associations

EHR electronic health record

EU European Union

FBI U.S. Federal Bureau of Investigation

FTC Federal Trade Commission

GPS global positioning system

GRC governance, risk management, and compliance

HIE Health Information Exchange

HIPAA Health Insurance Portability and Accountability Act

HITECH Act Health Information Technology for Economic and Clinical Health Act

HR human resources

IdP identity provider

IoT Internet of Things

IP Internet Protocol

IT information technology

LHNCBC Lister Hill National Center for Biomedical Communications

M2M machine to machine

MAC media access control

NBD-PWG NIST Big Data Public Working Group

NBDRA NIST Big Data Reference Architecture

NIEM National Information Exchange Model

NIST National Institute of Standards and Technology

OSS operations systems support

PaaS platform as a service

PHI protected health information

PII personally identifiable information

PKI public key infrastructure

SAML Security Assertion Markup Language

SDLC Systems Development Life Cycle

SIEM security information and event management

SKU stock keeping unit

SLA service-level agreement

STS Security Token Service

TLS Transport Layer Security

VM virtual machine

VPN virtual private network

XACML eXtensible Access Control Markup Language

1. References

This reference section needs to be consolidated, linked to text, and formatted.

REFERENCES FROM MARK’S VERSION 2 DOCUMENT. LAURIE TO RESOLVE.

ACM. (n.d.). The 2012 ACM Computing Classification System. Retrieved August 14, 2016, from https://www.acm.org/publications/class-2012

Atkinson, C., Stoll, D., & Bostan, P. (2010). Orthographic Software Modeling: A Practical Approach to View-Based Development BT - Evaluation of Novel Approaches to Software Engineering: 3rd and 4th International Conferences, ENASE 2008/2009, Funchal, Madeira, Portugal, May 4-7, 2008 / Milan, Italy, M. In L. A. Maciaszek, C. González-Pérez, & S. Jablonski (Eds.) (pp. 206–219). CHAP, Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-14819-4\_15

Aydal, E., Paige, R., Chivers, H., & Brooke, P. (2006). Security Planning and Refactoring in Extreme Programming. In P. Abrahamsson, M. Marchesi, & G. Succi (Eds.), *Extreme Programming and Agile Processes in Software Engineering SE - Lecture Notes in Computer Science* (Vol. 4044, pp. 154–163). CHAP, Berlin, Heidelberg: Springer Berlin / Heidelberg. http://doi.org/doi: 10.1007/11774129\_16

Boehm, B., Lane, J., Koolmanojwong, S., & Turner, R. (2014). *The Incremental Commitment Spiral Model: Principles and Practices for Successful Systems and Software*. BOOK, Addison-Wesley Professional.

Burger, E. (2014). *Flexible views for view-based model-driven development*. *Flexible Views for View-Based Model-Driven Development*. BOOK, Karlsruhe. Deutschland: KIT Scientific Publishing. http://doi.org/10.5445/KSP/1000043437

Celesti, A., Fazio, M., & Villari, M. (2013). SE CLEVER: A secure message oriented Middleware for Cloud federation. In *2013 IEEE Symposium on Computers and Communications (ISCC)* (pp. 35–40). CONF, IEEE. http://doi.org/doi: 10.1109/ISCC.2013.6754919

Chambakara, P. (2015, November 30). API-First Design: Dawn Of New Era In App Development. *Digital Doughnut*. INPR, London, UK: Digital Doughnut. Retrieved from citeulike-article-id:14074448

Chandramouli, R. (2016). *NIST Special Publication 800-125B Secure Virtual Network Configuration for Virtual Machine (VM) Protection*. Gaithersburg MD. http://doi.org/10.6028/NIST.SP.800-125B

Cheikes, B. A. (2015). *Forming Common Platform Enumeration ( CPE ) Names from Software Identification ( SWID ) Tags Forming Common Platform Enumeration ( CPE ) Names from Software Identification ( SWID ) Tags* (Vol. 8085). Gaithersburg, MD. Retrieved from http://csrc.nist.gov/publications/drafts/nistir-8085/nistir\_8085\_draft.pdf

CIO\_Council. (2012). *Recommendations for standardized implementation of digital privacy controls* (RPRT). Washington, DC: Federal Chief Information Officers Council. Retrieved from https://cio.gov/wp-content/uploads/downloads/2012/12/Standardized\_Digital\_Privacy\_Controls.pdf

Damiani, M., Bertino, E., Catania, B., & Perlasca, P. (2007). GEO-RBAC: A spatially aware RBAC. *ACM Trans. Inf. Syst. Secur.*, *10*(1), 2. JOUR. http://doi.org/doi: 10.1145/1210263.1210265

Das, P., Joshi, A., & Finin, T. (2016). Capturing policies for fine-grained access control on mobile devices. In *2016 IEEE 2nd International Conference on Collaboration and Internet Computing*. CONF, Piscataway NJ: IEEE. Retrieved from citeulike-article-id:14166569

Dhall, R. (2016). Performance Patterns in Microservices based Integrations. *Computing Now*. JOUR. Retrieved from citeulike-article-id:14166626

DISA. (2015). *DEPARTMENT OF DEFENSE (DoD) Cloud Computing Security Requirements Guide (SRG)* (RPRT). Fort Meade, MD: DISA. Retrieved from http://iase.disa.mil/cloud\_security/Documents/u-cloud\_computing\_srg\_v1r1\_final.pdf

Efrain Gonzalez. (2015). SENC Project: SABSA Enhanced NIST Cybersecurity Framework | sabsa.org. Retrieved August 15, 2015, from http://www.sabsa.org/node/176

Eric Chabrow. (2016). NIST Plans Cybersecurity Framework Update - GovInfoSecurity. Retrieved August 14, 2016, from http://www.govinfosecurity.com/interviews/nist-considers-cybersecurity-framework-update-i-3199#.V1jIbRyMY7E.twitter

ETSI. (2013). Cloud Standards Coordination Final Report.

Fazio, M., & Puliafito, A. (2011). Virtual Resource Management Based on Software Transactional Memory. In *Network Cloud Computing and Applications (NCCA), 2011 First International Symposium on* (pp. 1–8). CONF, Tempe, AZ: IEEE. http://doi.org/doi: 10.1109/NCCA.2011.8

Ferraiolo, D., Chandramouli, R., Ahn, G.-J., & Gavrila, S. (2003). The role control center: features and case studies. In *Proceedings of the eighth ACM symposium on Access control models and technologies* (p. 12–20 OR–ACM). CONF. Retrieved from citeulike-article-id:14114063

Ferrini, R., & Bertino, E. (2009). Supporting RBAC with XACML+OWL. In *Proceedings of the 14th ACM symposium on Access control models and technologies SE - SACMAT ’09* (pp. 145–154). CONF, New York, NY, USA: ACM. http://doi.org/doi: 10.1145/1542207.1542231

Frederic Guyomarc’h, Dekeyser, J.-L., & A. Wendell O. Rodrigues. (2013). An MDE Approach for Automatic Code Generation from UML/MARTE to OpenCL. *Computing in Science and Engineering*, *15*(1), 46–55. JOUR. http://doi.org/doi: http://doi.ieeecomputersociety.org/10.1109/MCSE.2012.35

Hickson, I. (2013). HTML Microdata. *W3C Working Group Note 29*. Retrieved from https://www.w3.org/TR/microdata/

Hickson, I., Kellogg, G., Tenisson, J., & Herman, I. (2014). Microdata to RDF – Second Edition. *W3C*. W3C. Retrieved from http://www.w3.org/TR/microdata-rdf/

Hu, V. C., Ferraiolo, D., Kuhn, R., Schnitzer, A., Sandlin, K., Miller, R., & Scarfone, K. (2014). Guide to attribute based access control (abac) definition and considerations. *NIST Special Publication*, *800*, 162. http://doi.org/10.6028/NIST.SP.800-162

INCITS. (2004). *ANSI INCITS 359-2004 Role Based Access Control Information Technology Industry Council*. Washington DC. Retrieved from http://profsandhu.com/journals/tissec/ANSI+INCITS+359-2004.pdf

Iqbal, M., & Rizwan, M. (2009). Application of 80/20 rule in software engineering Waterfall Model. In *2009 International Conference on Information and Communication Technologies* (pp. 223–228). CONF, Karachi, Pakistan: IEEE. http://doi.org/doi: 10.1109/icict.2009.5267186

Jansen, W., Grance, T., & Mell, P. (2011). Guidelines on Security and Privacy in Public Cloud Computing. *Director*, *144*(7), 800–144. http://doi.org/10.3233/GOV-2011-0271

Karmel, A., Chandramouli, R., & Iorga, M. (2016). *DRAFT Special Publication 800-180, NIST Definition of Microservices, Application Containers and System Virtual Machines*. *NIST Special Publication 800-180* (Vol. 800180). Gaithersburg, MD. Retrieved from http://csrc.nist.gov/publications/drafts/800-180/sp800-180\_draft.pdf

Kauffman\_Foundation. (2016, September). Welcome to EdWise - Education Data for Missouri. ELEC, Kansas City, MO: Kauffman Foundation. Retrieved from citeulike-article-id:14169722

Kent, K., Chevalier, S., Grance, T., & Dang, H. (2006). *Guide to Integrating Forensic Techniques into Incident Response Recommendations of the National Institute of Standards and Technology*. *Special Publication 800-86*. Gaitherburg, MD. Retrieved from http://csrc.nist.gov/publications/nistpubs/800-86/SP800-86.pdf

Kent, K., & Souppaya, M. (2006). *Guide to Computer Security Log Management Recommendations of the National Institute of Standards and Technology*. Gaithersburg, MD. Retrieved from http://csrc.nist.gov/publications/nistpubs/800-92/SP800-92.pdf

Kuhn, D. R., Coyne, E. J., & Weil, T. R. (2010). Adding Attributes to Role-Based Access Control. *Computer*, *43*(6), 79–81. http://doi.org/10.1109/MC.2010.155

Landers, G., Dayley, A., & Corriveau, J. (2016). Magic Quadrant for Structured Data Archiving and Application Retirement. Retrieved August 10, 2016, from https://www.gartner.com/doc/reprints?id=1-39B7753&ct=160613&st=sb

Lea, G. (2015, March 2). Notes from YOW! 2014: Scott Shaw on “Avoiding Speedbumps on the Road to Microservices.” ELEC, Graham Lea. Retrieved from citeulike-article-id:14169875

Lenz, K., & Oberweis, A. (2003). Inter-organizational Business Process Management with XML Nets. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 2472, pp. 243–263). Springer. http://doi.org/10.1007/978-3-540-40022-6\_12

Li, J.-S., Zhang, Y.-F., & Tian, Y. (2016). Medical Big Data Analysis in Hospital Information System. In *Big Data on Real-World Applications*. CHAP, Rijeka, Croatia: InTech. Retrieved from http://www.intechopen.com/books/big-data-on-real-world-applications/medical-big-data-analysis-in-hospital-information-system

Lin, X., Zhang, M., Zhao, H., & J., B. (2012). Multi-view of the ACM classification system. In *12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL ’12* (pp. 397–398). http://doi.org/10.1145/2232817.2232909

Maurya, A. (2012). *Running Lean: Iterate from Plan A to a Plan That Works (Lean Series)*. BOOK, Sebastopol CA: O’Reilly Media. Retrieved from http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&amp

Miles, A. (University of O., & Bechhofer, S. (University of M. (2009). SKOS Simple Knowledge Organization System Reference. Retrieved August 12, 2016, from https://www.w3.org/TR/skos-reference/

Mirkin, B., Nascimento, S., & Pereira, L. M. (2008). Representing a Computer Science Research Organization on the ACM Computing Classification System 1 ACM Computing Classification System Fits for Representing CS Research Activities. In P. Elkund & H. Ollivier (Eds.), *Supplementary Proceedings of the 16th International Conference on Conceptual Structures*. Toulouse, France: CEUR. Retrieved from http://ceur-ws.org/Vol-354/p19.pdf

Nelson, R. (2015). Big data analytics becomes strategic test tool. *Evaluation Engineering*. JOUR. Retrieved from citeulike-article-id:14169892

Newman, S. (2015). Building microservices : designing fine-grained systems. ELEC, Sebastopol CA: O’Reilly Media. Retrieved from http://www.worldcat.org/isbn/9781491950357

Niaksu, O. (2015). CRISP Data Mining Methodology Extension for Medical Domain. *Baltic Journal of Modern Computing*, *3*(2), 92–109. JOUR. Retrieved from http://www.bjmc.lu.lv/fileadmin/user\_upload/lu\_portal/projekti/bjmc/Contents/3\_2\_2\_Niaksu.pdf

Nicol, D. M. (2005). Modeling and simulation in security evaluation. *Security & Privacy, IEEE*, *3*(5), 71–74. JOUR. http://doi.org/doi: 10.1109/msp.2005.129

NIST. (2011). *NIST Special Publication 800-39, Managing Information Security Risk Organization, Mission, and Information System View*. *Nist Special Publication*. Gaitherburg, MD. Retrieved from http://csrc.nist.gov/publications/nistpubs/800-39/SP800-39-final.pdf

NIST National Institute of Standards, N. (2014). Framework for Improving Critical Infrastructure Cybersecurity, Version 1.0, National Institute of Standards and Technology, February 12, 2014. Retrieved from http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf

Obrst, L., Chase, P., & Markeloff, R. (2012). Developing an Ontology of the Cyber Security Domain. In P. Laskey & K. Laskey (Eds.), *Proceedings of the Seventh International Conference on Semantic Technologies for Intelligence, Defense, and Security* (pp. 49–56). CONF, Fairfax VA OR - CEUR. Retrieved from citeulike-article-id:14113899

Osterwalder, A., & Pigneur, Y. (2010). *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*. BOOK, Wiley. Retrieved from http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&amp

PCI Security Standards Council, V. S. I. G. (2011). *PCI DSS Virtualization Guidelines Standard: PCI Data Security Standard (PCI DSS) Virtualization Special Interest Group PCI Security Standards Council*. Wakefield, MA. Retrieved from https://www.pcisecuritystandards.org/documents/Virtualization\_InfoSupp\_v2.pdf

Proud-Madruga, D. (2016, May 9). Project Summary for Privacy, Access and Security Services (PASS) Healthcare Audit Services Conceptual Model. ELEC, Ann Arbor, MI OR - HL7: HL7. Retrieved from https://www.hl7.org/special/Committees/projman/searchableProjectIndex.cfm?action=edit&#38

Qasha, R., Cala, J., & Watson, P. (2015). Towards Automated Workflow Deployment in the Cloud Using TOSCA. In *Cloud Computing (CLOUD), 2015 IEEE 8th International Conference on* (pp. 1037–1040). CONF, IEEE. http://doi.org/doi: 10.1109/cloud.2015.146

Roche, J. (2013). Adopting DevOps Practices in Quality Assurance. *Queue*, *11*(9). JOUR. http://doi.org/doi: 10.1145/2538031.2540984

Ruan, K., & Carthy, J. (2013). Cloud Forensic Maturity Model. In M. Rogers & K. C. Seigfried-Spellar (Eds.), *Digital Forensics and Cyber Crime* (pp. 22–41). CHAP, Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-39891-9\_2

Ryoo, J., Kazman, R., & Anand, P. (2015). Architectural Analysis for Security. *IEEE Security & Privacy*, *13*(6), 52–59. JOUR. http://doi.org/doi: doi.ieeecomputersociety.org/10.1109/MSP.2015.126

Schaffer, J., Giridhar, P., Jones, D., Höllerer, T., Abdelzaher, T., & O’Donovan, J. (2015). Getting the Message?: A Study of Explanation Interfaces for Microblog Data Analysis. In *Proceedings of the 20th International Conference on Intelligent User Interfaces SE - IUI ’15* (pp. 345–356). CONF, New York, NY, USA: ACM. http://doi.org/doi: 10.1145/2678025.2701406

Sean Captain. (2016). With Mapbox Deal, IBM Watson Will Learn A Lot More About Where Things Are Happening | Fast Company | Business + Innovation. Retrieved August 9, 2016, from http://www.fastcompany.com/3062635/with-mapbox-deal-ibm-watson-will-know-where-things-are-happening

Spinellis, D. (2014). Service Orchestration with Rundeck. *IEEE Software*, *31*(4), 16–18. http://doi.org/10.1109/MS.2014.92

Steer, J., & Popli, A. (2008). Building secure business applications at Microsoft. *Information Security Technical Report*, *13*(2), 105–110. JOUR. http://doi.org/doi: 10.1016/j.istr.2008.04.001

Sun, Y., Meng, X., Liu, S., & Pan, P. (2005). An approach for flexible RBAC workflow system. In *Computer Supported Cooperative Work in Design, 2005. Proceedings of the Ninth International Conference on* (Vol. 1, p. 524–529 Vol. 1). CONF. Retrieved from citeulike-article-id:1204995

Tom Nolle. (2016a). Infrastructure as code complicates hybrid, multiple cloud management (Part 2 of 2). *Search Cloud Computing*. Retrieved from http://searchcloudcomputing.techtarget.com/tip/Infrastructure-as-code-complicates-hybrid-multiple-cloud-management

Tom Nolle. (2016b). Separating DevOps from the future-driven cloud orchestration. Retrieved August 16, 2016, from http://searchcloudcomputing.techtarget.com/tip/Separating-DevOps-from-the-future-driven-cloud-orchestration

US Dept of Homeland Security, D. (2015). Critical Infrastructure Cyber Community C3 Voluntary Program | Homeland Security. Retrieved August 14, 2016, from https://www.dhs.gov/ccubedvp

Versteden, A., Pauwels, E., & Papantoniou, A. (2015). An Ecosystem of User-facing Microservices Supported by Semantic Models. In B. Berendt, L. Dragan, L. Hollink, M. Luczak-Rösch, E. Demidova, S. Dietze, … J. Breslin (Eds.), *USEWOD-PROFILES@ESWC SE - CEUR Workshop Proceedings* (Vol. 1362, pp. 12–21). CONF, CEUR-WS.org. Retrieved from citeulike-article-id:14127922

Voas, J. (2016). *Networks of “Things”* (RPRT). Gaithersburg MD OR - NIST: NIST. http://doi.org/doi: 10.6028/NIST.SP.800-183

Zaslavsky, A., Perera, C., & Georgakopoulos, D. (2012). Sensing as a Service and Big Data. In *Proceedings of the International Conference on Advances in Cloud Computing* (Vol. abs/1301.0). CONF, Bangalore India. Retrieved from citeulike-article-id:14110474

|  |  |
| --- | --- |
| [1] | The White House Office of Science and Technology Policy, "OSTP Blog: Big Data is a Big Deal," [Online]. Available: http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal. [Accessed 21 February 2014]. |
| [2] | D. Spinellis, "Service Orchestration with Rundeck," *IEEE Software,* vol. 31, no. 4, p. 16–18, 2014. |
| [3] | A. P. C. &. G. D. Zaslavsky, "Sensing as a Service and Big Data," in *Proceedings of the International Conference on Advances in Cloud Computing*, Bangalore, India, 2012. |
| [4] | S. Captain, "With Mapbox Deal, IBM Watson Will Learn A Lot More About Where Things Are Happening | Fast Company | Business + Innovation," 2016. [Online]. Available: http://www.fastcompany.com/3062635/with-mapbox-deal-ibm-watson-will-know-where-things-are-happening. [Accessed 9 August 2016]. |
| [5] | R. Chandramouli, "NIST Special Publication 800-125B Secure Virtual Network Configuration for Virtual Machine (VM) Protection," NIST, Gaithersburg MD, 2016. |
| [6] | D. J.-L. &. A. W. O. R. Frederic Guyomarc’h, "An MDE Approach for Automatic Code Generation from UML/MARTE to OpenCL," *Computing in Science and Engineering,* vol. 15, no. 1, p. 46–55, 2013. |
| [7] | EMC2, "Digital Universe," [Online]. Available: http://www.emc.com/leadership/programs/digital-universe.htm. [Accessed 21 February 2014]. |
| [8] | B. D. W. G. Cloud Security Alliance, "Expanded Top Ten Big Data Security and Privacy Challenges," 2013. |
| [9] | S. c. w. J. G. K. (IBM), 2014. |
| [10] | A. B.-L. F. H. S. Weitzner, "Information Accountability," MIT, 2008. [Online]. Available: http://informatics.mit.edu/blog/2014/03/can-you-have-privacy-and-big-data-too-%E2%80%94-comments-whitehouse . [Accessed 2015]. |
| [11] | M. D. O. S. V. A. W. CACM. Altman, "Big Data Study: Request for Information," MIT , 2014. [Online]. Available: http://informatics.mit.edu/blog/2014/03/can-you-have-privacy-and-big-data-too-%E2%80%94-comments-whitehouse. [Accessed 2015]. |
| [12] | Cloud Security Alliance, Big Data Working Group, "Top 10 Challenges in Big Data Security and Privacy," November 2012. [Online]. Available: http://www.isaca.org/Groups/Professional-English/bigdata/GroupDocuments/Big\_Data\_Top\_Ten\_v1.pdf. [Accessed 2016]. |
| [13] | K. W. R. C. a. P. S. Y. Benjamin Fung, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys (CSUR),* vol. 42, no. 4, p. 14, 2010. |
| [14] | M. M. A. H. D. R. D. (. W. M. W. D. (. NIST, "NIST Special Publication 800-160, Systems Security Engineering". |
| [15] | C. K. K. a. J.-P. S. Onur Aciicmez, *Predicting Secret Keys Via Branch Prediction,* vol. 4377, M. Abe, Ed., San Francisco, CA: Springe, 2007, pp. 225-242. |
| [16] | D. B. a. M. K. Franklin, "Identity-based encryption from the Weil pairing," in *Lecture Notes in Computer Science*, Santa Barbara, CA, August 19-23, 2001. |
| [17] | J. C. a. H. Wee, "Fully, (almost) tightly secure IBE and dual system groups," in *Advances in Cryptology - CRYPTO 2013, Part II*, Santa Barbara, CA, August 18-22, 2013. |
| [18] | C. S. J. a. A. Roy, "Shorter quasi-adaptive NIZK proofs for linear subspaces," in *Advances in Cryptology - ASIACRYPT 2013, Part I*, Bengalore, India, December 1-5, 2013. |
| [19] | A. S. a. B. R. Waters, "Fuzzy identity-based encryption," in *Advances in Cryptology - EUROCRYPT 2005*, Aarhus, Denmark, May 22-26, 2005. |
| [20] | O. P. A. S. a. B. W. Vipul Goyal, "Attribute-based encryption for fine-grained access control of encrypted data," in *ACM CCS 06: 13th Conference on Computer and Communications Security*, Alexandria, Virginia, October 30 - November 3, 2006. |
| [21] | A. S. a. B. W. John Bethencourt, "Ciphertext-policy attribute-based encryption," in *2007 IEEE Symposium on Security and Privacy*, Oakland, California, May 20-23, 2007. |
| [22] | D. B. a. B. Waters, "Conjunctive, subset, and range queries on encrypted data," in *TCC 2007: 4th Theory of Cryptography Conference*, Amsterdam, The Netherlands, February 21-24, 2007. |
| [23] | S. J. C. S. J. H. K. M.-C. R. a. M. S. David Cash, "Highly-scalable searchable symmetric encryption with support for boolean queries," in *Advances in Cryptology - CRYPTO 2013, Part I*, Santa Barbara, CA, August 18-22, 2013. |
| [24] | D. M. Nicol, "Modeling and simulation in security evaluation," *Security & Privacy, IEEE,* vol. 3, no. 5, p. 71–74, 2005. |
| [25] | R. Dhall, "Performance Patterns in Microservices based Integrations," *Computing Now,* FEB 16, 2016. |
| [26] | Microsoft, "Deploying Windows Rights Management Services at Microsoft," 2015. [Online]. Available: http://technet.microsoft.com/en-us/library/dd277323.aspx. [Accessed 2016]. |
| [27] | The Nielsen Company, "Consumer Panel and Retail Measurement," 2015. [Online]. Available: www.nielsen.com/us/en/nielsen-solutions/nielsen-measurement/nielsen-retail-measurement.html. [Accessed 2016]. |
| [28] | SAFE-BioPharma Association, "Welcome to SAFE-BioPharma," SAFE-BioPharma, [Online]. Available: http://www.safe-biopharma.org/. [Accessed 3 March 2015]. |
| [29] | T. Weida, "HL7 Committee working note," 7 May 2014. [Online]. Available: http://wiki.hl7.org/images%2Fa%2Fae%2FEHR\_Action\_Verbs\_and\_Security\_Operations\_May\_2014\_HL7\_WGM.pptx. [Accessed 2016]. |
| [30] | Microsoft, "How to set event log security locally or by using Group Policy in Windows Server 2003," Microsoft, 7 January 2017. [Online]. Available: http://support.microsoft.com/kb/323076. [Accessed 2017]. |
| [31] | DefenseSystems, "UAV video encryption remains unfinished job," DefenseSystems, 31 October 2012. [Online]. Available: http://defensesystems.com/articles/2012/10/31/agg-drone-video-encryption-lags.aspx. |
| [32] | D. o. D. M. f. D. CIO, "Department of Defense Cloud Computing Strategy," Department of Defense, July 2012. [Online]. Available: http://1.usa.gov/1E0UTXT. |
| [33] | A. a. F. Lamberti, "Advances in target detection and tracking in Forward-Looking InfraRed (FLIR) imagery," *Sensors (Basel, Switzerland),* vol. 14, no. 11, pp. 20 297-20 303, 2014. |
| [34] | A. B. L. K. S. Z. Y. J. L. R. P. T. J. a. K. J. R. K. A. G. Fisher, "Quantum computing on encrypted data 5," *Nature Communications,* no. January, 2015. |
| [35] | J. Cartledge, "US Lawmakers Pledge to Close Air Cargo Security ‘Loophole’," Post&Parcel, 1 November 2010. [Online]. Available: http://postandparcel.info/35115/news/us-lawmakers-pledge-to-close-air-cargo-security-%E2%80%9Cloophole%E2%80%9D/. [Accessed 2016]. |
| [36] | J. T. J. M. R. B. J. M. L. B. a. D. L. Fang Liu, *NIST Cloud Computing Reference Architecture, SP500-292,* National Institute of Standards and Technology, 2011. |
| [37] | W. Jansen and T. Grance, "Guidelines on Security and Privacy in Public Cloud Computing," December 2011. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-144.pdf. [Accessed Februray 2017]. |
| [38] | J. M. a. B. Anderson, Preventing Good People From Doing Bad Things: Implementing Least Privilege, Berkeley, CA: Apress, 2011. |
| [39] | S. Patig, "Model-Driven Development of Composite Applications," in *Model-Based Software and Data Integration. Communications in Computer and Information Science*, Berlin, Heidelberg, 2008. |
| [40] | C. J. A. C. E. C. a. E. M. M. López-Sanz, "Modelling of Service-Oriented Architectures with UML," *Theoretical Computer Science,* vol. 194, no. 4, p. 23–37, 2008. |
| [41] | L. B. S. C. M. C. a. B. P. D. Ardagna, "A Service-Based Framework for Flexible Business Processes," *IEEE Software,* vol. 28, no. 2, pp. 61 - 67, 2011. |

Reference from Mark during 1/24/17 NBD-PWG meeting: <https://www.ise.gov/resources/document-library/ise-privacy-guidelines>

1. “Contributors” are members of the NIST Big Data Public Working Group who dedicated great effort to prepare and substantial time on a regular basis to research and development in support of this document. [↑](#footnote-ref-1)
2. Typically such supporting SnP 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 SnP for specific domains, with machine learning, ontologies and other specialized resources that may be beyond the capabilities of Build architects. [↑](#footnote-ref-2)
3. Gunderson, "Drone patrol: Unmanned craft find key role in U.S. border security," Minnesota Public Radio, Feb. 2015. [Online]. Available: http://www.mprnews.org/story/2015/02/19/predator-drone [↑](#footnote-ref-3)
4. US Department of Justice, “Guidance on Domestic Use of Unmanned Aircraft Systems,” www.justice.gov/file/441266/download, undated. [↑](#footnote-ref-4)
5. Source: <http://www.catnmsplan.com/web/groups/catnms/@catnms/documents/appsupportdocs/cat_nms_security_requirements_032416.pdf> [↑](#footnote-ref-5)
6. https://bitbucket.org/openid/heart/wiki/Alice\_Shares\_with\_Physicians\_and\_Others\_UMA\_FHIR [↑](#footnote-ref-6)
7. 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)). [↑](#footnote-ref-7)
8. 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. [↑](#footnote-ref-8)
9. CISSP is a professional computer security certification administered by (ISC)).2. (<https://www.isc2.org/cissp/default.aspx>) [↑](#footnote-ref-9)