**NIST Big Data**

**Security and Privacy Requirements**

**Version 1.0**

September 27, 2013

**Security & Privacy Subgroup**

**NIST Big Data Working Group (NBD-WG)**

**September, 2013**

# Table of Contents

[Table of Contents 2](#_Toc368145351)

[Executive Summary 5](#_Toc368145352)

[1 Introduction 5](#_Toc368145353)

[Background 5](#_Toc368145354)

[Objectives 6](#_Toc368145355)

[How This Report Was Produced 6](#_Toc368145356)

[Structure of This Report 7](#_Toc368145357)

[2 Big Data Security and Privacy 8](#_Toc368145358)

[Introduction 8](#_Toc368145359)

[Scope 9](#_Toc368145360)

[Actors 9](#_Toc368145361)

[Classification of Security and Privacy Topics 10](#_Toc368145362)

[3 Use Cases 12](#_Toc368145363)

[Retail/ Marketing 12](#_Toc368145364)

[3.1.1 Scenario 1: Modern Day consumerism 12](#_Toc368145365)

[3.1.2 Scenario 2: Nielsen Homescan 12](#_Toc368145366)

[3.1.3 Scenario 3: Web Traffic Analytics 13](#_Toc368145367)

[Healthcare 13](#_Toc368145368)

[3.1.4 Scenario 1: Health Information Exchange 13](#_Toc368145369)

[3.1.5 Scenario 2: Genetic Privacy 14](#_Toc368145370)

[3.1.6 Scenario 3: Pharma Clinic Trial Data Sharing [3] 15](#_Toc368145371)

[Cyber-security 16](#_Toc368145372)

[3.1.7 Scenario 16](#_Toc368145373)

[Government 16](#_Toc368145374)

[3.1.8 Scenario 1: Military (Unmanned Vehicle sensor data) 16](#_Toc368145375)

[3.1.9 Scenario 2: Education (“Common Core” Student Performance Reporting) 17](#_Toc368145376)

[Industrial: Aviation 18](#_Toc368145377)

[3.1.10 Scenario 18](#_Toc368145378)

[4 Abstraction of Requirements 18](#_Toc368145379)

[Privacy of data 18](#_Toc368145380)

[Provenance of data 19](#_Toc368145381)

[System Health 19](#_Toc368145382)

[5 Internal Security Practices 20](#_Toc368145383)

[Internal Access control rules for general industry 20](#_Toc368145384)

[6 Taxonomy of Security and Privacy Topics 22](#_Toc368145385)

[Privacy 22](#_Toc368145386)

[Provenance 23](#_Toc368145387)

[System Health 24](#_Toc368145388)

[7 Security Reference Architecture Components 25](#_Toc368145389)

[Interface of Data Providers $\rightarrow $ Big Data Application Provider 26](#_Toc368145390)

[Interface of Big Data Application Provider $\rightarrow $ Data Consumer 26](#_Toc368145391)

[Interface of Application Provider $\leftrightarrow $ Big Data IT Provider 27](#_Toc368145392)

[Internal to Big Data IT Provider 27](#_Toc368145393)

[General Components 27](#_Toc368145394)

[8 Mapping Use Cases to Reference Architecture 28](#_Toc368145395)

[Cargo Shipping 28](#_Toc368145396)

[Nielsen Homescan 30](#_Toc368145397)

[Pharma Clinical Trial Data Sharing 31](#_Toc368145398)

[Large Network Cybersecurity SIEM 32](#_Toc368145399)

[Consumer Digital Media Usage 33](#_Toc368145400)

[Unmanned Military Vehicle Sensor Systems 35](#_Toc368145401)

[Common Core K-12 Student Reporting 36](#_Toc368145402)

[Web Traffic Analytics 37](#_Toc368145403)

[Health Information Exchange 38](#_Toc368145404)

[9 References 42](#_Toc368145405)

[Appendix 1. Actors (Draft) 43](#_Toc368145406)

[Appendix 2. Classification of Security and Privacy Topics (Draft) 44](#_Toc368145407)

# Executive Summary

# Introduction

## Background

There is a 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 term used to describe the deluge of data in our networked, digitized, sensor-laden, information driven world. The availability of vast data resources carries the potential to answer questions previously out of reach. Questions like: How do we reliably detect a potential pandemic early enough to intervene? Can we predict new materials with advanced properties even before these materials have been synthesized? How can we reverse the current advantage of the attacker over the defender in guarding against cybersecurity threats?

However, there is also broad agreement on the ability of Big Data to overwhelm traditional approaches. The rate at which data volumes, speeds, and complexity are growing is outpacing scientific and technological advances in data analytics, management, transport, and more.

Despite the widespread agreement on the opportunities and current limitations of Big Data, a lack of consensus on some important, fundamental questions is confusing potential users and holding back progress. What are the attributes that define Big Data solutions? How is Big Data different from the traditional data environments and related applications that we have encountered thus far? 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?

At the NIST Cloud and Big Data Forum held in January 15-17, 2013, the community strongly recommended that NIST create a public working group for the development of a Big Data Technology Roadmap. This roadmap would help to define and prioritize requirements for *interoperability*, *portability*, *reusability*, and *extensibility* for big data usage, analytic techniques and technology infrastructure in order to support secure and effective adoption of Big Data.

On June 19, 2013, the NIST Big Data Public Working Group (NBD-PWG) was launched with overwhelmingly strong participation from industry, academia, and government across the nation. The scope of the NBD-PWG is to form a community of interests from all sectors including industry, academia, and government, with the goal of developing a consensus in definitions, taxonomies, secure reference architectures, and a technology roadmap. Such a consensus would therefore create a vendor-neutral, technology- and infrastructure-agnostic framework which would enable Big Data stakeholders to pickand hoose appropriate analytics tools for their processing and visualization requirements on the most suitable computing platform and cluster while allowing for value to be added by Big Data service providers.

NBD-PWG has created five subgroups: namely, the Definitions and Taxonomies, Use Case and Requirements, Security and Privacy, Reference Architecture, and Technology Roadmap. These subgroups will help to develop the following set of preliminary consensus working drafts by September 27, 2013:

1. Big Data Definitions
2. Big Data Taxonomies
3. Big Data Requirements
4. Big Data Security and Privacy Requirements
5. Big Data Reference Architectures White Paper Survey
6. Big Data Reference Architectures
7. Big Data Security and Privacy Reference Architectures
8. Big Data Technology Roadmap

Due to time constraints and dependencies between subgroups, the NBD-PWG hosted two hours weekly telecon meeting from Mondays to Fridays for the respective subgroups. Every three weeks, NBD-PWG called a joint meeting for progress reports and document updates from these five subgroups. In between, subgroup co-chairs met for two hours to synchronize their respective activities and identify issues and solutions.

## Objectives

**Scope**

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

**Tasks**

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

**Deliverables**

* + - 1. Produce a working draft for Big Data Security and Privacy Requirements Document
			2. Produce a working draft Big Data Security and Privacy Reference Architecture

## How This Report Was Produced

[Wo will put more thoughts into this section; each subgroup will be slightly different]

## Structure of This Report.

[Subgroup will draft this section and hope it will be consistent with other subgroup approaches]

# Big Data Security and Privacy

## Introduction

We attempted to identify security and privacy issues particular to Big Data. Variety, Volume and Velocity are key elements to Big Data, and, where possible, aspects of these properties directed our attention.

To standardize on a security reference architecture there is a need to leverage specialists in diverse domains. Application domains include health care, drug discovery, finance and many others from both the private and public sectors. Examples of scenarios within these application domains include health exchanges, clinical trials, mergers and acquisitions, device telemetry, and international anti-Piracy. Security technology domains include identity, authorization, audit, network and device security, and federation across trust boundaries. For instance, when using cloud service providers and federating across trust boundaries, there is a clear need for other services such as cryptography to strengthen security.

Just as it would be difficult to align schemas and protocols within implementations of a technical domain, it would be quite difficult to standardize, align, or even provide a mapping between terms across a variety of application and technology domains. To effectively communicate across domains, there is a need to socialize the usage of terms related to security and compliance. We hope to do this in a straightforward way that encourages others to engage without losing sight of complex and difficult security and privacy issues particular to Big Data.

**Variety and GRC** Within an industry vertical, a common security vocabulary may take hold because of underlying Governance, Risk management and Compliance (GRC) requirements. But there are disparate regulatory bodies across sovereign boundaries; often the entities that certify are distinct from those that audit or adjudicate. When methods to accommodate GRC can grow organically, there is a trend toward consistent blueprints for policy authoring, decision and enforcement, and for federation of demands, claims and obligations across organizations. On the other hand, consistency is difficult to maintain as variety increases dramatically.

Big Data may not be implemented using cloud technology, but most expect that cloud will play a big role. Clouds and federation tend to complicate things for application and the technology domains, and security mechanisms are often a previous generation of enterprise solutions that are being repurposed inappropriately for a radically different threat model. Also, the legal standing of data can change when it moves from an enterprise data center to the data center of a service provider; hence laws pertaining to discovery can be interpreted in unpredictable ways. Variety of data sources could lead to data providers and service providers in separate legal jurisdictions.

**Security and GRC Tradeoffs** Big Data for some projects could entail tradeoffs between security and compliance. Even if all participants were to have state of the art security, it is possible that they would be in violation of compliance requirements if they did not federate identity, authorization and audits in a manner that would enable each party to address compliance obligations. Big Data participants must communicate technical security requirements but also their broader GRC intent.

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

**Access to Privacy Data** Both HIPAA and the draft FTC Privacy Bill of Rights specify consumer access to data, and recourse to remedy data errors that are of consequence. These requirements need to shape some facets of Big Data, such as consumer-facing portals.

## Scope

Firstly, a distinction needs to be made between fault tolerance and security. Fault tolerance is resistance to unintended accidents, while security is resistance to malicious actions. Secondly, we need to understand how Big Data security concerns arise out of the defining characteristics of Big Data and how it is differentiated from traditional security concerns.

1. Big Data may be gathered from diverse end-points. There may be more types of actors than just Provider and Consumers – viz. Data Owners: e.g., mobile users, social network users.
2. Data aggregation and dissemination have to be made securely and inside the context of a formal, understandable framework. This could be made part of a contract with Data Owners.
3. Availability of data to Data Consumers is often an important aspect in Big Data, possibly leading to public portals and ombudsman-like roles for data at rest.
4. Data Search and Selection can lead to privacy or security policy concerns. What capabilities are provided by the Provider in this respect?
5. Privacy-preserving mechanisms will probably be needed, but these could add to system complexity or hinder certain types of analytics.
6. Since there may be disparate processing steps between Data Owner, Provider and Data Consumer, the integrity of data coming from end-points must be ensured. End-to-end information assurance practices for Big Data, e.g., for verifiability, are not dissimilar from other systems, but must be designed on a larger scale.

## Actors

Some consumers may naively believe themselves immune to the powerful force that is Big Data. For them, a tour of the many roles an individual can play in Big Data systems may be useful. Ordinary citizens assume roles as Big Data actors through a variety of touch points in retail, government, education, finance, insurance, employment and health care.

This material is still under development, but can be found in draft form in Appendix 1.

## Classification of Security and Privacy Topics

The set of topics was initially adapted from the scope of the CSA BDWG charter, organized according to the classification in [1]. Security and Privacy concerns are classified in four categories:

1. Infrastructure Security
2. Data Privacy
3. Data Management
4. Integrity and Reactive Security

Developing subcategories with adequate supporting explanation proved to be a large undertaking. Because the group believes a detailed look at these topics will be of interest to a larger community, it is premature to present the topic list as a final list. Rather than a prescriptive document at this stage of the Work Group’s effort, we present drafts representing the current state of community discussion. Later versions will refine as needed.

A brief outline of this work is reproduced in the table below.

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

A draft of the ongoing work is presented in Appendix 2.

# Use Cases

## Retail/ Marketing

## Scenario 1: Modern Day consumerism

**Scenario Description**: Consumers, with the help of smart devices have become very conscious of price, convenience and access before they make decision on a purchase. Content owners license data for usage by consumers through presentation portals, e.g., Netflix, iTunes, etc.

Comparative pricing from different retailers, store location and/or delivery options and crowd sourced rating have become common factors for selection. On the flip side, retailers, to compete, are keeping close watch on consumer locations, interests, spending patterns etc. to dynamically create deals and sell them products that consumers don’t yet know that they want.

**Current S&P**: Individual data is collected by several means such as Smart Phone GPS/ Location, Browser use, Social Media, Apps on smart devices, etc.

1. Privacy: Most means described above offer weak privacy controls, however consumer unawareness and oversight allows 3rd parties to “legitimately’” capture information. Consumers can have limited to no expectation of privacy in this scenario.
2. Security: Controls are **inconsistent and/ or not established appropriately** to
	1. Isolate, containerize and encrypt data,
	2. Monitor and detect threats,
	3. Identify users and devices for data feed
	4. Interfacing with other data sources, etc.
	5. Anonymization: Some data collection and aggregation uses anonymization techniques, however individual users can be re-identified by leveraging other public ‘big-data’ pools
	6. Original DRM model not built to scale to meet demand for forecast use for the data.

**Current Research**: Limited research in enabling Privacy and security controls that protect individual data (Whether anonymized or non-anonymized).

## Scenario 2: Nielsen Homescan

**Scenario description**: This is a subsidiary of Nielsen that collects family level retail transactions. A general transaction has a checkout receipt, contains all SKUs purchased, time, date, store location, etc. It is currently implemented using a statistically randomized national sample. As of 2005 this was already a multi-terabyte warehouse for only a single F500 customer’s product mix, mostly with structured data set. Data is in-house but shared with customers who have partial access to data partitions through web portions using columnar databases. Other Cx only receive reports, which include aggregate data, but which can be drilled down for a fee.

**Current S&P:**

1. Privacy: There is considerable amount of PII data. Survey participants are compensated in exchange for giving up segmentation data, demographics, etc.
2. Security:
	1. Traditional access security with group policy, implemented at the field level using the DB engine.
	2. No Audit and opt out scrubbing.

**Current Research**: TBD

## Scenario 3: Web Traffic Analytics

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

**Current S&P**:

1. Non-EU: Opt-in defaults are relied upon to gain visitor consent for tracking. IP address logging enables potential access to geo-coding to potentially block-level identification. MAC address tracking enables device ID which is a form of PII.
2. Some companies allow for purging of data on demand, but it’s unlikely to expunge previously collected webserver traffic.
3. EU has more strict regulations regarding collection of such data, which is treated as PII and is to be scrubbed (anonymized) even for multinationals operating in EU but based in the US.

**Current research**: TBD

## Healthcare

## Scenario 1: Health Information Exchange

**Scenario Description**: Health Information Exchanges (HIEs) aspire to facilitate sharing of healthcare information that might include Electronic Health Records (EHRs) such that they are accessible to relevant Covered Entities, but in a manner that enables Patient Consent.

HIEs under construction tend to be federated, where the respective Covered Entity retains custodianship of their data, which poses problems for many scenarios such as Emergency. This is for a variety of reasons that include technical (such as inter-operability) business, and security concerns.

Cloud enablement of HIEs through strong cryptography and key management that meets the HIPAA requirements for PHI, ideally without requiring the cloud service operator to sign a Business Associate Agreement, would provide several benefits that would include patient safety, lowered healthcare costs, regulated accesses during emergencies that might include break the glass and CDC scenarios.

Some preliminary scenarios proposed are:

1. 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 to get immediate access to relevant patient records. Using cryptographically enhanced key lifecycle management we can provide a sufficient level of visibility and nonrepudiation that would enable tracking violations after the fact.
2. Informed Consent: Often when there is a transfer of EHRs between Covered Entities and Business Associates, it would be desirable and necessary for the patient to be able to convey their approval, but also to specify what components of their EHR can be transferred (for instance, their Dentist would not need to see their psychiatric records.) Through cryptographic techniques we could leverage the ability to specify the fine-grain cipher text policy that would be conveyed.
3. Pandemic Assistance: There will be situations when public health entities, such as the CDC, and perhaps other NGOs 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 in a manner that facilitates the scenario requirement, but does this through authorization and audits

**Current and/or proposed S&P**:

1. Security:
2. Light-weight but secure off-cloud encryption: Need the ability to perform light-weight but secure off-cloud encryption of an EHR that can reside in any container that ranges from a browser, to an enterprise server, that leverages strong symmetric cryptography.
3. Homomorphic Encryption.
4. Applied Cryptography: Tight reductions, realistic threat models, and efficient techniques.
5. Privacy:
6. Differential Privacy: Techniques for guaranteeing against inappropriate leakage of PII
7. HIPAA

Current research: Homomorphic Encryption, Off-cloud Encryption.

## Scenario 2: Genetic Privacy

**Scenario Description**: A consortium of policy makers, advocacy organizations, individuals, academic centers and industry have 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 and 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, University of California San Francisco (UCSF), 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 families, friends, and clinicians can better understand what the mutation means. Working together to build this resource means working towards a better understanding of disease, higher quality patient care, and improved human health.

**Current S&P**:

1. Security:
	1. SSL based authentication and access control. Basic user registration with low attestation level
	2. Concerns over data ownership and custody upon user death
	3. Site administrators may have access to data- Strong Encryption and key escrow recommended.
2. Privacy:
3. Strict privacy which lets user control who can see information, and for what purpose.
4. Concerns over data ownership and custody upon user death.

**Current research**:

1. Under what circumstances can the data be shared with private sector?
2. Under what circumstances can the user data be shared with government?

## Scenario 3: Pharma Clinic Trial Data Sharing [3]

**Scenario Description**: Companies routinely publish their clinical research, collaborate with academic researchers, and share clinical trial information on public web sites at the time of patient recruitment, after new drug approval, and when investigational research programs have been discontinued.

Biopharmaceutical companies will apply these Principles for Responsible Clinical Trial Data Sharing as a common baseline on a voluntary basis, and we encourage all medical researchers, including those in academia and in the government, to promote medical and scientific advancement by adopting and implementing the following commitments

1. Enhancing data sharing with researchers
2. Enhancing public access to Clinical Study Information
3. Sharing results with Patients who participate in clinical trials
4. Certifying procedures for sharing trial information
5. Reaffirming commitments to publish clinical trial results

**Current and Proposed S&P**:

1. Security:
2. Longitudinal custody beyond trial disposition unclear, especially after firms merge or dissolve
3. Standards for data sharing unclear
4. Need for usage audit and Security
5. Publication restrictions : additional security will be required to ensure rights of publishers, e.g. Elsevier or Wiley
6. Privacy:
7. Patient-level data disclosure - elective, per company.
8. The association mentions anonymization (“re-identification”) but mentions issues with small sample sizes
9. Study Level data disclosure – elective, per company

**Current Research**: TBD

## Cyber-security

## Scenario

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

This scenario highlights two sub-scenarios:

1. Security for big-data
2. Big-data for security

**Current S&P**:

1. Security:
2. Traditional Policy-type security prevails, though temporal dimension and monitoring of policy modification events tends to be nonstandard or unaudited
3. Cyber-security apps themselves run at high levels of security and thus require separate audit and security measures
4. No cross-industry standards exists for aggregating data beyond operating system collection methods
5. Desired security characteristics for such systems are: data governance, encryption/ key management, tenant data isolation/ containerization,
6. Privacy:
7. Enterprise authorization for data release to state/ national organizations
8. Protection of PII data

**Current research**: Vendors are adopting big data analytics for mass scale log correlation and incident response.

## Government

## Scenario 1: Military (Unmanned Vehicle sensor data)

**Scenario Description**: Unmanned vehicles (“drones”) and their onboard sensors (e.g., streamed video) can produce petabytes of data that must be stored in nonstandardized formats. These streams are often not processed in real time, but DoD is buying technology to do this. Because correlation is key, GPS, time and other data streams must be co-collected. Security breach use case: Bradley Manning leak.

**Current S&P**:

1. Separate regulations for agency responsibility apply. For domestic surveillance, FBI. For overseas, multiple agencies including CIA and various DoD agencies. Not all uses will be military; consider NOAA.
2. Military security classifications are moderately complex and based on “need to know.” Information Assurance practices are followed, unlike some commercial settings.

**Current Research**:

1. Usage is audited where audit means are provided, software is not installed / deployed until “certified,” and development cycles have considerable oversight, e.g., see [4].
2. Insider Threat (a la Snowden, Manning or spies) is being addressed in programs like DARPA CINDER. This research and some of the unfunded proposals made by industry may be of interest.

## Scenario 2: Education (“Common Core” Student Performance Reporting)

**Scenario Description**: A number of states (45) have decided to unify standards for K-12 student performance measurement. Outcomes are used for many purposes, and the program is incipient, but will obtain longitudinal Big Data status. The datasets envisioned include student level performance across their entire school history, across schools and states, and taking into account variations in test stimuli.

**Current S&P**:

1. Data is scored by private firms and forwarded to state agencies for aggregation. Classroom, school and district tagging remains. Status of student PII is unknown; however it’s known that teachers receive classroom-level performance feedback. Do students/parents have access?
2. According to some reports [5], parents can opt students out, so that data must be collected.

**Current Research**:

1. Longitudinal performance data would have value for program evaluators if data scales up.
2. Data-driven learning content administration [6] will require access to performance data at learner level, probably more often than at test time, and at higher granularity, thus requiring more data. Example enterprise: Civitas Learning [7] Predictive analytics for student decision-making.

## Industrial: Aviation

## Scenario

**Scenario Description**: Most commercial airlines are equipped with hundreds of sensors to constantly capture engine and/or aircraft health information during a flight. For a single flight, the sensors may collect multiple GB’s 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, plane manufacturer company etc., and data may be shared across these companies. The aggregated data is analyzed for maintenance scheduling, flight routines, etc. One common request from airline companies is to secure and isolate their data from competitors, even when data is being streamed to the same analytics system. Airline companies also prefer to keep control of how/ when / to whom the data is shared, even for analytics purpose. Most of these analytics systems are now being moved to Infrastructure cloud providers.

**Current and desired S&P**:

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

**Current Research**:

1. Virtualized infra layer mapping on cloud provider
2. Homomorphic Encryption
3. Quorum based Encryption
4. Multi-party computational capability
5. Device PKI

# Abstraction of Requirements

## Privacy of data

**Retail**: Nielsen Homescan data: family level retail transactions. Data is in-house, but shared with customers who have partial access to data partitions. Others only receive reports, which include aggregate data, but which can be drilled down for a fee. Access security is traditional group policy, implemented at the field level using the DB engine. PII data is considerable. Survey participants are compensated in exchange for giving up segmentation data, demographics, etc.

**Healthcare**: An employer should never have access to an employee’s medical history, nor the medical records of family members. Need for HIPAA-capable cryptographic controls and key management. Doximity is a secure way for doctors to share research, clinical trial data, and patient records in the cloud. It is widely adopted already. Data ownership issues for genetic data have to be looked at. Usage audit and security is required.

**Media and Communications**: Collection and monitoring of PII data has to be looked at. Currently, policy based access control is enforced. Cybersecurity apps themselves run at high levels of security and thus require separate audit and security measures.

**Military**: Insider threat and leakage of military data, such as those collected by unmanned vehicles.

**Education**: Privacy of student performance data needs to be preserved. Access control policy among teachers, students and parents needs to be specified. On the other hand, this data is useful for predictive analytics for enhancing learning and decision making.

**Marketing**: Visit level webserver logs are used in correlation with other data sources. Opt-in consent is in place for tracking Non-EU visitors. EU regulations are stricter. IP address reveals geographical data. MAC address tracking can be construed as PII.

## Provenance of data

**Retail**: Fraudulent padding of data and double counting has to be prevented.

**Media**: Digital Rights Management.

## System Health

**Media**: Data collection, monitoring and correlation for gauging health of the system.

# Internal Security Practices

## Internal Access control rules for general industry

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

In addition, access control for big data covers more than access to the data.  The security of the account that is used for access needs to be considered.  If an account is shared amongst different systems and environments then there is a possibility that access control can be easily compromised.  Moreover, data may be accessed multiple ways.  For example, the same data may be accessed by a user, administrator, another system, etc.  In addition, it may be accessed externally or copies of it transmitted, transferred, etc. so there may be duplicates of full or partial data in multiple locations.  Therefore, knowing who / what accesses the data becomes increasingly important.  Of course, very restrictive measures to control accounts would be impractical.  However, there are best practices, such as protection based on classification of the data, least privilege and separation of duties that can reduce the risks.

**General**

* 1. Least privileges – access to big data should be based on the minimum amount of privilege need to perform the function.
	2. If one of the data elements is protected because of its classification (for example – PII, HIPAA, PCI, etc.), then all of the data that it is sent with it, will inherit that classification.  That way if the data is joined to / associated with other data that may cause a privacy issue, then all of that data is already protected.
	3. If data is accessed from, transfer to, or transmitted to the cloud, internet or another external entity, then the data should be protected based on its classification.
	4. There should be an indicator / disclaimer on the display of the user, if privacy or sensitive data is being accessed or viewed.
	5. All accounts (except for system related accounts) should be reviewed at a minimum of one year to insure that they are still required.
	6. All accounts (except for system related accounts) that have not been used within 180 days should be deleted.  If the system will not allow deletion of the account then the account should be disabled.
	7. Access to privacy of sensitive data should be logged.  The minimum logging requirements should be timestamp, account number.
	8. Role-based-access to big data should be based on roles.  Each role should be assigned the least privileges needed to perform the function.
	9. Roles should be reviewed a minimum of every 2 years to insure that they are still valid and to insure that the accounts assigned to them are still valid.

**User**

* 1. Each user should have his or her personal account.  Shared accounts should not be used unless there is a systems limitation.
	2. A user account should not be a multipurpose account.  For example, the user account should not be used as an administrative account or to run production jobs.

**System**

* 1. In case of system to system access, there should not be shared accounts.
	2. Access for a system that contains big data needs to be approved by the data owner or their representative.  The representative should not the system administrator, since that may cause a separation of duties issue.
	3. The same type of data stored on different systems should the same classification and rules for access controls to ensure that it has the same level of protection.

**Administrative**

* 1. 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.
	2. The same administrative account should not be used for access to the production and non-production (test, development, QA, etc.) systems.

# Taxonomy of Security and Privacy Topics

## Privacy

1. **Communication Privacy**: Confidentiality of data in transit, enforced for example by using TLS.
2. **Confidentiality**: Confidentiality of data at rest
	1. **Access Policies**: Policies to access data based on credentials
		1. **Systems**: Policy enforcement by using systems constructs like ACLs, VM boundaries.
		2. **Crypto Enforced**: Policy enforcement by using cryptographic mechanisms, such as PKI, identity/attribute-based encryption.
3. **Computing on Encrypted Data**
	1. **Searching and Reporting**: Cryptographic protocols that support searching and reporting on encrypted data – any information about the plaintext not deducible from the search criteria are guaranteed to be hidden.
	2. **Fully Homomorphic Encryption**: Cryptographic protocols that support operations on the underlying plaintext of an encryption – any information about the plaintext is guaranteed to be hidden.
4. **Secure Data Aggregation**: Aggregating data without compromising privacy.
5. **Key Management**

## Provenance

1. **End-point Input Validation:** Mechanism to validate whether input data is coming from an authenticated source, such as digital signatures.
	1. **Syntactic:** Validation at a syntactic level.
	2. **Semantic:** Semantic validation is an important concern. Generally, semantic validation would validate typical business rules such as a due date. This capability can extend to the relationship of the particular file format and that the constructs which may be based on a standard can have a variance that may poise a denial of service in the form of a lockup of an application. This may also be observed even when using data translators which may not recognize the particular variant. Protocols and data formats may be altered by a vendor using reserved data field for example that will allow their products to have capabilities that differentiate their products from other products. This may also appears in differences in versions of systems for consumer devices, such as mobile devices such as cell phones.
	 The semantic of a message and the data that is to be transported should be validated to assure at minimum conformant with standards. The use of digital signatures will be important to provide assurance that the data has been verified using a validator or data checker. This may be important to provide assurance that data from a sensor or from the data provider. This capability is important particularly if the data is to be Transformed or in the Curation of the data. If that data fails to meet the requirements it may be discarded and if that 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 able to be disseminated to consumers. The Big Data system is a special case where use of digital signatures will be very important.
2. **Communication Integrity:** Integrity of data in transit, enforced for example by using TLS.
3. **Authenticated Computations on Data**: Ensuring that computations taking place on critical fragments of data are indeed the expected computations.
	1. **Trusted Platforms:** Enforcement through the use of Trusted Platforms, such as TPMs.
	2. **Crypto Enforced:** Enforcement through the use of cryptographic mechanisms.
4. **Granular Audits:** Enabling audit at high granularity.
5. **Control of Valuable Assets**
	1. **Lifecycle Management**
	2. **Retention, Disposition, Hold**
	3. **Digital Rights Management**

## System Health

1. **Security against DoS**
	1. **Construction of cryptographic protocols proactively resistant to DoS**
2. **Big Data for Security**
	1. **Analytics for Security Intelligence**
	2. **Data-driven Abuse Detection**
	3. **Large-scale and Streaming Data Analysis**
	4. **Event Detection**
	5. **Forensics**

# Security Reference Architecture Components

Security and Privacy considerations form a fundamental aspect of the Big Data Reference Architecture. This is geometrically depicted in Figure 1 by having a Security and Privacy fabric around the General Big Data Reference Architecture components, since it touches all of the components. This way the role of S&P is depicted in the right relation to the components and at the same time does not explode into finer details. In addition to the Application and Framework Providers, we also decided to include the Data Provider and Data Consumer into the fabric since at the least they have to agree on the security protocols and mechanisms in place.



Figure . The General Big Data Reference Architecture

In this section, we propose the Security Reference Architecture (Figure 2) where we provide details on the Security and Privacy considerations at the interface of and internal to the General Reference Architecture components.



Figure . Big Data Security Reference Architecture

We now describe the Security Reference Architecture in detail.

Each section reflects an Architectural Component of the Security Reference Architecture.

## Interface of Data Providers $\rightarrow $ 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. Hence real time security monitoring is useful. Data discovery and classification has to be performed in a privacy respecting manner.

## Interface of Big Data Application Provider $\rightarrow $ Data Consumer

It has to be ensured that data or aggregate results going out to Data Consumers are privacy preserving. Data being accessed by third parties or other entities has to confirm to legal regulations like HIPAA. Access to sensitive data by the government and potential undermining of freedom of expression is a concern.

## Interface of Application Provider $\leftrightarrow $ Big Data IT Provider

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

## Internal to Big Data IT Provider

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

## General Components

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

# Mapping Use Cases to Reference Architecture

## Cargo Shipping

The following use case defines the overview of a Big Data application related to the shipping industry (i.e. FedEx, UPS, DHL, etc.). The shipping industry represents possibly the largest potential use case of Big Data that is in common use today. It relates to the identification, transport, and handling of item (Things) in the supply chain. The identification of an item begins with the sender to the recipients and for all those in between with a need to know the location and time of arrive of the items while in transport. A new aspect will be status condition of the items which will include sensor information, GPS coordinates, and a unique identification schema based upon a new ISO 29161 standards under development within ISO JTC1 SC31 WG2. The data is in near real-time being updated when a truck arrives at a depot or upon delivery of the item to the recipient. Intermediate conditions are not currently know, the location is not updated in real-time, items lost in a warehouse or while in shipment represent a problem potentially for homeland security. The records are retained in an archive and can be accessed for xx days.



**Mapping to the Security Reference Architecture**

|  |  |  |
| --- | --- | --- |
| RA Component | Security & Privacy Topic | Use Case Mapping |
| Sources → Transformation | End-Point Input Validation | Ensuring integrity of data collected from sensors |
| Real Time Security Monitoring | Sensors can detect abnormal temperature/environmental conditions for packages with special requirements. They can also detect leaks/radiation. |
| Data Discovery and Classification |  |
| Secure Data Aggregation | Aggregating data from sensors securely |
|  |  |  |
| Transformation → Uses | Privacy-preserving Data Analytics | Sensor collected data can be private and can reveal information about the package and geo-information. Revealing such information needs to be privacy preserving. |
| Compliance with Regulations |  |
| Govt access to data and freedom of expression concerns | Department of Homeland Security may monitor suspicious packages moving into/out of the country. |
|  |  |  |
| Transformation ↔ Data Infrastructure | Data Centric Security such as identity/policy-based encryption |  |
| Policy management for access control | Private, sensitive sensor data, package data should only be available to authorized individuals. Third party companies like LoJack have low level access to the data. |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption |  |
| Audits |  |
|  |  |  |
| Data Infrastructure | Securing Data Storage and Transaction logs | Logging sensor data is essential for tracking packages. They should be kept in secure data stores. |
| Key Management | For encrypted data. |
| Security Best Practices for non-relational data stores | Diversity of sensor types and data types may necessitate use of non-relational data stores. |
| Security against DoS attacks |  |
| Data Provenance | Meta-data should be cryptographically attached to the collected data, so that the integrity of origin and progress can be ensured. |
|  |  |  |
| General | Analytics for security intelligence | Anomalies in sensor data can indicate tampering/fraudulent insertion of data traffic. |
| Event detection | Abnormal events like 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 post facto. |

## Nielsen Homescan

Family level retail transactions and associated media exposure utilizing a statistically valid national sample. A general description [8] is provided by the vendor. This project description is based on the 2006 architecture.

**Mapping to the Security Reference Architecture**

|  |  |  |
| --- | --- | --- |
| RA Component | Security & Privacy Topic | Use Case Mapping |
| Sources → Transformation | End-Point Input Validation | Device-specific keys from digital sources; receipt sources scanned internally and reconciled to family ID . (Role issues) |
| Real Time Security Monitoring | None |
| Data Discovery and Classification | Classifications based on data sources (e.g.,retail outlets, devices, paper sources) |
| Secure Data Aggregation | Aggregated into demographic crosstabs. Internal analysts had access to PII. |
|  |  |  |
| Transformation → Uses | Privacy-preserving Data Analytics | Aggregated to (sometimes) product-specific statistically valid independent variables |
| Compliance with Regulations | Panel data rights secured in advance & enforced through organizational controls |
| Govt access to data and freedom of expression concerns | N/A |
|  |  |  |
| Transformation ↔ Data Infrastructure | Data Centric Security such as identity/policy-based encryption | Encryption not employed in place; only for data center to data center transfers. XML cube security mapped to Sybase IQ, 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, process step audits |
|  |  |  |
| Data Infrastructure | Securing Data Storage and Transaction logs | Project-specific audits secured by infrastructure team  |
| Key Management | Managed by project CSO. Separate key pairs issued for customers, internal users |
| Security Best Practices for non-relational data stores | Regular data Integrity checking via XML schema validation |
| Security against DoS attacks | Industry standard webhost protection provided for query subsystem.  |
| Data Provenance | Unique  |
|  |  |  |
| General | Analytics for security intelligence | No project-specific initiatives |
| Event detection | N/A |
| Forensics | Usage, cube-creation, device merge audit records were retained for forensics & billing. |

## Pharma Clinical Trial Data Sharing

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

**Mapping to the Security Reference Architecture:**

|  |  |  |
| --- | --- | --- |
| RA Component | Security & Privacy Topic | Use Case Mapping |
| Sources → Transformation | End-Point Input Validation | Opaque – company-specific |
| Real Time Security Monitoring | None  |
| Data Discovery and Classification | Opaque – company-specific |
| Secure Data Aggregation | 3rd party aggregator |
|  |  |  |
| Transformation → Uses | Privacy-preserving Data Analytics | Data to be reported in aggregate but preserving potentially small-cell demographics |
| Compliance with Regulations | Responsible developer & 3rd party custodian |
| Govt access to data and freedom of expression concerns | None considered: research limited community use; possible future public health data concern. Clinical Study Reports only, but possibly selectively at study-, patient-level |
|  |  |  |
| Transformation ↔ Data Infrastructure | Data Centric Security such as identity/policy-based encryption | TBD |
| Policy management for access control | Internal roles; 3rd party custodian roles; researcher roles; participating patients’ physicians |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | TBD |
| Audits | Release audit by 3rd party |
|  |  |  |
|  | Securing Data Storage and Transaction logs | TBD |
| Key Management | Internal varies by firm; external TBD |
| Security Best Practices for non-relational data stores | TBD |
| Security against DoS attacks | Unlikely to become public |
| Data Provenance | TBD – critical issue |
|  |  |  |
| General | Analytics for security intelligence | TBD |
| Event detection | TBD |
| Forensics |  |

## Large Network Cybersecurity SIEM

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

**Mapping to the Security Reference Architecture:**

|  |  |  |
| --- | --- | --- |
| RA Component | Security & Privacy Topic | Use Case Mapping |
| Sources → Transformation | End-Point Input Validation | Software-supplier specific; e.g., [9] |
| Real Time Security Monitoring |  |
| Data Discovery and Classification | Varies by tool, but classifies based on security semantics, sources |
| Secure Data Aggregation | Varies: subnet, workstation, server |
|  |  |  |
| Transformation → Uses | Privacy-preserving Data Analytics | Platform-specific; example: Windows groups  |
| Compliance with Regulations | Applicable, but regulated events not readily visible to analysts |
| Govt access to data and freedom of expression concerns | NSA, FBI access on demand |
|  |  |  |
| Transformation ↔ Data Infrastructure | Data Centric Security such as identity/policy-based encryption | Usually feature of O.S. |
| Policy management for access control | E.g.: Windows group policy for event log |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | Vendor, platform-specific |
| Audits | Complex – audits possible throughout |
|  |  |  |
| Data Infrastructure | Securing Data Storage and Transaction logs | Vendor, platform-specific |
| Key Management | CSO, SIEM product keys |
| Security Best Practices for non-relational data stores | TBD |
| Security against DoS attacks | N/A |
| Data Provenance | E.g., how know an intrusion record was actually associated w/ specific workstation |
|  |  |  |
| General | Analytics for security intelligence | Feature |
| Event detection | Feature |
| Forensics | Feature |

## Consumer Digital Media Usage

Content owners license data for usage by consumers through presentation portals, e.g., Netflix, iTunes, etc. Usage is Big Data, including demographics at user level, patterns of use such as play sequence, recommendations, content navigation.

**Mapping to the Security Reference Architecture:**

|  |  |  |
| --- | --- | --- |
| RA Component | Security & Privacy Topic | Use Case Mapping |
| Sources → Transformation | End-Point Input Validation | Varies, vendor-dependent. Spoofing is possible. E.g., Protections afforded by securing Microsoft Rights Management Services [10]. S/MIME |
| Real Time Security Monitoring | Content creation security  |
| Data Discovery and Classification | Discovery / classification possible across media, populations, channels |
| Secure Data Aggregation | Vendor-supplied aggregation services – security practices opaque |
|  |  |  |
| Transformation → Uses | Privacy-preserving Data Analytics | Aggregate reporting to content owners |
| Compliance with Regulations | PII disclosure issues abound |
| Govt access to data and freedom of expression concerns | Various issues, e.g, playing terrorist podcast, illegal playback |
|  |  |  |
| Transformation ↔ Data Infrastructure | Data Centric Security such as identity/policy-based encryption | unknown |
| Policy management for access control | User, playback admin, library maintenance, auditor |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | Unknown |
| Audits | Audit DRM usage for royalties |
|  |  |  |
| Data Infrastructure | Securing Data Storage and Transaction logs | unknown |
| Key Management | unknown |
| Security Best Practices for non-relational data stores | unknown |
| Security against DoS attacks | N/A? |
| Data Provenance | Traceability to right entities to be preserved. (Add’l use case: Wikipedia privacy issues when distributing data to researchers) |
|  |  |  |
| General | Analytics for security intelligence | Machine intelligence for unsanctioned use/access |
| Event detection | “Playback” granularity defined |
| Forensics | Subpoena of playback records in legal disputes |

## Unmanned Military Vehicle Sensor Systems

Unmanned vehicles (“drones”) and their onboard sensors (e.g., streamed video) can produce petabytes of data that must be stored in nonstandard formats. Refer to DISA large data object contract [11] for exabytes in DoD private cloud.

**Mapping to the Security Reference Architecture:**

|  |  |  |
| --- | --- | --- |
| RA Component | Security & Privacy Topic | Use Case Mapping |
| Sources → Transformation | End-Point Input Validation | Need to secure sensor (e.g., camera) to prevent spoofing/stolen sensor streams. New transceivers, protocols in DoD pipeline. Sensor streams to include smartphone, tablet sources |
| Real Time Security Monitoring | On-board & 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 [12] unsecured, unaggregated. |
|  |  |  |
| Transformation → Uses | Privacy-preserving Data Analytics | Geospatial constraints: cannot surveil beyond a UTM. Military secrecy: target, point of origin privacy. |
| Compliance with Regulations | Numerous. Also standards issues. |
| Govt access to data and freedom of expression concerns | See Google lawsuit over Street View. |
|  |  |  |
| Transformation ↔ Data Infrastructure | Data Centric Security such as identity/policy-based encryption | Policy-based encryption, often dictated by legacy channel capacity/type |
| Policy management for access control | Transformations tend to be made within DoD-contractor devised system schemes.  |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | Sometimes performed within vendor-supplied architectures, or by image-processing parallel architectures. |
| Audits | CSO, IG audit |
|  |  |  |
| Data Infrastructure | Securing Data Storage and Transaction logs | The usual, plus data center security levels are tightly managed (e.g., field vs. battalion vs. HQ) |
| Key Management | CSO – chain of command |
| Security Best Practices for non-relational data stores | Not handled differently at present; this is changing in DoD. |
| Security against DoS attacks | DoD anti-jamming e-measures. |
| Data Provenance | Must track to sensor point in time configuration, metadata. |
|  |  |  |
| General | Analytics for security intelligence | DoD develops specific field of battle security software intelligence – event driven, monitoring – often remote. |
| Event detection | E.g.: target identification in a video stream, infer height of target from shadow. Fuse data from satellite IR with separate sensor stream. |
| Forensics | Used for AAR (after action review) – desirable to have full playback of sensor streams. |

## Common Core K-12 Student 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.

**Mapping to the Security Reference Architecture:**

|  |  |  |
| --- | --- | --- |
| RA Component | Security & Privacy Topic | Use Case Mapping |
| Sources → Transformation | End-Point Input Validation | Application-dependent. Spoofing is possible. |
| Real Time Security Monitoring | Vendor-specific monitoring of tests, test-takers, administrators & data. |
| Data Discovery and Classification | unknown |
| Secure Data Aggregation | Typical: Classroom level  |
|  |  |  |
| Transformation → Uses | Privacy-preserving Data Analytics | Various: e.g., teacher level analytics across all same-grade classrooms. |
| Compliance with Regulations | Parent-, student-, taxpayer disclosure & privacy rules apply |
| Govt access to data and freedom of expression concerns | Yes. May be required for grants, funding, performance metrics for teachers, administrators, districts. |
|  |  |  |
| Transformation ↔ Data Infrastructure | Data Centric Security such as identity/policy-based encryption | Support both individual access (student) & partitioned aggregate  |
| Policy management for access control | Vendor (e.g., Pearson) controls, state level policies, federal level policies; probably 20-50 roles? |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | unknown |
| Audits | Support 3rd party audits by unions, state agencies, resp to subpoenas |
|  |  |  |
| Data Infrastructure | Securing Data Storage and Transaction logs | Large enterprise security, trx controls – classroom to Feds |
| Key Management | CSO’s from classroom level to national |
| Security Best Practices for non-relational data stores | unknown |
| Security against DoS attacks | standard |
| Data Provenance | Traceability to measurement event requires capturing tests @ point in time |
|  |  |  |
| General | Analytics for security intelligence |  |
| Event detection |  |
| Forensics |  |

## Web Traffic Analytics

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

**Mapping to the Security Reference Architecture:**

|  |  |  |
| --- | --- | --- |
| RA Component | Security & Privacy Topic | Use Case Mapping |
| Sources → Transformation | End-Point Input Validation | Device-dependent. Spoofing often easy. |
| Real Time Security Monitoring | Webserver monitoring |
| Data Discovery and Classification | Some geospatial attribution |
| Secure Data Aggregation | Aggregation to device, visitor, button, web event, others |
|  |  |  |
| Transformation → Uses | Privacy-preserving Data Analytics | IP anonymizing, timestamp degrading. Content-specific opt-out. |
| Compliance with Regulations | Anonymization may be required for EU compliance. Opt-out honoring. |
| Govt access to data and freedom of expression concerns | Yes.  |
|  |  |  |
| Transformation ↔ Data Infrastructure | Data Centric Security such as identity/policy-based encryption | Varies depending on archivist. E.g., Adobe Omniture |
| Policy management for access control | System-, application-level access controls |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | unknown |
| Audits | Customer audits for accuracy, integrity supported |
|  |  |  |
| Data Infrastructure | Securing Data Storage and Transaction logs | Storage archiving – big issue |
| Key Management | CSO + 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 DOM, point-in-time marketing events |
|  |  |  |
| General | Analytics for security intelligence | Access to web logs often requires priv elevation. |
| Event detection | Can infer e.g., numerous sales, marketing & overall web health events |
| Forensics | See SIEM use case. |

## Health Information Exchange

Health Information Exchange data aggregated from various data providers that might include covered entities such as hospitals, and CROs identifying participation in clinical trials. The data consumers would include emergency room personnel, the CDC, and other authorized health (or other) organizations. Since any city, or region might implement its own HiE, these might also serve as data consumers and data providers for each other.

**Mapping to the Security Reference Architecture:**

|  |  |  |
| --- | --- | --- |
| RA Component | Security & Privacy Topic | Use Case Mapping |
| Sources → Transformation | End-Point Input Validation | Strong authentication, perhaps through X.509v3 certificates, potential leverage of SAFE bridge in lieu of general PKI. |
| Real Time Security Monitoring | Validation of incoming records to ensure integrity through signature validation, and HIPAA privacy through ensuring PHI is encrypted. May need to check for evidence of Informed Consent. |
| Data Discovery and Classification | Leverage 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 policies of data source, or HiE Service Provider. |
| Secure Data Aggregation | Clear text columns can be de-duplicated, perhaps columns with deterministic encryption. Other columns may have cryptographic metadata for facilitating aggregation and de-duplication. We assume retention rules, but no disposition rules in the related areas of Compliance.  |
|  |  |  |
| Transformation → Uses | Privacy-preserving Data Analytics | Searching on Encrypted Data, Proofs of Data Possession. Identification of potential adverse experience due to Clinical Trial Participation. Identification of potential Professional Patients. Trends and epidemics, co-relations of these to environmental and other effects. Determine if drug to be administered will generate an adverse reaction, without breaking the double blind. Patient 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. To facilitate this, and the logging and alerts, will require federated identity integration with Data Consumers. |
| Govt access to data and freedom of expression concerns | CDC, Law Enforcement, Subpoenas and Warrants. Access may be toggled on based on occurrence of a pandemic (ex: CDC) or receipt of a warrant (Law Enforcement).  |
|  |  |  |
| Transformation ↔ Data Infrastructure | Data Centric Security such as identity/policy-based encryption | Row-level and Column-level Access Control. |
| Policy management for access control | Role-based and Claim-based. Defined for PHI cells.  |
| Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption | Privacy preserving access to relevant events, anomalies and trends, to CDC and other relevant health organizations.  |
| Audits | Facilitate HIPAA readiness, and HHS audits. |
|  |  |  |
| Data Infrastructure | Securing Data Storage and Transaction logs | Need to be protected for integrity and for privacy, but also for establishing completeness, with an emphasis on availability.  |
| Key Management | Federated across Covered Entities, with need to manage key lifecycles 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 but without compromising patient privacy. |
| Security against DoS attacks | Mandatory – Availability is Compliance Requirement. |
| Data Provenance | Completeness and integrity of data with records of all accesses and modifications. This information could be as sensitive as the data, and is subject to commensurate access policies.  |
|  |  |  |
| General | Analytics for security intelligence | Monitoring of Informed Patient consent; authorized and unauthorized transfers, 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. |

# References

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

[2] Retail research and benchmarking and design of an integrated solution a person centered design “A problem or do you have a wicked problem?”

[3] Phrma.org, “Principles for Responsible Clinical Trial Data Sharing”, <http://phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf>

[4] Army Regulation 25-2, <http://www.apd.army.mil/jw2/xmldemo/r25_2/main.asp>

[5] Wall Street Journal, “Before Tougher State Tests, Officials Prepare Parents”, [http://blogs.wsj.com/metropolis/2013/04/15/before-tougher-state-tests-officials-prepare-parents/](http://blogs.wsj.com/metropolis/2013/04/15/before-tougher-state-tests-officials-prepare-parents/?mod=dist_smartbrief)

[6] InformationWeek, “Common Core Meets Aging Education Technology”,
<http://www.informationweek.com/big-data/news/common-core-meets-aging-education-techno/240158684>

[7] Civitas Learning, <http://www.civitaslearning.com/about/>

[8] Nielsen, “Consumer Panel and Retail Measurement”, <http://www.nielsen.com/us/en/nielsen-solutions/nielsen-measurement/nielsen-retail-measurement.html>

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

[10] Microsoft, “Deploying Windows Rights Management Services at Microsoft”, <http://technet.microsoft.com/en-us/library/dd277323.aspx>

[11] GCN, “DISA plans for exabytes of drone, satellite data”, <http://gcn.com/articles/2013/04/12/disa-plans-exabytes-large-data-objects.aspx>

[12] DefenseSystems, “UAV video encryption remains unfinished job”, <http://defensesystems.com/articles/2012/10/31/agg-drone-video-encryption-lags.aspx>

##  Actors (Draft)

A person has relationships with many applications and sources of information in a Big Data system. The diagram below illustrates these system interactions and resulting roles.



We describe a number of instances to exemplify this assertion. (Note that each of these roles is “point in time.”) :

1. A retail organization refers to a person as a consumer before a purchase; afterwards, they are deemed a customer.
2. A person has a customer relationship with a financial organization in either prepaid or personal banking services.
3. A person may have a car or auto loan with a different or same financial institution.
4. A person may have a home loan with a different or same bank as a personal bank or each may be different organizations for the same person.
5. A person may be “the insured” on health, life, auto, homeowner or renters insurance.
6. 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.
7. A person has been educated by one or more public or private schools
8. A person may be an employee, temporary worker, contractor or third party employee for one or more private or public enterprises.
9. A person may be underage and have special conditions around collection of data.

##  Classification of Security and Privacy Topics (Draft)

The set of topics were initially adapted from the scope of the CSA BDWG charter, organized according to the classification in [1]. Security and Privacy concerns are classified in 4 categories:

1. Infrastructure Security
2. Data Privacy
3. Data Management
4. Integrity and Reactive Security

In this section, we describe the topics in detail. Rather than a prescriptive document at this stage, this text reproduces community discussion. Later versions will refine and organize as needed.

**Infrastructure Security**

1. Review of technologies and frameworks that have been primarily developed for performance, scalability and availability; e.g., Apache Hadoop, MPP databases, etc.
2. High-availability
	1. Security against Denial-of-Service (DoS) attacks.

**Data Privacy**

1. Impact of social data revolution on security and privacy of big data implementations. [*discussion*]
	1. Unknowns of Innovation - When a perpetrator, abuser or stalker misuses technology to target and harm a victim, there are various criminal and civil charges that might be applied to ensure accountability and promote victim safety. There are a number of U.S. federal and state/territory/tribal laws that might apply. To support the safety and privacy of victims, it is important to take technology-facilitated abuse and stalking seriously. This includes assessing all ways that technology is being misused to perpetrate harm, and, considering all charges that could or should be applied.
	2. Identify laws that address violence and abuse. Identify where they explicitly or implicitly include the use of technology and electronic communications:
2. Stalking and cyberstalking (felony menacing by, via electronic surveillance, etc.)
3. Harassment, threats, assault
4. Domestic violence, dating violence, sexual violence, sexual exploitation
5. Sexting and child pornography: electronic transmission of harmful information to minors, providing obscene material to a minor, inappropriate images of minors, lascivious intent
6. Bullying and cyberbullying
7. Child abuse
	1. Identify possible criminal or civil charges related to technology, communications, privacy and confidentiality:
8. Unauthorized access, unauthorized recording/taping, Illegal interception of electronic communications, illegal monitoring of communications, surveillance, eavesdropping, wiretapping, unlawful party to call
9. Computer and Internet crimes: fraud, network intrusion
10. Identity theft, impersonation, pretexting
11. Financial fraud, telecommunications fraud
12. Privacy violations
13. Consumer protection laws
14. Violation of no contact, protection and restraining orders
15. Technology misuse: Defamatory libel, slander, economic or reputational harms, privacy torts
16. Burglary, criminal trespass, reckless endangerment, disorderly conduct, mischief, obstruction of justice.
17. Data-centric security to protect data no matter where it is stored or accessed
	1. For example, attribute-based encryption, format-preserving encryption
18. Big data privacy and governance
	1. Data discovery and classification
	2. Policy management for accessing and controlling Big Data
		1. For example, new policy language frameworks specific to Big Data architectures
	3. Data masking technologies: anonymization, rounding, truncation, hashing, differential privacy
		1. It is important to consider how these approaches degrade performance or hinder delivery all together. Often these solutions are proposed and then cause an outage at the time of the release forcing the removal of the option. [*discussion*]
	4. Data monitoring
	5. Compliance with regulations such as HIPAA, EU data protection regulations, APEC Cross-Border Privacy Rules (CBPR) requirements, and country-specific regulations
		1. Regional data stores enable regional laws to be enforced
			1. Cybersecurity Executive Order 1998 - assumed data and information would remain within the region.
		2. People centered design makes the assumption that private sector stakeholders are operating ethically and respecting the freedoms and liberties of all Americans. [*discussion*]
			1. Litigation, including class action suits, could follow increased threats to Big Data security, when compared to other systems
				1. People before profit must be revisited to understand the large number of Executive Orders overlooked [*discussion*]
				2. People before profit must be revisited to understand the large number of domestic laws overlooked. [*discussion*]
			2. Indigenous and Aboriginal people and privacy of all associated vectors and variables must be excluded from any big data store, in any case a person must opt in. [*discussion*]
				1. all tribal land is an exclusion from any image capture and video streaming or capture.
				2. Human Rights
	6. Government access to data and freedom of expression concerns
		1. Polls show that U.S. citizens less concerned about loss of privacy than Europeans, but both are concerned about data misuse or inability to govern private and public sector use.
			1. In Cisco’s Internet of Everything, a project directly dependent on Big Data, a survey shows respondents worry over “threats to data (loss) and fear for physical safety.”



* 1. Potentially unintended/unwanted consequences or uses
		1. Appropriate uses of data collected or data aggregation and problem management capabilities must be enabled. [*discussion*]
		2. Mechanisms for the appropriate secondary or subsequent data uses.
			1. Filtered upon entry processed and presented in the inbound framework.
	2. Issues surrounding permission to collect data, consent and privacy
		1. If Facebook or Google permissions are marked ONLY MY FRIENDS, ONLY ME or ONLY MY CIRCLES the assumption must be that the person believes the setting in Facebook and Google control all content presented through Google and Facebook. How should this problem be addressed? Is it a Big Data issue? [*discussion*]
			1. Permission based on clear language and not forced by preventing users to access their online services.
			2. People do not believe the government would allow business people to take advantage of their rights. [*discussion*]
	3. Data Deletion: Responsibility to purge data based on certain criteria and/or events
		1. Examples include legal rulings that affect an external data source. Let’s say that Facebook loses a legal challenge and one of the outcomes is that Facebook must purge their databases of certain private information. Is there then a responsibility for downstream data stores to follow suit and purge their copies of the same data? Absolutely, the provider, producer, collector or social media supplier or the host must inform and remove all versions. Enforcement? Verification? [*discussion*]

e. Computing on encrypted data

1. De-duplication of encrypted data
2. Searching and reporting on the encrypted data
3. Fully homomorphic encryption
4. Anonymization of data (no linking fields to reverse identify)
5. De- identification of data (individual centric)
6. Non-identifying data (individual and context centric)
7. Secure data aggregation
8. Data Loss Prevention
9. Fault Tolerance - recovery for zero data loss
	1. Aggregation in end-to-end scale of resilience, record and operational scope for integrity and privacy in a secure or better risk management strategy.
	2. Fewer applications will require fault tolerance with clear distinction around risk and scope of the risk.

**Data Management**

1. Securing data stores
	1. Communication protocols
		1. DBLINKS
		2. ACL
		3. API
		4. Channel segmentation
		5. Federated (eRate) migration to cloud
	2. Attack surface reduction
2. Key management, and ownership of data
	1. Providing full control of the keys to the data owner
	2. Transparency of data lifecycle process: acquisition, uses, transfers, dissemination, destruction
	3. Maps to aid a non-technical person in seeing who and how their data is being used, including custody over time

**Integrity and Reactive Security**

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