**NIST Big Data**

**Security and Privacy Requirements**

**Version 1.0**

August 13, 2013

**Security & Privacy Subgroup**

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

**August, 2013**

[Executive Summary 3](#_Toc364192093)

[1 Introduction 3](#_Toc364192094)

[1.1 Objectives 3](#_Toc364192095)

[1.2 How This Report Was Produced 4](#_Toc364192096)

[1.3 Structure of This Report 4](#_Toc364192097)

[2 Scope 5](#_Toc364192098)

[2.1 General 5](#_Toc364192099)

[2.2 Infrastructure Security 5](#_Toc364192100)

[2.3 Data Privacy 5](#_Toc364192101)

[2.4 Data Management 7](#_Toc364192102)

[2.5 Integrity and Reactive Security 7](#_Toc364192103)

[3 Use Cases 9](#_Toc364192104)

[3.1 Retail (consumer) 9](#_Toc364192105)

[3.2 Healthcare 10](#_Toc364192106)

[3.3 Media 11](#_Toc364192107)

[2.1.1 Social Media 11](#_Toc364192108)

[2.2.1 Communications 11](#_Toc364192109)

[3.4 Government 12](#_Toc364192110)

[2.3.1 Military 12](#_Toc364192111)

[2.4.1 Justice Systems 13](#_Toc364192112)

[2.5.1 Healthcare 13](#_Toc364192113)

[2.6.1 Education 13](#_Toc364192114)

[2.7.1 Transportation 14](#_Toc364192115)

[2.8.1 Environmental 14](#_Toc364192116)

[2.9.1 Housing 14](#_Toc364192117)

[2.10.1 Labor 14](#_Toc364192118)

[2.11.1 Private Sector 14](#_Toc364192119)

[2.12.1 Financial Market 14](#_Toc364192120)

[3.5 Marketing 14](#_Toc364192121)

[4 Internal Security Practices 16](#_Toc364192122)

[5 Taxonomy of Security and Privacy Topics 17](#_Toc364192123)

[5.1 Taxonomy of Technical Topics 17](#_Toc364192124)

[5.2 Privacy 17](#_Toc364192125)

[5.3 Provenance 17](#_Toc364192126)

[5.4 System Health 17](#_Toc364192127)

[6 Security Reference Architecture 18](#_Toc364192128)

[6.1 Architectural Component: Interface of Sources $\rightarrow $ Transformation 18](#_Toc364192129)

[6.2 Architectural Component: Interface of Transformation $\rightarrow $ Uses 18](#_Toc364192130)

[6.3 Architectural Component: Interface of Transformation $\leftrightarrow $ Data Infrastructure 18](#_Toc364192131)

[6.4 Architectural Component: Internal to Data Infrastructure 18](#_Toc364192132)

[6.5 Architectural Component: General 18](#_Toc364192133)

[7 References 19](#_Toc364192134)

# Executive Summary

# Introduction

## Objectives

1. A distinction needs to be made between fault tolerance and security.
	1. Fault tolerance is resistance to unintended accidents.
	2. Security is resistance to malicious actions.
2. Big data is gathered from diverse end-points. So there are more types of actors than just Provider and Consumers – viz. Data Owners: for example, mobile users, social network users and so on.
	1. *A person has relationships with many applications and sources of information in a big data system.*
		1. *A retail organization refers to a person who “may” buy goods or services as a consumer, before the purchase and a customer after a purchase.*
		2. *A retail organization may use a social media platform as a channel for their online store.*
		3. *The person may be a patron at a food and beverage organization or as few as none and as many as 3 before a warning may need to be triggered.*
		4. *A person has a customer relationship with a financial organization in either prepaid or personal banking services.*
		5. *A person may have a car or auto loan with a different or same financial institution.*
		6. *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.*
		7. *A person may be “the insured” on health, life, auto, homeowner or renters insurance.*
			1. *A person may be the beneficiary or future insured person by an employer payroll deductions through a payroll service in the private sector employment development department in the public sector.*
		8. *A person has been educated by many or few educational organizations in either public or private schools for the first 15=20 years of their childhood making the right decisions.*
3. Data aggregation and dissemination have to be made securely and inside the context of a formal, understandable framework. This should part of the contract that has to be provided to Data Owners.
4. Availability of data to Data Consumers is an important aspect in Big Data.
	1. Availability can be maliciously affected by Denial of Service (DoS) attacks.
5. Searching and Filtering of Data is important since all of the massive amount of data need not be accessed.
	1. What are the capabilities provided by the Provider in this respect?
6. The balance between privacy and utility needs to be thoroughly analyzed. Big Data is most useful when it can be analyzed for information.
	1. However, privacy would restrict the form and availability of data to analytics technologies.
7. Since there is a separation between Data Owner, Provider and Data Consumer, the integrity of data coming from end-points has to be ensured.
	1. Data poisoning has to be ruled out.

## How This Report Was Produced

## Structure of This Report

# Scope

This initial list is 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

## General

a. Risk and threat models for big data

## Infrastructure Security

a. Review of technologies and frameworks that have been primarily developed for performance, scalability and availability. (e.g., Apache Hadoop, MPP databases, etc.,)

b. High-availability

i. Security against Denial-of-Service (DoS) attacks.

## Data Privacy

a. Impact of social data revolution on security and privacy of big data implementations.

b. Flexible policy management for accessing and controlling the data

i. For example, language framework for big data policies

c. Data-centric security to protect data no matter where it is stored or accessed in the cloud

i. For example, attribute-based encryption, format-preserving encryption

d. Big data privacy and governance

i. Data discovery and classification

ii. Data masking technologies: anonymization, rounding, truncation, hashing, differential privacy

iii. Data monitoring)

iv. 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. *Cyber-security Executive order 1998 - assumed data and information would remain within the region.*
			2. *People centered design makes the assumption that private sector stakeholders are being good Americans.*
				1. *If a presidential order is not enough to stop private sector stakeholders from putting Americans’ information in the hands of foreign threats.*

v. Government access to data and freedom of expression concerns

* + - 1. *I believe people in general are not nearly concerned about the freedom of expression as they are about misuse or inability to govern private sector use.*

vi. Potentially unintended/unwanted consequences or uses

vii. Appropriate uses of data collected or data in possession.

* + - * 1. *No way to enforce this even if we could define it.*

viii. Mechanisms for the appropriate secondary or subsequent data uses.

vii. Permission to collect data, (opt in/opt out), consent,

* + - 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.*
				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.*

 viii. 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. Isn’t there then a responsibility for downstream data stores to follow suit and purge their copies of the same data?

e. Computing on the encrypted data

i. De-duplication of encrypted data

ii. Searching and reporting on the encrypted data

iii. Fully homomorphic encryption

iv. Anonymization of data (no linking fields to reverse identify)

* + - 1. *Any use case supplied infers the ability to apply Java and perform create actions on data after download.*

v. De- identification of data (individual centric)

vi. Non-identifying data (individual and context centric)

* + - 1. Requires a person centered design strategy

f. Secure data aggregation

* + 1. *API in itself removes any pretense of securing data by aggregation.*
			1. *A db link behavior - full access to all information in the db or table*

g. Data Loss Prevention

* + 1. 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.*

iv. Anonymization of data (no linking fields to reverse identify)

* + - 1. *Any use case supplied infers the ability to apply Java and perform create actions on data after download.*

h. Data end of life (right of an individual to be forgotten)

* + - 1. *A double edge sword*

## Data Management

a. Securing data stores

i. Communication protocols

* + - 1. *DBLINKS*
			2. *ACL*
			3. *API*
			4. *Channel segmentation*
				1. *Federated (eRate) migration to cloud*

ii. Attack surface reduction

* + - 1. Fault tolerant

b. Key management, and ownership of data

i. Providing full control of the keys to the data owner

ii. Transparency of data lifecycle process: acquisition, uses, transfers, dissemination, destruction

* + - 1. *Maps to aid a non-technical person in seeing who and how the data is being used.*
				1. *No more - anonymous users stalking people on social media platforms.*

*LinkedIn*

## Integrity and Reactive Security

a. Big data analytics for security intelligence (identifying malicious activity) and situational awareness (understanding the health of the system)

* + 1. *Qualification and certification limited access privileges*
			1. *Specifically preventing private sector “experts” or monitoring through API’s*
			2. *Our county was unqualified to match fingerprints*
				1. *Bio-metric uses and abuses*

i. Data-driven abuse detection

* + - 1. *Very limited in who has the credentials*
			2. *Very limited on who makes assumptions*
			3. *Economically equal person’s*
				1. *an income median population in higher income groupings has little “actual” knowledge and no wisdom to make assumptions about a low median population.*
				2. *Southern states family services remove countless children from their families homes based on different “norms”.*

*A person with the same background should be consulted first, before allowing others to apply stereotypes or profile another citizen.*

ii. 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.*

iii. Streaming data analytics

* + - * 1. *REQUIRES a virtual machine and secure channel*
				2. *This is a low level requirement or phase iii*

*roadmap*

*priority on secure and return on investment must be done to move to this degree of maturity*

b. Event detection

* + 1. *Re-act upon deep analysis due to the API and ability for physical table source>work>target during work the information is transformed and reloaded.*
	1. Forensics
		1. *an API using a virtual server on a personal computer*
			1. *three directories created by an unknown person*
			2. *the three directories were non-view or unknown to the users*
				1. *bouncing many from foreign countries with china accused but clock reading middle eastern time zone.*

d. Security of analytics results

* + 1. *What and who defines SECURITY for this purpose?*

make a disc backup to recovery “if no java was inserted and no sql

# Use Cases

## Retail (consumer)

* 1. Scenario 1
		1. Current Method for Security and Privacy
			1. none in place
		2. Gaps, if any
			1. Retailers are padding their performance
				1. point of sales process
				2. Double counting in financial sector
		3. [Current Research](https://docs.google.com/file/d/0B7-jqY_f5N_1TVNSbzN6V3hBbFU/edit?usp=sharing)
			1. *no privacy nor security beyond traditional types*
				1. *new threats the API feeds into many*
				2. *new threats or unknown threats if we no longer require two forms of ID.*

*more fraud with less safeguards*

*in response to known risk showing up on credit reports and when asking for audits of the erroneous charges*

*many consumers are left to argue with their creditors*

* + - * 1. *Grouping by education sector to ensure age appropriate content “ethical advertising and marketing”*
				2. *Social responsibility*
	1. Scenario 2
	2. Retail and customer footprint with segmentation
		1. Scenario: Nielsen Homescan [MAU]
			1. Scenario description: This is a subsidiary of Nielsen that collects family level retail transactions. Transaction = checkout receipt, containing all SKUs purchased, time, date, store location. 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.
			2. Current S&P
				1. Data is in-house but shared with customers who have partial access to data partitions through web portions using columnar data bases. Other Cx 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.
				2. PII data is considerable. Survey participants are compensated in exchange for giving up segmentation data, demographics, etc.
			3. Gaps
				1. Opt-out scrubbing and custody audit not provided.
			4. Current Research: TBD

## Healthcare

* 1. Scenario 1 ‘Health Information Exchange’
		1. Current Method constraints to federation as needed
			1. *Private sector implementations*
				1. *an employer should NEVER have access to an employee's medical history nor the medical records of family members*

*Suddenly genetic diseases were cured mysteriously when an employee has had a heart condition since birth.*

*The same employee has a son with a severe sleep apnea condition that was suddenly cured, he has a neurological form not obstructive.*

*There is no cure only methods to make it less uncomfortable.*

*He mysteriously was cured according to his records.*

* + 1. Need for HIPAA-capable cryptographic controls and key management
			1. *virtual directories through machine to machine connectivity enables a threat to take ownership of a virtual drive one simply needs to copy the key file,*
				1. *When API begins use the key to extract well beyond the intended use.*

*appears to be a different person entirely*

* + 1. Sketchy, but technologies are ready for practical implementation
		2. Possible analogous scenario: Doximity, “a secure way for doctors to share research, clinical trial data, and patient records in the cloud.” Widely adopted already.
	1. Scenario 2 ‘Genetic Privacy’
		1. Data ownership
			1. *Who owns the data, the user who enters the information?*
				1. *Then when they mark “only my friends, only me on the account settings, do not present conflicting terms forcing a person to grant permission or prevent access to the platforms.*
		2. data uses
			1. *What use cases can be presented to ever allow private sector company use?*
			2. *What use cases can be presented to ever allow government use?*
		3. *Sub-scenario:* [*FreetheData*](http://www.geneticalliance.org/free-the-data)
	2. Scenario: Pharma Clinical Trial Data Sharing. ([Details](http://onphr.ma/18Ssl2u)) [MAU]
		1. Scenario Description: Under an industry trade group proposal, clinical trial data will be shared outside intra-enterprise warehouses. The EU and others have made competing proposals that differ in significant S&P respects.
		2. Current S&P
			1. The proposed S&P will require secured access by a public review board which will be different for each dataset owner.
			2. Custody is restricted to approved use, hence need for usage audit and security.
			3. Patient-level data disclosure - elective, per company (details in p. 3). The association mentions anonymization (“re-identification”) but mentions issues with small sample sizes.
			4. Study-level data disclosure - elective, per company (details, p. 3
			5. Publication restrictions: additional security will be required to ensure rights of publishers, e.g., Elsevier or Wiley
			6. Cloud vs. self-hosted data? Up to each firm.
		3. Gaps
			1. Standards for data sharing unclear
			2. Access by patients and public patient groups/advocates is a different use case than currently described
			3. Longitudinal custody beyond trial disposition unclear, especially after firms merge or dissolve
		4. Current Research: TBD

## Media

### Social Media

* + 1. Big Data Brokering
			1. Who authorized these private sectors to assume this role without considerations for our human rights?
		2. Entertainment

### Communications

* + 1. *Telecommunications*
		2. *Networks*
			1. Scenario: CyberSecurity [MAU]
				1. 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.
				2. Current S&P

Protections for intra-enterprise privacy and security are not generally honored, and perhaps not needed, but it is collected and thus if aggregated would contain employee / vendor / customer PII.

Traditional policy-type security prevails, though temporal dimension and monitoring of policy modification events tends to be nonstandard or unaudited.

Cybersecurity apps themselves run at high levels of security and thus require separate audit and security measures.

* + - * 1. Gaps

No cross-industry standards exist for aggregating data, beyond operating system collection methods.

* + - * 1. Current Research: TBD
	1. Marketing
	2. Scenario: Digital Media Usage by Consumers [MAU]
		1. Scenario Description: 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.
		2. Current method for security and privacy
			1. Silos exist within proprietary provider and owner networks. Protections within providers are conventional single-auth password. Protections between owners and provider networks are unknown to me.
		3. Gaps
			1. Standards unclear. Original DRM model not built to scale to meet demand for forecast use for the data. Data itself likely to be commercialized if not already, so patterns for privacy and security protection likely not addressed but will be important.
		4. Current research: TBD

## Government

### Military

* + 1. Scenario: Unmanned Vehicle Sensor Data [MAU]
			1. Scenario Description: Unmanned vehicles (“drones”) and their onboard sensors (e.g., streamed video) can produce petabytes of data which 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.
			2. Current Method for 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.
			3. 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 [Army guidelines](http://www.apd.army.mil/jw2/xmldemo/r25_2/main.asp).
				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.

### Justice Systems

* + 1. *Private Sector role*

### Healthcare

* + 1. *Private Sector role*

### Education

* + 1. Scenario: “Common Core” Student Performance Reporting [MAU]
			1. 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.
			2. 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](http://www.smartbrief.com/04/15/13/nyc-parents-told-brace-lower-test-scores-under-common-core-1#.UgJc4ZJ9DHQ), parents can opt students out, so that data must be collected.
			3. Research: TBD
				1. Longitudinal performance data would have value for program evaluators if data scales up.
				2. [Data-driven learning content administration](http://www.informationweek.com/big-data/news/common-core-meets-aging-education-techno/240158684) will require access to performance data at learner level, probably more often than at test time, and at higher granularity, thus requiring more data.
				3. Example enterprise: [Civitas Learning](http://www.civitaslearning.com/about/) Predictive analytics for student decision-making
		2. *Private Sector Role*

### Transportation

* + 1. Private Sector Role

### Environmental

* + 1. Energy
			1. Private Sector Role

### Housing

* + 1. *Private Sector role*

### Labor

* + 1. *Private Sector role*

### Private Sector

### Financial Market

## Marketing

* 1. Scenario: Web Traffic Analytics [MAU]
		1. 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.
		2. 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.
				1. Some companies allow for purging of data on demand, but it’s unlikely to expunge previously collected webserver traffic.
			2. 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.
		3. Current research: TBD

# Internal Security Practices

# Taxonomy of Security and Privacy Topics

## Taxonomy of Technical Topics

## Privacy

## Provenance

## System Health

# Security Reference Architecture

## Architectural Component: Interface of Sources $\rightarrow $ Transformation

## Architectural Component: Interface of Transformation $\rightarrow $ Uses

## Architectural Component: Interface of Transformation $\leftrightarrow $ Data Infrastructure

## Architectural Component: Internal to Data Infrastructure

## Architectural Component: General

# 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?”

1.