NIST BDWG Security and Privacy

Use Cases

# Consumer Digital Media Usage

**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).

**Mapping to 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 |

# 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

**Mapping to 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. |

# 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

**Mapping to the 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

**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.

**Mapping to the 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. |

# 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?

# 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

**Mapping to the 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 |  |

# Cyber-security

**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.

**Mapping to the 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 |

# 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.

**Mapping to the 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. |

# 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.

**Mapping to the 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 |  |

# Aviation Industry

**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**:

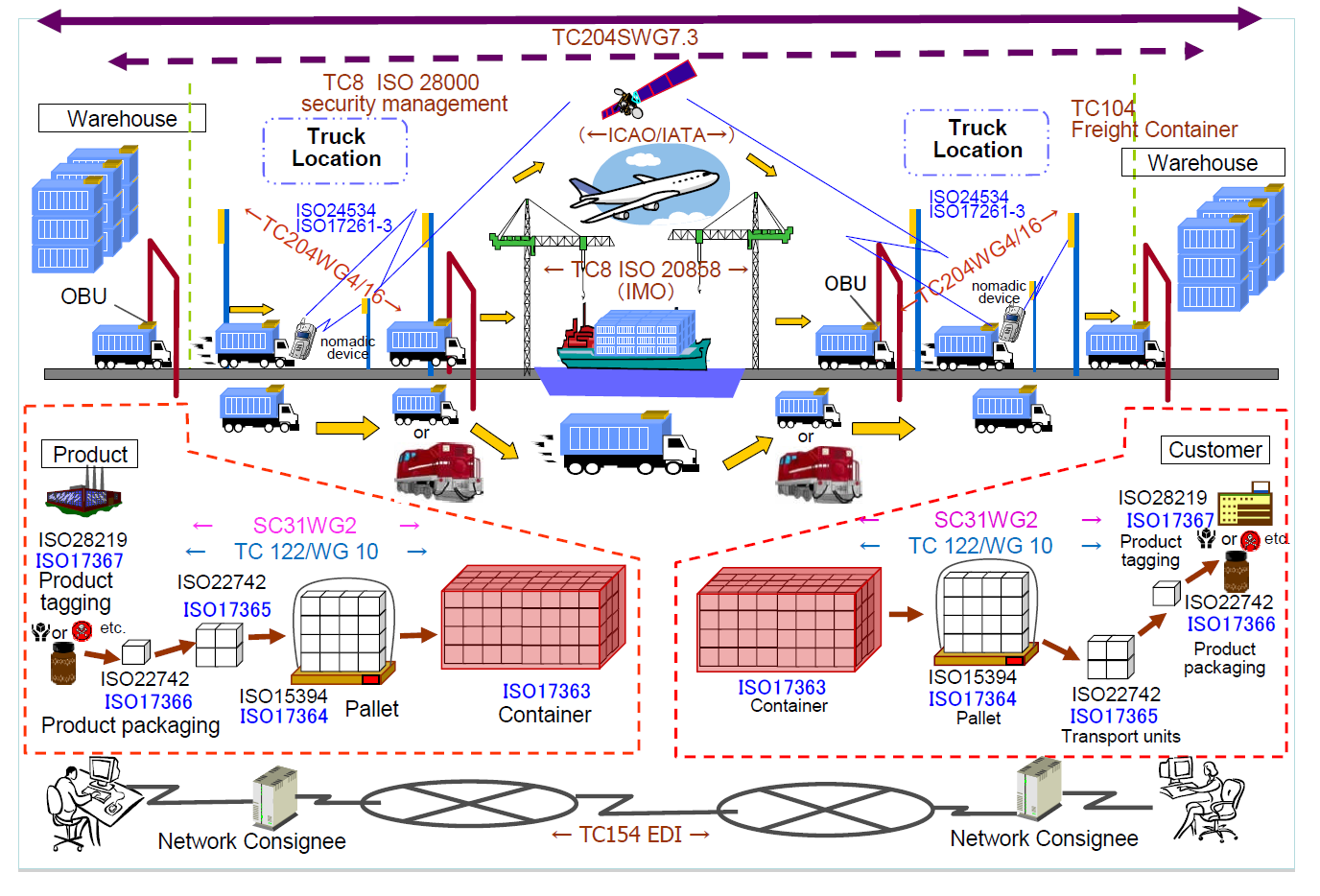
1. Encryption at rest: Big data systems need to encrypt data stored at the infra layer so that cloud storage admins cannot access that data.
2. Key management: The encryption key management should be architected such end customers (airliners) have sole/shared control on the release of keys for data decryption.
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

# 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 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. |

# 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/

[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

[13] Fishman, L. F. Correspondence from American Hospital Association to Patrick Gallagher (NIST), 2013-12-11, http://assets.fiercemarkets.com/public/newsletter/fiercehealthit/131211-cl-cybersecurity.pdf.

[14] Health Information Technology for Economic and Clinical Health (HITECH) Act, Public Law 111-5, dated February 17, 2009, http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/hitechact.pdf, retrieved 2013-12-01.

[16] P. Chen, B. Plale, and M. S. Aktas, "Temporal representation for mining scientific data provenance," Future Generation Computer Systems, Oct. 2013. [Online]. Available: http://dx.doi.org/10.1016/j.future.2013.09.032.

[17] S. S. Sahoo, A. Sheth, and C. Henson, "Semantic provenance for eScience: Managing the deluge of scientific data," Internet Computing, IEEE, vol. 12, no. 4, pp. 46-54, 2008. [Online]. Available: http://dx.doi.org/10.1109/MIC.2008.86

[18] N. B. Sayles, Ed., Health information management technology: an applied approach, 4th ed. American Health Information Management Assoc (AHIIMA), 2013.

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

[20] Distributed Common Ground System – Army, web, http://dcgsa.apg.army.mil/ .

[21] National Information Exchange Model, https://www.niem.gov/Pages/default.asp, accessed 2013-12-18.

[22] M. Kokar and G. Ng, "High-level information fusion and situation awareness," Information Fusion, vol. 10, no. 1, pp. 2-5, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.inffus.2008.05.003

[23] S. Auer, "The emerging web of linked data," in Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications, ser. ISWSA '11. New York, NY, USA: ACM, 2011. [Online]. Available: http://dx.doi.org/10.1145/1980822.1980823