Security and Privacy Requirements
March 18, 2014

Arnab Roy, Fujitsu
On behalf of the NIST BDWG S&P Subgroup
Process

The CSA Big Data Working Group Top 10 S&P Challenges
Googledoc with initial set of topics and solicitation of use cases
Taxonomy of topics
Input from Reference Architecture Group
Security Reference Architecture overlaid on RA
Mapping use cases to the SRA
Editorial phase
Current Working Draft (M0110)
CSA BDWG: Top Ten Big Data Security and Privacy Challenges

1) Secure computations in distributed programming frameworks
2) Security best practices for non-relational datastores
3) Secure data storage and transactions logs
4) End-point input validation/filtering
5) Real time security monitoring
6) Scalable and composable privacy-preserving data mining and analytics
7) Cryptographically enforced access control and secure communication
8) Granular access control
9) Granular audits
10) Data provenance
Top 10 S&P Challenges: Classification

**Infrastructure security**
- Secure Computations in Distributed Programming Frameworks
- Security Best Practices for Non-Relational Data Stores

**Data Privacy**
- Privacy Preserving Data Mining and Analytics
- Cryptographically Enforced Data Centric Security
- Granular Access Control

**Data Management**
- Secure Data Storage and Transaction Logs
- Granular Audits
- Data Provenance

**Integrity and Reactive Security**
- End-point validation and filtering
- Real time Security Monitoring
# Taxonomy – conceptual axis

## Privacy
- Communication Privacy
  - Data Confidentiality
    - Access Policies
  - Computing on Encrypted Data
    - Searching and Reporting
    - Fully Homomorphic Encryption
  - Secure Data Aggregation
  - Key Management

## Provenance
- End-point Input Validation
  - Syntactic Validation
  - Semantic Validation
- Communication Integrity
- Authenticated Computations on Data
  - Trusted Platforms
  - Crypto Enforced
- Granular Audits
- Control of Valuable Assets
  - Lifecycle Management
  - Retention, Disposition, Hold
  - Digital Rights Management

## System Health
- Security against DoS
  - Construction of cryptographic protocols proactively resistant to DoS
- Big Data for Security
  - Analytics for Security Intelligence
  - Data-driven Abuse Detection
  - Event Detection
  - Forensics
## Taxonomy – operational axis

### Big Data Security and Privacy

<table>
<thead>
<tr>
<th>Registration, Security Model and Policy Enforcement</th>
<th>Identity and Access Management</th>
<th>Data Governance</th>
<th>Visibility and Infrastructure Management</th>
<th>Risk and Accountability</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Device, User, Asset, Services, Applications registration</td>
<td>- Virtualization Layer Identity</td>
<td>- Encryption and Key Management (including Multi-Key)</td>
<td>- Threat and Vulnerability Management</td>
<td>- Accountability</td>
</tr>
<tr>
<td>- Security Metadata Model</td>
<td>- Application Layer Identity</td>
<td>- Isolation/Containerization</td>
<td>- Monitoring, Alerting</td>
<td>- Compliance</td>
</tr>
<tr>
<td>- Policy Enforcement</td>
<td>- End User Layer Identity Management</td>
<td>- Storage Security</td>
<td>- Mitigation</td>
<td>- Forensics</td>
</tr>
<tr>
<td></td>
<td>- Identity Provider</td>
<td>- Data Loss Prevention, Detection</td>
<td>- Configuration Management</td>
<td>- Business Risk Model</td>
</tr>
<tr>
<td></td>
<td>- Additional XACML Concepts</td>
<td>- Web Services Gateway</td>
<td>- Logging</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Data Transformation</td>
<td>- Malware Surveillance and Remediation</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Data Lifecycle Management</td>
<td>- Network Boundary Control</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>- End Point Input Validation</td>
<td>- Resiliency, Redundancy and Recovery</td>
<td></td>
</tr>
</tbody>
</table>
Use Cases

• Retail/Marketing
  – Modern Day Consumerism
  – Nielsen Homescan
  – Web Traffic Analysis

• Healthcare
  – Health Information Exchange
  – Genetic Privacy
  – Pharma Clinical Trial Data Sharing

• Cyber-security

• Government
  – Military
  – Education
Big Data Security Reference Architecture

- End-Point Input Validation
- Real Time Security Monitoring
- Data Discovery and Classification
- Secure Data Aggregation

- Data Centric Security such as identity/policy-based encryption
- Policy management for access control
- Computing on the encrypted data: searching/filtering/deduplicate/fully homomorphic encryption
- Granular audits
- Granular access control

- Privacy preserving data analytics and dissemination
- Compliance with regulations such as HIPAA
- Govt access to data and freedom of expression concerns

Security and Privacy
### Interface of Data Providers -> BD App Provider

<table>
<thead>
<tr>
<th>S&amp;P Consideration</th>
<th>Health Info Exchange</th>
<th>Military UAV</th>
</tr>
</thead>
<tbody>
<tr>
<td>End-Point Input Validation</td>
<td>Strong authentication, perhaps through X.509v3 certificates, potential leverage of SAFE bridge in lieu of general PKI</td>
<td>Need to secure sensor to prevent spoofing/stolen sensor streams</td>
</tr>
<tr>
<td>Real Time Security Monitoring</td>
<td>Validation of incoming records. May need to check for evidence of Informed Consent.</td>
<td>On-board &amp; control station secondary sensor security monitoring</td>
</tr>
<tr>
<td>Data Discovery and Classification</td>
<td>Leverage HL7 and other standard formats opportunistically, but avoid attempts at schema normalization.</td>
<td>Varies from media-specific encoding to sophisticated situation-awareness enhancing fusion schemes.</td>
</tr>
<tr>
<td>Secure Data Aggregation</td>
<td>Clear text columns can be deduplicated, perhaps columns with deduplication.</td>
<td>Fusion challenges range from simple to complex.</td>
</tr>
</tbody>
</table>
### Interface of BD App Provider -> Data Consumer

![Diagram showing the interface of BD App Provider to Data Consumer]

### S&P Consideration

<table>
<thead>
<tr>
<th>Privacy preserving data analytics and dissemination</th>
<th>Health Info Exchange</th>
<th>Military UAV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searching on encrypted data. Determine if drug administered will generate an adverse reaction, without breaking the double blind.</td>
<td>HIPAA security and privacy will require detailed accounting of access to HER data.</td>
<td>Geospatial constraints: cannot surveil beyond a UTM. Military secrecy: target, point of origin privacy.</td>
</tr>
<tr>
<td>Compliance with regulations</td>
<td></td>
<td>Numerous. Also standards issues.</td>
</tr>
<tr>
<td>Govt access to data and freedom of expression concerns</td>
<td>CDC, Law Enforcement, Subpoenas and Warrants. Access may be toggled based on occurrence of a pandemic or receipt of a warrant.</td>
<td>Google lawsuit over streetview.</td>
</tr>
</tbody>
</table>

### Security and Privacy
### Interface of BD App Provider -> BD Framework Provider

![Diagram showing the interface between Big Data Application Provider and Big Data Framework Provider]

<table>
<thead>
<tr>
<th>S&amp;P Consideration</th>
<th>Health Info Exchange</th>
<th>Military UAV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy based encryption</td>
<td>Row-level and Column-level Encryption</td>
<td>Policy-based encryption, often dictated by legacy channel capacity/type.</td>
</tr>
<tr>
<td>Policy management for access control</td>
<td>Role-based and claim-based</td>
<td>Transformations tend to be made within DoD-contractor devised system schemes.</td>
</tr>
<tr>
<td>Computing on encrypted data</td>
<td>Privacy preserving access to relevant events, anomalies and trends.</td>
<td>Sometimes performed within vendor-supplied architectures, or by image-processing parallel architectures.</td>
</tr>
<tr>
<td>Audits</td>
<td>Facilitate HIPAA readiness, and HHS audits</td>
<td>CSO, IG audit.</td>
</tr>
</tbody>
</table>
## Security and Privacy

### Internal to BD Framework Provider

<table>
<thead>
<tr>
<th>S&amp;P Consideration</th>
<th>Health Info Exchange</th>
<th>Military UAV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Securing Data Stores and Transaction Logs</td>
<td>Need to be protected for integrity and for privacy, but also for establishing completeness, with an emphasis on availability.</td>
<td>The usual, plus data center security levels are tightly managed (e.g., field vs. battalion vs. HQ).</td>
</tr>
<tr>
<td>Security Best Practices for non-relational data</td>
<td>End-to-end encryption.</td>
<td>Not handled differently at present; this is changing in DoD.</td>
</tr>
<tr>
<td>Security against DoS attacks</td>
<td>Mandatory – availability is a compliance requirement.</td>
<td>DoD anti-jamming e-measures.</td>
</tr>
<tr>
<td>Data Provenance</td>
<td>Completeness and integrity of data with records of all accesses and modifications</td>
<td>Must track to sensor point in time configuration, metadata.</td>
</tr>
</tbody>
</table>
Next Steps

- Taxonomy to Reference Architecture Mapping
- Scope for Standards
Next steps: Cryptography and Privacy Enhancing Technologies

• BIG
  – Scale up existing solutions for volume, variety and velocity
  – Retarget to Big Data infrastructural shift

• DATA
  – Balance privacy and utility
  – Enable analytics and governance on encrypted data
  – Reconcile authentication and anonymity
Top 10 Challenges in Crypto and PET identified by CSA BDWG

Infrastructure
- Communication protocols
- Key management

Encryption
- Access policy based encryption
- Searching / filtering encrypted data
- Secure outsourcing of computation

Privacy
- Secure dissemination
- Secure data collection / aggregation
- Secure collaboration

Data Governance
- Data integrity
- Proof of data storage
Thank you!