Security and Privacy Requirements
May 13, 2014

Arnab Roy, Fujitsu
On behalf of the NIST BDWG S&P Subgroup
S&P Requirements Emerging due to Big Data Characteristics

• Variety:
  – Traditional encryption schemes hinder organization of data based on semantics.

• Volume:
  – Threat models for multi-tiered data storages are complex and evolving.

• Velocity:
  – Distributed computing infrastructures and non-relational data storages require retargeting of traditional security mechanisms.
S&P Requirements Emerging due to Big Data Characteristics

• Veracity:
  – Keeping track and ensuring integrity of the ownership, source and other metadata of individual data is a complex and sophisticated requirement, given the movement of data between nodes, entities and geographical boundaries.

• Volatility:
  – Indefinitely persistent data requires evolving S&P considerations. With the passage of time, roles may evolve and governance may shift depending on the merger and disappearance of responsible organizations.
CSA BD 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

<table>
<thead>
<tr>
<th>Privacy</th>
<th>Provenance</th>
<th>System Health</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Communication Privacy</strong></td>
<td><strong>End-point Input Validation</strong></td>
<td><strong>Security against DoS</strong></td>
</tr>
<tr>
<td><strong>Data Confidentiality</strong></td>
<td><strong>Syntax Validation</strong></td>
<td><strong>Analytics for Security Intelligence</strong></td>
</tr>
<tr>
<td><strong>Computing on Encrypted Data</strong></td>
<td><strong>Semantic Validation</strong></td>
<td><strong>Data-driven Abuse Detection</strong></td>
</tr>
<tr>
<td><strong>Secure Data Aggregation</strong></td>
<td><strong>Authenticated Computations on Data</strong></td>
<td><strong>Event Detection</strong></td>
</tr>
<tr>
<td><strong>Key Management</strong></td>
<td><strong>Trusted Platforms</strong></td>
<td><strong>Forensics</strong></td>
</tr>
<tr>
<td><strong>Communication Integrity</strong></td>
<td><strong>Crypto Enforced</strong></td>
<td><strong>Big Data for Security</strong></td>
</tr>
<tr>
<td><strong>Authenticated Computations on Data</strong></td>
<td></td>
<td><strong>Access Policies</strong></td>
</tr>
<tr>
<td><strong>Granular Audits</strong></td>
<td></td>
<td><strong>Systems</strong></td>
</tr>
<tr>
<td><strong>Control of Valuable Assets</strong></td>
<td></td>
<td><strong>Construction of cryptographic protocols proactively resistant to DoS</strong></td>
</tr>
<tr>
<td><strong>Security against DoS</strong></td>
<td></td>
<td><strong>Crypto Enforced</strong></td>
</tr>
<tr>
<td><strong>System Health</strong></td>
<td></td>
<td><strong>Big Data for Security</strong></td>
</tr>
<tr>
<td><strong>Big Data for Security</strong></td>
<td></td>
<td><strong>Retention, Disposition, Hold</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Digital Rights Management</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Construction of cryptographic protocols proactively resistant to DoS</strong></td>
</tr>
</tbody>
</table>
## 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>
</tbody>
</table>

- Thwart and Vulnerability Management
- Monitoring, Alerting
- Mitigation
- Configuration Management
- Logging
- Malware Surveillance and Remediation
- Network Boundary Control
- Resiliency, Redundancy and Recovery

### Data Governance

- Encryption and Key Management (including Multi-Key)
- Isolation/Containerization
- Storage Security
- Data Loss Prevention, Detection
- Web Services Gateway
- Data Transformation
- Data Lifecycle Management
- End Point Input Validation
- Digital Rights Management

### Visibility and Infrastructure Management

- Threat and Vulnerability Management
- Monitoring, Alerting
- Mitigation
- Configuration Management
- Logging
- Malware Surveillance and Remediation
- Network Boundary Control
- Resiliency, Redundancy and Recovery
Use Cases

- **Retail/Marketing**
  - Consumer Digital Media Usage
  - Nielsen Homescan: Family level Retail Transactions
  - Web Traffic Analysis
- **Healthcare**
  - Health Information Exchange
  - Genetic Privacy
  - Pharma Clinical Trial Data Sharing
- **Cyber-security**
  - Network Protection
- **Government**
  - Military
  - Education
- **Industrial**
  - Aviation: Sensor Data Storage and Analytics
  - Transportation: Cargo Shipping
Big Data Security Reference Architecture

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

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

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

Securing Data Storage and Transaction logs
- Key Management
- Security Best Practices for non-relational data stores
- Security against DoS attacks
- Data Provenance

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

<table>
<thead>
<tr>
<th>S&amp;P Consideration</th>
<th>Health Info Exchange</th>
<th>Military UAV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy preserving data analytics and dissemination</td>
<td>Searching on encrypted data. Determine if drug administered will generate an adverse reaction, without breaking the double blind.</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>HIPAA security and privacy will require detailed accounting of access to HER data.</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>
Interface of BD App Provider -> BD Framework Provider

**S&P Consideration**

Policy based encryption
- Row-level and Column-level Encryption
- Policy-based encryption, often dictated by legacy channel capacity/type.

Policy management for access control
- Role-based and claim-based
- Transformations tend to be made within DoD-contractor devised system schemes.

Computing on encrypted data
- Privacy preserving access to relevant events, anomalies and trends.
- Sometimes performed within vendor-supplied architectures, or by image-processing parallel architectures.

Audits
- Facilitate HIPAA readiness, and HHS audits
- CSO, IG audit.
## 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
### Emerging Cryptographic Technologies

<table>
<thead>
<tr>
<th>Utility of Encrypted Client Data</th>
<th>Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>No operation possible at Cloud</td>
<td>Standard Encryption</td>
</tr>
<tr>
<td>Controlled results visible at Cloud</td>
<td>Searchable Encryption</td>
</tr>
<tr>
<td></td>
<td>- Symmetric</td>
</tr>
<tr>
<td></td>
<td>- Asymmetric</td>
</tr>
<tr>
<td>Policy-based Access Control</td>
<td>Identity-based Encryption</td>
</tr>
<tr>
<td></td>
<td>Attribute-based Encryption</td>
</tr>
<tr>
<td>Transformations possible, but results not visible to Cloud</td>
<td>Homomorphic Encryption</td>
</tr>
</tbody>
</table>
Thank you!
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 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