Networking and Information Technology
Research and Development

***The government seeks individual input; attendees/participants may provide individual advice only*.**

Middleware and Grid Interagency Coordination (MAGIC) Meeting Minutes[1]
May 6, 2020, 12-2 pm ET

Virtual

Participants

Tom Barton (UChicago)	Neal Magee (UVA)
Richard Carlson (DOE/SC)	Jennifer Mahalingappa (DOE/HQ)
Dhruva Chakravorty (TAMU)	David Martin (ANL)
Vipin Chaudhary (NSF)	Deep Medhi (NSF)
Michael Corn (UCSD)	Sharad Mehrotra (UCI)
Martin Doczkat (FCC)	Linden Mercer (NRL)
Rick Downs (UVA)	Michael Roy Nelson (Carnegie)
Sharon Broude Geva (UMich)	Donald Petravic (NCSA)
Margaret Johnson (NCSA)	Steve Petruzza (Utah)
Padma Krishnaswamy (FCC)	Birali Runesha (UChicago)
Eric Lancon (BNL)	Arjun Shankar (ORNL)
Joyce Lee (NCO)	Suhas Somnath (ORNL)
Zhengchun Liu (ANL)	Andrew Theissen (NTIA)
Miron Livny (UW-Madison)	Sean Wilkinson (ORNL)
Jay Lofstead (Sandia)	

Proceedings

This meeting was chaired by Richard Carlson (DOE/SC) and Vipin Chaudhary (NSF).

Guest Speakers:

Ronald R. Hutchins, Vice President for Internet Technology, University of Virginia, *Protecting Data in the Time of Open Science*

What's the problem?

World is changing (foreign influence, hacking, etc.)

- Need to protect data on research side – even if it is publicly, open data (e.g. weather data);
- Business confidentiality (reputation issues), personal privacy; controlled, unclassified data; export technologies

[1] Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program.

- Data confidentiality – not one standard for each set of tools; controls/standards need to meet are different. E.g., HIPAA similar but not same as controlled unclassified (CUI requires separation by project). “Walled garden” concept facilitates research

Single platform for both HIPAA and CUI works

- But researchers with only HIPAA constraints need to do more to get onboarded - need to address.

Goal: open up data for research while sufficiently protecting it.

- Use REDCap and TriNetX (allows running of clinical trials across multiple places)

<u>How are we addressing it?</u>

<u>Single platform</u>- Efforts to build single platform for HIPAA/CUI compliance for use by Virginia universities ; note that single attorney general across state universities and shared cyber insurance facilitate this model

<u>Containerization standpoint</u> – enable portability in code and building environments with goal of certifying

- Efficiency POV as well
- Custom solutions
- Security by design (securing up front):
- Security piece
    - Code-driven and code-based. Researchers draw from pre-invented and pre-approved catalog of containers.
    - Only allow customization for packages (cpan, library support for languages)
    - Tight ingress controls/ no outside internet access
    - HIPAA, CUI environment: VM-based; create custom VM
        - Research computing environment is purpose-driven, focused and efficient
            - Containers – more efficiency out of hardware and learning research patterns; perhaps build automated pipelines - containers do work

3rd party audit- implementing process - connecting security folks with technical folks to solve problems

<u>Discussion</u>

- Heart of problem- working on (same no matter which health care provider?) Academic medical centers vs. Hospitals
- iThrive (integrated Translational health Research Institute of Virginia) – INOVA Health System, UVA, Virginia Tech, Carilion Clinic collaborative project. No one wants to share PII across universities
- However, note COVID-19: want to share data
- For compliance reasons, medical data must be held in tight control of medical center– note hybrid institutional constraints (university and hospital)
- Hospital models (Walled garden notion) for securing data may be antiquated
- HIPAA: Locally interpreted – research trumps compliance in some way (manage risk according to HIPAA does not require “walled garden” approach)
- Data sharing – releasing PHI to other places with BAA. Other health care institutions using red cap

What's next

Push out a few things using federation to bringing other schools without creating UVA accounts (use InCommon and COmanage tools)

IThrive

- Partnership – clinical trials using REDCap, etc.
- Also building iThrive Data Commons modeled after NIH's Data Commons model (set of data with MD on top and catalogue). Distributed data and common catalog through a portal. Machine sits underneath and data object identifiers inside portal will point to data.

Data management is final frontier:

- Administrative side: Inventory of data needed for compliance
- Research side: working on building management for data

CC* NSF award for science DMZ and data transfer node

- Critical for UVA getting its current architecture. DMZ white lists entrants.
- DTN is central , remote mounted, and read-only to IVY (protected machine).
- Local storage available. Globus - transfer practice.

Simplifying Operations- Efforts to balance security and access

- Limit initial access control process to the appropriate elements
- Automate registration processes- Down from 2 weeks to 5 minute
- Limit steps in science workflow necessary to begin research based on constraints on the data
- Limit manual data copies internal to the system
- Limit administrative functions not performed by the researcher (scale)
- Make the process easier (and more secure) than could be accomplished by the individual researcher without our help

Discussion

- Data ownership issue: comes into play. Need to ensure providence, ownership and ability to credit creators of data (e.g., index akin to H-index) Creating repository across different projects issue.
- Data access policy – controlled by owners or university?
    - CUI data controlled by sponsor, in conjunction with PI
    - HIPAA data access: IRB process. Need to be on IRB and obtain access through PI.

Hakizumwami Birali Runesha, Assistant Vice President for Research Computing; Director of the Research Computing Center, The University of Chicago, *Hosting and Working with Sensitive data on High Performance Computing environments*

- Focus: HPC computing side
- RCC - >600 faculty and >4k end users; Need to support sensitive data at RCC
- Researchers are increasingly working with data requiring compliance with a variety of data security and privacy standards.

Observations: lack standardized security, privacy and operating procedures.

- Interpretations based on university and different internal units. Environment: deluge of data, and AI tools; increased data security breaches

Overwhelming variety of regulations and standards (ITAR, HIPAA, CUI, GDPR, FISMA, DFAR)

- no guidance or way to approach. Own control and way of supporting – difficult for end user when conducting research

Data Privacy Considerations

Ensure policies are in place and align them with evolving laws, regulations and standards

Data privacy – ethical as well as legal/regulatory POVs

Protecting data- depends on regulatory controls, DUA negotiations, etc.

- External data providers (require Data Use Agreement and perhaps IRB):
- Data generated on campus may require IRB protocol
- Procured data

Many faculty navigated these contracts on their own.

- While tend to focus on last stage (Storage and computing), many critical services and processes come into the picture.

Secure Research Data Strategy (SRDS): Assist University managing and storing sensitive research data

- University-wide approach and collaboration, meet bi-weekly to provide secure computing environments for UChicago researchers to access, store and analyze sensitive research data
- Partnership among RCC, IT services, Legal Counsel and University Research Administration
- Strategy and Approach
    - Develop University Edition CyberSecurity and Data Privacy policies – template
    - Introduced Secure Data Enclave (SDE) – secure, centralized service to provide
        - access and storage of sensitive data; has HPC capability;
        - consultation services for data compliance and handling sensitive data and
        - computing resources to analyze, manage and report on such data
    - Created workflow for risk assessment
    - Create governance (to drive activities)
    - SOPs
    - SRDS approved environment
- SDE Resources
    - MidwayR - HPC system within SDE that provides access to secure platform environment for workloads requiring high end computing or high performance storage (Slide)
    - Commodity-Virtual machines – as not all of needs can be supported by HPC cluster

    - Secure Rooms (off network) - not everything can be accessed on the network

RCC HPC Ecosystem

- Trying to mirror Usual HPC cluster open to public for sensitive environment
- As support different researchers (HIPAA, non-HIPAA), need to separate sensitive data by level of needed protection (Low, moderate high levels) to avoid hindering research productivity
- Secure room off network

Controls (e.g., Access control through firewall; media control)

Governance- guiding access to sensitive data includes many layers (Slide, diagram)

Created Privacy & Security Council for University Research (govern execution of strategy at university level) - also SRC Oversight Committee; Board of Computing Activities and Services and Research Computing Oversight Committee; Also working groups

Looking ahead: Need the following

- Management systems and collaborative tools – to drive information flow
- Develop vocabulary for data obligations
- Common interpretation of security standards
- SOP to streamline procedures
- Reference architecture and replicable processes
- Training end user

Sharad Mehrotra, Professor of Information and Computer Science, University of California, Irvine, ***Secure Data Outsourcing over the years***

Shift to cloud computing over the years – no infrastructure. Build up costs, no system management headaches, cost amortization

Key Challenge: Loss of Control
Cloud is common resource – lose control of data security personnel, policies or enforcement, to those managing data in cloud. Gov. Subpoenas

Adversarial Cloud model - most common model

- “Honest-but-curious" adversary – wish to learn about data vs. Malicious (sabotage)
- Passive (making inferences based on passive observations) vs. Active Adversary (launching attack on data)

Solution: Can encrypt data before uploading to cloud – but how to process data processing application?

Approaches: Download and process locally, but limited benefit from cloud (using for storage)

(1) Cryptographic solutions – can process certain amount on encrypted domain itself; appropriate encryption techniques

- Data Processing over encrypted data
    - Fully homomorphic encryption – inefficient, time-consuming

  - Partially homomorphic – add or multiply, not both. Limits the type of possible applications
  - Searchable encryption – allow comparisons; enabled property-preserving encryptions – preserves certain characteristics, but many techniques developed for such encryptions are “broken” - thus, no protection
- Multi Party Computation and Secret shared data– break up secrete share data and send to different users to compute the portion received. Secure against stronger adversaries, but very costly
- Cryptographic Solution Landscape - many solutions with tradeoffs regarding level of security provided, efficiency, supportable queries (e.g., selection), amount of work client must do, some techniques not work with updates

(2) Exploiting Trusted Computing hardware

- Encrypt data and store in cloud side, and conduct data processing, which may require encryption, enclave, for example, could be in cloud.
- When processing requires data to be decrypted, push to enclave and decryption occurs in secure hardware. server and enclave execute data processing needs and send back results.
- Secure trusted hardware at cloud acts as trusted agent of data provider
- Problem
  - Performance:
    - Entry and exit very costly
    - Memory is encrypted and physical memory is limited, which significantly slows downs (e.g., simple key value storage)
  - Security
    - Enclave only assumes that very small part of application is in secure execution environment, but many (side-channel) attacks possible
  - Bottomline: no silver bullets but ongoing research to protect against vulnerabilities.
    - Some problems will be fixed in hardware (cache, branch prediction)
    - Some problems too hard to fix completely (unacceptable overheads – e.g., enclave memory access patterns)\

Other issues

Computation Cost of security - compare techniques with execution time (see graph)

Security threats – Want security model (Full Download)

Despite all this, industry and academia building encryption-based and secret-sharing-based systems. Testimony of importance of problem.

Key question: Can we design an outsourcing solution that is simultaneously efficient and secure?

Way forward – sizing the solution to the problem

Traditional binary view point regarding data protection (all or no data needs protection)

In contrast, real-world may be more gray: As only some organizational data is sensitive, could outsource nonsensitive data in plain text manner.

Work on understanding the security and performance implications of using multiple cryptographic techniques simultaneously

- E.g. If store data across multiple cryptographic domains - data partly stored using MPC, property-preserving encryptions, and plain text form and do data processing in this partitioned world, what are the implications
- Problem is just starting to get attention

Jennifer Mahalingappa, Patent Attorney, DOE Office of the Assistant General Counsel for Technology Transfer and Intellectual Property, *Data Confidentiality at the Department of Energy*

Acquisition and use of data sets- data sensitivity and confidentiality implications

<u>DOE Data Confidentiality Discussion</u>

Data protection spectrum (different types of data and associated rights, limitations on use and protection)

- Classified or unclassified information (OUO – still sensitive)
- Federally Funded information (Unlimited Rights Data) vs. Privately Funded Information (Limited Rights Data)
- Privacy – PII
- PHI (HIPAA, Privacy Act – special protections based on how/why data is collected)
    - e.g., ORNL and other DOE national labs: HIPAA and Privacy Act-compliant data enclaves; attorneys ensure that enclaves are sufficiently secure for data owner (conforms with relevant NIST standards). Also requires data use agreement and Authority to Operate (ATO) specifies certifications have for enclave and access-controls - must conform to data originating sources standards for access and security. Audit and inspection provision usually included in ATO.
- IRB- related approvals: business associate agreement (if acting on behalf of someone for HIPAA, they have right to transfer data to you as a business associate); Alliance agreement – so IRB can provide approvals for use of enclaves; set forth protocols for data use.
- FOIA request for government record – Mark data as government agency in order to clarify how information should be treated. If government record, fit under FOIA exemptions?
    - Trade secrets – limited rights data - must be privately funded information falling under definition of trade secret pursuant to Trade Secret Act. N/a to research performed with federal funding.
    - Confidential business/financial information - must be marked as such to be withheld from disclosure; otherwise default to government record (unlimited rights). Must mark data for protections.
    - *Argus leader* (Supreme Court case)- if customarily treated as confidential information, not matter if marked if government makes promise (even implied) that information is confidential, can be withheld under FOIA

Discussion

If specialized computing infrastructures (e.g.,HPC), experience with nontechnical auditors?

- Internal auditing team learned about technical issues
- Mapping compliance rules into technical solution
    - No direct interpretation from compliance rules into technical solution. So unclear and auditors have their own interpretation.
    - Challenge – NIST 800.871 audit – audit uses ISO standards and tries to map back to NIST. Often different interpretations of what auditor required.

Research computing and enclaves – CUI and CMMC, Secure in place

- Wetlands and other environments; shared environments that can't be moved into HPC- success stories/advice?
- HIPAA audit of research computing environment- using NIST RMF and 866. Found adequate.

Auditing

- emerging need because underlying organization not trusted? Ensure adaptations can be automated to certain degree.
- Automated audit for HIPAA is impossible.
- Many lack understanding of some of these systems

Research data – built infrastructure to capture sensor data (including location of people). No privacy legislature or policies related to data at the time at UCI. In last 4 years, built up policies. Where are we headed as smart campuses built up?

- UVA – much data is sensitive (while publish energy use, etc.) - HVAC specifics etc.
- UChicago – Array of Things project collects data; may be a resource

Resources – beyond individual resources

- Outside 3rd party for CUI certification (see DoD list)
- UVA -Internal folks met monthly (IRB, technical, etc.) to document issues
- UChicago – standing meetings as well to brainstorm and document issues.
- Sharing best practices moving forward could be helpful

Meetings:

July 26 – 30, 2020, PEARC20 Meeting, Portland, OR (February 17, 2020 deadline for submissions)

Next Meeting: June 3, 2020 (12 noon ET); continuation of confidential data discussion