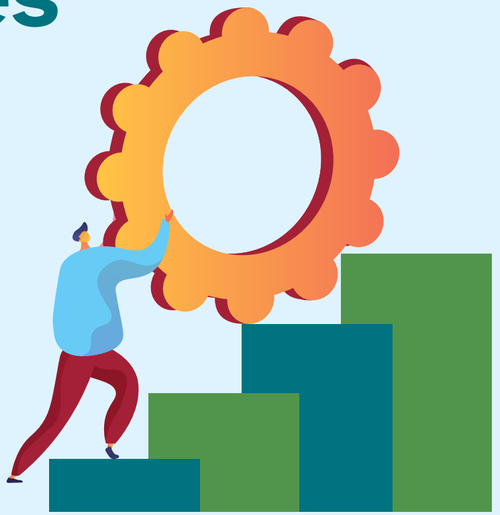
Challenges for Online Research Data Enclaves

John E Marcotte, PhD, University of Michigan (jemarcot@umich.edu)



Online or “Virtual” Data-Enclaves for analyzing restricted research data are growing in popularity because the sensitive data stay on the server and access can be suspended at any time. These enclaves have advantages over non-networked computers, the long time standard for research data because they enable researchers to work remotely while maintaining security. Moreover, the enclaves act as collaboratories where projects can share space for programs, documents, and data. Team Science is an important component of NIH funded research.

- An enclave must have encrypted connections, two-factor authentication and prevent files from being copied out of the enclave.
- Although enclaves have clear advantages for providing access to sensitive data, enclaves still face important challenges.

Identity Management & Accounts

Some research data have restrictions on simultaneous access to prevent linking. Enclaves must provide access to data from source 1 or data from source 2 but not data from sources 1 and 2 simultaneously (XOR). Different login credentials are contrary to best practices for identity management at most universities and organizations.

High Performance Computing

Enclaves must incorporate High Performance Computing into the security enclave. Not all projects require HPC, so the challenge is to provide HPC as needed and at a scalable price. Any HPC must maintain security requirements including preventing end-users from copying files from the system.

Queuing Availability

Enclaves have fewer seats than users. Researchers typically receive a seat on a first come basis. Enclaves need waiting queues that inform users when a seat has opened. Cable and wireless providers often provide these type of queues for customer service.

Batch Processing

Computation can extend beyond an interactive session. Batch processing would allow the scheduling of longer jobs and potentially make seats available. Batch processing could provide access to more resources.

UNIVERSITY OF MICHIGAN

EDUCAUSE ANNUAL CONFERENCE 2021

Challenges for Online Research Data Enclaves

John E Marcotte (jemarcot@umich.edu)

Online or "Virtual" Data-Enclaves for analyzing restricted research data are growing in popularity because the sensitive data stay on the server and access can be suspended at any time. These enclaves have advantages over non-networked computers, the longtime standard for research data because they enable researchers to work remotely while maintaining security. Moreover, the enclaves act as collaboratories where projects can share space for programs, documents, and data. Team Science is an important component of National Institutes of Health (NIH) funded research. The data in these enclaves are typically labelled as sensitive, confidential, and restricted; moreover, these data have inferential disclosure risk which require security controls. Remote enclaves have been particularly important during the COVID pandemic of 2020-21. Despite these advantages, enclaves still face important challenges. We identify four key challenges:

(1) Identity Management and Accounts

Some research data have restrictions on simultaneous access. In some circumstances, researchers must not be able to access data from multiple sources at the same time. The restriction prevents data from being merged or combined. The challenge is that most systems operate under authentication and authorization. A system must provide access to data from source 1 or data from source 2 but not data from sources 1 and 2 simultaneously (XOR). In these circumstances, the paradigm of Authentication, Authorization and Audit is not adequate for some research circumstances. Different login credentials are contrary to best practices for identity management at most universities and organizations. Setting up different secure systems for each data source is typically not practical or cost effective. Virtualizing every security permutation will make maintenance difficult. One possible solution is a secondary sign in for each project, but that may not provide sufficient safeguards. Now, some enclaves provide researchers with different login credentials for each project.

(2) High Performance Computing

Enclaves must incorporate High Performance Computing into the security enclave. Not all projects require HPC, so the challenge is to provide HPC as needed and at a scalable price. Any HPC must maintain security requirements including preventing end-users from copying files from the system. Leveraging services from cloud providers such as AWS (Amazon Web Services) is one option; however, researchers must already have secured funding. One of the goals of many research centers at universities is to support junior faculty, new researchers and students who have not yet obtained funding. A research track record is needed to earn a grant. Research centers want to develop capacities to support affiliated researchers at a variety of levels. The special needs to access and analyze sensitive research data requires some extra accommodations by computing services.

(3) Queuing Availability

Enclaves have fewer seats than users. Researchers typically receive a seat on a first come basis. Enclaves need waiting queues that inform users when a seat has opened. Cable and wireless providers often provide these types of queues for customer service. Any limited resource should provide a way for users to queue up for the service. Computer labs often have a signup process. Online enclaves are like these labs.

(4) Batch Processing

Batch Processing is a throwback to a bygone era. Nevertheless, batch processing does allow the scheduling of jobs with respect to available resources. Computation can extend beyond an interactive session. Batch processing would allow the scheduling of longer jobs and potentially make seats available. Batch processing could provide access to more resources but only during limited times.

In addition to these four challenges, software updates and special software requests can also be difficult. Enclaves must decide how often to update and develop a process for determining if requested software will be added.