Summary Report – SCE Environment Scan

August 15, 2024

Report Author and Point of Contact:
Kurt Rohloff, PhD.
CTO and Co-founder, Duality Technologies
krohloff@dualitytech.com

America's DataHub Consortium (ADC), a public-private partnership, implements research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). These results document research funded through ADC and are being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed above do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. This product has been reviewed for unauthorized disclosure of confidential information under NCSES-DRN24-080.

Executive Summary

The objective of this project was to better understand the requirements and needs of a Secure Computing Environment (SCE) within a future National Secure Data Service (NSDS). SCEs are environments that enable researchers, analysts, economists, statisticians and other data scientists to access, link, and analyze data in a protected manner and generate analytic results without unduly exposing sensitive information.

SCEs are essential infrastructure for a future NSDS because SCEs are designed to support data sharing and data utilization, all while respecting and maintaining privacy and confidentiality restrictions. An SCE can support efficient data collaboration on sensitive data by enabling the linking and analysis of these data in an environment that has established policies and procedures to protect data and offer services to researchers who may not have existing mechanisms to analyze these types of sensitive data. This then leads to better-informed policy decisions by using data for evidence building that may not have been previously available to researchers.

To understand the requirements and needs of SCEs within a future NSDS, we performed a scan of SCE usage in the federal data community. Our specific goals were to gather diverse perspectives within the federal environment and to identify considerations from experienced and likely SCE users as a component of a future NSDS. Our overarching goal was to identify, through interviews with members of the federal data community, considerations relevant to six areas of interest, including computing, security, technical, legal, policy, and related requirements of current and likely SCE users.

To accomplish our goals, we approached 75 members of the federal data community, inclusive of data producers, data consumers, data analysts, IT staff, and other staff who support adjacent roles. Targeted and recommended individuals included those from the 13 federal principal statistical agencies who we identified as being currently or likely users of SCEs and select members of the Federal Chief Data Officer (CDO) Council and the Federal Chief Information Office (CIO) Council.

We generated an interview protocol to understand staff members' experiences with SCEs, and the potential benefits and perceived challenges with a future SCE in an NSDS.   Of those that were approached for interviews, 21 members of the federal data community agreed to participate.

A summary of the interview observations is below:

- With some notable exceptions, the majority of current or potential SCE users' data and analysis needs could be met by commercial cloud computing environments.
    - Most users run computing workloads with relatively low resource requirements and using legacy tools.
    - The core exceptions are driven by very large scale and very low-latency applications.
- Data protection requirements oblige the use of SCEs, and SCEs could scale to support heavy data analysis workloads based on observations shared by interviewees.
- The principal statistical agencies have data protection requirements driven by Confidential Information Protection and Statistical Efficiency Act (CIPSEA) regulations.
    - CIPSEA establishes uniform confidentiality protections for information collected for statistical purposes by U.S. statistical agencies.

  - Existing technical solutions for SCE needs are likely to address these CIPSEA-driven security requirements.
  - A small subset of users have much more restrictive requirements driven either by more extensive legal and regulatory obligations or national security concerns. Smaller federal statistical agencies with smaller budgets and fewer staff members are currently hindered in their day-to-day work by administrative overhead, including the need for extensive Memorandum of Understanding (MOUs) and other similar agreements to access and collaborate on data. There was excitement that the NSDS would hopefully reduce the costly administrative burdens of adopting and using SCEs.
  - Interviewees seemed to be general cautious around acceptance of the use of SCEs in the NSDS that would encourage the use of a future SCEs more broadly, not just in the NSDS.
- Most existing SCE users are proficient in the use of computing tools but may not have the ability to bring additional technical staff to support the broader adoptions of SCE needs.
  - There was high awareness and excitement for broader adoptions of SCEs across the federal statistical community, but interviewees outside of the federal statistical community had less awareness but were intrigued by the concept of using SCEs. The primary value noted by interviewees in SCEs was to support the ability to join data from multiple sources with easier compliance around policies and regulations that required data privacy and data security measures by better protecting data that are used in analysis, so that results can be obtained without exposing the source data.
  - Several of the larger federal statistical data agencies (ex: Census, BEA, etc.) had recent relevant experience with secure data computing.

Based on the observations from interviewees, we observed the following high-level considerations for the future adoption of an SCE for the NSDS:

- Smaller federal statistical agencies are less aware of how to apply and adopt SCEs and are likely to receive more benefit from a shared SCE. It could be beneficial to increase engagement and outreach to smaller agencies to ensure a future SCE in an NSDS could support their needs.
- There was general and consistent feedback around the importance of policy and regulatory acceptance for security and legal concerns around the use of SCEs, but uncertainty exists about eventual acceptance of an SCE for the NSDS by administrative and policy authorities. Additional policy and regulatory clarity around the SCE could support engagement and adoption.
- Potential SCE users seem to have consistent interest in relatively simple capabilities that are currently available commercially, either from commercial software vendors or commercial cloud environments, specifically including the ability to join data privately and then run simple operations and generating descriptive statistics and building regression models. These simple capabilities could be a starting point to services offered in the SCE.

- Potential users were hesitant to change the way they currently work to use advanced SCEs, and many were most receptive to the advanced SCEs if there was support for tools they currently use or are moving to, specifically Python. The ability to support operations transparently and with privacy protections was critical.