Protecting Sensitive Data with Secure Data Enclaves

Mark Howison*

Research Improving People's Lives, Providence, RI, USA

Mintaka Angell

Research Improving People's Lives, Providence, RI, USA

Justine S. Hastings

Research Improving People's Lives, Providence, RI, USA

Department of Economics, University of Washington, Seattle, WA, USA

National Bureau of Economic Research, Cambridge, MA, USA

A Secure Data Enclave is a system that allows data owners, such as governments and private firms, to control data access and ensure data security while facilitating approved uses of data by other parties. This model of data use offers additional protections and technical controls for the data owner compared to the more commonly used approach of transferring data from the owner to another party through a data sharing agreement. Under the data use model, the data owner retains full transparency and auditing over the other party's access, which can be difficult to achieve in practice with even the best legal instrument for data sharing. We describe the key technical requirements for a Secure Data Enclave, provide a reference architecture for its implementation on the Amazon Web Services platform using managed cloud services, and describe four use cases of this architecture in partnerships with state governments to control access to sensitive administrative data.

CCS CONCEPTS • **Security and privacy** → **Information accountability and usage control** • **Information systems** → **Cloud based storage**

Additional Keywords and Phrases: Data governance, data management, data sharing, security controls, managed services, remote workspaces, scientific collaboration, public policy

ACM Reference Format:

Mark Howison, Mintaka Angell, and Justine S. Hastings. 2021. Protecting Sensitive Data with Secure Data Enclaves. *OSF Preprints*. DOI:<https://doi.org/10.31219/osf.io/jmd7t>

1 INTRODUCTION

Data is increasingly a driver of innovation and improved product and service delivery. Advances in data-intensive methods like machine learning and artificial intelligence have fueled innovation in the private sector, improving product

* Contact: connect@ripl.org

Version 3 (2023-02-18)

Copyright © 2021-2023 Innovative Policy Lab d/b/a Research Improving People's Lives ("RIPL"), Providence, RI. All Rights Reserved.

quality, customer experience, and lowering production costs [1]. The public sector has similarly begun exploring the use of big data, machine learning, and artificial intelligence to solve important policy challenges [2-6].

Many organizations, both private and public, require external expertise to fully leverage the value of their data or to apply cutting-edge methods like machine learning and artificial intelligence. Organizations may, for example, contract with consultants who specialize in data analysis or engage academic researchers who can bring a rigorous and scientific framework to solving problems with data. Sharing sensitive data with other parties, however, creates many risks for organizations, especially when it involves transferring data to a system outside the organization. Moreover, collaborations with academic researchers often involve an academic responsibility to openly publish findings through scientific peer review, which can conflict with the restrictions surrounding sensitive data [7].

Traditionally, data sharing is accomplished with legal agreements that govern acceptable uses of the data, transfer liability, and provide remedies for data breaches or other unacceptable data use [8]. However, once the data leave the custody of an organization it can be difficult in practice to enforce these terms or audit the actual use of the data by other parties.

In contrast, data use is a model in which an organization retains custody of its data at all times while allowing approved users to access the data. The benefit of this approach is that the organization has full transparency into the use of the data and can terminate access at any time. However, this approach requires that the organization maintain a secure system with appropriate technical controls to securely host their data and grant access to other parties.

Many previous systems have been created with these goals in mind and we collectively refer to them as Secure Data Enclaves, following terminology already in use.¹ A related concept is a Research Data Center, which is a secure research facility that requires physical access for analyzing sensitive federal data, run in conjunction with the US Census Bureau or National Center for Health Statistics [9-10]. Because of the practical challenges and constraints around physical access, remote-access variants of these secure systems have been developed, many of which are hosted on-site at research universities [11].

Today, organizations of any size can implement a Secure Data Enclave in a cost-effective way using best practices for security and compliance through managed cloud services. In this paper, we describe a cloud-based approach to implementing a Secure Data Enclave, along with a standardized set of technical and governance controls and a reference architecture for the Amazon Web Services cloud platform. We describe four existing use cases of this solution, including our organization’s recent use of a cloud-based Secure Data Enclave for processing Pandemic Unemployment Assistance claims as part of the State of Rhode Island’s emergency response to the COVID-19 pandemic [12]. That system followed the reference architecture described here and allowed the state to safely share sensitive claims data with contractors for claims processing and fraud detection.

2 TECHNICAL DESIGN

A Secure Data Enclave implements both technical controls (denoted with “T”) that enforce constraints on the use of the system by data users, as well as governance controls (denoted with “G”) that provide best practices and policies for how data owners manage use of the system. Most importantly, the Secure Data Enclave is owned and managed by the data owner. Data never leave the data owner’s custody. The data owner retains control over who can access the data, how the data are used, and what analysis products can be exported from the system. Depending on the specific requirements of the data owner, many or all the controls listed in Table 1 can be implemented in a Secure Data Enclave.

We describe one possible implementation of the technical controls in Amazon Web Services (AWS) using managed cloud services, summarized visually in Figure 1. However, a Secure Data Enclave can be implemented with many cloud provider’s services or even in an on-premise system (although this would negate many of the advantages for security, scalability, resiliency, and cost optimization offered by cloud services).

¹ The term “Secure Data Enclave” is used to describe managed systems at the University of Chicago (<https://securedata.uchicago.edu/>), the University of California – Los Angeles (<https://ccpr.ucla.edu/services/computing/support/secure-data/>), and Columbia University (<https://cuit.columbia.edu/sde>).

Table 1: Technical and governance controls for a Secure Data Enclave

SDE T1	No outbound access to the Internet.
SDE T2	Inbound remote access through a controlled client only.
SDE T3	No copy/paste out from the remote client.
SDE T4	Enforce a password lock screen after 5 minutes of inactivity in the remote client.
SDE T5	Software installation is controlled centrally.
SDE T6	Software is kept up-to-date with security patches applied on a regular schedule.
SDE T7	New software is scanned for viruses and malware at installation time.
SDE T8	All data is encrypted in transit.
SDE T9	All data is encrypted at rest.
SDE T10	Data cannot be downloaded from the system by data users.
SDE T11	Comprehensive audit logging of all access to the system.
SDE G1	Access approval by user/project with security pledge (see Appendix A.1 for an example).
SDE G2	Review and documentation of all data exports by the data owner.
SDE G3	Periodic review by the data owner of all approved projects and users (e.g. quarterly).

2.1 Virtual Desktops

The key managed service for our reference architecture is Amazon WorkSpaces, which provides managed Windows or Linux virtual desktops. AWS provides a remote desktop access client and manages the security, patching, and maintenance of gateway and access client. The managed workspaces run within a virtual private cloud (VPC) that is cordoned off from the Internet and all other systems.

The VPC is configured without an Internet Gateway, meaning that no traffic routes to the Internet, implementing SDE T1. In some cases, limited Internet access may be required to pre-specified locations (for example, to import data into the environment). That use case can be implemented by adding Internet and NAT Gateways to the VPC and modifying the built-in WorkSpaces security group (firewall) to allow egress to only the approved destinations, or through a more feature-rich stateful firewall implemented with the AWS Network Firewall service. There are Amazon VPC Endpoints that are created within the VPC, providing private and secure access to only the required AWS services, for example Amazon S3.

Amazon WorkSpaces requires integration with Microsoft Active Directory (AD) infrastructure, which can be managed with AWS Directory Service. This directory provides the identity management for approved users of the system, implementing SDE G1. The following options are currently recommended by AWS [13]:

1. Integration with an existing Azure AD or on-site directory through the AWS AD Connector service.
2. Creating a stand-alone directory for the Secure Data Enclave using the AWS Managed Microsoft AD service.
3. Installing a self-managed stand-alone directory on cloud-hosted virtual machines.

In addition to being configured without Internet routing, the workspaces are only accessible using approved Amazon WorkSpaces clients for Windows, Linux, Mac, Android, iOS, web browsers, or Chrome OS, which can be enabled or disabled for any combination of platforms, implementing SDE T2. Additionally, group policies can be configured in the Active Directory to control the directionality of copy and paste in the WorkSpaces client and require a password-protected lock screen, implementing SDE T3 and SDE T4.

Figure 1: Architecture diagram of a Secure Data Enclave in Amazon Web Services.

The software stack available in the workspaces is managed and updated via a secured build workspace that is independent of the Secure Data Enclave and located in a separate AWS account with Internet access, implementing SDE T5. This build workspace can be updated on a regular schedule to incorporate security patches, implementing SDE T6, and can also be updated on-demand to install new analysis or statistical software and packages as required by the data users. To create a new software release, the build workspace is scanned for viruses and malware, implementing SDE T7, and imaged, then the image is shared across AWS accounts for use as the base image with the Secure Data Enclave workspaces. Amazon WorkSpaces provides a mechanism for rebuilding a data user's workspace in-place against an updated image while preserving the user's locally stored data [14].

2.2 Data Transfer, Storage, and Access

Sensitive data can be ingested into the Secure Data Enclave using either the native S3 transfer protocol or the Secure File Transfer Protocol (supported through the AWS Transfer for SFTP managed service [15]), both of which provide in-transit encryption to implement SDE T8. All S3 buckets and the storage volumes underlying the workspaces are encrypted with AWS Key Management Service (KMS) customer-managed keys, implementing SDE T9.

The S3 buckets are configured so that the *sensitive data bucket* is writable by the data owner from outside the system using S3 or SFTP as described above. It is read-only from the workspaces of the data users. The *import bucket* is writeable by data users from outside the system and read-only from the workspaces. It allows public data sets, analysis code, and other non-sensitive resources to be imported into the Secure Data Enclave. Finally, the *export bucket* is writable by data users from the workspaces but can only be read outside the system by the data owner. It allows the data owner to export approved analysis products from the Secure Data Enclave.

We use a combination of S3 bucket policies (e.g. the resource policies on each of the S3 buckets), the resource policy on the VPC Endpoint for S3, and the IAM roles assigned to the data users to implement these read/write permissions. This configuration prevents sensitive data from being downloaded by data users outside of the VPC, implementing SDE T10, while allowing data users to import non-sensitive public data and analysis code in a self-service way without requiring excessive administration and intervention from the data owner. Furthermore, the export bucket allows the data owner to implement SDE G2 and review all data exports from the system.

Data users can read directly from the sensitive data bucket from inside the workspace and store analysis results to local workspace storage. However, this requires using the command line interface or a third-party client to manually interact with S3 and sync files to the local volume. Also, local storage has limited backup and snapshot settings and cannot be shared among data users within the system. There are several additional storage options, compared in Table 2, which can be deployed to provide shared network-attached storage to the workspaces internally within the VPC.

2.3 Account Security

The Secure Data Enclave should be deployed within an AWS account that is configured according to security best practices, such as those recommended by the Center for Internet Security's AWS Benchmarks [16]. Compliance with these benchmarks can be continuously monitored within the account using AWS Security Hub. At a minimum, the account should be configured to audit all account activity and S3 bucket access with AWS CloudTrail [17], implementing SDE T11.

Table 2. Comparison of storage options for the Secure Data Enclave.

Managed Service	Pros	Cons
AWS Storage Gateway Acts as an abstraction layer between WorkSpaces and an S3 bucket, removing the need to interact with S3 through the command line.	Mountable as a network file system in WorkSpaces using AD credentials and standard protocols (SMB or NFS). Provides SSD-backed and memory-backed caching for better throughput on repeated data retrieval.	Requires persistent server resources (through EC2 instances in the VPC) that incur additional costs. Files written directly to the S3 bucket need to be discovered by Storage Gateway, although it can be scheduled to refresh regularly.
Amazon FSx for Windows File Server A serverless network-attached file system that runs within the VPC.	Like Storage Gateway, provides a mountable SMB network file system in WorkSpaces that fully supports Windows access control lists. A fully managed service with no servers to manage.	Requires pre-provisioning the available storage allocation and throughput.
Amazon Athena A serverless query service that treats big data as a SQL-like database. Athena offers flexible computation of complex queries without requiring significant resources within the VPC.	Offloads the computational resources needed to run SQL-like queries against structured data in an S3 bucket. Works well with a variety of big data formats including CSV data files and Apache Parquet columnar data files.	Access controls are not as fine-grained as IAM policies for S3 buckets or Windows access control lists. Most analysis software (such as Power BI, Python, or R) will require specialized packages and an ODBC driver to directly access Athena.

3 USE CASES

Since 2020, the Secure Data Enclave architecture described here has been deployed to meet a variety of policy needs in the public sector, including the following four use cases:

1. **Rapidly delivering unemployment assistance.** The COVID-19 public health emergency caused widespread economic shutdown and unemployment. The resulting surge in Unemployment Insurance (UI) claims threatened to overwhelm the legacy systems state workforce agencies rely on to collect, process, and pay claims. Research Improving People's Lives (RIPL) partnered with the Rhode Island Department of Labor and Training and Amazon Web Services (AWS) to deploy the Secure Data Enclave as part of a scalable cloud solution to collect and process Pandemic Unemployment Assistance (PUA) claims. RIPL's system was developed, tested, and deployed within just 10 days, making Rhode Island the first state in the nation to collect, validate, and pay PUA claims [12]. The Rhode Island Department of Labor and Training (RIDLT) is now launching new tech-based solutions using their Secure Data Enclave to help displaced workers reskill for economic recovery through career recommendations powered by data and artificial intelligence [18].
2. **Securely exchanging data between state agencies to reduce benefit processing time.** To ensure that PUA and UI benefits were paid quickly during the pandemic, Rhode Island needed to verify claims against claimants' individual tax data from federal or state tax authorities on a real-time basis, as opposed to the traditional verification against UI tax records collected from employers. RIPL partnered with the Rhode Island Department of Taxation to deploy a Secure Data Enclave architecture for verifying 2018/2019 state income tax records against the adjusted gross income reported in PUA claims. Using separate enclaves connected through a private VPC link, the Department of Taxation and RIDLT exchanged data securely without disclosing Personally Identifiable Information to staff at either department. This automated verification allowed Rhode Island to "pre-validate" PUA claims and avoid downstream costs and delays associated with

manual verification, contributing to Rhode Island being the first state in the country to pay PUA claims [12]. This use case also illustrates “evaluation as a service,” a model enabled by the Secure Data Enclave architecture where other parties can partner with government to use administrative data to evaluate the impact of policies and programs while maintaining data confidentiality between all parties.

3. **Helping students connect, learn, and earn rewards for academic improvement during the summer.** During the COVID-19 public health emergency, the Rhode Island Department of Education (RIDE) needed a way to encourage students to stay on the path to academic success and reduce remediation following school closures and lost in-person classroom time. RIPL partnered with RIDE to deploy the free Summer Academy for Interactive Learning (SAIL), which provided virtual courses to support transitions to high school and college [19]. Using a Secure Data Enclave augmented with the Amazon Pinpoint and Lex managed services, RIPL deployed a chatbot in under four weeks to deliver texting-based behavioral nudges and financial incentives that encouraged students to attend virtual classes, submit weekly classwork, and prepare for the 2020-2021 school year. RIDE is now using the Secure Data Enclave to partner with RIPL researchers to measure the causal effects of nudge-based educational programs like SAIL and the Rhode2College [20] program on connecting students to successful pathways to college.
4. **Streamlining and securing data use across the Commonwealth of Virginia.** The Commonwealth of Virginia has more than 1,400 data systems used to track and measure policy and programs. In 2020, the Office of the Governor issued an Executive Order to establish a permanent data sharing and analytics structure for the Commonwealth to promote increased data sharing and leverage government data for evidence-based policy. RIPL partnered with the Office of the Chief Data Officer and the Center for Innovative Technology to deploy a Secure Data Enclave for accessing and analyzing data approved through DataSAGE, Virginia’s Secure Analytics and Governance Environment, which launched in August 2020. DataSAGE uses a Secure Data Enclave to facilitate secure data use by approved researchers under a data governance structure where government agencies review and approve access for important policy and research projects [21].

4 CONCLUSION

The comprehensive scope and availability of managed services in the cloud allow organizations to rapidly deploy cost-effective systems for securing and controlling access to sensitive data. We have described the key security controls to consider when implementing such a system, as well as a reference cloud architecture. Four use cases involving public sector partnerships illustrate how this architecture has been applied in practice to emergent public policy applications that require access to sensitive data.

ACKNOWLEDGMENTS

We thank Michael Hicklen and Kyle Hancock for their contributions to the AWS reference architecture. We thank the many organizations and individuals who have partnered with us to prototype, test, and refine the Secure Data Enclave concept as well as its practical implementation: Amazon Web Services Inc., the Rhode Island Department of Labor and Training, the Rhode Island Department of Education, the Rhode Island Department of Taxation, the Rhode Island Office of the Governor, the Center for Innovative Technology, the Commonwealth of Virginia’s Office of the Chief Data Officer, and Qlarion Inc.; Eric Schwenter, Casey Burns, Chris Johnson, Chuck Fuller, Scott Jensen, Jim Lucht, Abby McQuade, Amelia Roberts, Stephen Osborn, Scott Gausland, Elizabeth Texeira, Santiago Guerrero, Carlos Rivero, David Ihrle, Lynn McDaniel, Dawn Virts, Adam Roy, and Robert Reynolds.

Parts of this work were supported by National Science Foundation Convergence Accelerator (1937061) and RAPID (2029746) awards to Justine S. Hastings.

REFERENCES

- [1] Y. LeCun, Y. Bengio, G. Hinton. 2015. Deep learning. *Nature* 521: 436–444. doi:10.1038/nature14539
- [2] Office of the Press Secretary, The White House. 2016. Announcing Over \$80 million in New Federal Investment and a Doubling of Participating Communities in the White House Smart Cities Initiative. Accessed July 7, 2021 from: <https://obamawhitehouse.archives.gov/the-press-office/2016/09/26/fact-sheet-announcing-over-80-million-new-federal-investment-and>
- [3] Commission on Evidence-Based Policymaking. 2017. The Promise of Evidence-Based Policymaking. Accessed November 19, 2021 from: <https://www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf>
- [4] Schatsky D, Chauhan R. 2018. How CDOs can promote machine learning in government. Deloitte Insights. Accessed November 19, 2021 from: <https://www2.deloitte.com/us/en/insights/industry/public-sector/chief-data-officer-government-playbook/five-uses-machine-learning-government-for-cdos.html>
- [5] Hastings JS, Howison M, Lawless T, et al. 2019. Unlocking Data to Improve Public Policy. *Communications of the ACM* 62(10): 48–53. doi:10.1145/3335150
- [6] Hastings JS. 2019. Fact-Based Policy: How Do State and Local Governments Accomplish It? The Hamilton Project (Brookings Institution), Policy Proposal 2019-01. Accessed November 19, 2021 from: https://www.hamiltonproject.org/assets/files/Hastings_PP_web_20190128.pdf
- [7] Plale BA, Dickson E, Kouper I, et al. 2019. Safe Open Science for Restricted Data. *Data and Information Management* 3(1): 50–60. doi:10.2478/dim-2019-0005
- [8] Wilson M, Crompton S, Matthews B, Orlov A. 2011. Enforcing Scientific Data Sharing Agreements. In *Proceedings of the 2011 IEEE Seventh International Conference on eScience*, 2011, pp. 271–278, doi: 10.1109/eScience.2011.45. doi: 10.1109/eScience.2011.45
- [9] US Census Bureau. Federal Statistical Research Data Centers. Accessed July 20, 2021 from: <https://www.census.gov/about/adrm/fsrdc.html>
- [10] Centers for Disease Control and Prevention. Research Data Center. Accessed July 20, 2021 from: <https://www.cdc.gov/rdc/index.htm>
- [11] Lane J, Shipp S. 2007. Using a Remote Access Data Enclave for Data Dissemination. *International Journal of Digital Curation* 2(1): 128–34. doi:10.2218/ijdc.v2i1.20
- [12] Angell M, Gold S, Howison M, et al. 2020. Delivering Unemployment Assistance in Times of Crisis. *Digital Government: Research and Practice* 2(1): 5:1-5:11. doi:10.1145/3428125
- [13] Amazon Web Services. Active Directory Domain Services on AWS – Quick Start. Accessed July 26, 2021 from: <https://aws.amazon.com/quickstart/architecture/active-directory-ds/>
- [14] Amazon Web Services. Rebuild a Workspace – Amazon WorkSpaces. Accessed August 4, 2021 from: <https://docs.aws.amazon.com/workspaces/latest/adminguide/rebuild-workspace.html>
- [15] Amazon Web Services. AWS Transfer Family. Accessed August 4, 2021 from: <https://aws.amazon.com/aws-transfer-family/>
- [16] Center for Internet Security. Amazon Web Services Benchmarks. Accessed July 26, 2021 from: https://www.cisecurity.org/benchmark/amazon_web_services/
- [17] Amazon Web Services. How CloudTrail Works – AWS CloudTrail. Accessed July 26, 2021 from: <https://docs.aws.amazon.com/awscloudtrail/latest/userguide/how-cloudtrail-works.html>
- [18] Rhode Island Office of the Governor. Rhode Island to Launch Virtual Career Center powered by Google Cloud as part of “Back to Work RI” Initiative. Accessed September 2, 2021 from: <https://www.ri.gov/press/view/39624>
- [19] Rhode Island Department of Education. Rhode Island Students SAIL Into Summer Learning. Accessed September 2, 2021 from: <https://www.ride.ri.gov/InsideRIDE/AdditionalInformation/News/ViewArticle/tabid/408/ArticleId/696/Rhode-Island-Students-SAIL-Into-Summer-Learning.aspx>
- [20] Rhode Island Department of Education. Rhode Island Launches Statewide Incentive Program to Increase the Number of Students Prepared for College. Accessed October 13, 2021 from: <https://www.ride.ri.gov/InsideRIDE/AdditionalInformation/News/ViewArticle/tabid/408/ArticleId/511/Rhode-Island-Launches-Statewide-Incentive-Program-to-Increase-the-Number-of-Students-Prepared-for-College.aspx>
- [21] Office of the Chief Data Officer of Virginia. Virginia Secure Analytics and Governance Environment (SAGE). Accessed September 2, 2021 from: <https://www.cdo.virginia.gov/resources/datasage/>

A APPENDICES

A.1 Sample Security Pledge

Security Pledge for the Use of Confidential Data from <<INSERT DATA OWNER HERE>> (“Data Owner”)

I, _____, (“Approved User”) through my involvement with and work on << Insert Approved Project Title Here >> will have access to the sensitive and confidential data provided by the Data Owner to be used in producing research and analysis results. I understand that access to this data carries with it the responsibility to guard against unauthorized use and the possibility of unauthorized access or use. To treat information as confidential means not to divulge it to anyone who is not an Approved User, or to cause it to be accessible to anyone who is not an Approved User.

I understand that disclosing confidential information directly or allowing non-authorized access to such information may subject me to criminal prosecution and/or civil recovery.

I agree to fulfill my responsibilities on this project in accordance with the following guidelines:

1. I agree not to permit anyone access to these data, either electronically or in hard copy, unless to another Approved User.
2. I agree not to attempt to identify individuals, families, or households.
3. I agree that in the event an identity of an individual, family, or household is discovered inadvertently, I will (a) make no use of this knowledge, (b) advise << Insert Data Owner Point of Contact >> of the incident, (c) safeguard or destroy the information as directed by << Insert Data Owner Point of Contact >>, (d) not inform any other person of the discovered identity.

Approved User:

Witness:

Name: _____

Name: _____

Signature: _____

Signature: _____

Date: _____

Date: _____