Security in the Administrative Data Research Facility

ADRF Project Overview

The Administrative Data Research Facility (ADRF) provides a secure, cloud-based computing environment built to analyze confidential micro-data for evidence-based policymaking. The ADRF also includes features for data stewardship to authorize access and use and permit agencies to monitor the appropriate use of sensitive data across agencies in federal, state, and local government.

Agencies need to be able to share data across state and jurisdictional lines in order to respond to many social problems. For example, examining the impact of access to jobs and neighborhood characteristics on the earnings and employment outcomes of ex-offenders and social benefit recipients on their subsequent recidivism or retention on welfare requires data from at least four different agencies (Corrections, Human Services, Labor and Housing) – ideally from multiple states. The same holds true for describing the earnings and employment outcomes of different education pathways, since students may get jobs in multiple states. To make that research possible, data from multiple organizations must be linked – with authorizations for data access based on requests at an individual, per-person level.

The design of the ADRF makes such linkage possible. The ADRF established a secure environment in the Amazon Web Services GovCloud. Within AWS GovCloud, the ADRF uses well-known secure open source software including Project Jupyter, Docker, Kubernetes, and other contemporary infrastructure tools for collaborative projects hosted on scalable, secure computing environments.

The ADRF has achieved moderate FedRAMP certification[1], has received Authorization to Operate (AO) from the Census Bureau, and won a national innovation award in the process. It is available on the FedRAMP marketplace for agency use. The design means that each agency can put its data into its own secure environment within the secure FedRAMP boundaries, and control both access and use (see Figure 1). To date, the ADRF platform has provided secure access for approved projects to over 50 confidential government datasets from over 20 different agencies at all levels of government.

[1] FedRAMP is a government-wide program that provides a standardized approach to security assessment, authorization, and continuous monitoring for cloud products and services. FedRAMP created and manages a core set of processes to ensure effective, repeatable cloud security for the government.

Security Model and Compliance

The security model for ADRF is based on The Five Safes framework, a widely-accepted approach for making effective use of sensitive data. Figure 1 provides an overview of the design. ADRF security considerations for The Five Safes are:

Safe Projects

Is this use of the data appropriate, lawful, ethical and sensible? The design of the ADRF ensures that only approved projects which follow the relevant legal and ethical considerations are allowed; subsequent user interactions only occur through access-controlled project workspaces. In Figure 1, work occurs in the green, approved, workspace; approval is typically time-limited and hence temporary in nature.

Safe People

Can the researchers be trusted to use it in an appropriate manner?

Access is only allowed for individuals who have been approved by the data stewards. They must have completed Security Awareness training and their use is subject to careful review, management, and tracking. In Figure 1, the individuals who are allowed access are only those who are approved.

Safe Data

Does the data itself contain sufficient information to allow confidentiality to be breached?

Extensive import review processes guide restricted data into a secure, cloud-based data research facility, where only the datasets required for approved uses are allowed. In addition, personally identifiable information (PII) is typically hashed using a hash-based Message Authentication Code (HMAC) algorithm. The data in Figure 1 represent data that have been processed according to agency rules.

Safe Settings

Does the access facility limit unauthorised use or mistakes?

All data access and use must comply with a FedRAMP Moderate rating, based on implementation of standard security protocols, with GovCloud infrastructure. The boundary in Figure 1 represents the secure, FedRAMP, boundary.

FedRAMP (Federal Risk and Authorization Management Program) is a government-wide program that provides a standardized approach to security, facilitating a shift from insecure, tethered, tedious IT toward more secure, mobile, nimble, and rapid practices. It also creates a marketplace of solutions, within which ADRF is listed. FedRAMP requires monthly scans and reporting to demonstrate the continuous monitoring required for any software platform in the program to maintain its certification. Those scans ensure that ADRF adheres to all FedRAMP controls based on NIST 800–53 Revision 4 standards plus additional controls specific to cloud computing. All required security updates must be applied within a designated period: 7 days for critical-risk updates, 30 days for high-risk updates.

Safe Outputs

Are the statistical results non-disclosive?

The FedRAMP boundary ensures that data cannot be downloaded, since the environment is isolated from the internet. Any data exported from the environment must go through disclosure review to ensure that no confidential data gets exported.

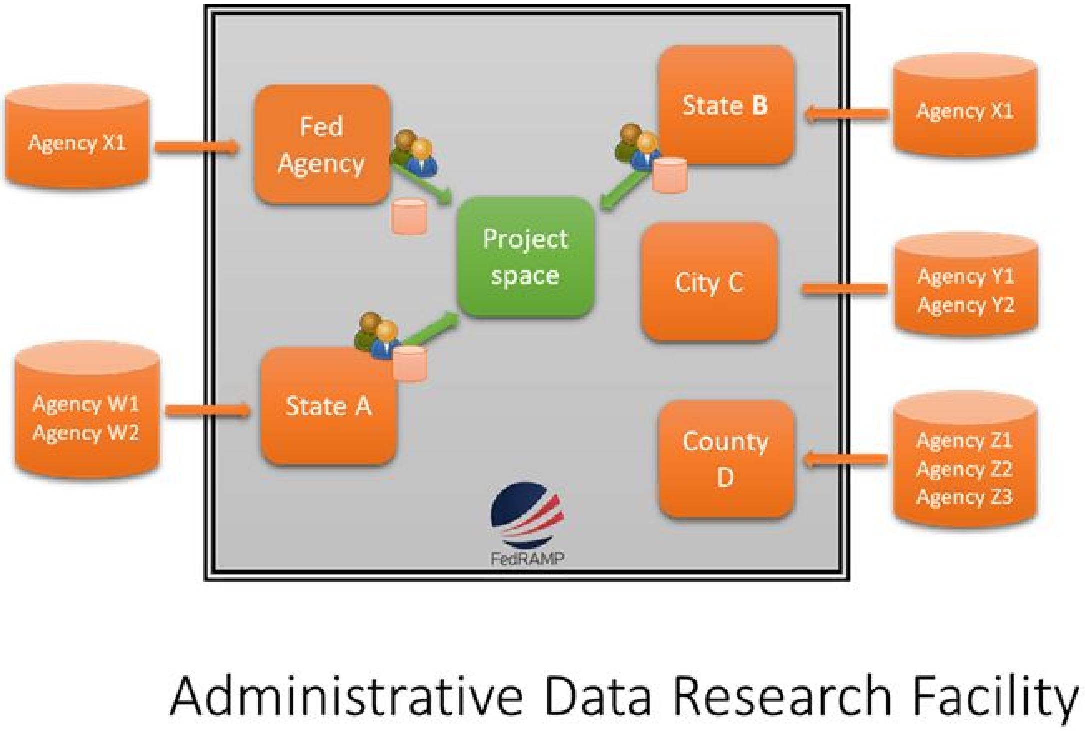


Figure 1: ADRF Design

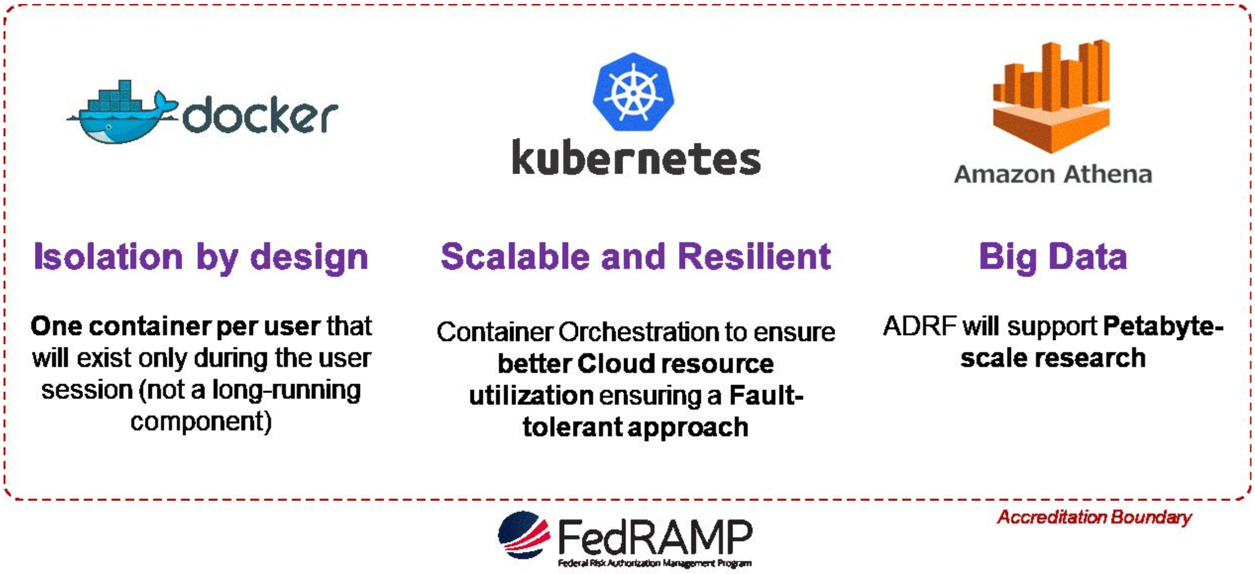
Secure Workspaces based on Cloud and Open Source

At its foundation, ADRF is built atop Amazon GovCloud, which is designed to host sensitive data, run regulated workloads, and address even the most stringent US government security requirements, including FedRAMP compliance that allows customers to host sensitive Controlled Unclassified Information (CUI). Internally, GovCloud includes facilities for Virtual Private Cloud (VPC) which closely resembles a traditional network operated within a private data center, with the benefits of the scalable infrastructure of AWS. Externally, GovCloud provides AWS Shield as an always-on protection service that safeguards applications from Distributed Denial of Service (DDoS) attacks.

As of June 1, 2019, ADRF launched its version 2.0 release, a major upgrade which now leverages open source container-based infrastructure to ensure that there is *no shared environment* between projects, i.e., complete isolation and controlled access to resources. These open source frameworks guarantee:

- *isolation* – runs one unique container per user/project session, destroyed once that session ends
- *scalable fault-tolerance* – ensures reliably run containers that make more efficient use of cloud resources, while reducing potential attack surface
- *well-defined configuration* – ensures that auditable best practices are followed as containers get deployed on the cloud

On the one hand, these open source frameworks – including Jupyter, Docker, Kubernetes, and Helm – represent contemporary "cloud-native" best practices for leveraging computing resources on GovCloud. On the other hand, they are widely used throughout industry, including mission-critical use cases at scale for Amazon, Google, Microsoft, IBM, and so on. Source code for these frameworks is therefore subject to extensive testing, code reviews, audits, and other scrutiny to identify and resolve flaws or potential security issues.



Isolation by design

One container per user that will exist only during the user session (not a long-running component)

Scalable and Resilient

Container Orchestration to ensure **better Cloud resource utilization** ensuring a **Fault-tolerant approach**

Big Data

ADRF will support **Petabyte-scale research**

Figure 2: ADRF Version 2.0

ADRF leverages components of Project Jupyter as its foundation for collaboration. This includes JupyterHub for launching collaborative cloud-based environments atop Kubernetes, and JupyterLab as an extensible, next-generation web-based user interface for remote computing based on Jupyter notebooks. Moreover, the ADRF team is working jointly with Project Jupyter to include features for enhancing security measures even further.

Security During Usage

A *workspace* in ADRF consists of a remote desktop session where a user can have access to JupyterLab, a network file system, a database, plus other related services. These resources are sufficient for most data science workloads. However, when work with large datasets is required, ADRF also makes use of Amazon Athena for serverless interactive big data queries with durable storage based on Amazon S3, which supports petabyte-scale research on GovCloud.

At the start of a user/project session, the IAM authentication and user management service in GovCloud enforces securely controlled access to AWS services. This allows ADRF administrators centralized management over users, security credentials (e.g., access keys), and permissions to control which AWS resources users and applications can access. Authentication features in turn integrate with how organizations manage individuals, groups, and resources through directory services such as Active Directory or LDAP, and single sign on procedures such as OAuth.

ADRF bases all data access on use of a data stewardship module (white paper forthcoming), which facilitates the data access approval process, reduces administrative time, and improves resource utilization. Rich, automated workflows support the data access request and approval process, along with features for user-generated metadata moderation, data access control, and reporting. Moreover, all data fields that represent personally identifying information (PII) are hashed to be anonymized for “Safe Data” use.

User access to the workspaces runs through VNC, which is a graphical remote desktop-sharing system that prevents the copying of data. Also, all workspace usage has *event logging* enabled, to support operational monitoring, intrusion detection, compliance auditing, and a-posteriori analysis of platform usage. The event logs get indexed in Elasticsearch and are easily searchable, providing a trail of auditable telemetry.

Long-Term Outlook

The ADRF team, in collaboration with Project Jupyter and others, is developing AI capabilities to enhance both the security and utility of cross-agency data science work in support of evidence-based policymaking. The long-term goal is to build user interfaces that present *rich context* to users about the datasets as they work within secure environments while adhering to privacy and confidentiality standards, such as GDPR. In other words, ADRF collects metadata about dataset usage: who else has used the data, for what purpose, and how it was accessed and analyzed data in their work. Machine learning, knowledge graph, and other AI technologies can then be leveraged to suggest appropriate uses of data, and ostensibly to support enhanced security. Through the telemetry enhancement to the Jupyter stack, expected in 2020, analysis performed in the ADRF will be auditable at as detailed a level as individual data elements when required.