AAA Assessment Draft

Center for Data Intensive Science
University of Chicago

December 29, 2017
Draft 1.0

Contents

Acronyms

ARM	App Reroute Method
AWS	Amazon Web Services
CBIIT	Center for Biomedical Informatics and Information Technology
CLI	Command-Line Interface
CR	Cloud Resource
DCF	Data Commons Framework
FISMA	Federal Information Security Management Act
GCP	Google Cloud Platform
GDC	Genomic Data Commons
IAM	Identity and Access Management
ISB	Institute for Systems Biology
ISB-CGC	ISB Cancer Genomics Cloud
JWT	JSON Web Token
NCI	National Cancer Institute
NIH	National Institutes of Health
SA	Service Account
SBG	Seven Bridges Genomics
SSP	System Security Plan

1 Introduction

The purpose of this document is to describe the current infrastructure and process for authentication and authorization for Seven Bridges Genomics (SBG), Institute for Systems Biology (ISB), and Broad Institute (Broad) Cloud Resources (CRs) and assess the pros and cons for each approach. Portions of this report were developed with feedback from the three Cloud Resources.

A cloud-agnostic approach or an approach with a cloud resource using another commercial cloud like Azure may need additional retooling for an optimal solution. The recommendations following are for implementing the Data Commons Framework (DCF) with the Cloud Resources and their cloud native approaches. As such, the recommendations as documented below are equivalent to documenting the current state for authentication, authorization, and data access. The three approaches may eventually have use cases across each of the resources, although one use case is expected to be dominant for each resource. These use cases are summarized as:

1. Pre-signed URLs, mainly for Seven Bridges Genomics.
2. A token vending machine, mainly for Broad Institute.
3. An application reroute method, only for the Institute for Systems Biology.

This report will document benefits, risks, and challenges associated with the implementation for each Cloud Resource.

2 Recommended Implementations

It is important to note that the Data Commons Framework team uses an agile-informed development methodology. Agile values iteration, which means that while the general shape of the recommendations are expected to remain the same, some changes are expected. Technology partners will change, and the DCF wants to change accordingly to help serve the mission as given by the NCI.

The goal of the authorization and authentication services provided by DCF is to provide a standard interface for users as well as cloud resources to access controlled data that is located in different storage locations. In broad strokes: cloud resources will access GDC data in the commercial clouds by using DCF's single sign on endpoint, which redirects users straight to NIH or eRA ID Login. Authorization information will be stored in a relational database and synced from dbGaP every 6 hours. Users have two possible roles: submitter and downloader, and they can have access to multiple projects and programs.

Cloud resources will use the OAuth2 flow to login through DCF and obtain JSON Web Tokens (JWTs), an `access_token` and a `refresh_token`. The `access_token`'s signed payload includes a minimal piece of authorization information about the user.

The Single Sign On endpoint is cloud agnostic. However, it is recommended that additional cloud native options be implemented in order to leverage fully the power of each cloud. We want to grant user access in a way that make users and a CRs previous activities work seamlessly. This means that we want to support whatever is the best approach in each cloud. This principle inevitably implies that the solution will not be the same for each cloud.

2.1 Pre-Signed URLs

The DCF recommends pre-signed URLs for object access. The end user will communicate the objects desired, and the CR will allow the user to connect to the DCF authorization and authentication API to obtain a pre-signed URL. The pre-signed URL allows for download of data into the users VM on the CR. The pre-signed URL implementation is still under development, and thus does not have a defined workflow at this moment. The workflow should be completely or nearly invisible to the end user, as the interactions are largely between the CR and the DCF.

Pre-signed URLs are temporary, valid for only one use, and access can be revoked if needed. If a user runs a workflow that lasts longer than the URL's validation period, they can ping the system for a new one.

Google Cloud Platform (GCP) and Amazon Web Services (AWS) support pre-signed URLs. Seven Bridges has used pre-signed URLs within their Federal Information Security Management Act (FISMA) moderate environment, serving as a precedence for FISMA Moderate compliance. Upon review by the Cloud Operations team at the University of Chicago's Center for Data Intensive Science, pre-signed URLs are a secure, compliant implementation for object access.

Seven Bridges Genomics and the Broad Institute accept this recommendation as a reasonable execution within their use cases. SBG currently plans only to use this authentication and authorization implementation from the DCF. Broad will use it for a subset of their authentication and authorization use cases. The Institute for Systems Biology has expressed no current plans to use pre-signed URLs for their users.

To summarize, pre-signed URLs are recommended as the best option for Cloud Resources operating on AWS, and one of the options that will be supported for Cloud Resources operating on GCP, with the understanding that other options will also be implemented to support additional use cases required by the Cloud Resources operating on the GCP.

2.2 Token Vending Machine

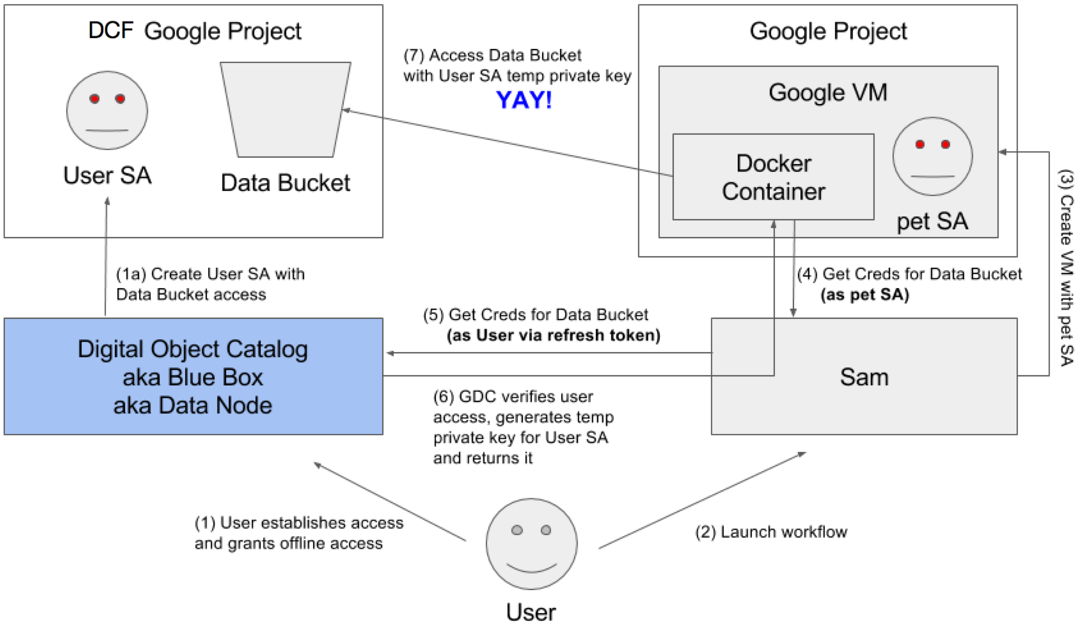

Figure 1: The token vending machine workflow proposed by Broad, allowing users to receive tokens for retrieving objects from the DCF Google project. [4]

Whereas pre-signed URLs are best for AWS, Cloud Resources leveraging the services of Google Cloud Platform may also want to use a "token vending machine" for authentication and authorization. This recommended implementation allows end users to receive tokens to use for retrieving objects from the GDC storage buckets in the GCP. Figure 1 displays the workflow as proposed by Broad. The process in the image below shows a user connecting their dbGaP credentials to the account they use in FireCloud, the DCF and Broad associating a service account with the user, and, finally, a token allowing the user to bring genomic data from the GDC on GCP to the docker container on the Google Project associated with the user in FireCloud.

Broad is migrating to this mechanism in early 2018. This approach allows users to have two distinct identities: their Google identity and a personal Service Account called a pet Service Account to authenticate and act as a user. Google resources are not shared directly with either of these identities. Google resources are shared with a Google group known as a proxy group, the only members of which are the user's Google identity and pet Service Account. Access to protected data is managed via a Google group that is synchronized every 8 hours with the dbGaP whitelist. Users must link their Google identity with their era commons identity. The Google group contains proxy groups and not user Google identities directly.

Tokens have precedence for FISMA moderate compliance in the Genomic Data Commons. Users of the Bionimbus Protected Data Cloud have brought in data from GDC using the access tokens. Both Bionimbus and GDC are FISMA Moderate environments. This particular implementation allows the CR to obtain the token on behalf of the user and place the data in the VM for the user. At no point does the CR act as the user, but rather acts as a conduit for a user to achieve authentication and authorization for GDC data in GCP.

This method is recommended for GCP Cloud Resources alongside the pre-signed URL option. Broad is currently developing an Identity and Access Management (IAM) service called SAM that will interact with the DCFs token vending machine. More information on SAM can be found at [5]. The ISB-CGC may interact with this implementation in the future, but has not expressed plans to do so in the near future.

2.3 App Reroute Method (ARM)

The final implementation is recommended only for the Google Cloud Platform and only for the ISB Cancer Genomics Cloud (ISB-CGC). It is an alternative method in order for the ISB-CGC to continue providing access to GDC data in GCP without developing a new service or method to interact with the DCF.

Instead of the CR interacting with the DCF's authentication and authorization API, users will be redirected from the ISB-CGC application to a landing page for the DCF. The only current use case for this landing page is to be the authentication and authorization arm of the ISB-CGC, hence "ARM."

1. User needs to register their GCP project with ISB-CGC
2. DCF asks user to login through NIH
3. User needs to add DCF Service Account to their GCP project so that DCF can monitor project membership
4. DCF checks with Google for members in the project

5. User provides the identifier for the service account
6. DCF checks if the Google project is under organization (only ISB) whitelist, or the project has to be not attached to any organization.
7. DCF checks the full list of members of the project, disallow a project that has anything thats not a user/service account with a "role" in the project, disallow SAs from other projects except for DCF SA and Google system SAs.
8. DCF checks all users have the same set of different authorized datasets—for the datasets that the SA is intended to access
9. Disallow user managed keys for all service accounts
10. Disallow any roles attached to service accounts.
11. Cron job per minute to do steps 6–10 and revoke SA access when any of them fail and email the project owner
12. Add validated service account to authorized datasets google groups.

The Center for Biomedical Informatics and Information Technology (CBIIT) and the DCF will need to ask the GDC for permission to host the ARM on GDC on-prem, in their FISMA Moderate environment. The site will maintain compliance with FISMA through regular scans as noted in GDC's System Security Plan (SSP).

After authentication and authorization, ISB-CGC adds users to authorized groups for access to particular objects. NCI has granted the Trusted Partner status and certified ISB-CGC as compliant with FISMA standards.[1] However, it is recommended that object access is trackable upon a per user basis. If a user is compromised, access can be removed, but neither ISB-CGC, DCF, nor GDC will be able to know what objects were accessed by the compromised user or whether any breaches of protected information actually happened. Furthermore, compliance with the Trusted Partner data security policy could be at risk from a strong audit.[2]

Since this implementation requires a front end, the user journey will be impacted. Support for the ARM will not only be between DCF and the CR, but also include the user who is aware of the additional party. CBIIT, DCF, and ISB-CGC need to discuss user support for ARM and make a recommendation to GDC, who will need to be involved since the ARM will be hosted by GDC.

Although this solution could be implemented with future Cloud Resources, it is recommended that a new Cloud Resource would develop their system by working directly with the DCF API.

3 Benefits

3.1 Seven Bridges Genomics

1. Pre-signed URLs are cloud agnostic, if a CR operates on both AWS and GCP and use pre-signed URLs to distribute data, user experience will be the same regardless of the underlying cloud provider.

3.2 Broad Institute

1. Able to track usage of Service Account
2. Revoking one users access doesn't interrupt other users activities
3. The service account is owned by Broad, so the CR has full control of how the service account can be used.
4. Compared to ISBs approach to grant users SA directly, this approach decoupled the dependencies between the data owner and the CR, making the communication much cleaner.

3.3 Institute for Systems Biology

1. Compared to pre-signed URLs, Service Accounts are considered best practices in the Google Cloud. Virtual Machines are assigned a Service Account at startup, and Service Accounts fit in with the rest of Googles infrastructure.
2. Since Service Account is owned by the user, the user experience for users who are already familiar with GCP is seamless.
3. Users can configure it to have the all roles they need. The Virtual Machine can operate using just this Service Account for both data retrieval and using other GCP resources.

4 Challenges and Risks

4.1 Seven Bridges Genomics

1. There is more work on the CR side to develop both client-side tools and server-side tools to handle pre-signed URLs. On the client side, CR needs to develop tools to exchange user token with pre-signed URLs; in workflow environment, CR needs to develop tool to handle high-throughput data transfer with pre-signed URLs. SBG provides a CLI with support for data transfer and a fuse system to mount data using pre-signed URLs, both are not open sourced.
2. Some of the cloud native services/tools will be hard to use if they are not designed to work with pre-signed URLs.
3. Since generating pre-signed URLs do not need users to have any account in particular cloud provider, billing can't be passed directly to users' cloud accounts.

4.2 Broad Institute

1. This approach requires some work on the CR side to handle token exchanges and passing tokens/keys to compute engines.
2. Because of (1), it is harder for users to use the data outside of a CR environment.
3. Because the credentials handling on both the CR side and data owner/DCF side is encapsulated, passing data access billing (if needed) from data owner/DCF to users will also be a challenge.

4.3 Institute for Systems Biology

1. Data owner needs visibility into user project. This means prompting the user to add an ISB-CGC service account to their project with an Editor role. This has a clear security downside.
2. The new "organization" concept introduces some new problems for GCP projects that have a "parent organization" from which IAM roles can be inherited.

3. Need for Polling. Currently, polling is the only way to insure that a GCP project that is using Service Accounts continues to meet the requirements that have been confirmed to be in place when the service account is initially approved for use.

4. Scalability. The 60-second cron job needs to grind through all projects with active Service Accounts, and take appropriate actions. The resources to complete this will grow as user base grows. Also, if the actual processing takes longer than 60 seconds, concurrency issues will come to the fore and require more in-depth engineering to avoid multiple cron jobs from stepping on each other.

5. Non-trivial setup. Membership in a Google Project needs to be tailored so that all members have permissions for the data the Service Account will access. If a user is added to the project before they have linked their Google and eRA Commons identities, or if they are added while not being on every whitelist for all the data sets the Service Account is using, the account will get kicked off.

Notes

1. Users register the GCP project and add an ISB-CGC service-account as a project Editor to their project to allow for the ISB to check project membership. They register a Service Account to access the controlled data associated with one or more "programs" (e.g. TCGA, TARGET, etc). This step adds the Service Account to the Google groups that are on the ACL for the relevant controlled access buckets.

   When the user requests that this Service Account be given access to controlled-data, ISB looks at the project membership to confirm that all members have the correct authorization (regardless of what precise IAM roles they have in the project). If this does not check out, the attempt to register the SA for access to controlled-data is rejected.

   An "end-user" GCP project may have been set up by ISB-CGC and "given" to the end-users, along with some NCI-funded cloud credits, or the GCP project may have originated from outside of ISB-ISB-CGC. An end-user GCP project may have any number of associated Service Accounts.

2. The policy states: "at a minimum, policies, procedures, controls, and standards comparable to the HHS Information Security Program to ensure the integrity, con-

fidentiality, and availability of Federal information systems must be adopted and implemented" [1].

The cybersecurity policies of the HHS have specific policies for reporting breaches of information. A breach involving a compromised user would be not be auditable in this implementation. See also [2] and [3].

References

[1] US Department of Health and Human Services. URL: `http://www.hhs.gov/ocio/index.html` (accessed 2017-12-26).

[2] US Department of Health and Human Services. URL: `https://www.hhs.gov/sites/default/files/hhs-ocio-policy-2013-0002.pdf` (accessed 2017-12-26).

[3] US Department of Health and Human Services. URL: `https://www.hhs.gov/about/agencies/asa/ocio/cybersecurity/policy-it-security-and-privacy-incident-reporting-and-response/index.html` (accessed 2017-12-26).

[4] Broad Institute. Image courtesy of Broad, edited to change "GDC Google Project" to "DCF Google Project." URL: `https://github.com/broadinstitute/sam/blob/develop/data_access.png` (accessed 2017-12-26).

[5] Broad Institute. URL: `https://github.com/broadinstitute/sam` (accessed 2017-12-26).