SevenBridges

KEEPING GENOMIC DATA SAFE ON THE CLOUD

KEEPING GENOMIC DATA SAFE ON THE CLOUD

Cloud computing is a powerful tool to deal with the exponential growth of storage and processing requirements of genomic data. However, cloud computing also raises concerns around keeping this data safe.

Seven Bridges has been offering genomic data processing on the Amazon Web Services *(AWS)* cloud since early 2012, processing Petabytes of genomic data for thousands of users.

INTRODUCTION

Security and privacy are essential when dealing with genomic data and its analysis. Patients are rightfully concerned about the privacy of their genetic information. For researchers in academic and commercial environments, the safekeeping of valuable intellectual property derived from genetic information becomes an additional goal.

The general concerns around genomic data security naturally become amplified when this data is kept and processed in a cloud computing environment, i.e., an environment that is by design shared between different parties and entities. And even though we would argue that, objectively, "local" server structures are no safer than a well-managed cloud *(quite to the contrary in some cases we've observed)*, recent political events have only contributed to instilling a general sense of distrust toward cloud environments.

To gain the trust of customers to store and process genomic data with Seven Bridges on the cloud, we have designed and implemented a comprehensive security framework for keeping this data safe.[1] While our framework certainly needs to ensure compliance with current data protection standards *(such as HIPAA regulations in the US and the European Data Protection Directive)*, it also offers concrete implementation suggestions in areas where existing standards do not yet provide certainty with respect to specific security requirements for genomic data and the global nature of cloud environments.

CURRENT REGULATORY ENVIRONMENT

On a global scale, the regulatory space for data protection is vast and cannot be sufficiently explored within the scope of this whitepaper. However, several regulatory frameworks are commonly named in discussions around genomic data security and privacy:

1) Regulations issued pursuant to the US Health Insurance Port*ability* and Accountability Act *(HIPAA)*, which aim to protect all *"Protected Health Information" (PHI)* and consist of four parts:

- **Privacy Rule**, which protects the privacy of medical records and other personal health information;
- **Security Rule**, which sets national standards for the security of electronic protected health information;
- **Breach Notification Rule**, which requires covered entities and business associates to provide notification following a breach of unsecured protected health information; and the confidentiality provisions of the
- **Public Safety Rule**, which protect identifiable information being used to analyze patient safety events and improve patient safety.

2) The Clinical Laboratory Improvement Amendments *(CLIA)*, a set of US federal regulatory standards that apply to all clinical laboratory testing performed on humans in the United States, except clinical trials and basic research.

3) The Data Protection Directive, a European Union directive that regulates the processing of personal data within the European Union. *(Will be superseded by the European General Data Protection Regulation)*.

There are also general IT security frameworks to consider such as SOC 2, CoBIT, and ISO 27001 which we will not address specifically in this whitepaper.

All of these frameworks provide comprehensive guidance for IT and health data security in general. However, they do not *(yet)* provide specific guidelines as to dealing with genomic data or some specific challenges of cloud environments. With the exception of some specific national regulations — Norway and Brazil being two prominent examples — the current regulatory frameworks and compliance requirements therefore seem to be solid foundation, but not entirely sufficient to deal with security and privacy issues around keeping genomic data safe in a cloud environment.

THE SEVEN BRIDGES SECURITY FRAMEWORK

We believe that it's our job to provide users with end-to-end security and control over their data and analysis so that they can focus on work rather than having to deal with complex setups, compliance headaches, and security. To achieve this, we have designed a comprehensive security framework for processing genomic data in the cloud that covers three main areas:

1) **DATA SECURITY:** Ensuring that all sensitive data is kept safe during its full life-cycle. This includes data encryption and secure user authentication.

2) **PLATFORM AND INFRASTRUCTURE SECURITY:** Ensuring that the software platform and its underlying infrastructure *(server and network)* support the secure architecture.

3) **SECURITY CONTROLS:** Ensuring security of the system by implementing administrative, technical, and other security controls, while at the same time ensuring compatibility with a broad range of trusted information security frameworks and compliance requirements.

The following three sections present Seven Bridges' take on each of these three areas and provide concrete implementation examples from the Seven Bridges cloud platform. Since the cloud version of the Seven Bridges Platform is built on pre-already existing cloud infrastructure, such as that provided by Amazon Web Services (AWS) and Google Cloud Platform, we use provider AWS terminology for the remainder of this paper, such as "S3" to denote storage buckets and "EC2" to denote computation instances. Please refer to http://aws.amazon.com/ or https://cloud.google.com/ for details.

1. DATA SECURITY

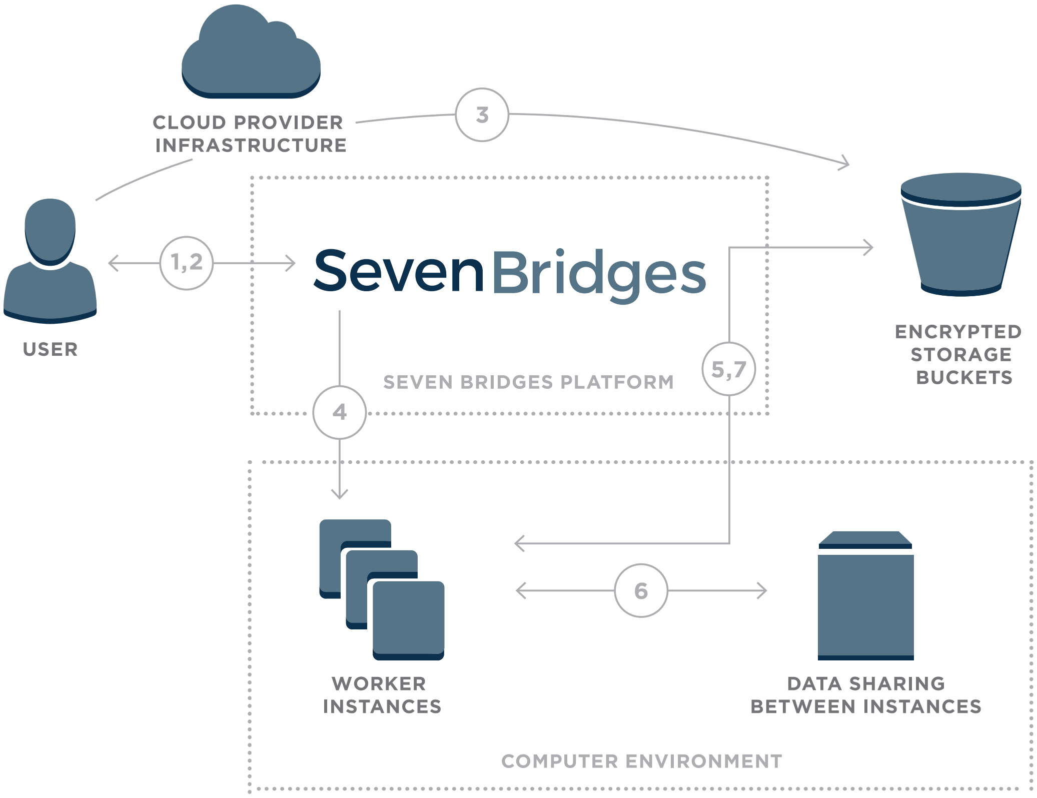


Figure 1 - Overview of data flow on the Seven Bridges Platform

1) User logs onto the platform
2) Platform creates a unique signed URL for the user
3) Using signed URL, data is uploaded to an encrypted storage bucket
4) Once the user starts a compution, the Seven Bridges platform calculates the optimal execution plan and starts up task worker instances
5) Worker instances securely pull data from cloud storage
6) Worker instances are able to securely share intermediate data
7) Final results are uploaded to cloud storage

At the end of the data life-cycle, a strict data purging policy ensures that all data is safely deleted if it is no longer needed on ephemeral storage or when an authorized user chooses to delete data on the platform.

While encryption and the safe purging of data are key to protect data in the cloud, they are only effective with appropriate authorization and access controls.[2]

Access controls on the Seven Bridges Platform have been implemented in a very fine-grained manner. Rather than establishing principal “file owners,” access permissions are set on a per-user-per-project basis, meaning that a user’s access permissions to a given file can depend on the context *(project)* in which this file is being used. This includes sharing of data, which can only be performed via the platform itself unless users have the appropriate permissions to download a file. Seven Bridges keeps secure audit logs for all data access for six years to ensure regulatory compliance.

By default, users authenticate on the platform through a username and secure password. However, the Seven Bridges Platform can support more strict data access control with client-encryption, two-factor authentication, and integration of external key management.

2. PLATFORM & INFRASTRUCTURE SECURITY

One clear advantage of running a platform in the cloud is that the cloud provider will usually offer a broad spectrum of built-in compliance and security features for the underlying infrastructure. For example, both Amazon Web Services and Google Cloud Platform provides a broad spectrum of security features[3] and standards compliance,[4] ranging from physical datacenter security and network infrastructure security to secure media handling and data encryption.

Naturally, compliance of the cloud provider does not imply compliance of the overall platform and ecosystem, but it is a solid foundation on which to build. Some regulations such as HIPAA even require the compliance of all individual providers within an environment (*which, in the case of HIPAA, are linked to each other through so-called Business Associate Agreements).*

In addition, Seven Bridges secures its infrastructure in a number of ways:

1) All AWS computation instances run within Virtual Private Clouds (VPC). VPCs are logically isolated networks within the AWS cloud and kept only minimally open for the necessary external and internal access.

2) All Google Cloud (GCE) computation instances run on separate networks, similar in functionality to AWS VPC with similar firewall configuration, allowing only necessary services for external and internal access.

3) Users can choose to isolate all computation resources through an "Instance Lockdown" mode that disables any access during the computation, even by platform admin staff.

4) Best security practices regarding tenancy, such as AWS Architecting for HIPAA are followed.

5) Computation instance reuse is limited to the same user and project to prevent data leakage. Bioinformatics apps run within LXC containers managed by Docker software with restricted Linux capabilities and strict firewall implemented.

6) Bioinformatics apps run within LXC containers managed by Docker software with restricted Linux capabilities and strict firewall implemented.

Access to the production and development environments are secured through Virtual Private Networks as shown in Figure 2.

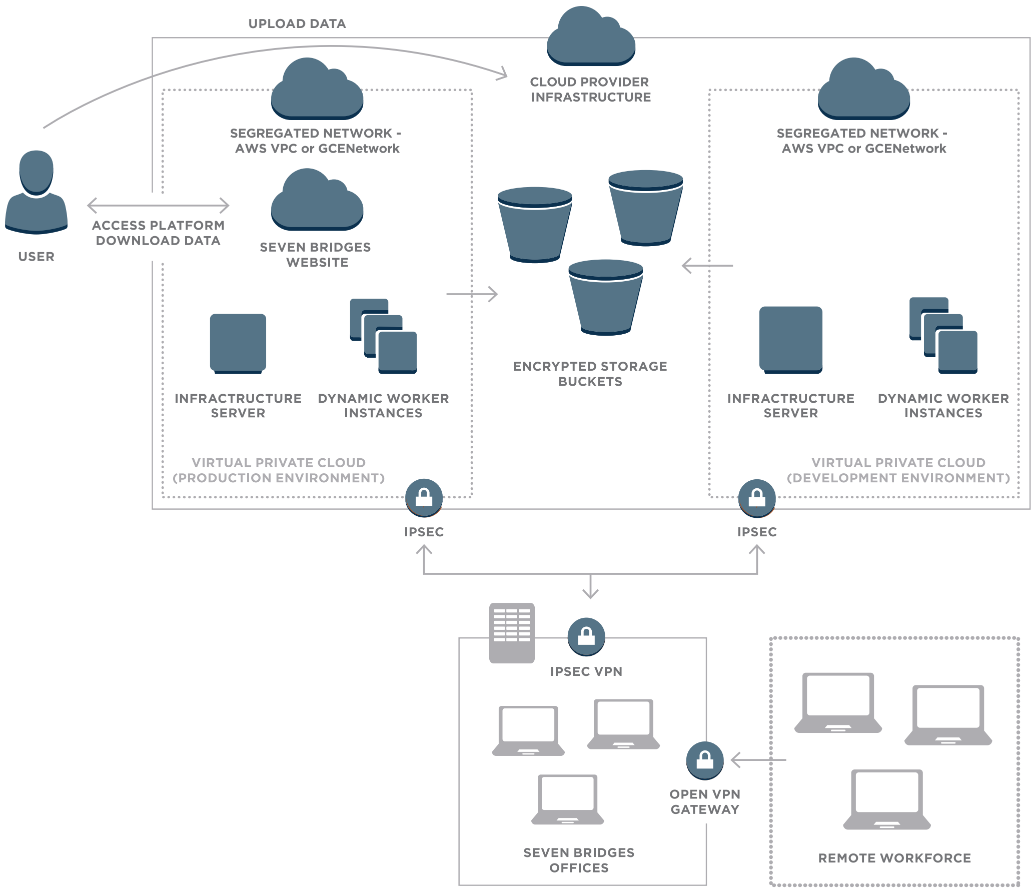
A second pillar to ensure that the Seven Bridges Platform environment is always secure is to constantly monitor and improve our security by following best-practices of infrastructure stability and security, including

1) regular software and infrastructure vulnerability assessments to discover vulnerabilities and remediate them,

2) regular penetration tests to discover vulnerabilities in the system which may not be noticed in a regular vulnerability assessment,

3) regular audit log analysis and system-level inspection to look for suspicious behavior, potential attacks, and security breaches,

4) a strict patch-management policy and regular server updates *(depending on criticality, the response/fix time is between a few hours and 7 working days)*, and restriction of access for technical staff to resources on a per-need basis.



UPLOAD DATA
CLOUD PROVIDER INFRASTRUCTURE
SEGREGATED NETWORK - AWS VPC or GCENetwork
SEGREGATED NETWORK - AWS VPC or GCENetwork
USER
ACCESS PLATFORM DOWNLOAD DATA
SEVEN BRIDGES WEBSITE
ENCRYPTED STORAGE BUCKETS
INFRACTRUCTURE SERVER
DYNAMIC WORKER INSTANCES
INFRACTRUCTURE SERVER
DYNAMIC WORKER INSTANCES
VIRTUAL PRIVATE CLOUD (PRODUCTION ENVIRONMENT)
VIRTUAL PRIVATE CLOUD (DEVELOPMENT ENVIRONMENT)
IPSEC
IPSEC
IPSEC VPN
OPEN VPN GATEWAY
SEVEN BRIDGES OFFICES
REMOTE WORKFORCE

Figure 2 - Overview of network security on the Seven Bridges Platform

3. SECURITY CONTROLS

As in any IT security framework, security of the system must also be ensured by implementing administrative, technical, and other security controls. The Seven Bridges security framework aims to establish these in such a way that compatibility with a broad range of trusted information security frameworks and compliance requirements *(such as HIPAA and ISO27001)* is ensured at the same time.

A general aim should be to make all software and infrastructure fully compliant with the NIST 800-53A moderate profile, which Seven Bridges considers to be the most complete set of controls which can be easily mapped to the majority of accepted information security and compliance frameworks.[5] These controls cover all areas of information security, namely access control, security awareness and training, auditing and accountability, security authorizations, configuration management, contingency planning, authentication, incident response, dealing with equipment maintenance, secure media handling, physical and environmental security, risk management and security planning, personnel security, systems and network security, dealing with supply chain security, and system and information integrity. Proper implementation of these controls requires a broad range of policies that need to be effectively implemented as shown in Table 1.

PURPOSE	POLICY /CONTROL	MAIN CONTENT
Ensure proper controls	Risk Assessment and Management	How to treat risk and what controls to implement in order to support the required level of information security and minimize risk to business operations and customer data
Keep data safe	Information Security	Detailing the information security stance, regarding infrastructure and customer data
	Encryption	How to use encryption to secure data, both in transit and at rest. How to properly manage secret encryption keys
	Workplace Security	How to secure the workstation, clean desk policy, expected employee behavior
Ensure compliance	Security Awareness and Training	How to train staff and make them aware of security issues and procedures
	Sanctions	How to deal with employees and subcontractors who violate the policies and procedures
Maintain ongoing platform security	Change Management	How to manage changes in software and infrastructure to minimize information security and operations risk
	Asset Management	How to manage information assets, maintain a good inventory and information system border
	Vulnerability Management	How to manage vulnerabilities in software and infrastructure, namely discover and remediate them without disrupting business operations
	Logs and Auditing	What to audit, how and why; how to analyze logs and provide reporting
Keep infrastructure safe	Network Security	How to secure the network, how to connect remote systems and users, which protocols to use and how in order to maintain proper level of security
	Antimalware	How to defend from malware
	Acceptable Use	Details employee and subcontractor obligations
	Patch Management	How to manage patches to software and underlying infrastructure
Plan ahead	Security Incident Response	How to treat security incidents and potential breaches; how to notify customers of potential breaches
	Disaster Recovery *(including Recovery Plan)*	How to manage backups and how to recover infrastructure and customer data in a case disaster strikes

Table 1 – Policies required for implementation of comprehensive security controls

STATUS AND OUTLOOK

The explosion of genomic data and rise of cloud computing are recent developments. Currently Seven Bridges has implemented all infrastructure, policies and controls to ensure that:

1) Customers who enter into a Business Associate Agreement with Seven Bridges can process Protected Health Information *(PHI)* on the Seven Bridges Platform in full compliance with HIPAA
2) Customers who include EU-approved model contract clauses in their agreement with Seven Bridges may upload EU data subjects' personal data to the Seven Bridges Platform in compliance with the EU Data Protection Directive

Over the coming months and years, we expect regulatory standards on a national and international level to change and evolve to reflect the specific security and privacy challenges that are now emerging.

Likewise, we expect our security framework to iterate and constantly evolve as well.

CONTACT

Please feel free to check back regularly with us at www.sevenbridges.com to learn about our recent developments and obtain the most recent version of this whitepaper. For any questions or concerns, please feel free to reach out to security@sevenbridges.com.

REFERENCES

1. This development was done in close collaboration with various security experts within our customers' organizations and in close collaboration with Amazon Web Services.
2. Some go so far as to make the argument that proper user authentication is much more important than encryption since, from a risk perspective, an access breach is by comparison much more likely than a physical security breach.
3. See http://aws.amazon.com/security/ ; https://cloud.google.com/security/
4. See http://aws.amazon.com/compliance/ ; https://cloud.google.com/security/compliance
5. For some frameworks such as HIPAA, there have been debates as to their applicability to genomic data. In spite of these debates, we believe that information security controls should certainly follow the more general NIST 800-53A control recommendations with respect to HIPAA/PHI.

EXTERNAL REFERENCES

Amazon Web Services Compliance Center: http://aws.amazon.com/compliance/

Amazon Web Services Compliance Center: http://aws.amazon.com/compliance/

Amazon Web Services Risk and Compliance Whitepaper: https://d0.awsstatic.com/whitepapers/compliance/AWS_Risk_and_Compliance_Whitepaper.pdf

Amazon Web Services Security Center: https://aws.amazon.com/security/

Architecting for HIPAA Security and Compliance on Amazon Web Services: https://d0.awsstatic.com/whitepapers/compliance/AWS_HIPAA_Compliance_Whitepaper.pdf

Google Cloud Security Whitepaper: https://cloud.google.com/security/whitepaper

Google Cloud Platform provides support for HIPAA covered entities: https://cloudplatform.googleblog.com/2014/02/google-cloud-platform-provides-support-for-hipaa-covered-entities.html

NIST SP 800-30 Rev 1 "Guide for Conducting Risk Assessments": http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf

TEAM@SEVENBRIDGES.COM
SEVENBRIDGES.COM

SevenBridges