Architecting for Genomic Data Security and Compliance in AWS

Executive Overview

Angel Pizarro
Chris Whalley
Carina Veksler

December 2014

Contents

Overview

Individual-level genotype and phenotype research continues to identify breakthrough treatments for a variety of health issues. However, as this data continues to grow in volume and utility, the availability of adequate data processing, storage, and security technologies has become a critical constraint on genomic research.

Cloud computing provides a simple way to access servers, storage, databases and a broad set of application services over the Internet—and the global research community is recognizing the practical benefits of Amazon Web Services (AWS).

When evaluating a cloud platform, researchers need to consider security best practices for human genomic data and controlled access datasets such as those from National Institutes of Health (NIH) repositories like Database of Genotypes and Phenotypes (dbGaP) and genome-wide association studies (GWAS). One of the most basic architectural considerations for dbGaP compliance in AWS is whether the architected system will run entirely on AWS, or as a hybrid deployment with a mix of AWS and non-AWS resources. This whitepaper focuses on the control areas for AWS resources when architecting hybrid deployments.

Genomic Data Privacy and Security in Human Research

Researchers often have questions about security and compliance as they assess AWS for running genomic sequence data. Specifically, they need to understand how to meet guidelines and best practices set by government and grant funding agencies such as the National Institutes of Health. This presents scientific investigators, institutional signing officials, IT directors, ethics committees and data access committees with a number of common questions including:

- Is data protected on secure servers?
- Where is the data located?
- How is access to data controlled?
- Are data protections appropriate for the Data Use Certification?

These considerations are not new, nor are they cloud-specific. The essential considerations for human genomic data are the same regardless of where data resides: in an investigator lab; an institutional network; an agency-hosted data repository; or the AWS cloud.

When planning research systems, data protection and security controls need to be clearly defined, and then architected into the system design. This is particularly important when evaluating a shared responsibility model.

The AWS Approach to Shared Security Model

AWS delivers a robust web services platform with features that enable research teams around the world to create and control their own private area *in* the AWS cloud with rapid access to flexible and low cost IT resources. And with cloud computing, it is not necessary to make large upfront investments in hardware, or spend time maintaining systems and facilities. However, since AWS does not access or manage the customer's private AWS environment or the data in it, customers retain both the responsibility and accountability for the configuration and security controls implemented in their AWS account. Customer accountability over their private AWS environment is fundamental to understanding the respective roles of AWS and our customers with regard to data protection and security practices for human genomic data.

The responsibility for security in the AWS cloud is shared between clients and AWS.

- AWS secures the underlying infrastructure.
- Researchers secure anything you put on the infrastructure (such as operating systems, platforms, and data) with a view of meeting your specific regulatory and business compliance requirements for your data.

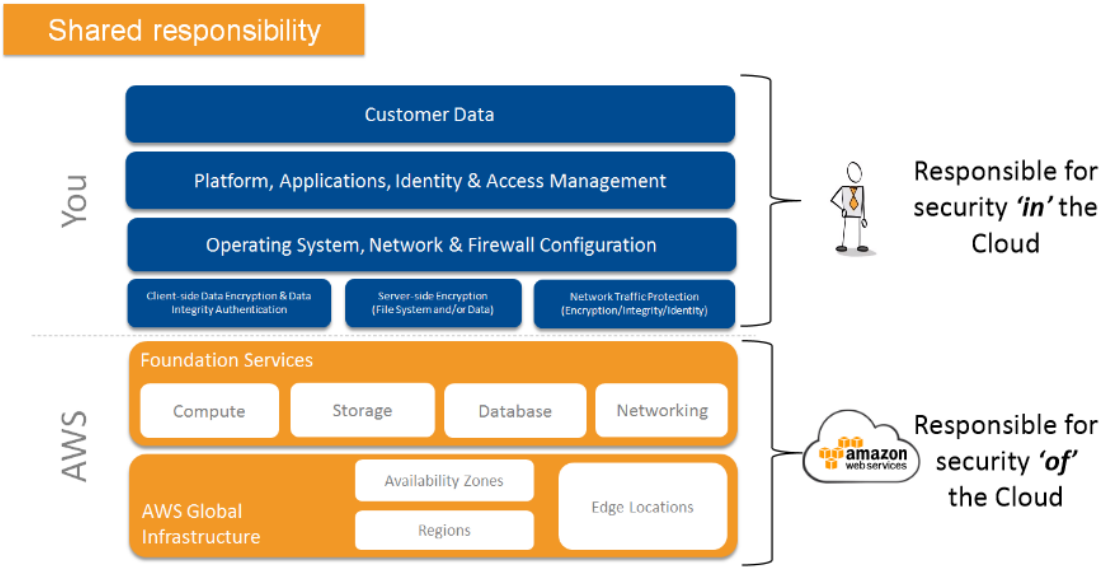

Figure 1: Shared Responsibility Model

This shared responsibility model also provides researchers with flexibility and control to meet industry-specific certification requirements for application deployment. And it is possible for researchers to enhance security and meet stringent compliance requirements by leveraging AWS technology such as host-based firewalls, host-based intrusion detection/prevention, encryption, and key management.

Architecting for Security and Compliance in AWS

The dbGaP security best practices specify that researchers should download data to a secure computer or server, and avoid unsecured network drives or servers.[1] The remainder of the dbGaP security best practices fall within a set of three IT security control domains to meet this principle.

Security Control Domains	Description
Physical Security	Securing physical access to resources, whether located in a data center, a researcher's desk, or through remote administrative access.
Electronic Security	Securing the configuration and use of networks, servers, operating system, and application level resources that hold and analyze dbGaP data.
Data Access Security	Managing user authentication and authorization of access to data, how data copies are tracked and managed, and policies and processes to manage the data lifecycle.

AWS Global Infrastructure

AWS is organized into regions and Availability Zones that that allow for high throughput and low-latency communication between the zones. Researchers with location-specific requirements or regional data privacy policies can establish and maintain their private AWS environment within appropriate locations. In addition, customers can also choose to replicate and backup content in more than one region, as configured by the researcher.

Physical Server Access

AWS's physical servers and network hardware are housed in highly secure, state-of-the-art datacenters. This meets the scope of AWS's independent third party security assessments for ISO 27001, Service Organization Controls 2 (SOC 2), NIST's federal information system security standards, and other security accreditations. Physical access to AWS datacenters and hardware is based on the least privilege principle.

[1] http://www.ncbi.nlm.nih.gov/projects/gap/pdf/dbgap_2b_security_procedures.pdf

Access is only authorized for essential personnel who have experience in cloud computing operating environments, and who are required to maintain the physical environment.

Internet, Networking, and Data Transfers

Although the AWS cloud is inherently a set of web services delivered over the Internet, data within each customer's private AWS account is only exposed to the Internet if the customer specifically configures it. This is a critical element of compliance with dbGaP security best practices and the AWS cloud has a number of built-in features that prevent direct Internet exposure of genomic data.

- **Amazon Elastic Compute Cloud (EC2).** When researchers create new Amazon EC2 instances for downloading and processing genomic data, these instances are only accessible by authorized users within the private AWS account. They are not discoverable or directly accessible on the Internet unless the researcher configures it otherwise.
- **Data Storage.** Genomic data is typically stored in Amazon Simple Storage Service (S3) or Amazon Elastic Block Store (EBS). Alternative storage services include Amazon Relational Database Service (RDS), Amazon Redshift, Amazon DynamoDB, and Amazon ElastiCache. Like Amazon EC2, all of these storage and databases services default to least privilege access and are not discoverable or directly accessible from the Internet unless configured by the customer.
- **Amazon Virtual Private Cloud (VPC).** Researchers can create private, isolated networks within the AWS cloud where they retain complete control over the virtual network environment.

Security Considerations

When designing a hybrid deployment model, researchers also need to consider the following components in the architectural design.

Security Control Domains	Description
Portable Storage Media	Whenever data is downloaded to a portable device such as a laptop or smartphone, the data should be encrypted and hardcopy printouts controlled.
User Accounts, Passwords, and Access Control Lists	Managing user access under dbGaP requirements relies on a principle of least privilege to ensure that individuals and/or processes are granted only the rights and permissions to perform their assigned tasks and functions, but no more.
Data Encryption	Within AWS there are several options for encrypting genomic data, ranging from completely automated AWS encryption solutions (server-side) to manual, client-side options. As researchers architect their system for controlled access datasets, it's important to identify each AWS service and encryption model they will use with the genomic data.

File Systems and Storage Volumes	Within the private AWS account, researchers need to configure storage services and security features to ensure access is limited to authorized users.
Operating Systems and Applications	Researchers are responsible for configuring and maintaining their operating systems and applications in the associated AWS services such as Amazon EC2 and Amazon S3 in accordance with standards such as NIST 800-53, dbGaP Security Best Practices Appendix A or other regionally accepted criteria.
Auditing, Logging, and Monitoring	The dbGaP security recommendations suggest the use of security auditing and intrusion detection software that regularly scans and detects potential data intrusions. Within the AWS ecosystem, researchers have the option to use built in monitoring tools such as Amazon CloudWatch or AWS CloudTrail, as well as a rich partner ecosystem of security and monitoring software specifically built for AWS cloud services. The AWS Partner Network lists a variety of system integrators and software vendors that can help researchers meet security and compliance requirements.
Authorizing Access to Data	Researchers must obtain user access approval from the Data Access Committee (DAC) or within the terms of the researcher's existing Data Use Certification (DUC in order to control access to datasets.
Cleaning Up Data and Retaining Results	In AWS, data deletion and retention operations are under the control of a researcher. Researchers can comply with these dbGaP security recommendations in AWS through a combination of data encryption and other standard operating procedures, such as resource monitoring and security audits.

Customer Examples

Baylor College of Medicine

Baylor College of Medicine in Houston, Texas is home to the Human Genome Sequencing Center (HGSC), one of three federally funded sequencing centers in the US. One of the projects HGSC is involved with is the Cohorts for Heart and Aging Research in Genomic Epidemiology project (CHARGE), a consortium of more than 200 scientists across 5 institutions worldwide who are working to identify genes that contribute to aging and heart disease. Over the last century, a number of studies have followed patients throughout their lives to determine how people develop certain conditions or diseases. With the development of DNA sequencing tools, as well as the ability to manage vast sets of data, the results from these studies are now being re-analyzed as part of the CHARGE project. CHARGE scientists all over the world are making use of data to research the causes and prevention of disease.

With more than 430 TB of data in play on the CHARGE project, Baylor needed a cost-efficient, easily maintainable solution that would enable it to provide safe, effective worldwide collaboration without delays caused by setting up a physical infrastructure.

After moving to AWS, Baylor completed its first analysis in 10 days—five times faster than with the local infrastructure—and was able to share the findings quickly.

AWS scalability helps CHARGE scientists gain more predictive power over the conditions they are studying with unlimited compute and data storage resources. They can also identify "protective" genes that may help shield a person from developing a condition—and they can do so quickly and securely.

Penn State Biological Engineering Department

Penn State's Biological Engineering Department wanted to give biotech researchers an easy way to share research methods and data and run computationally intensive simulations on DNA. By moving its research portal to AWS, Penn State made it easy for 6,000 researchers worldwide to design more than 50,000 synthetic DNA sequences, using Penn State's design methods and optimization algorithms.

- Eliminated need to purchase ongoing equipment, saving both time and money.
- Flexibility to turn compute nodes on and off as needed to meet changing computing requirements.
- Researchers no longer have to wait in line to access design algorithms, with services available when needed.
- Using a web interface, researchers now have access to information.

Claritas Genomics

Claritas is a genetic diagnostic laboratory that has the goal of providing the highest quality testing services for diagnosis of pediatric disorders. Originating as the in-house genetic testing lab at Boston Children's Hospital, Claritas was launched as a stand-alone entity in February 2013.

However, after spinning out of Boston Children's Hospital, Claritas no longer had access to the data centers at Boston Children's Hospital or Harvard Medical School. After evaluating their options, Claritas chose to use the AWS Cloud.

- Avoided a $5M data center expense
- Cost effective, resulting in a 36% decrease in monthly expenses
- Decreased clinical turnaround times from 4-6 months to 4-6 weeks

AWS allowed Claritas to process its systems in a manner that facilitated HIPAA compliance, and all data is encrypted as it moves in and out of the AWS Cloud infrastructure. By using AWS, Claritas is now able to put its development staff first, with an agile infrastructure that easily scales to meet their computing requirements.

Conclusion

The security best practices enable researchers to meet the requirements of the National Institutes of Health, in order to work on controlled-access genomic sequencing data in a secure, scalable, and cost effective environment on Amazon Web Services.

More Information

For additional information, please consult the following sources:

- "Overview of Security Processes" whitepaper.[2]
- AWS Life Science Partner web page[3].
- IAM: IAM documentation[4], IAM Best Practices[5], and Multi-Factor Authentication[6].
- VPC: Amazon VPC whitepaper[7], VPC documentation[8], and VPC Connectivity Options Whitepaper[9]
- Encryption: Securing Data at Rest with Encryption Whitepaper[10], AWS CloudHSM[11].
- A3 Security: Access Control[12], Using Data Encryption[13], Amazon S3 Developer Guide[14].
- AWS Security Overview: Amazon Web Services: Overview of Security Processes[15]
- Amazon EBS security features: Amazon EBS Encryption[16], Amazon Elastic Block Store[17].

---

[2] http://media.amazonwebservices.com/pdf/AWS_Security_Whitepaper.pdf
[3] http://aws.amazon.com/partners/competencies/life-sciences/
[4] http://aws.amazon.com/documentation/iam/
[5] http://docs.aws.amazon.com/IAM/latest/UserGuide/IAMBestPractices.html
[6] http://aws.amazon.com/iam/details/mfa/
[7] https://d36cz9buwru1tt.cloudfront.net/Extend_your_IT_infrastructure_with_Amazon_VPC.pdf
[8] http://aws.amazon.com/documentation/vpc/
[9] https://media.amazonwebservices.com/AWS_Amazon_VPC_Connectivity_Options.pdf
[10] https://media.amazonwebservices.com/AWS_Securing_Data_at_Rest_with_Encryption.pdf
[11] https://aws.amazon.com/cloudhsm/
[12] http://docs.amazonwebservices.com/AmazonS3/latest/dev/UsingAuthAccess.html
[13] http://docs.amazonwebservices.com/AmazonS3/latest/dev/UsingEncryption.html
[14] http://docs.amazonwebservices.com/AmazonS3/latest/dev/
[15] http://awsmedia.s3.amazonaws.com/pdf/AWS_Security_Whitepaper.pdf
[16] http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EBSEncryption.html
[17] http://aws.amazon.com/ebs/

Notices