Securing OHDSI on AWS for HIPAA and Research Data Management Compliance

Michael Lubke[1], Tapati Mazumdar[1], Murat Sincan, M.D.[1,2*], Catherine Hajek, M.D.[1,2*]
[1]Sanford Imagenetics, Sioux Falls, South Dakota, 2 Sanford School of Medicine, University of South Dakota, Sioux Falls, SD

Background

With cloud services and cloud storage adoption becoming more widespread in the healthcare industry, the need for securing these environments and the health data within is of utmost importance. The Observational Health Data Sciences and Informatics

(OHDSI)onAWS project provides an enterprise-level solution for enabling advanced analytics and outcome prediction on this observational health data with the ultimate goal of improving population and individual health. With this project comes the need for developing customized cloud solutions to process various research study data files, and for that reason it is important that every step of the data processing conforms to HIPAA requirements as well as the data management policies outlined in each research study protocol. Enforcing strict security measures in the cloud environment is not only performed to mitigate the financial risks associated with noncompliance, but it is also done to implement additional layers of security with the purpose of preventing the patient health data from being compromised in a data breach.

Methods

The default OHDSIonAWS environment was analyzed, and the following were identified as areas needing additional security:

- Restrict access to OHDSI instance to the trusted network
- Ensure all health data is encrypted at rest and in-transit
- Restrict access to study data to applicable persons and systems outlined in research data management plan

A site-to-site Virtual Private Network (VPN) solution was implemented that creates a tunnel between the on-premises trusted network at Sanford Health and the Virtual Private Cloud (VPC) on AWS. This combined with leveraging the AWS Certificate Manager to handle the generation and application of SSL certificates on the OHDSI on AWS environment enforce HIPAA compliance requirements by ensuring that the communication between all system components within the VPC are encrypted in-transit.

Data access policies were configured for each AWS Simple Storage Service (S3) bucket restricting accessing the data to individual members of each research study. Furthermore, in order to meet HIPAA requirements, all S3 buckets are encrypted at-rest. Policies were created for AWS System Roles that allow the components of the custom Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) Extract, Transform, Load (ETL) pipeline to decrypt the contents of the S3 buckets. The

transformed research data is then inserted directly into the corresponding study schema on Amazon Redshift via a JDBC connection, which ensures that the data is once again encrypted in-transit.

Results

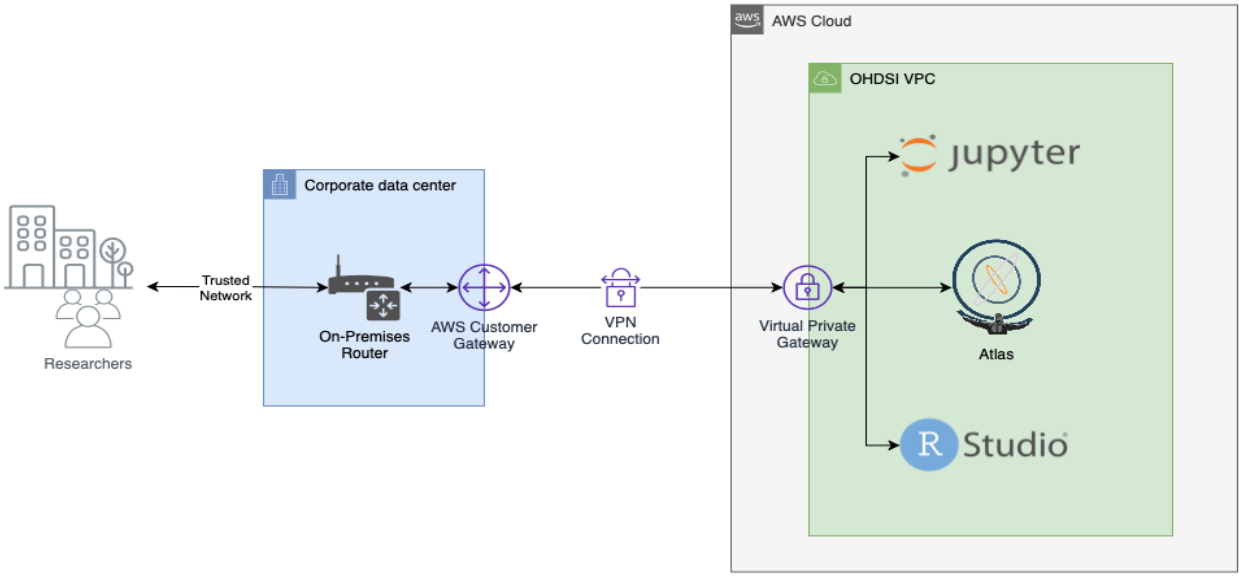

Figure 1. Site-to-Site VPN

Figure 1 illustrates the site-to-site VPN solution that was established between the AWS VPC and a routing device at the Sanford data center. This configuration resulted in the default public-facing applications in the OHDSI on AWS environment stack (Atlas, RStudio, Jupyter) being accessible via the Sanford trusted network only.

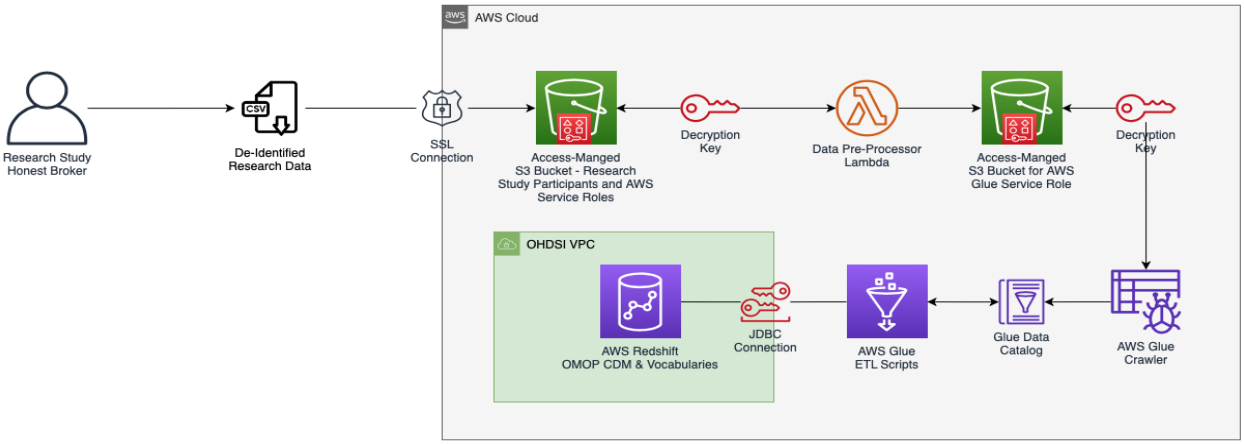

Figure 2. Research Study ETL Pipeline Dataflow Diagram

Figure 2 Research Study ETL Pipeline Dataflow Diagram shows the final state of the ETL pipeline after establishing data access permissions per HIPAA and data management policy requirements. The steps

are summarized as follows:

- The research study honest broker uploads de-identified study data into an access-managed AWS S3 bucket, accessible only to those listed in the study's data management policy.
- New data event triggers the lambda function, which first decrypts the contents of the bucket, then stores the resulting processed data into a subsequent encrypted access-managed S3 bucket.
- AWS Glue Crawler automatically decrypts the processed data files and stores the contents in an AWS Glue Data Catalog

AWS Glue ETL scripts conform the data into the OMOP CDM and insert the records directly into the corresponding study's schema in Amazon Redshift via an encrypted JDBC connection.

Conclusion

It is imperative that every interaction with patient health data hosted in a cloud environment is thoroughly analyzed to ensure that HIPAA and research study compliance requirements are met. Each component added to the cloud environment must be carefully implemented to ensure that the health data is secured regardless of where it exists. This process can be daunting to undertake when an entire data processing pipeline has already been established without initially configured for necessary access controls and encryption practices. Implementing best practices in accordance with HIPAA and data management compliance requirements at the onset of development can provide a safeguard for the cloud environment as well as the data stored within.

References/Citations

1. https://github.com/OHDSI/OHDSIonAWS
2. AWS Site-to-Site User Guide [Internet]. Amazon Web Services; 2011 Sep 29 [updated 2020 Oct 29; cited 2021 Jun 16]. Available from: https://docs.aws.amazon.com/vpn/latest/s2svpn/VPC_VPN.html.
3. Architecting for HIPAA Security and Compliance on Amazon Web Services: AWS Whitepaper [Internet]. Amazon Web Services; 2016 Oct [updated 2020 Oct 29; cited 2021 Jun 16]. Available from: https://d1.awsstatic.com/whitepapers/compliance/AWS_HIPAA_Compliance_Whitepaper.pdf
4. Guidance on HIPAA & Cloud Computing [Internet]. U.S. Department of Health and Human Services [updated 2020 Nov 24; cited 2021 Jun 16]. Available from: https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html