Title

Protecting and Anonymizing PHI/PII data with AWS

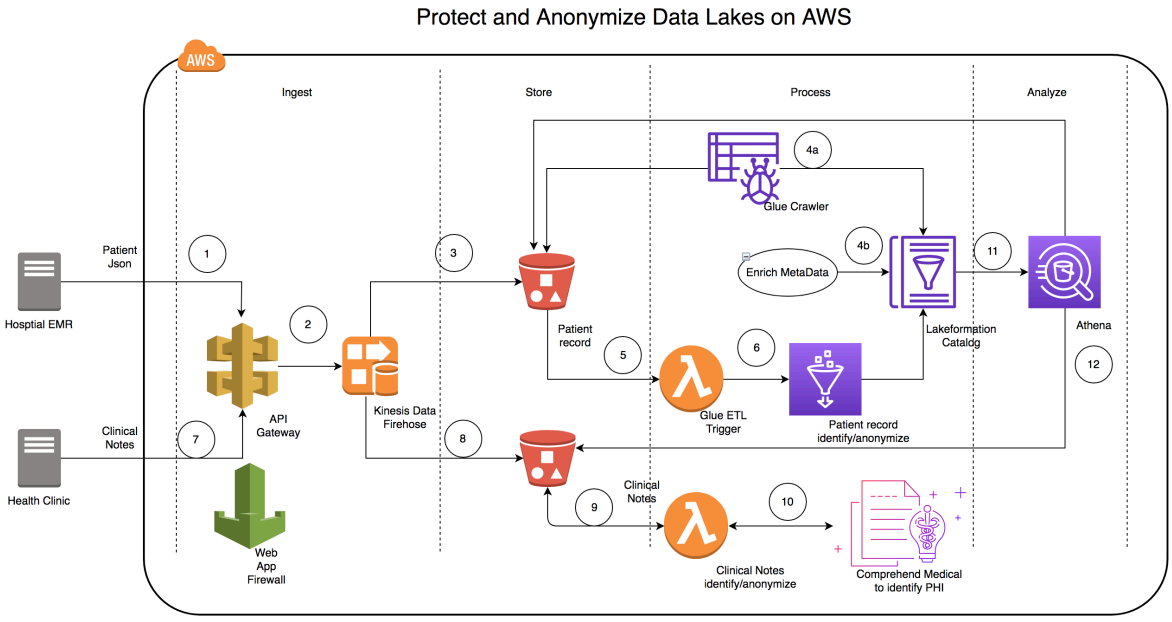
Abstract

Today, many organizations such as healthcare and financial organizations have to exchange and store data. Often times that data may contain Protected Health Information(PHI) or Personally Identifiable Information(PII) that are highly sensitive in nature.

It is job zero for such organization to protect PHI/PII data, prevent data loss and comply with regulations such as HIPAA.
These organizations are looking for new ways to protect their PHI/PII data.

We present an approach on how you can protect and anonymize PHI/PII elements in your structured data such as patient records and your unstructured data such as clinical notes.

Architecture



Use Cases

The above architecture provides a framework for protecting and anonymizing your sensitive data in AWS Data Lake. We will illustrate how the framework operates using the following use cases.

Anonymizing Patient Records

1. Patient records sent from Hospital EMR systems will be received by AWS API Gateway service. The data source will be validated using the integrated Web Application Firewall.
2. If Patient data is received from legitimate sources, API Gateway service will proxy that request to Kinesis Data Firehose Delivery Stream.
3. Firehose Delivery Stream will save that data into a raw folder in a S3 bucket
4. Use Glue Crawler and LakeFormation to enrich the metadata. This step only needs to be executed when there is new or change to existing schema.
    a. A scheduled Glue Crawler will scan the S3 bucket and create the schema in the LakeFormation Data Catalog
    b. Use the LakeFormation service to enrich the schema by adding custom column properties to columns that are sensitive.
5. When the patient dataset is saved in S3 bucket, a Lambda will trigger the Glue ETL job.
6. Glue ETL job will read the patient records from S3 bucket, retrieve the enriched schema from LakeFormation catalog, anonymize the sensitive fields and save the anonymized patient records into a S3 bucket.

Anonymizing Clinical Notes

7. Clinical notes sent from Medical Clinics will be received by AWS API Gateway service. If Clinical Notes is received from legitimate sources, API Gateway service will proxy that request to Kinesis Data Firehose Delivery Stream.
8. Firehose Delivery Stream will save that data into a raw folder in a S3 bucket
9. When the clinical notes is saved in S3 bucket, a Lambda for processing the notes will be triggered
10. Lambda will send notes to Comprehend Medical service that will identify the PHI fields in the notes. Lambda will anonymize those PHI fields and write the anonymize notes to a S3 bucket.

Query and Visualization

11. You can run query on Athena to visualize the anonymized data. Athena integrates with LakeFormation Catalog to discover the schema
12. Athena reads the data from the S3 buckets and displays the results

Business Benefits

1. Eliminate data breaches in your data lakes by proactively anonymizing sensitive data elements.
2. Generate synthetic data for testing in lower environments by anonymizing your production data