Data Security Macie Project Overview

Problem Statement

AWS Global Financial Services customers use Macie to discover Personal Identifiable Information (PII), such as credit card numbers stored in their S3 buckets. Once Macie identifies sensitive data, data engineers need to review the findings. They must identify if the storage of this data was intentional, proper data protections exist, and etc. In order to troubleshoot Macie findings, data engineers may require access to the metadata of the sensitive data to help with their investigation. The challenge is how to provide a secure mechanism for access to the sensitive data.

Solution Summary

The project will provide a solution that takes Macie's findings and generate an event, which will trigger a computer logic that will interrogate the finding to retrieve metadata of the Personal Identifiable Information (PII). Next, the function will store the metadata in a highly secure manner to provide a way for data engineers to mitigate the problem. Then, data engineers can query the relevant metadata from a client.

Solution Detail

Amazon Macie is a fully managed data security and data privacy service that uses Machine Learning and Pattern Matching to discover and protect customer's sensitive data in AWS. However, Macie can not remove or modify the sensible data by itself. Once a customer stores a sensitive data - Personal Identifiable Information (PII) - in Amazon S3 buckets, Macie will find the PII and create a Cloudwatch event that is received by Amazon Eventbridge. Next, Amazon Eventbridge will trigger a lambda function that will parse the event with Macie's finding's details. The script will retrieve the details of the PII to extract the sensible data. Then, the code will encrypt the extracted data with Amazon KMS to store the encrypted metadata into an Amazon DynamoDB table. The original object containing PII in the S3 bucket will be tagged based on the type of PII. The data engineers can now mitigate their potential in sensitive data exposure in S3 without the possibility of data exposure.

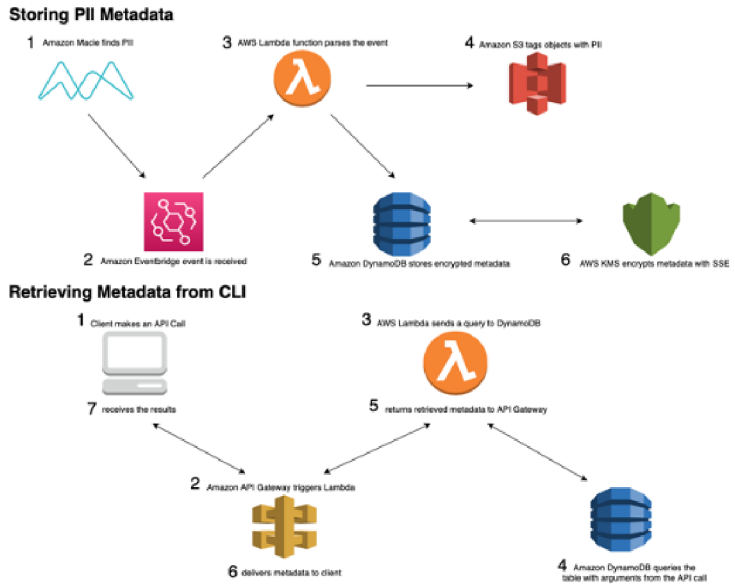
Scenario

MoonTrust is a bank that has been a loyal AWS customer. Amazon Macie found sensitive data (PII) in a S3 bucket that stores MoonTrust's data. The PII is bank account numbers of MoonTrust's customers. Because this PII is not encrypted end-to-end, data protection is critical. Macie generates a CloudWatch event within EventBridge, which triggers a lambda function that parses the event with findings details. During parsing, the code retrieves the name of the object and location of PII data within the S3 bucket, and it accesses the object to extract the bank account numbers. The function also calls AWS KMS (Key Management Service) to encrypt the bank account numbers and stores the encrypted numbers information in a DynamoDB table. The lambda code makes an API call that will query the DynamoDB table to find the sensitive data, to provide a secure way for data engineers to view the PII data via an authenticated web portal. Finally, the lambda function will destroy or manipulate PII in the original S3 bucket.

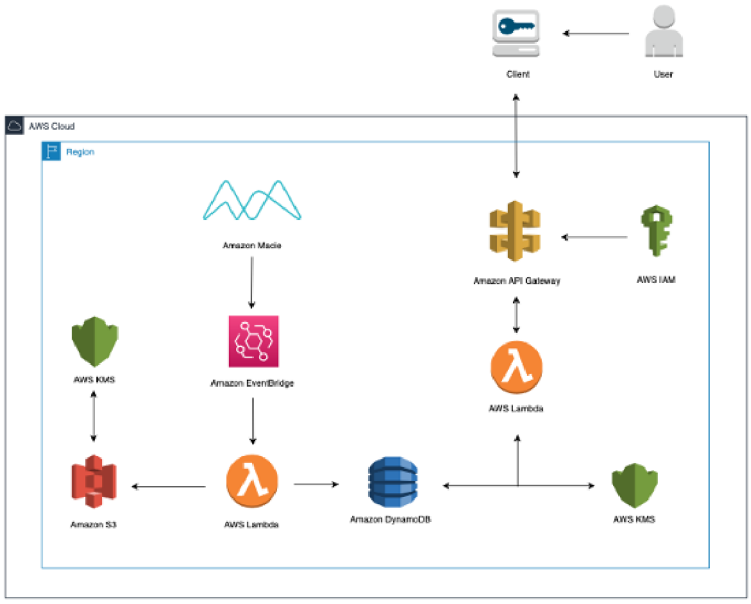
Project Requirements

- All access to PII data must be via secure a API, because human users should not have direct access to the sensitive PII data in S3 buckets.
- access to metadata should be properly authenticated and authorized | must be logged and audited.
- The metadata stored in DynamoDB must be secure with encryption.

Project Workflow



Architecture Diagram



Pseudo Code

Storing Metadata in Amazon DynamoDB

import statements

call Amazon S3 and DynamoDB
parse through Macie's findings
Put the relevant metadata into DynamoDB
Tag PII objects in S3

REST API

import statements
call Amazon DynamoDB
parse through the REST API call
get item from DynamoDB with passed in arguments
Format return statement

Design Considerations / Q&A

Macie

- Provide a 200 level overview of Macie
- Why do we need Macie to identify PII data within S3 objects, what other alternatives can we possibly use, if any?
- How does Macie know if something is PII?
- Can Macie generate a false positive?
- Can Macie miss PII data?

S3

- Provide a 200 level overview of S3
- How's S3 different from EBS storage that is attached to EC2?
- What is the SLA for S3?
- What do we mean when we say that S3 is highly durable?
- Why do you think Macie supports scanning of S3 objects and not for example EBS volumes or databases such as RDS or DynamoDB?

Lambda

- Provide a 200 level overview of Lambda
- How's Lambda different from EC2 instance?
- Why are we using Lambda in this architecture and not EC2 instance or containers?
- What is the pricing model for Lambda?
    - Estimate the cost of Lambda for this POC

KMS

- Provide a 200 level overview of KMS
- How does envelope encryption work?
- What do we mean when we say that S3 is integrated with KMS?
- What is a KMS key policy and how does it work?

DynamoDB

- Provide a 200 level overview of DynamoDB
- How's NoSQL database different from a relational database?

- Why are we using DynamoDB in this architecture instead of Aurora?
- Why did you chose what you chose for the partition key for your table?
- How does DynamoDB server side encryption work?

Glue

- Provide a 200 level overview of Glue
- What is meant by a data catalog?
- What is meant by schema on read?
- What is ETL? Please provide an example.