
AI-Powered Health Data Masking Implementation Guide



AI-Powered Health Data Masking: Implementation Guide

Copyright © Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

Table of Contents

Welcome	1
Overview	2
Cost	2
Architecture	3
Components	4
API Interface	4
AWS AI Services	4
Considerations	5
Regional Deployment	5
Template	6
Deployment	7
Prerequisites	7
What We'll Cover	7
Step 1. Launch the Stack	7
Step 2. Create an IAM Policy to Access the API	8
Grant access to the entire API	8
Grant access to masking functions	9
Create the IAM policy	9
Security	10
Logging	10
Authorization and Authentication	10
Encryption	10
Resources	11
Appendix A: API Description	12
Appendix B: Testing the API	13
Mask Text	7
Mask Image	7
Source Code	15
Revisions	16
.....	16

Identify and mask health data in images or text

Publication date: *August 2019*

This implementation guide discusses architectural considerations and configuration steps for deploying AI-Powered Health Data Masking on the Amazon Web Services (AWS) Cloud. It includes links to an [AWS CloudFormation](#) template that launches and configures the AWS services required to deploy this solution using AWS best practices for security and availability.

The guide is intended for IT infrastructure architects, data scientists, administrators, and DevOps professionals who have practical experience with health data architecting on the AWS Cloud.

Overview

Healthcare organizations generate large amounts of health data such as medical images and patient information and send that data to different applications, including population health management and electronic health records. The challenge medical professionals and developers face is using medical information in applications while meeting their compliance obligations for health data, such as protected health information (PHI).

Currently, there are multiple methods to mask data and each organization has their own approaches based on internal risk assessments. AWS recommends you consult risk assessment specialists for your organization's specific implementation process.

The AI-Powered Health Data Masking solution helps customers identify and mask health data in images or text. This solution uses [Amazon Comprehend Medical](#) to detect health data in a body of text, [Amazon Rekognition](#) to identify text in an image, [Amazon API Gateway](#) and [AWS Lambda](#) to provide an API interface for this functionality, and [AWS Identity and Access Management \(IAM\)](#) to authorize API requests.

This solution was designed to be used as part of a set of mitigating controls in your environment, and does not guarantee alignment to any regulatory framework.

Important

If subject to HIPAA, you must have an AWS Business Associate Addendum (BAA) in place, and follow its configuration requirements, before running protected health information (PHI) workloads on AWS. You should not use your AWS account in connection with PHI until you have accepted the AWS BAA and configured your AWS account(s) as required by the AWS BAA. Under HIPAA regulations, covered entities and business associates are responsible for putting in place a business associate agreement between themselves and each of their business associates. You are solely responsible for determining whether you and your organization need a business associate agreement with AWS. If you determine you need a business associate agreement with AWS, you can [accept the AWS BAA through a self-service portal in AWS Artifact](#). It is your responsibility to obtain a BAA from AWS. For more information about the AWS BAA, please visit the [AWS HIPAA Compliance webpage](#).

This solution does not address state-specific laws that may apply to you. This solution only addresses requirements set forth under HIPAA, a U.S. federal law. Many individual states have adopted rules that are different and, in some cases, stricter than those that are federally mandated under HIPAA.

This solution will not, by itself, make you HIPAA-compliant. The information contained in this solution package is not exhaustive, and must be reviewed, evaluated, assessed, and approved by you in connection with your organization's particular security features, tools, and configurations. However, it is the sole responsibility of you and your organization to determine which HIPAA regulatory requirements are applicable to you, and to ensure that you comply with those applicable requirements. Importantly, most of the requirements under HIPAA are not technical but administrative (that is, people- and process-oriented).

Note that it is your responsibility to ensure the outputs generated by this solution comply with any legal or other requirements applicable to your organization.

Cost

You are responsible for the cost of the AWS services used while running this solution. As of the date of publication, the cost for running this solution with default settings in the US East (N. Virginia) Region is approximately **\$0.015 per text** and **\$0.01 per image** for masking health data. After the model is trained,

the cost to process data from the example dataset is less than **\$0.01 per hour**. Pricing assumes a 1000-character text document and an image with 500 characters of text returned by Amazon Rekognition.

Prices are subject to change and may also be less if you qualify for [AWS Lambda](#) or [Amazon Comprehend Medical](#) free tiers. For full details, see the pricing webpage for each AWS service you will be using in this solution.

Architecture Overview

Deploying this solution builds the following environment in the AWS Cloud.

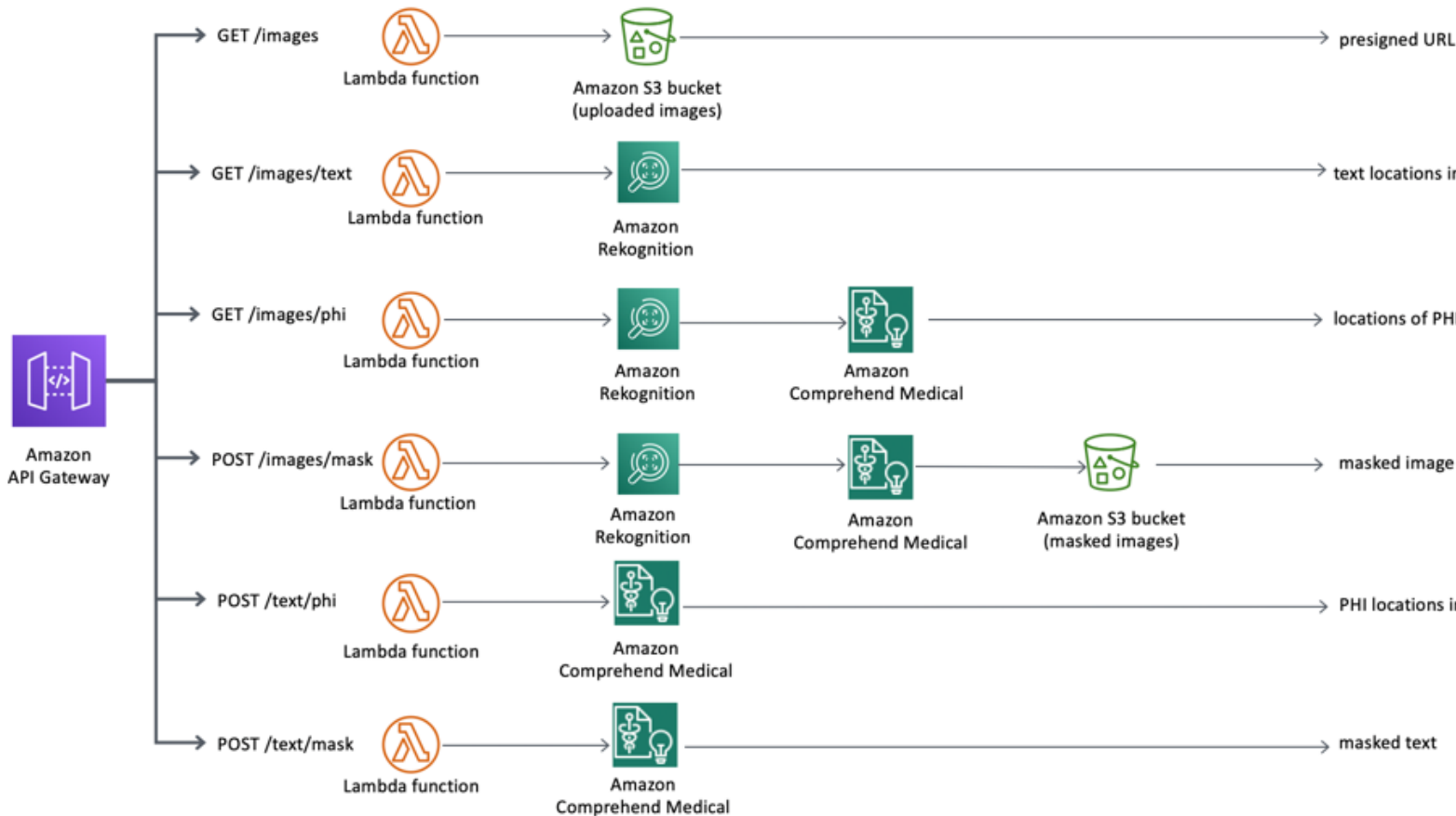


Figure 1: AI-Powered Health Data Masking architecture on AWS

The AWS CloudFormation template deploys an Amazon API Gateway to invoke the microservices (AWS Lambda functions). The microservices provide the business logic to manage preprocessing configuration and logic, and identify and mask health data. The microservices interact with Amazon Rekognition to identify text in an uploaded medical image, and the Amazon Comprehend Medical protected health information data extraction and identification (PHId) API to identify health data in text.

Additionally, the template deploys an [Amazon Simple Storage Service](#) (Amazon S3) bucket for storing raw and masked images, [AWS CloudTrail](#) to log API actions, and [Amazon CloudWatch Logs](#) to log errors within the AWS Lambda functions. By default, log files are encrypted over HTTPS.

For more information on API calls created by the solution, see [Appendix A \(p. 12\)](#).

Solution Components

API Interface

AI-Powered Health Data Masking uses Amazon API Gateway with AWS Lambda proxy to mask health data. By default, access to the API is governed by AWS Identity and Access Management (IAM). However, the solution does not automatically create an IAM policy, role, group, or user to access the API. For information on creating an IAM policy to access the API, see [Step 2 \(p. 8\)](#).

AWS AI Services

AI-Powered Health Data Masking uses Amazon Rekognition to detect text in a medical image and Amazon Comprehend Medical to identify health data in text. Both of these artificial intelligence (AI) services do not require you to use any infrastructure, but are accessed using API calls which can be governed by IAM. Additionally, AWS Lambda accesses these two services using IAM roles scoped to the specific API actions. For more information, see [Authorization and Authentication \(p. 10\)](#).

Considerations

Regional Deployment

AI-Powered Health Data Masking uses Amazon Comprehend Medical and Amazon Rekognition, which are currently available in specific AWS Regions only. Therefore, you must launch this solution in an AWS Region where these services are available. For the most current availability by region, see [AWS service offerings by region](#).

AWS CloudFormation Template

This solution uses AWS CloudFormation to automate the deployment of the AI-Powered Health Data Masking solution on the AWS Cloud. It includes the following AWS CloudFormation template, which you can download before deployment:

[View
Template](#)

ai-powered-health-data-masking.template: Use this template to launch the solution and all associated components. The default configuration deploys Amazon API Gateway, AWS Lambda, AWS Identity and Access Management (IAM), Amazon Simple Storage Service (Amazon S3), AWS CloudTrail, and Amazon CloudWatch, but you can also customize the template based on your specific network needs.

Automated Deployment

Before you launch the automated deployment, please review the architecture, configuration, prerequisites, post-deployment instructions, and other considerations discussed in this guide. Follow the step-by-step instructions in this section to configure and deploy the AI-Powered Health Data Masking solution into your account.

Time to deploy: Approximately two minutes

Prerequisites

Before deploying this solution, verify that you have requested any necessary limit increases for your account. This solution uses Amazon API Gateway, AWS Lambda, Amazon Comprehend Medical, and Amazon Rekognition. For more information about limit increases, see the [Service Limits](#) page for each service.

What We'll Cover

The procedure for deploying this architecture on AWS consists of the following steps. For detailed instructions, follow the links for each step.

[Step 1. Launch the Stack \(p. 7\)](#)

- Launch the AWS CloudFormation template into your AWS account.
- Enter values for required parameter: **Stack Name**

[Step 2. Create an IAM Policy to Access the API \(p. 8\)](#)

- Create or update the IAM policy to access the solution-created API

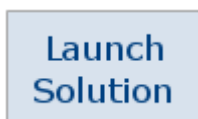
Step 1. Launch the Stack

This automated AWS CloudFormation template deploys AI-Powered Health Data Masking on the AWS Cloud. Verify that you've increased any service limits as needed before launching the stack.

Note

You are responsible for the cost of the AWS services used while running this solution. See the [Cost \(p. 2\)](#) section for more details. For full details, see the pricing webpage for each AWS service you will be using in this solution.

1. Sign in to the AWS Management Console and click the button below to launch the `ai-powered-health-data-masking` AWS CloudFormation template.



You can also [download the template](#) as a starting point for your own implementation.

2. The template is launched in the US East (N. Virginia) Region by default. To launch this solution in a different AWS Region, use the region selector in the console navigation bar.

Note

This solution uses Amazon Comprehend Medical and Amazon Rekognition services, which are currently available in specific AWS Regions only. Therefore, you must launch this solution in an AWS Region where these services are available. For the most current service availability by region, see the [AWS service offerings by region](#).

3. On the **Create stack** verify that the correct template URL shows in the **Amazon S3 URL** text box and choose **Next**.
4. On the **Specify stack details** page, assign a name to your solution stack.
5. Choose **Next**.
6. On the **Configure stack options** page, Choose **Next**.
7. On the **Review** page, review and confirm the settings. Be sure to check the box acknowledging that the template will create AWS Identity and Access Management (IAM) resources.
8. Choose **Create stack** to deploy the stack.

You can view the status of the stack in the AWS CloudFormation console in the **Status** column. You should see a status of **CREATE_COMPLETE** in approximately two minutes.

Step 2. Create an IAM Policy to Access the API

This solution does not automatically create an AWS Identity and Access Management (IAM) policy to invoke the created API. Follow at least one of the two procedures below to implement an IAM policy for access to the API.

Grant access to the entire API

Use the following procedure to grant access to the entire API to mask images and text in the JSON document. Note that using this procedure will allow a user to mask health data and view all information.

1. In the following JSON document, replace *us-east-1* with the AWS Region you are deploying in.
2. Replace *123456789012* with your account ID.
3. Replace *ab12cd3efg* with your API Gateway ID. You can find your ID in the Outputs tab of the AWS CloudFormation stack deployment.
4. Replace *prod* with the name of your staging environment. Note that if you did not change this in the mappings section of the AWS CloudFormation template when deploying, you can leave as is.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "execute-api:Invoke",
        "apigateway:PUT",
        "apigateway:POST",
        "apigateway:GET"
      ],
      "Resource": [
        "arn:aws:execute-api:<us-east-1>:<123456789012>:<ab12cd3efg>/<prod>/*",
        "arn:aws:apigateway:<us-east-1>::/restapis/<ab12cd3efg>/resources/*"
      ]
    }
  ]
}
```

```
}  
]  
}
```

Grant access to masking functions

To grant access to only the functions that mask images and text use the following procedure to modify the JSON document below. Note that using this procedure will allow a user to mask health data without viewing specific information.

1. Replace `us-east-1` with the applicable AWS Region.
2. Replace `123456789012` with your account ID.
3. Replace `ab12cd3efg` with your API Gateway ID. You can find your ID in the Outputs tab of the AWS CloudFormation stack deployment.
4. Replace `prod` with the name of your staging environment. Note that if you did not change this in the mappings section of the AWS CloudFormation template when deploying, you can leave as is.

```
{  
  "Version": "2012-10-17",  
  "Statement": [  
    {  
      "Effect": "Allow",  
      "Action": [  
        "execute-api:Invoke",  
        "apigateway:PUT",  
        "apigateway:POST",  
        "apigateway:GET"  
      ],  
      "Resource": [  
        "arn:aws:execute-api:<us-east-1>:<123456789012>:<ab12cd3efg>/<prod>/POST/  
image/mask",  
        "arn:aws:execute-api:<us-east-1>:<123456789012>:<ab12cd3efg>/<prod>/POST/  
text/mask",  
        "arn:aws:apigateway:<us-east-1>::/restapis/<ab12cd3efg>/resources/*"  
      ]  
    }  
  ]  
}
```

Create the IAM policy

Use the following procedure to create the access policies above.

1. Navigate to [AWS Identity and Access Management console](#).
2. In the navigation pane, select **Policies**. Then, select the **Create policy** button.
3. Navigate to the **JSON** tab.
4. Copy and paste the modified JSON document you modified in the previous section for the access policy you want to create.

See [Appendix B \(p. 13\)](#) for details on how to test the API.

Security

When you build systems on AWS infrastructure, security responsibilities are shared between you and AWS. This shared model can reduce your operational burden as AWS operates, manages, and controls the components from the host operating system and virtualization layer down to the physical security of the facilities in which the services operate. For more information about security on AWS, visit the [AWS Security Center](#).

Logging

AI-Powered for Health Data Masking uses Amazon CloudWatch to capture Amazon API Gateway actions in your environment and AWS CloudTrail to log information from the AWS Lambda functions and Amazon API Gateway. All AWS CloudTrail logs are archived to the solution-created Amazon S3 log bucket.

Authorization and Authentication

In concordance with the principles of least privilege and separation of concerns, each AWS Lambda function operates with a separate AWS Identity and Access Management (IAM) role and policy. For example, only Lambda functions that use Amazon Rekognition for detecting text in an image are authorized to make the `DetectText` API call to Amazon Rekognition.

Amazon API Gateway uses IAM to control access for invoking the deployed API. This guide provides an example IAM policy that you can create after you have deployed the solution. For more information about creating an IAM policy, see [Step 2 \(p. 8\)](#).

Encryption

All internal and external communications for AI-Powered Health Data Masking are over HTTPS. For example, Amazon API Gateway only accepts communication over HTTPS and not HTTP, and all AWS API calls made by AWS Lambda are over HTTPS. Additionally, the solution-created Amazon S3 buckets have default server-side encryption enabled, and automatically encrypt any object uploaded into an S3 bucket.

Additional Resources

AWS services

- [AWS CloudFormation](#)
- [Amazon API Gateway](#)
- [AWS Lambda](#)
- [Amazon Comprehend Medical](#)
- [Amazon Rekognition](#)
- [Amazon CloudWatch Logs](#)
- [AWS CloudTrail](#)
- [Amazon S3](#)
- [AWS Identity and Access Management](#)

Appendix A: API Description

The AI Powered Health Data Masking solution creates the following API calls:

- **Get Image (GET):** Returns a pre-signed URL for the image based on the Amazon Simple Storage Service (Amazon S3) bucket and key. To return a successful link, the query parameters must only include the bucket and key, and the AWS Identity and Access Management (IAM) role associated with the AWS Lambda function must have the appropriate permissions to generate the pre-signed URL.
- **Get Image Text (GET):** Returns the location of all text in an image. This is a wrapper for the `DetectText` feature in Amazon Rekognition. To return a successful response, the query parameters must only include the bucket and key, and the IAM role associated with the AWS Lambda function must have the appropriate permissions to make calls using Amazon Rekognition.
- **Get Image PHI (GET):** Returns the location of potential protected health information (PHI) in a medical image. This is a wrapper for the `DetectText` feature in Amazon Rekognition, and is sent to the `DetectPHI` API in Amazon Comprehend Medical. To return a successful response, the query parameters must only include the bucket and key and an optional PHI threshold parameter, and the IAM role associated with the Lambda function must have the appropriate permissions to make calls with Amazon Rekognition and Amazon Comprehend Medical.
- **Mask PHI in Image (POST):** This function contains the same functionality as **Get Image PHI** but includes the feature for generating a new masked image based on the PHI identified using Amazon Comprehend Medical and the locations of the text using Amazon Rekognition. New images are written to Amazon S3 and the location of the images is returned.
- **Get PHI in Text (POST):** Returns the location of the PHI within a body of text. Only allowed parameters are the body of text and an optional parameter of PHI detection threshold. This is a wrapper for the `DetectPHI` call for Amazon Comprehend Medical.
- **Mask PHI in Text (POST):** Masks the PHI in a body of text identified by Amazon Comprehend Medical, and replaces the identified entity with the identified attribute. For example, NAME.

Appendix B: Testing the API

The AI-Powered Health Data Masking API receives images and text as input. Before testing the API, you must complete [Step 2: Create an IAM Policy to Access the API \(p. 8\)](#) and attach the policy to an IAM role.

Use the following procedures to use the API image and text masking capabilities:

Note

Verify that the IAM role you assume has an attached IAM policy that restricts the image and text masking functions.

Mask Text

Use the following procedure to test the solution API text masking capability:

1. In the [AWS CloudFormation console](#), select the deployed solution stack.
2. In the **Outputs** tab, copy the **ApiGatewayId** and the **TextMaskResourceId**.

Using a portion of a fictional medical note provided by the Amazon Comprehend Medical team, you can test the solution API for masking text using AWS CLI or your AWS SDK of choice.

```
PERSON INFORMATION
Name: SALAZAR, CARLOS
MRN: RQ36114734
ED Arrival Time: 11/12/2011 18:15

Sex: Male
DOB: 2/11/1961
```

The example POST request using the example note above, will de-identify the request message and look like the following: `{"text": "PERSON INFORMATION\nName: SALAZAR, CARLOS\nMRN: RQ36114734\nED Arrival Time: 11/12/2011 18:15\nSex: Male\nDOB: 2/11/1961"}`

3. Open a Python terminal and paste in the following code sample. You must update the **api_id** and **resource_id** you copied from the **Outputs** tab above.

```
import boto3
import json

client = boto3.client('apigateway')
api_id = YOUR_API_ID
resource_id = YOUR_RESOURCE_ID
payload = {"text": "PERSON INFORMATION\nName: SALAZAR, CARLOS\nMRN: RQ36114734\nED
Arrival Time: 11/12/2011 18:15\nSex: Male\nDOB: 2/11/1961"}

response = client.test_invoke_method(
    restApiId=api_id,
    resourceId=resource_id,
    httpMethod='POST',
    headers={"Content-Type": "application/json"},
    body=json.dumps(payload)
)

print(response['body'])
```


If the masking was completed successfully, you will see the following message:

```
{"maskedText": "PERSON INFORMATION\nName: NAME\nMRN: ID\nADDRESS Arrival  
Time: DATE 18:15\nSex: Male\nDOB: DATE"}
```

Mask Image

Use the following procedure to test the solution API image masking capability.

1. Download the example [chest x-ray image](#) from a dataset made available by the NIH Clinical Center, and upload the image into the created Amazon S3 bucket.
2. Copy the Amazon S3 bucket name and key.
3. In the **Outputs** tab, copy the **ApiGatewayId** and the **ImageMaskResourceId**.
4. Open a Python terminal and paste in the following code sample. Note that you must update the **api_id** and **resource_id** you copied from the **Outputs** tab above, and the S3 bucket **name** and **key**.

```
import boto3
import json

client = boto3.client('apigateway')
api_id = YOUR_API_ID
resource_id = YOUR_RESOURCE_ID
s3_bucket = YOUR_S3_IMAGE_BUCKET
s3_key = YOUR_S3_IMAGE_KEY
destination_key = 'masked/' + s3_key
payload = {"s3Bucket": s3_bucket, "s3Key": s3_key, "destinationBucket": s3_bucket,
          "destinationKey": destination_key}

response = client.test_invoke_method(
    restApiId=api_id,
    resourceId=resource_id,
    httpMethod='POST',
    headers={"Content-Type": "application/json"},
    body=json.dumps(payload)
)

print(response['body'])
```

5. Once the image is successfully masked, navigate to the response path to view the image. The path will be returned in the JSON document (`response['body']`).

Source Code

Visit our [GitHub repository](#) to download the templates and scripts for this solution, and to share your customizations with others.

Document Revisions

Date	Change	
August 2019	Initial release	

Notices

Customers are responsible for making their own independent assessment of the information in this document. This document: (a) is for informational purposes only, (b) represents current AWS product offerings and practices, which are subject to change without notice, and (c) does not create any commitments or assurances from AWS and its affiliates, suppliers or licensors. AWS products or services are provided "as is" without warranties, representations, or conditions of any kind, whether express or implied. The responsibilities and liabilities of AWS to its customers are controlled by AWS agreements, and this document is not part of, nor does it modify, any agreement between AWS and its customers.

The AI-Powered Health Data Masking solution is licensed under the MIT No Attribution at <https://spdx.org/licenses/MIT-0.html>