aws professional services

Secure Shield

PII Redaction Pipeline

Miniar Jabri (she/her)

Professional Services Consultant Intern @AWS

08/06/2024

Agenda

1. About me!

2. Context

- Customer
- Customer’s challenge
- Our solution

3. Dive Deep

- Technology Stack
- Architecture
- Frontend
- Step function workflow:
  - Text Extraction
  - PII/PHI Detection
  - Redaction
- Monitoring and Observability
- Security
- Limitations
- Next steps

4. Lessons Learned

About Me !

About me

- I’m Miniar, (Meen - yaar) a first time Amazonian
- School: 2nd year engineering student at UPSSITECH in Toulouse.
- I graduate in august 2026

During my AWSome Time here….

- I passed the SAA Certificate
- I passed the AI practionier Certificate
- I completed a a Gen-AI course on Machine Learning University
- Participated to an instructive firehose on security
- Won the second place in a JAM
- I joined @women_in_amazon Group

Our Customers

HealthTech Analytics

A HEALTHCARE DATA ANALYTICS COMPAGNY

Customer Processes large volumes of patient records and healthcare data to deliver insights for improving patient care and operational efficiency.

Goal

Protect Sensitive information while maintaining data utility for analysis

WHY IT MATTERS for AWS (Global Context)

The global Healthcare & Life spends $23.6 billion annually on data analytics, with expected CAGR of 21.4% until 2030.
=> This highlights the significant market opportunity for building scalable solution

Source : *https://w.amazon.com/bin/view/APNServiceAcceleration/AnalyticsHome/Programs/ASAP/PLHCLSSolutions/*

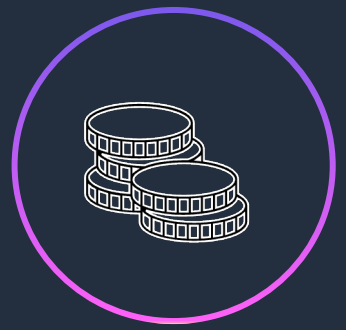
Data Analytics **Healthcare** **HIPPAA Compliance** **Protect PII/PHI**

The Problem / Challenge

Current Challenge

Time-consuming: 30-45 minutes per document

Error prone: Human error rate: up to 7%



Expensive: Dedicated staff for manual review
-Labor cost: $25-35/Hour
- Monthly cost (1,000 documents): $18,000

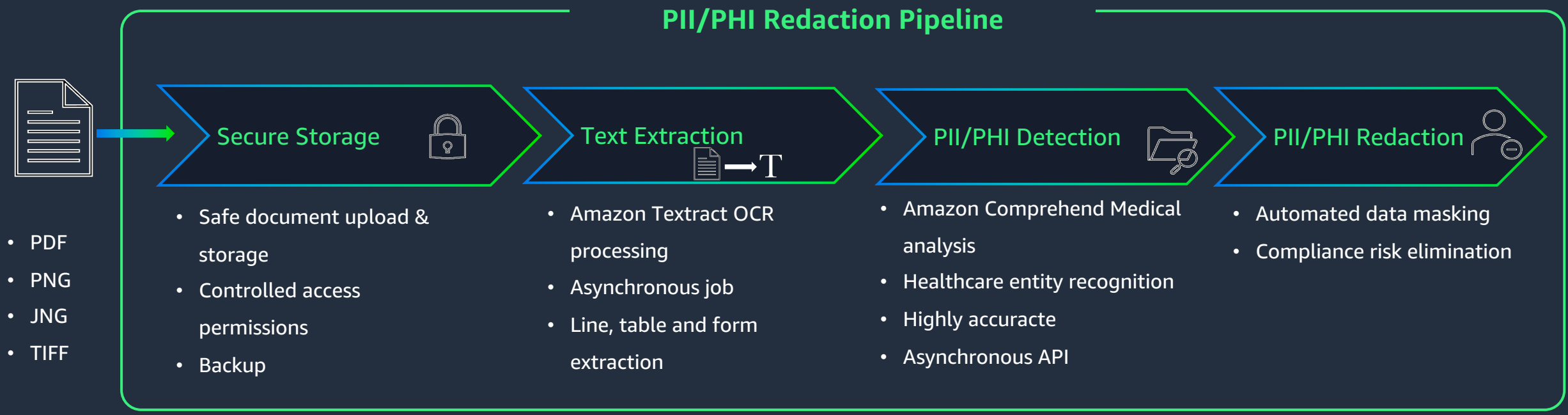
Manual Process ➡️ Processing Delays ➡️ Delayed Analytics ➡️ Slower Decision Making ➡️ Revenue Loss

Faster document processing shortens decision cycles; this enables you to serve customer and have people do higher value task

Our Solution

Our Solution - High-level Workflow

How it works:



CONTEXT

PII Redaction Data Pipeline. – Business Value

AWS makes Redaction better, faster, and more cost-effective.

Simple Operation

- Quick setup with IaC
- Pre-packaged solution that accelerates deployment and saves time

Enables **One-click deployment.**

Cost-Effective

- Cuts costs by reducing manual labor and errors.
- Pay-per-use AWS Services

Reduces costs by **70%** compared to manual redaction, **$0.04/page** and **$450/month** (1000 documents)

Batch Processing

- Processes thousands of documents per minute.
- Optimized for businesses of any size.

Process **up to 400 Documents** simultaneously with auto-scaling and takes up to **8 minutes.**

Intuitive User-Interface

- Intuitive dashboard for document upload
- Real-time workflow status tracking.

Streamline User Experience with secure document handling."

“ Privacy is not negotiable. In healthcare, it's the foundation of the doctor-patient relationship and the cornerstone of quality care. “

Kathleen Sebelius

Former U.S. Secretary of Health and Human Services

Dive Deep...

Tech Stack

Frontend:

- React

Backend:

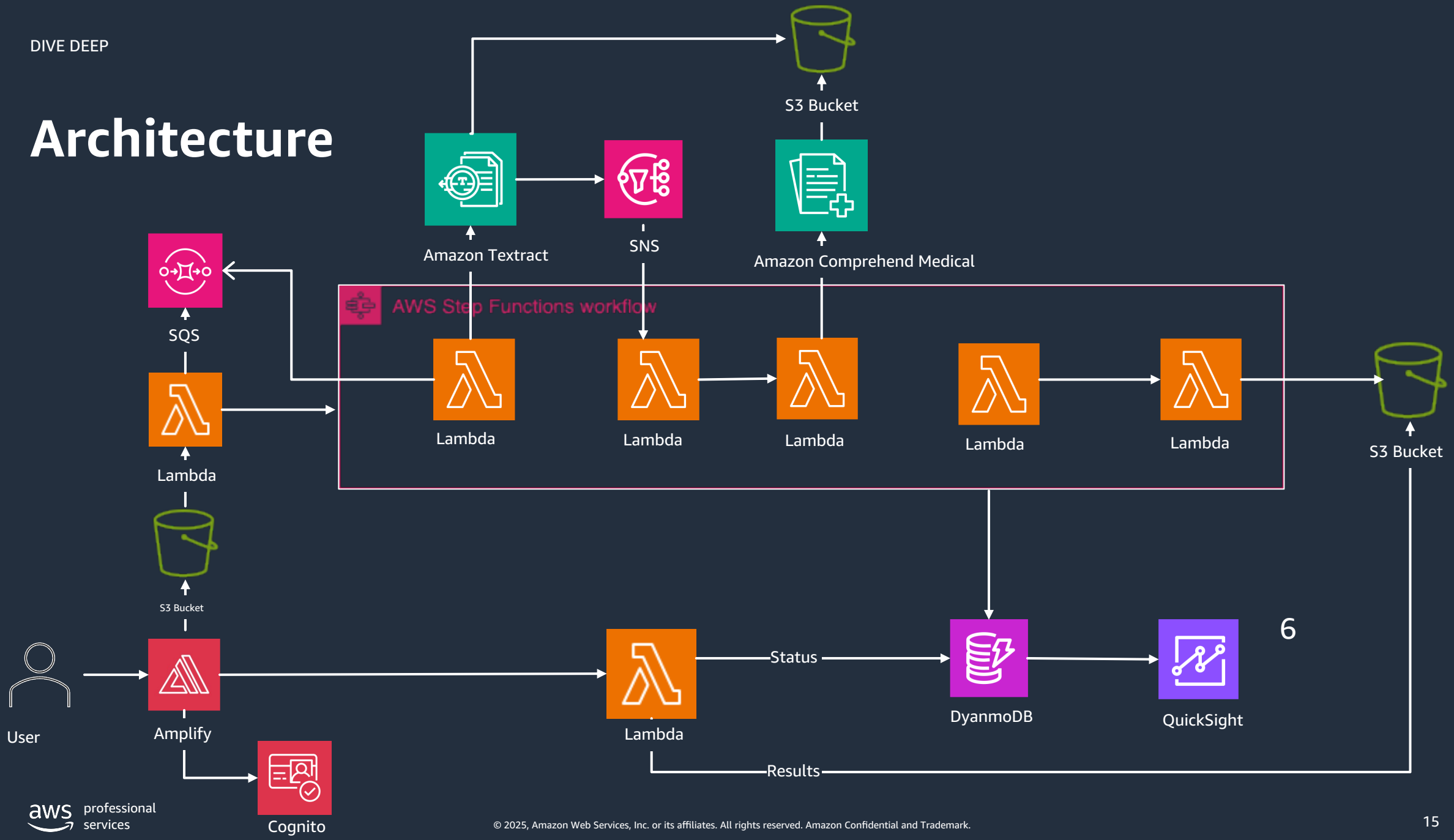
- AWS CDK (IaC)
- Javascript (cdk)
- Python (Lambda)

Version Control:

- GitLab:

https://gitlab.aws.dev/aws-proserve-intern-program/2025-projects/emea/miniarja-2025-intern-PII-Redaaction-Pipeline/-/tree/my-project?ref_type=heads

Architecture

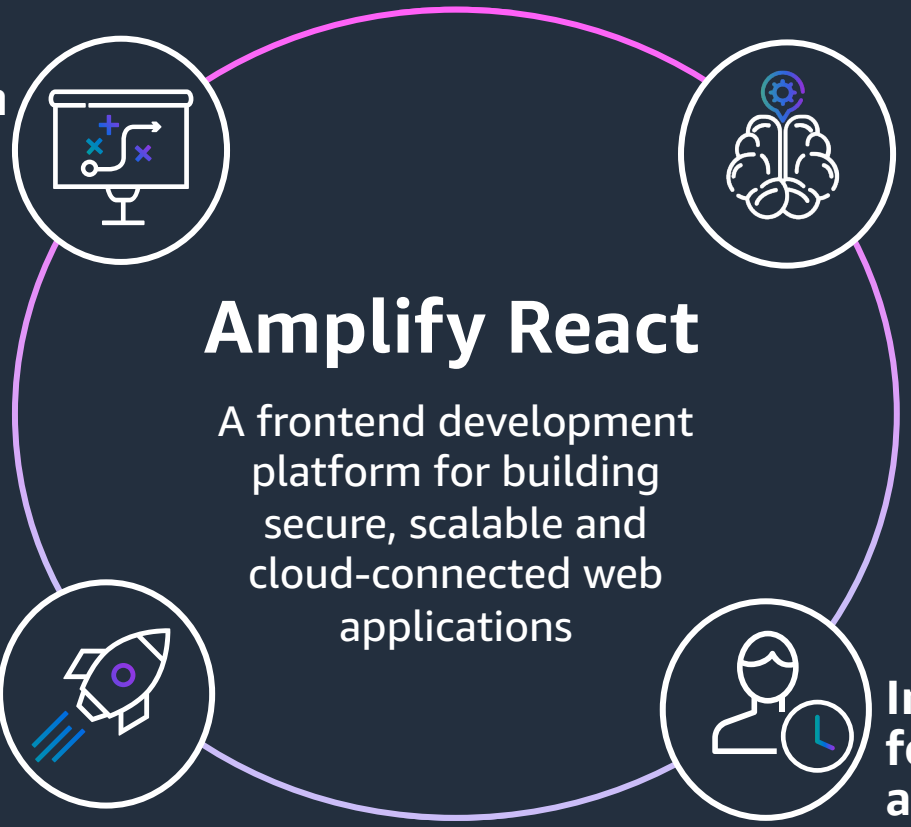


Frontend – User Interface

User Interface

Seamless connection to AWS services and backend

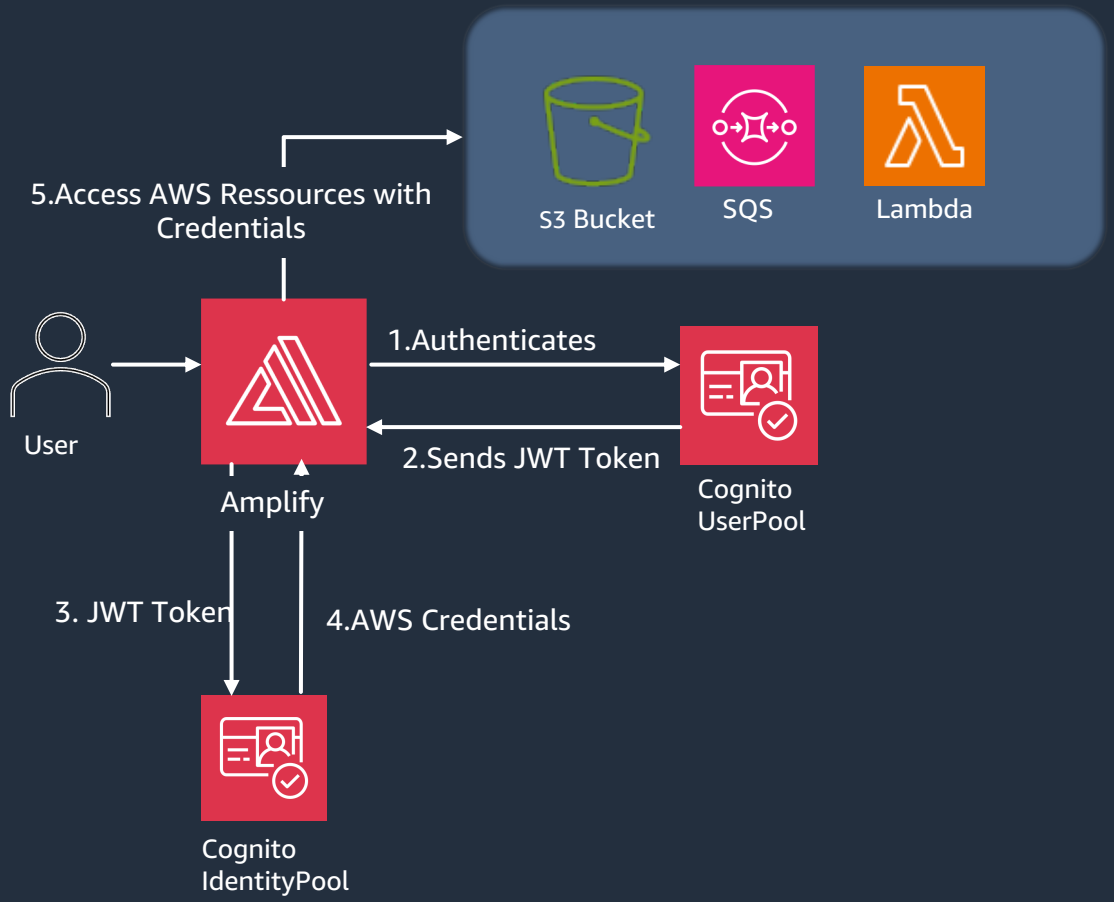
Developer-friendly with React-based tools



Rapid deployment with pre-built UI components

Integrated with Cognito for secure authentication

Authentication and Authorization



Amazon Cognito

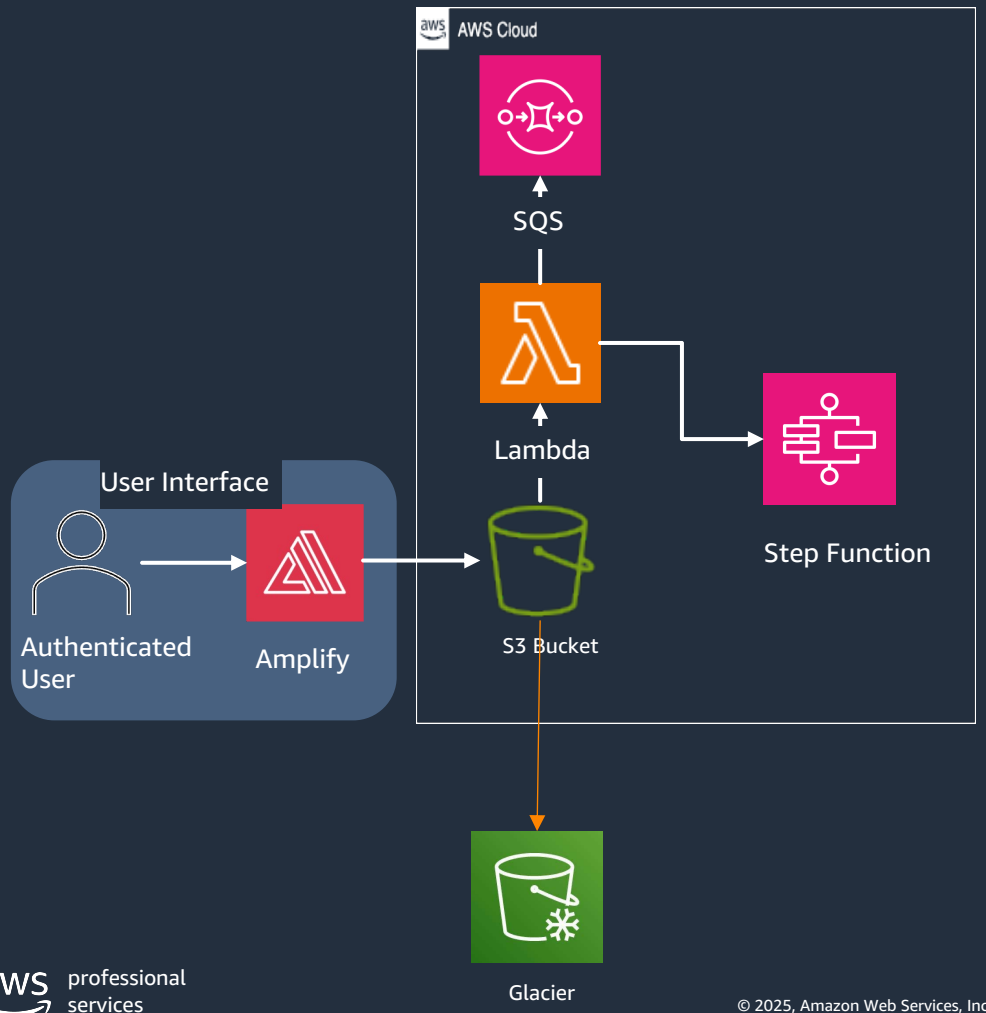
A fully managed AWS service that provides authentication, authorization, and user management for web and mobile applications.

- **User Pools:** handles authentication (WHO you are)
- **Identity Pools :** handles the authorization (WHAT you can access)

Keys features:

- Zero credential storage in frontend.
- Temporary credentials
- No sign up, only sign in
- Role-based access control with least privilege principle:
  - Admin Group: can access the PII
  - Customer Group: can’t access PII

File Upload



Keys features:

- Up to 400 documents at once
- Supported file types: .pdf, .png, .jpeg, .tiff
- 500 MB/ file
- Secure pre-signed URLs with 2-hour expiration for enhanced security
- S3 versioning
- Automatic transition to Glacier storage after 90 days
- SQS messages contain:
  - Workflow ID
  - Document name
  - Input path

Live Demo...

Deliver Results

Step Function workflow:

- Text Extraction
- PII/PHI Detection
- Redaction

1. Text Extraction – more than just OCR...

Structured data extraction
Extracts txt, forms, tables, lines.
From: pdf, JPEG,PNG

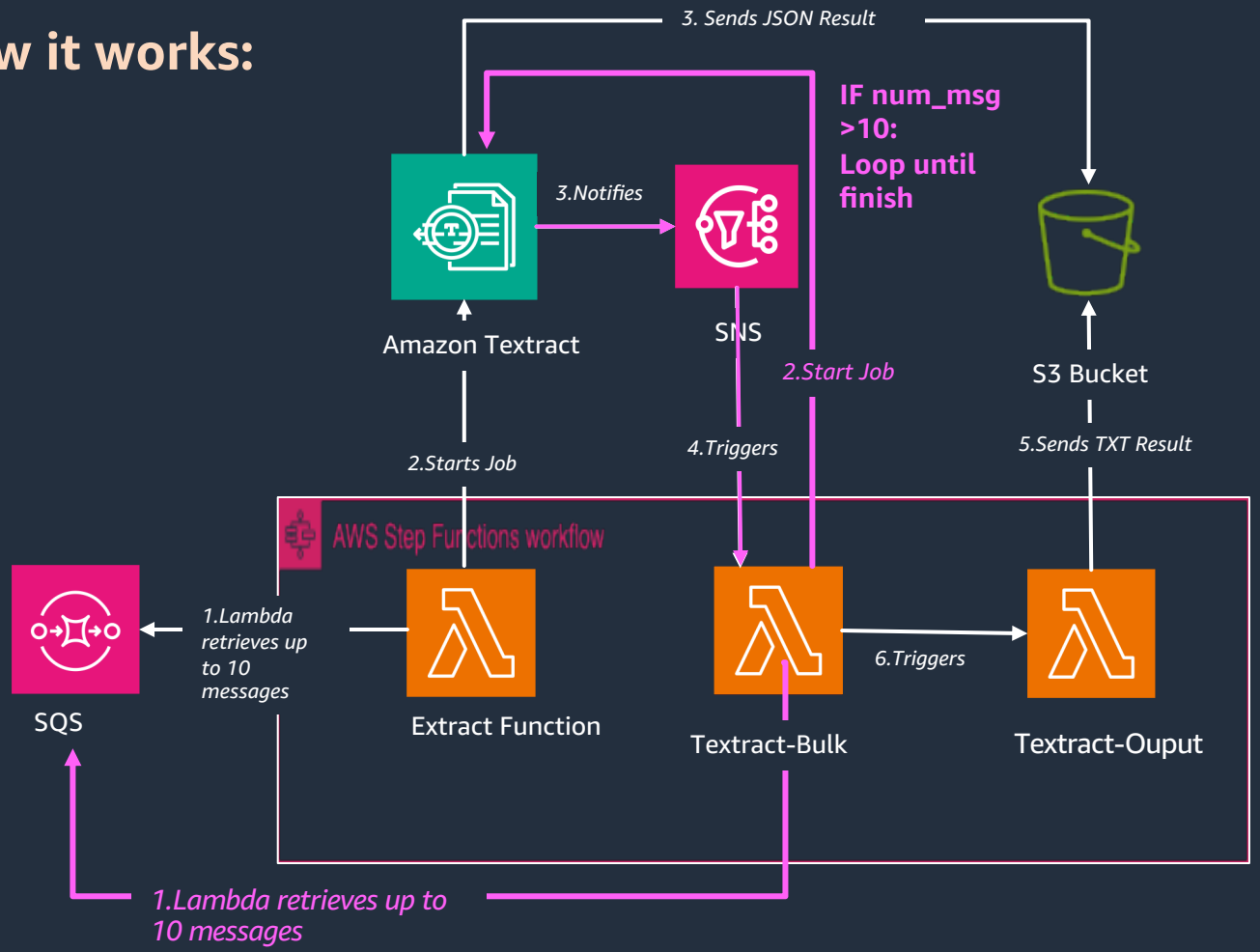
Asynchronous processing

- Highly accurate
- High confidence score

No ML training required

1. Text Extraction

How it works:



Text extraction Workflow:

- Max SQS receive_message=10 (Prevents overwhelming the system)
- Lambda retrieves up to 10 messages from SQS :
- Using SNS Eliminates resource-intensive polling

Textract-bulk: Orchestrates the workflow

  - Triggers Textract-Output
  - If num_msg >10: Loop until all documents are processed

- **Textract-output:** Transforms raw Textract JSON to TXT

Code Snippets..

Dive deep

Code Snippets... Dive deep

```
txrct_response = textract.start_document_analysis(
                    DocumentLocation={
                        'S3Object': {
                            'Bucket': env_vars['INPUT_BKT'],
                            'Name': f"public/{doc['input_path']}{doc['document_name']}",
                        }
                    },
                    FeatureTypes=['TABLES','FORMS'],
                    JobTag=doc['workflow_id'],
                    NotificationChannel={
                        'SNSTopicArn': env_vars['SNS_TOPIC'],
                        'RoleArn': env_vars['SNS_ROLE']
                    },
                    OutputConfig={
                        'S3Bucket': env_vars['INPUT_BKT'],
                        'S3Prefix': f"public/output/{doc['workflow_id']}"
                    }
                )
```

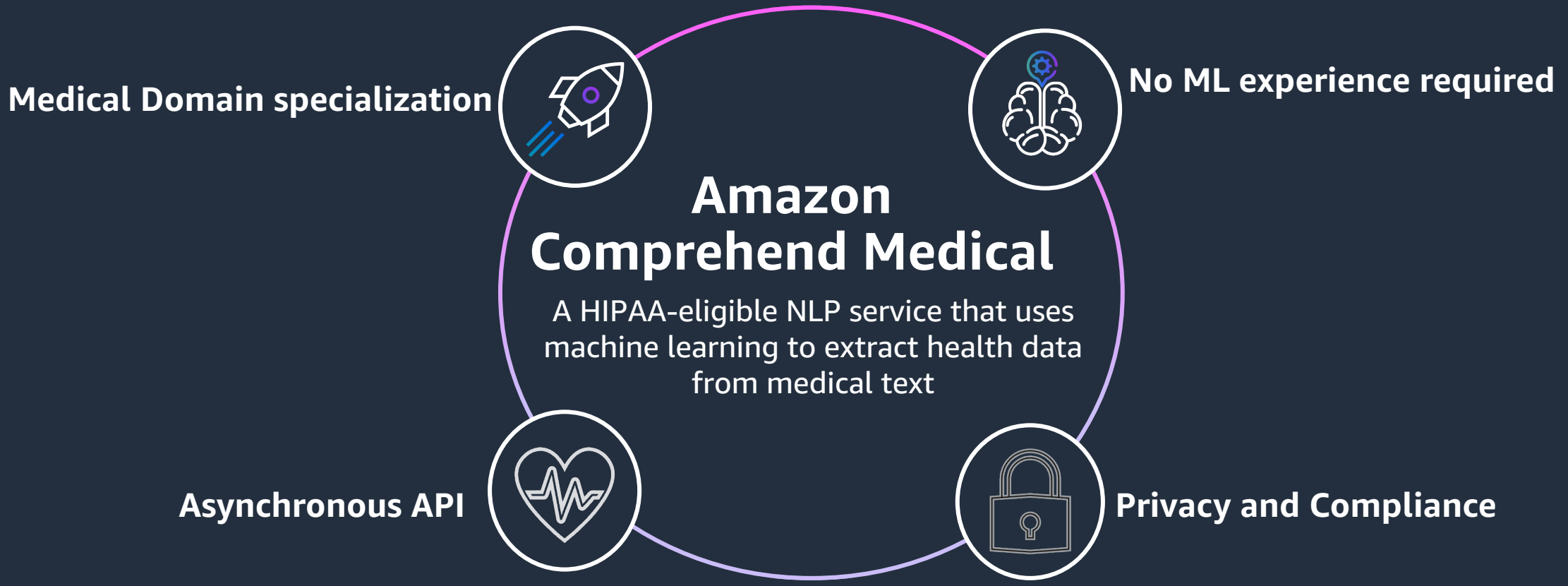
Initiating Textract job

Code Snippets... Dive deep

```
# Some documents remain to be submitted to Textract
if len(processed_files) < len(files):      You, 3 weeks ago • stacks and ressources
    # Documents remained to be processed or are being processed
    event["bucket"] = bucket
    jobs = get_msg_submit(event, env_vars, 10)
    logger.debug(f"Submitted Jobs : {json.dumps(jobs)}")
    return jobs
else:
    #Post to state machine that workflow is done
    select = f"SELECT \"workflow_token\" FROM \"{env_vars['PII_TABLE']}\" WHERE part_key=? AND sort_key=?"
    ddb_response = ddb.execute_statement(Statement=select, Parameters=[
                                            {'S': workflow_id},
                                            {'S': f"input/{workflow_id}/"}
                                        ])
    deserialized_document = {k: deserializer.deserialize(v) for k, v in ddb_response['Items'][0].items()}
    sm_token = deserialized_document['workflow_token']
    smresponse = sfn.send_task_success(taskToken=sm_token,
                                       output=json.dumps({
                                            "Payload": {
                                                "workflow_id": workflow_id,
                                                "bucket": bucket,
                                                "tmp_process_dir": f"{root_prefix}/temp/{workflow_id}",
                                                "phi_input_dir": f"{root_prefix}/phi-input/{workflow_id}"
                                            }
                                        }))
    logger.debug(smresponse)
```

Loop

2. PII/PHI Detection



2. PII Detection – Amazon Comprehend Medical

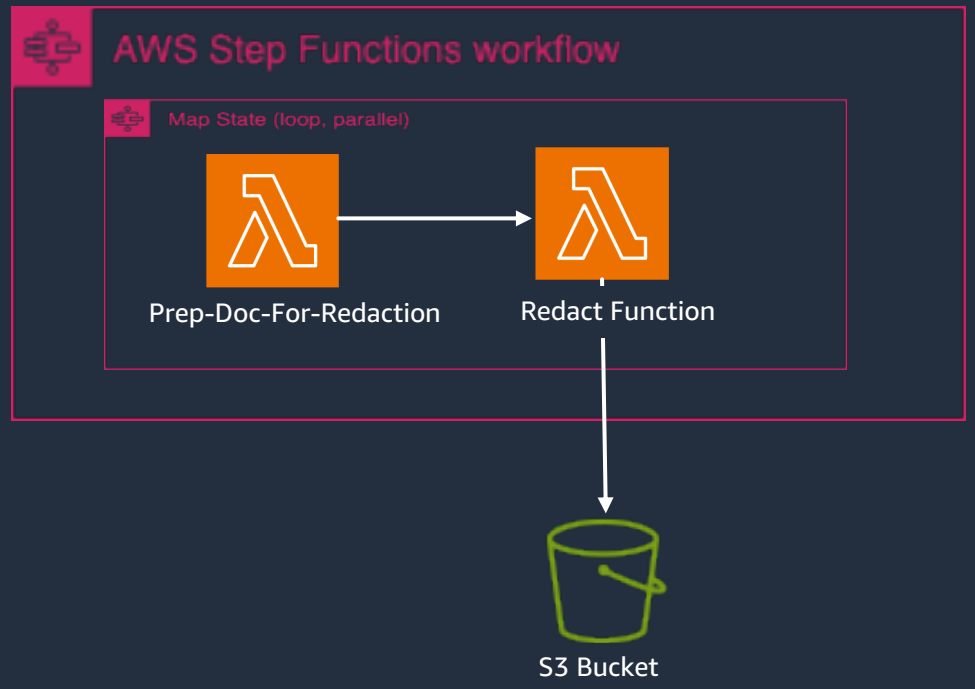
How it works:

Key Features:

- Automated Detection & Redaction
- asynchronous API
- Parallel processing of multiple documents
- HIPAA-compliant

3. PII Redaction - Lambda

How it works:



Map State:
- Maximum scale: 40 parallel executions × 10 documents each = 400 documents simultaneously

1.Prep-Doc-For-Redaction :
maps between:
- Original document location
- Textract output
- Amazon Comprehend Medical Output

2.Redaction: :
- Visual Redaction Application:
  - Use Textract bounding boxes
  - Draws black rectangles precisely over PHI entities (detected by comprehend medical) on origal files

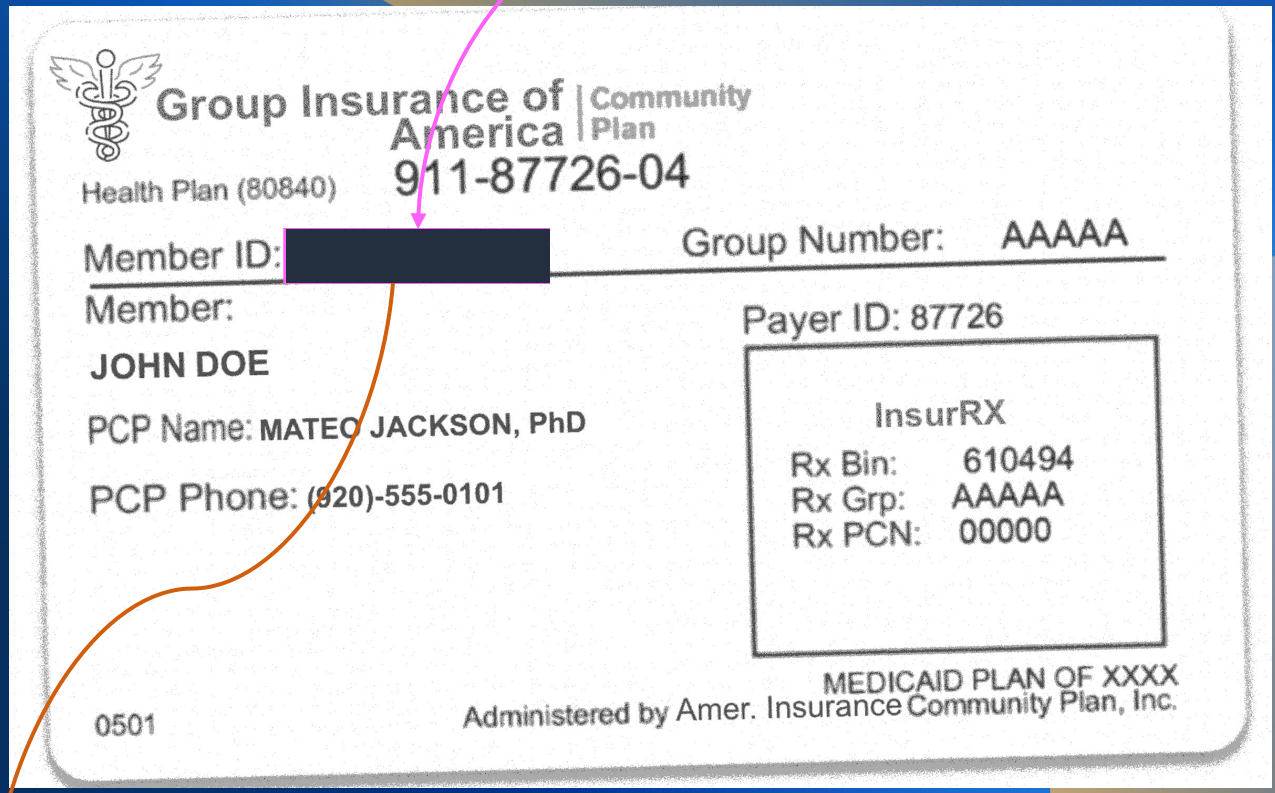
- Map state updates workflow status in DynamoDB

Redaction:

Redact.py lambda function

Amazon Textract

Bounding box



PII

Amazon Comprehend Medical

Code Snippets..

Dive deep

Code Snippets... Dive deep

1. Getting the bounding boxes for all text lines in the document

```
# Set up overlay for text lines
logger.debug("Setting overlay")
overlay=[Textract_Types.LINE]

# Get bounding boxes for text from Textract JSON
logger.debug("Getting bounding boxes")
bounding_box_list = get_bounding_boxes(textract_json=textract_json, document_dimensions=document_dimension, overlay_features=overlay)
```

2. Getting the PHI entities from Comprehend Medical

```
for entity in comprehend_json['Entities']:
    entity_text = entity['Text']
    entities.append(entity_text)
    # Add lowercase entity to lookup dictionary
    entity_lookup[entity_text.lower()] = True
```

Code Snippets... Dive deep

```
for idx, bbox in enumerate(bounding_box_list):
    if idx in processed_boxes:
        continue
    bbox_text_lower = bbox.text.lower()

    for entity in entities:
        entity_lower = entity.lower()
        if entity_lower in bbox_text_lower or bbox_text_lower in entity_lower:
            redactions.append(bbox)
            processed_boxes.add(idx)
            break
```

For every bounding box

For every PHI

matching text entities from Comprehend Medical with their exact physical locations from Textract.

Code Snippets... Dive deep

```
# Draw black rectangles over bounding boxes that contain PHI entities
for idx,img in enumerate(images):
    draw = ImageDraw.Draw(img)
    page_num = idx + 1
    for box in redactions:
        if box.page_number == page_num:
            draw.rectangle(xy=[box.xmin, box.ymin, box.xmax, box.ymax], fill="Black")
```

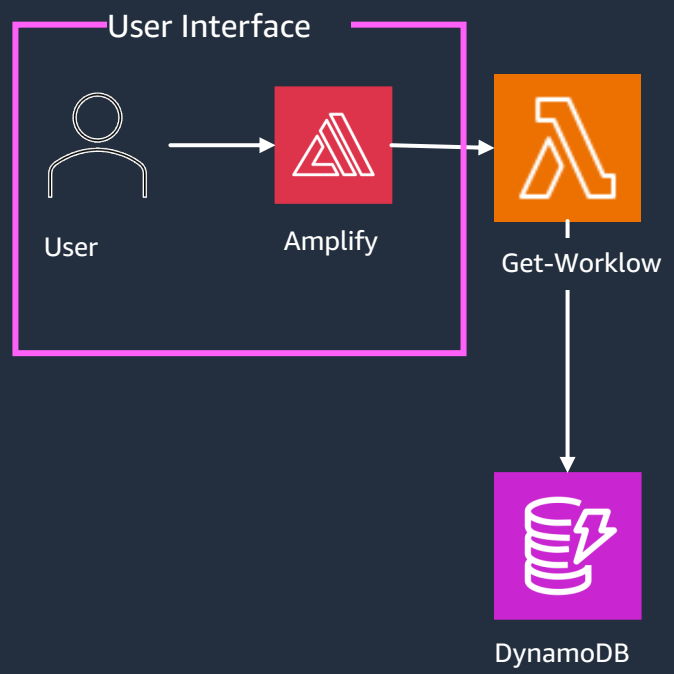
The final snippet applies the actual redactions by drawing black rectangles precisely over sensitive information.

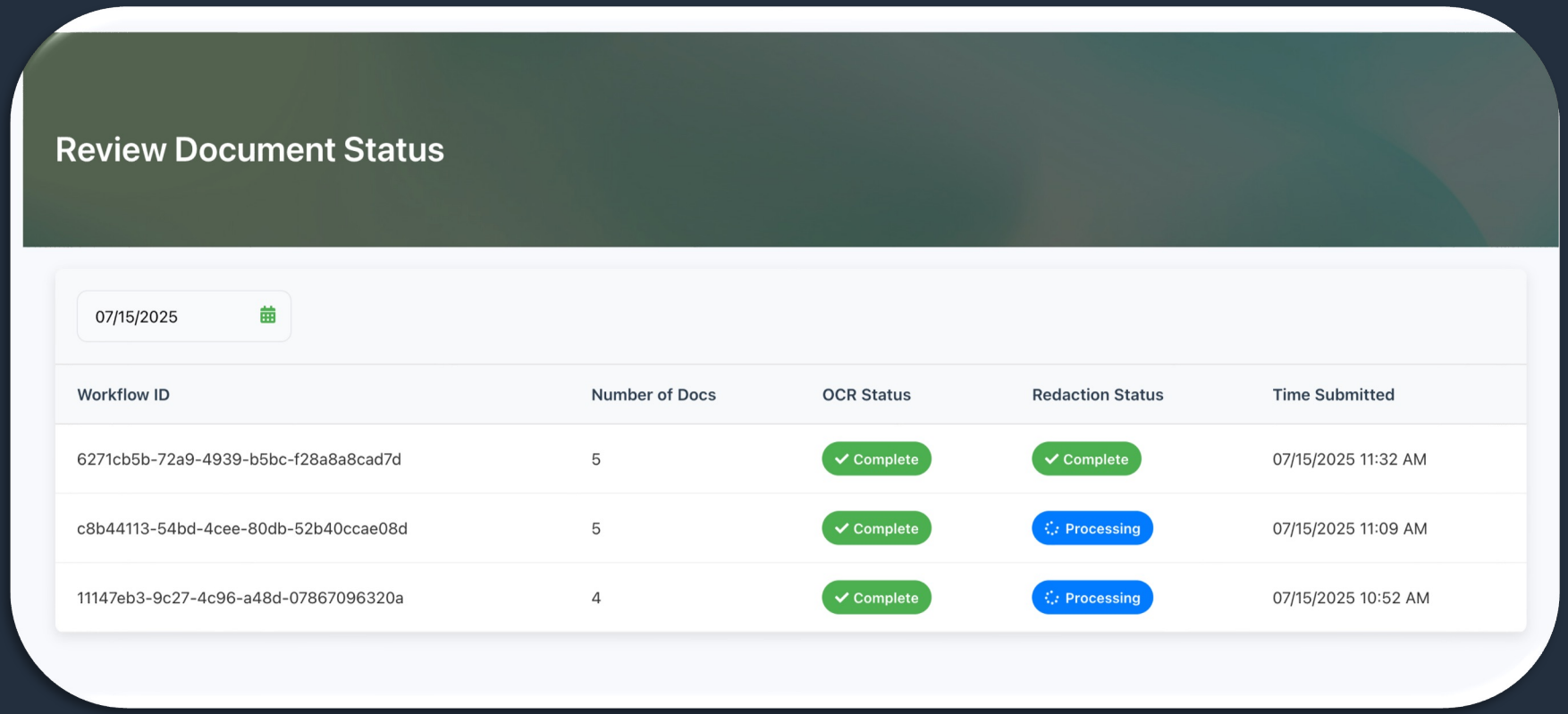
Observability and Monitoring

Insist on High Standards...

FrontEnd – Status Tracking

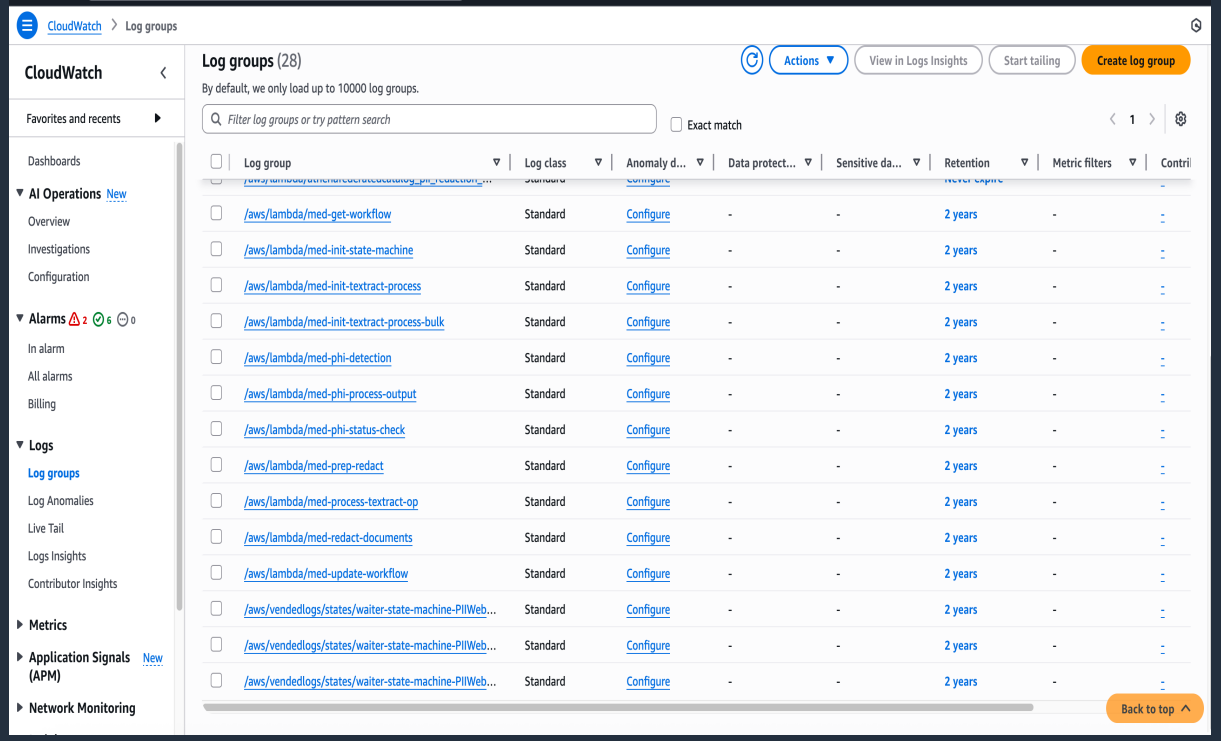
How it works:



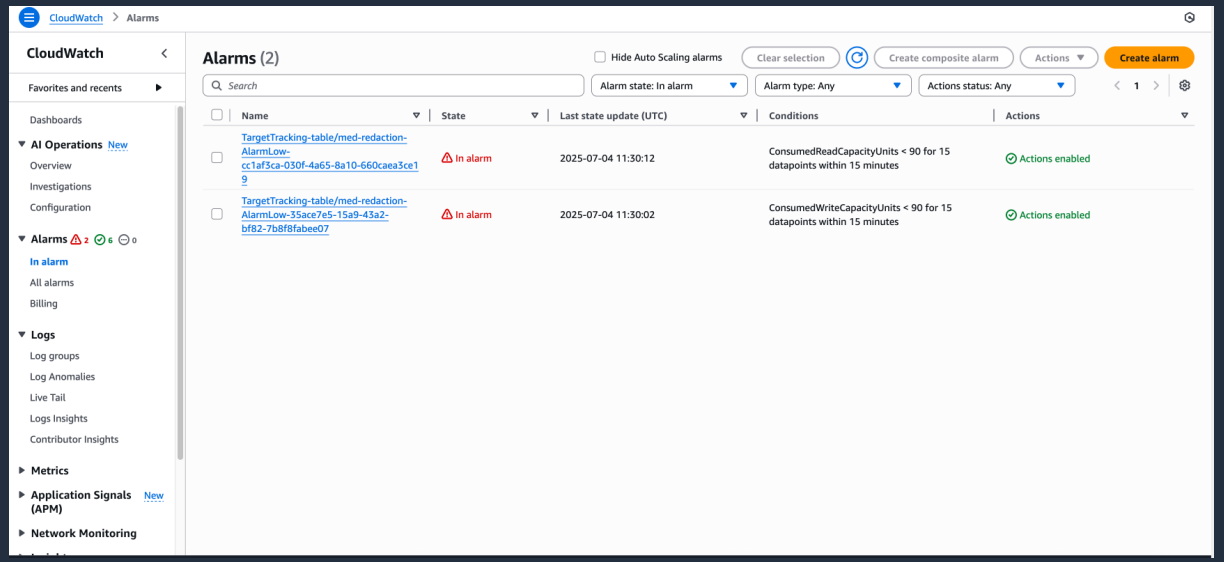


Monitoring and audit – CloudWatch

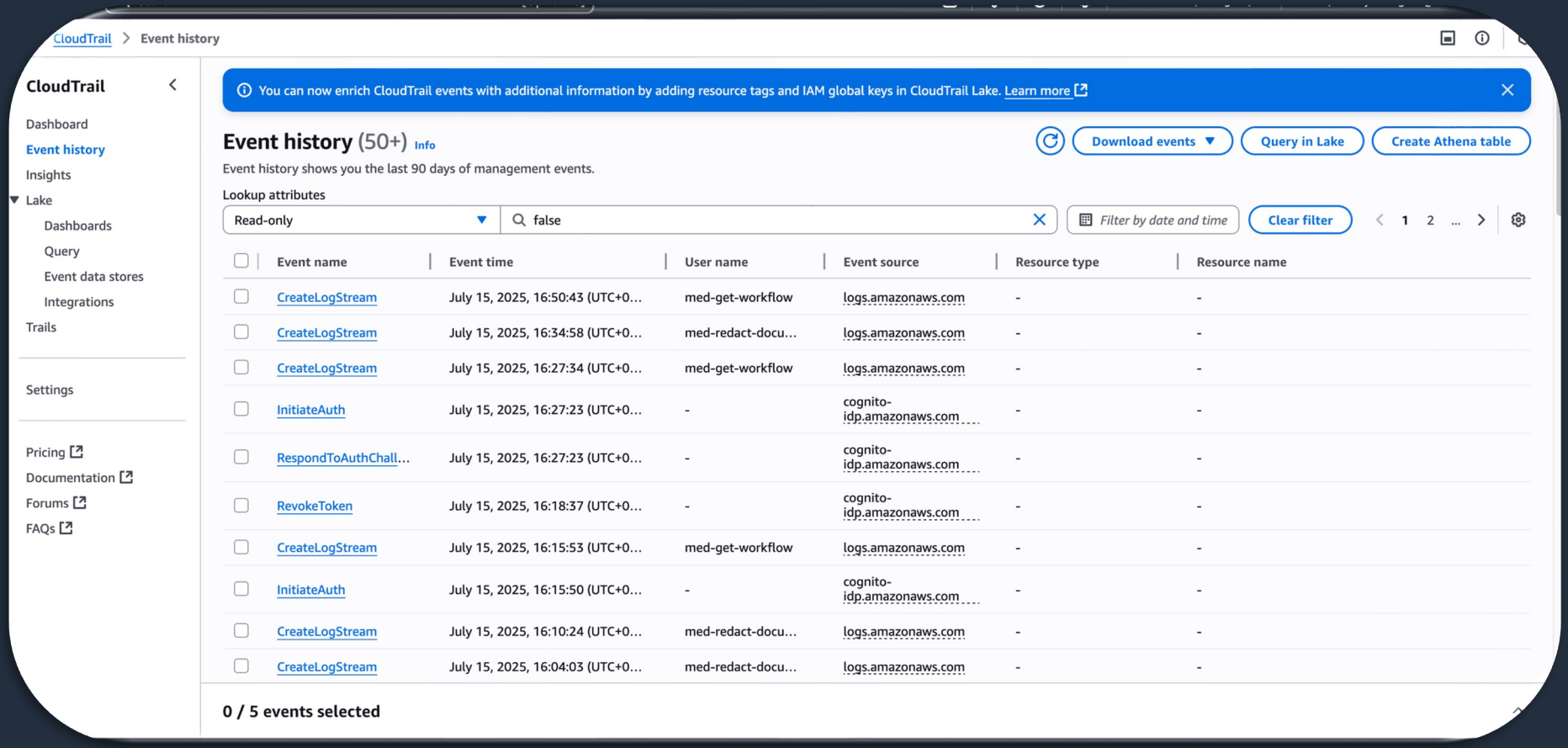
Log Groups



Alarms



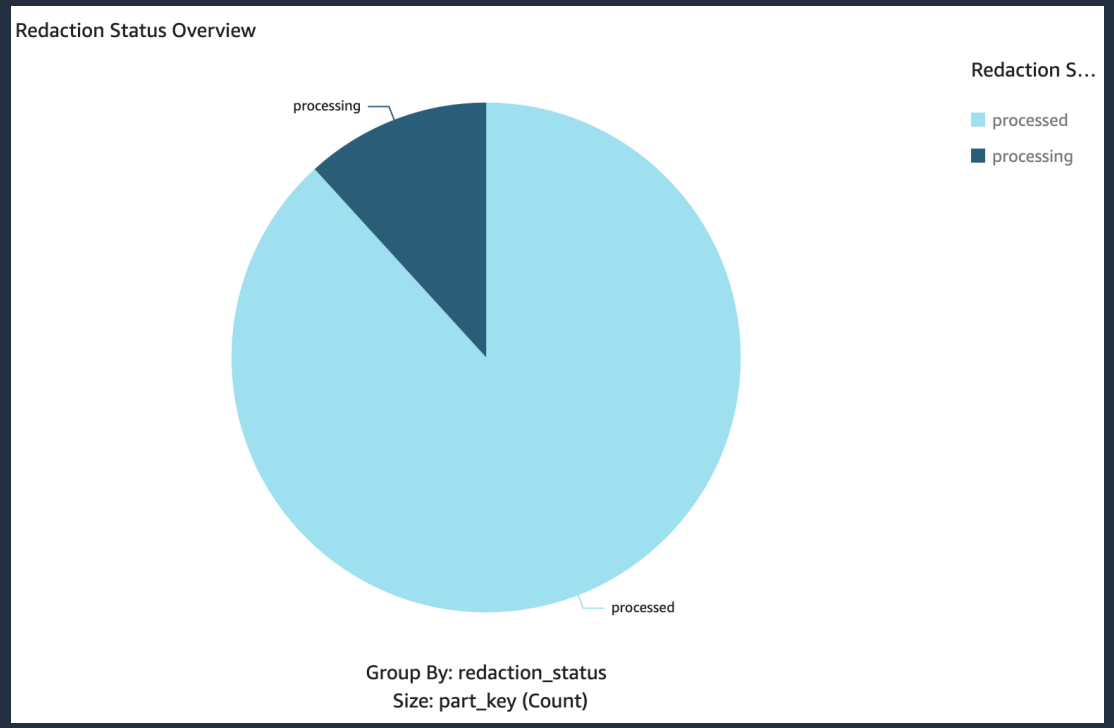
Monitoring – CloudTrail (API)



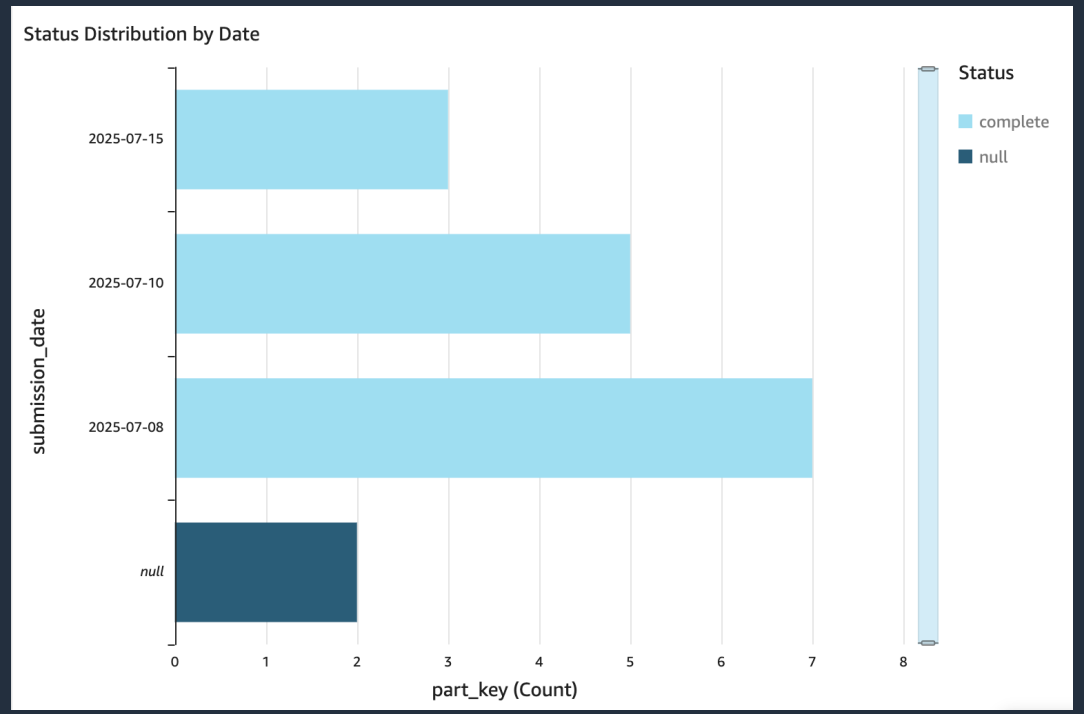
BI (Business Intelligence) - QuickSight

High Level:

QuickSight Dashbord – Pie Chart

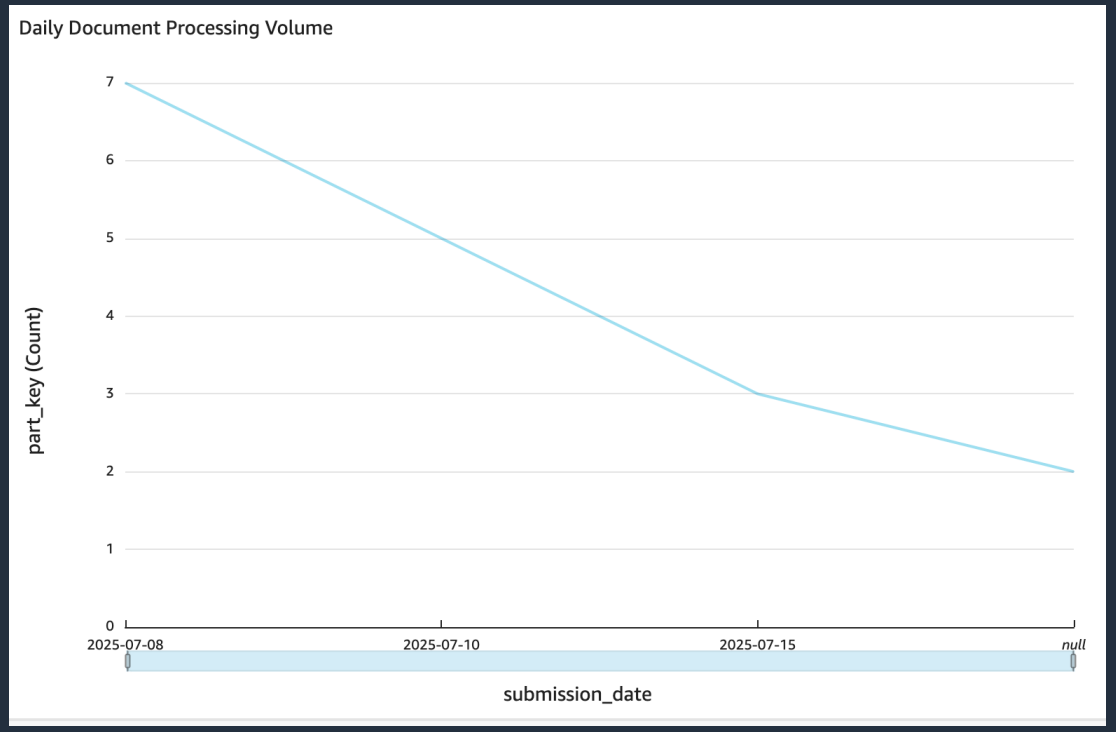


Pie Chart

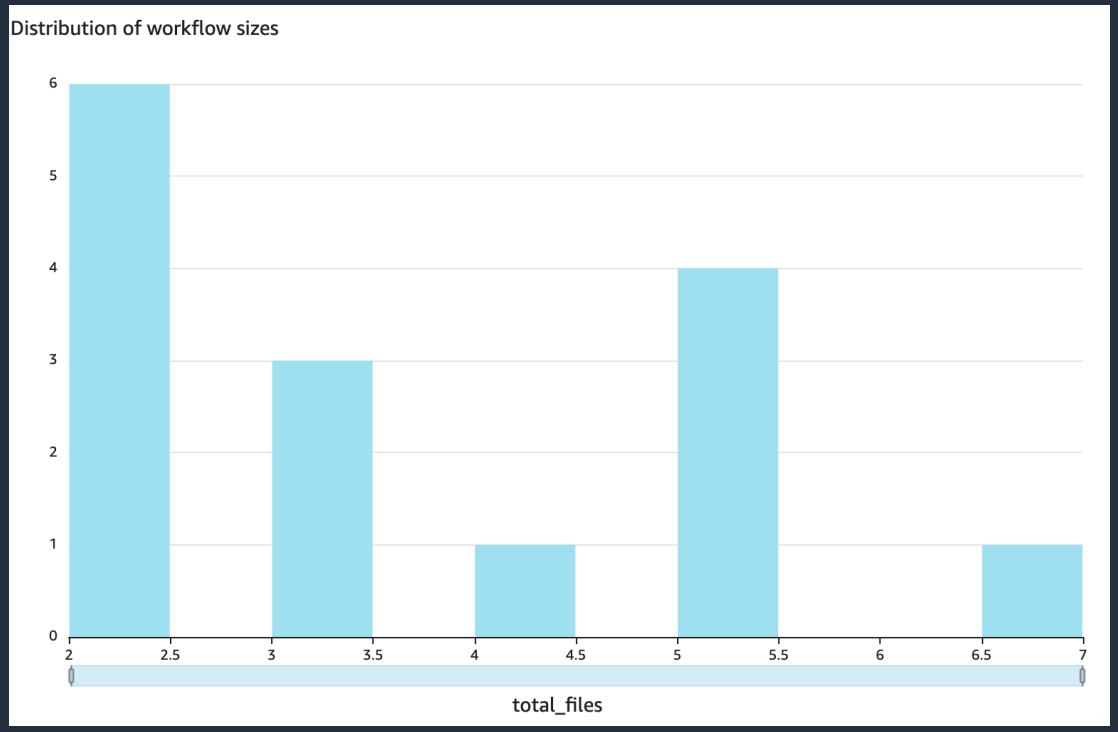


Stacked bar chart

QuickSight Dashbord



Line Chart



histogram

Security

Insist on High Standards…

Amazon’s TOP Priority !

Security

Secure storage
with encryption
at rest / in transit

IAM roles

Permissions
Least Privilege Principle

KMS Keys

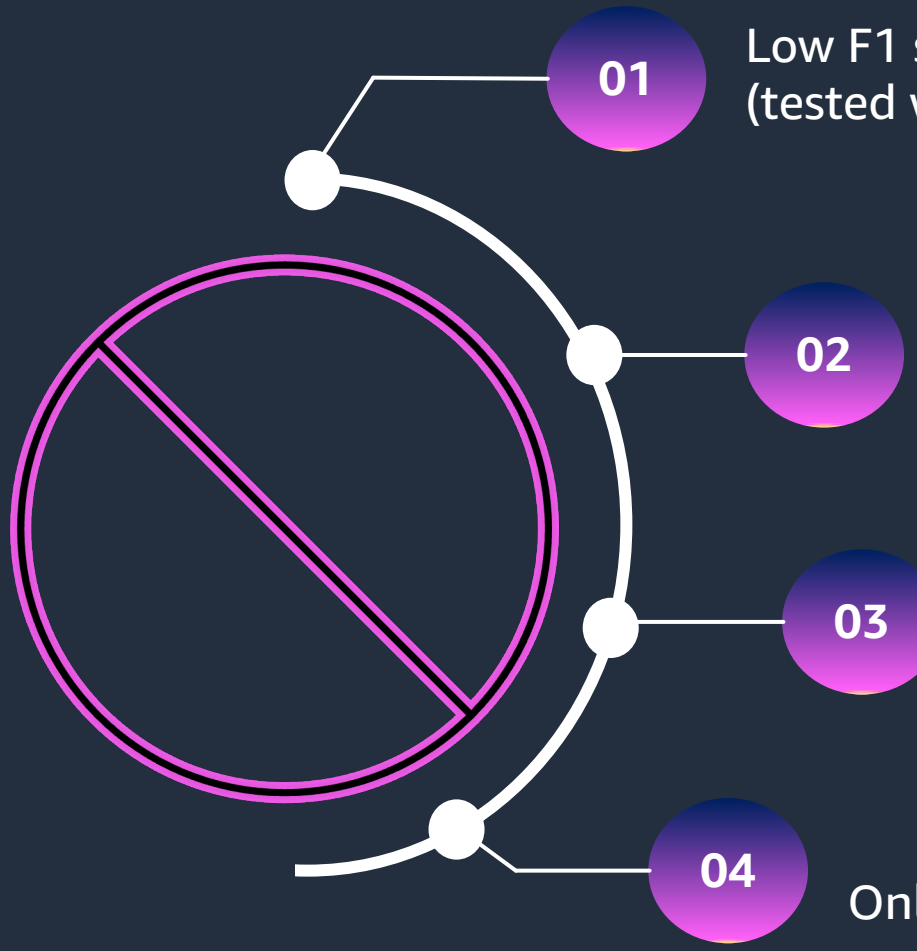
RBAC

Back to the Live Demo...

Deliver Results

Limitations

Limitations



01 Low F1 score Amazon Comprehend Medical of 0.77 (tested with 20 documents)

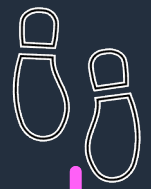
02 English-Language Documents

03 Batch Processing takes 8 minutes

04 Only accepts PDF, PNG, JPEG, TIFF document types

Next Step

Think Big



Next Steps

- Multi-region
- CloudFront
- Route 53 domain name
- Multi industries not just Healthcare
- Extend support for other format types (conversion)
- Implement custom entity recognizers using Amazon SageMaker to supplement Comprehend Medical
- Integrate Amazon Translate as a pre-processing step for non-English documents

Lessons Learned

Challenges

- Thinking Big
- Bias for Action
- Insist on high standards
- Working through Ambiguity

Victories

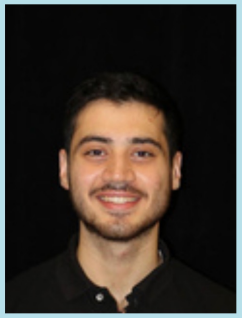
- Taking Ownership
- Learning and being curious
- Diving deeper
- Customer Obsession

Takeaways

- Networking
- Earning trust
- Discipline

Acknowledgements

ziados

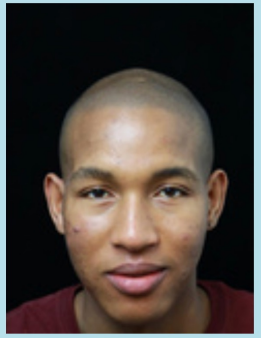


Ziad
Osman

mmmccrac

Michael
McCracken

sadoubah



Thierno-Sadou
Bah

aws professional services

Thank you!

You are now welcome to ask any questions.

Miniar Jabri
@miniarja