aws professional services

Service Suitability – PII Redaction Pipeline

Miniar JABRI

Professional Services Intern
@AWS

Agenda

1. Customer problem
2. Solutions overview
3. Evaluation criteria
4. Candidate services
5. Service comparison
6. Recommendations
7. Next steps

1.Customer Problem

Who is the customer?

-Healthcare Providers
-Health Insurance Companies
-Medical Research Institutions
-Healthcare Analytics Companies
-other healthcare organizations

What do they need?

-A system to detect & redact PII in diverse healthcare documents
-An accurate solution
-A scalable solution
-HIPAA compliant

2.Solutions Overview

End goal:

Build a fully automated and serverless redaction pipeline

AWS tools in scope:

- Amazon Textract
- Amazon Comprehend
- Amazon Comprehend Medical
- Amazon Bedrock
- Amazon SageMaker
- Amazon Macie

Today's focus: Choosing the right AI/ML service for PII redaction

3.Evaluation Criteria

What “suitability” means

Accuracy

Ability to detect sensitive entities (PHI/PII) correctly

Service maturity

Proven and supported in production

Customizability

Can we define custom rules?

Cost

Pay-per-use vs training overhead

Scalability

Handles large volumes of documents

Compliance

Highlight the most important content in the presentation and on each slide

Integration

Handles large volumes of documents

Today’s focus: Selecting the most **suitable** AI/ML service for redacting PII

4.Candidate Services

AI/ML services

- **Amazon Textract**

  Text extraction (OCR) from PDFs, scanned documents

- **Amazon Comprehend**

  Non-medical PII detection and redaction.

- **Amazon Comprehend Medical**

  Clinical PII/PHI detection

- **Amazon SageMaker**

  Fully custom ML models

- **Amazon Bedrock**

  LLM-based PII detection (experimental)

- **Amazon Macie**

  Managed sensitive data discovery and PII detection in S3 buckets

Today's focus: Selecting the right **AI/ML service** for redacting PII

5.Service Comparison

Comparison table : ML/AI services for PII detection

	Accuracy	Custom rules	cost	HIPAA Ready	Integration	Verdict
Comprehend	✓✓	✓	$	✓	✓✓	✓Best for PII
Comprehend Medical	✓✓✓	🚫	$$	✓✓	✓✓	✓For Healthcare
Bedrock	✓✓✓(LLM)	✓✓✓	$$$	🚫	! (medium, evolving)	🚫Overkill
SageMaker	✓(if trained)	✓✓✓ (if trained)	$$$	✓(manual)	✓✓ (complex)	🚫Heavy setup
Macie	✓ (S3 only)	✓ (managed types)	$$	✓✓	✓ (S3 only)	✓Good for S3 audits
Textract	No Entity detection, OCR only.					

Selected AI/ML Service

Amazon Comprehend Medical

- Medical purpose
- Fully managed
- Serverless and fully integrated
- Compliant and Secure
- Cost effective
- Simple and fast
- Customizable (custom PII)
- 99,999% accurate

Comparison table : OCR Services

	Textract	Rekognition Text Detection	Amazon Comprehend	Amazon Bedrock	SageMaker	Other OCRs (Google, Azure..)
PDF	✓	🚫	Not OCR, text analysis only	! complex prompt engineering	✓(if trained)	✓ (varies)
DOCX	! (convert)	🚫			✓(if trained)	! (varies)
Text	✓	🚫			✓(if trained)	✓
Images (PNG,JPEG,TIFF)	✓	✓ (text in images only)			✓(if trained)	✓
AWS Integration	✓	✓		✓	✓	! (depends)
Custom training	🚫	🚫		✓	✓✓	🚫
Cost	$$ (pey-per-use)	$(low cost)		$$$	$$$$	Depends

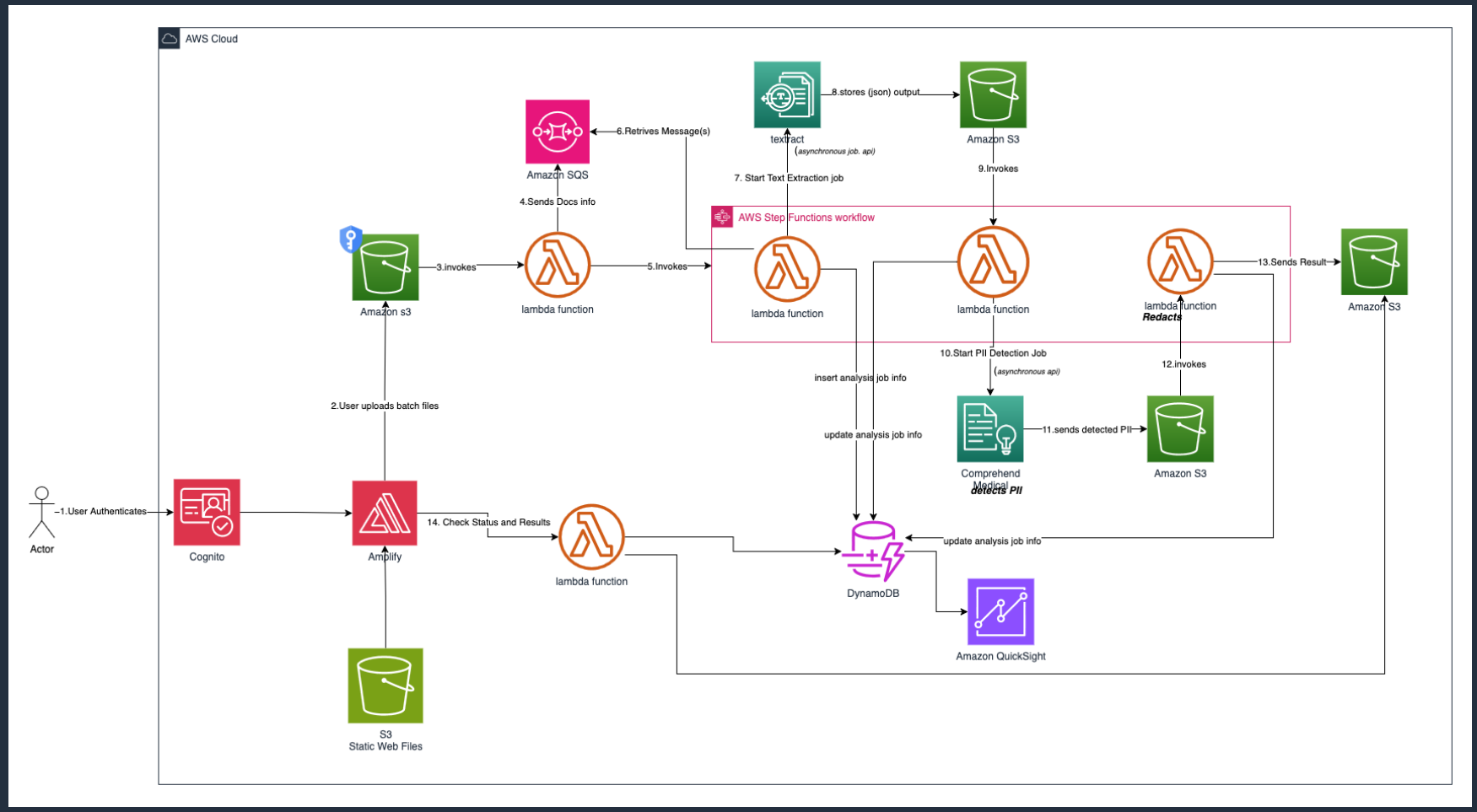
Selected OCR Service

Amazon Textract

- Native support for healthcare documents
- Supports multiple types of documents
- High layout preservation
- Fully managed
- Serverless and fully integrated
- Compliant and Secure
- Cost effective
- Simple and fast

6.Recommendation

Architecture Overview



aws professional services

Thank you!

Miniar JABRI

ProServe Intern @AWS France