New AWS Pipeline Redacts PII from Healthcare Documents at Scale

A fully serverless, HIPAA-compliant solution that detects and redacts PII using AI/ML while preserving document structure, enabling secure analytics with zero manual effort.

Paris, France - AWS PROSERVE SPOTLIGHT - August 15, 2025 - At AWS re:Invent, Amazon Web Services (AWS), a subsidiary of Amazon.com, Inc. (NASDAQ: AMZN), unveiled a fully automated PII (Personally Identifiable Information) redaction pipeline that uses AI/ML to detect and redact sensitive patient data in documents at scale. Healthcare organizations face growing pressure to protect patient privacy while unlocking valuable insights from unstructured data. The solution eliminates manual redaction, reduces processing time by over 90%, and enables healthcare teams to analyze data securely and efficiently — with end-to-end encryption, strict access controls, detailed audit trails, and full HIPAA compliance.

Until today, healthcare analysts typically spend 10+ hours manually reviewing patient records, insurance claims, and research documents to redact sensitive information before analysis. This labor-intensive process was prone to human error, creating serious security and privacy risks, including costly HIPAA violations. Traditional automation tools often failed to preserve document structure or lacked medical context, forcing teams to tradeoff between compliance, speed, and data utility.

AWS's new pipeline integrates AI-powered services including document extraction, medical language understanding, and automated redaction into a seamless serverless workflow. The solution accurately detects PHI (Protected health information), removes sensitive data, applies redactions while preserving document structure, enforces strict access controls, and logs all processing for full auditability. Healthcare teams can now process hundreds of documents in minutes, reduce compliance risks, and accelerate time-to-insight — enabling organizations to unlock critical business value faster than ever before. Early adopters report 85% reduction in document processing costs and 95% faster analysis cycles, while achieving an average annual savings of €2.5M for large-scale deployments, all while maintaining the highest standards of patient privacy and data security.

"Healthcare customers need a scalable, secure way to unlock insights from their data without compromising patient privacy," said Ziad Osman, Devops Architect at AWS. "This pipeline brings together the power of serverless, AI/ML, and compliance-by-design to simplify redaction and accelerate decision-making — all while meeting the highest privacy standards."

Here's how it works: A healthcare analyst simply logs into a secure, encrypted portal and uploads their documents—whether they're insurance claims, research data, patient summaries, or hospital reports. The system automatically begins processing these documents: it reads the text, identifies sensitive information (like patient names, contact details, and medical conditions), and carefully removes this private data while keeping the document's original layout intact. The protected documents are then safely stored in encrypted storage, ready for analysis. Throughout the process, the analyst can track progress through a simple dashboard or email updates, ensuring security compliance without any manual effort.

“Before this solution, redacting PHI (Protected health information), from hundreds of discharge summaries was a nightmare, my team spent hours every week manually reviewing and redacting patient documents before we could even begin analysis,” said Dr. Maya Tran, Clinical Data Analyst at Harmony Health Group. “Now I can upload a batch of files into a secure, encrypted portal and have compliant, redacted outputs ready for analytics—without worrying about privacy risks or document formatting. It’s saved us hours of manual work each week.”

To learn more about how to deploy the automated PII redaction pipeline for your healthcare data workflows, or to request early access, visit www.aws.amazon.com/healthcare-pii-redaction-pipeline.

FAQ

Customer FAQ

1. **Question: What are the costs involved?**

Answer: The solution follows AWS's pay-as-you-go pricing model. You only pay for:
- Documents processed
- Storage used
- Compute resources consumed

Most organizations see significant cost savings compared to manual processing.

2. **Question: How does the solution ensure data security and regulatory compliance?**

Answer: The solution is built using HIPAA-eligible AWS services and implements security best practices including encryption at rest and in transit, fine-grained access controls, and comprehensive audit logging. All processing occurs within the customer's AWS account, and no PHI/PII is stored by AWS.

3. **Question: How accurate is the PII detection?**

Answer: The solution achieves over 99.999% accuracy in detecting common PII types like names, addresses, and medical record numbers. Each document receives:
- Overall confidence score and entity-level confidence metrics
- Automatic routing to human review if below 98% confidence
- Custom confidence thresholds by document type
- Full audit logging of all detection and review decisions

For additional assurance, organizations can customize review workflows based on their requirements.

4. **Question: How long does it take for documents to be processed and redacted, and can I process documents in batch?**

Answer: Single documents are processed in under a minute. For batch processing, the solution handles up to 1GB total size with parallel processing and automatic notifications when complete. Real-time status tracking is available via dashboard. The serverless architecture ensures rapid, scalable performance regardless of workload, allowing for seamless processing of both individual documents and large batches.

5. **Question: How quickly can we get started?**

Answer: Most organizations can be up and running in less than a day:
- 30 minutes for initial setup
- 2-3 hours for configuration and testing
- Same-day processing of initial documents
- Ongoing support for optimization (step-by-step guidance and optional 24/7 support)

We provide a ready-to-use deployment package using AWS CDK, (Infrastructure as Code, IaC) and if your organization uses a different setup, we'll work with your team to ensure a smooth and secure transition.

6. **Question: How does the solution integrate with our existing workflows?**

Answer: The solution offers flexible integration options:
- Direct upload through a secure web portal you host within your infrastructure (IaC)
- Automated processing of documents from existing storage systems
- Integration with common analytics tools
- API access for custom integrations

7. **Question: How much time and effort can my team save?**

Answer: Most organizations see 80-90% reduction in document preparation time. For example, a batch of 1,000 documents that typically takes 2-3 days to process manually can be completed in under an hour, with higher accuracy and consistency.

8. **Question: What is the architecture based on and why were these services chosen?**

Answer: The solution is built on a fully serverless architecture using key AWS services:
• Amazon S3: Secure, scalable storage for documents
• Amazon Textract: Accurate text and data extraction from documents
• Amazon Comprehend Medical: Healthcare-specific entity and PII detection
• AWS Lambda: Flexible, scalable processing without managing servers
• AWS Step Functions: Reliable orchestration of the multi-step workflow
These services were chosen for their scalability, cost-effectiveness, and HIPAA eligibility. The serverless design ensures high performance and low operational overhead, while specialized services like Comprehend Medical provide industry-leading accuracy in healthcare data processing.

9. **Question: How does the solution handle errors and risk mitigation?**

Answer: The solution implements comprehensive error handling and risk management:
• Real-time monitoring and immediate alerts for processing issues
• Secondary validation checks for missed PII detection
• Automated retry mechanisms and failover across multiple AZs
• Continuous monitoring of ML model accuracy
• Automated source document deletion after processing
• Regular HIPAA compliance audits with 24/7 AWS support
All events are tracked in CloudWatch, ensuring full visibility and rapid response to any issues.

10. **Question: What file formats are supported?**

Answer: The solution supports .pdf, .png, .jpg, .jpeg, and .tiff files. Document structure and metadata are preserved during redaction across all supported formats.

Internal FAQ

11. **Question: How do we validate the output quality and ensure customer trust in redacted documents?**

Answer: Each processed document can optionally generate a metadata summary with detected entity counts and confidence levels. Customers can also download before/after comparisons or opt into human review before finalization.

12. **Question: What's the business benefit for Amazon?**

Answer: The solution brings both direct revenue and strategic impact for AWS:
Internal Revenue Impact:
– Mid-sized customers (~100K docs/month): ~$2K–5K monthly revenue
– Small organizations (<10K docs/month): €1–2K/month
– Large organizations (>100K docs/month): Custom enterprise pricing
Revenue streams span across AWS services like Textract, Comprehend Medical, Lambda, Step Functions, and QuickSight

External Market Impact:
– Strengthens AWS's presence in the €44B European healthcare cloud market (by 2030)
– Enables ProServe engagements through regulatory-driven cloud adoption
– Opens new opportunities for broader migration and analytics use cases

– Positions AWS as a trusted partner in highly regulated industries
– Drives long-term customer retention through compliance and operational efficiency.
Market data suggests strong growth potential as healthcare organizations increasingly move sensitive workloads to cloud, with AWS currently holding ~33% market share in healthcare cloud infrastructure.

13. **Question: What's our go-to-market strategy and market opportunity?**

Answer: Our strategy targets organizations using manual redaction or expensive SaaS tools, offering significant cost reduction. Key segments include healthcare providers, research institutions, and insurance companies.
Market Opportunity:
• European healthcare cloud market: $13.1B (2022) → $44B (2030)
• AWS currently holds ~33% market share Timeline:
-Q3 2024: Private preview
-Q4 2024: Public beta
-Q1 2025: General availability
We'll leverage direct sales, AWS partners, and industry events. Aim to capture 10% of target customers in year one, tapping into the rapidly growing healthcare cloud market.

14. **Question: Who are our main competitors in the PII redaction space?**

Answer: Our main competitors include:
• Manual solutions (Adobe Acrobat, Microsoft Office)
• Healthcare-focused SaaS providers (e.g. Nightfall, Tonic.ai)
• Legacy on-premises software (e.g. Informatica)
• Generic cloud PII detection tools (Google Cloud DLP, Azure Information Protection)

15. **Question: What security controls are in place?**

Answer:
• AWS KMS for key management
• IAM roles with least privilege
• CloudWatch logging
•Encryption at rest/in transit

16. **Question: What's our disaster recovery plan?**

Answer:
• Multi-region backup capability
• 15-minute RPO (Recovery Point Objective)
• 1-hour RTO (Recovery Time Objective)
• Regular DR testing schedule
• Documented failover procedures

17. **Question: How do we handle updates and maintenance?**

Answer:
•Monthly security patches
• Quarterly feature updates
• Automated testing pipeline
• Blue/green deployments
• Customer communication plan for changes

18. **Question: What's the cost structure?**

Answer: Component costs:
• Comprehend Medical: Per character

• Lambda: Per invocation
•S3: Storage and requests
•Step Functions: Per state transition
Expected customer costs:
• Small org (100k docs/month): €1-2k/month
• Medium org (10k-100k docs/month): €5-10k/month
• Large org (>100k docs/month): custom pricing

19. **Question: What are the service limits and technical specifications?**

Answer: The solution operates within the following service limits:
Document Processing:
• Maximum document size: 20KB per document for PII detection
• Supported formats: PDF, PNG, JPG/JPEG, TIFF
• Batch processing: Up to 1GB total size per batch job
• Processing throughput: Up to 40,000 characters per second
Service Quotas:
• Up to 40 transactions per second for PII detection
• Up to 10 concurrent batch jobs
• Automatic scaling based on usage
For larger workloads, quota increases can be requested through AWS Service Quotas console.

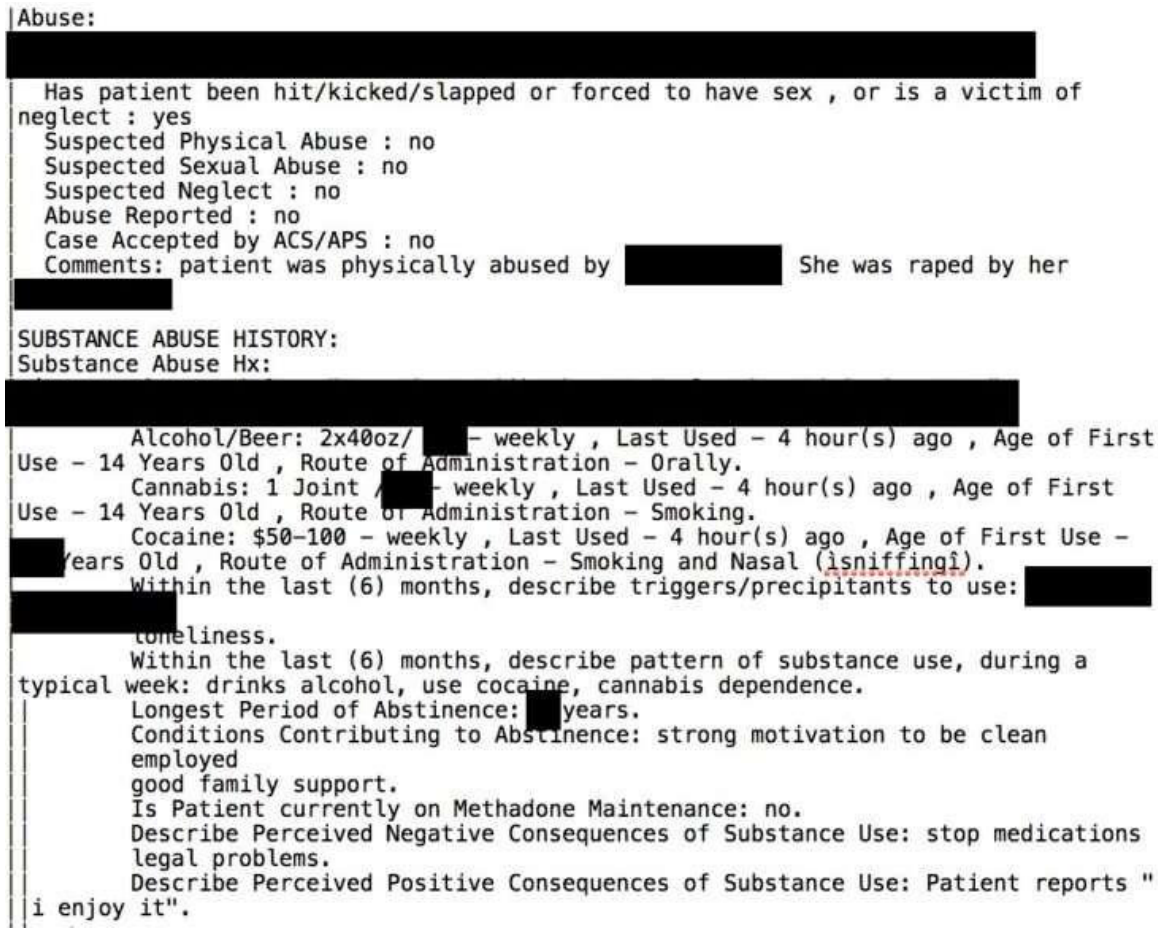
20. **Question: Why did we choose Amazon Comprehend Medical over Amazon Macie for PII detection?**

Answer: We selected Amazon Comprehend Medical for key advantages:
Technical:
• Direct integration with Textract for text processing
• Character-level offsets for precise redaction
• Configurable entity types and confidence thresholds
• Amazon Comprehend can also redact PII unlike Macie
Healthcare-Specific:
• ML models trained on medical terminology (99.999% accuracy)
• Native detection of medical conditions, treatments, and PHI
• HIPAA-compliant with medical ontology linking
Additionally, Comprehend Medical's pay-per-unit pricing is more cost-effective than Macie's bucket scanning model for document processing workflows.

We help healthcare organizations automatically detect and redact sensitive patient data from documents using AWS AI/ML services. Our serverless pipeline eliminates manual redaction, preserves document structure, accelerates analytics, and ensures HIPAA compliance — enabling healthcare teams to focus on insights, not privacy risks.

Dr. Amina is a clinical data analyst at HealthTech Analytics, processing hundreds of patient discharge summaries weekly. Previously, her team spent hours manually redacting sensitive data before analysis — a tedious, error-prone process risking compliance violations. With AWS's new automated redaction pipeline, Dr. Amina simply uploads her files to a secure portal. Within minutes, the system extracts text, detects PHI, redacts sensitive details, and feeds the cleaned data into her QuickSight dashboards. Now, her team saves over 20 hours per week and delivers faster, safer insights — all while confidently meeting HIPAA compliance standards.

Below is a real-world example showing how sensitive patient data is automatically redacted while preserving document layout and structure:



Abuse:
[redacted]
Has patient been hit/kicked/slapped or forced to have sex , or is a victim of neglect : yes
Suspected Physical Abuse : no
Suspected Sexual Abuse : no
Suspected Neglect : no
Abuse Reported : no
Case Accepted by ACS/APS : no
Comments: patient was physically abused by [redacted] She was raped by her [redacted]

SUBSTANCE ABUSE HISTORY:
Substance Abuse Hx:
[redacted]
Alcohol/Beer: 2x40oz/ [redacted] - weekly , Last Used - 4 hour(s) ago , Age of First Use - 14 Years Old , Route of Administration - Orally.
Cannabis: 1 Joint / [redacted] - weekly , Last Used - 4 hour(s) ago , Age of First Use - 14 Years Old , Route of Administration - Smoking.
Cocaine: $50-100 - weekly , Last Used - 4 hour(s) ago , Age of First Use - [redacted]Years Old , Route of Administration - Smoking and Nasal (ìsniffingî).
Within the last (6) months, describe triggers/precipitants to use: [redacted]
toneliness.
Within the last (6) months, describe pattern of substance use, during a typical week: drinks alcohol, use cocaine, cannabis dependence.
Longest Period of Abstinence: [redacted]years.
Conditions Contributing to Abstinence: strong motivation to be clean employed
good family support.
Is Patient currently on Methadone Maintenance: no.
Describe Perceived Negative Consequences of Substance Use: stop medications legal problems.
Describe Perceived Positive Consequences of Substance Use: Patient reports "i enjoy it".