Automated PDF Processing Using AWS Serverless Architecture

Course: Introduction to Cloud Computing

Group: 2

Team Members:

Siddharth | Saif | Ayush | Ujjawal

GitHub Repository:

https://github.com/siddharth404/icc-aws-pdf-textract-project/

Submission Date: 2026-02-12

1. Executive Summary

This project delivers a scalable, event-driven solution to automate the extraction of structured data from unstructured PDF resumes. By leveraging AWS Serverless technologies (Lambda, Textract, SQS, SNS), Group 2 has engineered a system that eliminates manual data entry, providing HR departments with immediate, queryable insights while optimizing for cost and operational overhead.

2. Business Context

Recruitment workflows are bottlenecked by manual data entry, processing thousands of resumes weekly. This introduces latency (5-10 mins/doc) and errors. Our solution transforms 'dark data' (PDFs) into analytics-ready CSVs significantly faster (<5s compute time), enabling real-time talent acquisition.

3. Cloud Service Model Mapping

The architecture utilizes the following cloud service models:
- FaaS (Function-as-a-Service): AWS Lambda for event-driven compute.
- SaaS (Software-as-a-Service): Amazon Textract for AI/ML document analysis.
- PaaS (Platform-as-a-Service): Amazon SQS/SNS for messaging and S3 for object storage.

4. Architecture Overview

The system follows a Decoupled Asynchronous Pattern (See Attached Diagram):
1. Ingestion: User uploads to S3 Incoming Bucket.
2. Event Trigger: S3 invokes SubmissionLambda.
3. Async Integration: Lambda triggers Textract (StartDocumentAnalysis) and exits.
4. Decoupling: Textract notifies SNS -> SQS Queue.
5. Processing: ProcessingLambda polls SQS, retrieves results, and writes to S3 Processed/Archive.

5. Key Design Decisions

- Asynchronous Textract: Chosen over synchronous API to support multi-page documents and avoid Lambda 60s/15m timeout limits.
- SQS Load Leveling: Acts as a buffer for burst traffic, ensuring downstream systems are not overwhelmed.
- Dead Letter Queue (DLQ): Captures failed events after 3 retries, ensuring zero data loss.

6. Monitoring & Security

- Observability: CloudWatch Logs for execution tracing; CloudWatch Metrics for Queue Depth.

- Security: IAM Least Privilege applied to all roles. S3 Encryption (SSE-S3) enabled at rest. TLS 1.2 in transit.

7. Cost Analysis

Based on 10,000 pages/month (us-east-1):
- Amazon Textract: $150.00 ($15 per 1,000 pages)
- AWS Lambda: ~$0.00 (Free Tier: 400,000 GB-s)
- Amazon S3/SQS/SNS: ~$0.50

Total Monthly Estimate: ~$150.50. Cost is 99% dominated by AI value services.

8. Scalability & Performance

Throughput T = (Concurrency * 1000) / Duration.
With a default concurrency of 1,000 and 2s processing time:
Max Throughput = 500 documents/second.
Daily Capacity > 4 Million documents. The system easily meets the project requirement of 10,000/day.

9. Limitations

- OCR Quality: Scans <150 DPI result in low confidence extraction.
- Handwriting: Cursive text extraction relies on confidence thresholds which may lead to null values.

10. Conclusion

Group 2 has successfully deployed a production-grade serverless pipeline. The system is robust, cost-effective, and fully automated.