Automated PDF Processing Using AWS Serverless Architecture

Introduction to Cloud Computing

Date: 2026-02-12

1. Executive Summary

This project implements a scalable, event-driven solution regarding the extraction of structured data from unstructured PDF resumes. By synthesizing AWS Textract, AWS Lambda, and Amazon SQS/SNS, the system achieves a fully decoupled architecture capable of handling burst workloads.

2. Business Problem

Recruitment processes involve processing thousands of resumes. Manual data entry introduces latency and errors. This solution automates the pipeline, converting 'dark data' (PDFs) into queryable insights.

3. Architecture Overview

The system follows a Decoupled Asynchronous Pattern:
- Ingestion: S3 Bucket (incoming/)
- Event Trigger: S3 Event -> SubmissionLambda
- Async Integration: StartDocumentAnalysis (Textract) -> SNS -> SQS
- Processing: ProcessingLambda -> S3 (processed/) + DynamoDB

4. Key Design Decisions

- Async Textract: Selected over synchronous API to support multi-page documents and avoid Lambda timeouts.
- SQS Buffering: Acts as a load leveler for burst traffic.
- Dead Letter Queue: Ensures fault tolerance for failed messages.

5. Cost Analysis

Based on 10,000 pages/month:
- Textract: $150.00 ($15 per 1k pages)
- Lambda: ~$0.00 (Free Tier)
- S3: ~$0.37
Total: ~$150.37/month. Cost is dominated by AI services.

6. Scalability Analysis

Throughput T = (Concurrency * 1000) / Duration.
With 1000 concurrent Lambdas and 2s processing time, T = 500 docs/sec.

Daily Capacity > 4 Million documents, far exceeding the 10,000/day requirement.

7. Limitations

- OCR Integrity: Dependent on scan quality (>150 DPI).
- Handwriting: Variable confidence scores.

8. Conclusion

The project successfully demonstrates a cloud-native, serverless approach to document processing, achieving high scalability and cost-efficiency.