AWS Cost Analysis: Intelligent Document Processing Application

Executive Summary

This document provides a comprehensive cost analysis for the Intelligent Document Processing (IDP) application built on AWS serverless architecture. The analysis covers three usage scenarios (Low, Medium, High) and includes all AWS services identified in the technical design.

Architecture Overview

The IDP application uses the following AWS services: - **API Gateway**: REST API endpoints - **Lambda Functions**: 6 functions for processing logic - **S3**: Document storage - **DynamoDB**: Metadata and results storage - **Textract**: OCR processing - **Bedrock**: Claude model for classification and summarization - **CloudWatch**: Logging and monitoring

Usage Scenarios

Low Usage Scenario

- **Documents processed**: 100 per month
- **Average document size**: 2 MB
- **API requests**: 500 per month
- **Target users**: Small business or pilot project

Medium Usage Scenario

- **Documents processed**: 2,000 per month
- **Average document size**: 2 MB
- **API requests**: 10,000 per month
- **Target users**: Medium enterprise department

High Usage Scenario

- **Documents processed**: 10,000 per month
- **Average document size**: 2 MB
- **API requests**: 50,000 per month
- **Target users**: Large enterprise or high-volume processing

Detailed Cost Breakdown

1. API Gateway Costs

Pricing Model: $3.50 per million API calls + data transfer

Scenario	API Calls/Month	Monthly Cost
Low	500	$0.002
Medium	10,000	$0.035
High	50,000	$0.175

2. Lambda Functions Costs

6 Lambda Functions: - UploadHandler (512 MB, 5s avg) - OCRProcessor (1024 MB, 30s avg) - DocumentClassifier (512 MB, 10s avg) - DocumentSummarizer (512 MB, 15s avg) - StatusHandler (256 MB, 1s avg) - ResultsHandler (256 MB, 2s avg)

Pricing: $0.0000166667 per GB-second + $0.20 per 1M requests

Scenario	Total Invocations	Compute Time (GB-s)	Monthly Cost
Low	700	1,225	$0.20
Medium	14,000	24,500	$4.08
High	70,000	122,500	$20.42

3. S3 Storage Costs

Storage: $0.023 per GB/month (Standard) **Requests**: $0.0004 per 1,000 PUT requests, $0.0004 per 10,000 GET requests

Scenario	Storage (GB)	PUT Requests	GET Requests	Monthly Cost
Low	0.2	100	300	$0.005
Medium	4.0	2,000	6,000	$0.09
High	20.0	10,000	30,000	$0.47

4. DynamoDB Costs

On-Demand Pricing: $1.25 per million write requests, $0.25 per million read requests **Storage**: $0.25 per GB/month

Scenario	Write Requests	Read Requests	Storage (GB)	Monthly Cost
Low	600	400	0.001	$0.001
Medium	12,000	8,000	0.02	$0.017
High	60,000	40,000	0.1	$0.085

5. Amazon Textract Costs

Pricing: $1.50 per 1,000 pages processed

Scenario	Pages/Month	Monthly Cost
Low	100	$0.15
Medium	2,000	$3.00

High	10,000	$15.00

6. Amazon Bedrock (Claude) Costs

Claude 3 Sonnet Pricing: - Input tokens: $3.00 per 1M tokens - Output tokens: $15.00 per 1M tokens

Assumptions: - Classification: 2,000 input + 50 output tokens per document - Summarization: 3,000 input + 200 output tokens per document

Scenario	Input Tokens (M)	Output Tokens (M)	Monthly Cost
Low	0.5	0.025	$1.88
Medium	10.0	0.5	$37.50
High	50.0	2.5	$187.50

7. CloudWatch Costs

Log Ingestion: $0.50 per GB **Log Storage**: $0.03 per GB/month

Scenario	Log Volume (GB)	Monthly Cost
Low	0.1	$0.053
Medium	2.0	$1.06
High	10.0	$5.30

Total Monthly Cost Summary

Service	Low Usage	Medium Usage	High Usage
API Gateway	$0.002	$0.035	$0.175
Lambda	$0.20	$4.08	$20.42
S3	$0.005	$0.09	$0.47
DynamoDB	$0.001	$0.017	$0.085
Textract	$0.15	$3.00	$15.00
Bedrock	$1.88	$37.50	$187.50
CloudWatch	$0.053	$1.06	$5.30
TOTAL	**$2.29**	**$45.76**	**$228.95**

Annual Cost Projections

Scenario	Monthly Cost	Annual Cost
Low	$2.29	$27.48
Medium	$45.76	$549.12
High	$228.95	$2,747.40

Cost Optimization Recommendations

1. Reserved Capacity

- Consider DynamoDB reserved capacity for predictable workloads
- Potential savings: 20-40% on DynamoDB costs

2. S3 Intelligent Tiering

- Implement S3 Intelligent Tiering for documents older than 30 days
- Potential savings: 30-50% on storage costs

3. Lambda Optimization

- Right-size Lambda memory allocation based on actual usage
- Implement connection pooling for database connections
- Potential savings: 15-25% on compute costs

4. Bedrock Model Selection

- Evaluate Claude 3 Haiku for simpler classification tasks
- Use Claude 3 Sonnet only for complex summarization
- Potential savings: 40-60% on AI/ML costs

5. Monitoring and Alerting

- Set up CloudWatch billing alerts
- Implement cost anomaly detection
- Regular cost review and optimization

Key Cost Drivers

1. **Amazon Bedrock (82% of high usage costs)**: Largest cost component due to AI processing
2. **Lambda Functions (9% of high usage costs)**: Scales with document volume
3. **Amazon Textract (7% of high usage costs)**: OCR processing costs
4. **Other Services (2% of high usage costs)**: Infrastructure and storage

Assumptions Made

1. **Document Complexity**: Average complexity requiring standard OCR processing
2. **Processing Success Rate**: 95% success rate, minimal reprocessing
3. **Data Retention**: 1-year retention policy for processed documents
4. **Geographic Region**: US East (N. Virginia) pricing
5. **Network Transfer**: Minimal data transfer costs (within same region)
6. **Development/Testing**: Production costs only, excludes dev/test environments

Risk Factors

1. **Token Usage Variability**: Bedrock costs can vary significantly based on document complexity
2. **Processing Failures**: Failed processing attempts still incur costs
3. **Seasonal Variations**: Document volume may fluctuate seasonally
4. **Service Pricing Changes**: AWS pricing subject to change

Conclusion

The Intelligent Document Processing application demonstrates a cost-effective serverless architecture with predictable scaling characteristics. The primary cost driver is the AI/ML processing through Amazon Bedrock, which provides significant value through automated document classification and summarization.

For organizations starting with document processing, the low usage scenario at $27.48 annually provides an excellent entry point. As volume scales, the cost per document decreases due to the serverless architecture's efficiency.

Regular monitoring and optimization can achieve 20-40% cost reductions through right-sizing and intelligent service selection.