# AWS Pricing Analysis: Metadata Extraction, Summarization, and PII Detection

## Assumptions:

1. Document Distribution: 10,000 documents

- Large: 4,000 documents @ 20 MB each = 80,000 MB (~78.125 GB)

- Medium: 3,000 documents @ 12 MB each = 36,000 MB (~35.156 GB)

- Small: 3,000 documents @ 4 MB each = 12,000 MB (~11.719 GB)

- Total Document Size: 126,000 MB (~123.047 GB)

2. Amazon Comprehend PII Detection Pricing: $0.0001 per unit (1 unit = 100 characters).

3. LLM-Based Metadata Extraction and Summarization:

- LLM pricing: $0.002 per 1,000 tokens (average)

- 1 Token ≈ 4 characters

4. SNS Notification Costs: $0.50 per 1 million notifications; each document triggers 1 notification.

## Cost Calculations

### 1. Amazon Comprehend PII Detection

Units Calculation:

- Large Documents: 20 MB ≈ 20,000,000 characters = 200,000 units/document

Total for 4,000 documents: 4,000 × 200,000 = 800,000,000 units

- Medium Documents: 12 MB ≈ 12,000,000 characters = 120,000 units/document

Total for 3,000 documents: 3,000 × 120,000 = 360,000,000 units

- Small Documents: 4 MB ≈ 4,000,000 characters = 40,000 units/document

Total for 3,000 documents: 3,000 × 40,000 = 120,000,000 units

Grand Total: 800M + 360M + 120M = 1,280,000,000 units

Cost: 1,280,000,000 × $0.0001 = $128,000.00

### 2. LLM-Based Metadata Extraction & Summarization

Token Calculation:

- Large Documents: 20 MB ≈ 5,000,000 tokens/document

Total for 4,000 documents: 4,000 × 5,000,000 = 20,000,000,000 tokens

- Medium Documents: 12 MB ≈ 3,000,000 tokens/document

Total for 3,000 documents: 3,000 × 3,000,000 = 9,000,000,000 tokens

- Small Documents: 4 MB ≈ 1,000,000 tokens/document

Total for 3,000 documents: 3,000 × 1,000,000 = 3,000,000,000 tokens

Grand Total: 20B + 9B + 3B = 32,000,000,000 tokens

Cost: 32,000,000,000 ÷ 1,000 × $0.002 = $64,000.00

### 3. SNS Notification Costs

Notifications: 10,000 documents

Cost: 10,000 ÷ 1,000,000 × $0.50 = $0.005

## Summary of Costs

|  |  |
| --- | --- |
| Service | Cost |
| Amazon Comprehend PII Detection | $128,000.00 |
| LLM-Based Metadata Extraction | $64,000.00 |
| SNS Notifications | $0.005 |
| Total for LLM Workflow | $64,000.01 |

## Analysis

1. Comprehend PII Detection:

- PII detection is expensive because it requires scanning every character in the document to find sensitive information.

- At $0.0001 per 100 characters, the cost scales linearly with document size.

- Total cost: $128,000 for 10,000 documents.

2. LLM-Based Metadata Extraction and Summarization:

- LLM pricing is token-based, which is cheaper per unit compared to Comprehend.

- Even when processing the full document to generate summaries, the cost is significantly lower: $64,000.

- Optimizations like token reduction (summarizing key sections) can further reduce costs.

3. SNS Notifications:

- Costs for notifications are negligible at $0.005 for 10,000 messages.

4. Overall Comparison:

- Using an LLM-based approach for metadata extraction and summarization reduces costs by 50% compared to Amazon Comprehend for full-document analysis.

- The total for the LLM workflow, including notifications, is $64,000.01.

## Recommendations

1. Use LLMs for metadata extraction and summarization, as they offer significant cost savings over Amazon Comprehend for full-document analysis.

2. If PII detection is still required, consider using a hybrid approach:

- Preprocess the documents to reduce text size (e.g., extract relevant sections).

- Run PII detection only on extracted sections instead of the full document.

3. Explore further optimizations, such as prompt tuning and token reduction strategies, to reduce LLM costs further.