

# Healthcare Management System - AWS Cost Estimation

---

## Cost Calculator

<https://calculator.aws/#/estimate?id=af1af2243a65549ed9e32644fb573b06b5f2b979>

## Frontend Component

### AWS Amplify

Item	Usage	Monthly Cost
Build Minutes	20	\$0*
Data Transfer	277.96GB	\$41.69*
Storage	18.8MB	\$0*
<b>Total</b>		\$41.69
<b>Total per medic</b>		\$0.004
<b>Total per appointment</b>		\$0.000087

### Detail

- All 10000 medics are daily active users
- Web app size is 4.7MB (from demo)
- Users visit 1 page
- Single page is 1MB in size
- App takes 5 minutes to build (from demo)
- WAF Firewall is 15 USD + WAF costs
- No build cost if they build outside Amplify
- if free tier applies: no cost on storage or build time + 15 GB free (2.25 USD)

### Calculations

- Build Minutes: 1 build weekly  $\times$  5 minutes per build  $\times$  4 weeks = 20 min - 1000 min per month free = 0 min
- Data Transfer: 10000 daily active users  $\times$  1MB per webpage  $\times$  30 days = 292.96GB - 15GB per month free = 277.96GB
- Storage: 4.7MB per build  $\times$  1 build weekly  $\times$  4 weeks = 18.8 MB - 5GB free tier storage = 0MB

## Amazon Cognito

Item	Usage	Monthly Cost
Monthly Active Users	10000	0*

Item	Usage	Monthly Cost
<b>Total</b>	0	
<b>Total per medic</b>	0	
<b>Total per appointment</b>	0	

#### Detail

- Cost jumps to 149.25USD if the 10000 users are SAML/OIDC

#### Calculations

- Monthly Active Users: 10000 MAUs \* 0.015 per MAU - 10000 MAUS free tier = 0

### Frontend Cost Summary

Item	Monthly Cost
Amazon Cognito	\$0*
Amplify Hosting	\$41.69*
<b>Total</b>	<b>\$41.69</b>
<b>Total per medic</b>	<b>\$0.004169</b>
<b>Total per appointment</b>	<b>\$0.00000868125</b>

### Frontend Alternative Scenarios

Scenario	Monthly Cost	Notes
SAML/OIDC Authentication	+\$149.25	If using enterprise SSO instead of standard Cognito
With AWS WAF	+\$15.00	Basic WAF protection (excludes usage-based charges)
Without Free Tier	+\$2.25	Storage + 15GB data transfer if free tier exhausted

### Backend Component

#### Amazon API Gateway

Item	Usage	Monthly Cost
Virtual Assistant Solution	0 requests	\$0.00
Baseline (Other Apps - 30K RPM)	288M requests	\$287.00
<b>Total</b>		<b>\$287.00</b>
<b>Total per medic</b>		<b>\$0.0287</b>
<b>Total per appointment</b>		<b>\$0.00005979166</b>

Item	Usage	Monthly Cost
Alternative (REST API for baseline)	288M requests	\$1,004.50

#### Detail

- **Virtual assistant solution does NOT use API Gateway** - all traffic bypasses it:
  - Frontend → AgentCore (direct)
  - AgentCore → Lambda tools (direct invocation)
  - No API Gateway in the solution architecture
- **Baseline represents OTHER existing applications** using the same infrastructure:
  - 30K RPM capacity requirement for other apps
  - 288M requests monthly from other systems
  - NOT part of the virtual assistant solution cost
- Using HTTP API pricing (\$1.00 per million) - cheaper than REST API (\$3.50 per million)

#### Calculations

##### Virtual Assistant Solution:

- API Gateway Usage: \$0.00 (solution does not use API Gateway)
- All communication is direct: Frontend ↔ AgentCore ↔ Lambda

##### Baseline (Other Apps - 30K RPM with HTTP API):

- HTTP API Requests:  $\$1.00 \times (288M - 1M \text{ free tier}) \div 1M = \$287.00$
- Data Transfer:  $\$0.09 \times (2,880\text{GB} - 100\text{GB free tier}) = \$250.20$
- Total Baseline: \$537.20

##### Alternative (REST API for baseline):

- REST API Requests:  $\$3.50 \times (288M - 1M \text{ free tier}) \div 1M = \$1,004.50$
- Data Transfer: Same = \$250.20
- Total: \$1,254.70 (2.3x more expensive than HTTP API)

#### AWS Lambda

Item	Usage	Monthly Cost
API Requests	240K GB-s	\$4.16
Agent Tools	1.2M GB-s	\$15.27
<b>Total</b>		<b>\$19.43</b>
<b>Total per medic</b>		<b>\$0.001943</b>
<b>Total per appointment</b>		<b>\$0.00000404791</b>
Baseline (30K RPM capacity)	14.4M GB-s	\$240.00

#### Detail

- Realistic usage:
  - API Lambdas: 4.8M requests (0.4s each, 128MB)
  - Tool Lambdas: 24M tool invocations (0.4s each, 128MB)
  - Total: 28.8M invocations, 1.44M GB-seconds
- Baseline capacity: 288M API requests at 30K RPM (0.4s each, 128MB)

## Calculations

### Realistic Usage:

- API Lambda Requests:  $\$0.0000002 \times 4.8M = \$0.96$
- API Lambda Compute:  $\$0.0000166667 \times (4.8M \times 0.4s \times 0.125GB) = \$4.00$
- Tool Lambda Requests:  $\$0.0000002 \times 24M = \$4.80$
- Tool Lambda Compute:  $\$0.0000166667 \times (24M \times 0.4s \times 0.125GB) = \$20.00$
- Total: \$29.76 (after free tier: ~\$19.43)

### Baseline (30K RPM):

- Requests:  $\$0.0000002 \times 288M = \$57.60$
- Compute:  $\$0.0000166667 \times (288M \times 0.4s \times 0.125GB) = \$240.00$
- Total: \$297.60 (after free tier)

## Amazon RDS Aurora

Item	Usage	Monthly Cost
Storage	43.2 GB	\$4.32
Storage I/Os	52.8M I/Os	\$10.56
Compute	600 ACU-hours	\$72.00
RDS Data API	52.8M requests	\$18.13
<b>Total</b>		<b>\$105.01</b>
<b>Total per medic</b>		<b>\$0.010501</b>
<b>Total per appointment</b>		<b>\$0.00002187708</b>
Baseline (30K RPM capacity)	336M I/Os	\$67.28

## Detail

- Realistic usage: 52.8M I/O operations monthly (91.67 IOPS average)
  - API calls: 4.8M I/Os
  - Appointments: 24M I/Os (5 I/Os per appointment)
  - Document processing: 24M I/Os (5 I/Os per document)
- Baseline capacity: 30K RPM = 500 IOPS sustained for 160 hours = 288M I/Os
- ACU sizing: 0.5-5 ACUs (450-650 IOPS per ACU baseline)
- All database access via RDS Data API (no VPC data transfer costs)

## Calculations

## Realistic Usage:

- Storage:  $43.2\text{GB} \times \$0.10 = \$4.32$ 
  - Patient data:  $6\text{KB} \times 4.8\text{M appointments} = 28.8\text{GB}$
  - Document vectors:  $3\text{KB} \times 4.8\text{M documents} = 14.4\text{GB}$
- I/O Operations:  $52.8\text{M} \times \$0.0000002 = \$10.56$ 
  - Baseline (100/day):  $2,377 \text{I/Os} \times \$0.0000002 = \$0.0005$
  - Peak ( $91.67 \text{IOPS} \times 160\text{h}$ ):  $52.8\text{M I/Os} \times \$0.0000002 = \$10.56$
- Compute (ACUs):
  - Capacity estimation:
  - 8 ACUs approx= db.r8g.large with 3600 IOPS and 4 ACUs approx=db.r8g.medium = 2500 IOPS Base -> 450-650 IOPS per ACU
  - ACUs in peak hours approx=  $2 \text{ACUs} \times 160 \text{hours} \times 0.12 \text{USD per ACU/Hour} = \$38.40$
  - ACUs rest of the time =  $0.5 \text{ACUs} \times 560 \text{hours} \times 0.12 \text{USD per ACU/Hour} = \$33.60$
  - Total cost: \$72.00
  - 24\*7 availability
- RDS Data API:  $(52.8\text{M} - 1\text{M free tier}) \times \$0.35/\text{M} = \$18.13$

## Baseline (30K RPM):

- I/O Operations:  $336\text{M} \times \$0.0000002 = \$67.28$ 
  - Peak ( $500 \text{IOPS} \times 160\text{h}$ ):  $288\text{M I/Os}$
  - Appointments + Documents:  $48\text{M I/Os}$
  - Total:  $336\text{M I/Os}$
- RDS Data API:  $(336\text{M} - 1\text{M free tier}) \times \$0.35/\text{M} = \$117.25$
- Total Baseline I/O + API: \$184.53

## Backend Cost Summary

Item	Monthly Cost
Amazon API Gateway	\$0.00
AWS Lambda	\$19.43
Amazon RDS Aurora	\$105.01
<b>Total</b>	<b>\$124.44</b>
<b>Total per medic</b>	<b>\$0.012444</b>
<b>Total per appointment</b>	<b>\$0.00002592</b>

## Backend Baseline Costs (Other Applications)

Item	Monthly Cost	Notes
Amazon API Gateway (30K RPM)	\$287.00	HTTP API for other apps
API Gateway Data Transfer	\$250.20	2,880GB after free tier
AWS Lambda (30K RPM)	\$240.00	Baseline capacity compute

Item	Monthly Cost	Notes
Amazon RDS I/O (30K RPM)	\$67.28	Additional I/O operations
RDS Data API (30K RPM)	\$117.25	Additional API calls
<b>Baseline Total</b>	<b>\$961.73</b>	Infrastructure for other systems

## Backend Alternative Scenarios

Scenario	Monthly Cost	Notes
REST API instead of HTTP API	+\$717.50	\$1,004.50 vs \$287.00 for baseline

## Document Workflow Component

### Amazon S3

Item	Usage	Monthly Cost
Standard Storage (New)	12.48TB	\$286.70
PUT Requests (Upload)	4.8M requests	\$24.00
PUT Requests (Metadata)	33.6M requests	\$168.00
GET Requests	14.4M requests	\$5.76
<b>Total</b>		<b>\$484.46</b>
<b>Total per medic</b>		<b>\$0.048446</b>
<b>Total per document</b>		<b>\$0.0001</b>
Baseline (Existing Documents)	15TB	\$345.00

### Detail

- **New storage from solution:** 12.48TB monthly
  - New documents:  $4.8\text{M docs} \times 2.55\text{MB} = 12.24\text{TB}$
  - Processed data:  $4.8\text{M docs} \times 50\text{KB} = 240\text{GB}$
- **Baseline (existing documents):** 15TB pre-existing medical documents (not part of solution cost)
- **Document uploads:** 4.8M PUT requests (1 per appointment)
- **Metadata updates:** 33.6M PUT requests
  - Input documents:  $4.8\text{M} \times 2 \text{ updates (upload + processing complete)} = 9.6\text{M}$
  - Output documents:  $4.8\text{M} \times 5 \text{ data automation outputs} \times 1 \text{ update each} = 24\text{M}$
  - Total: 33.6M metadata updates
- **GET requests:** 14.4M retrievals (3x rate: upload access + processing + knowledge base)
- Document lifecycle: Upload → Metadata Update → Process → Metadata Update → Store → Retrieve

### Calculations

#### Solution Cost:

- New Storage:  $\$0.023 \times 12,480\text{GB} = \$286.70$ 
  - Documents:  $12,240\text{GB} \times \$0.023 = \$281.52$
  - Processed data:  $240\text{GB} \times \$0.023 = \$5.52$
- PUT Requests (Upload):  $\$0.005 \times (4.8\text{M} \div 1\text{K}) = \$24.00$
- PUT Requests (Metadata):  $\$0.005 \times (33.6\text{M} \div 1\text{K}) = \$168.00$ 
  - Input metadata:  $9.6\text{M updates} \times \$0.005/\text{1K} = \$48.00$
  - Output metadata:  $24\text{M updates} \times \$0.005/\text{1K} = \$120.00$
- GET Requests:  $\$0.0004 \times (14.4\text{M} \div 1\text{K}) = \$5.76$
- Total: \$462.86

### **Baseline (Existing Documents):**

- Existing Storage:  $\$0.023 \times 15,000\text{GB} = \$345.00$

### **AWS Lambda (Document Processing)**

<b>Item</b>	<b>Usage</b>	<b>Monthly Cost</b>
Job Launch	600K GB-s	\$8.96
Document Processing	1.8M GB-s	\$24.96
<b>Total</b>		\$33.92
<b>Total per medic</b>		\$0.03392
<b>Total per document</b>		\$0.00000706666

### **Detail**

- Document processing: 4.8M documents monthly (1 per appointment)
- Processing time: 3 seconds per document with 128MB memory
- Extracts 50KB structured data per 5-page document
- Triggers knowledge base ingestion after processing if not busy

### **Calculations**

- Document Processing: 4.8M documents 2 lambda calls (3s data extraction, 128MB) + (1s bda launch, 128MB)
  - bda launch calculator estimation: Amount of memory allocated:  $128\text{ MB} \times 0.0009765625\text{ GB in a MB} = 0.125\text{ GB}$  Amount of ephemeral storage allocated:  $512\text{ MB} \times 0.0009765625\text{ GB in a MB} = 0.5\text{ GB}$  Pricing calculations  $4,800,000\text{ requests} \times 1,000\text{ ms} \times 0.001\text{ ms to sec conversion factor} = 4,800,000.00\text{ total compute (seconds)}$   $0.125\text{ GB} \times 4,800,000.00\text{ seconds} = 600,000.00\text{ total compute (GB-s)}$  Tiered price for:  $600,000.00\text{ GB-s} \times 0.0000133334\text{ USD} = 8.00\text{ USD}$  Total tier cost = 8.00 USD (monthly compute charges) Monthly compute charges: 8.00 USD  $4,800,000\text{ requests} \times 0.0000002\text{ USD} = 0.96\text{ USD}$  (monthly request charges) Monthly request charges: 0.96 USD 0.50 GB - 0.125 GB (no additional charge) = 0.00 GB billable ephemeral storage per function Monthly ephemeral storage charges: 0 USD 8.00 USD + 0.96 USD = 8.96 USD Lambda cost (monthly): 8.96 USD
- data extraction calculator estimation

Amount of memory allocated: 128 MB x 0.0009765625 GB in a MB = 0.125 GB Amount of ephemeral storage allocated: 512 MB x 0.0009765625 GB in a MB = 0.5 GB Pricing calculations 4,800,000 requests x 3,000 ms x 0.001 ms to sec conversion factor = 14,400,000.00 total compute (seconds) 0.125 GB x 14,400,000.00 seconds = 1,800,000.00 total compute (GB-s) Tiered price for: 1,800,000.00 GB-s 1,800,000 GB-s x 0.0000133334 USD = 24.00 USD Total tier cost = 24.0001 USD (monthly compute charges) Monthly compute charges: 24.00 USD 4,800,000 requests x 0.0000002 USD = 0.96 USD (monthly request charges) Monthly request charges: 0.96 USD 0.50 GB - 0.5 GB (no additional charge) = 0.00 GB billable ephemeral storage per function Monthly ephemeral storage charges: 0 USD 24.00 USD + 0.96 USD = 24.96 USD Lambda cost (monthly): 24.96 USD

## Amazon Bedrock Data Automation

Item	Usage	Monthly Cost
Document Analysis	24M pages	\$960000
<b>Total</b>		<b>\$960000</b>
<b>Total per medic</b>		<b>\$96</b>
<b>Total per document</b>		<b>\$0.2</b>

### Detail

- Document processing: 4.8M documents × 5 pages average = 24M pages monthly
- Bedrock Data Automation extracts structured data from medical documents
- Generates 50KB processed data per document for knowledge base ingestion
- Pricing based on page count for document analysis

### Calculations

- Document Analysis: \$0.040 per page on custom output × 24M pages = \$960000
- Alternative: \$0.010 per page on standard output × 24M pages = \$240000 -> 24 usd per medic -> 0.05 per document

## Amazon EventBridge

- From aws service to aws service normal events in the same account: 0 USD

## Document Workflow Cost Summary

Item	Monthly Cost
Amazon S3	\$484.46
AWS Lambda (Document Processing)	\$33.92
Amazon Bedrock Data Automation	\$960,000.00
Amazon EventBridge	\$0.00
<b>Total</b>	<b>\$960,518.38</b>

Item	Monthly Cost
Total per medic	\$96.05183
Total per appointment	\$0.20010799

## Document Workflow Baseline Costs

Item	Monthly Cost	Notes
S3 Storage (Existing Documents)	\$345.00	15TB pre-existing medical documents

## Document Workflow Alternative Scenarios

Scenario	Monthly Cost	Savings	Notes
BDA Standard Output	\$240,518.38	-\$720,000	\$0.01/page vs \$0.04/page, less structured data, not that viable with medical data
Textract + Comprehend (5 pages)	\$1,600,358.38	+\$639,840	Full processing with tables, more expensive than BDA
Without BDA (Manual KB Ingestion)	\$614.38	-\$959,904	Requires custom parsing, limited multimodality and manual PII/PHI detection and masking
Custom Parsing with Nova Pro	\$20,014.38	-\$940,504	LLM-based parsing instead of BDA
Without BDA + Custom Parsing	\$96.00	-\$96,000	KB ingestion only, 4.8B tokens vs 24B tokens

## Textract + Comprehend Alternative Breakdown

### Workload Assumptions

- 4,800,000 documents per month
- 5 pages per document average (matching baseline)
- Total pages: 24M pages monthly

### Amazon Textract Costs (5 pages per document = 24M pages)

Feature	Pages	Price per 1K	Monthly Cost
DetectDocumentText (Basic OCR)	24M	\$1.50	\$36,000
AnalyzeDocument Forms	24M	\$50.00	\$1,200,000
AnalyzeDocument Tables	24M	\$15.00	\$360,000
<b>Textract Total</b>			<b>\$1,596,000</b>

### Amazon Comprehend Costs

Feature	Documents	Price per Unit	Monthly Cost
Standard Entity Detection	4.8M	\$0.0001 per unit	\$480
Custom Entity Detection	4.8M	\$0.0003 per unit	\$1,440
Custom Classification	4.8M	\$0.0003 per unit	\$1,440
PII Detection	4.8M	\$0.0001 per unit	\$480
<b>Comprehend Total</b>			<b>\$3,840</b>

#### Total Cost Comparison

Scenario	Textract	Comprehend	S3 + Lambda	Total	vs BDA Custom
Textract + Comprehend (5 pages)	\$1,596,000	\$3,840	\$518.38	<b>\$1,600,358.38</b>	+\$639,840
BDA Custom Output (5 pages)	-	-	\$518.38	<b>\$960,518.38</b>	baseline

#### Key Insights

- **Bedrock Data Automation is significantly more cost-effective**
  - BDA saves \$639,840/month (40% cheaper) vs Textract + Comprehend
  - At 5 pages per document, BDA's integrated approach provides better value
- **Textract Forms analysis is the most expensive component**
  - \$1.2M monthly for forms extraction alone
  - \$50 per 1K pages vs BDA's integrated \$40 per page
- **Comprehend costs are minimal (\$3,840) compared to document processing**
- **Development complexity is higher** with multiple services vs single BDA solution
  - Textract requires separate API calls for OCR, forms, and tables
  - Comprehend requires additional processing after Textract extraction
  - BDA provides integrated processing in a single service
- **Additional considerations:**
  - Textract doesn't include built-in PII/PHI redaction (requires Comprehend)
  - BDA includes multimodal processing (images, tables, text) in single pass
  - Textract + Comprehend requires custom orchestration logic
  - BDA provides better integration with Knowledge Bases for medical data

---

## Virtual Assistant Component

- 5 interactions per appointment
- Conversion rate (lorem ipsum and bedrock token counter as references)
  - 140 tokens = 446 characters = 1 paragraph = 0.446 text units
- 3 paragraphs of medic input per appointment
  - 3 paragraphs × 0.446 text units = 1.338 text units per appointment
  - 480 appointments × 1.338 = 642.24 text units per medic monthly

- 420 tokens = 1,338 characters per appointment
- 201,600 tokens = 642,240 characters per medic monthly
- 10 paragraphs of assistant output per appointment
  - 10 paragraphs × 0.446 text units = 4.46 text units per appointment
  - 480 appointments × 4.46 = 2,140.8 text units per medic monthly
  - 1,400 tokens = 4,460 characters per appointment
  - 672,000 tokens = 2,140,800 characters per medic monthly
- Total per appointment:  $1.338 + 4.46 = 5.798$  text units
- Total per medic:  $642.24 + 2,140.8 = 2,783.04$  text units

## Amazon Bedrock Guardrails

Item	Usage	Monthly Cost
GuardRails Evaluation	27.83M units	\$9,740.64
<b>Total</b>		<b>\$9,740.64</b>
<b>Total per medic</b>		<b>\$0.97</b>
<b>Total per appointment</b>		<b>\$0.00203</b>

### Detail

- Guardrails process all interactions for multiple checks:
  - Contextual Grounding (\$0.10 per 1K text units)
  - Sensitive Information PII (\$0.10 per 1K text units)
  - Content Policy (\$0.15 per 1K text units)
- Text units: Input + Output =  $2,783.04$  text units per medic × 10K medics = 27.83M text units monthly
- Regex patterns (6 LATAM patterns) are FREE

### Calculations

- Text Units Processed:  $642.24$  input +  $2,140.8$  output =  $2,783.04$  text units per medic per month
  - Contextual Grounding:  $\$0.10$  per 1K text units ×  $2.78304$  =  $\$0.278304$
  - Sensitive Information PII:  $\$0.10$  per 1K text units ×  $2.78304$  =  $\$0.278304$
  - Content Policy:  $\$0.15$  per 1K text units ×  $2.78304$  =  $\$0.417456$
  - Total:  $\$0.974064$  per medic per month × 10K =  $\$9,740.64$
  - Total per appointment:  $\$9,740.64 \div 4.8M = \$0.00203$

## Amazon Bedrock Model Calls

Item	Usage	Monthly Cost
Input Tokens	2.016B tokens	\$1612
Output Tokens	6.72B tokens	\$21504
<b>Total</b>		<b>\$23116</b>
<b>Total per medic</b>		<b>\$2.3116</b>

Item	Usage	Monthly Cost
<b>Total per appointment</b>		<b>\$0.00481</b>

#### Detail

- Input: 420 tokens per appointment  $\times$  4.8M appointments = 2016000000 tokens
- Output: 1,400 tokens per appointment  $\times$  4.8M appointments = 6720000000 tokens
- No caching in estimate

#### Calculations

##### On Amazon Nova Pro:

- Input Tokens:  $\$0.0008 \times (2016000000 \div 1000) = \$1612$
- Output Tokens:  $\$0.0032 \times (6720000000 \div 1000) = \$21504$
- **Total: \$23116**

##### On Claude 4.5 Haiku:

- Input Tokens:  $\$0.001 \times (2016000000 \div 1000) = \$2016$
- Output Tokens:  $\$0.005 \times (6720000000 \div 1000) = \$33600$
- **Total: \$35616, \$3.5616 per medic, \$0.00742 per appointment**

##### On Amazon Nova Lite

- Input Tokens:  $\$0.00006 \times (2016000000 \div 1000) = \$120.96$
- Output Tokens:  $\$0.00024 \times (6720000000 \div 1000) = \$1612.8$
- **Total: \$1733.76, \$0.173376 per medic, \$0.0003612 per appointment**
- *calculator shows a more expensive estimate due to accounting for weekend usage*

#### Amazon Bedrock AgentCore

Item	Usage	Monthly Cost
AgentCore Runtime	0.1vCPU + 0.52 GB	\$8427.2
AgentCore Gateway	6 tools	\$135.0012
AgentCore Identity	IAM Auth	\$0
<b>Total</b>	<b>\$8562.2012</b>	
<b>Total per medic</b>	<b>\$0.85622012</b>	
<b>Total per appointment</b>	<b>\$0.00178379191</b>	

#### Detail

- Agentcore Runtime: session terminates on 15 minutes of inactivity or after 8 hours of session, appointments last 20 minutes with 5 interactions each so over a workday per medic:

- 5 interactions × 3 appointments per hour per medic × 8 hours of work = 120 interactions in 8 hours per medic = 4 minutes in between interactions on average = NOT enough inactivity for a session reset
- 1 session per day per medic × 10K medics × 20 days worked monthly = 200000 sessions per month
- 95% I/O Wait time
- 28800 seconds of session duration
- 0.1 vCPU (average from observability and cloudwatch metrics)
- 0.51 GB Memory usage (average from observability and cloudwatch metrics)

## Calculations

- session = 1 medic day × 10K medics × 20 days worked = 200000 sessions per month
- 28800 seconds per session - 95% response waiting = 1440s = 0.4 Hr effective utilization
- vCpu =  $0.1 \times 0.4 \text{ Hours} = 0.04 \text{ vCPU-H} \times \$0.0895 \text{ per vCPU-H} = \$0.00358 \text{ per session}$
- Memory =  $0.51 \text{ GB} \times 8 \text{ Hours} = 4.08 \text{ GB-H} \times \$0.00945 \text{ per GB-H} = \$0.038556 \text{ per session}$
- Total Runtime Cost:  $\$0.00358 + \$0.038556 \times 200\text{K sessions per month} = \$8427.2$
- 3 tools searches per session (1 per agent)
  - 600K searches per month × \$0.025 per 1K searches = \$15
- 6 tools × \$0.0002 per indexed tool = \$0.0012
- 2.4K tool invocations per medic monthly × 10K medics × \$0.005 per 1K invocations = \$120
- Total Gateway Cost: \$135.0012

~~one event created per interaction 120 interactions per session × 200K sessions × \$0.25 per 1K sessions = \$6000 \* Calculator shows a more expensive cost since it auto-accounts long term storage~~

- No memory usage due to knowledge base

## Amazon Bedrock Knowledge Bases

Item	Usage	Monthly Cost
Vector Queries	192M tokens	\$3.84
Ingestion	24MM tokens	\$480
<b>Total</b>		<b>\$483.84</b>
<b>Total per medic</b>		<b>\$0.048384</b>
<b>Total per document</b>		<b>\$0.0001008</b>

## Detail

- Only cost is vectorization of queries (model usage), vectorization of documents (model usage) and storage (included in RDS).
- For queries:
  - 2 of every 5 interactions triggers a KB Query =  $2 \text{ per appointment} \times 480 \text{ per medic monthly} = 960 \text{ queries per medic monthly} \times 10K \text{ medics} = 9.6M \text{ monthly}$
  - token usage for queries is very short since it's done by the agent: 20 tokens per query is 192M monthly token usage for query embeddings.
- For documents, every document is ingested from the data automation outputs:
  - $4.8M \text{ documents} \times 5 \text{ data automation outputs} = 24M \times 1000 \text{ Tokens avg per document} = 24000M \text{ tokens total}$
- Alternatively: not using data automation yields:
  - $4.8M \text{ documents} \times 1000 \text{ tokens per doc average} = 4800M \text{ tokens total}$

## Calculations

- Storage: storage in rds aurora as a vector store, included in RDS estimate = \$0
- Vector Queries: embedding calls with Amazon Titan Text Embeddings V2
  - $\$0.00002 \text{ per 1000 tokens} \times 192M \text{ tokens} / 1000 \text{ per 1000 tokens} = \$0.00002 \times 192000 = 3.84 \text{ USD}$
- Document ingestion: embedding calls with Amazon Titan Text Embeddings V2
  - Ingestion:  $\$0.00002 \text{ per 1000 tokens} \times 24000M \text{ tokens} / 1000 \text{ per 1000 tokens} = \$0.00002 \times 24M = \$480$ 
    - No additional multimodal storage or custom parsing requirements = \$0
  - Alternative:  $\$0.00002 \text{ per 1000 tokens} \times 4800M \text{ tokens} / 1000 \text{ per 1000 tokens} = \$0.00002 \times 4.8M = \$96$  at less precision
    - may require custom parsing (by LLM or data automation), limited multimodality, custom PII/PHI redaction detection and implementation
      - Custom parsing by Amazon Nova pro would be  $4800M \text{ Tokens} \times \$0.0008 \text{ per 1000 input tokens} + 4800M \text{ Tokens} \times \$0.0032 \text{ per 1000 output tokens} = \$19200$
- Total is queries + ingestion =  $\$480 + \$3.84 = \$483.84$

## Virtual Assistant Cost Summary

Item	Monthly Cost
Amazon Bedrock Guardrails	\$9,740.64
Amazon Bedrock Model Calls (Nova Pro)	\$23,116.00
Amazon Bedrock AgentCore	\$8,562.20
Amazon Bedrock Knowledge Bases	\$483.84
<b>Total</b>	<b>\$41,902.68</b>
<b>Total per medic</b>	<b>\$4.190268</b>
<b>Total per appointment</b>	<b>\$0.00872972</b>

## Virtual Assistant Alternative Scenarios

Scenario	Monthly Cost	Difference	Notes
Claude 4.5 Haiku Model	\$55,518.68	+\$13,616.00	\$35,616 for model vs \$23,116 for Nova Pro
Amazon Nova Lite Model	\$20,519.68	-\$21,383.00	\$1,733.76 for model vs \$23,116 for Nova Pro
Without BDA (Manual KB)	\$41,518.68	-\$384.00	4.8B tokens (\$96) vs 24B tokens (\$480) for ingestion

## Other Services

Service	Usage	Monthly Cost
AWS CDK	Infrastructure as Code	\$0.00
CloudFormation	Stack operations	\$0.00
AWS KMS	AWS managed keys (free tier)	\$0.00
SSM Parameter Store	Standard parameters	\$0.00
AWS IAM	Identity management	\$0.00
CloudWatch Logs	Default Lambda logging (free tier)	\$0.00
CloudWatch Metrics	Default AWS metrics (free tier)	\$0.00
Secrets Manager	2 database secrets	\$0.80
<b>Total</b>		<b>\$0.80</b>
<b>Total per medic</b>		<b>\$0.000008</b>
<b>Total per appointment</b>		<b>\$0.00000017</b>

### Detail

- **Secrets Manager:** 2 database credential secrets (RDS Aurora connection strings)
  - 2 secrets × \$0.40 = \$0.80
  - API calls: Cached by Lambda, minimal retrieval (within free tier of 10K calls)
- **CloudWatch:** Using only default/included monitoring
  - Lambda logs: Within free tier (5GB ingestion, 5GB storage, sufficient retention policy for initial phase and keep on free tier)
  - Metrics: Only default AWS service metrics (free)
  - No custom metrics or log insights queries
- **CloudTrail:** Management events only (free), no S3 data events enabled
- **EventBridge:** AWS-to-AWS events in same account (free)

## Other Services Alternative Scenarios

Category	Scenario	Monthly Cost	Difference	Notes
Frontend	SAML/OIDC Authentication (Cognito)	+\$149.25	+\$149.25	Enterprise SSO instead of standard auth
Frontend	AWS WAF with Amplify	+\$15.00	+\$15.00	Basic firewall protection (excludes usage charges)
Frontend	Without Free Tier (Amplify)	+\$2.25	+\$2.25	Storage + 15GB data transfer
Backend	REST API instead of HTTP API	+\$717.50	+\$717.50	For baseline infrastructure only
Document Workflow	BDA Standard Output	\$240,518.38	-\$720,000.00	\$0.01/page vs \$0.04/page
Document Workflow	Textract + Comprehend (5 pages)	\$1,600,358.38	+\$639,840.00	Full processing, 40% more expensive than BDA
Document Workflow	Without BDA (Manual KB)	\$614.38	-\$959,904.00	Custom parsing required
Document Workflow	Custom Parsing (Nova Pro)	\$20,014.38	-\$940,504.00	LLM-based document parsing
Document Workflow	Manual KB (4.8B tokens)	\$96.00	-\$96,000.00	Reduced ingestion tokens
Virtual Assistant	Claude 4.5 Haiku Model	\$55,518.68	+\$13,616.00	Alternative to Nova Pro
Virtual Assistant	Amazon Nova Lite Model	\$20,519.68	-\$21,383.00	Cheaper model option
Virtual Assistant	Without BDA KB Ingestion	\$41,518.68	-\$384.00	4.8B vs 24B tokens
Monitoring	CloudWatch Custom Metrics	+\$3.00	+\$3.00	10 custom application metrics
Monitoring	CloudTrail S3 Data Events	+\$19.20	+\$19.20	19.2M S3 data event logging
Monitoring	CloudWatch Logs (beyond free)	+\$0.02	+\$0.02	Additional log ingestion/storage

## Cost Summary

Total Monthly Costs

Component	Current Scale (10K)	Per Medic	Per Appointment
-----------	---------------------	-----------	-----------------

<b>Component</b>	<b>Current Scale (10K)</b>	<b>Per Medic</b>	<b>Per Appointment</b>
<b>Frontend</b>	\$41.69	\$0.004169	\$0.00000868
<b>Backend</b>	\$124.44	\$0.012444	\$0.00002592
<b>Document Workflow</b>	\$960,518.38	\$96.051838	\$0.20010799
<b>Virtual Assistant</b>	\$41,902.68	\$4.190268	\$0.00872972
<b>Other Services</b>	\$0.80	\$0.00008	\$0.00000017
<b>TOTAL</b>	<b>\$1,002,587.99</b>	<b>\$100.26</b>	<b>\$0.20887</b>

#### Baseline Costs (Existing Infrastructure)

<b>Component</b>	<b>Monthly Cost</b>	<b>Notes</b>
Backend Baseline	\$961.73	30K RPM capacity for other applications
Document Storage Baseline	\$345.00	15TB pre-existing medical documents
<b>Baseline Total</b>	<b>\$1,306.73</b>	Not included in solution cost

#### Key Assumptions

- Number of doctors: 10,000
- All doctors are daily active users
- Single page visits per session
- Weekly deployment cycles
- Standard authentication (not SAML/OIDC)