ISSN : 2455-135X \

Machine Learning for Smart Grid forecasting and Energy Optimisation

Ishita Bahamnia
bahamniaishita@gmail.com

Part 1: Complete System Architecture

The architecture follows a microservices-based, cloud-native approach to ensure scalability, maintainability, and independent deployment of modules.

1. Presentation Layer (Client)

- Web Application: React.js or Next.js (for SEO and server-side rendering of public pages).
- Mobile Application: React Native (for iOS and Android) to maintain a single codebase.
- Chrome Extension: Optional, for lawyers to use the tool while browsing court websites.

2. API Gateway & Load Balancing

- Gateway: AWS API Gateway or Kong.
- Purpose: Manages authentication (JWT/OAuth2), rate limiting, request routing, and SSL termination.
- Load Balancer: AWS Application Load Balancer (ALB) to distribute traffic across microservices.

3. Microservices (Business Logic)

This is the core of the system. Each module in your overview becomes a microservice:

- User Service: Registration, firm management, subscription plans.
- AI Orchestration Service: Manages prompts, routes to LLMs, and handles RAG logic.
- Drafting Service: Handles the workflow for generating legal documents (Civil, Criminal, Commercial).
- Property Due Diligence Service: OCR processing, title chain analysis, and report generation.

ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

- Crawler Service: Manages scheduled scrapers for court websites (handles retries, IP rotation, CAPTCHA solving).
- Search Service: Abstracts the search logic (legal, case law, documents).
- Payment Service: Manages subscriptions (Razorpay/Stripe).
- Notification Service: Email, SMS, and in-app notifications.

4. Data & AI Layer

- Relational Database (SQL): PostgreSQL (hosted on AWS RDS).
  - *Stores:* Users, firms, billing info, user-generated drafts, metadata.
- Vector Database: Pinecone (managed) or Weaviate (self-hosted on EKS).
  - *Stores:* Embeddings of judgments, acts, and clauses for semantic search.
- Search Engine: Elasticsearch (hosted on AWS OpenSearch Service).
  - *Stores:* Indexed court data (case status, cause lists) for fast keyword search.
- Object Storage: AWS S3.
  - *Stores:* Raw PDFs (judgments, uploaded deeds), parsed JSON data, crawled court PDFs.
- Caching: Redis (AWS ElastiCache).
  - *Stores:* Session data, frequently searched case laws, LLM response caching to reduce costs.

5. AI & ML Pipeline

- LLM Layer: Use a mix of models.
  1. *Primary:* GPT-4o (for complex reasoning and drafting) via Azure OpenAI (for enterprise compliance).
  2. *Fine-tuned/Open Source:* Llama 3 (for high-volume, low-cost inference like data extraction from PDFs).
- RAG Pipeline:
  1. Query comes in.
  2. Convert query to vector.
  3. Retrieve relevant context from Pinecone (legal embeddings).
  4. Enrich with metadata (Section, Act, Date) from PostgreSQL/Elasticsearch.
  5. Send context + prompt to LLM.

ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

- OCR & Parsing: Azure Document Intelligence (formerly Form Recognizer) or AWS Textract. These are superior to Tesseract for complex legal PDF layouts.
- Embedding Model: `text-embedding-3-large` (OpenAI) or `all-MiniLM-L6-v2` (for self-hosted).

6. DevOps & Infrastructure

- Containerization: Docker.
- Orchestration: AWS EKS (Kubernetes) or AWS ECS (simpler).
- CI/CD: GitHub Actions or GitLab CI.
- Monitoring: Prometheus + Grafana, AWS CloudWatch.

Part 2: Tech Stack Summary

Layer	Technology Choice	Rationale
Frontend	React.js / Next.js, React Native	Cross-platform consistency, large ecosystem for UI components (Ant Design for admin panels).
Backend	Python (FastAPI)	Best ecosystem for AI/ML integration, async support for crawlers, easy to maintain.
Database	PostgreSQL, Elasticsearch, Pinecone	Covers relational data, search, and vector similarity respectively.

ISSN : 2455-135X \

Cloud	AWS (or Azure)	AWS has strong AI services (Bedrock, Textract). Azure is strong for OpenAI compliance.
AI/LLM	GPT-4o (Orchestrator), Llama 3 (Fine-tuned), Claude (Legal Analysis)	Hybrid approach: High-quality reasoning for drafting, fine-tuned for specific legal tasks.
Crawling	Scrapy (Python), Puppeteer (JS)	Scrapy for static pages, Puppeteer for dynamic JS-heavy court sites.
Storage	AWS S3, CloudFront CDN	For storing PDFs, judgments, and serving static assets quickly.

Part 3: Approximate Development Cost

Estimating costs for a project of this scale is complex and depends heavily on location (hiring in India vs. US), seniority of engineers, and whether you build the AI from scratch or use APIs.

Assumptions:

- Development team based in India (Hybrid/Remote).
- Duration: 12-18 months for a full-feature launch (V1 + V2 modules).
- Scope: Building all modules listed, including the custom crawlers and RAG pipeline.

A. Team Composition & Monthly Burn

ISSN : 2455-135X \

Role	Seniority	Count	Monthly Salary (INR)	Monthly Salary (USD)	Rationale
Project Manager	Senior	1	₹1,50,000	$1,800	Managing legal domain complexities.
Frontend Dev	Senior	2	₹1,20,000 x2	$2,900	Web + React Native.
Backend Dev	Senior	3	₹1,50,000 x3	$5,400	Microservices, API, Integration.
AI/ML Engineer	Senior	2	₹1,80,000 x2	$4,300	RAG pipeline, Fine-tuning, Prompt engineering.
Data Engineer	Mid	2	₹1,00,000 x2	$2,400	Crawlers, ETL pipelines, Data cleaning.
DevOps	Senior	1	₹1,50,000	$1,800	Kubernetes, CI/CD, Cloud infra.
QA Engineer	Mid	1	₹80,000	$960	Automated testing, security testing.

ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

UI/UX Designer	Senior	1	₹1,00,000	$1,200	Complex legal workflow design.
Total Monthly		13	₹17,30,000	~$20,800	

Total Development Cost (12 Months): ₹2.07 Crores (~$250,000 USD)

Note: This is the *burn rate*. The actual capitalization cost will be higher due to:

- Infrastructure (Cloud): ₹2-3 Lakhs/month during development for dev/staging environments, databases, and LLM API keys.
- Data Acquisition: Legal databases (Manupatra/SCC Online) may require licensing for commercial use (if not scraping public sites).
- Contingency: 15-20% buffer.

B. Module-wise Cost Distribution (Approx.)

If you want to go MVP first (as suggested in your document), you can reduce the initial scope:

1. MVP (Months 1-6):
    - Core: User Auth, AI Orchestration, Basic RAG (Case Law Search), Simple Drafting Engine, Property Due Diligence (OCR).
    - Team: Reduce to 8 core members.
    - Cost: ₹80 Lakhs – ₹1 Crore ($100k - $120k).
2. Full Platform (Months 7-12):
    - Add: Crawlers (Court data), Tax/Corporate Engine, Evidence Builder, Advanced Drafting Templates, Mobile Apps.
    - Cost: ₹1.2 – ₹1.5 Crores ($145k - $180k).

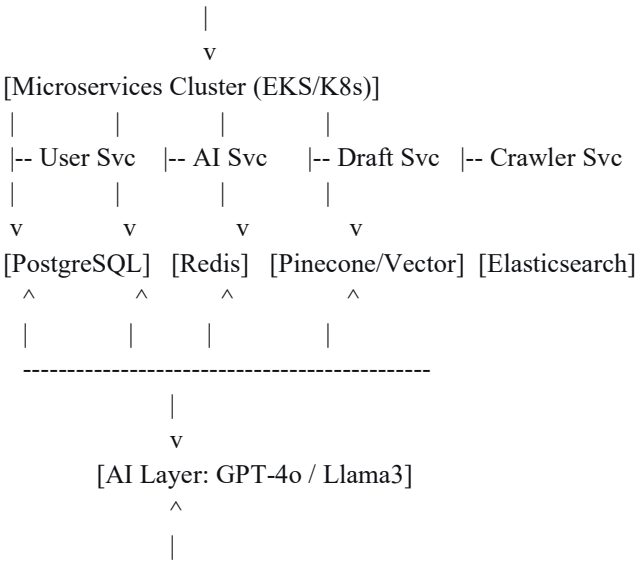
Part 4: Critical Considerations for Cost Optimization

ISSN : 2455-135X \

1. Hybrid AI Model:
    - Don't: Use GPT-4 for every single request (e.g., parsing a PDF). It will burn cash.
    - Do: Use fine-tuned smaller models (e.g., Llama 3-8B fine-tuned on Indian judgments) for extraction and classification tasks. Use GPT-4o only for final drafting and complex reasoning.
2. Crawler Maintenance:
    - Indian court websites change frequently. Dedicate 1 engineer permanently to "Data Engineering & Maintenance." Court crawlers are not a "build once" feature; they require constant updates.
3. Legal Accuracy (The "Hallucination" Risk):
    - Cost: You must invest in a "Citation Verification" microservice. After the LLM generates a draft citing "Section 420 IPC" or "2023 SC 123," the system must query the actual database to verify that case exists and the citation is correct. This adds development complexity but is non-negotiable for trust.
4. Cloud vs. On-Premise:
    - For vector databases and embedding models, consider running them on your own GPU instances (AWS EC2 G4dn instances) rather than managed Pinecone if you have high volume, to save costs in the long run. However, managed services are faster for launch.

5. Summary of Architecture Diagram (Text Representation)

text



ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

[OCR: AWS Textract / Azure AI Doc Intel]

Conclusion

Building a comprehensive AI Legal Chamber Platform as described is a $250,000–$300,000 project over 12-14 months if built in India, or double that in the US/EU.

Recommended First Step:
Start with the MVP focused on RAG-based Case Law Search + Property Due Diligence. Property due diligence (reading sale deeds, generating reports) has immediate monetization potential and solves a very specific pain point for both lawyers and real estate firms, allowing you to generate revenue while building the more complex drafting modules.

This response is AI-generated, for reference only.

where to deploy

Deployment Strategy for AI Legal Chamber Platform

Given the complexity, data sensitivity, and AI-heavy nature of your platform, I'll provide a comprehensive deployment strategy with multiple options based on your priorities (cost, compliance, scalability).

Option 1: AWS (Recommended - Best Balance)

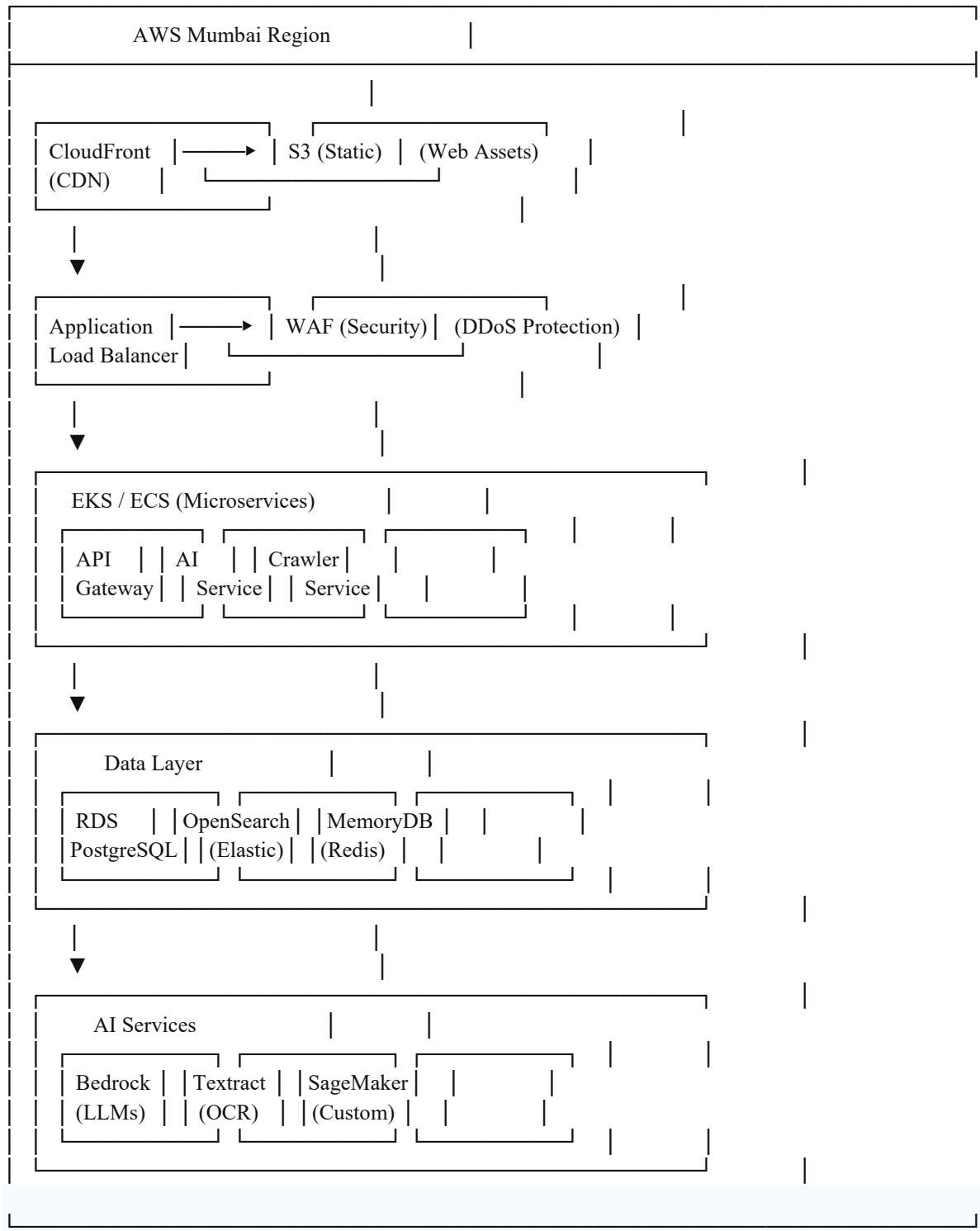
Why AWS is Ideal for Legal AI Platforms:

- AI Services Maturity: Bedrock (LLMs), Textract (Legal OCR), Comprehend (NLP)
- Compliance: HIPAA eligible, ISO 27001, essential for legal data
- Indian Presence: AWS Mumbai region (ap-south-1) ensures data localization

ISSN : 2455-135X \

Deployment Architecture on AWS:

text



ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

AWS Monthly Cost Breakdown (Production):

Service	Configuration	Monthly Cost (USD)
Compute (EKS)	5-8 nodes (t3.xlarge)	$800 - $1,500
RDS PostgreSQL	db.r6g.xlarge (Multi-AZ)	$400 - $600
OpenSearch	3 nodes (r6g.xlarge)	$600 - $900
Redis (MemoryDB)	Cluster with shards	$300 - $500
S3 + CloudFront	Storage + CDN	$200 - $400
AI Services		
- Bedrock (LLM)	GPT-4o / Claude usage	$1,000 - $5,000*
- Textract (OCR)	10,000 pages/day	$1,500 - $3,000
- SageMaker	Custom models	$500 - $1,500
Vector DB	Pinecone (managed) or PGVector	$400 - $1,000

ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

Networking	NAT Gateway, Data Transfer	$200 - $400
Backup & Security	Snapshots, WAF, GuardDuty	$300 - $600
Total	Monthly Recurring	$6,200 - $15,000

LLM costs vary significantly based on usage volume

Option 2: Azure (Best for Enterprise & OpenAI)

Why Choose Azure:

- Preferred OpenAI Partner: Native GPT-4 with enterprise compliance
- Strong India Presence: Central India (Pune) and South India (Chennai) regions
- Legal Industry Trust: Many large law firms use Microsoft stack

Azure-Specific Services:

- Azure OpenAI Service: For GPT-4, fine-tuning capabilities
- AI Document Intelligence: For legal document OCR (better than AWS for complex Indian legal formats)
- Azure Kubernetes Service (AKS): Managed K8s
- Azure SQL/PostgreSQL: Managed databases

Azure Cost Estimate:

- Similar to AWS base infrastructure ($5,000 - $8,000)

ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

- Azure OpenAI is slightly more expensive but offers better compliance
- Total: $7,000 - $18,000/month

Option 3: Hybrid Cloud (For Data Sovereignty)

When to Choose Hybrid:

- If you need to keep sensitive legal data on-premise
- For high-volume document processing (to avoid cloud egress costs)
- If you plan to fine-tune large models on sensitive data

Architecture:

```
text

[On-Premise / Colocation]
├── GPU Servers (for fine-tuned Llama models)
├── Vector Database (Qdrant/Weaviate)
├── Crawler Infrastructure
└── Document Storage (MinIO)

[Cloud (AWS/Azure)]
├── Web Frontend
├── User Management
├── Public API Gateway
└── Backup & DR
```

Cost Structure:

- CapEx: $50,000 - $100,000 for GPU servers (4x NVIDIA A100)
- Colocation: $1,000 - $3,000/month per rack
- Cloud: $2,000 - $4,000/month (reduced workload)
- Total Monthly: $3,000 - $7,000 + amortized hardware

ISSN : 2455-135X \

Option 4: Specialized Indian Cloud (For Cost Optimization)

Providers:

- Netmagic (NTT): Indian data centers, good for compliance
- CtrlS: Tier-4 data centers in multiple Indian cities
- DigitalOcean: Simple, cost-effective for early stage

Cost Comparison:

Provider	Monthly Base Infra	AI Services	Total
DigitalOcean	$2,000 - $4,000	Self-managed LLMs	$4,000 - $8,000
Netmagic	$3,000 - $5,000	Self-managed	$5,000 - $10,000
AWS/Azure	$5,000 - $8,000	Managed AI services	$6,000 - $15,000

Recommended Deployment Strategy (Phased)

Phase 1: MVP Launch (Months 1-3)

Target: 100-200 beta users

ISSN : 2455-135X \

Component	Deployment Choice	Monthly Cost
Frontend	Vercel/Netlify	$0 - $100
Backend	DigitalOcean Kubernetes	$200 - $400
Database	DigitalOcean Managed PostgreSQL	$50 - $100
Vector DB	Pinecone (Starter)	$100 - $300
LLM	OpenAI API (pay-per-use)	$500 - $1,000
Storage	DigitalOcean Spaces	$50
Total		$900 - $2,000

Phase 2: Growth Stage (Months 4-8)

Target: 1,000+ users, enterprise clients

Component	Deployment Choice	Monthly Cost
Infrastructure	AWS Mumbai (EKS)	$3,000 - $5,000

ISSN : 2455-135X \

AI Services	Bedrock + Textract	$2,000 - $4,000
Databases	RDS + OpenSearch	$1,500 - $2,500
CDN & Security	CloudFront + WAF	$500 - $1,000
Total		$7,000 - $12,500

Phase 3: Enterprise Scale (Month 9+)

Target: 10,000+ users, law firms, corporate clients

Component	Deployment Choice	Monthly Cost
Infrastructure	AWS Multi-Region (Mumbai + Singapore)	$8,000 - $12,000
AI Services	Mix of Bedrock + Fine-tuned Models on SageMaker	$5,000 - $10,000
Compliance	HIPAA/ISO certified setup	$2,000 - $3,000
Support	24/7 Enterprise Support	$1,500 - $3,000
Total		$16,500 - $28,000

ISSN : 2455-135X \

Critical Deployment Decisions

1. Data Residency & Compliance

- Requirement: Indian legal data must often stay within India
- Solution: Use AWS Mumbai or Azure Central India regions
- Cost Premium: ~10-15% over US regions

2. LLM Deployment Strategy

Option	Pros	Cons	Cost Impact
OpenAI API	Best quality, easy setup	Data privacy concerns, high per-token cost	High variable cost
AWS Bedrock	Enterprise-ready, data privacy	Limited model selection	Medium variable
Fine-tuned Llama on GPU	Full control, lower long-term cost	Requires ML expertise, CapEx	High fixed, low variable
Mix (Recommended)	Best of both worlds	Complex orchestration	Balanced

Recommended Mix:

ISSN : 2455-135X \ © 2025 International Scientific and Academic Research (ISAR) Publisher

- Drafting/Reasoning: OpenAI GPT-4o or Claude (via Bedrock)
- Extraction/Parsing: Fine-tuned Llama 3 (on GPU instances)
- Embeddings: Open-source models on your infrastructure

3. Database Location Strategy

```
text

Hot Data (Active cases, Users)    → Managed Cloud (RDS)
Warm Data (Recent judgments)      → Vector DB (Pinecone/Weaviate)

Cold Data (Archived cases, PDFs)  → S3 Glacier (Deep Archive: $1/TB/month)
```

4. Disaster Recovery

- RPO (Recovery Point Objective): 15 minutes
- RTO (Recovery Time Objective): 4 hours
- Setup: Multi-AZ within Mumbai region + cross-region backup to Singapore
- Cost: Additional 20-30% of base infrastructure

Specific Recommendations

For Your Legal AI Platform:

1. Start with AWS Mumbai (Phase 1-2)
    - Best balance of services, compliance, and Indian presence
    - Use AWS Bedrock for LLMs to maintain data privacy
    - Leverage AWS Textract for legal document OCR
2. Implement Cost Controls Early
    - Set up budget alerts at $5,000, $10,000, $15,000
    - Use Lambda functions for auto-scaling based on demand

ISSN : 2455-135X \

- Implement Redis caching for LLM responses (reduces API calls by 40-60%)

3. Security-First Deployment
   - Deploy within VPC with no public access to databases
   - Use AWS WAF with custom rules for Indian legal traffic patterns
   - Enable AWS GuardDuty for threat detection
4. Monitor These Key Metrics
   - LLM token usage (biggest cost driver)
   - OCR page processing volume
   - Crawler bandwidth usage
   - Database connection pools

Summary: Where to Deploy

Stage	Recommended Platform	Monthly Cost Range	Timeline
MVP/Prototype	DigitalOcean + OpenAI API	$1,000 - $2,500	Months 1-3
Beta/Growth	AWS Mumbai (EKS + Bedrock)	$5,000 - $10,000	Months 4-8
Production	AWS Mumbai (Multi-AZ)	$10,000 - $20,000	Months 9+
Enterprise	AWS Multi-Region + Hybrid	$20,000 - $40,000	Year 2+

Final Recommendation:

ISSN : 2455-135X \

Deploy on AWS Mumbai (ap-south-1) for production, starting with a moderate setup ($5,000-7,000/month) and scaling based on user adoption. This gives you:

- Data localization compliance
- Access to all required AI services (Bedrock, Textract, SageMaker)
- Ability to scale to enterprise requirements
- Cost predictability with reserved instances (savings of 30-40% on compute)