**Section 1: RAG System Design Complete Solution**

**1.1 RAG Architecture Diagram:**

Document Sources → Ingestion Pipeline → Vector DB → RAG Pipeline → User Interface

↓ ↓ ↓ ↓ ↓

PDF/Word/PPT → Text Extraction → Pinecone → Retrieval → Web Chat

Emails/Reports → Chunking → → Generation → Mobile App

→ Embedding → → → Source Citations

**1.2 Technology Stack (AWS):**

Cloud Provider: AWS (US Regions)

Compute:

- AWS Lambda (serverless functions)

- ECS Fargate (persistent services)

Storage:

- Amazon S3 (raw documents)

- Amazon RDS PostgreSQL (metadata)

Vector Database: Pinecone Serverless

Search: Amazon OpenSearch (with k-NN)

LLM: Anthropic Claude-3-Haiku via AWS Bedrock

Authentication: AWS Cognito

API: Amazon API Gateway

CDN: Amazon CloudFront

**1.3 Security Architecture:**

User → AWS Cognito → API Gateway → Microservices → Data Layers

↓ ↓ ↓ ↓

SSO/MFA Rate Limiting IAM Roles Encryption

RBAC WAF VPC KMS Encryption

**1.4 Scaling Strategy:**

Phase 1: 500 users, 100K documents

Phase 2: 1000 users, 1M documents (add read replicas)

Phase 3: 5000+ users, 10M+ documents (sharding + CDN)  
  
**1.6 Implementation Phases:**

Phase 1 (Months 1-2):

- Basic PDF ingestion

- Simple vector search

- Web interface

- AWS Cognito auth

Phase 2 (Months 3-4):

- All document types (Word, PPT, Email)

- Source citations

- Mobile app

- Advanced permissions

Phase 3 (Months 5-6):

- Real-time updates

- Advanced analytics

- Performance optimization