Cloud Test – AWS RAG Architecture

Cloud Test – RAG System Design (AWS)

1. Assumptions
- 250k docs; formats: PDF, Word, PPT, Email
- 200 new docs/week; real-time ingestion
- 500 employees; 500 concurrent users peak
- US-only data residency; strict ACL
- AWS familiarity: S3, IAM, Lambda. New to Bedrock/OpenSearch.
- LLM: Bedrock Claude 3 + Titan Embeddings

2. High-Level Architecture
[ASCII diagram]
(Document sources → S3 → Lambda → Textract → Bedrock Embeddings → OpenSearch Vector DB → Bedrock LLM

3. Ingestion & Indexing Pipeline
- S3 stores raw docs
- Lambda normalizes formats, extracts text, metadata
- Textract OCR for scanned PDFs
- Chunking (200–500 tokens)
- Bedrock Titan embeds chunks
- OpenSearch Serverless stores vectors + metadata

4. Retrieval & Response Logic
- API Gateway + Cognito auth
- Embed query using Titan
- k-NN search (k=8–12)
- Permission filtering
- Optional re-ranking
- Claude 3 generates answer + citations

5. User Interface & Application Layer
- React web/mobile UI
- CloudFront CDN
- API Gateway as single entry
- Lambda backend for RAG pipeline
- Signed URLs for document previews

6. Security Architecture
[ASCII diagram]
(Cognito → API Gateway+WAF → Lambda → VPC Private Endpoints → OpenSearch+S3+Bedrock → CloudTrail)

Key controls:
- Federation via SAML/OIDC
- JWT-based access
- VPC-only access to Bedrock/OpenSearch
- KMS encryption
- CloudTrail auditing
- IAM least privilege

7. Scaling Strategy
- Serverless auto-scaling (Lambda, API Gateway)
- OpenSearch Serverless autoscaling
- Bedrock handles concurrency
- S3 infinite storage
- Multi-AZ design ensures 99.5%+ uptime

8. Cost Strategy (<$8k/month)
- Bedrock LLM: $3.5k–$4k
- OpenSearch: $1.2k–$1.8k
- Textract: $500–$1k
- S3: $150–$300
- Lambda: $100–$250
- Others: $300–$500
Optimizations:
- Query caching
- Cheaper model tiering (Haiku)
- Smarter ingestion filtering

- Vector compression

9. Risks, Tradeoffs, Alternatives
Risks:
- LLM hallucination
- Permission leaks if metadata incorrect
- Index cost growth
Tradeoffs:
- Serverless over clusters
- AWS-native Bedrock vs external LLMs
- OpenSearch vs Pinecone
Alternatives:
- Azure OpenAI + Cognitive Search
- GCP Document AI + Vertex RAG
- Hybrid search with Kendra