**Scenario**

A mid-size consulting firm (500 employees) wants to build an internal AI-powered knowledge management system. They have 10+ years of project documents, proposals, contracts, and research reports scattered across different systems. They want employees to ask questions in natural language and get accurate answers with source citations.

**Example Use Cases:**

* "What were the key challenges in our healthcare projects from 2022-2023?"
* "Show me all contract templates for government clients"
* "What pricing models did we use for similar AI implementations?"

**Your Task**

Design a secure, cloud-based RAG (Retrieval-Augmented Generation) system that addresses:

**Technical Requirements:**

* **Document Ingestion**: Process PDFs, Word docs, PowerPoints, emails
* **Vector Database**: Store and search document embeddings efficiently
* **RAG Pipeline**: Retrieve relevant context and generate accurate responses
* **User Interface**: Clean, intuitive chat interface (web + mobile)
* **Source Citations**: Always show which documents informed each answer
* **Real-time Updates**: New documents automatically indexed

**Security Requirements:**

* **Company-Only Access**: Only employees can use the system
* **Document Permissions**: Respect existing access controls (some docs are confidential)
* **Audit Trail**: Log all queries and document access
* **Data Residency**: All data must stay in US-based cloud regions

**Business Constraints:**

* Initial budget: $8,000/month cloud spend
* Must handle 500 concurrent users during peak hours
* 99.5% uptime requirement

**Deliverables**

1. **RAG Architecture Diagram** - Show data flow from document ingestion to user response
2. **Technology Stack** - Specific cloud services, vector databases, LLM choices
3. **Security Architecture** **Diagram** - Authentication, authorization, and data protection layers
4. **Scaling Strategy** - How to handle growing document corpus and user base
5. **Cost Strategy** – How can we minimize cost for development
6. **Implementation Phases** - What gets built first vs. later features