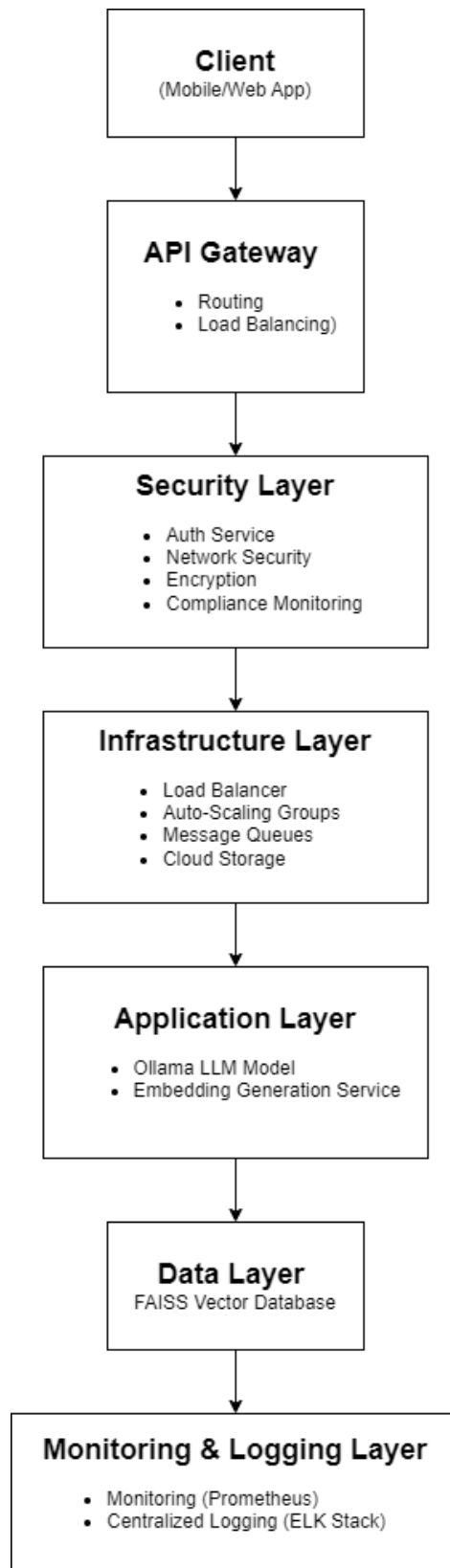


Deployment Diagram:



Explanation:

1. **Clients:** The users interact with the system through web or mobile applications.
2. **API Gateway:** Routes requests from clients to the appropriate services in the application layer and provides load balancing.
3. **Security Layer:**
 - **Auth Service:** Manages authentication and authorization.
 - **Network Security:** Includes firewalls and VPC configurations to protect the system.
 - **Encryption:** Ensures data is encrypted both in transit and at rest.
 - **Compliance Monitoring:** Ensures the system adheres to relevant regulations.
4. **Infrastructure Layer:**
 - **Load Balancer:** Distributes incoming network traffic across multiple servers.
 - **Auto-Scaling Groups:** Automatically adjusts the number of instances in response to the load.
 - **Message Queues:** Manages asynchronous processing tasks.
 - **Cloud Storage:** Stores large datasets and other resources.
5. **Application Layer:**
 - **LLM Model:** Handles natural language processing tasks using a Ollama model.
 - **Embedding Generation Service:** Generates embeddings for the document chunks.
6. **Data Layer:**
 - **Vector Database:** Stores embeddings and facilitates fast similarity searches.
7. **Monitoring & Logging Layer:**
 - **Monitoring:** Tracks system performance and health metrics.
 - **Centralized Logging:** Collects and stores logs for analysis and troubleshooting.