ChatGPT

Enterprise RAG Platform Blueprint: Scaling RAGitify.ai for the Enterprise

---

Vision

Transform a RAG-based application (like RAGitify.ai) into an enterprise-grade platform that provides secure, scalable, and customizable document intelligence through Retrieval-Augmented Generation (RAG).

Core Enterprise Additions

1. Multi-Tenant SaaS Architecture

- Isolated document indices per tenant
- Subdomain routing (e.g., `acme.ragitify.ai`)
- Scoped API tokens and tenant-aware auth sessions

2. Enterprise Security & Compliance

- SOC 2 / HIPAA / GDPR readiness
- SAML 2.0 / SSO integration
- SCIM for user provisioning
- Role-Based Access Control (RBAC)
- Audit logging & anomaly detection

3. Deployment Flexibility

- Public cloud SaaS
- Self-hosted (Helm/Terraform packages)
- Private VPC support (AWS, Azure, GCP)

RAG Engine Enhancements

Advanced Retrieval Controls

- Dense + sparse hybrid retrieval
- Metadata filtering (e.g., `{doc_type: policy}`)
- Query rewriting & semantic re-ranking

Ingestion Framework

- Auto-chunking & version control
- Scheduled syncing (webhooks, cron)

- OCR + layout-preserving embedding (for scanned docs)

Citation & Transparency

- Answer-level citations with in-doc highlighting
- Confidence scores & feedback collection

Knowledge Management

Data Connectors

- Google Drive, SharePoint, Confluence, Notion
- Salesforce, HubSpot, GitHub, Box
- SQL/NoSQL databases, REST APIs

Governance & Compliance

- Document lifecycle (draft → review → published)
- Embedding expiration / reindexing rules
- Legal hold & access logging for compliance

Enterprise Collaboration Layer

User Collaboration

- Shared conversations
- Team bookmarks
- Context-aware chat history

Admin Tools

- User access control
- Heatmaps & usage insights
- Retrieval quality dashboards

Customization & Extensibility

Model/Embedding Flexibility

- Bring Your Own LLM (OpenAI, Claude, Mistral)
- Bring Your Own Embeddings (OpenEmbed, HuggingFace)
- Switchable retrieval strategy per workspace

Plugin Framework

- Actions (trigger Slack alerts, file tickets, draft emails)
- Domain agents (LegalBot, ComplianceCoach, etc.)
- REST/GraphQL SDKs for third-party app integration

📊 Analytics & Monitoring

Usage Intelligence

- Query volume, success, fallback rates
- Top documents & terms per workspace
- User engagement dashboards

Retrieval Quality Control

- Feedback loops for tuning retrievers
- Drift detection for embeddings & documents

Bonus Differentiators

Feature	Description
Memory Graphs	Concept maps between documents and answers
Auto-RAG Feedback Tuner	User thumbs-up/down guides retrieval tuning
Multilingual RAG	Indexing & querying across languages
Query Templates	Industry-specific retrieval templates (legal, health, finance)

📈 Monetization Tiers

Plan	Features
Team	10 users, SaaS-only, limited connectors
Business	Full connectors, usage analytics, RBAC
Enterprise	Self-hosted, BYO LLM, custom SLAs, plugin/agent support

---

Example Tech Stack

Frontend: Angular

Backend: Django (Python), PostgreSQL

RAG Pipeline: - LangChain / LlamaIndex / Haystack - FAISS / Weaviate / Qdrant for vector search - OpenAI / Azure OpenAI / Anthropic Claude

Infra: - Docker, Kubernetes, Prometheus + Grafana - CI/CD with GitHub Actions / GitLab - Vault / SOPS for secrets management

---

First 30-Day Build Plan

Week	Deliverables
1	UI wireframes + ingestion pipeline prototype
2	RAG flow MVP with hybrid retrieval & LLM interface
3	Document citation, metadata filtering, source tracing
4	Multi-tenant setup, SSO auth, usage logging dashboard

🏛 Closing Thought

To become the "Notion for Knowledge + ChatGPT", this platform must blend the elegance of UI, depth of enterprise tooling, and flexibility of AI plumbing.

RAG is not the product. **Knowledge is.**

Let the devs build the pipes, so the users see magic.