# HR Intelligent Chatbot

A RAG-Powered Knowledge Assistant for Safran SED

| | |
|---:|:---|
| **Hackathon** | Think To Deploy — Phase 2 |
| **Client** | Safran SED, Casablanca |
| **Scope** | HR Department — Knowledge Access |
| **Status** | Proof of Concept — Pitched to Safran HR |
| **Year** | 2025 |

# Contents

# Chapter 1

# Executive Summary

**Real-World Project**

This report describes a project developed as part of the *Think To Deploy* hackathon. The solution was designed, built as a Proof of Concept, and **officially pitched to the HR department of Safran SED (Casablanca)**. It addresses a concrete operational problem: employees and HR staff spend significant time manually searching across five disconnected systems to find basic HR information.

**Impact Snapshot**

- Estimated reduction of repetitive HR queries directed at the HR team: **30–50%**
- Daily automated refresh of **5 knowledge systems** via a J-1 Airflow pipeline
- **Zero personal data exposure** at any stage, enforced by irreversible anonymisation before AI processing

The proposed solution is an **HR Intelligent Chatbot** powered by a RAG pipeline. Any authenticated SED employee types a question in natural language and receives an accurate, source-cited answer drawn from official HR documentation — in seconds, 24 hours a day.

The pipeline is fully containerised (Docker), orchestrated (Apache Airflow), and designed to operate within Safran's C2 confidentiality tier with irreversible anonymisation of all personal data before any AI processing takes place.

The solution architecture, anonymisation strategy, and end-to-end RAG pipeline were designed and implemented by the project team during the final phase of the Think To Deploy hackathon — from problem framing through to a working, containerised POC.

# Chapter 2

# Business Context & Problem Statement

## 2.1  Organisational Context

Safran SED Casablanca is an industrial aerospace site with several hundred employees. The HR department manages onboarding, procedural compliance, administrative notes (nominations, information memos), and knowledge distributed across five internal platforms. Finding the right information requires knowing *which* system holds *which* content — a non-trivial task for both newcomers and seasoned staff.

## 2.2  The Five Source Systems

| System | Content | Role |
|--------|---------|------|
| **Prisma** | Process rules, scope definitions | Process governance |
| **Intranet** | Employee handbook, org chart, procedure links | Primary self-service portal |
| **DFS** | Employee directory / roster | Identity reference |
| **SharePoint** | Procedure tutorials, how-to guides | Step-by-step guidance |
| **SELIA (LMS)** | Nomination notes, information memos | Formal HR communications |

## 2.3  The Problem: Fragmented Knowledge Access

No unified interface exists across these five systems. Every query requires manually navigating multiple platforms, generating three tangible costs:

1. **Employee friction:** particularly during onboarding, when employees have no institutional knowledge of which system to consult.

2. **HR team overload:** repetitive, low-value queries consume HR bandwidth that should be directed at higher-value tasks.

3. **No 24/7 availability:** shift workers outside business hours have no way to access procedural information.
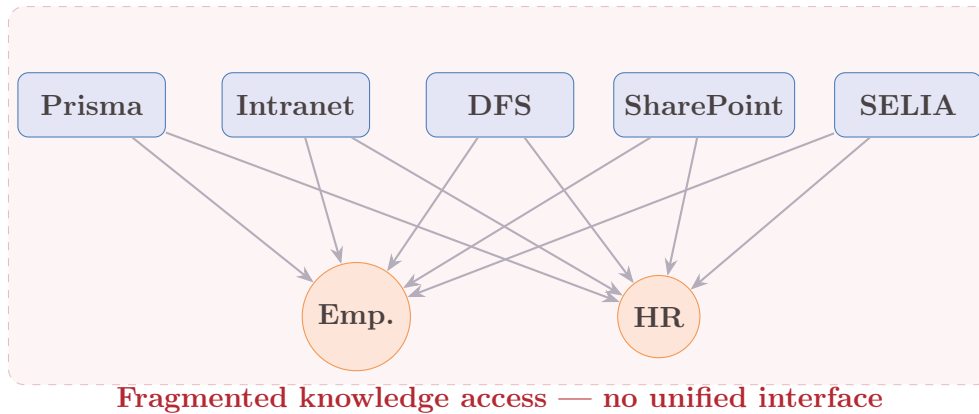


Figure 2.1: Current fragmented HR knowledge access model at Safran SED

## 2.4   Why an AI Chatbot?

A RAG-powered chatbot solves all three problems: it provides a **single conversational entry point** to all five systems, handles unlimited concurrent queries at near-zero marginal cost, and is available around the clock. Unlike a generic LLM, RAG grounds every answer in retrieved official documents, eliminating hallucination risk and providing full source traceability. All processing stays within Safran-approved infrastructure, and no personal data ever reaches the AI model.

# Chapter 3

# Solution & Technical Pipeline

## 3.1  What the Chatbot Does

Employees authenticate with their employee ID and Active Directory password, then type any HR question in natural language. The system retrieves the most relevant excerpts from the official HR corpus and generates a concise, cited answer. The chatbot is scoped to Casablanca SED staff, accessible on professional PCs and dedicated tablets, and exposes **no salary, personal, or sensitive financial data**.

Sample questions the chatbot handles:

- *"What is the procedure for requesting annual leave?"*
- *"What are the onboarding steps for a new technician?"*
- *"Where can I find the nomination note template?"*

## 3.2  End-to-End Pipeline

The system transforms raw HR documents from five source systems into a queryable, AI-ready knowledge base through five automated stages, all orchestrated by Apache Airflow on a daily batch schedule (J-1).

## 3.3  Stage 1 — Raw Ingestion (Airflow)

The DAG `dfs_data_ingestion_dag.py` runs daily, fetching documents from all five source systems. PDFs are parsed with `pdfplumber`, Word files with `python-docx`. Data flows are **unidirectional and read-only**.

## 3.4  Stage 2 — Anonymisation

A Named Entity Recognition model detects PII spans (name, address, phone, employee ID, email) and replaces them with neutral tokens such as `[PERSON]`. No reverse mapping is retained anywhere in the pipeline. An anonymisation report is generated per document for audit purposes.
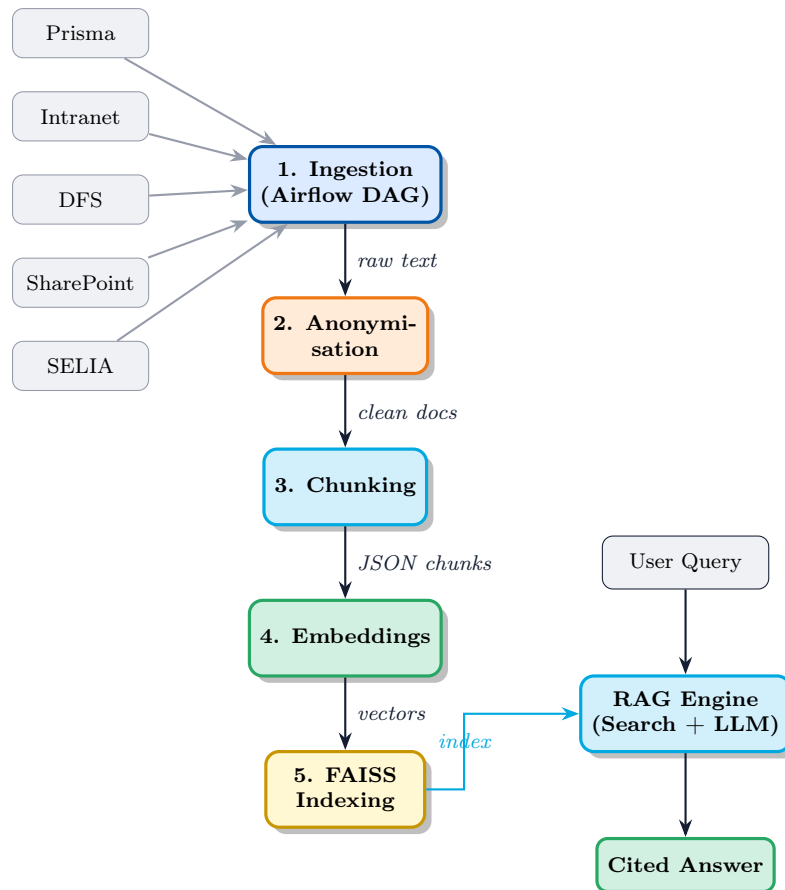
Figure 3.1: End-to-end HR chatbot pipeline — from source documents to cited answers

> **Why irreversibility matters**
>
> Under GDPR Recital 26, truly anonymous data falls outside the scope of data protection law. Irreversible anonymisation means the chatbot corpus is legally and technically free of personal data — satisfying both GDPR and Safran's C2 classification simultaneously.

## 3.5   Stages 3–5 — Chunking, Embeddings, FAISS

Documents are split into 512-token segments with 50-token overlap and stored as JSON chunks. Each chunk is then encoded into a 384-dimensional dense vector using the sentence transformer `all-MiniLM-L6-v2`. These vectors are loaded into a FAISS index (one per source system, then merged into a single `faiss.index` by `merge_faiss_indexes.py`), enabling sub-millisecond semantic search across the entire document corpus.

## 3.6   RAG Inference

When a user submits a question: (1) the question is encoded into a query vector; (2) FAISS returns the top-$k$ most semantically similar chunks; (3) a structured prompt is assembled from those chunks and the question; (4) the LLM generates a concise, cited answer. The model is instructed to answer *only* from the retrieved context, preventing hallucination.

|  | Generic LLM | RAG (this project) |
| --- | --- | --- |
| **Knowledge freshness** | Frozen at training date | Updated daily |
| **Hallucination risk** | High | Low — grounded in docs |
| **Source traceability** | None | Every answer cites source |
| **Data sovereignty** | External API risk | On Safran infrastructure |

# Chapter 4

# System Architecture & Deployment

## 4.1 Full-Stack Overview

The application is structured in four layers: a **Frontend** web chat UI (Streamlit / Flask, `app.py`); a **Backend REST API** managing authentication and session routing; a **RAG Engine** microservice combining FAISS search with LLM inference; and a **Data Layer** of Airflow DAGs, the FAISS index, and chunked JSON files refreshed daily. All four are containerised via `docker-compose.yaml` for reproducibility and portability to Safran's cloud infrastructure.

## 4.2 Security & Compliance

| Principle | Implementation |
| --- | --- |
| **Anonymisation** | Irreversible PII removal before any AI processing |
| **Data flow** | Unidirectional, read-only from source systems |
| **Authentication** | Matricule + AD password; professional devices only |
| **Access scope** | All SED staff; no salary or personal records surfaced |
| **Hosting** | Safran-approved cloud (C2-compliant, to confirm with IT) |
| **Monitoring** | Pipeline logs, email incident alerts, access audit trail |
| **GDPR** | Data processed in anonymised form in accordance with Recital 26 |

# Chapter 5

# Quality, Testing & Governance

## 5.1 Automated Testing

`test_rag.py` covers three dimensions: **retrieval precision** (relevant chunks must appear in top-$k$ for a golden query set), **answer faithfulness** (generated answers must be grounded in the retrieved context only), and **latency** (end-to-end response time benchmarks).

## 5.2 Monitoring

Pipeline logs capture per-run metrics (duration, document counts, anonymisation stats, error rates). Email alerts fire automatically on DAG failures or latency anomalies.

## 5.3 Governance

> ### HR Validation Before Go-Live
> Chatbot responses are reviewed by the HR team in a pre-production environment before deployment. A designated **HR referent** owns ongoing content governance, with scheduled retrospective reviews once in production to keep answer quality high and the corpus current.

# Chapter 6

# Deployment Plan & Success Criteria

## 6.1 Roadmap

1. **Phase 1 — Ideation & Pre-selection:** Problem scoping and concept validation with the Think To Deploy jury.

2. **Phase 2 — POC & Pitch (current):** Functional POC built on a curated HR corpus, pitched to Safran SED HR department.

3. **Phase 3 — Restricted Pilot:** Deployment to 10 simultaneous terminals with live HR team feedback.

4. **Phase 4 — Industrialisation:** Full rollout to ≈200 terminals; MLOps via Dataiku; potential extension to other Safran sites.

## 6.2 Success Criteria

| Dimension | Criterion |
|---|---|
| **Business relevance** | HR team validates $\geq 80\%$ of responses as correct and useful |
| **Data sovereignty** | Zero PII leakage; all processing within Safran infrastructure |
| **Added value** | Measurable reduction in repetitive HR queries |
| **Robustness** | Uptime $\geq 99\%$ during POC; graceful failure handling |
| **Integration** | Daily refresh from all five sources without manual intervention |

# Chapter 7

# Key Learnings

- **Governance-first design is non-negotiable in regulated environments.** Building anonymisation into the ingestion stage — before any embedding or indexing — was the only compliant approach. Retrofitting privacy controls after the fact is both technically risky and legally insufficient.

- **RAG substantially reduces hallucination versus standalone LLMs.** Grounding every answer in retrieved, source-traceable document chunks makes the system auditable and trustworthy — critical for HR contexts where incorrect answers have real consequences for employees.

- **Orchestration and observability are as important as the model.** A daily Airflow pipeline with structured logging, alerting, and anonymisation reports is what makes the system *deployable* — not just *impressive in a demo*. Industrial AI requires operational rigour.

- **Communicating AI to non-technical stakeholders requires business framing.** The HR department evaluated the solution on business relevance, data sovereignty, and operational impact — not model accuracy metrics. Pitching at the right level of abstraction was as important as the architecture itself.

# Chapter 8

# Conclusion

The HR Intelligent Chatbot replaces five disconnected manual lookups with a single conversational interface, grounded in official Safran HR documentation and refreshed daily. The solution is production-ready in design, compliant by construction, and extensible to additional source systems at minimal cost. It was developed as a real project and pitched to the HR department of Safran SED as part of the Think To Deploy Phase 2 hackathon.

> **Next Steps**
>
> **Short term:** POC validation on 10 SED terminals; HR team feedback on answer quality.
>
> **Medium term:** Safran Cloud deployment; full corpus expansion; Dataiku MLOps integration.
>
> **Long term:** Full-site rollout to all 200 terminals; replication at other Safran Group sites.