AI-Driven Enterprise Search Blueprint

Robert Wood Johnson Foundation

June 28, 2025

1 AI-Driven Enterprise Search Blueprint – Robert Wood Johnson Foundation

Version 1.6 – CAG Enhancements, OSS RAG Readiness, and UX Rationalisation
Date: 28 June 2025
Author: Harindha Fernando – Enterprise Architect, NCINGA

1.1 Executive Summary – Why Modernise Now?

RWJF's legacy enterprise search relies on Raytion connectors, a technology that has now been placed on an end-of-life path following Raytion's recent acquisition. This creates an imminent risk of search outage across Oracle, Adobe AEM and SharePoint unless a replacement architecture is adopted. At the same time, RWJF is transitioning grants management onto Salesforce Nonprofit Cloud and intends to leverage Salesforce Data Cloud as its primary data layer while maintaining a substantial Microsoft 365 and Azure footprint.

The proposed architecture therefore serves two urgent purposes: 1. Preserve continuity of enterprise search as Raytion support winds down. 2. Modernise the experience to deliver AI-driven, semantically rich answers that span Salesforce, Microsoft and on-premises content — all under a single, governed interface.

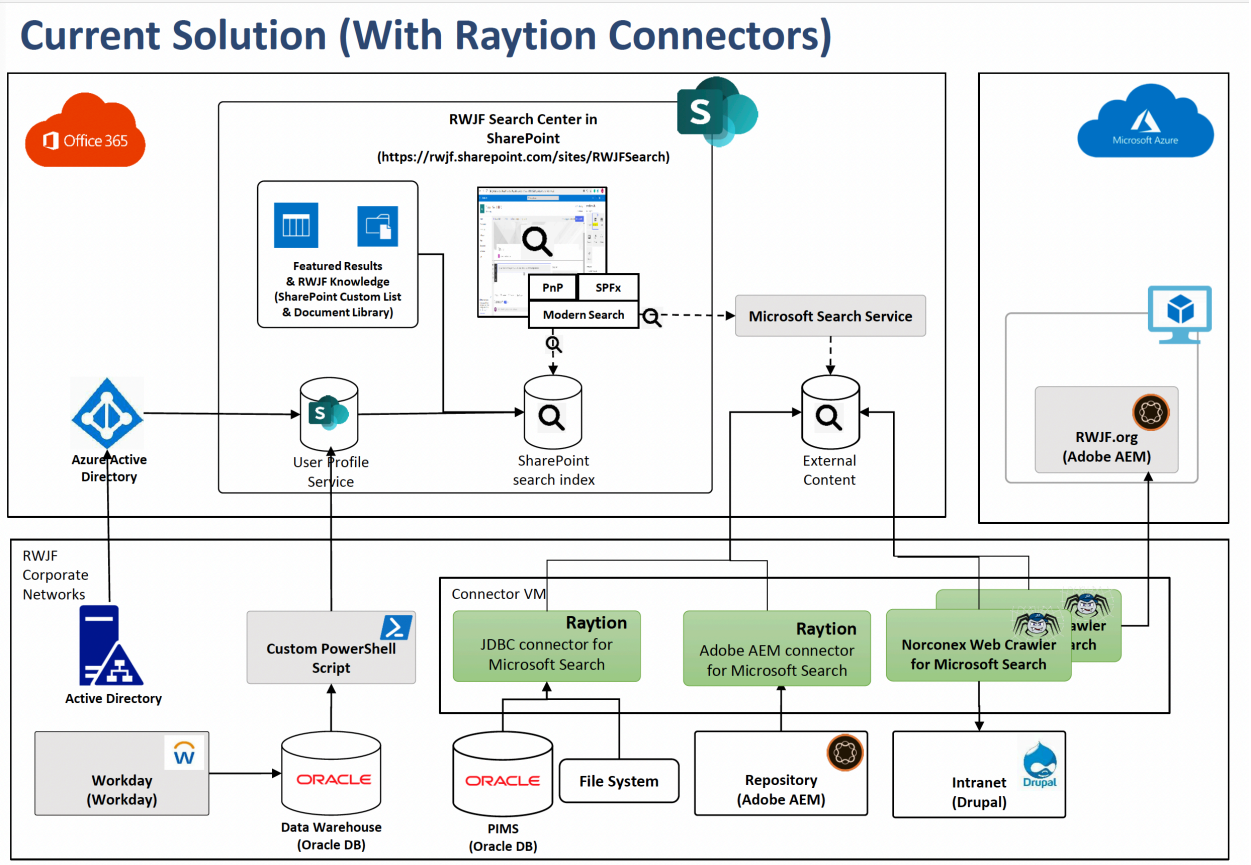
NCINGA's blueprint replaces brittle connectors with a dual-layer integration model: Denodo virtualises structured sources while UIB, our secure API gateway, brokers unstructured content. Sentiyo provides the conversational AI layer, and Databricks supplies lakehouse analytics and vector storage for Retrieval-Augmented Generation (RAG). This design aligns with RWJF's hybrid-cloud strategy, enabling phased adoption without disruptive data migrations.

In discussions with RWJF leadership, two design paths — Salesforce-centric and Azure-centric — were explored. The blueprint supports either path, ensuring continuous value regardless of future ecosystem emphasis.

1.2 Existing Architecture and Challenges

RWJF's current enterprise search solution is built on Raytion connectors integrated with Oracle databases, Adobe Experience Manager (AEM), Microsoft SharePoint, and internal content repositories such as rwjf.org. This system has been in place for several years and has served as the core enterprise knowledge access mechanism for internal staff.

Figure 2.1 – Technology Architecture



1.2.1 Technologies in Use

- **Oracle PIMS DB** – Primary structured data source for grants and program data.
- **Adobe AEM** – Hosts foundation publications, policy documents, and reports.
- **SharePoint Online and Microsoft 365** – Used widely for internal collaboration and documentation.
- **Drupal CMS** – Manages portions of the rwjf.org site for public content.
- **Raytion Connectors** – Facilitate search federation across these sources.

1.2.2 Technical Limitations and Risks

- **Vendor Discontinuation**: Raytion has been acquired, and support for the existing connector stack is being deprecated, creating high risk of future downtime or patch gaps.
- **Data Silos**: Content is indexed individually from each source, without unified semantic metadata or governance enforcement.
- **No Interoperability with Salesforce**: As RWJF adopts Salesforce Nonprofit Cloud and Data Cloud, the current system has no native compatibility with these platforms.

- **Limited AI Capabilities**: The current search lacks semantic awareness, natural language support, and contextual understanding.
- **Compliance Gaps**: GDPR and PII controls are fragmented, with no centralized policy enforcement layer.

1.2.3 Business Impact

- **User Frustration**: Inconsistent or irrelevant results reduce productivity.
- **Support Overhead**: Legacy connectors require manual intervention and frequent troubleshooting.
- **High TCO**: Licensing and maintenance costs for Raytion and Oracle-based architecture are significant.
- **Innovation Barrier**: Lack of AI extensibility hinders future innovation in analytics, decision support, and generative insights.

1.3 Business Drivers & Success Criteria – Why This Matters to RWJF

- **Mission alignment:** Faster insight loops directly advance RWJF's public-health mission.
- **Operational efficiency:** Target a 70 % reduction in time-to-information by unifying search.
- **Risk & compliance:** A single policy plane ensures GDPR-grade protection of donor and staff data.
- **Cost optimisation:** ROI will be quantified jointly; section 9 outlines the data we will gather.

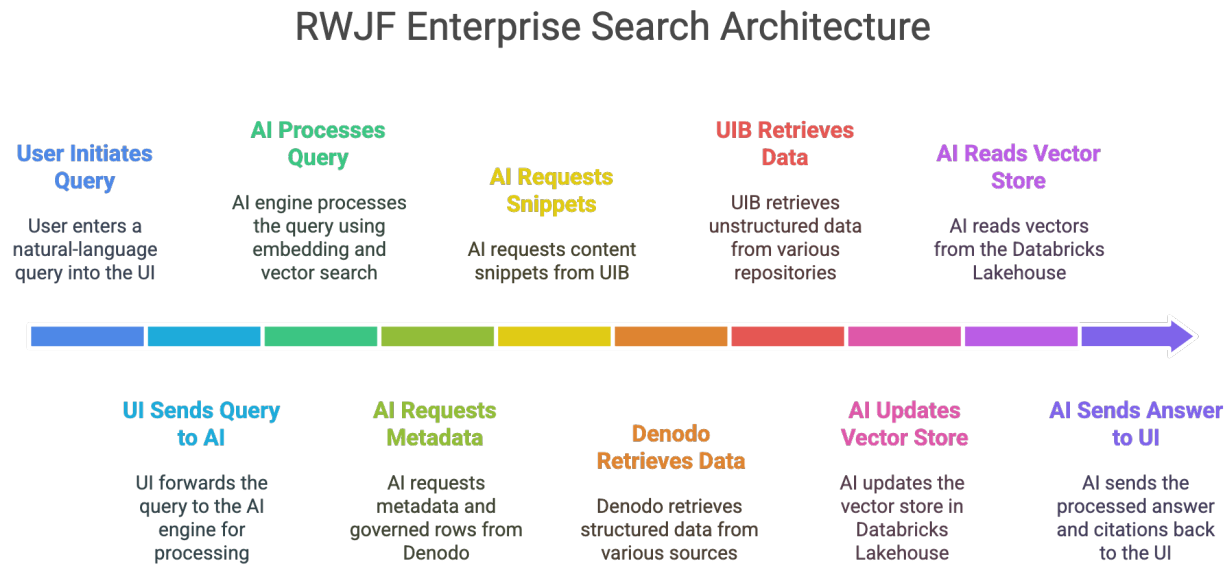
1.4 Architectural Principles – How We Keep the Solution Durable

1. **Governance-First** – access checked before data leaves its source.
2. **Virtualise, Don't Migrate** – move data only when analytics demands it.
3. **Composable over Monolithic** – each layer replaceable without lock-in.
4. **Observable by Design** – logs and metrics exported to SOC tooling.
5. **User-Centric UX** – one conversational interface; plumbing is invisible to users.

1.5 Reference Architecture – What We Are Building and Why

The architecture is deliberately layered. Users interact solely with Sentiyo's conversational interface. Sentiyo interprets intent, retrieves governed structured data via Denodo and secure document snippets via UIB, then composes answers backed by Databricks vector search. This separation maximises agility and simplifies policy enforcement.

Figure 5.1 – Technology Architecture



1.6 User Journey Flow

The user journey begins with a natural language query — for example, "Show me all grants awarded to education programs in New Jersey in the last 5 years." This query, whether entered via the Sentiyo interface or federated via intranet/SharePoint, is authenticated through Azure AD and routed to the intelligent orchestration layer.

1.6.1 Step-by-Step Flow:

1. **Authentication** – The user signs in using corporate SSO (Azure AD) with policy-enforced multifactor authentication.
2. **Intent Interpretation** – Sentiyo's agentic AI interprets the query and determines the semantic domain (structured or unstructured).
3. **Access Authorization** – Role-based access control, data masking, and policy checks are applied based on user context.
4. **Data Retrieval** – Depending on the source type:
    - Structured queries are routed through Denodo for virtualized, policy-enforced access.
    - Unstructured queries route through UIB for snippet-level extraction from SharePoint, AEM, Drupal, or rwjf.org.

5. **Answer Synthesis** – Sentiyo generates an explainable answer using Retrieval-Augmented Generation (RAG), combining structured values with referenced source snippets.
6. **Result Delivery** – The response is returned through Sentiyo's interface with citations, or redirected to Power BI for data exploration.

The following subsections expand on the intelligent orchestration and architectural decomposition behind this flow.

1.7 Agentic AI Execution Model – Modular Intelligence in Action

At the heart of the Sentiyo search layer is a modular **Agentic AI Execution Model**. Instead of monolithic AI models, the platform delegates responsibilities across autonomous, cooperating agents. This provides auditability, explainability, and the ability to evolve components independently. Sentiyo's architecture has evolved to support both Context-Aware Generation (CAG) and Cache-Augmented Generation. This enables real-time personalization with memory of prior queries, while reducing latency and cost through intelligent caching of responses.

1.7.1 Key Agents and Responsibilities:

- **Intent Agent**
  Parses the natural language query, identifies semantic scope, entities, and query type (e.g., fact lookup vs. document retrieval). It formulates an intermediate search plan.
- **Access Agent**
  Consults the user's JWT token, associated RBAC policies, and organizational masking rules. It validates permissions before triggering any downstream execution.
- **Retrieval Agent**
  Orchestrates data access:
  - For structured data: issues SQL or Denodo view calls.
  - For unstructured: issues secure requests via UIB to repositories like SharePoint, AEM, or public websites. Now includes full-text indexing (via Azure AI Search or Elasticsearch) for rich relevance-based retrieval.
- **Response Agent**
  Synthesizes a complete, human-readable response using vector embeddings and Retrieval-Augmented Generation (RAG), referencing both structured values and content snippets.
- **Context-Aware Agent (CAG)**
  Maintains session-level context and multi-turn memory, allowing refined follow-up queries and personalized search behavior.

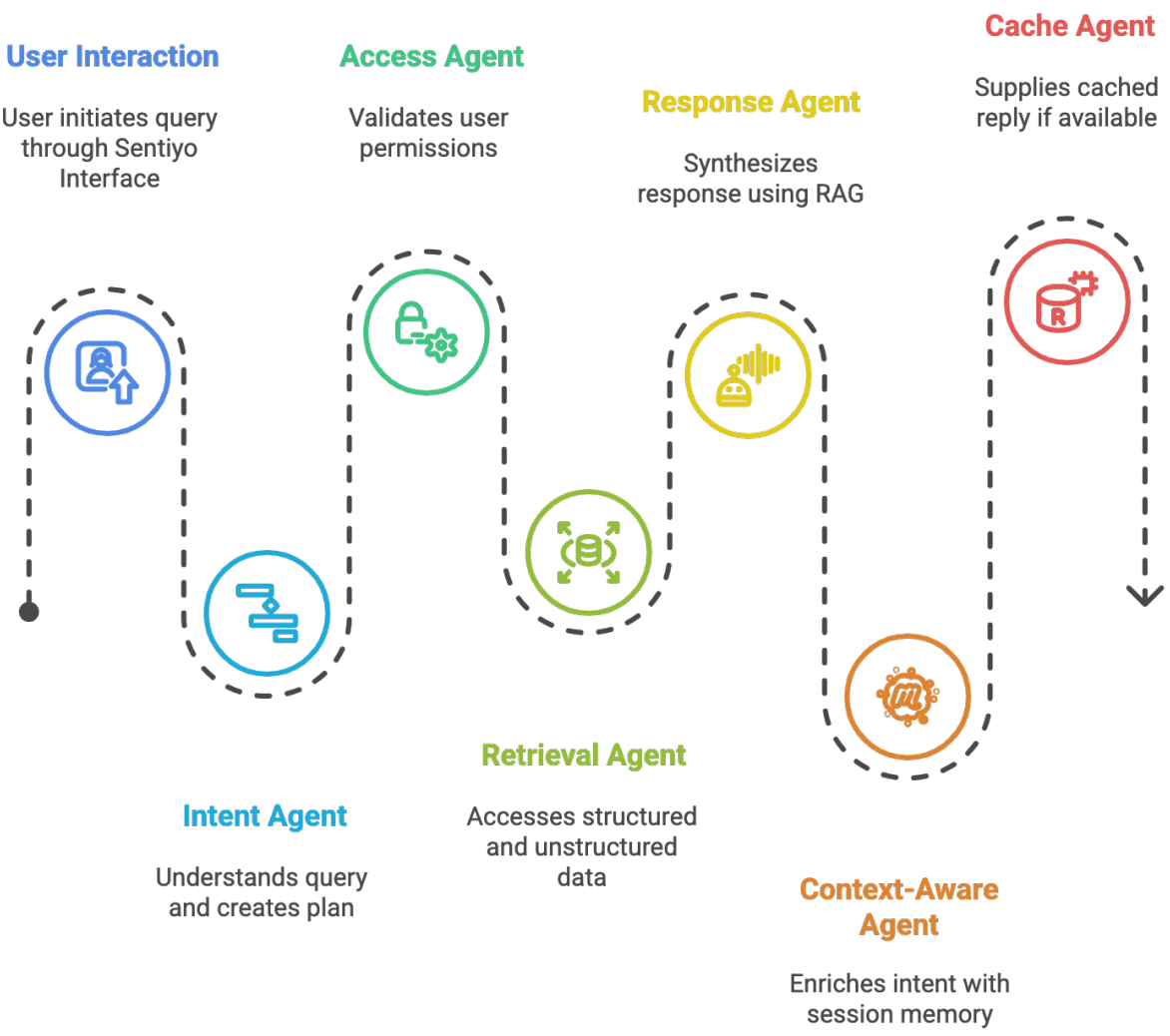
- **Cache Layer (CAG2)**
  Identifies and retrieves previous responses for repeated or semantically similar queries to avoid redundant LLM calls.

This hybrid CAG model enhances response speed, cost-efficiency, and user relevance.

1.7.2 Agentic Benefits:

- **Separation of concerns**: Each agent is self-contained and focused on one task.
- **Observability**: Logs and metrics are granular, improving traceability and debugging.
- **Upgradability**: Each agent can evolve independently (e.g., plugging in GPT-5.5 instead of GPT-4-turbo).
- **Compliance**: Easier enforcement of data governance rules via the Access Agent.

Figure 6.1 – Agentic AI Execution Model with Dual CAG



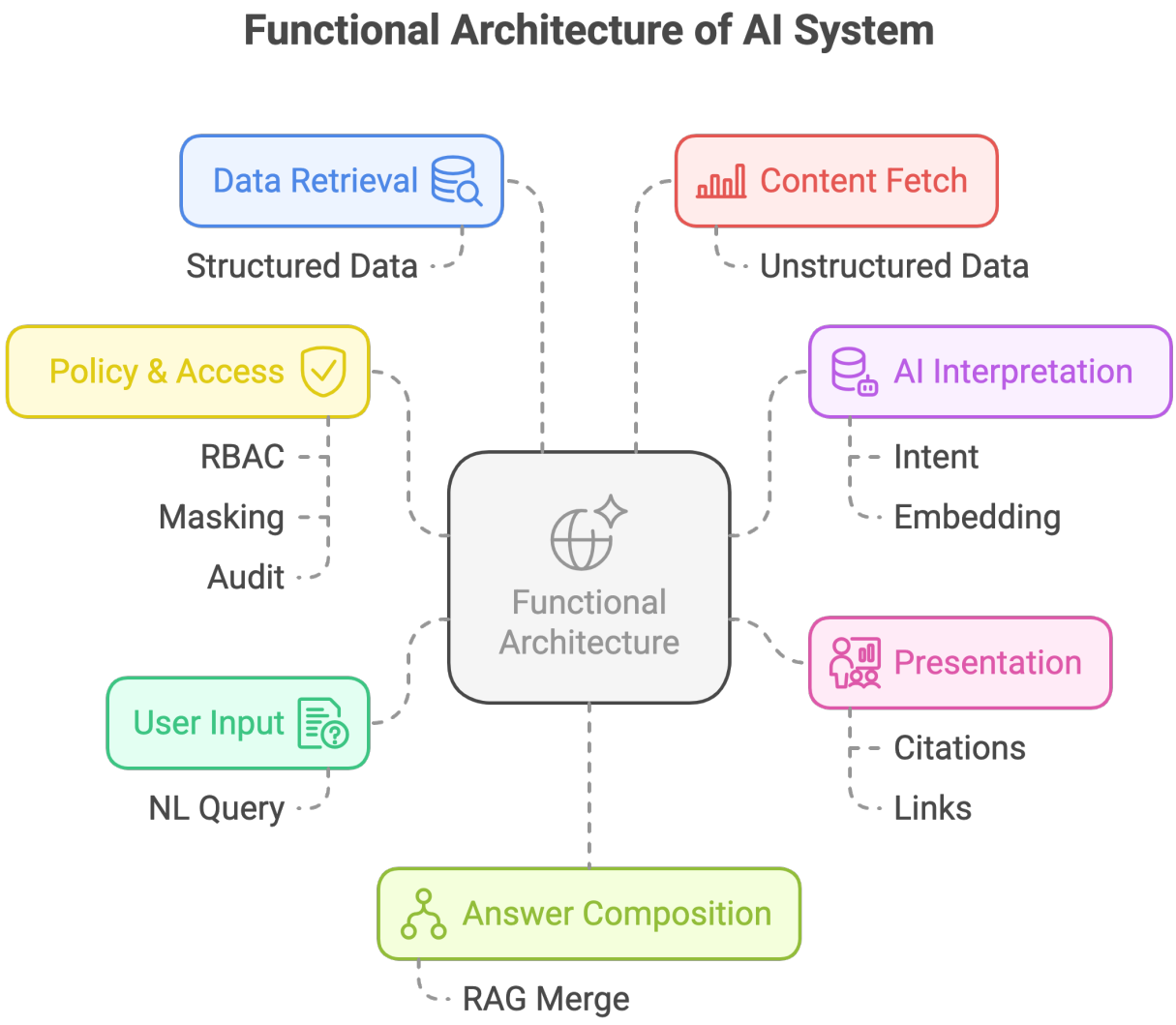
1.8 Functional Layers – Responsibilities Over Vendors

This architecture separates concerns cleanly across logical layers, ensuring each function is modular, testable, and replaceable. Below is a vendor-agnostic breakdown of the system's responsibilities:

- **Input Layer**
  Captures user query (natural language) via Sentiyo's UI. No preprocessing or assumptions occur here.
- **Intent Understanding Layer**
  Interprets query intent, determines if it targets structured or unstructured data, and formulates a retrieval plan. Performed by the Sentiyo AI Engine's "Intent Agent".

- **Policy & Access Control Layer**
  Evaluates user's JWT context, access rights, and PII masking requirements before allowing any query execution. Enforced jointly by Denodo (structured) and UIB (unstructured).
- **Retrieval Layer – Structured**
  Executes virtual SQL queries via Denodo Views, providing governed access to live transactional data without migration.
- **Retrieval Layer – Unstructured**
  Leverages full-text indexing via Azure AI Search or open-source alternatives (e.g., Elasticsearch, Apache Solr) to support deep document retrieval. Snippets are extracted securely via UIB once documents are located.
- **Answer Composition Layer**
  Combines retrieved data using Retrieval-Augmented Generation (RAG), with vector support via Databricks, to produce grounded, explainable responses.
- **Presentation Layer**
  Sentiyo returns the final response to the user, with inline citations, and optionally enables hand-off to Power BI for deep analytics.

Figure 6.2 – Functional Layers (Vendor-Agnostic)



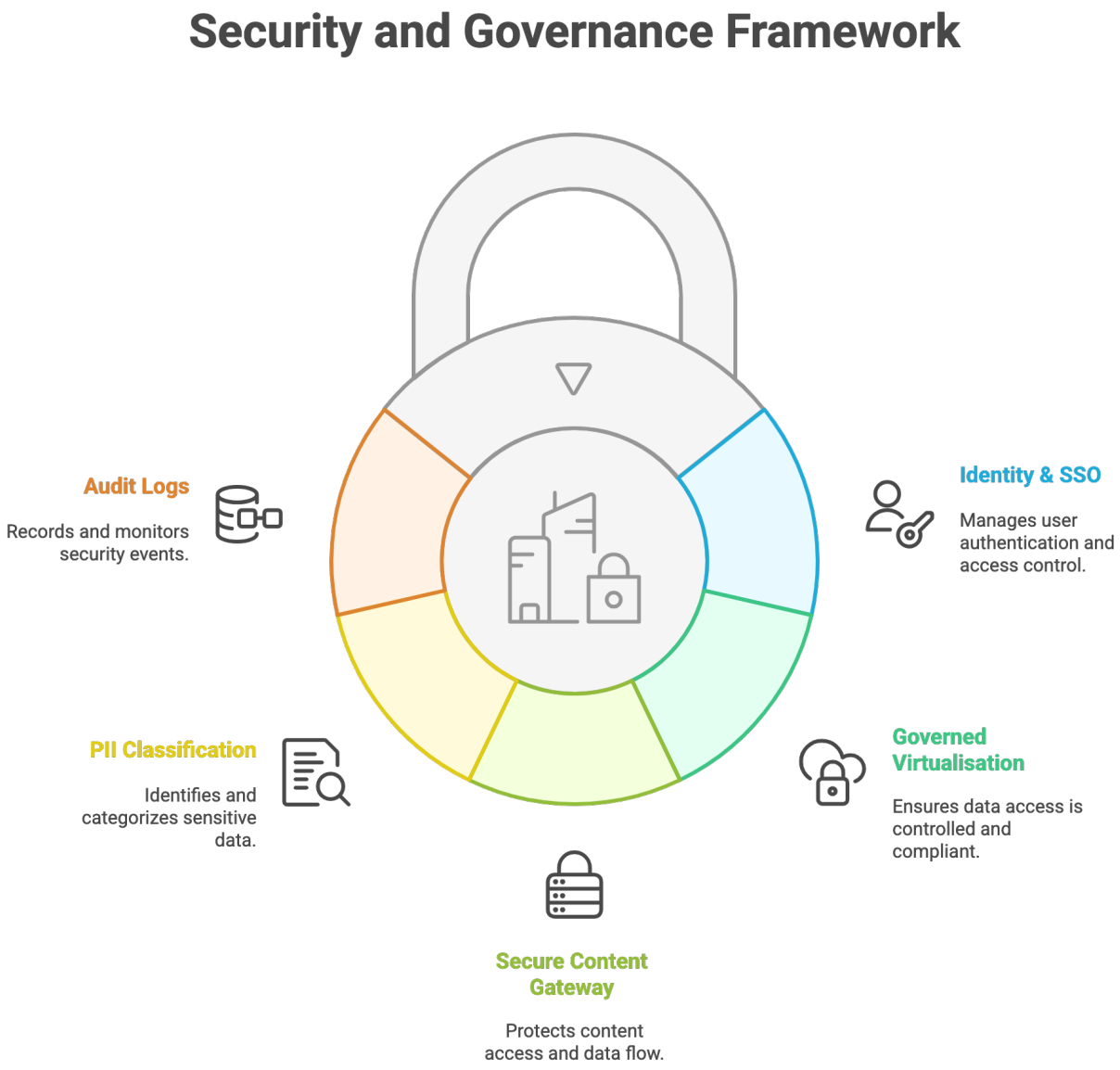
1.9 Security & Compliance – Trust by Architecture, Not Afterthought

Security in this solution is embedded from design through operations. Rather than bolted on post-facto, security controls are composed into each layer — from data classification to access policy, observability, and automated response.

- **Data Classification**
  Azure Purview automatically tags sensitive fields (e.g., PII, grant recipient info), which Denodo and UIB enforce at runtime via masking, row-level security, and audit tagging.
- **Access & Encryption**
  TLS 1.3 is enforced end-to-end across microservices. All access is governed by role-based policies, backed by Azure AD. UIB inspects every request for token validity and policy compliance.

- **SIEM & Sentinel Automations**
  Logs, traces, and events from UIB, Denodo, Sentiyo, Databricks, and infrastructure are streamed into **Microsoft Sentinel**, RWJF's native cloud SIEM.
  Sentinel correlates signals to detect threats, such as:
  - Unusual access patterns
  - Excessive RAG queries hitting PII
  - Query injection attempts into Denodo views
    Automated playbooks respond by notifying SOC teams, disabling sessions, or opening service tickets in RWJF's ITSM tool.
- **DevSecOps Integration**
  CI/CD pipelines include policy-as-code, container image scanning, and endpoint hardening checks. Code failing security tests cannot progress to staging or production.
- **SRE Observability**
  All components emit structured logs, metrics, and health checks. Alerts are routed to Azure Monitor and Grafana dashboards. Quarterly chaos drills validate that the system fails closed and recovers gracefully.
- **Compliance Archiving**
  Immutable audit logs are retained for a minimum of 7 years, meeting IRS reporting and GDPR obligations for donor and stakeholder data.

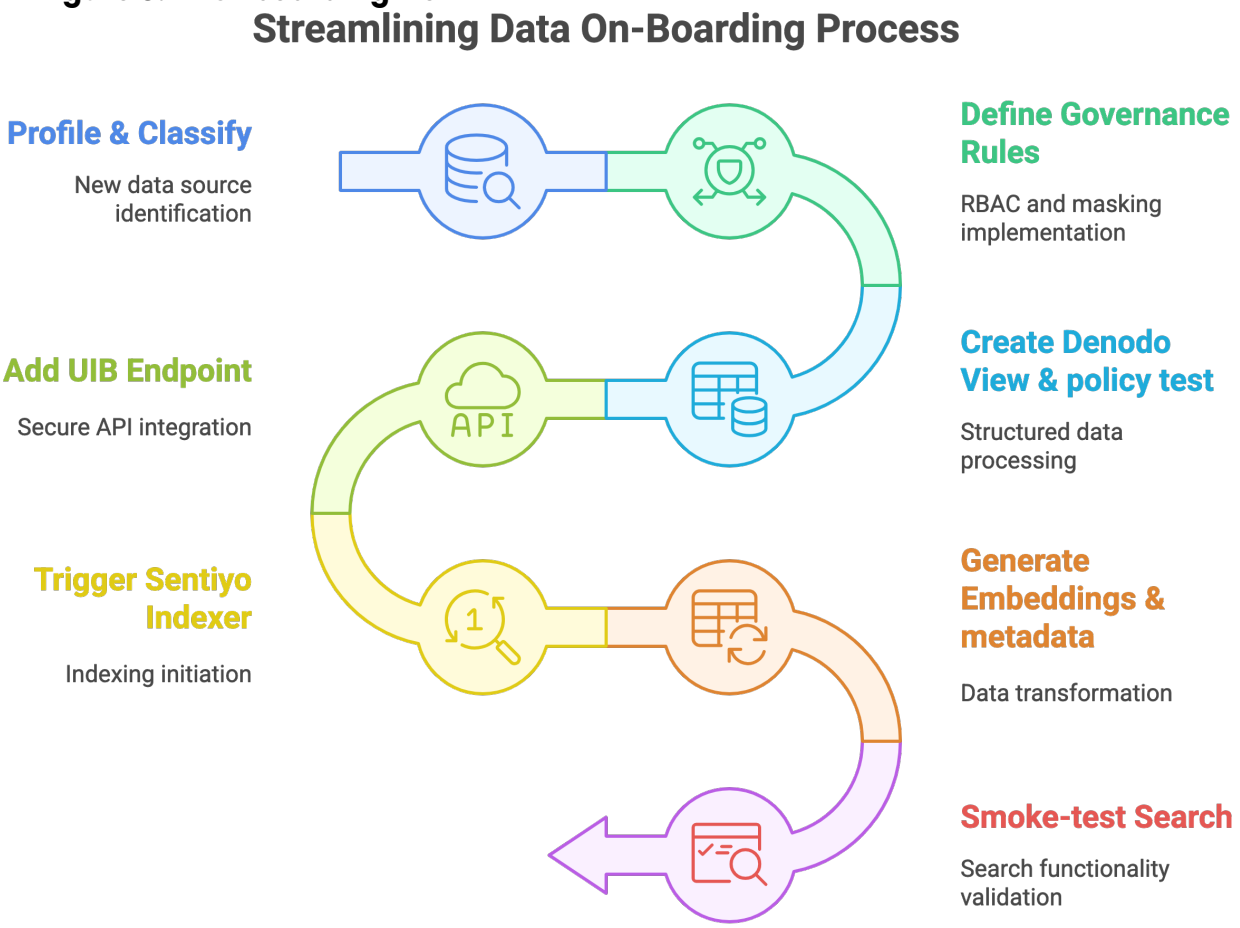
Figure 7.1 – Security & Compliance Layers



1.10 Deployment & Onboarding – How New Systems Join the Platform

1. Profile & classify – data stewards tag sensitivity and ownership
2. Govern – RBAC and masking rules created once in Denodo or UIB
3. Connect – API or SQL view established, policies smoke-tested
4. Index – Sentiyo embeds metadata and snippets
5. Validate – Pilot queries confirm relevance and redaction before go-live

Figure 8.1 – Onboarding Flow



1.11 Operational Excellence – Keeping the Lights On, Securely

CI/CD pipelines deploy all components via blue/green containers. Logs, traces and metrics stream into Azure Monitor; Sentinel automations raise alerts. Quarterly chaos drills validate fail-closed security behaviour.

1.12 Cost Optimisation & ROI – Structured Next Step

Rather than applying speculative ROI figures, we propose forming a joint RWJF–NCINGA working group to collect: - Legacy platform spend (Oracle, Raytion, bespoke connectors)
- Productivity baselines (time-to-information, ticket volume)
- Licence and support contract data
- User-satisfaction metrics
- Storage redundancy and replication overhead

Within four weeks the group will deliver a baseline TCO, a defensible board-level ROI projection and an insights report to guide future budgeting.

The true value of the transformation lies not only in cost savings but in enabling faster insights, fewer escalations and stronger compliance.

1.13 Road-map & Phased Delivery – Risk-Managed Transformation

- **Phase 0 – Mobilise (2 weeks):**
  Finalize project charter, RACI structure, cloud landing zone, and compliance onboarding.
- **Phase 1 – Foundation (6 weeks):**
  Deploy Denodo for structured virtualization, harden UIB API Gateway, integrate Azure AD SSO, and set up initial DevSecOps pipelines.
- **Phase 2 – Pilot (8 weeks):**
  - Connect **Salesforce Data Cloud** and **Nonprofit Cloud** to Denodo and AI pipeline.
  - Enable **SharePoint** and full-text index integration (using Azure AI Search or OSS).
  - Activate Sentiyo with **CAG + RAG support** for unified structured and unstructured search.
- **Phase 3 – Expansion (12 weeks):**
  - Onboard **Adobe AEM, Drupal, Workday**, and **rwjf.org** via UIB.
  - Retire legacy Oracle PIMS system post-successful data migration to Salesforce.
- **Phase 4 – Optimise (Ongoing):**
  - Continuous fine-tuning of RAG relevance using user feedback.
  - Enhance dashboards and analytics via Power BI and Databricks.
  - Enable additional use cases (e.g., Grants Assistant, Contracts Summarizer) on Sentiyo.
