AWP Agentic Workflow Platform

Architecture Study

Mohamad NABAA

20 Janvier 2026

Sommaire

Index

01 Introduction

02 Logical Architecture

03 Physical Architecture

04 RAG Strategies

05 Advantages

06 TCO & Roadmap

07 Conclusion

Introduction

Use Case

LUXE Sector client with 25000 internal users, are using a static chatbot that provides information but lack the capability to perform autonomous operational actions.

Vision

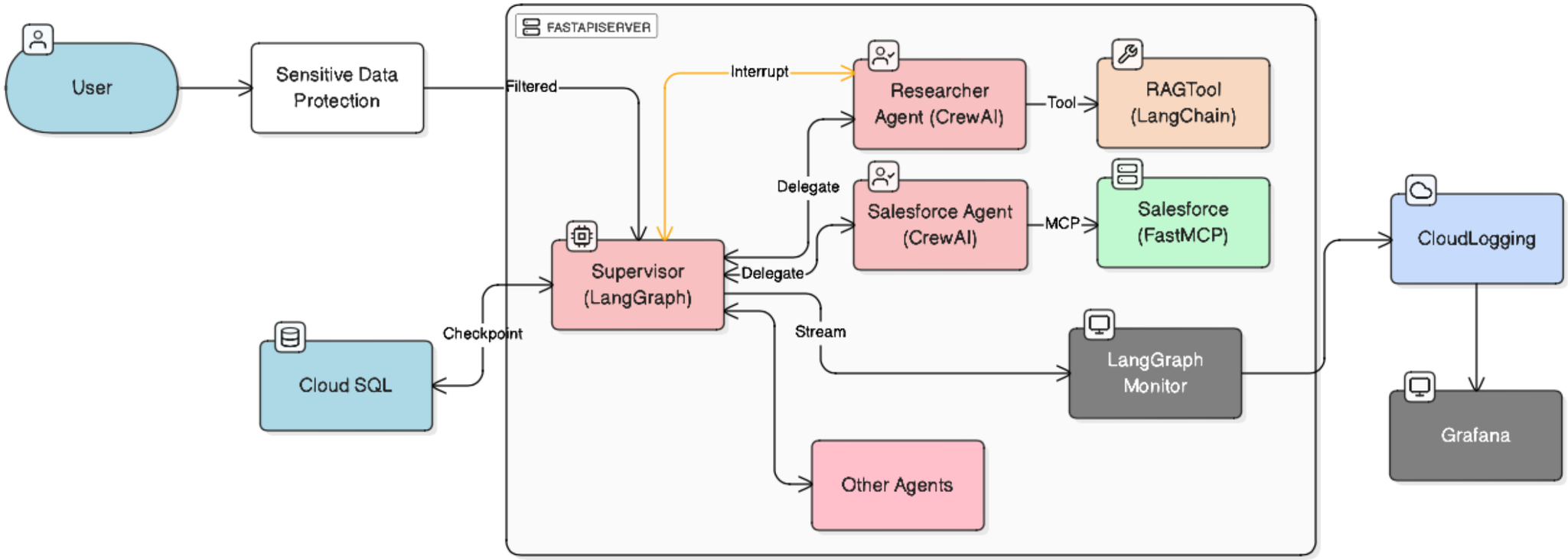
From Static Chatbot to Sovereign Agentic Platform

Objective

Create a Sovereign Agentic-powered "Virtual Team" that executes complex tasks while keeping the human in the loop for actions.

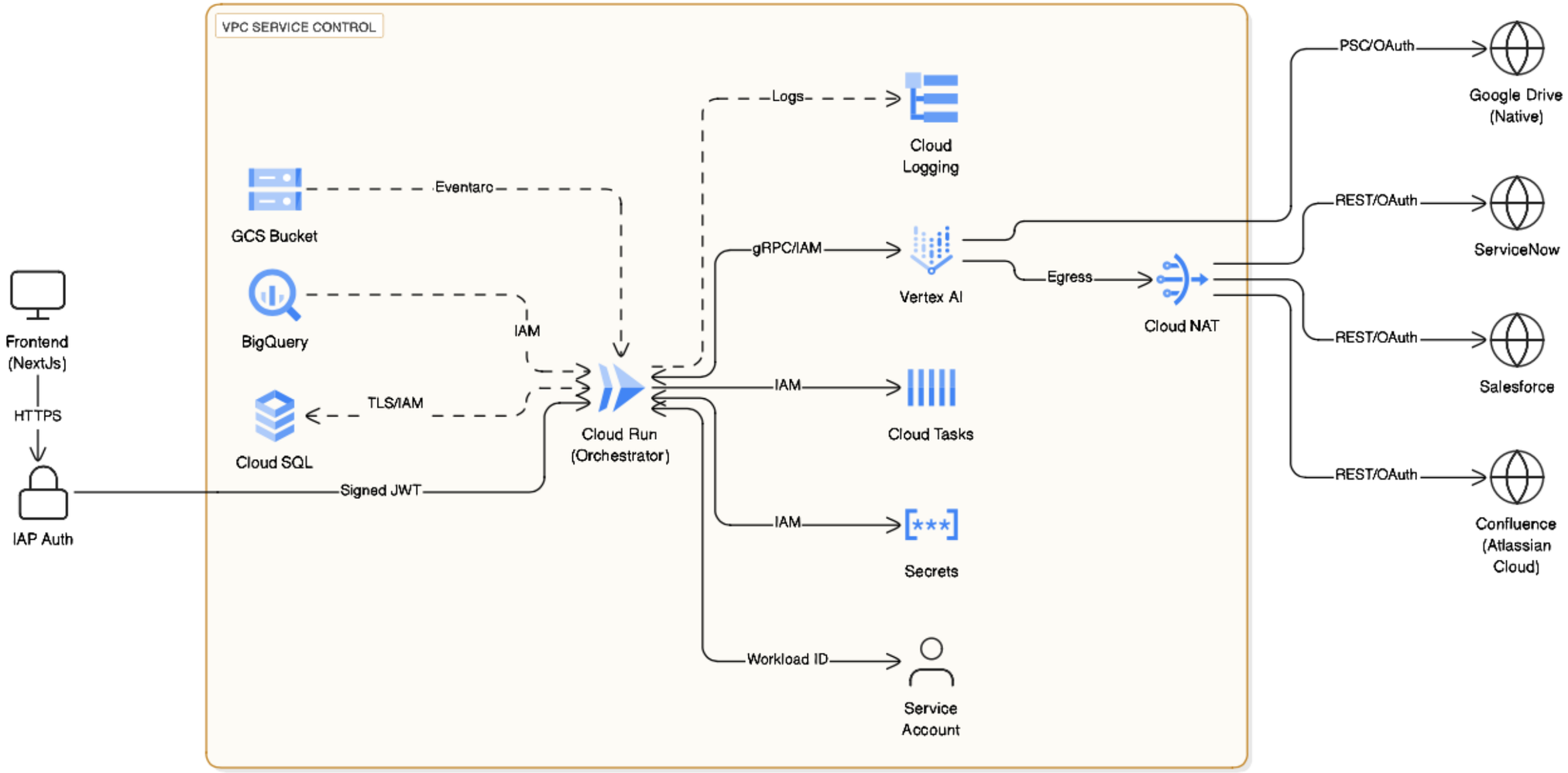
Logical Architecture (The "Virtual Team")

Stateful Multi-Agent Orchestration with End-to-End Observability



Physical Architecture

Zero-Trust Serverless Enclave for Enterprise GenAI (Paris-west-9)



Architecture

Asynchronous Python native stack

Here's a breakdown of our technical architecture:

Stack

- **LangGraph** for graph orchestration, **LangChain** for native tools
- **CrewAI** for agent implementation
- **FastAPI** for the async server runtime
- **FastMCP** for MCP server/client
- **Terraform**: Infrastructure as a Service

Observability

LangGraph Monitoring traces every graph node state transition, exported to Cloud Logging and Cloud Trace (Grafana) for production monitoring.

Governance

Sensitive Data Protection: (GCP managed service) for PII redaction at ingress

Interrupt-based approval flow for high-risk actions (Human-in-Loop validated via IAP identity).

RAG Strategies

Multiple sources of KB seamlessly integrated in the pipeline

Corporate Knowledge (Confluence, Drive)

- Vertex AI Agent Builder
- Secure with Private Service Connect

Structured Business Data (BigQuery)

- Semantic Text-to-SQL (Langchain Tool) exposed to agent
- text-embedding-004
- Vector Search on preingested (worker based) structured schema embeddings
- DRY RUN with authorized functions with Read-Only via Service Account

Transient Uploads

- Presigned URL Upload
- Vertex AI Layout-Aware Chunking
- text-embedding-004
- Vertex AI RAG Engine configured for Semantic Chunking

Analyse Comparative

Advantages

Advantage	Static Chatbot (Current)	Agentic Platform Capability	Architectural Solution
Operational Efficiency	High L1/L2 support costs (€1M+ annually) for 25k users.	Agents autonomously offload L1/L2 support and operational actions.	**Supervisor (LangGraph) & CrewAI Agents**
Reduced MTTR	Response and resolution times are bottlenecked by human intervention.	Faster response times with minimal human intervention through automated workflows.	**Supervisor (LangGraph) & CrewAI Agents**
Data Sovereignty	Reliance on "Black Box" SaaS with vague GDPR and data residency controls.	100% EU Data Residency with Zero-Trust security in a serverless enclave.	**VPC Service Control**(Paris-west-9)
System Ownership	Renting a platform where the reasoning logic is a vendor secret.	Full ownership of the "AI Brain," specialized logic, and system connectors.	**FastAPI Native Stack& FastMCP Connectors**
Security & Privacy	Limited ability to filter or redact sensitive data before LLM processing.	Integrated PII redaction and identity propagation for secure operations.	**Cloud SDP & IAP Authentication**
Scalable TCO	Inflexible scaling; costs don't align perfectly with actual usage.	Horizontal, demand-driven scaling with usage-dependent serverless costing.	**Cloud Run & Serverless Scaling**

Planification

TCO & Roadmap

Build Phase

- CapEx: Estimated Cost: €350K ->€500K
- 3-6 months, Requires 4-5 FTE Squad:
  - Cloud Architect
  - Senior AI Engineer
  - Data Engineer
  - DevOps

Run Phase

- OpEx: Infrastructure €6,500 - €8,000 / month
- Cloud Run, Vertex AI, Langchain, LLM Tokens for ~1k users

Operational Team

- Requires ~2.5 FTE (Platform Engineer + Data Steward + FTE) to maintain connectors and governance.

Conclusion

Agentic Workflow Platform

Total replacement of a static chatbot. Activates autonomous agents with a human in the loop while preserving observability and governance. Scalable on demand.

Budget

Development budget replaces recurring 1 year at most of L1/L2 support cost, rest is maintenance cost which is a lot lower.

Ownership

We own the "Brain" and the "Connectors" (the Black Box SaaS).