AWP Agentic Workflow

Platform

Architecture Study

Mohamad NABAA

20 Janvier 2026

Sommaire

Index

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

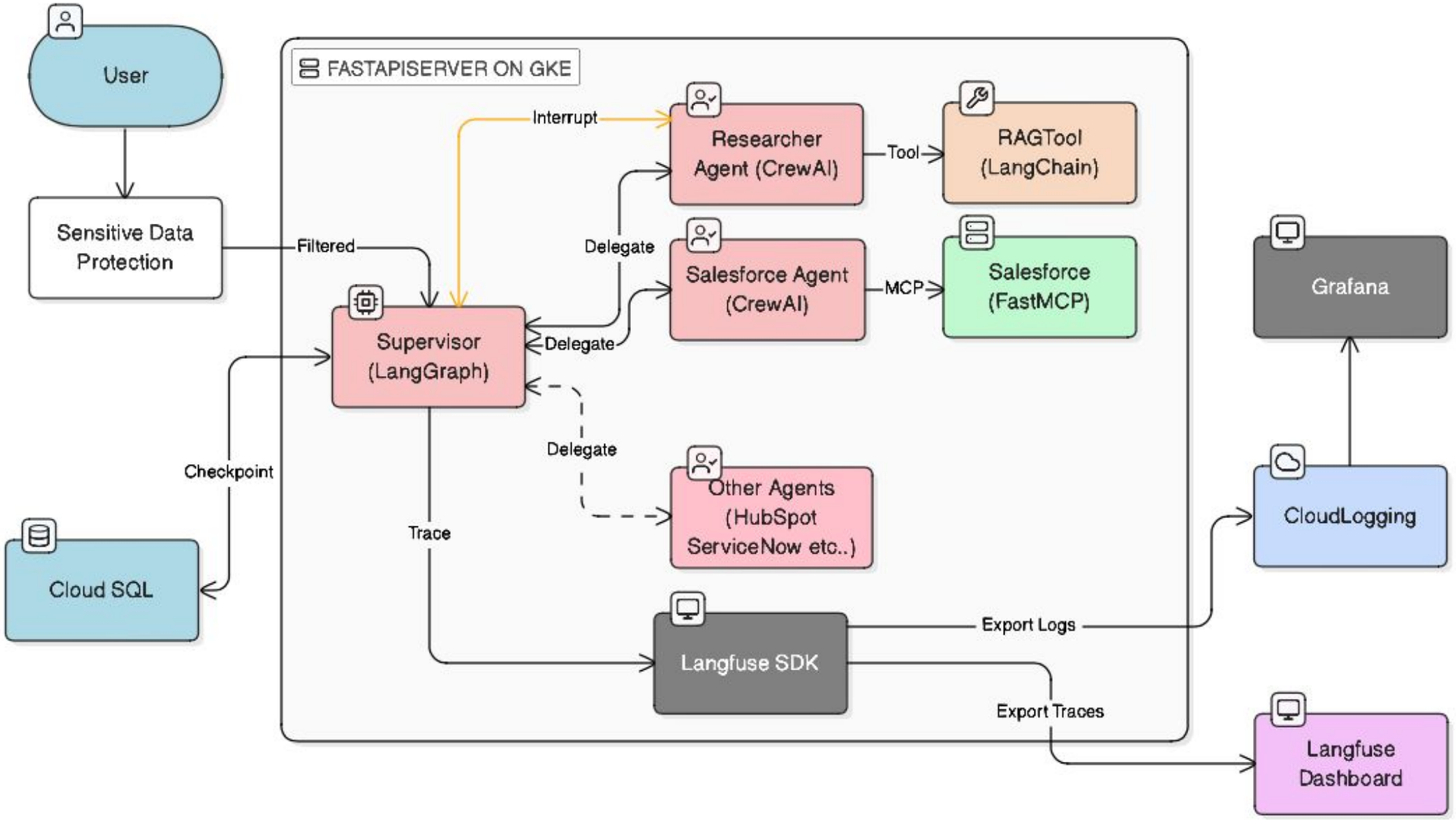
Transform from Static Chatbot to Sovereign Agentic Platform

Objective

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

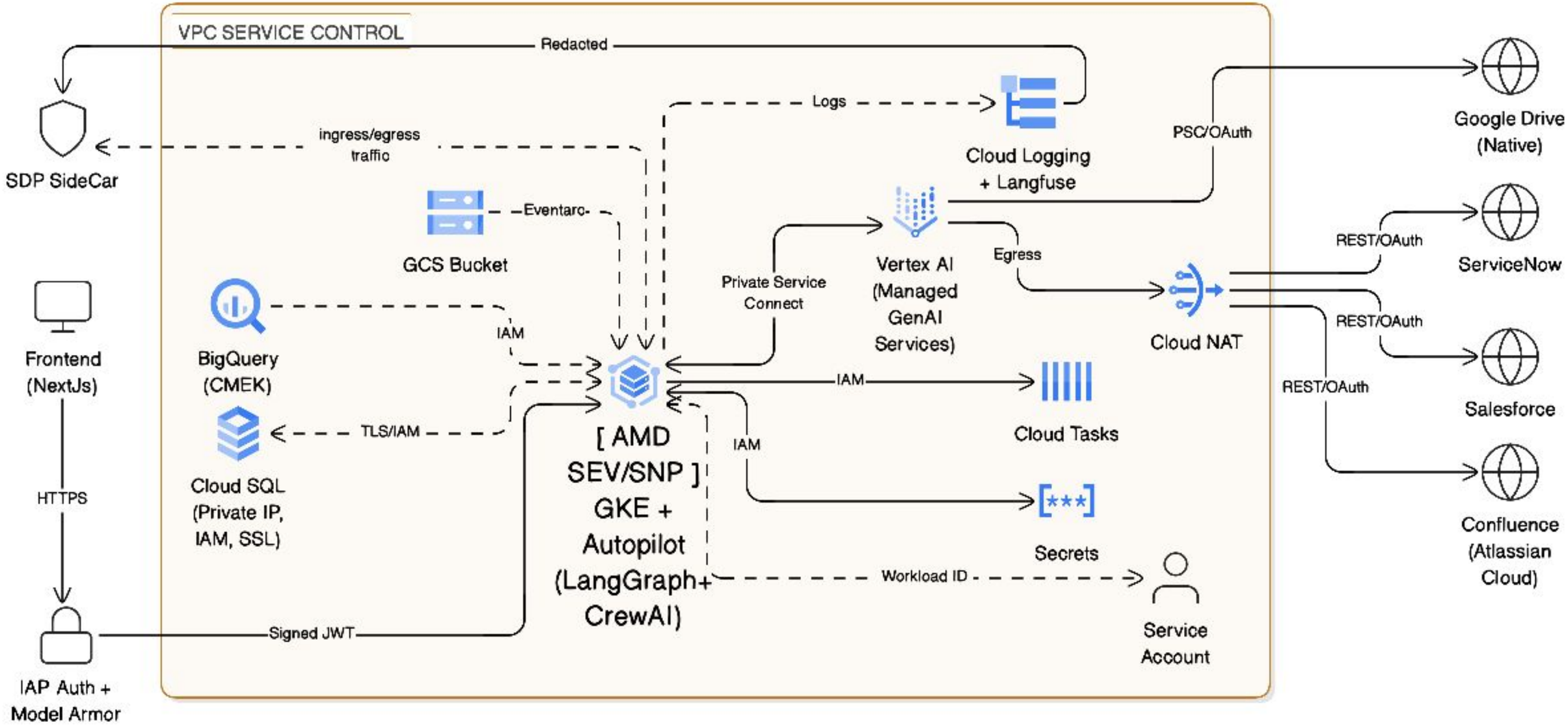
Logical Architecture (The "Virtual Team")

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



Physical Architecture

Zero-Trust Serverless Enclave for Enterprise GenAI (europe-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 services
- **Terraform**: Infrastructure as Code

Observability

Langfuse Monitoring traces every graph node state transition, exports logs to **Cloud Logging** and **Grafana** or (**Cloud Trace**) for production monitoring.

Governance

Sensitive Data Protection sidecar: (GCP managed service) for PII redaction at ingress and industry standard guard rails

Model Armor: Intent Validation, blocking malicious "jailbreak" commands or prompt injections before they reach the LLM orchestrator.

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

RAG Strategies

Multi-Modal Agentic RAG

Autonomous Multi-Source Intelligence: Orchestrating Hybrid Retrieval with Self-Correcting Logic and Enterprise-Grade Governance

Pattern A: Adaptive Semantic Search (Unstructured)

Vector Engine: Vertex AI RAG Engine using text-embedding-005 for high-dimensional accuracy.

Chunking Strategy: Layout-Aware & Semantic Chunking to preserve document hierarchy.

Reranking: Cross-Encoder Rerankernode to score the most relevant 3-5 passages, reducing "lost in the middle" hallucinations.

Pattern B: Verified Analytical Query (Structured)

Engine: LangChain Semantic Text-to-SQL feeding into BigQuery.

Governance: Mandatory Dry Run Validation via a Read-Only Service Account to ensure zero data-leakage.

Few-Shot examples: Uses a vector store of "Golden Query-SQL" pairs to guide the agent in generating high-accuracy queries for complex luxury inventory metrics.

Pattern C: Hybrid Agentic Orchestrator (Cross-Source)

Logic: LangGraph Supervisor decomposes complex multi-source requests (e.g., "Compare sales trends against last month's policy update").

Self-Correction Loop: If the initial retrieval is irrelevant, a Query Transformation node automatically rewrites the prompt and triggers a second search attempt.

Comparative Analysis

Advantages

Advantage	Agentic Platform Capability
Operational Efficiency	Agents offload L1/L2 support and operational cost
Reduced MTTR	Automated workflows vs manual processes
Control	100% EU Data Residency with Zero-Trust security in a serverless enclave.
Sovereignty	Full ownership of the "AI Brain," specialized logic, and system connectors.
Scalability	Horizontal, demand-driven scaling with usage-dependent serverless costing.

Planification

TCO & Roadmap

Build Phase

- CapEx: Estimated Cost: €480K→€660K
- 3-6 months, Requires 6-7 FTE Squad:
  - 1 Cloud Architect
  - 2 Senior AI Engineer
  - 1 Data Engineer
  - 1 DevSecOps / Security Architect
  - 1 UX/Product Lead
  - 1 Frontend Dev

Run Phase

- OpEx: Infrastructure €8,000 - €10,000 / month in pilot phase
- GCP Infrastructure + LLM Tokens for ~1k users

Operational Team

- Requires 2.5 FTE → (Platform Engineer + FinOps/Security Analyst) + 0.5 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 and secure.

Budget

Cost Offset

Development budget replaces recurring 1 year for most of L1/L2 support cost.

Value

Ownership

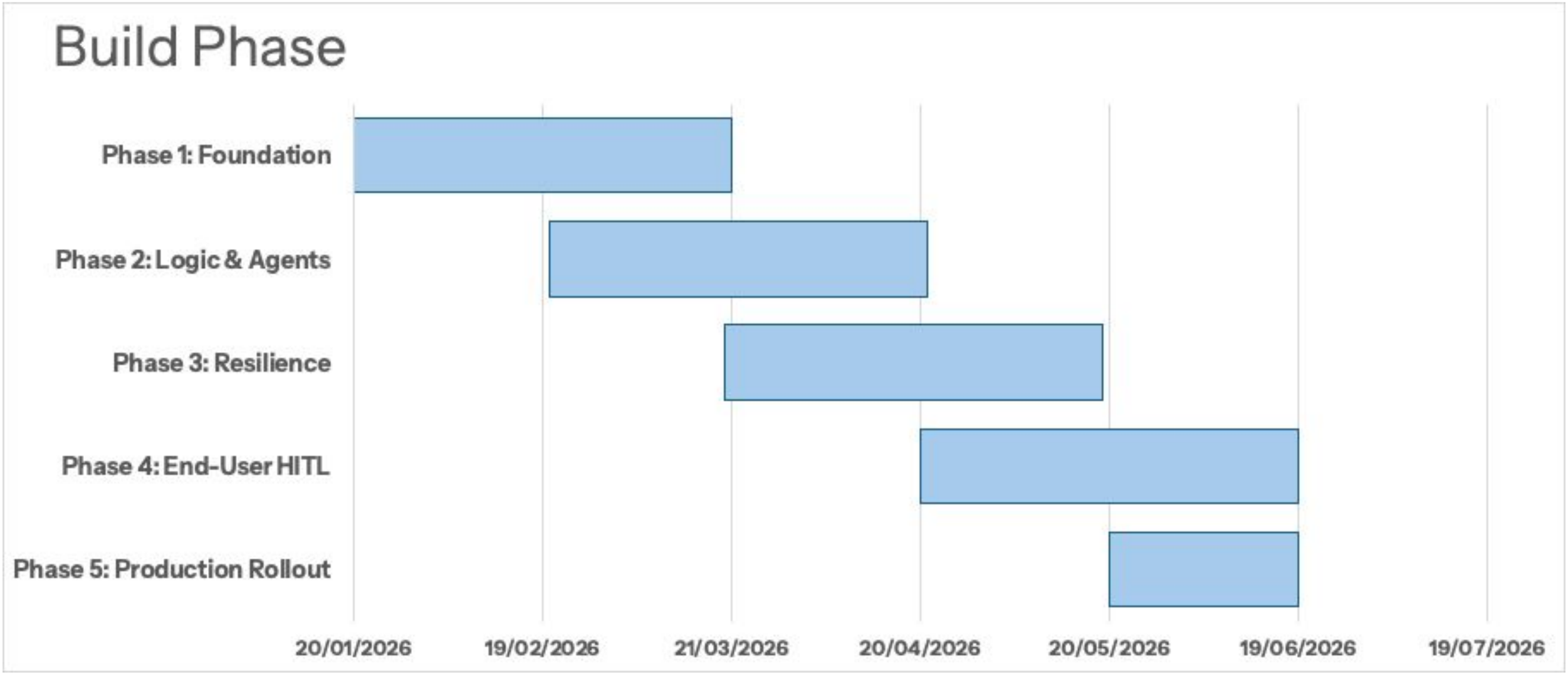
We own the "Brain" and the "Connectors" (no Black Box SaaS).

Merci

Q/A

Annex

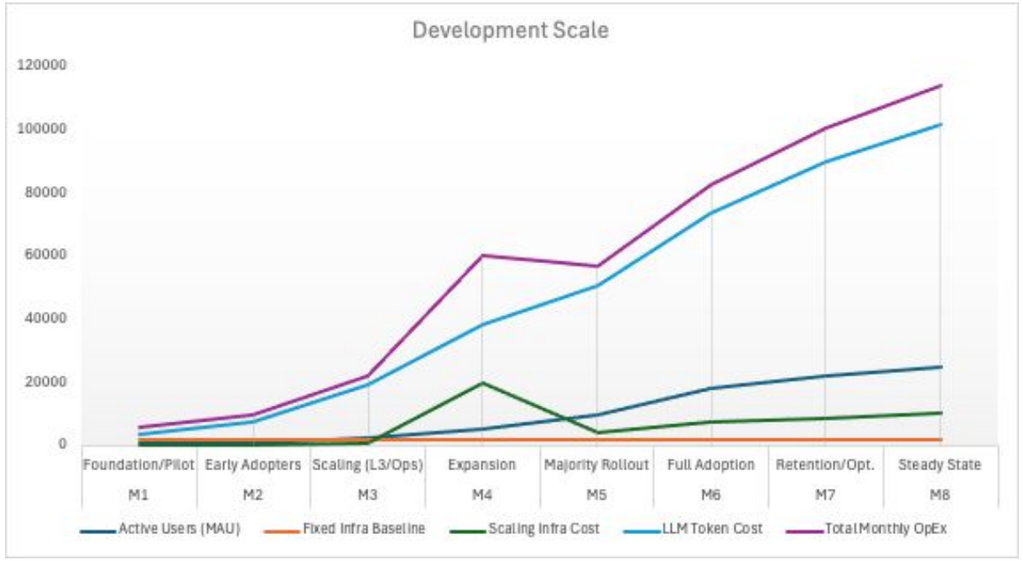
Build Phase


Phase 1: Foundation
Phase 2: Logic & Agents
Phase 3: Resilience
Phase 4: End-User HITL
Phase 5: Production Rollout
20/01/2026
19/02/2026
21/03/2026
20/04/2026
20/05/2026
19/06/2026
19/07/2026

Zero Trust

Phase	Timeline	Key Activities	Technical Focus
Foundation	Month 1-2	Setup VPC Service Controls & Confidential GKE Cluster.	**Security:** Establishing the Zero-Trust serverless enclave in europe-west-9
Logic & Agents	Month 2-3	Develop the LangGraph Supervisor and specialized CrewAI agents (IT, Ops, Analyst).	**Orchestration:** Implementing stateful multi-agent flows with FastAPI and LangGraph.
Resilience (Queue)	Month 3-4	Implement Cloud Tasks for long-running (3–5 min) actions and Transient RAG processing.	**Connects the Confidential GKE orchestrator to asynchronous background services for higher reliability**
End-User HITL	Month 4-5	Integrating the **"Interrupt" UI** to facilitate Human-in-the-Loop (HITL) approvals for high-risk actions.	**Governance:** Using IAP identity to validate high-risk actions before execution
Run Phase	Month 6+	Production rollout to 25k users with 100% EU residency.	**Scalability:** Managed by horizontal, demand-driven serverless scaling via Confidential GKE Autopilot.

Run Phase Scaling



Service Category	Estimated Cost	Core Purpose & Governance
Confidential Compute	**€800**	Management fee for **GKE Autopilot** and the minimum "warm" pods for the **LangGraph Supervisor**. Includes the premium for **RAM Encryption**.
Knowledge Search Index	**€600**	Continuous hosting of the **Vertex AI Search** vector store to ensure RAG data (Confluence/Drive) is instantly available.
Stateful Data Tier	**€250**	**Cloud SQL (HA)** instance for stateful multi-agent checkpointing, ensuring conversation continuity.
Security & Perimeter	**€250**	Fixed costs for **VPC Service Controls**, **IAP Authentication**, **Cloud NAT**, and **Secret Manager**.
Networking & LB	**€470**	Global External HTTP(S) LB with US-based Anycast IP and cross-region data transfer .
TOTAL BASELINE	**€2370**	**Minimum "Sovereign Tax" for a Zero-Trust Enclave.**

Month	Phase	Active Users (MAU)	Fixed Infra Baseline	Scaling Infra Cost	LLM Token Cost	Total Monthly OpEx
M1	Foundation/Pilot	500	€2 370,00	€200	€3 850,00	**€6 420,00**
M2	Early Adopters	1000	€2 370,00	€400	€7 700,00	**€10 470,00**
M3	Scaling (L3/Ops)	2500	€2 370,00	€1 000,00	€19 250,00	**€22 620,00**
M4	Expansion	5000	€2 370,00	€20 000,00	*€38 500,00*	**€60 870,00**
M5	Majority Rollout	10000	€2 370,00	€4 050,00	€51 000,00	**€57 420,00**
M6	Full Adoption	18000	€2 370,00	€7 250,00	€73 500,00	**€83 120,00**
M7	Retention/Opt.	22000	€2 370,00	€8 900,00	€89 800,00	**€101 070,00**
M8	Steady State	25000	€2 370,00	€10 100,00	€102 000,00	**€114 470,00**

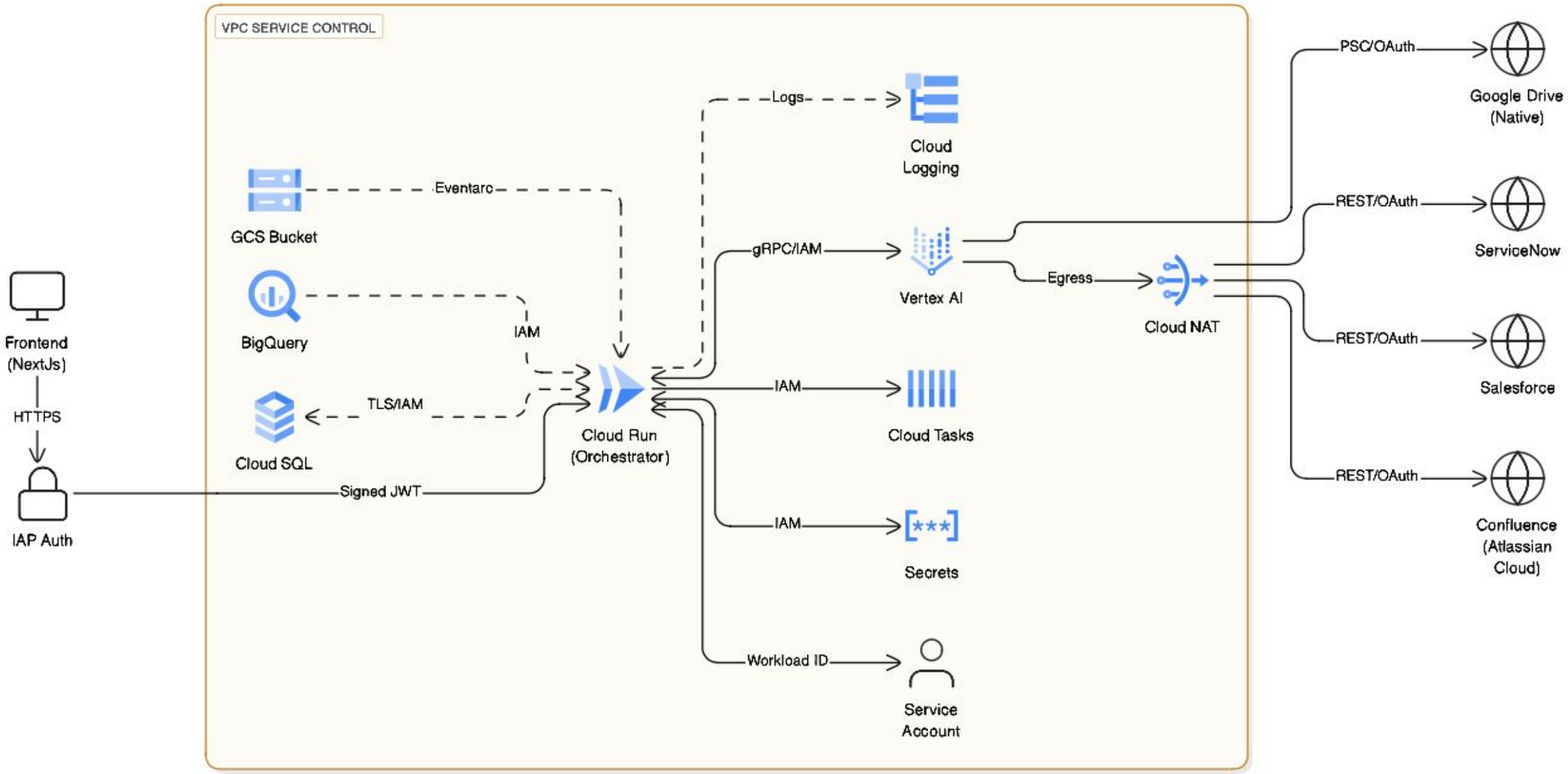
Operational Maturity (Day 2)

Capability	Implementation	Value
Observability	Distributed Tracing (Langfuse/opentelemtery)	User flagged response (*hallucination*) is traced back to the specific retrieved document or agent decision.
QA & Testing	"Golden Dataset" Regression + Continuous Benchmarking	CI/CD pipeline runs 100+ standard questions to ensure new code doesn't degrade answer quality.
Resilience	Circuit Breakers	Graceful degradation if a tool (e.g., ServiceNow) is slow or down.

Compute & Application Hosting	▪ Cloud Run (Services & Workers): ~€600 - €800 / month ▪ Cloud Run (Jobs - Connectors): ~€30 - €60 / month
Data & Storage	▪ Cloud SQL (PostgreSQL HA): ~€250 - €400 / month ▪ Vertex AI Search (Vector Store): ~€600 - €1,000 / month ▪ Cloud Storage (GCS): ~€50 - €100 / month ▪ Memorystore (Redis) - Optional: ~€40 - €150 / month
AI & Intelligence	▪ Vertex AI (Gemini Models): ~€200 - €500 / month ▪ Security & Networking (The "Enterprise" Premium) ▪ Cloud Armor + Load Balancer: ~€40 - €100 / month ▪ Cloud NAT: ~€50 / month ▪ Secret Manager & KMS: ~€50 - €80 / month
Observability & Operations	▪ Cloud Logging & Monitoring: ~€100 - €200 / month ▪ Dataplex (Data Quality): ~€100 - €200 / month

Physical Architecture

Contractual Trust Serverless Enclave for Enterprise GenAI (europe-west-9)



Analyse Comparative

Advantages*(Contractual)*

Advantage	Static Chatbot (Current)	Agentic Platform Capability	Architectural Solution
Operational Efficiency	High L1/L2 support costs (€1M+ annually) for 25k users.	Agents offload L1/L2 support and operational actions and their recurrent cost.	**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**(europe-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**

Contractual Trust

Phase	Timeline	Key Activities	Technical Focus
Foundation	Month 1-2	Setup GCP Landing Zone, VPC Service Controls, and IAP Auth)	**Security:** Establishing the Zero-Trust serverless enclave in europe-west-9
Logic & Agents	Month 2-3	Develop the LangGraph Supervisor and specialized CrewAI agents (IT, Ops, Analyst).	**Orchestration:** Implementing stateful multi-agent flows with FastAPI and LangGraph.
Resilience (Queue)	Month 3-4	Implement Cloud Tasks for long-running (3–5 min) actions and Transient RAG processing.	Connects the Cloud Run orchestrator to asynchronous background services for higher reliability
End-User HITL	Month 4-5	Integrating the **"Interrupt" UI** to facilitate Human-in-the-Loop (HITL) approvals for high-risk actions.	**Governance:** Using IAP identity to validate high-risk actions before execution
Run Phase	Month 6+	Production rollout to 25k users with 100% EU residency.	**Scalability:** Managed by horizontal, demand-driven serverless scaling via Cloud Run.

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

Service Heavy vs Asset Heavy

Composant GCP	Équivalent Kubernetes Natif / Souverain	Fonction Clé
VPC Service Control	**Cilium + Network Policies**	Isolation logique et segmentation L7.
IAP + Access Approval	**Keycloak + OAuth2 Proxy + Teleport**	Authentification forte et accès conditionnel.
Cloud NAT	**Cilium / Istio Egress Gateway**	Contrôle et traçabilité des flux sortants.
GKE Autopilot	**RKE2 ou Talos Linux**	Orchestration durcie et immuable.
Confidential GKE (SEV)	**Confidential Containers (CoCo)**	Chiffrement de la RAM au niveau matériel.