Rapport de Gestion de Projet

PULS-EVENTS

Transformation du POC en MVP Scalable : Architecture, Roadmap, Coûts et Risques pour une IA générative robuste.

STRATÉGIE D'INDUSTRIALISATION 2026

Analyse Stratégique

Contexte & Objectifs du MVP

LIMITES DU POC ACTUEL

Fragilité en production : Une "démo qui marche" mais manque de robustesse.

Absence de mémoire : Pas d'historique conversationnel entre les sessions.

Contexte géo limité : Manque de précision géographique pour les recommandations.

Observabilité faible : Difficulté à monitorer la latence et la qualité des réponses.

OBJECTIFS D'INDUSTRIALISATION

Scalabilité : Architecture AWS robuste basée sur ECS et RDS.

Personnalisation : Intégration d'une mémoire long-terme et des préférences.

Pertinence : Recherche hybride (RAG) couplée à la géolocalisation.

Fiabilité : Mise en place de Guardrails et monitoring temps réel.

Segmentation Marché

Cibles & Cas d'Usage

L'Explorateur

"Que faire ce soir près de moi ?"

Besoin de géolocalisation précise et de fraîcheur des données pour des décisions immédiates.

Le Planificateur

"Organiser mon week-end complet."

Besoin de filtres complexes (budget, dates) et de suggestions d'itinéraires logiques.

Le Passionné

"Tout savoir sur le Jazz à Paris."

Exigence forte sur la pertinence sémantique et la profondeur du catalogue thématique.

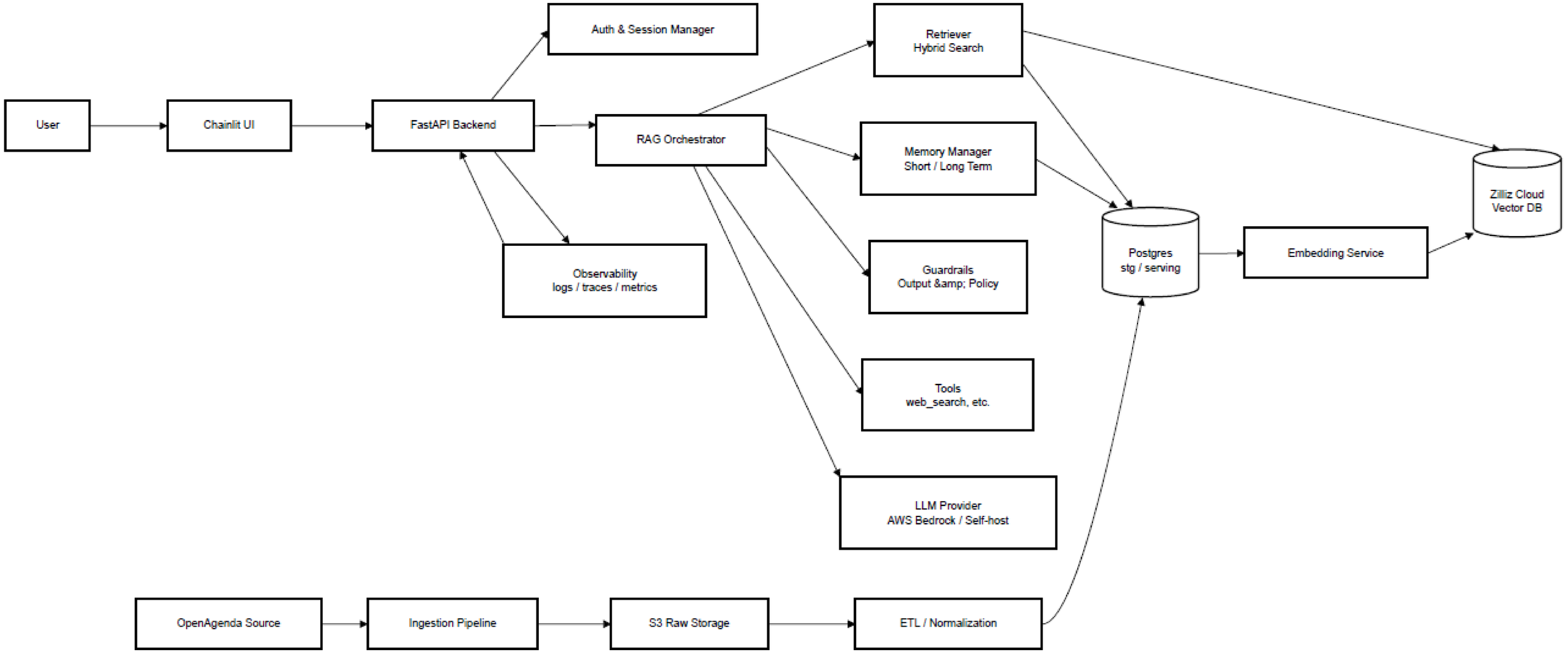
Architecture Technique (Détails)

Infrastructure AWS

- **Compute:** ECS Fargate pour l'API Backend, l'Orchestrateur RAG et les Workers d'ingestion.
- **Sécurité:** VPC Endpoints (PrivateLink) pour sécuriser le trafic vers Bedrock.
- **Stockage:** S3 pour le Data Lake, RDS (PostgreSQL) pour les données relationnelles.

Stack IA & RAG

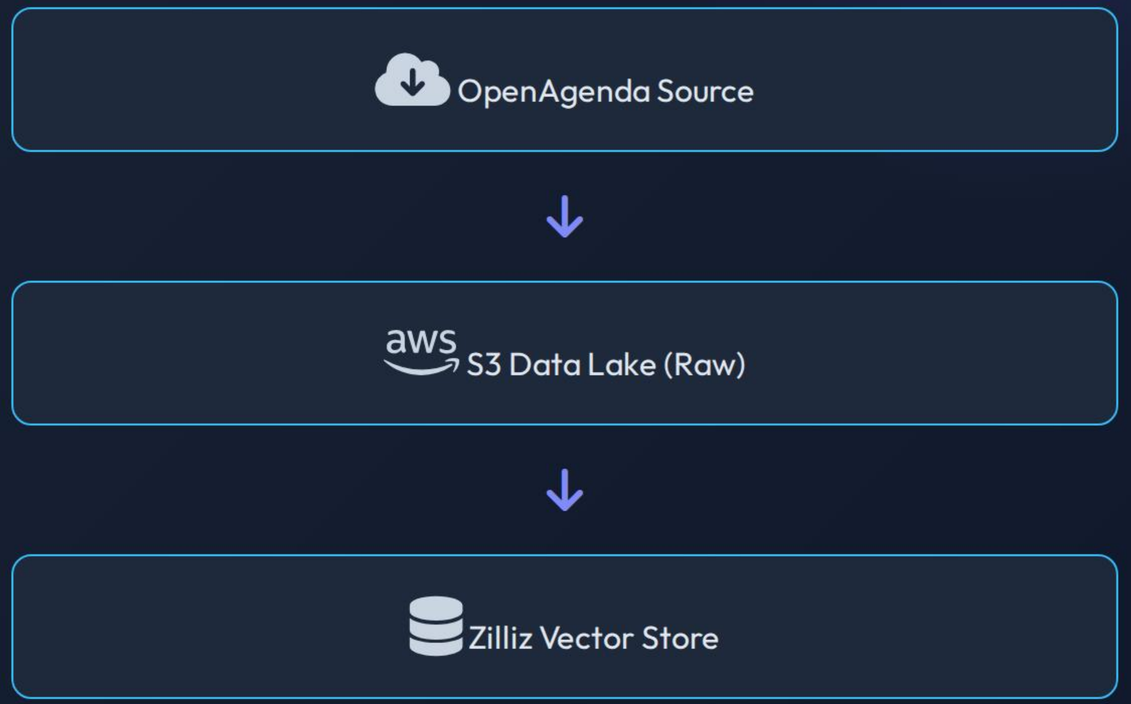
- **LLM:** Amazon Bedrock (modèle managé) pour la gouvernance et les métriques natives.
- **Vector DB:** Zilliz Cloud pour l'indexation et la recherche vectorielle scalable.
- **Cache:** Redis (ElastiCache) pour la mémoire court-terme et la performance.


User
Chainlit UI
FastAPI Backend
Auth & Session Manager
RAG Orchestrator
Retriever
Hybrid Search
Memory Manager
Short / Long Term
Guardrails
Output & Policy
Tools
web_search, etc.
LLM Provider
AWS Bedrock / Self-host
Observability
logs / traces / metrics
Postgres
stg / serving
Embedding Service
Zilliz Cloud
Vector DB
OpenAgenda Source
Ingestion Pipeline
S3 Raw Storage
ETL / Normalization

Pipeline de Données

De OpenAgenda à Zilliz

- **Ingestion (Raw):** Collecte des flux OpenAgenda vers S3 (datalake).
- **Normalisation (Staging):** Nettoyage, dédoublonnage des lieux (Venues) et formatage dates.
- **Vectorisation:** Création des embeddings et indexation dans Zilliz Cloud avec métadonnées.
- **Temps Réel:** Mises à jour fréquentes pour garantir la fraîcheur des événements.



Valeur Ajoutée

Fonctionnalités Clés du MVP

Mémoire & Personnalisation

SHORT-TERM (REDIS)

Gestion du contexte de la session active pour une fluidité conversationnelle immédiate.

LONG-TERM (POSTGRESQL)

Profil utilisateur consolidé, historique des interactions et préférences durables.

Intelligence Géographique

POSTGIS ENGINE

Calculs de distance précis et filtrage par rayon pour des recommandations locales.

RANKING HYBRIDE

Score final combinant pertinence sémantique (IA) et proximité géographique réelle.

Fiabilité & Sécurité

GUARDRAILS

Vérification factuelle systématique et citations obligatoires des sources d'événements.

MONITORING SODA

Plan de Projet (Synthèse)

Déploiement sur 12 semaines via méthodologie hybride agile.

S1-S3

Fondations Infra
& Ingestion Data

S4-S7

Core RAG
Embeddings & Chat UI

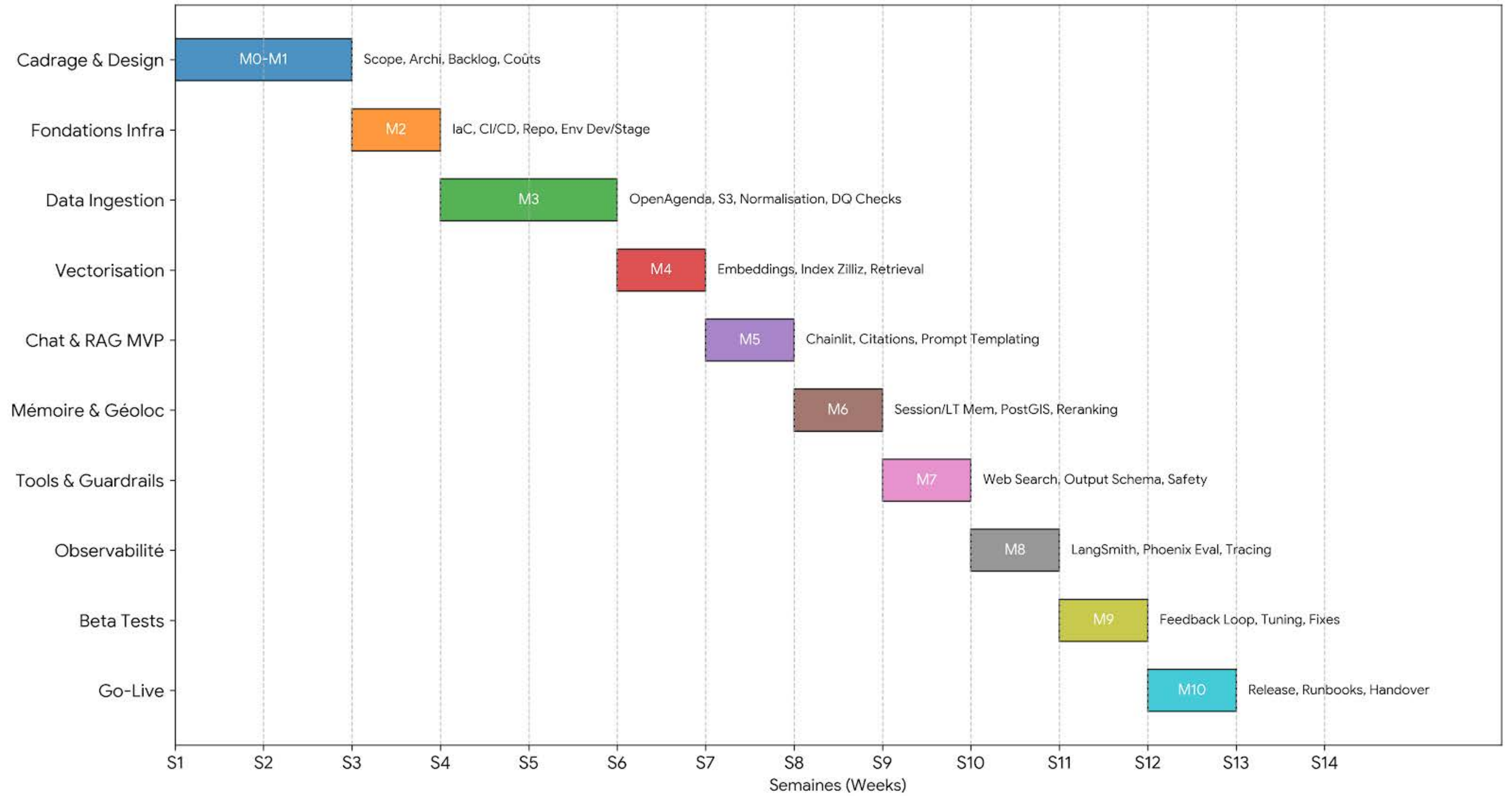
S8-S9

Intelligence
Mémoire & Géo

S10-S12

Go-Live
Observabilité & Beta

Diagramme de Gantt - Projet RAG Hybride Agile



Budget & Investissement

Estimation Financière (Build)

Budget Total Estimé

75k€

Design & Architecture	7 500 €
Développement (Ingestion + RAG)	27 000 €
Intelligence (Mémoire & Géo)	7 500 €
Provision pour Risques (15%)	9 795 €

La provision de 15% est calculée pour absorber les incertitudes liées à l'intégration des modèles LLM et à la scalabilité de l'infrastructure.

Trajectoire Financière

Coûts d'Exploitation (OPEX)

Poste de Coût	S1 (MVP / Faible)	S2 (Croissance)	S3 (Scale)
Utilisateurs / Jour	200	2 000	20 000
LLM (Bedrock Tokens)	~216 $	~4 320 $	~54 000 $
Infra (ECS, RDS, Redis)	~236 $	~763 $	~3 022 $
Vector DB (Zilliz)	~200 $	~800 $	~2 500 $
TOTAL ESTIMÉ	**~722 $ / mois**	**~6 163 $ / mois**	**~60 422 $ / mois**

Analyse des Coûts

Le coût des tokens LLM est le facteur principal d'augmentation. Stratégies de mitigation prévues : Model Routing (utilisation de modèles plus petits pour les tâches simples) et Semantic Caching.

Gouvernance & Qualité

Gestion des Risques et Qualité

RISQUES IDENTIFIÉS

HALLUCINATIONS

Réponses inventées ou hors contexte par le LLM, impactant la crédibilité.

LATENCE CRITIQUE

Temps de réponse supérieur à 3 secondes dégradant l'expérience utilisateur.

QUALITÉ DES DONNÉES

Informations sur les événements obsolètes ou incomplètes provenant des sources.

DÉRIVE DES COÛTS

Explosion de la facture liée à la consommation imprévue de tokens LLM.

STRATÉGIES DE MITIGATION

GUARDRAILS & CITATIONS

Vérification factuelle systématique et obligation de citer les sources officielles.

SEMANTIC CACHING

Optimisation de la latence et réduction des coûts via la mise en cache des requêtes.

MONITORING SODA

Contrôles automatiques de la fraîcheur et de l'intégrité des données en amont.

GOUVERNANCE FINOPS

Alerting budgétaire en temps réel et quotas d'utilisation par segment utilisateur.

Exécution & Engagement

Prochaines Étapes

Action 01

Validation Immédiate

Décision Go/No-Go sur le budget global (75k€) et validation finale du plan d'architecture technique.

Action 02

Lancement Infra

Mise en place de l'environnement AWS via Terraform (IaC) et initialisation du repository Git centralisé.

Action 03

Recrutement Team

Constitution de la Feature Team dédiée : 1 Data Engineer, 1 Backend Developer et 1 DevOps.

Transformer la vision Puls-Events en réalité opérationnelle.